Approches robustes pour la comparaison d'images et la reconnaissance d'objets

Julien Rabin

Thèse

présentée pour obtenir le grade de docteur de Télécom ParisTech

Spécialité : Traitement du Signal et des Images

Julien Rabin

Approches robustes pour la comparaison d'images et la reconnaissance d'objets

Soutenue le 9 décembre 2009 devant le jury composé de :

Rapporteurs	Renaud Keriven	École des Ponts ParisTech
	Michael Lindenbaum	Technion - Israel Institute of Technology
	Jean-Michel Morel	École Normale Supérieure de Cachan
Examinateurs	Henri Maître	Télécom ParisTech
	Lionel Moisan	Université Paris Descartes
	Patrick Pérez	INRIA, IRISA
Directeurs de thèse	Julie Delon	CNRS, Télécom ParisTech
	Yann Gousseau	Télécom ParisTech

À mes parents,

À Sophie

Remerciements

Je tiens à exprimer ici toute l'estime et l'admiration que je porte à Julie Delon et Yann Gousseau pour leurs qualités à la fois pédagogiques et scientifiques, mais avant tout humaines. C'est en effet avec un immense plaisir que j'ai travaillé sous leur direction durant mon stage de master puis mon doctorat. Profitant de leur infaillible soutien, j'ai pu découvrir le monde de la recherche et apprendre tout ce que je connais en traitement des images. Je les remercie pour leur gentillesse, leur optimisme, leur disponibilité, ainsi que leur patience parfois, ce qui m'a permis de réaliser avec sérénité ce doctorat.

Je remercie les membres du jury, Renaud Keriven, Michael Lindenbaum, Henri Maître, Lionel Moisan, Jean-Michel Morel et Patrick Pérez, d'avoir accepté d'évaluer mon travail et de l'avoir enrichi par leurs conseils et leurs remarques.

Je souhaite également remercier l'ensemble des personnes qui m'ont fait bénéficier de leur expertise, en particulier Gabriel Peyré, Pascal Monasse (qui en outre a activement participé à la relecture de ce manuscrit), et Lionel Moisan, avec qui j'ai eu la chance de pouvoir collaborer sur l'un des chapitres de ce manuscrit. Je remercie par ailleurs Saïd Ladjal et Isabelle Bloch pour leur disponibilité lorsque j'avais besoin d'aide.

Je remercie Patricia et Sophie-Charlotte pour leur gentillesse et leur efficacité à résoudre les problèmes administratifs et informatiques.

J'en profite pour saluer les anciens du C07 : Bin, Marie et Mihai, Xavier, Yvan, ainsi que les nouveaux : Adrian, Gui-Song, Julien, Payam et Vincent. Je n'oublie pas les thésards du C032 : Alex, Aymen, Gabrielle, Charles-Alban, Thibault, Vincent et du C029 : Camille, Goeffroy, Jean-Baptiste, Jérémy, Olivier. Je regretterai tout particulièrement les discussions avec Vincent, Gabrielle et Thomas.

Je souhaite enfin exprimer ma gratitude envers mes proches qui m'ont toujours encouragé, en particulier ma compagne, Sophie, qui m'a souvent conseillé, aidé et soutenu pendant ce doctorat. À vous tous, sans qui rien de tout cela n'aurait été possible, merci !

Résumé

La problématique générale de cette thèse est la comparaison d'images, que nous traitons à la fois de manière locale et globale *via* différentes applications.

Nous nous sommes principalement intéressés au problème de la reconnaissance d'objets entre différentes images à partir de descripteurs locaux. Nous proposons un système complet, robuste et automatique de reconnaissance d'objets multiples, dont la mise en œuvre repose principalement sur deux approches méthodologiques : la théorie de la décision *a contrario* [Desolneux *et al.* 2000] et la théorie du transport optimal de Monge-Kantorovich. Dans ce cadre, une mesure de dissimilarité est définie pour la comparaison de descripteurs de type SIFT [Lowe 1999] en fonction du coût de transport optimal entre histogrammes circulaires et unidimensionnels (Circular Earth Mover's Distance). Un critère de mise en correspondance de descripteurs locaux s'appuyant sur la théorie de la décision *a contrario* est par la suite introduit. Ce critère permet de s'affranchir du réglage du seuil de détection et de la restriction usuelle au plus proche voisin, autorisant les mises en correspondances multiples entre images. Toujours dans le cadre méthodologique *a contrario*, nous proposons un algorithme de type RANSAC (RANdom SAmple Consensus [Fischler et Bolles 1981]) pour le groupement de correspondances de descripteurs locaux. Cet algorithme, reposant sur une généralisation de l'approche introduite par [Moisan et Stival 2004], permet de détecter des groupes multiples. L'approche proposée permet également la sélection du modèle géométrique de la transformation rigide due au changement de point de vue et/ou au mouvement de l'objet détecté entre les différentes images.

Dans le cadre de la théorie du transport optimal, nous étudions par ailleurs l'intérêt de l'EMD (Earth Mover's Distance [Rubner 1998]) pour la comparaison globale d'images (indexation d'images). Nous proposons enfin une méthode de régularisation de la carte de transport s'inspirant des approches par filtrage non-local, en vue d'une application au transfert de caractéristiques entre images par spécification d'histogramme. Nous démontrons l'intérêt de cette méthode pour les applications de changement de contraste et de transfert de couleurs.

Abstract

The general topic of this dissertation is image comparison, which is treated both as a local and a global issue *via* various applications.

We mainly consider the object recognition task, based on local descriptors. We propose a complete, robust and automatic system for multiple object recognition, which relies on two methodological approaches : the *a contrario* detection theory [Desolneux *et al.* 2000] and the Monge-Kantorovich optimal mass transport theory. In the latter framework, a dissimilarity measure is introduced for the comparison of SIFT-like descriptors [Lowe 1999] relying on the optimal transportation cost between circular and one-dimensional histograms (Circular Earth Mover's Distance). A matching criterion of local descriptors based on the *a contrario* methodology is then introduced. This criterion, which does not require any detection threshold setting nor any usual nearest-neighbor restriction, enables multiple correspondences between images. Moreover, we propose an algorithm based on the RANSAC strategy (RANdom SAmple Consensus [Fischler and Bolles 1981]) for the grouping of local matches. This algorithm, generalizing the *a contrario* approach introduced by [Moisan and Stival 2004], handles multiple group detection. It also enables us to automatically select the best geometrical model describing the object pose modification, which can be due for instance to the viewpoint change between several images.

In addition, in the context of optimal mass transportation theory, we study the interest of the EMD (Earth Mover's Distance [Rubner 1998]) for global comparison of images (image retrieval). Eventually, a regularization approach for the transportation map is proposed, inspired from non-local filters, in the context of characteristics transfer between images by histogram specification. We show the interest of this approach for contrast modification and color transfer applications.

Table des matières

Introduction

La problématique générale de cette thèse est la comparaison d'images, que nous traitons à la fois de manière locale et globale. Dans un premier temps, nous abordons le problème de la reconnaissance d'objets à partir de descripteurs locaux. Nous proposons un système complet, robuste et automatique de reconnaissance d'objets multiples, dont la mise en œuvre repose principalement sur des algorithmes et des critères de décision qui sont présentés dans la première partie de cette thèse. Ce système utilise également une mesure de dissimilarité permettant la comparaison de descripteurs locaux qui est proposée dans la seconde partie de ce manuscrit, consacrée au transport optimal entre histogrammes. Dans ce cadre d'étude, nous étudions par ailleurs l'intérêt du transport pour la comparaison globale d'images, en vue d'une application à l'indexation d'images. Nous proposons enfin une méthode de régularisation du transport optimal dont nous démontrons l'intérêt pour le transfert de caractéristiques entre images.

Reconnaissance d'objets

La reconnaissance d'objets consiste à identifier les objets communs à plusieurs images. De nombreuses applications sont concernées par cette problématique, comme la construction de mosaïques d'images ou la recherche dans une base. Étant donnée une image représentant un objet requête, la reconnaissance de cet objet dans une nouvelle image nécessite plusieurs étapes : la *description* des attributs visuels des images analysées, la *comparaison* de ces attributs, la *détection* de la présence ou non de l'objet, et enfin l'*estimation* de la pose de cet objet dans l'image. C'est un problème difficile en raison de la présence éventuelle d'une occultation partielle de l'objet recherché, d'un changement de point de vue, d'un changement d'éclairage de l'objet, ou encore de « fouillis » (*clutter* en anglais). En plus de ces contraintes, nous nous sommes également intéressés au problème de la reconnaissance *automatique* d'objets *multiples*, qui consiste à reconnaître plusieurs objets différents ou bien les occurrences multiples d'un même objet, sans avoir à régler différents paramètres de détection.

Le cadre de travail le plus propice pour la reconnaissance d'objets est la représentation locale des images, où chaque image analysée est représentée par un ensemble de descripteurs locaux. Une approche classique consiste alors à mettre en correspondance les descripteurs locaux de plusieurs images. Ensuite, une étape de groupement de ces correspondances permet d'identifier la position de l'objet recherché. Cette approche requiert généralement le réglage de nombreux paramètres de détection, qui s'avèrent critiques en termes de performances. Pourtant, peu de méthodes ont été proposées pour rendre ce réglage automatique et pour permettre la reconnaissance de plusieurs objets à la fois. De plus, la comparaison des descripteurs locaux repose généralement sur une mesure de dissimilarité qui ne prend pas en compte les différentes perturbations auxquelles ils sont soumis (changement de point de vue et quantification par exemple).

Nous proposons dans cette thèse un système automatique et complet de reconnaissance d'objets utilisant une représentation locale des images. Nous utilisons des descripteurs de type SIFT [Low04] (Scale Invariant Feature Transform), qui sont connus pour leur grande robustesse. La mise en œuvre de ces descripteurs est détaillée en annexe B. Notre système repose sur trois contributions pour répondre aux différents objectifs précédemment définis :

- ▷ Nous proposons une nouvelle mesure de dissimilarité pour les descripteurs de type SIFT. Cette mesure est définie dans le cadre du transport optimal qui est étudié dans la seconde partie de cette thèse.

- ▷ Un critère de mise en correspondance fondé sur la méthodologie de détection *a contrario*, introduite par Desolneux, Moisan et Morel [DMM08], est défini pour la sélection robuste des correspondances locales entre les images. Cette méthode nous permet de sélectionner des correspondances multiples entre deux images, tout en contrôlant l'espérance du nombre de fausses détections.

- ▷ Un algorithme de groupement des correspondances de type RANSAC [FB81] (*RANdom SAmple Consensus*) est finalement mis en œuvre pour la reconnaissance d'objets multiples. Cet algorithme repose sur l'approche proposée par Moisan et Stival [MS04] dans le cadre de la géométrie épipolaire. Nous avons adapté cette approche à la problématique de la reconnaissance d'objets, de manière à détecter de manière robuste plusieurs objets distincts selon la géométrie plane. Cet algorithme nous permet également d'aborder le problème de la sélection de modèles géométriques pour l'estimation de la pose d'un objet.

Transport optimal entre histogrammes

L'autre problématique à laquelle nous nous sommes intéressés dans cette thèse est la comparaison des images *via* la mesure de la dissimilarité entre leurs histogrammes de caractéristiques. Nous traitons ce problème dans le cadre du transport optimal, introduit par Gaspard Monge dans son mémoire sur la théorie des déblais et des remblais [Mon81]. Il étudie l'optimisation du coût de transport lié à la répartition de tas de terre dans un ensemble de trous. Cette notion de transport est de nos jours très largement exploitée pour la comparaison d'images, notamment depuis les travaux de [RTG00]. Elle consiste à utiliser une mesure de dissimilarité entre histogrammes appelée *Earth Mover's Distance* (ou EMD).

Comme nous l'avons précédemment annoncé, nous avons été amenés à considérer le transport optimal entre histogrammes pour la comparaison robuste de descripteurs locaux de type SIFT. Ces descripteurs sont composés d'histogrammes d'orientation, qui ont la particularité d'être *circulaires*. Dans la seconde partie du manuscrit, nous étudions le coût du transport optimal entre des histogrammes unidimensionnels dans le cas circulaire. Une nouvelle mesure de dissimilarité pour les descripteurs de type SIFT est ensuite proposée dans ce cadre spécifique.

Nous proposons également dans cette thèse une analyse comparative générale de la distance de transport EMD avec les distances usuelles qualifiées de « bin-à-bin », telles que les distances L^p. Ce terme désigne les distances entre histogrammes reposant sur une stricte comparaison des cellules de quantification (ou bins) ayant la même position sur une grille. Nous illustrons ensuite, par une application à l'indexation d'images, les avantages et les inconvénients de l'EMD par rapport aux distances bin-à-bin.

En plus d'un coût de transport, le flot du transport optimal définit un transfert entre deux histogrammes. Ce transfert peut être utilisé pour le transfert de caractéristiques entre images, dont nous avons étudié deux applications récemment proposées dans [Del04] pour l'égalisation conjointe des couleurs de plusieurs images (histogramme mi-chemin) et dans [PKD07] pour le transfert de palette de couleurs. Une méthode de régularisation du transport est proposée pour réduire les problèmes liés à l'utilisation du transport.

Organisation du manuscrit

Le manuscrit est organisé en deux parties. La première est consacrée aux étapes de décision d'un système de reconnaissance d'objets.

Nous donnons dans le chapitre 1 un rapide aperçu des méthodes de représentation locale des images, puis nous présentons un état de l'art des méthodes de mise en correspondance de descripteurs locaux. Dans le chapitre 2 est introduit un nouveau critère de mise en correspondance qui s'inscrit dans le cadre théorique de la détection *a contrario*. Une validation expérimentale sur une base de 3 millions de descripteurs est utilisée pour démontrer l'intérêt de notre approche.

La problématique de la reconnaissance d'objets à partir de correspondances locales est présentée dans le chapitre 3. Nous réalisons ensuite un état de l'art des approches de groupement de correspondances, suivi par un état de l'art sur les critères de sélection de modèles géométriques pour l'estimation de la pose d'un objet. Un algorithme de groupement multiple de correspondances, désigné par l'acronyme MAC-RANSAC (Multiple A Contrario RANdom SAmple Consensus), est défini au chapitre 4. Une analyse expérimentale est ensuite proposée pour évaluer les différentes contributions, reposant à la fois sur des images synthétiques et réelles.

La seconde partie du manuscrit est dédiée au transport entre histogrammes, dont la problématique est introduite au chapitre 5.

Dans le chapitre 6 nous étudions le transport optimal entre des d'histogrammes circulaires et unidimensionnels. L'intérêt de la distance obtenue, appelée CEMD (Circular Earth Mover's Distance), est par la suite analysé comparativement à la distance bin-à-bin L^1. Les conclusions de cette analyse sont ensuite illustrées pour une application d'indexation d'images couleurs sur différentes bases.

Une nouvelle mesure de dissimilarité reposant sur la distance CEMD est proposée pour la comparaison des descripteurs SIFT au chapitre 7. Ses performances sont ensuite évaluées expérimentalement sur une base de données pour la reconnaissance d'objets, comparativement à d'autres distances proposées dans la littérature.

Le chapitre 8 est consacré au transfert de caractéristiques entre deux images. Nous étudions en particulier deux applications utilisant transport optimal entre histogrammes : l'égalisation de contraste et le transfert de palette de couleurs. Après en avoir décrit le principe, nous illustrons les limitations de ces méthodes. Une nouvelle approche de régularisation du transport est ensuite proposée. L'intérêt de cette méthode, appelée *Non Local Map Regularization*, est illustrée par de nombreux exemples.

En annexe A, nous rappelons le principe générique de la théorie de la détection *a contrario* qui est exploitée à plusieurs reprises dans notre système de reconnaissance d'objets.

Une mise en œuvre des descripteurs SIFT est proposée en annexe B.

Les modèles géométriques de transformations rigides utilisés pour la reconnaissance d'objets sont rappelés au chapitre C.

Première partie

Une méthode automatique de reconnaissance d'objets

Chapitre 1

État de l'art sur la reconnaissance d'objets par descripteurs locaux

Nous nous intéressons, dans la première partie de cette thèse, aux différentes outils mis en œuvre dans le cadre de la reconnaissance d'objets. Dans ce chapitre préliminaire, nous introduisons la problématique générale de la reconnaissance d'objets, puis nous donnons un bref aperçu des principales méthodes de représentation locale des images. Un état de l'art sur les différentes approches de mise en correspondance de descripteurs locaux est ensuite présenté.

1.1 Problématique

La reconnaissance d'objets consiste à *identifier* dans une image un objet dont on possède une ou plusieurs vues, puis à *estimer* sa pose. Dans l'exemple donné en figure 1.1, il s'agit de localiser le bâtiment de l'image 1.1(a) dans une seconde photographie (figure 1.1(b)).

(a) Où se trouve ce bâtiment ...

(b) ... dans cette image ?

FIG. 1.1 – *Où se trouve l'objet recherché ?*

De très nombreuses applications sont concernées par la problématique de la reconnaissance d'objets : reconstruction 3D [Bar07, LPK07] (*structure from motion*), recherche dans une base d'images [CPS+07, PSZ08] ou dans une vidéo [SZ06] (*object retrieval*), recalage d'images [YSST07] (*image registration*) et création de mosaïques d'images [BL07] (*Image Stitching*) notamment.

On distingue généralement ces applications de reconnaissance d'objets de celles du domaine de la *détection d'objets*. Contrairement à la reconnaissance, la détection d'objets consiste à identifier dans une image *la classe sémantique* des objets qu'elle contient. L'atteinte d'un tel objectif suppose un appren-

tissage des caractéristiques propres à une catégorie d'objets, ce qui nécessite une base d'apprentissage générique. Les applications les plus courantes visent à détecter des catégories d'objets tels que les piétons, les voitures, les visages [VJ02], ou encore à définir la thématique de l'image [BZM08].

Contraintes Les applications de reconnaissance d'objets soulèvent deux problèmes majeurs. Le premier problème concerne la *robustesse* de la détection et de l'estimation : le système doit être capable de localiser l'objet recherché en étant invariant à certains phénomènes, en particulier le changement d'éclairement, les variations de la pose 3D de l'objet ou encore l'occultation partielle de l'objet. Le second problème posé par la reconnaissance d'objets est celui de la *prise de décision* : il s'agit de valider ou non la présence de l'objet dans l'image analysée.

La première catégorie de méthodes à avoir été proposée pour la reconnaissance d'objets est celle des approches dites « globales ». Il s'agit de déterminer une transformation globale permettant de recaler l'image requête représentant l'objet et l'image analysée. La méthode la plus simple consiste alors à regarder la corrélation entre ces deux images [JD01] . Cependant, ces approches appelées *template matching* ne sont pas robustes à de nombreux phénomènes, tels que l'occultation et le mouvement 3D de l'objet par exemple. Par la suite ont été proposées des approches dites « locales », c'est-à-dire fondées sur la description partielle des images analysées. L'idée est de détecter indépendamment plusieurs sous-parties de l'objet afin d'accroître la robustesse de la détection. Cette stratégie a été rendue possible par le développement de représentations locales robustes, dont la plus populaire est sans doute la représentation SIFT (Scale Invariant Feature Transform) [Low04].

Mise en œuvre de la reconnaissance d'objets par une approche locale Le schéma classique de reconnaissance d'objets utilisant une représentation locale des images est le suivant :

Détection et codage de structures Cette première étape vise à représenter une image à partir de descripteurs locaux. Il faut tout d'abord détecter des structures saillantes dans l'image analysée (points, régions, ou contours), puis les coder. Nous rappelons dans la section suivante (§ 1.2) quelles sont les principales méthodes qui ont été proposées pour accomplir cette tâche. Dans cette thèse, nous utilisons des descripteurs de type SIFT, dont le principe et la mise en œuvre sont détaillés en annexe B.

Comparaison des descripteurs et mise en correspondance À ce stade, chaque image est représentée par un ensemble de descripteurs locaux. La reconnaissance de l'objet requête requiert une mise en correspondance de ses descripteurs locaux avec les descripteurs extraits de l'image analysée. Pour cela, une mesure de dissimilarité entre les descripteurs doit être définie. Afin d'améliorer la comparaison des descripteurs SIFT, nous proposons au chapitre 7 une nouvelle mesure de dissimilarité inspirée de la théorie sur le transport optimal.

Un critère de décision se basant sur cette mesure est ensuite utilisé pour sélectionner les correspondances les plus fiables. Les différents critères utilisés en reconnaissance d'objets sont étudiés en section 1.3. Un nouveau critère de correspondance est introduit dans le chapitre 2, se basant sur la théorie de la décision *a contrario* dont le principe est rappelé en annexe A.

Groupement des correspondances et estimation de la pose L'étape précédente de mise en correspondance peut être vue comme le moyen de réduire drastiquement le nombre d'hypothèses sur la position de l'objet dans l'image analysée. Une étape de groupement est ensuite nécessaire pour valider la détection de l'objet recherché et pour estimer sa pose. Un état de l'art lié à cette problématique est réalisé au chapitre 3. Un nouvel algorithme de groupement de correspondance est ensuite introduit dans le chapitre 4.

Notons que d'autres stratégies de reconnaissance d'objets se basant sur une représentation géométrique ont été proposées, à l'instar de l'algorithme *geometric hashing* [LW88], ou du groupement de lignes [HU90].

1.2 Représentation locale des images

Dans cette section nous revenons brièvement sur la notion de représentation locale des images, afin de donner un aperçu des méthodes ayant été proposées. Cette étape de représentation d'une image se déroule en deux temps : une phase de détection de structures d'intérêt, puis une phase de codage de ces structures.

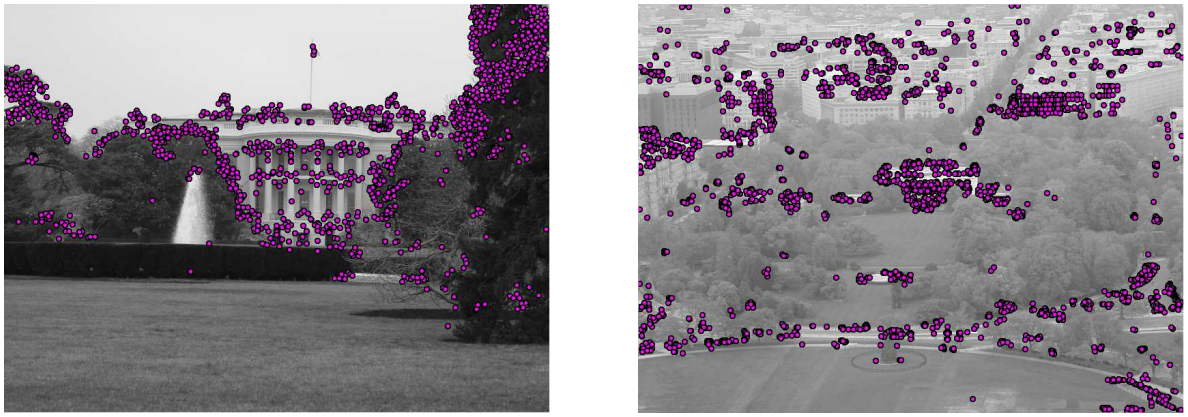
1.2.1 Détection de structures d'intérêt

Une première phase de détection vise à détecter les structures géométriques les plus importantes de l'image, de manière à restreindre la quantité d'information à coder. Différents types de structures peuvent ainsi être détectées : points saillants (coins, jonctions en T, *etc.*), régions, texture, ou encore contours.

De très nombreuses approches ont été proposées de manière à améliorer la robustesse et la répétabilité de la détection de points saillants. L'une des premières approches à avoir été largement utilisée est le détecteur de coins de Harris [HS88], invariant à l'orientation de la structure détectée. Les travaux de Lindeberg sur la représentation en espace-échelle linéaire des images lui ont ensuite permis de définir une famille de détecteurs de structures (*blob features*) invariantes au changement d'échelle [Lin98]. Les détecteurs de coins multi-échelle qui ont été introduits par la suite (Harris-Laplace [MS01] et DoG-Hessian [Low04] notamment) s'inspirent de cette approche. En définissant un point d'intérêt comme un extremum local de la représentation en espace-échelle, ces approches permettent d'attribuer à ce point une échelle caractéristique. Des approches *a contrario* ont été définies pour la détection de coins [Cao04] et de jonctions en T [Bél06], invariantes au changement de contraste.

D'autres approches s'appuient sur la détection de formes, à l'instar de [MSC+06, HGCS08] où des morceaux de lignes de niveau contrastées sont extraits.

Enfin, une autre stratégie consiste à détecter des régions d'intérêt, à l'image des MSER [MCUP02] (Maximally Stable Extremal Region), ou encore du détecteur proposé dans [MS02] (*affine covariant region detector*). Il est également possible de combiner plusieurs types de détecteurs, de façon à obtenir une description plus riche de l'image analysée (voir par exemple [MTS+05]).



(a) Extraction de points d'intérêt de l'image requête contenant l'objet recherché

(b) Extraction de points d'intérêt de l'image à analyser

FIG. 1.2 – *Illustration de l'extraction de points d'intérêt.*

À l'issue de cette première étape, seules quelques structures dites « d'intérêt » sont détectées dans chacune des images. La figure 1.2 montre les points d'intérêt détectés dans la paire d'images donnée en exemple dans la première section. Pour être capable de comparer ces points d'intérêt entre les deux images, une deuxième phase consiste à représenter ces points à l'aide de descripteurs compacts. Nous allons maintenant rappeler quelques méthodes de représentation du voisinage des points d'intérêt (ou des contours) détectés.

1.2.2 Représentation par descripteurs locaux

Une solution simple, mais peu robuste, consiste à extraire un patch centré sur le point d'intérêt. Deux méthodologies sont employées pour améliorer la robustesse de la description locale des points d'intérêt.

Une première approche consiste à calculer des coefficients invariants à des transformations géométriques et aux changements d'éclairement. Dans [SM97], Schmid et Mohr définissent des moments invariants par changement d'orientation et d'échelle, calculés au voisinage de points d'intérêt. Dans [Bau00], Baumberg propose une approche similaire pour définir des descripteurs invariants par transformation affine. Des descripteurs de formes fondés sur le calcul de moments sont également utilisés dans [MG98, MCUP02].

Une autre solution consiste à normaliser le voisinage du point considéré selon ses caractéristiques géométriques (échelle caractéristique par exemple). Sur ce principe, D. Lowe propose dans [Low99] une représentation locale des images appelée SIFT. Le descripteur SIFT est composé d'histogrammes d'orientation du gradient. Ces histogrammes sont estimés à partir de régions distinctes du voisinage normalisé et centré de chaque point d'intérêt considéré. Il a été montré dans [MS05] que ce type de descripteur est très robuste à différents phénomènes (bruit, compression JPEG, changement d'éclairement, rotation et changement d'échelle). Une analyse détaillée des descripteurs SIFTs est proposée dans [MY08]. C'est ce type de descripteur que nous utilisons dans cette thèse pour la reconnaissance d'objets. Une étude de leur mise en œuvre est détaillée dans l'annexe B. De nombreuses extensions de ce descripteur ont été proposées, dont une variante utilisant les MSER dans [FL07]. En particulier, Morel et Yu ont défini un descripteur invariant affine appelé ASIFT dans [MY09].

On trouve également dans la littérature d'autres représentations locales des images, dont le descripteur de formes de [MSC+06], les *Local Binary Pattern* [Pie05], ou encore les groupes de contours adjacents [FFJS08]. Soulignons également les travaux sur la détection [AM97] et sur la représentation [Sur07] des coins en espace échelle non linéaire (Affine Morphological Scale Space).

Une fois que les descripteurs locaux sont extraits de chacune des images à traiter, la reconnaissance d'objets consiste à étudier les correspondances possibles entre ces descripteurs. Dans la section suivante, nous étudions les différentes stratégies adoptées dans la littérature pour ce processus de mise en correspondance.

1.3 Mise en correspondance de descripteurs locaux

1.3.1 Notations et définition

Soit A l'image requête pour laquelle on a détecté N_Q points (ou régions) d'intérêt. On suppose que la base de recherche (constituée d'une seule image ou bien d'une base de données d'images), notée B, contient N_C points d'intérêt obtenus avec la même procédure de détection. Chaque point d'intérêt, qui possède éventuellement des attributs locaux (orientation principale ou échelle caractéristique par exemple), est représenté par un descripteur de son voisinage qui est normalisé selon les caractéristiques géométriques locales du point considéré. On note $\{a^i \ , \ i = 1, \ldots, N_Q\}$ et $\{b^j \ , \ j = 1, \ldots, N_C\}$ les ensembles de descripteurs de A et de B respectivement.

Le processus de mise en correspondance peut être considéré comme un classifieur, permettant de classer en deux catégories l'ensemble des correspondances possibles $N_Q \times N_C$: celles qui sont validées et considérées comme correctes (« positives »), et celles rejetées car considérées comme incorrectes (« négatives »). Pour évaluer la pertinence de cette classification, on utilise une « vérité-terrain » qualifiant de « vrai » un appariement de deux points d'intérêt qui représentent physiquement un même point sur un objet, tous les autres appariements étant qualifiés de « faux ». Cette classification est le résultat d'un classifieur dit « idéal ». Pour quantifier les performances d'un classifieur réel, on distingue quatre catégories de résultats (tableau 1.1) : les « vrai-positifs » et « vrai-négatifs » qui sont des appariements correctement classés, et les appariements « faux-positifs » et « faux-négatifs » qui sont des estimations incorrectes du classifieur réel. Selon l'application, la performance d'un classifieur est ainsi analysée en

fonction du score de ces différentes catégories. Nous reviendrons sur ces différentes notions dans la partie expérimentale (§ 2.2) du chapitre suivant .

TAB. 1.1 – *Taxinomie des différentes types de correspondances.*

Vérité \ Estimation	Positive	Négative
Correspondance Correcte	Vraie Positive (vp)	Fausse Négative (fn)
Correspondance Incorrecte	Fausse Positive (fp)	Vraie Négative (vn)

1.3.2 Principe

Ainsi que nous l'avons brièvement évoqué dans la problématique, le procédé de mise en correspondance de descripteurs locaux en vue de la reconnaissance d'un objet est crucial. Il s'agit à partir d'un ensemble de N_Q descripteurs requêtes de trouver des correspondances avec des descripteurs similaires dans une base de taille N_C, ainsi que l'illustre la figure 1.3. Les éventuelles correspondances ainsi obtenues permettent d'établir une similarité locale entre une région de l'image requête et une autre image de la base. Cette étape peut donc être vue comme un moyen de réduire drastiquement les hypothèses sur la présence de la requête dans la base de descripteurs : de l'ensemble des $N_Q \times N_C$ mises en correspondances possibles, on obtient typiquement à l'issue de ce processus $O(N_Q)$ correspondances (c'est-à-dire de l'ordre de grandeur de N_Q). On retrouve ce principe très général dans d'autres applications que le traitement d'images (traitement de la parole et analyse de document notamment).

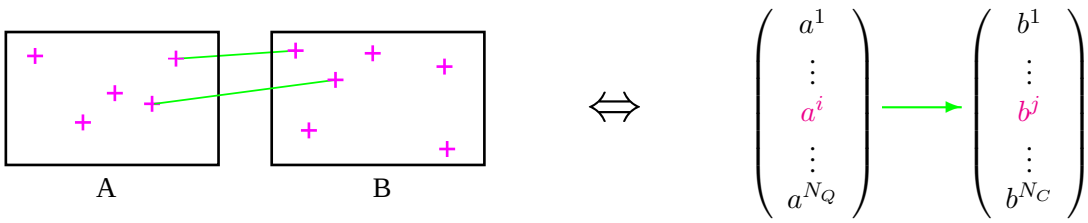


FIG. 1.3 – *A représente une image requête représentée à l'aide de N_Q descripteurs requêtes, et B une base de descripteurs contenant N_C candidats. Ce problème de mise en correspondance de points d'intérêt revient en pratique à apparier des descripteurs locaux.*

Baumberg insiste dans [Bau00] sur l'importance du contrôle des fausses détections (*mismatches*) lors de la procédure de mise en correspondance :

> *Robust statistical methods such as RANSAC [FB81] can be used to tolerate a significant fraction of mismatched features. However there is an associated computational cost that grows with the number of mismatched features.*

En effet, nous avons vu qu'une étape de post-traitement est généralement utilisée pour agglomérer les hypothèses qui ont été sélectionnées afin d'estimer la pose de l'objet recherché. Cette opération est le plus souvent réalisée à l'aide d'un algorithme de type RANSAC [FB81] ou de la transformée de Hough [Hou59]. Il s'agit de déterminer les mises en correspondances qui peuvent être expliquées par une même transformation géométrique entre deux images. Un état de l'art sur cette dernière phase de la reconnaissance d'objets est effectué au chapitre 3.

Le principe de mise en correspondance est illustré en figure 1.4 : l'utilisation d'un critère robuste permet d'apparier les points d'intérêt de l'objet recherché, tout en évitant de sélectionner un trop grand nombre de fausses correspondances. L'identification de ces couples de points d'intérêt rend alors possible l'estimation de la pose de l'objet dans la scène.

Perona et Moreels dans [MP07] résument la procédure de mise en correspondance en ces termes :

> *Important aspects of matching are **metrics** and **criteria** to decide whether two features should be associated, and **data structures and algorithms** for matching efficiently.*

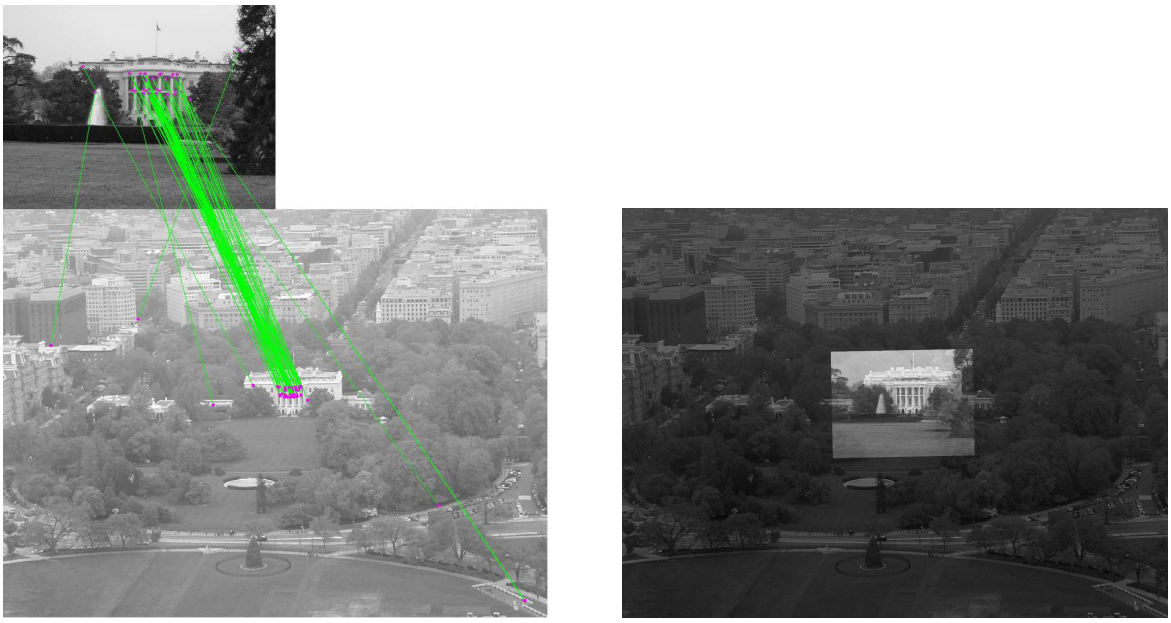


FIG. 1.4 – ***Illustration de l'intérêt du critère de mise en correspondance.*** *La sélection des correspondances de points d'intérêt les plus similaires permet de réduire considérablement le nombre d'hypothèses de départ. Ceci permet ensuite de mettre en œuvre une estimation robuste de la position de l'objet.*

Cela signifie que la mise en place d'un tel processus est critique à plusieurs titres :

Mesure de dissimilarité La mise en correspondance de points d'intérêt repose grandement sur la comparaison de leurs descripteurs selon une mesure de dissimilarité. Cette dernière permet en effet d'identifier pour une requête quels sont les éléments les plus similaires dans la base de descripteurs, parmi lesquels certains seront par la suite sélectionnés. Sa définition est primordiale car elle conditionne les performances du critère de mise en correspondance.

Critère de mise en correspondance Le critère de décision doit permettre la validation des correspondances les plus similaires au sens de la mesure précédemment définie, tout en assurant le rejet des autres correspondances. En ce sens, il peut être considéré comme un ensemble de classifieurs spécifiques à chaque requête. Cela signifie, entre autres, que chaque classifieur doit être idéalement optimal par rapport au descripteur requête pour lequel il est défini. La seconde difficulté réside dans le compromis classique entre limitation des fausses correspondances (faux-positifs) et maximisation des bonnes correspondances (vrai-positifs). Comme nous le verrons par la suite, quel que soit le critère choisi, il existe un paramètre global permettant de régler le seuil de validation. Si ce seuil est trop permissif, on risque d'obtenir trop de fausses correspondances. Pour les algorithmes de détection utilisés en amont, il sera donc très difficile de reconnaître l'objet recherché et d'estimer sa pose de manière précise, en raison du « bruit » créé par ces fausses correspondances. Au contraire, avec un seuil de validation trop strict, le nombre insuffisant de correspondances validées rend impossible la reconnaissance de l'objet recherché. Nous rappelons en section 1.3.3 quels sont les critères usuels de mise en correspondance utilisés en reconnaissance d'objets.

Stratégie de recherche Mis à part quelques cas particuliers qui seront détaillés ultérieurement (section 1.3.4), on ne dispose d'aucune d'information *a priori* sur les mises en correspondance que l'on souhaite valider, de sorte qu'aucune tentative de correspondance ne peut être rejetée sans avoir été au préalable examinée. Pour cette raison, nous avons considéré dans cette thèse la mise en correspondance de descripteurs locaux par comparaison *exhaustive*. Nous verrons en section 1.3.5, que définir une stratégie de mise en correspondance qui permette d'éviter une comparaison exhaustive des descripteurs requiert l'utilisation d'outils de recherche *approchée*, propres à la fouille de données (apprentissage non supervisé).

Bien qu'étant l'une des étapes les plus décisives de la chaîne de détection en terme de robustesse et de coût calculatoire, la question de la mise en correspondance n'a pourtant pas fait l'objet de beaucoup d'études. Dans cette thèse, nous nous sommes plus particulièrement intéressé aux deux premiers aspects de la mise en correspondance, que sont la définition d'une mesure de dissimilarité et d'un critère de mise en correspondances. Le chapitre 7, dans la seconde partie de ce manuscrit, est consacré à la définition d'une nouvelle mesure de dissimilarité pour les descripteurs de type SIFT. Cette mesure repose sur le coût de transport optimal sur le cercle, qui est auparavant étudié au chapitre 6. Une étude comparative est réalisée afin d'évaluer les performances obtenues selon la distance utilisée entre descripteurs. Par conséquent, nous considérons dans la suite du présent chapitre la question du choix de la mesure comme réglée. Dans les sections suivantes, nous verrons quelles sont les méthodes qui ont été proposées dans la littérature pour les deux autres aspects de la mise en correspondance.Les critères génériques de mise en correspondance de descripteurs locaux sont présentés en section 1.3.3. Nous proposons dans le prochain chapitre un nouveau critère, reposant sur la théorie de la détection *a contrario*, que nous comparons à ces différents critères. En section 1.3.4, nous présentons quelques cas particuliers de mise en correspondance utilisant des contraintes géométriques. La dernière section concerne les méthodes de comparaisons approchées, que nous n'avons pas considéré dans cette thèse.

1.3.3 Critères génériques de mise en correspondance

Dans un cadre général, on ne dispose d'aucune connaissance *a priori* sur les points d'intérêt ou leur descripteurs. La solution naturellement privilégiée est alors de prendre une décision reposant sur une comparaison exhaustive de l'ensemble des mises en correspondance (voir par exemple [JT08, FTG06, Low04, CM05, DZLF94, MCUP02, RLSP07, Bau00, KP06, BMP02, ZK06a]). Dans un premier temps, pour chaque descripteur requête a^i, les descripteurs $\{b^j \ , \ j = 1, \dots, N_C\}$ de la base sont donc tous examinés à l'aide d'une mesure de dissimilarité. On note dorénavant $D(a, b)$ la mesure de dissimilarité entre les descripteurs a et b. À l'issue de cette étape de comparaison, une **mesure de qualité** $Q(a^i, b^j)$ est associée à l'appariement des descripteurs a^i et b^j, fonction de l'ensemble des distances $\{D(a^i, b^j)\ , \ 1 \leq i \leq N_Q \ 1 \leq j \leq N_C\}$. Un **critère de validation** des mises en correspondance est ensuite utilisé afin de valider les appariements en fonction de leur mesure de qualité.

À notre connaissance, seuls deux critères de validation ont été proposés dans la littérature, tous deux reposant sur l'utilisation d'un seuil de détection global, fixé par l'utilisateur, de la mesure de qualité Q.

Seuil de validation sur la distance La mesure de qualité la plus naturelle pour une correspondance (a, b) est donnée par sa mesure de dissimilarité $Q(a, b) = D(a, b)$. Le critère de validation d'appariement repose ainsi sur le seuillage de la mesure de dissimilarité. On se référera à présent à l'acronyme DT pour désigner ce type de critère (Distance Threshold).

Définition 1 (Critère DT) *Pour chaque requête a^i, l'ensemble des descripteurs candidats ayant une mesure de dissimilarité plus petite que le seuil global t sont validés. On obtient l'ensemble des appariements suivant :*

$$\mathcal{C}_{DT} := \left\{(a^i, b^j)\ , \ i \in \{1, \dots, N_Q\} \ \textit{et} \ j \in \{1, \dots, N_C\} \ : \ D(a^i, b^j) \leq t\right\}$$

Comme nous aurons l'occasion de le constater expérimentalement dans le chapitre suivant (§ 2.2), ce premier critère est limité par deux inconvénients majeurs. Le premier est le **choix du seuil global** sur la distance pour toutes les requêtes. En effet, le seuil optimal qui permet de ne sélectionner que des appariements corrects (ou « vrai-positifs ») varie selon le type de descripteur requête et la base de descripteurs considérés. Cela signifie qu'il est impossible en pratique de définir un seuil satisfaisant pour l'ensemble des requêtes, mais également que les performances pour un même seuil varient d'une paire d'images à une autre. Une des conséquences les plus graves étant que l'utilisateur est alors obligé d'intervenir pour ajuster ce paramètre (calibration manuelle). La seconde limitation de cette méthode est liée au fait que le critère DT ne restreint pas le nombre de mise en correspondance. Les distributions de la mesure de dissimilarité étant différentes d'une requête à une autre, certains descripteurs vont être mis en correspondance

avec plusieurs descripteurs de la base : on qualifie de **mise en correspondance multiples** de tels appariements. En pratique, on observe que le nombre de mises en correspondance multiples augmente très fortement avec le seuil de détection t. Cependant, une large proportion de ces correspondances multiples sont des fausses correspondances (ou « faux-positifs »). Or, comme nous l'avons déjà évoqué, les algorithmes de détections en amont de la chaîne de traitement requièrent un certain nombre d'appariements corrects pour permettre la détection d'un objet (ainsi que l'estimation de sa pose par exemple), mais également un bon « rapport signal à bruit » (c'est-à-dire une proportion de correspondances correctes élevée par rapport à l'ensemble des correspondances validées).

Afin de maximiser le nombre de correspondances correctes tout en essayant de préserver le rapport signal à bruit, une solution pratique consiste à *restreindre la mise en correspondance de chaque requête à son plus proche voisin* dans la base de descripteurs. On désigne par NN (nearest neighbor) la restriction d'un critère au plus proche voisin. On trouve de nombreux exemples dans la littérature de l'utilisation d'une telle restriction (notamment [DZLF94, Bau00, JT08]).

Définition 2 (Critère NN-DT) *Pour chaque requête a^i, le descripteur candidat le plus proche est validé si sa mesure de dissimilarité est plus petite que le seuil global t. On obtient l'ensemble des appariements suivant :*

$$\mathcal{C}_{NN\text{-}DT} := \left\{ (a^i, b^{J(i)}) \ , \ i \in \{1, \ldots, N_Q\} \ : \ D(a^i, b^{J(i)}) \leq t \ \textit{avec} \ J(i) = \underset{j \in \{1,\ldots,N_C\}}{\arg\min} \ D(a^i, b^j) \right\}$$

Si en pratique cette restriction au plus proche voisin présente l'avantage de réduire considérablement le nombre de fausses correspondances, ce choix limite également le nombre de bonnes correspondances. En premier lieu, quand un objet est présent plusieurs fois dans la base d'images, la restriction au plus proche voisin limite de toute évidence le nombre de bonnes détections. Par ailleurs, la mesure de dissimilarité et le descripteur utilisés n'étant pas parfaits, il arrive en pratique que le "bon" candidat (selon la vérité-terrain) ne soit pas le plus proche voisin, mais le deuxième ou troisième seulement (par exemple, lorsque l'on compare deux photographies d'un objet selon des points de vue et des conditions photométriques très différents), ce qui limite d'autant plus le nombre de bonnes détections. C'est la raison pour laquelle un compromis différent entre quantité de bonnes détections et taux de fausses détections a été choisi pour certaines applications, où ce sont les k plus proches voisins qui sont potentiellement sélectionnés. À titre d'exemple, cette variante de NN-DT pour la mise en correspondance de descripteurs locaux est utilisée dans les travaux suivants, avec : $k = 3$ pour une mise en correspondance dense [FTG06], $k = 4$ pour la construction de panorama ([BL07]), et k entre 5 et 10 pour la mise en correspondance de séquences vidéo dans [RLSP07].

Néanmoins, la principale limitation des critères DT et NN-DT reste le problème du réglage du seuil global t qui est le même pour tous les appariements testés. Idéalement, un seuil optimal devrait être défini pour chaque requête en prenant en compte la structure représentée. Pour contourner la difficulté de définir de tels seuils adaptatifs, [Low04] a proposé une mesure de qualité alternative que nous allons décrire dans le paragraphe suivant.

> **Remarque 1 :**
> Notons qu'un autre type de variante, le critère de validation croisée, a été proposé [DZLF94] dans le contexte du recalage des images stéréoscopiques non calibrées (puis notamment repris dans [Sch96] et [Bau00]). Les auteurs de [DZLF94] définissent une mesure de qualité symétrique, ce qui a pour effet de restreindre les mises en correspondances aux appariements de descripteurs (a, b) pour lesquels on s'assure que b est le plus proche voisin de a, mais également que a est le plus proche voisin de b.

Seuil de validation sur le rapport des distances aux deux plus proches voisins Afin de permettre la définition d'un seuil global qui soit plus pertinent, Lowe propose une mesure de qualité alternative pour le critère de mise en correspondance au plus proche voisin NN-DT. Cette mesure est cette fois fondée sur la comparaison de la requête a^i à la fois au plus proche voisin b^{J1}, et au second plus

proche voisin b^{J2} dans la base de descripteurs. Elle consiste à regarder le rapport des distances de a^i avec respectivement b^{J1} et b^{J2}, soit :

$$Q(a^i, b^{J1}) = \frac{D(a^i, b^{J1})}{D(a^i, b^{J2})} \leq 1 \ .$$

Cette définition repose sur le constat suivant : lorsqu'une requête n'a pas de véritable correspondant dans la base de descripteurs, les distances $D(a^i, b^{J1})$ et $D(a^i, b^{J2})$ ont de grandes chances d'être très proches, et leur rapport d'être plus proche de 1. Au contraire, le candidat le plus proche b^{J1} est d'autant plus singulier dans la base de descripteurs que le rapport des distances est proche de 0. Cette nouvelle mesure de qualité permet de définir le critère de mise en correspondance suivant, auquel on se réfère désormais en utilisant l'acronyme **NN-DR** (Nearest Neighbor Distance Ratio) :

Définition 3 (Critère NN-DR) *Pour une requête a^i, on note b^{J1} son plus proche voisin et b^{J2} son second plus proche voisin dans la base de descripteurs. La correspondance de a^i avec b^{J1} est validée si le rapport de la distance $D(a^i, b^{J1(i)})$ sur la distance $D(a^i, b^{J2(i)})$ est plus petite que le seuil global r. On obtient ainsi l'ensemble des appariements suivant :*

$$\mathcal{C}_{\textit{NN-DR}} := \left\{ (a^i, b^{J1(i)}) \ , \ i \in \{1, \ldots, N_Q\} \ : \ \frac{D(a^i, b^{J1(i)})}{D(a^i, b^{J2(i)})} \leq r \ , \textit{ avec} \right.$$
$$\left. J1(i) = \underset{j \in \{1,\ldots,N_C\}}{\arg\min} \ D(a^i, b^j) \textit{ et } J2(i) = \underset{j \in \{1,\ldots,N_C\} \setminus J1(i)}{\arg\min} \ D(a^i, b^j) \right\}$$

Comparé au critère NN-DT, qui repose sur un seuillage des distances, le critère de mise en correspondance NN-DR présente deux avantages considérables. Tout d'abord, le choix du seuil de sélection du critère NN-DR est plus intuitif qu'un seuil sur les distances, car la quantité $Q(a^i, b^{J1})$ mesure à quel degré le plus proche voisin dans la base est unique. Ensuite, ce critère se révèle expérimentalement beaucoup plus robuste que le critère NN-DT lorsque l'objet recherché est présent exactement une seule fois dans la base d'images, comme cela a été démontré par [MP07] sur une grande base. C'est la raison pour laquelle ce critère est très largement mis en pratique pour la mise en correspondance de descripteurs locaux (par exemple [SSS08, CM05, MP07]), ou encore la recherche d'image (*content based image retrieval*).

Le critère NN-DR n'en garde pas moins certains défauts qui limitent ses performances. D'une part, au même titre que le critère NN-DT, seul le plus proche voisin peut être apparié avec le descripteur requête. Le seuil de détection requiert également d'être fixé par l'utilisateur selon les expériences réalisées, comme le montrent les différents seuils de détection utilisés dans la littérature selon le type d'image et l'application considérée : $r = 0.8$ dans [Low04], $r = 0.6$ dans [SSS08], $r = 0.95$ dans [CM05], ou r entre 0.56 et 0.7 dans [MP07] par exemple. Par ailleurs, ainsi que cela a été reporté dans [MS05], la mesure de qualité utilisée par le critère NN-DR entraîne le rejet à tort des appariements avec des éléments répétés de la base de descripteurs. Or, il existe de nombreuses situations pratiques où ce type de répétition se produit :

- quand un objet possède des *structures répétitives*, comme c'est le cas des objets texturés, ou encore des objets manufacturés qui présentent souvent de nombreuses auto-similarités. C'est particulièrement le cas des photographies de bâtiments dont le recalage pose problème avec le critère NN-DR [ZK06a] ;
- lorsqu'un même objet est présent plusieurs fois dans la base d'images (voir par exemple [DMA07]) ;
- en raison de la redondance du détecteur de points d'intérêt, qui entraîne la duplication de descripteurs dans la base (voir la section B.1.4 en annexe).

Notons qu'une variante de ce critère a été proposée par [BSW05] pour limiter ce phénomène dans le cadre de la construction de mosaïques d'images (panorama). Dans ce contexte, plusieurs images que l'on souhaite recaler représentent partiellement la même scène et par conséquent le critère NN-DR est tenu en échec. Plutôt que d'utiliser le critère NN-DT, les auteurs de cet article ont modifié la mesure de qualité du critère NN-DR, en calculant le rapport de la distance au plus proche voisin sur la distance *moyenne* au second plus proche voisin dans l'ensemble des images.

Dans le but de surmonter les limitations des critères précédemment décrits (restriction au plus proche voisin, seuils non adaptatifs, robustesse du critère), nous allons introduire au prochain chapitre un nouveau critère générique de mise en correspondance de descripteurs locaux.

Afin de compléter cet état de l'art, nous allons auparavant présenter d'autres stratégies d'appariement dans les deux paragraphes suivants. Ces méthodes permettent de prendre en compte les contraintes spécifiques de certaines applications, ou d'exploiter une structure de la base de descripteurs, inférée à partir d'un processus d'apprentissage non supervisé.

1.3.4 Utilisation de contraintes géométriques

Dans certains cas, le problème de la mise en correspondance est soumis à des contraintes qu'il convient de prendre en compte afin de rendre plus robuste le processus d'appariement.

Images recalées Pour certaines applications, on dispose de paires d'images pour lesquelles la transformation entre les deux vues est connue jusqu'à un certain degré. Un recalage –même grossier– peut donc être réalisé de manière à faciliter la mise en correspondance de points d'intérêt. En reconstruction 3D par exemple, un dispositif calibré permet d'obtenir des paires d'images stéréoscopiques, où un point dans une image est contraint d'appartenir à une ligne –dite « épipolaire »– dans l'autre image (voir le rappel sur la géométrie épipolaire en annexe C). Le mouvement du point selon cette ligne (ou « disparité ») est alors fonction de la distance entre ce point et la caméra. En ayant un *a priori* sur la scène (distance minimum et maximum à la caméra par exemple), la mise en correspondance de points peut être limitée à une zone très restreinte de l'image (un segment). Cela permet ainsi en pratique de réduire de manière spectaculaire le nombre de fausses correspondances et de mettre en oeuvre des méthodes de mise en correspondance dense extrêmement précises [SAM08]. On retrouve d'autres champs d'application où le processus d'appariement peut être contraint spatialement, par exemple les bases de photographies de visages [CJ07b], ou bien encore le suivi dans des séquences vidéo [ZYS09] où cette contrainte est à la fois spatiale et temporelle (selon le mouvement de l'objet entre deux trames successives).

Mise en correspondance guidées Dans un cadre plus général où la transformation entre deux images est inconnue, nous avons vu que c'est l'appariement puis le groupement de points d'intérêt qui permet d'inférer les transformations des différents objets constituant la scène entre plusieurs vues. Certains auteurs proposent de réaliser conjointement la correspondance et le groupement des points d'intérêt, en imposant de manière locale ou globale des contraintes géométriques sur les appariements. Pour la reconstruction 3D d'objets rigides, qui requiert entre autres la mise en correspondance dense de points d'intérêt, l'utilisation de contraintes épipolaires globales permet d'améliorer les performances et de diminuer la complexité du processus [ZK06a, KP06, LQ00]. D'autre travaux ont été proposés pour le cas des objets non rigides, notamment [CJ07a, FTG06] qui imposent seulement des contraintes locales. On retiendra également les travaux de [MS04], où est étudié le cas particulier du groupement de points d'intérêt sans mise en correspondance.

Un autre type d'approche a été proposée dans [SARK08], où le problème de la mise en correspondance est traité comme un problème de minimisation d'énergie. Cette énergie est définie en fonction de la similarité des points d'intérêt appariés (donnée comparaison de leurs descripteurs SIFT) mais également en fonction de la cohérence géométrique de la mise en correspondance des points d'intérêt voisins.

Notons enfin qu'un autre type de contrainte géométrique à été proposé par [Sch96] (et repris par exemple dans [SZ06]). Il s'agit de ne valider une mise en correspondance entre a^i et b^j que lorsque les descripteurs du voisinage de a^i sont également mis en correspondance avec des descripteurs du voisinage de b^j. Ce type de stratégie réduit cependant le nombre de mises en correspondances validées et nécessite l'emploi de nouveaux paramètres pour définir les contraintes sur le voisinage spatial et le nombre de points d'intérêt appariés dans ce voisinage.

1.3.5 Méthodes de comparaisons approchées de descripteurs locaux

Le problème évoqué en introduction de ce chapitre est le temps de calcul du processus de correspondance. C'est en effet, dans la plupart des systèmes de détection et de reconnaissance en vision, l'étape qui représente la part la plus importante en charge de calcul. De plus, la complexité de la mise en correspondance est intrinsèquement liée à la taille de la base de données. Pour diminuer cette complexité, de nombreuses études ont montré que l'on pouvait réduire la taille du descripteur par des méthodes d'ACP (analyse en composante principale), tout en limitant la détérioration des performances de mise en correspondance (voir par exemple [MM07]). Cependant, lorsqu'il s'agit de comparer une requête avec une base de plusieurs millions d'images comme c'est le cas dans [CPS+07], il est tout à fait irréalisable de faire une comparaison exhaustive de chaque descripteur avec l'ensemble des descripteurs de la base de recherche. Nous exposons brièvement dans cette section les outils qui permettent de réaliser de manière approchée de telles comparaisons dans un temps raisonnable.

Dans la littérature, il existe deux approches complémentaires afin de permettre la comparaison d'images *via* des descripteurs locaux avec de grandes bases d'images. La première, apparue très tôt pour la fouille de données, consiste à chercher de manière efficace le plus proche voisin d'un point dans un ensemble de données en exploitant sa structure. La seconde approche, plus récente, est spécifique à la comparaison d'images. Elle repose sur une représentation globale des images à partir de statistiques sur les descripteurs locaux.

Algorithme de recherche approchée La problématique est la suivante : étant donné un descripteur, comment trouver de manière rapide son plus proche voisin parmi un ensemble de descripteurs ? La réponse la plus simple est de partitionner l'ensemble de données en sous-ensembles (à l'aide par exemple d'un algorithme reposant sur les k-moyennes), dont on délimite le domaine spatial et auxquels on attribue un descripteur représentatif du sous-ensemble (*e.g.* les cellules de Voronoï et leurs centroïdes). Cela revient à créer une structure associée à la base de descripteurs, qui permet de limiter le nombre de comparaisons : lorsque l'on cherche le plus proche voisin d'un descripteur requête, on cherche d'abord à quel domaine il appartient, pour ensuite le comparer avec tous les éléments de ce sous-ensemble[1]. À partir de ce principe, de nombreuses méthodes ont été mises au point pour représenter de manière efficace une base de données et optimiser la recherche dans une telle structure. On retiendra notamment les représentations hiérarchiques en arbres de données obtenues par les algorithmes de type *kd-tree*. Cependant, l'inconvénient majeur de ces méthodes est qu'elles souffrent du problème classique de partitionnement : la malédiction de la dimension (*curse of dimensionality*). En effet, lorsque le descripteur est de grande dimension (typiquement 128 dans le cas des SIFTs [Low04]), le nombre de descripteurs peut être largement insuffisant pour donner des statistiques fiables et obtenir une représentation optimale de l'ensemble des éléments de la base [ML09]. Dès lors, on peut arriver à des situations paradoxales où ce type d'algorithme de recherche hiérarchique est moins rapide qu'une recherche exhaustive au delà d'une dimension critique du descripteur [MM07]. Ceci est d'autant plus vrai que le nombre d'images comparées est faible. De plus, pour chaque nouvelle paire d'images analysée, il est nécessaire de recalculer la structure des données. Dans [CK08], les auteurs montrent ainsi que le gain en temps de calcul est seulement d'un facteur 3 pour la comparaison de deux images. Ainsi que le résume [ML09] :

> *The classical kd-tree algorithm (Freidman et al., 1977) is efficient in low dimensions, but in high dimensions the performance rapidly degrades. To obtain a speedup over linear search it becomes necessary to settle for an approximate nearest-neighbor. This improves the search speed at the cost of the algorithm not always returning the exact nearest neighbors.*

En pratique, ce sont donc des algorithmes de recherche **approchés** qui sont mis en oeuvre pour de grandes bases d'images, afin d'obtenir un gain de temps suffisamment important, à titre d'exemple : *locality-sensitive hashing* (LSH) [DIIM04], *ε-approximate nearest neighbor* [AMN+98], ou encore *multiple randomized kd tree* [SAH08]. En plus de la nécessité d'une gestion efficace de la mémoire, il existe

[1] ainsi que les éléments des domaines voisins si le descripteur est proche de la frontière du sous-ensemble

en pratique de nombreux paramètres à régler afin d'optimiser les performances, en particulier afin de contrôler le taux de précision moyen avec lequel le plus proche voisin est effectivement trouvé par l'algorithme. En outre, le temps de calcul requis pour la construction de la structure hiérarchique est d'autant plus important que l'on souhaite diminuer le temps de recherche. Le compromis entre temps de calcul de la structure et temps de recherche dépend de l'application considérée : dans certains cas, la construction de l'arbre peut se faire *hors ligne* (offline), et l'on peut alors obtenir des gains en temps de calcul très conséquents par rapport à une comparaison exhaustive (jusqu'à trois ordres de grandeur) [ML09].

Représentation par sac de mots (Bag of features) Un autre type de stratégie à été suggéré plus récemment pour la mise en correspondance de descripteurs locaux. Jusqu'à présent, nous avons considéré chaque image comme un ensemble de descripteurs locaux, pour chacun desquels on cherche des appariements avec des descripteurs locaux d'une base d'images. Au contraire, l'approche « sac de mots » (l'étymologie venant du domaine de l'analyse textuelle dont elle est issue) consiste à représenter une image par un histogramme d'occurrences de mots [SZ06, CPS+07]. Pour cela, la base de données de descripteurs est tout d'abord partitionnée de manière non supervisée en sous-ensembles pour obtenir des classes sémantiques, qui constituent en quelque sorte des « mots visuels ». La comparaison de deux images peut alors se faire grâce à leurs histogrammes de vocabulaire visuels, c'est-à-dire sans prendre en compte la position spatiale des points d'intérêt, ce qui permet de rejeter à moindre coût les paires d'images différentes. Pour les images jugées similaires, cette approche permet en outre de limiter la comparaison de descripteurs locaux aux éléments appartenant à la même classe, réduisant par conséquent considérablement le temps de calcul. Afin d'améliorer les performances de cette approche, des travaux récents proposent de prendre en compte la disposition relative spatiale des points d'intérêt, à l'image de [Ved08].

Dans le chapitre suivant, nous allons présenter un nouveau critère de mise en correspondance pour des descripteurs locaux de type SIFT, inspiré des méthodes *a contrario*.

Chapitre 2

Mise en correspondance *a contrario* de descripteurs locaux

Ce chapitre est consacré à la mise en correspondance (ou appariement) de descripteurs locaux dont nous avons étudié le principe et les enjeux dans le chapitre précédent. Un nouveau critère de validation des mises en correspondance, fondé sur la méthodologie *a contrario*, est introduit. L'intérêt de notre approche est ensuite validée expérimentalement sur une base d'images dans une seconde partie.

Ces travaux ont fait l'objet d'une publication dans [RDG08a, RDG09].

2.1 Critère de mise en correspondance *a contrario*

Rappelons que notre objectif est de choisir des correspondances entre N_Q descripteurs requêtes a^i d'une image A et N_C descripteurs candidats b^j d'une base B constituée de plusieurs images. Etant donnée une mesure de dissimilarité D entre descripteurs, on souhaite donc savoir comment seuiller les mesures $D(a^i, b^j)$ de manière pertinente, afin de valider certaines mises en correspondance. Nous avons vu au paragraphe 1.3.3 qu'il existait plusieurs critères pour seuiller ces mesures de dissimilarité. Plus précisément, les trois principaux critères de la littérature sont :

- le critère DT, qui consiste à conserver toutes les correspondances (a^i, b^j) pour lesquelles $D(a^i, b^j)$ est sous un seuil global donné (indépendant de a^i) ;
- le critère NN-DT, qui consiste à appliquer le critère précédent tout en se restreignant pour chaque a^i à son plus proche voisin dans la base (une seule correspondance est autorisée par descripteur requête) ;
- le critère NN-DR, introduit par D. Lowe dans [Low04], qui consiste à seuiller globalement le rapport entre la distance au plus proche voisin et la distance au deuxième plus proche voisin.

En pratique, on voit que ces différents critères se contentent d'un seuillage sur une mesure de qualité utilisant au plus les deux plus proches voisins d'un descripteur dans la base de données. Le seuil de détection est fixé par l'utilisateur pour l'ensemble des correspondances examinées. Un critère idéal devrait, au contraire, pouvoir s'adapter à la diversité de la base de données et aux descripteurs requêtes considérés, en définissant des seuils de détection adaptatifs qui autorisent les correspondances multiples.

Nous allons voir dans ce chapitre comment la méthodologie dite *a contrario* permet de fixer de manière adaptative des seuils sur ces correspondances, afin d'assurer le rejet des mises en correspondance accidentelles. Pour plus de détails sur les principes généraux de cette méthodologie, introduite par Desolneux, Moisan et Morel [DMM08], on se réfèrera à l'annexe A. Précisons tout d'abord quels types de descripteurs et de mesure de dissimilarité D sont adaptés à ce cadre de travail.

2.1.1 Introduction

On s'intéresse au cas de la comparaison exhaustive de descripteurs : on suppose ainsi que, pour chaque requête a^i, l'ensemble des distances avec les éléments de la base $\{D(a^i, b^j)\ ,\ j \in \{1, \dots, N_C\}\}$ ont été calculées préliminairement. Aucune autre connaissance *a priori* n'est requise sur les caractéristiques des descripteurs à mettre en correspondance (caractéristiques géométriques ou sémantiques par exemple).

Descripteurs On suppose dans ce chapitre que les descripteurs utilisés pour représenter un point ou une région d'intérêt sont une collection de M vecteurs caractéristiques de dimension N, à l'image des SIFTs, qui peuvent être considérés comme un ensemble de M histogrammes d'orientation du gradient[1], et qui seront utilisés dans la partie expérimentale. D'autres descripteurs locaux possèdent également une telle structure, comme par exemple les *Shapes Context* [BMP02], et pourraient s'adapter à notre cadre de travail. On note désormais $\{a^i := (a^i_1, \dots, a^i_M)\}$ (et de manière analogue $\{b^j := (b^j_1, \dots, b^j_M)\}$) les descripteurs de A (respectivement B).

Mesure de dissimilarité On suppose également, sans trop perdre en généralité, que l'on peut exprimer la dissimilarité entre deux descripteurs a^i et b^j comme la somme

$$D(a^i, b^j) = \sum_{m=1}^{M} d(a^i_m, b^j_m)\ , \tag{2.1}$$

où $d(.,.)$ est une métrique laissée au choix de l'utilisateur. On remarquera qu'il est possible de se ramener à cette écriture pour la plupart des mesures de dissimilarité utilisées pour comparer des descripteurs locaux. C'est par exemple le cas des distances L^2 ou χ^2 qui sont généralement utilisées pour comparer les descripteurs SIFT. C'est également le cas de la nouvelle mesure de dissimilarité qui est introduite dans le chapitre 7. Cette mesure, définie à partir du coût de transport circulaire entre histogrammes d'orientation, sera utilisée dans la partie expérimentale.

Approche *a contrario* pour la mise en correspondance de caractéristiques La théorie de la détection *a contrario*, qui a été proposée par Desolneux, Moisan et Morel [DMM08], s'inspire des tests d'hypothèse pour détecter des groupements significatifs d'objets partageant des caractéristiques similaires. Ces groupes sont détectés s'ils ont fort peu de chance d'apparaître sous l'hypothèse que les caractérisiques que l'on observe sont indépendantes. Cette hypothèse d'indépendance est appelée *hypothèse nulle*, et le modèle selon lequel les caractéristiques des objets suivent cette hypothèse est appelé modèle de fond. Ces méthodes permettent également d'associer à chaque groupe observé un degré de significativité (ou de qualité), noté NFA, qui mesure à quel point le groupe rejette le modèle de fond.

Notons que, parmi l'ensemble des applications de la théorie de la décision *a contrario*, certaines ont déjà eu trait à la mise en correspondance de descripteurs locaux. Sur *et al.* ont les premiers utilisé ce cadre de travail afin d'apparier des morceaux de lignes de niveau [MSC+06] entre des paires d'images. Dans [CLM+08], un chapitre est consacré à la mise en correspondance *a contrario* de descripteurs locaux inspirés des SIFTs. Ces descripteurs sont une collection d'orientations du gradient, échantillonnés en certaines positions de patchs normalisés. La problématique de la mise en correspondance de ces descripteurs est ensuite posée de manière similaire à la détection de segments [DMM00]. Ce travail se distingue de notre approche par le fait que les descripteurs, la mesure de dissimilarité ainsi que le critère utilisés sont différents. Dans le cadre de la vision stéréoscopique, un critère de décision *a contrario* est proposé dans [SAM08] pour la mise en correspondance dense de points afin de construire une carte de disparité.

Nous proposons dans ce chapitre de définir un critère de mise en correspondance de descripteurs locaux fondé sur la théorie de la décision *a contrario*, définissant des seuils adaptatifs sur la mesure de

[1] Les SIFTs [Low04] sont des histogrammes 3D d'orientation du gradient, calculés en diverses régions d'une grille (voir en annexe B).

dissimilarité, et permettant les appariements multiples tout en limitant le nombre de fausses détections. Nous allons dans ce cadre de travail définir un *modèle de fond*. Les mises en correspondance validées seront celles qui rejettent ce modèle. Une *mesure de qualité* est également définie de manière à contrôler le nombre de fausses alarmes.

2.1.2 Modèle de fond

Définition de l'hypothèse nulle Nous souhaitons définir un critère de mise en correspondance qui puisse s'adapter aussi bien à la base de données qu'à chaque requête a^i considérée. Dans ce but, nous allons donc considérer dans ce qui suit l'appariement de a^i, descripteur de A, avec un descripteur aléatoire $\mathbf{b}$, composé de M caractéristiques aléatoires $(\mathbf{b}_1, \dots, \mathbf{b}_M)$.

Rappelons qu'en pratique, afin de mesurer le degré de dissimilarité des caractéristiques d'un couple de descripteurs (a, b), on utilise une mesure $D(a, b) = \sum_{m=1}^{M} d(a_m, b_m)$. Nous proposons alors de définir une hypothèse nulle pour la mise en correspondance de la requête a^i avec un descripteur aléatoire $\mathbf{b}$ en fonction de l'indépendance des distances $d(a^i_m, \mathbf{b}_m)$.

Définition 4 *Pour tout $i \in \{1, \dots, N_Q\}$, on dira qu'un descripteur aléatoire $\mathbf{b}$ satisfait la i-ième hypothèse nulle, notée $\mathcal{H}^i_0$, si les distances $\left\{d(a^i_m, \mathbf{b}_m)\right\}_{m\in\{1,\dots M\}}$ sont des variables aléatoires mutuellement indépendantes.*

Cette hypothèse nulle $\mathcal{H}^i_0$ va nous permettre de définir ce qu'est une mise en correspondance « sûre » : il s'agit d'un appariement de deux descripteurs (a^i, b^j) dont la mesure de dissimilarité est anormalement petite sous l'hypothèse que les distances observées $d(a^i_m, b^j_m)$ sont indépendantes. Pour savoir s'il faut valider ou non une telle mise en correspondance, il faut donc être capable de mesurer le degré de rejet de cette hypothèse nulle.

Avant de présenter la façon de réaliser cette mesure, nous allons brièvement discuter de la validité pratique de ce modèle de fond.

Sur la validité pratique des hypothèses nulles Dans ce qui suit, on notera $\mathcal{H}_0 = \{\mathcal{H}^i_0\}$ l'ensemble des hypothèses nulles. Un descripteur aléatoire $\mathbf{b}$ suit $\mathcal{H}_0$ si et seulement si il suit tous les H^i_0. Cette hypothèse nulle globale $\mathcal{H}_0$ doit permettre le rejet des fausses mises en correspondance (ou « faux-positifs »). Pour que $\mathcal{H}_0$ soit vérifié en pratique, il faut s'assurer par construction des descripteurs de l'indépendance entre les M caractéristiques d'un descripteur. Par exemple, pour la mise en correspondance de descripteurs d'orientation du gradient, les auteurs de [CLM+08] définissent une procédure d'échantillonnage visant à s'assurer de l'indépendance des échantillons.

Dans notre cas, nous supposons donc que les descripteurs sont construits de telle manière qu'ils satisfassent l'hypothèse globale $\mathcal{H}_0$. Nous vérifierons dans la partie expérimentale la validité de cette assertion.

2.1.3 Mesure de significativité d'une correspondance

Nous allons maintenant définir la mesure de significativité d'un appariement du descripteur a^i avec un élément de la base, à partir du modèle de fond.

Probabilité sous hypothèse nulle Un descripteur requête a^i est mis en correspondance avec un descripteur $\mathbf{b}$ si la distance $D(a^i, \mathbf{b})$ est suffisamment petite sous l'hypothèse nulle $\mathcal{H}^i_0$ que les variables aléatoires $d(a^i_m, \mathbf{b}_m)$ sont indépendantes. Or, sous cette hypothèse, la densité de probabilité de la variable $D(a^i, \mathbf{b})$ peut s'exprimer comme la convolution des M densités de probabilité des variables aléatoires $d(a^i_m, \mathbf{b}_m)$, ce que l'on note symboliquement $\underset{m=1}{\overset{M}{*}} p^i_m = p^i_1 * \dots * p^i_M$, où p^i_m désigne la densité de probabilité de $d(a^i_m, \mathbf{b}_m)$. On peut en déduire la proposition suivante :

Proposition 1 *Sous l'hypothèse $\mathcal{H}_0^i$, la probabilité que la distance entre a^i et $\mathbf{b}$ soit plus petite qu'un seuil donné δ s'exprime comme*

$$\mathbb{P}\left(D(a^i,\mathbf{b}) \leq \delta \,|\, \mathcal{H}_0^i\right) = \int_0^{\delta} \underset{m=1}{\overset{M}{*}} p_m^i(x)\, dx \ . \tag{2.2}$$

Suivant la méthodologie *a contrario*, nous allons maintenant définir une mesure de qualité d'une correspondance conditionnellement au modèle de fond. Cette mesure de qualité est évaluée grâce à la probabilité définie à l'équation (2.2).

Définition du NFA Selon la théorie de la détection *a contrario*, une mesure de qualité intitulée **NFA** (pour « nombre de fausses alarmes ») est définie pour contrôler globalement le nombre de fausses détections de l'ensemble du processus de détection. Son expression générique, qui est rappelée en annexe A, consiste à pondérer par un nombre de tests la probabilité d'observer un groupe sous l'hypothèse nulle. Dans notre cas, le nombre de tests est égal au produit $N_Q \times N_C$. En utilisant la proposition 1, on peut alors définir le NFA de la mise en correspondance d'une requête a^i avec un descripteur aléatoire $\mathbf{b}$ en fonction du seuil δ sur la mesure de dissimilarité $D(a^i,\mathbf{b})$.

Définition 5 *Étant donné un descripteur a^i, pour tout $\delta > 0$ on définit la quantité*

$$\mathit{NFA}(a^i,\delta) := N_Q\, N_C \cdot \mathbb{P}\left(D(a^i,\mathbf{b}) \leq \delta \,|\, \mathcal{H}_0^i\right) \ . \tag{2.3}$$

Cette définition du NFA va nous permettre, en tant que mesure de qualité, de tester et de valider les mises en correspondance qui rejettent significativement le modèle de fond.

Estimation du NFA Considérons maintenant l'évaluation pratique de l'appariement (a^i,b^j) à partir de la mesure de dissimilarité observée $D(a^i,b^j)$ et de la définition du NFA (2.3). En choisissant $\delta = D(a^i,b^j)$, le NFA mesure les chances d'observer des correspondances vérifiant le modèle de fond qui ont un degré de dissimilarité plus faible que $D(a^i,b^j)$. Pour reprendre les notations du chapitre précédent, cela revient à définir la mesure de qualité de la manière suivante :

$$Q(a^i,b^j) = \text{NFA}(a^i, D(a^i,b^j)) = N_Q N_C \cdot \mathbb{P}\left(D(a^i,\mathbf{b}) \leq D(a^i,b^j) \,|\, \mathcal{H}_0^i\right) \ .$$

Pour estimer la valeur de cette fonction, il faut exprimer numériquement la probabilité définie à l'équation (2.2)) pour n'importe quelle valeur de $\delta = D(a^i,b^j)$. Pour cela, il est nécessaire de connaître *a priori* les lois marginales des variables $d(a_m^i,\mathbf{b}_m)$, pour chaque $i \in \{1,\dots N_Q\}$ et chaque $m \in \{1,\dots M\}$.

Or, nous voulons que le critère de mise en correspondance s'adapte à la fois au descripteur requête a^i, mais également aux descripteurs candidats de l'image B. Pour ces raisons, nous avons choisi d'**estimer empiriquement** ces densités de probabilité marginales à partir des distances calculées entre les M vecteurs du descripteur $a^i = (a_1^i,\dots,a_M^i)$ et l'ensemble des vecteurs de même indice dans la base $(b_1^j,\dots,b_M^j)_{1\leq j\leq N_C}$. On exprime ainsi les lois p_m^i comme les histogrammes de réalisation des distances observées :

$$\forall i \in \{1,\dots,N_Q\}, \quad p_m^i(x) \ = \ \frac{1}{N_C}\, \#\left\{\ j\ ,\ d(a_m^i,b_m^j) = x\ \right\} \ .$$

De ce fait, nous pouvons à présent associer pour chaque couple de descripteurs une mesure de qualité reposant sur un modèle de fond appris directement sur la base. Il ne nous reste plus qu'à définir un critère de mise en correspondance fondé sur le seuillage du NFA.

2.1.4 Critère de validation automatique des mises en correspondance

2.1.4.1 Seuillage du NFA

L'une des propriétés fondamentales du NFA est de permettre le contrôle du nombre de fausses détections (erreur de type I) par l'utilisation d'un simple seuil global, que l'on note ε, fixé pour l'ensemble du processus de correspondance et pour toutes les expériences. Cela revient à définir un critère de mise en correspondance *a contrario* que l'on désignera désormais par **critère AC**.

Définition 6 (Critère AC) *Une correspondance (a^i, b^j) est validée si la quantitée $NFA(a^i, D(a^i, b^j))$ est inférieure au seuil ε. On obtient l'ensemble des correspondances suivant :*

$$\mathcal{C}_{AC} := \Big\{ (a^i, b^j) \ , \ NFA\big(a^i, D(a^i, b^j)\big) \leq \varepsilon \ \ \forall i \in \{1, \dots, N_Q\} \ \textit{et} \ \forall j \in \{1, \dots, N_C\} \Big\}$$

On appelle ε***-significatives*** l'ensemble des correspondances qui sont ainsi validées à l'aide du seuil ε.

L'appellation « nombre de fausses alarmes » peut prêter à confusion. Il ne s'agit pas du nombre de faux-positifs qui sont effectivement validées par le critère ainsi défini. Le NFA représente une borne sur l'espérance du nombre de correspondances vérifiant le modèle de fond, qui auraient été validées en réalisant le même nombre de tests.

Proposition 2 *Soient $\{a^i\}$ un ensemble de N_Q descripteurs requête et $\{\mathbf{b}^j\}$ un ensemble de N_C descripteurs aléatoires vérifiant l'ensemble des hypothèses $\mathcal{H}_0^i$ (c'est-à-dire l'hypothèse globale $\mathcal{H}_0$). Alors, l'espérance du nombre de correspondances ε-significatives est plus petit que ε.*

Preuve de la proposition 2 On définit $F_i(\delta) := \mathbb{P}_{\mathcal{H}_0^i}(D(a^i, \mathbf{b}) \leq \delta)$ la fonction de répartition de la variable aléatoire $D(a^i, \mathbf{b})$ lorsque le descripteur aléatoire $\mathbf{b}$ suit l'hypothèse nulle $\mathcal{H}_0^i$. Rappelons que si X est une variable aléatoire de fonction de répartition F_X, alors on a $\mathbb{P}(F_X(X) \leq \alpha) \leq \alpha$ pour tout α positif. Par suite, on peut écrire : $\mathbb{P}_{\mathcal{H}_0^i}(F_i(D(a^i, \mathbf{b})) \leq \alpha) \leq \alpha$. On note $\mathbb{1}$ l'indicatrice des évènements, et $\mathbb{E}_{\mathcal{H}_0}$ l'espérance suivant l'hypothèse nulle globale $\mathcal{H}_0$.

$$\begin{aligned}
\mathbb{E}_{\mathcal{H}_0}\Big[\sum_{i=1}^{N_C} \sum_{j=1}^{N_Q} \mathbb{1}\big\{ (a^i, \mathbf{b}^j) \text{ est } \varepsilon\text{-significatif} \big\} \Big] &= \sum_{i=1}^{N_C} \sum_{j=1}^{N_Q} \mathbb{P}_{\mathcal{H}_0^i}\Big[\text{NFA}(a^i, D(a^i, \mathbf{b}^j)) \leq \varepsilon \Big] \\
&= \sum_{i=1}^{N_C} \sum_{j=1}^{N_Q} \mathbb{P}_{\mathcal{H}_0^i}\Big[F_i\big(D(a^i, \mathbf{b}^j)\big) \leq \frac{\varepsilon}{N_Q N_C} \Big] \\
&\leq \sum_{i=1}^{N_C} \sum_{j=1}^{N_Q} \frac{\varepsilon}{N_Q N_C} \ = \ \varepsilon \ .
\end{aligned}$$

□

2.1.4.2 Définition des seuils sur la mesure de dissimilarité

Nous avons défini aux paragraphes précédents une mesure de qualité d'une correspondance (NFA) et un critère de décision à l'aide d'un simple seuil de sélection ε. Pour interpréter le fonctionnement de ce nouveau critère, nous allons à présent en donner une définition alternative.

Pour cela, nous introduisons la quantité $\delta_i(\varepsilon)$, correspondant à la mesure de dissimilarité maximale entre un descripteur a^i et un descripteur aléatoire vérifiant le modèle de fond, telle que le couple soit ε-significatif :

$$\delta_i(\varepsilon) = \arg\max_{\delta} \big\{ \text{NFA}(a^i, \delta) \leq \varepsilon \big\} \ . \tag{2.4}$$

Dès lors, pour chaque descripteur requête a^i, un appariement entre a^i et un descripteur b^j sera ε-significatif si la mesure de dissimilarité est plus petite que le seuil $\delta_i(\varepsilon)$ (voir l'illustration 2.1).

Ainsi, en définissant la mesure de qualité d'un couple comme la mesure de dissimilarité $Q(a^i, b^j) = D(a^i, b^j)$, le critère AC peut donc être vu comme la construction automatique, à partir d'un unique paramètre ε, d'un ensemble de seuils $\{\delta_i(\varepsilon)\}_{1 \leq i \leq N_C}$ sur cette mesure de dissimilarité. Ce processus de mise en correspondance est présenté sous la forme d'un algorithme en table 2.1.

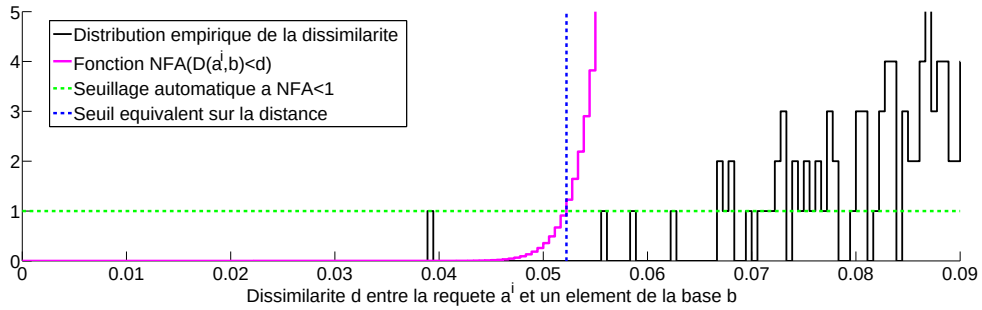


FIG. 2.1 – *Illustration de la sélection automatique de seuil sur la mesure de dissimilarité en fonction du NFA.*

TAB. 2.1 – *Procédure de mise en correspondance A Contrario (AC).*

Algorithme 2.1 Seuillage automatique de la distance.

Entrées : N_Q descripteurs requêtes $\{a^i\}$, N_C descripteurs candidats $\{b^j\}$, et le paramètre ε.

Pour chaque descripteur requête a^i, $i = 1, \ldots, N_Q$:

1) Calcul des distances $d_m(a^i, b^j)$ pour tous les $m = 1, \ldots, M$ et $j = 1, \ldots, N_C$;
2) Estimation des densités de probabilité : pour chaque m, la densité p^i_m est calculée à partir de la distribution empirique de $d(a^i_m, b^j_m)$, pour l'ensemble des b^j_m de la base B ;
3) Calcul de la probabilité $\mathbb{P}\left(D(a^i, \mathbf{b}) \leq \delta \,|\, \mathcal{H}^i_0\right)$ à partir de la définition (2.2) ;
4) Calcul du seuil $\delta_i(\varepsilon)$ à partir de la formule (2.4) ;
5) Appariement de a^i avec les descripteurs b^j tels que $D(a^i, b^j) \leq \delta_i(\varepsilon)$;

Sortie : Liste des correspondances $\mathcal{C}_{AC}$.

2.1.4.3 Comportement asymptotique du seuil ε

Contrairement aux seuils utilisés par les critères DT, NN-DT et NN-DR définis au chapitre précédent, le réglage du seuil global ε est très intuitif. Il correspond au nombre moyen de fausses détections, avec un nombre de tests équivalent, pour des descripteurs aléatoires.

Dans le paragraphe suivant nous étudions le comportement asymptotique des seuils sur la distance $\delta_i(\varepsilon)$ en fonction du paramètre ε. Comme c'est souvent le cas des applications de la méthode *a contrario*, l'influence du paramètre ε est mieux exprimée sur une échelle logarithmique. Pour cette raison, les valeurs du paramètre ε utilisées en partie expérimentale pour illustrer le comportement du critère de mise en correspondance AC seront exprimées en puissance de 10.

Condition suffisante de significativité Soit a un descripteur quelconque donné, et $\mathbf{b}$ un descripteur aléatoire tel que les distances $\mathbf{d}_m = d(a_m, \mathbf{b}_m)$ soient indépendantes et identiquement distribuées (hypothèse nulle $\mathcal{H}_0$). Supposons que leurs densités de probabilité peuvent être approchées par des distributions gaussiennes. La convolution de ces M lois marginales est alors une distribution gaussienne, dont la moyenne et l'écart-type sont respectivement notés μ et σ.

Dans ce qui suit, nous montrons que si $\varepsilon \leq \dfrac{N_Q N_C}{2e\sqrt{\pi}}$ et si $D(a,b) < \mu - \sigma\sqrt{2}\sqrt{\log \dfrac{N_Q N_C}{2\sqrt{\pi}\varepsilon}}$[2], alors la correspondance de a et b est ε-significative.

Démonstration À partir de la définition de la significativité (critère 6), la mise en correspondance de a et $\mathbf{b}$ est ε-significative si $\mathrm{NFA}(a, \delta) \leq \varepsilon$. Sous les hypothèses précédentes, le NFA défini à l'équation

[2] où log désigne le logarithme népérien et $e = \exp(1)$.

(2.3) peut être réécrit à partir de la fonction d'erreur gaussienne :

$$\text{NFA}(a,\delta) = \frac{N_Q N_C}{2}\left(1+\text{erf}(\frac{\delta-\mu}{\sqrt{2}\sigma})\right),$$

où la fonction d'erreur gaussienne est définie comme

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt = \frac{2}{\sqrt{\pi}}\int_{-\infty}^x e^{-t^2}dt - 1\ .$$

Pour $x<0$, on peut définir la borne supérieure suivante

$$\text{erf}(x) \leq \frac{2}{\sqrt{\pi}}\int_{-\infty}^x e^{-t^2}\left(1+\frac{1}{2t^2}\right)dt - 1 = \frac{1}{\sqrt{\pi}}\frac{e^{-x^2}}{|x|} - 1\ .$$

Pour $\delta<\mu$, on peut donc écrire

$$\text{NFA}(a,\delta) \leq \frac{N_Q N_C}{2\sqrt{\pi}}\frac{e^{-\left(\frac{\delta-\mu}{\sqrt{2}\sigma}\right)^2}}{\left|\frac{\delta-\mu}{\sqrt{2}\sigma}\right|}\ . \tag{2.5}$$

Par ailleurs, si $D(a,b) < \mu - \sigma\sqrt{2}\sqrt{\log\frac{N_Q N_C}{2\sqrt{\pi}\varepsilon}}$, alors

$$\frac{D(a,b)-\mu}{\sigma\sqrt{2}} < -\sqrt{\log\frac{N_Q N_C}{2\sqrt{\pi}\varepsilon}}\ \text{, et}\ \left(\frac{D(a,b)-\mu}{\sigma\sqrt{2}}\right)^2 > \log\frac{N_Q N_C}{2\sqrt{\pi}\varepsilon}\ .$$

En supposant $\varepsilon \leq \frac{N_Q N_C}{2e\sqrt{\pi}}$, alors $\log\frac{N_Q N_C}{2\sqrt{\pi}\varepsilon} > 1$, et on en déduit que $\frac{|D(a,b)-\mu|}{\sigma\sqrt{2}} > 1$.

Ainsi,

$$\left(\frac{D(a,b)-\mu}{\sigma\sqrt{2}}\right)^2 + \log\left|\frac{D(a,b)-\mu}{\sigma\sqrt{2}}\right| > \log\frac{N_Q N_C}{2\sqrt{\pi}\varepsilon}\ .$$

En combinant le dernier résultat et l'équation (2.5) pour définir une borne sur $\text{NFA}(a, D(a,b))$, on obtient finalement :

$$\text{NFA}(a, D(a,b)) \leq \varepsilon\ .$$

□

2.1.4.4 *Principe de maximalité*

On s'intéresse ici à la notion de *maximalité* souvent utilisée avec la méthode de décision *a contrario*, et parfois nécessaire selon les applications considérées. Transposé dans notre cadre de travail, le principe de maximalité consiste à ne pouvoir valider pour chaque descripteur requête que la correspondance avec l'élément de la base ayant le plus faible NFA. Or, le NFA étant une fonction croissante de la mesure de dissimilarité, ce principe revient donc à une restriction au plus proche voisin (NN). On se réfère désormais au critère NN-AC comme à la restriction au plus proche voisin du critère AC.

Définition 7 (Critère NN-AC) *Pour chaque requête a^i, le descripteur candidat le plus proche est validé si le NFA de l'appariement est plus petit que le seuil global ε. On obtient l'ensemble des correspondances suivant :*

$$\mathcal{C}_{\textit{NN-AC}} := \left\{ (a^i, b^{J(i)})\ ,\ \textit{NFA}\left(a^i, D(a^i, b^{J(i)})\right) \leq \varepsilon\ \ \forall\, i \in \{1,\ldots,N_Q\}\ \textit{avec}\ J(i) = \underset{j\in\{1,\ldots,N_C\}}{\arg\min}\ D(a^i,b^j) \right\}\ .$$

Cette définition va permettre la comparaison de notre approche avec les critères usuels restreints au plus proche voisin (critères NN-DT et NN-DR). Nous discuterons de l'intérêt réel d'une telle restriction dans la section expérimentale suivante.

2.2 Évaluation expérimentale

Cette section est consacrée à la validation expérimentale du critère *A Contrario* qui vient d'être présenté. Cette validation repose sur une analyse comparative de ses performances sur une large base d'images pour différents types de scénarios. Nous présentons par la suite quelques expériences pour illustrer les conclusions de cette analyse.

2.2.1 Mise en œuvre

Avant de présenter notre protocole expérimental, nous allons tout d'abord détailler la mise en œuvre de notre critère de mise en correspondance. Le paragraphe suivant est consacré à l'étape préliminaire de comparaison des descripteurs locaux. Nous étudierons ensuite la complexité de calcul du critère AC, puis la validité du modèle de fond utilisé.

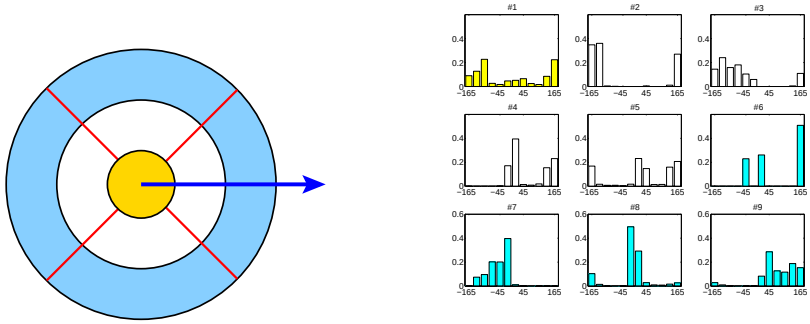
2.2.1.1 Mesure de dissimilarité et descripteurs utilisés

Descripteurs Nous rappelons ici brièvement le procédé de détection et de construction des descripteurs locaux utilisés dans l'ensemble de cette partie expérimentale, qui est décrit en détail en annexe B.

Des points d'intérêt sont tout d'abord détectés dans chacune des images A et B, en utilisant un détecteur de coins de Harris adapté à l'analyse en espace-échelle linéaire (détecteur de « Laplace-Harris », en section B.1). Chaque point d'intérêt possède alors une échelle caractéristique σ. Une analyse de l'histogramme de l'orientation du gradient dans le voisinage du point d'intérêt est ensuite réalisée pour définir une orientation principale θ.

Nous utilisons les descripteurs de type SIFT décrits en section B.2, obtenus de manière très similaire à l'approche originale [Low04]. Des histogrammes d'orientation du gradient sont extraits d'un masque circulaire découpé selon une grille polaire en $M = 9$ régions distinctes (figure 2.2), et centré sur le point d'intérêt considéré. La taille du masque est proportionnelle à l'échelle σ du point d'intérêt, de telle sorte que le rayon du masque est égal à 12σ. L'orientation du masque est en outre recalée selon l'orientation principale θ du point d'intérêt. Les histogrammes d'orientation sont quantifiés sur $N = 12$ bins. La dimension du descripteur SIFT ainsi obtenu est alors de $M \times N = 108$. Les M histogrammes sont finalement normalisés de telle sorte que la norme L^1 du descripteur SIFT soit égale à 1.

Mesure de dissimilarité utilisée Le critère AC tel qu'il a été défini dans ce chapitre suppose seulement que la distance entre deux descripteurs a et b peut s'écrire comme la somme des distances entre les histogrammes d'indice m : $D(a,b) = \sum_{m=1}^{M} d(a_m, b_m)$. N'importe quelle mesure de dissimilarité d peut donc être utilisée pour comparer les histogrammes a_m et b_m, par exemple la distance L^2 [Low04], L^1 ou la distance du χ^2 [BMP02, FL07]. Pour plus de détails, on pourra se référer au chapitre 7 de cette thèse, qui est consacré à l'étude de différents types de distances pour la comparaison de descripteurs locaux. Or, nous montrons que l'utilisation d'une distance de transport circulaire entre les histogrammes d'orientation permet d'accroître la robustesse et le pouvoir discriminant des descripteurs SIFT. Dans la suite des expériences, nous utilisons donc cette mesure de dissimilarité, notée D_{CEMD}, pour comparer les histogrammes a_m et b_m.



(a) Masque circulaire découpé en 9 régions.

(b) 9 histogrammes d'orientation du gradient correspondant aux différentes régions du masque.

FIG. 2.2 – *Illustration du descripteur SIFT.* *À gauche : un masque circulaire, découpé en $M = 9$ régions distinctes, est centré sur le point d'intérêt. Sa taille et son orientation dépendent des caractéristiques du point (échelle et orientation principale). À droite : 9 histogrammes d'orientation du gradient, quantifiés sur $N = 12$ bins, sont extraits de chacune des régions du masque.*

L'expression générale de cette distance est la suivante :

$$D_{\text{CEMD}}(a_m, b_m) = \frac{1}{N} \min_{k \in \{1,\dots,N\}} \|A_m^k - B_m^k\|_1 \ , \text{ avec } \sum_m ||a_m||_1 = \sum_m ||b_m||_1 = 1 \ ,$$

où $||.||_1$ désigne la norme L^1. A_m^k et B_m^k représentent les histogrammes cumulés *circulairement* depuis le bin d'indice k des histogrammes a_m et b_m respectivement[3], ce qui donne pour A_m^k (la définition de B_m^k est la même en remplaçant A par B)

$$\forall i, k \in \{1, \dots, N\} \ , \ A_m^k[i] = \begin{cases} A_m[i] - A_m[k-1] & \text{si } i \geq k \\ A_m[i] - A_m[k-1] + A_m[N] & \text{si } i \leq k-1 \end{cases} \ ,$$

avec la convention $A[0] = 0$.

2.2.1.2 Coût algorithmique

Afin d'analyser la complexité du critère A Contrario (AC) proposé dans ce chapitre, nous étudions les différentes étapes de sa mise en œuvre, qui sont énumérées en table 2.1. On exprime la complexité par le nombre d'opérations élémentaires réalisées (somme, multiplication, *etc.*). Par la suite, nous nous intéressons seulement au terme prépondérant (c'est-à-dire, celui ayant l'ordre le plus élevé).

La première étape de calcul (table 2.1) est commune à tous les critères de mise en correspondance qui requièrent une comparaison exhaustive. Sa complexité dépend de la mesure de dissimilarité utilisée. Avec la distance D_{CEMD}, la complexité de cette étape de comparaison exhaustive est $\mathcal{O}(N_Q N_C M N^2)$.

En comparaison des critères DT et NN-DR, notre approche nécessite les étapes supplémentaires 2, 3 et 4. La complexité des étapes 2 et 4 correspond à celle d'un parcours de tableau, qui est largement négligeable par rapport à celle de l'étape 1. L'étape 3 est cependant beaucoup moins négligeable en temps de calcul, car elle requiert le calcul de $M - 1$ convolutions des marginales calculées à l'étape 2. La complexité de cette étape dépend grandement du nombre de cellules de quantification utilisées pour calculer l'histogramme empirique des distributions marginales. On note Δ ce nombre de cellules. La convolution des deux premiers histogrammes de taille Δ requiert ainsi Δ^2 multiplications et l'histogramme obtenu a une taille $2\Delta - 1$. À chaque nouvelle convolution, la taille de l'histogramme obtenu augmente de $\Delta - 1$.

[3] les deux histogrammes a_m et b_m ne sont pas nécessairement de poids égal. Pour plus de détails sur la question de la normalisation, voir le paragraphe 7.2.

À la n-ième convolution, on doit donc convoluer un histogramme h de taille $n(\Delta-1)+1$ avec un histogramme g de taille Δ, ce que l'on peut écrire sous la forme :

$$\forall i \leq (n+1)(\Delta-1),\ h*g[i] = \sum_{j=0}^{n(\Delta-1)} h[j]g[i-j]\mathbb{1}\{0 \leq i-j \leq \Delta-1\} = \sum_{j=\max\{0,i+1-\Delta\}}^{\min\{n(\Delta-1),i\}} h[j]g[i-j].$$

La complexité de calcul de la n-ième convolution, exprimée comme le nombre de multiplications réalisées, pour l'ensemble des valeurs $i \in \{0,\dots,(n+1)(\Delta-1)\}$ est donc de :

$$\begin{aligned}
\mathcal{M}_n &= \sum_{i=0}^{(n+1)(\Delta-1)} \sum_{j=\max\{0,i+1-\Delta\}}^{\min\{n(\Delta-1),i\}} 1 = \sum_{i=0}^{\Delta-1}\sum_{j=0}^{i} 1 + \sum_{i=\Delta}^{n(\Delta-1)}\sum_{j=i+1-\Delta}^{i} 1 + \sum_{i=n(\Delta-1)+1}^{(n+1)(\Delta-1)}\sum_{j=i+1-\Delta}^{n(\Delta-1)} 1 \\
&= \sum_{i=0}^{\Delta-1}(i+1) + \sum_{i=\Delta}^{n(\Delta-1)}\Delta + \sum_{i=n(\Delta-1)+1}^{(n+1)(\Delta-1)}\{(n+1)(\Delta-1)-i+1\} \\
&= \frac{\Delta(\Delta+1)}{2} + (n+1)(\Delta-1)\Delta + \sum_{k=1}^{\Delta-1} k \;=\; n\Delta^2-(n-1)\Delta\,.
\end{aligned}$$

Au terme de la convolution des M histogrammes, l'histogramme final de taille $M(\Delta-1)+1$ a donc nécessité en nombre de multiplications :

$$\sum_{n=1}^{M-1}\mathcal{M}_n = \sum_{n=1}^{M-1} n\Delta^2-(n-1)\Delta = \frac{M(M-1)}{2}\Delta(\Delta-1)+(M-1)\Delta \approx \frac{M^2\Delta^2}{2}\,.$$

Ce qui donne finalement une complexité en $\mathcal{O}(M^2\Delta^2)$ pour l'estimation du seuil de validation. Le tableau suivant compare la complexité des deux étapes critiques de la mise en correspondance.

TAB. 2.2 – *Complexité des principales étapes du critère* AC.

Étape	Complexité
1) Calcul des distances avec D_{CEMD}	$\mathcal{O}(N_Q N_C M N^2)$
3) Calcul de la probabilité	$\mathcal{O}(N_Q M^2\Delta^2)$

Le surcoût de calcul lié à notre approche dépend en particulier du paramètre Δ, comme le montre le tableau 2.3. Du point de vue des performances néanmoins, ce paramètre n'est pas critique : il correspond seulement à l'approximation de la probabilité utilisée pour le calcul du NFA. Nous avons observé que quelques dizaines de bins sont suffisants pour obtenir de bon résultats. Nous avons fixé la taille des histogrammes empiriques à $\Delta = 100$ bins pour l'ensemble des expériences, de sorte que les résultats ne soient pas trop biaisés par l'approximation du NFA.

TAB. 2.3 – *Temps de calculs moyen pour la mise en correspondance de $N_Q = N_C = 10^3$ descripteurs de taille $M = 9$ et $N = 12$, avec la mesure de dissimilarité D_{CEMD}.*

Critère utilisé	Temps moyen d'exécution en secondes
AC avec $\Delta = 100$	13s
AC avec $\Delta = 50$	10s
AC avec $\Delta = 10$	8s
DT, NN-DR ou NN-DT	7s

Remarque 1 :
Notons que le calcul de convolution peut être rendu plus rapide en utilisant la transformée de Fourier rapide (FFT). Le signal étant supposé périodique, il convient cependant de réaliser un « bourrage de zéros » (*zero-padding*) pour éviter que le calcul de la convolution soit circulaire. Pour cela, l'histogramme empirique de chaque marginale est remplis de zéros pour avoir une taille $M\Delta$. M transformées de Fourier discrètes sont ensuite calculées. Après avoir calculé le produit des M vecteurs de taille $M\Delta$, l'histogramme final est obtenu par calcul de la transformée de Fourier discrète inverse. La complexité est alors en $\mathcal{O}(\, N_Q M^2 \Delta \log(M\Delta)\,)$.

2.2.1.3 Sur le choix du seuil de détection

Les performances du critère AC dépendent du seuil de détection ε, dont nous discutons ici le réglage pratique.

Le modèle de fond est-il valide ? Intéressons-nous tout d'abord à la validité du modèle de fond (défini au paragraphe 2.1.2) qui nous permet de rejeter les fausses alarmes. Ces dernières sont définies comme des correspondances telles que les distances entre les paires d'histogrammes de mêmes indices sont *mutuellement indépendantes*. Comme nous l'avons auparavant souligné, cela signifie que l'on doit s'assurer, par construction des M caractéristiques des descripteurs, que cette hypothèse d'indépendance est bien vérifiée pour de fausses détections.

Prenons le cas extrême où le descripteur est construit à partir d'une seule caractéristique qui est ensuite dupliquée M fois. En testant des correspondances de descripteurs aléatoires (a, b), on observera alors en pratique beaucoup plus de petites distances $D(a, b)$ que le modèle de fond ne le suggère, car il suppose l'indépendance de ces M caractéristiques.

Or, la représentation des images par des descripteurs SIFT se prête bien à ce type d'hypothèse d'indépendance. En effet, le descripteur SIFT est connu pour être très discriminant : il représente à la fois la structure détectée par le détecteur de points d'intérêt mais également le voisinage de cette structure. Cela signifie que les M régions du descripteur SIFT peuvent représenter des structures très différentes.

Pour illustrer cette propriété, nous réalisons une expérience très simple qui consiste à mettre en correspondance des descripteurs SIFT provenant de deux images différentes. Dans une première expérience, les descripteurs SIFT sont calculés à partir de points d'intérêt détectés de manière conventionnelle. La figure 2.3(a) montre les points ($N_Q = 2500$, $N_C = 2800$) obtenus par le détecteur Laplace-Harris, qui correspondent à des structures saillantes de l'image. Le résultat de la mise en correspondance avec le critère AC est donné en figure 2.3(c), où quatre fausses correspondances ont été validées (représentées par des segments verts).

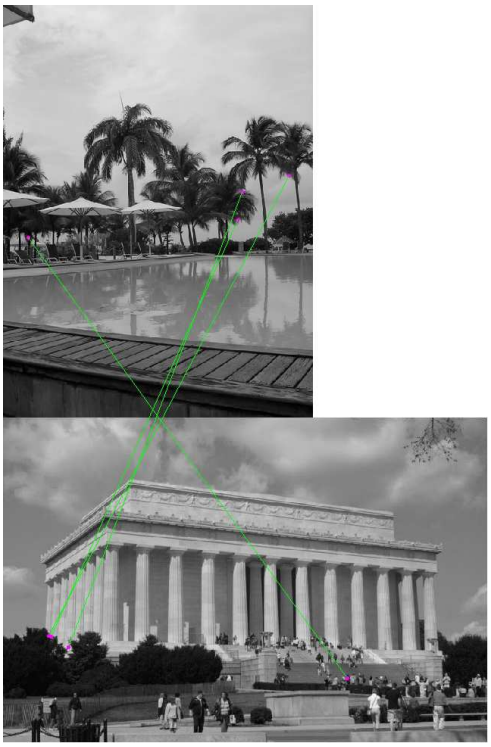
Dans la seconde expérience, les points d'intérêt sont cette fois obtenus par un échantillonnage aléatoire uniforme en espace-échelle dans chacune des images. Une orientation aléatoire leur est également affectée. La figure 2.3(b) montre les points ($N_Q = 2000$, $N_C = 2500$) ainsi obtenus. Les descripteurs SIFT ne sont donc plus calculés à partir d'une structure réelle : une majorité de points sont tirés dans des zones homogènes de chaque image, dont on va extraire M caractéristiques quasiment identiques. La figure 2.3(d) ci-après représente les mises en correspondance obtenues avec le critère AC, en utilisant le même seuil de détection que pour l'expérience précédente. Le nombre d'appariements dépasse cette fois le millier, illustrant le fait que l'hypothèse d'indépendance est trop facilement rejetée. En particulier, on observe un très grand nombre de correspondances entre les points d'intérêt dans les régions homogènes. Au contraire, il n'y a quasiment pas de correspondances avec les zones texturées de la seconde image.

Faut t-il choisir $\varepsilon = 1$? Nous avons vu dans le cadre méthodologique *a contrario* que le seuil ε permet en théorie de contrôler l'espérance du nombre de fausses alarmes. Le seuil de détection est donc usuellement fixé à $\varepsilon = 1$ pour les applications utilisant cette approche *a contrario*.

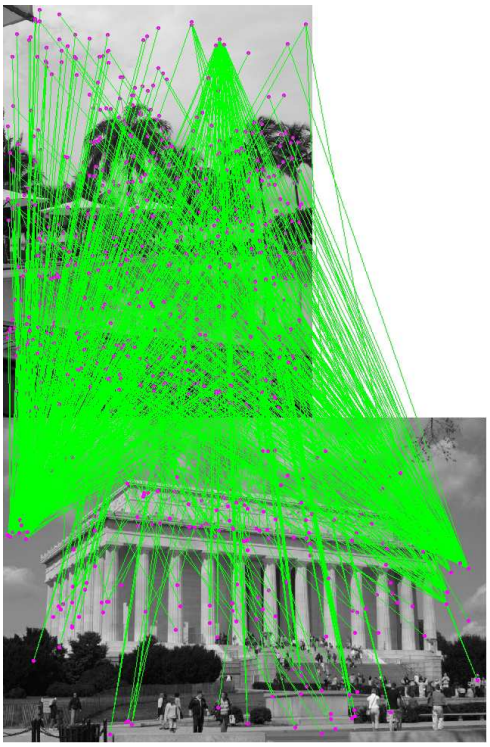
En pratique, nous avons cependant observé que fixer le seuil à $\varepsilon = 10^{-1}$ ou $\varepsilon = 10^{-2}$ donnait des résultats plus satisfaisants. D'après l'analyse asymptotique réalisée au paragraphe 2.1.4.3, l'influence du seuil ε s'exprime sur une échelle logarithmique. En conséquence, le choix de $\varepsilon = 10^{-2}$ modifie

(a) Paire d'images avec points détectés par Laplace-Harris

(b) Paire d'images avec points tirés aléatoirement



(c) Critère AC avec des descripteurs SIFT construits de manière conventionnelle



(d) Critère AC avec des descripteurs SIFT construits en des points d'intérêt aléatoires

FIG. 2.3 – *Illustration de l'importance du détecteur de points d'intérêt pour le critère de correspondance.*

peu le résultat en comparaison de $\varepsilon = 1$, mais le nombre de fausses correspondances validées est plus satisfaisant.

Ce décalage de ε vers des valeurs plus faibles suggère donc que le modèle de fond est plus facilement rejeté qu'il ne le devrait. L'une des hypothèses les plus plausibles pour expliquer ce résultat est la dépendance entre les différentes régions du masque utilisé pour calculer le descripteur SIFT. En effet, nous avons vu qu'il était important de construire le descripteur de manière à rendre ses différentes composantes statistiquement indépendantes, ce que semble permettre la procédure de construction des SIFTs d'après l'expérience précédente. Cependant, nous n'avons pas évoqué jusqu'à présent le fait que les descripteurs sont construits à partir d'une représentation en espace-échelle de l'image, par convolutions successives avec un noyau gaussien de taille croissante. En raison de ce flou gaussien, les orientations du gradient pour des régions limitrophes du masque sont très corrélées. Par voie de conséquence, les M histogrammes construits à partir de régions distinctes ne sont donc pas totalement indépendants, ce qui peut expliquer ce léger décalage que l'on observe pour le nombre de fausses alarmes. Une autre explication pourrait venir du fait que les structures détectées présentent souvent une symétrie [DZM+07]. Dans ce cas, les régions opposées du masque sont potentiellement corrélées.

2.2.2 Analyse expérimentale sur une base de données

Dans cette partie de la section expérimentale, nous allons présenter une analyse de performance des différents critères sur une grande base d'images. Pour évaluer la robustesse d'un algorithme, il est en effet important de le tester sur des données variables et en quantité suffisante pour en observer le comportement « moyen ».

2.2.2.1 Présentation de la base

Nous avons utilisé une base de 732 photographies variées, représentant des scènes extérieures et intérieures, dans différentes conditions d'éclairage (jour, nuit, lumière naturelle ou non), de 800×600 pixels[4]. Quelques images de cette base sont visibles en figure 2.4. Nous avons extrait au total plus de 3.10^6 descripteurs, soit environ 4000 points d'intérêt par image. Cette base a été collectée par nos soins, car il n'existe pas de grande base standard pour évaluer la performance de critères de mise en correspondance. La taille de cette base d'images est du même ordre de grandeur que celle utilisée pour l'étude de [MP07] pour la comparaison de descripteurs locaux, contenant 100 photographies d'objets 3D requêtes ainsi que 535 images indépendantes, pour un total avoisinant 10^5 descripteurs locaux. À titre de comparaison, la base d'images proposée par [MS05] pour l'évaluation comparative de descripteurs locaux ne contient que 8 catégories images. Nous allons en effet illustrer par la suite le fait que les performances d'un critère peuvent varier d'une paire d'images à une autre, mais qu'une large base d'images permet d'en analyser les caractéristiques principales.

2.2.2.2 Protocoles expérimentaux

Afin d'évaluer les différents aspects d'un critère de mise en correspondance, nous avons élaboré différents « protocoles » correspondant à trois situations :

- reconnaissance d'un objet qui est présent exactement une fois dans la base de données ;
- reconnaissance d'un objet présent une fois ou non dans la base de données ;
- reconnaissance multiple d'un objet présent plusieurs fois dans la base de données.

La comparaison des différents critères sur la base d'images se fera en deux temps. Tout d'abord, nous comparerons les critères restreints au plus proche voisin (NN-DT, NN-DR et NN-AC) avec différents protocoles, pour lesquels l'objet d'intérêt apparaît au plus une fois dans la base de données. Ensuite, nous étudierons le comportement des critères de mise en correspondance multiple DT et AC lorsque l'objet apparaît potentiellement plusieurs fois.

Nous allons maintenant détailler les différents protocoles utilisés.

Protocole $\mathbf{A} \rightarrow \mathbf{A}'$ Le premier protocole consiste à mettre en correspondance des points d'intérêt entre une image A et une image A' représentant exactement la même scène mais dans des conditions de prise de vue différentes. Dans le prochain paragraphe est détaillée l'obtention de A' en fonction de A pour définir la vérité-terrain. Ce protocole classique, que l'on intitule $A \rightarrow A'$, est usuellement utilisé pour les études comparatives de descripteurs locaux (voir par exemple [Low04, MS05, MP07]). C'est néanmoins le protocole le plus simpliste dans le cadre d'évaluation des performances d'un critère de mise en correspondance. En effet, il correspond au cas très particulier où l'objet requête de l'image A est exactement présent une fois dans la base d'images. Cette expérience simple va néanmoins nous permettre d'illustrer la robustesse de la mesure de qualité utilisée.

Protocole $\mathbf{A} \rightarrow \{^{\mathbf{A}'}_{\mathbf{B}}$ Afin de se placer dans des conditions d'évaluation plus réalistes, nous proposons une simple extension du protocole précédent. En effet, dans de nombreuses applications, le critère que l'on utilise pour les appariements de descripteurs doit être capable de gérer des situations dans lesquelles l'objet requête n'est pas présent dans certaines images de la base de données (par

[4] Cette base est disponible à l'adresse : http://www.tsi.enst.fr/~rabin/matching/

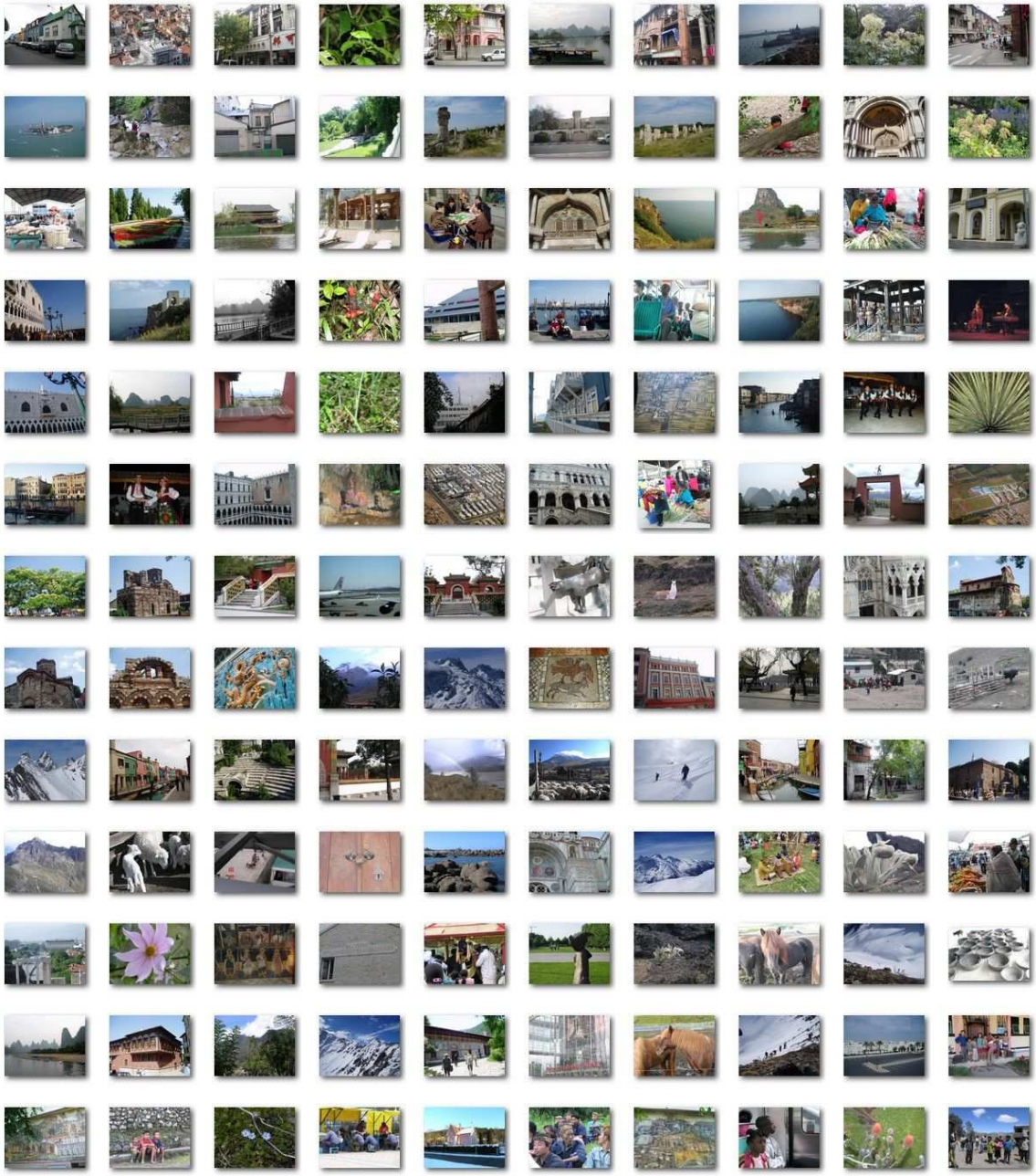


FIG. 2.4 – *Quelques images de la base de 732 photographies de scènes variées.*

exemple, la construction d'une mosaïque à partir de plusieurs images). Pour quantifier ce phénomène, l'image A est cette fois comparée (en plus de l'image A') à une autre image B complètement différente de A, **en utilisant le même seuil de détection**. Ceci permet d'évaluer l'aptitude du critère à ne valider que les bonnes correspondances (vraies-positives) de l'image A, tout en rejetant les correspondances avec B (limitation du nombre de fausses-positives). Le protocole ainsi défini, que l'on désigne symboliquement par $A \rightarrow \{^{A'}_{B}$, est analogue à celui proposé dans [MP07], où l'image requête est comparée à des images complètement différentes dans le but de mesurer la faculté du critère à limiter le nombre de fausses correspondances lorsque l'objet d'intérêt n'est pas présent dans la base.

Une extension du protocole $A \rightarrow \{^{A'}_{B}$ sera également expérimentée, où l'image A est à la fois

comparée à A' et à l'ensemble des images de la base (privée de l'image A). Ce protocole, désigné par $A \rightarrow \{^{A'}_{\text{Base}\setminus A}$, vise à analyser le cas extrême où l'objet requête apparaît très rarement dans la base de données.

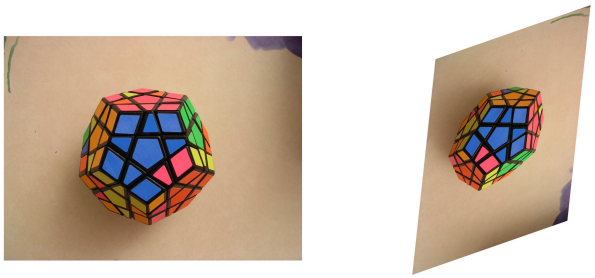
Protocole A $\rightarrow \{^{\mathbf{A'+A''}}_{\mathbf{B}}$ Ce dernier protocole va nous permettre d'évaluer la faculté des critères non restreints au plus proche voisin (AC et DT) à valider les correspondances multiples d'un descripteur dans la base, tout en contrôlant le nombre de fausses détections. Pour réaliser cela, l'image A est comparée à une image comptant deux fois l'objet d'intérêt avec différentes transformations (image notée $A' + A''$), et à une image différente B. On désigne ce protocole par : $A \rightarrow \{^{A'+A''}_{B}$.

2.2.2.3 *Vérité-terrain et évaluation des performances*

Obtention de l'image A' Dans [MS05], des photographies de la même scène sous différents angles de vues sont utilisées dans le but de mesurer la robustesse du descripteur. Dans l'étude de [MP07], une approche similaire est adoptée pour des objets 3D, sur une large base de 100 objets. Dans ces deux cas, la vérité-terrain requiert la calibration de la caméra utilisée, et la connaissance de son mouvement lors des différentes prises de vue, ce qui est très fastidieux pour constituer une large base de données. Or, dans notre cadre d'évaluation, l'objectif n'est pas d'évaluer le pouvoir discriminant de notre descripteur local (de type SIFT) mais d'analyser la robustesse et la fiabilité de notre critère en terme de classification.

Nous avons donc choisi de construire l'image A' à partir d'une dégradation synthétique de l'image A, à l'instar de [Low04]. Ce type de transformation synthétique nous permet, d'une part, d'obtenir aisément une vérité-terrain pour l'évaluation des performances ; cela nous permet également de « perturber » les descripteurs locaux pour simuler les conditions réelles de reconnaissance d'objets. Pour cela, nous appliquons une transformation affine à l'image A, ainsi qu'un bruit additif.

Dans une analyse récente et approfondie de l'invariance du descripteur SIFT, les auteurs de [MY09] ont montré que ce type de descripteur était effectivement invariant à la similitude. Ils ont également montré que les SIFTs étaient robustes pour une classe de transformations affines. Le paramètre critique pour désigner cette classe de transformations est appelé *tilt* (voir en annexe C). Il est ainsi montré expérimentalement que les SIFTs sont grossièrement invariants pour des transformations affines dont le tilt n'excède pas 2. Nous avons également observé expérimentalement une valeur critique du même ordre (voir le chapitre 7 pour une analyse de l'influence du tilt). De manière à perturber les SIFTs, nous avons donc utilisé une transformation affine de tilt égal à 2.5, qui est illustrée par la figure 2.5. Un bruit blanc gaussien est ensuite ajouté, avec un écart-type de $\sigma = 5$ pour des images quantifiées sur 256 niveaux par canal.



A A'

FIG. 2.5 – *Transformation affine utilisée avant l'ajout de bruit.*

Vérité-terrain Pour mesurer les performances, une vérité-terrain est nécessaire pour distinguer les correspondances correctes (vraies-positives) des fausses correspondances (fausses-positives)[5]. Suivant le protocole utilisé par [MS05], une correspondance entre deux points d'intérêt est considérée comme

[5] selon la classification introduite dans le tableau 1.1.

correcte si l'erreur de superposition est plus petite que 50%. Soient R_a et R_b les régions utilisées pour construire les descripteurs a et b dans l'image A. L'erreur de superposition d'une correspondance entre les descripteurs a et b, notée $\mathcal{E}$, est définie comme le rapport entre l'aire d'intersection et l'aire d'union des régions R_a et R_b. En notant $|R|$ l'aire de la région R, l'erreur de superposition s'écrit :

$$\mathcal{E} = 1 - |R_a \cap R_b| / |R_a \cup R_b| \, .$$

> **Remarque 2 :**
> Bien que le procédé ait été repris à de nombreuses reprises dans la littérature, cette définition de la vérité-terrain possède néanmoins quelques limitations. La première est donnée par les auteurs de [MS05] qui remarquent que la définition de l'erreur n'est pas invariante à l'échelle des points d'intérêt comparés. De plus, la définition de l'erreur suppose qu'un bon appariement ne peut correspondre qu'à la détection du même objet physique dans les deux images. Cela signifie que les mises en correspondance avec des structures répétées sont systématiquement considérées comme fausses. Un critère de mise en correspondance tel que le critère NN-DR, qui rejette facilement les appariements de structures répétées, s'en trouve alors avantagé.

Courbes ROC La courbe ROC (*Receiver Operating Characteristic*) est couramment employée pour représenter graphiquement les performances d'une procédure de décision (reconnaissance et détection d'objets, ou encore indexation d'images). Il en existe différentes définitions dans la littérature, suivant les normalisations et les conventions utilisées. Dans le domaine de la reconnaissance et de la détection d'objets, les courbes ROC sont généralement tracées en représentant le *taux de rappel* en fonction du *taux de fausses alarmes*. Si l'on reprend la terminologie empruntée au domaine de la classification (voir le tableau 1.1), le taux de rappel est défini comme la proportion de correspondances correctes sélectionnées (vp) parmi l'ensemble des correspondances correctes existantes (vp+fn). Le taux de fausses alarmes désigne la proportion de correspondances incorrectes sélectionnées (fp) parmi l'ensemble des correspondances sélectionnées (vp+fp). La courbe est tracée en faisant varier le seuil de détection du critère de décision analysé. Si l'on note s le seuil de détection $s \in [0,1]$, $s = 1$ représentant le seuil maximum pour lequel toutes les correspondances sont validées, les taux de fausses alarmes et de rappel sont définis de la manière suivante :

$$\begin{cases} \text{taux de rappel}(s) = & \dfrac{\#\{\, \text{vp}(s) \,\}}{\#\{\, \text{vp}(1) + \text{fn}(1) \,\}} \, , \\ \text{taux de fausses alarmes}(s) = & \dfrac{\#\{\text{fp}(s) \,\}}{\#\{\, \text{fp}(1) + \text{vp}(1) \,\}} \, . \end{cases}$$

La courbe ROC, pour un classifieur idéal, correspond à une droite confondue avec l'axe des ordonnées jusqu'au point de coordonnées $(0,1)$ (toutes les correspondances correctes sont détectées en premier), puis à une droite parallèle à l'axe des abscisses jusqu'au point $(1,1)$ (les fausses correspondances ne sont validées qu'ensuite). Un classifieur aléatoire, au contraire, a une courbe moyenne qui correspond à la ligne médiane passant par l'origine et le point $(1,1)$. Un critère de décision est alors considéré comme d'autant plus performant que sa courbe est proche du classifieur idéal.

Pour chaque critère de mise en correspondance et pour chaque image A de la base d'images, selon les différents protocoles définis précédemment, nous définissons une courbe ROC. Six images de la base sont illustrées en figure 2.6. Pour chacune, deux courbes ROC sont tracées, correspondant aux protocoles $A \to A'$ et $A \to \{^{A'}_{B}$. Sur ces courbes ROC, les critères NN-AC, NN-DT et NN-DR sont respectivement représentés en rouge, bleu et vert. Rappelons que tous les critères sont évalués avec la *même mesure de dissimilarité*, ce qui signifie que le nombre maximal de correspondances correctes sélectionnées est identique pour toutes les méthodes. Comme nous pouvons le constater, les courbes de performance issues d'un même protocole varient beaucoup selon l'image utilisée. Il est donc tout à fait inutile de vouloir conclure en la supériorité d'une méthode à partir de quelques images. C'est la raison pour laquelle nous utilisons une grande base d'images.

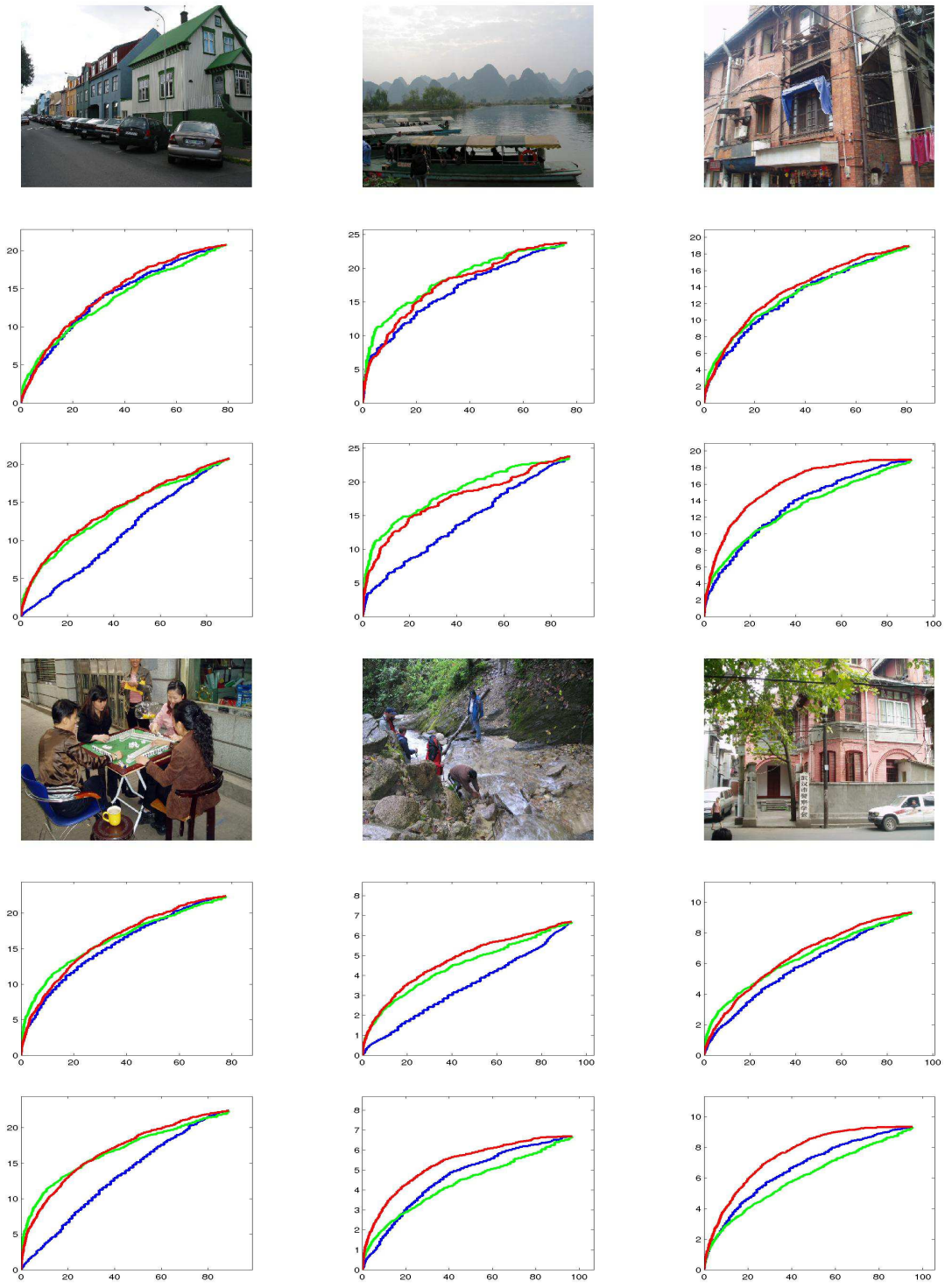


FIG. 2.6 – *Six photographies provenant de la base d'images et leurs courbes ROC, suivant deux protocoles expérimentaux.* *Seuls les critères restreints au plus proche voisin sont ici représentés,* NN-AC *étant tracé en rouge,* NN-DT *en bleu et* NN-DR *en vert. La seconde et la cinquième rangée de courbes correspondent au protocole $A \rightarrow A'$, où l'image A est mise en correspondance avec sa version dégradée A'. Les courbes de la troisième et sixième rangées sont obtenues avec le protocole $A \rightarrow \{^{A'}_{B}$, où l'image A est à la fois comparée à A' et à B, une image différente. Cette comparaison des différentes courbes sur quelques images illustre la variabilité des performances de chaque critère selon l'image utilisée.*

Synthèse par courbes ROC globales Dans le but de faire la synthèse de l'ensemble des courbes ROC obtenues, nous allons par la suite présenter les résultats sur la base par des **courbes ROC globales**, en nous inspirant de [MP07]. Une telle courbe est obtenue en traçant le nombre total de bonnes et de mauvaises correspondances, en utilisant le **même seuil de détection** sur l'ensemble des images de la base. Ce type de courbe de performance, calculée à partir de plusieurs images, présente alors l'intérêt majeur de tester la stabilité d'un seuil de validation d'une image à une autre.

> **Remarque 3 :**
> Comme nous souhaitons par la suite comparer les performances de différents critères, avec et sans restriction au plus proche voisin, nous n'avons pas normalisé les courbes ROC. Nous traçons ainsi directement le nombre de correspondances, au lieu du taux, ce qui ne modifie pas l'allure des courbes. Ce choix va nous permettre par la suite de visualiser les différences de performance lorsque le nombre de mises en correspondance possibles n'est pas le même entre les différents critères comparés.

2.2.2.4 Présentation des résultats

Notre critère de mise en correspondance *a contrario* est comparé en deux temps aux critères présentés dans le chapitre précédent.

Dans un premier temps, nous comparons les critères restreints au plus proche voisin (NN), où chaque descripteur requête peut seulement être mis en correspondance avec le descripteur le plus proche dans la base de données. Il s'agit des critères NN-DT et NN-DR, qui sont les plus utilisés en pratique afin de limiter le nombre de fausses alarmes. Pour que leurs performances puissent être comparées avec notre approche, le critère *a contrario* est utilisé avec une limitation au plus proche voisin (NN-AC). Afin d'évaluer leurs performances dans un cadre où la restriction au plus proche voisin fait sens, nous utilisons les deux protocoles $A \rightarrow A'$ et $A \rightarrow \{^{A'}_{B}$, pour lesquels l'objet apparaît au plus une fois dans la base de données. Ces expériences nous permettent principalement d'illustrer la stabilité du seuil de détection suivant le critère utilisé.

Ensuite, nous étudions le cas plus général correspondant aux critères de mise en correspondance DT et AC sans restriction au plus proche voisin, avec le protocole $A \rightarrow \{^{A'+A''}_{B}$. Le nombre d'appariements potentiellement validés étant beaucoup plus important, cette expérience illustre l'intérêt de notre approche pour le contrôle du nombre de fausses alarmes. Nous allons ensuite revenir sur l'expérience $A \rightarrow \{^{A'}_{B}$ pour montrer la pertinence des seuils automatiques obtenus avec notre méthodologie.

Comparaison des critères de correspondance au plus proche voisin – protocole $A \rightarrow A'$ Rappelons tout d'abord que le critère NN-DR utilise un seuil sur le rapport des mesures de dissimilarité au premier et second plus proches voisins. Le critère NN-DT consiste à utiliser un seuil fixe sur la mesure de dissimilarité pour valider les appariements ; avec notre approche NN-AC, ce seuil est estimé automatiquement pour chaque descripteur requête. Dans ce premier protocole, le nombre de tests utilisés pour estimer la fonction de NFA (équation (2.3)) est exprimé à l'aide de $N_Q = N_C = N_A$, où N_A est le nombre de descripteurs de l'image A.

Les courbes ROC globales pour le protocole $A \rightarrow A'$ sont données en figure 2.7. La courbe en trait continu rouge correspond à notre critère NN-AC, en bleu au critère NN-DT, et en vert au critère NN-DR. Les courbes ROC ont un aspect très « lisse » car elles sont obtenues à partir de millions de mises en correspondance.

La première conclusion que l'on peut tirer de cette expérience est que le critère NN-DT est très instable, car sa courbe ROC est très proche de la ligne médiane. En effet, si le critère NN-DT peut offrir en pratique des performances correctes (cependant en deça des autres critères, comme l'illustre les courbes obtenues sur la figure 2.6), c'est à la condition de choisir le seuil optimal pour chaque paire d'images donnée. Si l'on souhaite utiliser le même seuil sur différents types d'images, on voit que les performances globales sont très mauvaises : ceci explique le succès du critère NN-DR qui offre en comparaison de meilleures performances. Remarquons combien les résultats que l'on obtient sont proches de ceux de l'étude proposée dans [MP07], où un protocole similaire est utilisé. Les auteurs obtiennent des courbes

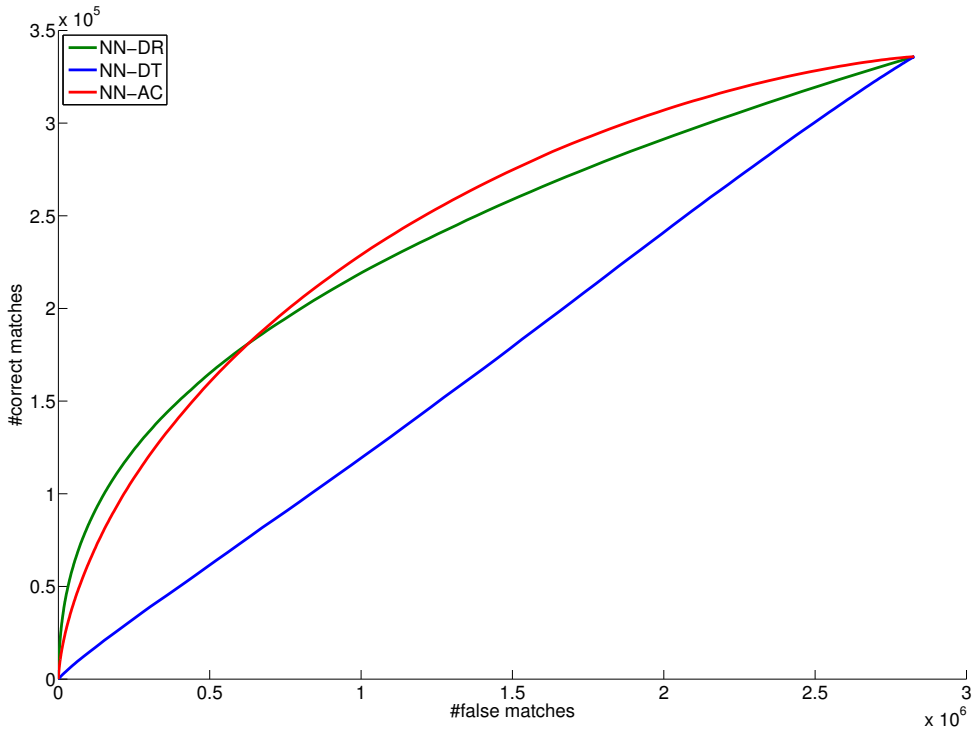
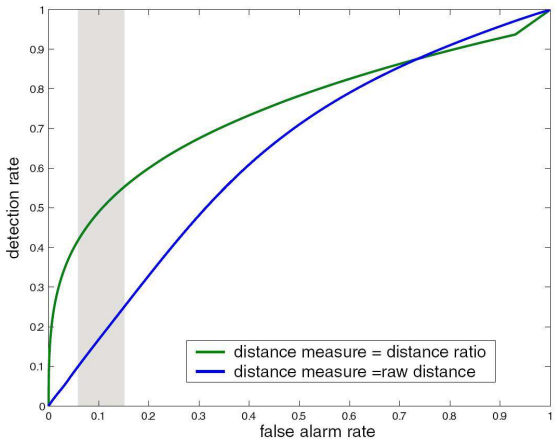


FIG. 2.7 – *Courbes ROC globales obtenues en utilisant le même seuil de détection sur toute la base, avec le protocole $A \rightarrow A'$, où chaque image est mise en correspondance avec sa version dégradée A'. Trois critères sont testés avec la restriction au plus proche voisin :* NN-AC *en rouge,* NN-DT *en bleu et* NN-DR *en vert.*

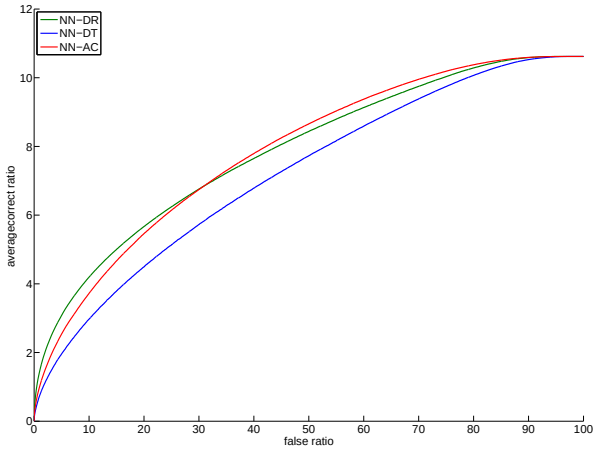
ROC globales qui sont ici reproduites en figure 2.8(a). La courbe du critère NN-DT, tracée en bleu et intitulée « raw distance »), est très proche de la ligne médiane. La courbe ROC globale du critère NN-DR est tracée en vert et intitulée « distance ratio ». Ces deux courbes montrent une fois de plus la supériorité du critère NN-DR par rapport au critère NN-DT en terme de stabilité du seuil de détection. Cette analogie entre nos résultats et les leurs confirme l'intérêt d'une large base d'images pour établir des courbes de performance.

> **Remarque 4 :**
> Le fait que la courbe ROC globale du critère NN-DT soit très proche de la ligne médiane ne signifie pas en pratique que ce critère ne vaut pas mieux que le hasard. Si l'on calcule les courbes ROC moyennes (selon le procédé détaillé au chapitre 7), on obtient les courbes tracées en figure 2.8(b) : on peut constater que le critère NN-DT fait effectivement mieux en moyenne que le hasard. La comparaison avec la courbe globale de la figure 2.7 illustre le degré de variabilité de la mesure de dissimilarité entre descripteurs. Ceci explique le manque de robustesse du seuil sur la distance et l'intérêt d'une procédure de seuillage automatique telle que la nôtre.

La deuxième observation concerne les performances relativement similaires des critères NN-AC et NN-DR. Cela illustre le fait que, même dans le cas limité au plus proche voisin, notre méthode de sélection automatique des seuils est bien plus pertinente qu'un simple seuil fixe. Par ailleurs, notre méthode n'offre pas d'avantages significatifs comparé au critère NN-DR. En effet, dans ce cas de figure très particulier où chaque requête apparaît exactement une fois dans l'autre image, le test réalisé par le critère NN-DR est très bien adapté. Ce test qui consiste à étudier le rapport des distances au premier et au second plus proches voisins peut être interprété comme un test statistique très simple, mais qui a ici du sens. Nous verrons avec le protocole $A \rightarrow \{^{A'}_{B}$ comment se comporte ce test en présence de distracteurs.



(a) Courbes ROC globales provenant de l'étude [MP07]



(b) Courbes ROC moyennes obtenues pour le protocole $A \rightarrow A'$

FIG. 2.8 – *Fig. 2.8(a) : Courbes ROC globales provenant de [MP07] (pour un protocole similaire à $A \rightarrow A'$ sur une large base d'images) pour la comparaison de deux critères :* NN-DT *en trait bleu et* NN-DR *en trait vert. Fig. 2.8(b) : Courbes ROC moyennes obtenues à partir des* 732 *courbes ROC de la base, avec le protocole $A \rightarrow A'$. Trois critères sont testés avec la restriction au plus proche voisin :* NN-AC *en rouge,* NN-DT *en bleu et* NN-DR *en vert.*

> **Remarque 5 :**
> Pour les très faibles seuils de détection, notre critère se comporte légèrement moins bien que le critère NN-DR. Une analyse plus poussée des résultats nous a amenés à constater que, dans ce protocole, les correspondances liées à des structures répétitives sont fortement pénalisées. Prenons l'exemple de la première photographie de la figure 2.6, qui possède certaines répétitions (les fenêtres par exemple). Du fait de la transformation qui permet d'obtenir l'image A', le meilleur candidat pour une structure de A peut correspondre à une répétition de cette structure dans A', ce qui est considéré comme une fausse correspondance en utilisant l'erreur de superposition. Avec notre critère, ce type de correspondance peut avoir une très bonne mesure de qualité car nous n'avons pas d'*a priori* sur la répétition des structures. C'est tout le contraire avec le critère NN-DR, qui associe à ces correspondances une très mauvaise mesure de qualité, rendant plus difficile leur mise en correspondance.

Comparaison des critères de correspondance au plus proche voisin – protocole $A \rightarrow \{^{A'}_{B}$

Dans les applications de vision par ordinateur, on est souvent amené à analyser plusieurs images différentes, où l'objet recherché est absent de certaines images, et caché parmi du « fouillis » (ou *clutter*) dans d'autres. L'extension $A \rightarrow \{^{A'}_{B}$ du protocole précédent va nous permettre d'évaluer la capacité d'un critère de mise en correspondance à gérer ce genre de situations, en proposant un cadre d'expérimentation plus réaliste. L'image A est ainsi mise en correspondance avec une image B complètement différente, en utilisant le même seuil de détection que pour la mise en correspondance avec l'image A'. Les correspondances validées entre A et B étant toutes fausses, leur nombre est ajouté à la quantité de fausses détections obtenues avec le protocole $A \rightarrow A'$. Le nombre de tests du critère *a contrario* est alors défini à partir de $N_Q = N_A$ et $N_C = N_A + N_B$, N_B étant le nombre de descripteurs de l'image B. En faisant varier le seuil de détection et en utilisant toutes les images de la base de données, on obtient ainsi les courbes ROC globales tracées en figure 2.9.

On constate cette fois que les performances du critère NN-DR diminuent clairement, comparé à celles du critère NN-AC : pour un nombre de bonnes correspondances fixé, notre approche donne un nombre de fausses alarmes plus faible que le critère NN-DR. Cela démonte la faculté de la méthode de décision *a contrario* à mieux discriminer les cas où l'objet d'intérêt est présent de ceux où il est absent. Les performances du critère NN-DT, quant à elles, sont quasiment inchangées par rapport à l'expérience précédente.

Afin de confirmer la robustesse de notre approche dans les cas où un descripteur est rarement présent dans la base de données, nous proposons de comparer l'image A à l'ensemble des 731 autres images de notre base d'images. Le nombre de descripteurs candidats testés pour ce protocole $A \rightarrow \{^{A'}_{\text{Base}\setminus A}$ est alors pour chaque descripteur requête de $N_C = \sum N_B \approx 3.10^6$. Comme le nombre de paires d'images est très grand pour réaliser ce test (un demi million environ), nous avons réduit le nombre d'images A à 100 (impliquant la mise en correspondance de plus de 70000 images). La figure 2.10 montre les courbes ROC globales obtenues par ce protocole expérimental.

L'amélioration apportée par le critère *a contrario* vis-à-vis des autres critères, en particulier NN-DR, indique que notre critère est beaucoup plus apte à faire face aux situations où seule une faible portion des descripteurs de la base correspondent à l'objet recherché.

Comparaison des critères de correspondance multiples – protocole $A \rightarrow \{^{A'+A''}_{B}$ Nous considérons cette fois les critères de mise en correspondance *sans restriction sur le nombre de correspondances*. Le protocole $A \rightarrow \{^{A'+A''}_{B}$ reprend le protocole $A \rightarrow \{^{A'}_{B}$ où l'image A' est remplacée par une image $A' + A''$ contenant deux fois l'objet d'intérêt. Cette expérience permet de mesurer l'aptitude d'un critère à valider les mises en correspondances multiples correctes, tout en limitant le nombre de fausses détections. Pour ce protocole, le nombre de descripteurs requêtes est toujours $N_Q = N_A$, et le nombre de descripteurs candidats $N_C = N_B + 2N_A$. Nous utilisons les critères AC et DT qui sont les seuls à permettre ce type de mises en correspondance multiples. Il n'existe pas à notre connaissance d'extension possible du critère NN-DR pour valider des correspondances multiples entre deux images.

La figure 2.11 montre que le critère AC surpasse le critère DT sur la base d'images en termes de contrôle de fausses alarmes. Cela illustre la raison pour laquelle la restriction au plus proche voisin est nécessaire en pratique lorsque l'on utilise le critère de seuil sur la distance. Ce n'est pas le cas de notre critère qui, en définissant des seuils automatiques sur la mesure de dissimilarité, permet de limiter considérablement le nombre de fausses détections. La question à laquelle nous souhaitons maintenant répondre est la suivante : le contrôle du nombre de fausses détections du critère A Contrario est-il suffisamment fiable pour se passer en pratique de la restriction au plus proche voisin ?

La limitation au plus proche voisin est t-elle nécessaire ? Pour répondre à cette question, nous proposons d'utiliser le protocole $A \rightarrow \{^{A'}_{B}$ avec les critères AC et DT pour comparer leurs performances avec et sans restriction au plus proche voisin. La figure 2.12 montre les courbes ROC globales pour les critères AC et DT en trait continu rouge et bleu respectivement. Les courbes correspondant aux mêmes critères lorsqu'ils sont restreints au plus proche voisin[6], NN-AC et NN-DT, sont tracées en trait interrompu rouge et bleu respectivement.

Rappelons que le cas présent correspond à la situation où l'objet recherché apparaît au plus une seule fois dans la base d'images. Comme on pouvait s'y attendre, les performances du critère DT diminuent fortement par rapport à NN-DT, en raison de l'explosion du nombre de fausses alarmes. Par contre, les critères AC et NN-AC donnent des résultat similaires, bien que le critère AC n'ait aucune restriction sur le nombre de correspondances par descripteur requête. En particulier, leurs courbes ROC globales sont quasiment identiques pour les valeurs les plus faibles du paramètre de détection ε. Pour cette plage de valeur de ε, cela signifie que les seuils $\{\delta_i(\varepsilon)\}$ sur la mesure de dissimilarité sont automatiquement définis de manière à ne valider en moyenne que le plus proche voisin parmi les descripteurs candidats. Il est donc superflu de restreindre le nombre de mises en correspondance avec le critère AC.

[6] précédemment vues à la figure 2.9

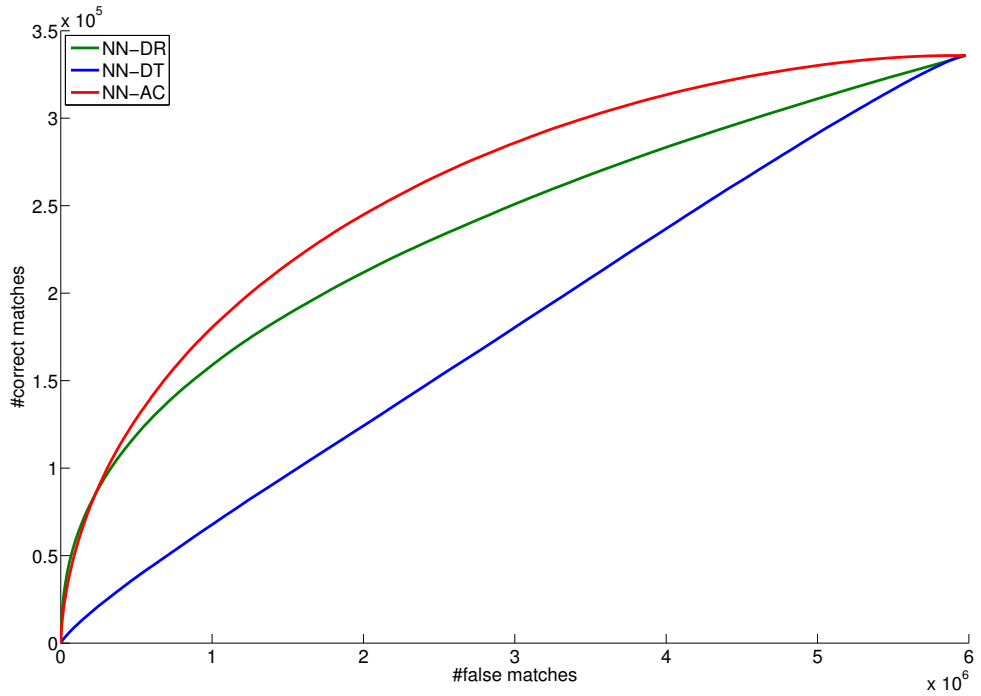


FIG. 2.9 – *Courbes ROC globales obtenues en utilisant le même seuil de détection sur toute la base, avec le protocole* $A \rightarrow \{^{A'}_{B}$*, où chaque image est mise en correspondance avec sa version dégradée* A' *et une autre image* B*. Trois critères sont testés avec la restriction au plus proche voisin :* NN-AC *en rouge,* NN-DT *en bleu et* NN-DR *en vert.*

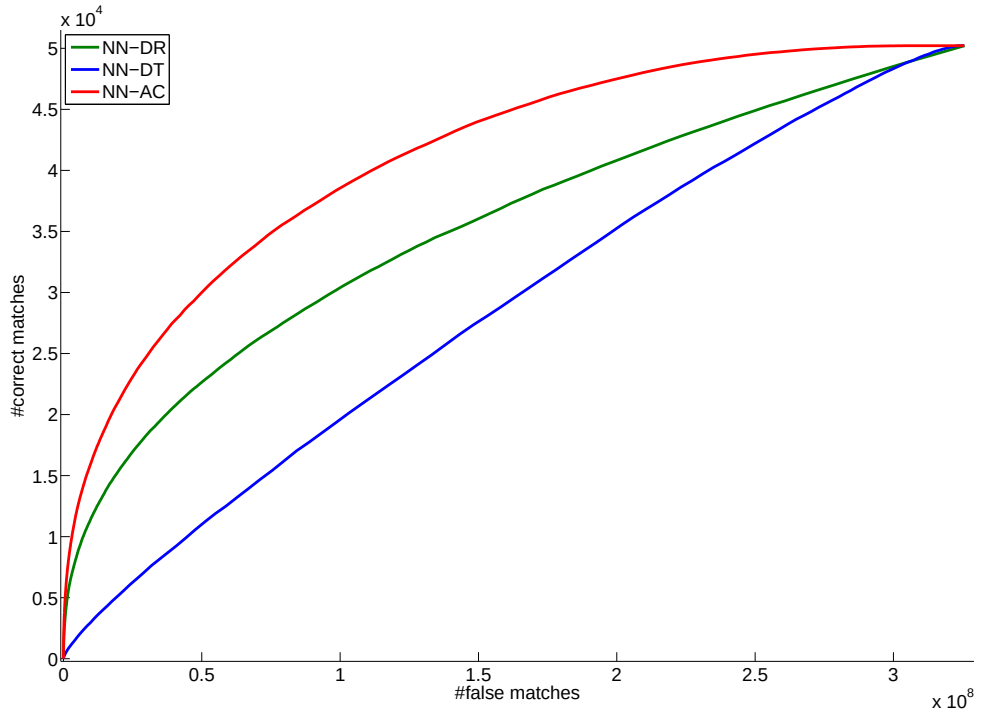


FIG. 2.10 – *Courbes ROC globales obtenues en utilisant le même seuil de détection sur toute la base, avec le protocole* $A \rightarrow \{^{A'}_{Base \setminus A}$ *sur* 100 *images, où chaque image* A *est mise en correspondance avec sa version dégradée* A' *et toutes les autres images de la base de données (soit* 731 *images). Trois critères sont testés avec la restriction au plus proche voisin :* NN-AC *en rouge,* NN-DT *en bleu et* NN-DR *en vert.*

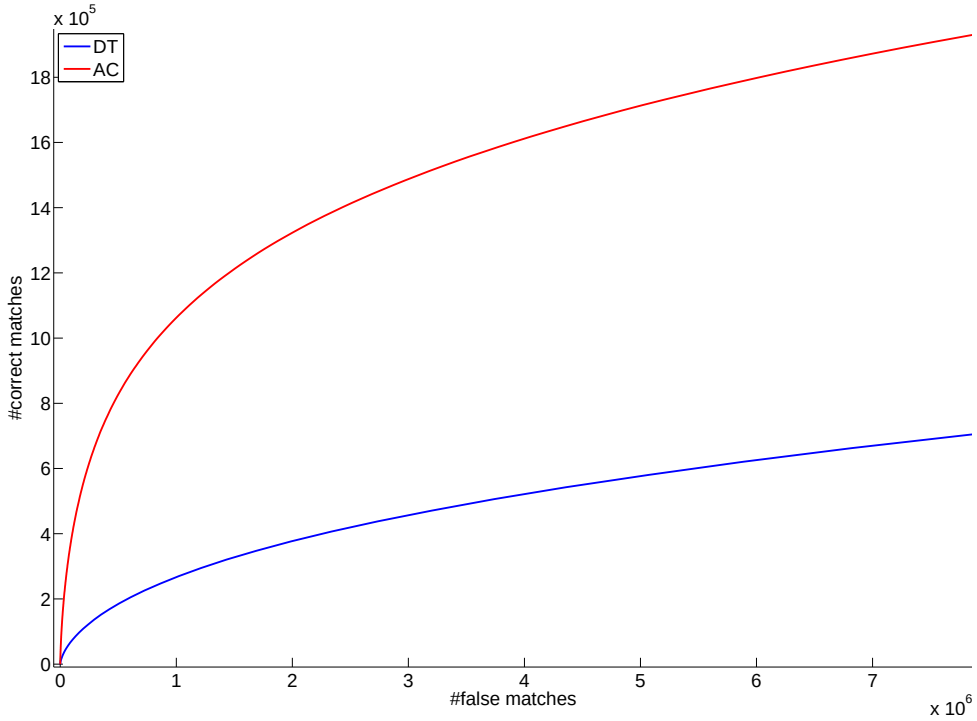


FIG. 2.11 – *Courbes ROC globales obtenues avec le protocole* $A \rightarrow \{_B^{A'+A''}$*, où l'image* $A' + A''$ *contient deux fois l'objet requête. Deux critères sont testés sans restriction au plus proche voisin :* AC *en rouge et* DT *en bleu.*

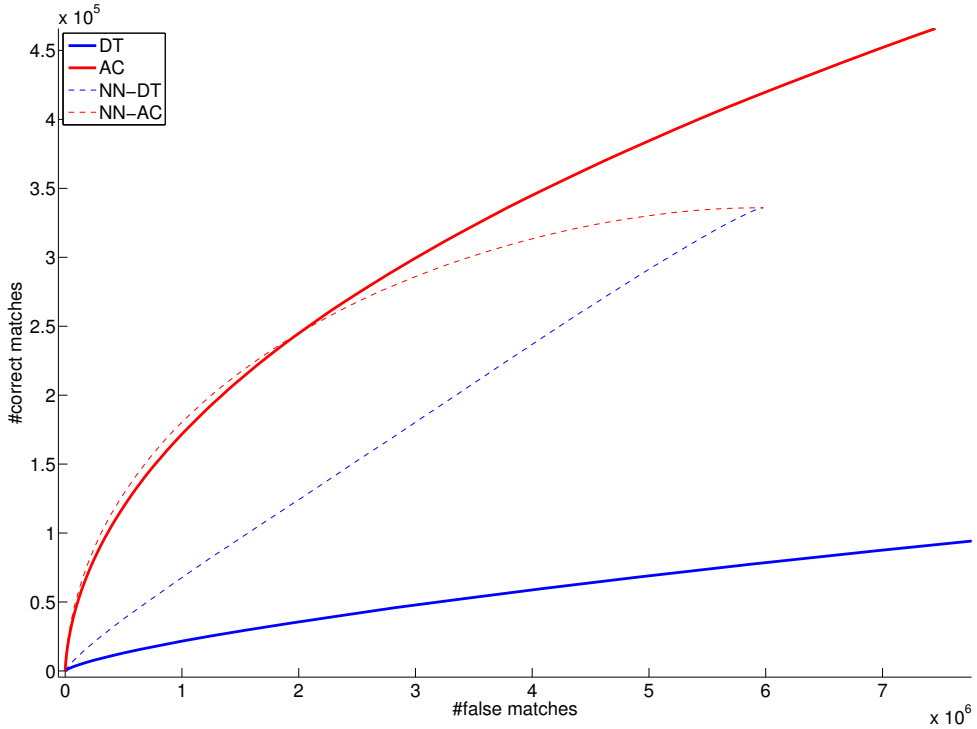


FIG. 2.12 – *Courbes ROC globales obtenues en utilisant le même seuil de détection sur toute la base, avec le protocole* $A \rightarrow \{_B^{A'}$*, où chaque image est mise en correspondance avec sa version dégradée* A' *et une autre image* B*. Deux critères sont testés avec et sans restriction au plus proche voisin. En trait continu, les critères de mise en correspondance multiple :* AC *en rouge et* DT *en bleu. En trait interrompu, les critères restreints au plus proche voisin :* NN-AC *en rouge et* NN-DT *en bleu.*

2.2.3 Autres exemples

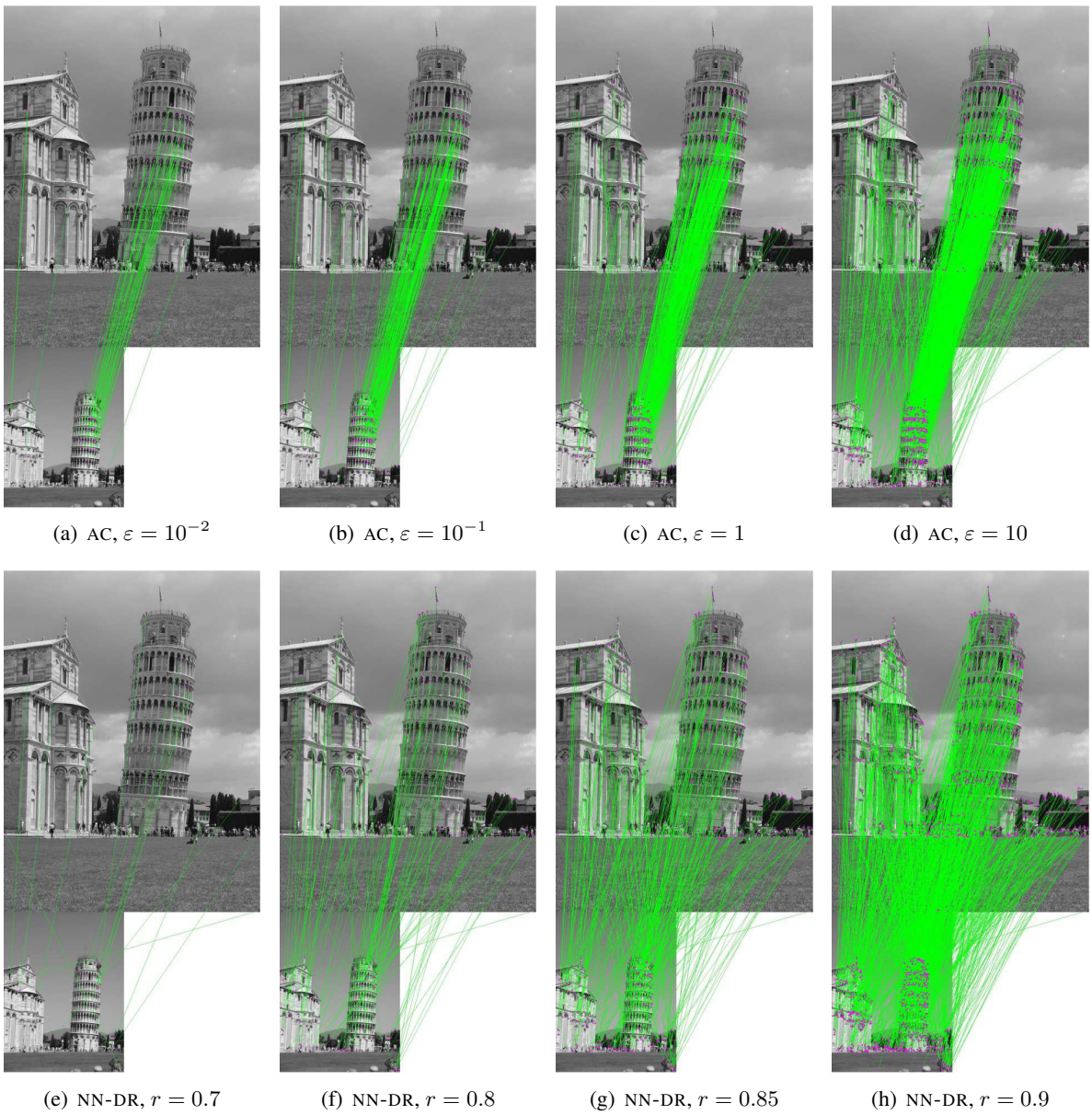
Dans le but d'illustrer de manière visuelle l'intérêt de la méthode proposée, nous présentons dans cette section des expériences supplémentaires sur quelques paires d'images.

Structures auto-similaires Dans l'exemple suivant, nous illustrons le comportement de notre critère lorsque l'on fait varier le seuil de détection, pour une scène contenant de nombreuses structures répétées. Il a été montré dans [ZK06a] que c'est un cas de mise en correspondance particulièrement difficile. La figure 2.13 montre les résultats de l'appariement de deux photographies de la tour de Pise (dans des conditions d'éclairement et des prises de vue différentes) avec les critères NN-DR et AC. La seconde rangée d'images (figures 2.13(e), 2.13(f), 2.13(g) et 2.13(h)) montre les résultats obtenus par le critère NN-DR avec les seuils $r = 0.7$, $r = 0.8$, $r = 0.85$ et $r = 0.9$ respectivement. La proportion de correspondances correctes est visiblement très faible quel que soit le seuil utilisé. Il est en effet très difficile de mettre en correspondance des structures répétées avec un tel critère qui suppose que l'objet d'intérêt apparaît au plus une seule fois dans la base de données. Lorsque l'objet recherché est texturé ou présente une forte auto-similarité, un descripteur requête peut avoir plusieurs descripteurs candidats très similaires, ce qui contredit l'hypothèse sur laquelle repose le critère NN-DR.

Au contraire, le critère de correspondance AC – qui n'est pas restreint au plus proche voisin – donne des mises en correspondance multiples entre les colonnes et les arches similaires. De plus, en observant les résultats pour différentes valeurs de seuil (figures 2.13(a), 2.13(b), 2.13(c) et 2.13(d)), on peut constater que la proportion de fausses détections est très limitée jusqu'à $\varepsilon = 1$. En augmentant ε de 10^{-2} à $\varepsilon = 1$, on obtient ainsi de plus en plus de correspondances entre les structures similaires de la tour, sans que le nombre de fausses détections explose pour autant. Bien entendu, tous ces appariements ne sont pas corrects au sens où ils ne correspondent pas physiquement à la même structure, mais ils sont cependant satisfaisants car représentatifs d'une réelle auto-similarité. Lorsque ε est plus grand que 1, le nombre de fausses détections commence à devenir très important, illustrant le fait que $\varepsilon = 1$ est bien le seuil critique défini par la théorie. Cependant, comme nous l'avons auparavant remarqué, il y a un décalage pratique entre l'ordre de grandeur du nombre observé de fausses détections et l'ordre de grandeur de ε : si l'on souhaite valider très peu de fausses détections, les valeurs $\varepsilon = 10^{-2}$ ou $\varepsilon = 10^{-1}$ sont généralement plus adéquates.

Occurences multiples Nous avons illustré dans l'exemple précédent l'intérêt de notre approche pour la mise en correspondance d'objets ayant des structures répétées. Un autre cas de figure où le critère AC présente un atout considérable est celui des objets apparaissant plusieurs fois dans la base de données. Dans l'expérience 2.14, nous utilisons la photographie de la figure 2.14(a) comme image requête ; elle représente une canette de soda dont le logo apparaît plusieurs fois (28 exactement) dans la seconde photographie (figure 2.14(a)). Avec le critère AC (figure 2.14(c)), de nombreuses mises en correspondance multiples sont automatiquement validées avec le seuil $\varepsilon = 10^{-1}$, tout en contrôlant le nombre de fausses correspondances (seuls quelques appariements incorrects entre des canettes différentes sont visiblement obtenus). C'est tout le contraire du critère NN-DR, qui ne valide que très peu de correspondances correctes avec un seuil à $r = 0.8$ (figure 2.14(d)).

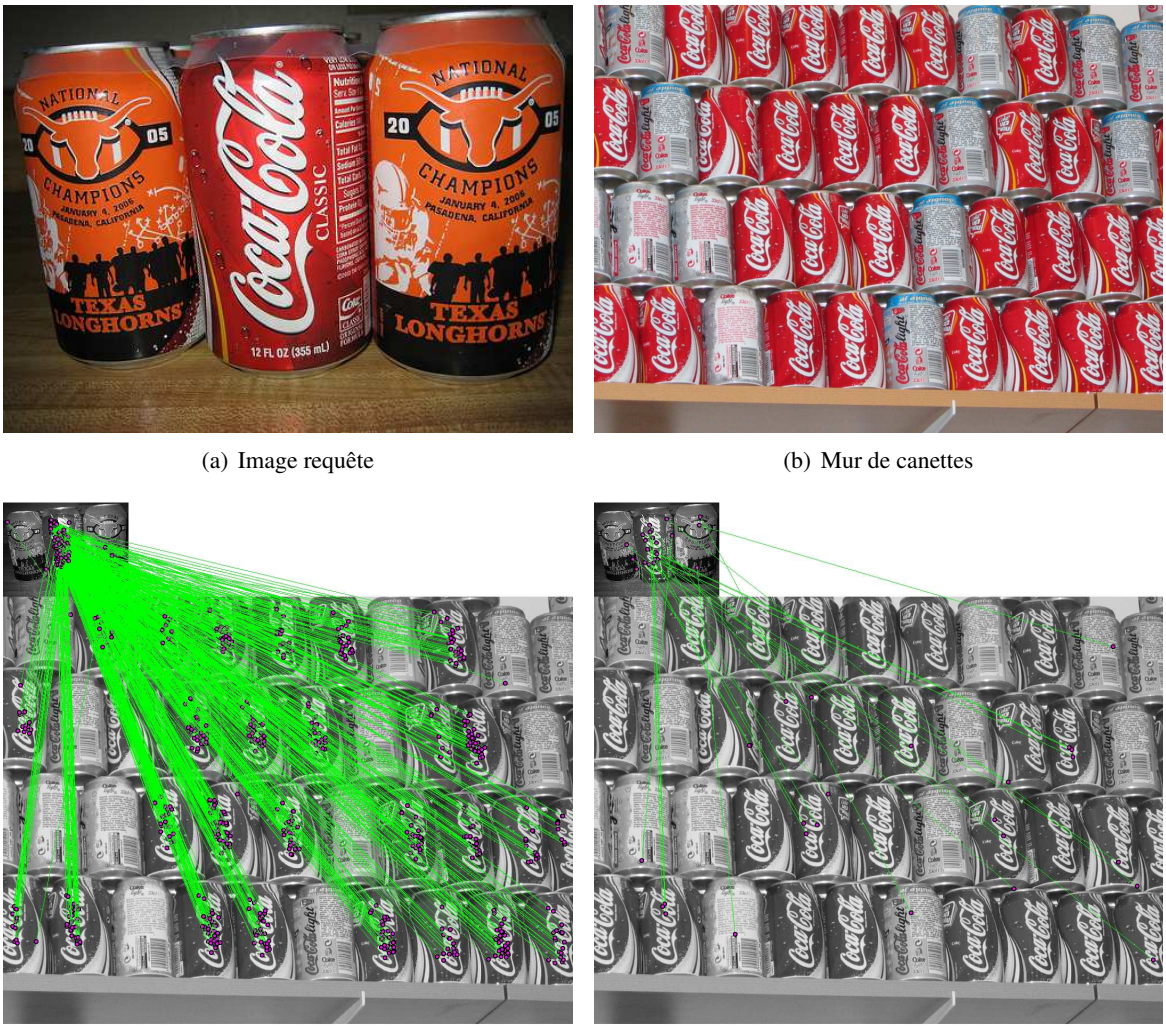
Robustesse du seuil de validation L'exemple suivant va nous permettre d'illustrer les différents points vus au moment de l'analyse des performances sur une grande base de données à l'aide de courbes ROC. Cette fois, une image requête est appariée avec chacune des 8 images d'une base de données, où l'objet requête (une boîte de conserve) apparaît une seule fois, plusieurs fois, ou pas du tout selon les images. Les neuf photographies utilisées pour cette expérience sont données en figure 2.15(a). L'image requête, avec un cadre bleu, est au centre des 8 autres images. Les 4 photographies dans les coins n'ont rien de commun avec l'image centrale. Les 4 autres images représentent la même boîte de conserve que celle présente dans l'image centrale, en présence de fouillis et dans des conditions de prise de vue et d'éclairement différentes. La boîte apparaît ainsi une unique fois dans les images immédiatement à



(a) AC, $\varepsilon = 10^{-2}$ (b) AC, $\varepsilon = 10^{-1}$ (c) AC, $\varepsilon = 1$ (d) AC, $\varepsilon = 10$

(e) NN-DR, $r = 0.7$ (f) NN-DR, $r = 0.8$ (g) NN-DR, $r = 0.85$ (h) NN-DR, $r = 0.9$

FIG. 2.13 – *Mise en correspondance d'un objet avec des structures répétitives : la tour de Pise.* *Dans cet exemple, deux critères de mise en correspondance de descripteurs locaux de type SIFT sont utilisés avec différents seuils de détection. Les traits verts entre les deux photographies représentent les correspondances validées pour le seuil et le critère indiqué en légende. La première rangée illustre le résultat obtenu à l'aide du critère de mise en correspondance A Contrario (noté* AC*) introduit dans ce chapitre, sans aucune restriction sur le nombre d'appariements par requête. Les mises en correspondances entre structures répétées (les colonnes et les arches de la tour en particulier) sont de plus en plus nombreuses en augmentant le seuil de significativité ε, mais le nombre de fausses correspondances entre des objets différents reste limité. La seconde rangée montre les résultats du critère* NN-DR *qui restreint chaque descripteur requête à son plus proche voisin. Il existe très peu de correspondances correctes entre les deux vues de la tour en raison de la répétition des structures.*



(a) Image requête (b) Mur de canettes

(c) AC avec $\varepsilon = 10^{-1}$ (d) NN-DR avec $r = 0.8$

FIG. 2.14 – ***Mise en correspondance d'un objet avec des occurrences multiples : canette de soda.*** *(Photographies de Frédéric Sur). Les deux photographies de la première rangée représentent l'image requête (2.14(a)) et l'image utilisée comme de base de données (2.14(b)). Le logo de la canette de soda au centre de la première image apparaît de manière similaire* 28 *fois dans la seconde image. Le critère* AC *valide automatiquement les mises en correspondances multiples correspondant au même objet, tout en contrôlant le nombre de fausses détections. À cause de ces occurrences multiples, le critère* NN-DR *ne permet pas d'avoir autant de bonnes mises en correspondances, et le taux de fausses détections est très élevé (plus de* 50%*).*

gauche et à droite de l'image centrale, et elle est présente trois fois dans les images au dessus et en dessous.

Les mises en correspondance validées par le critère AC en utilisant la mesure de dissimilarité D_{CEMD} sont illustrées en figure 2.15(b) selon la représentation usuelle (traits verts tracés entre les points d'intérêt appariés). Les figures 2.16(a) et 2.16(b) montrent respectivement les résultats des critères NN-DR et NN-DT avec la distance euclidienne, afin de montrer le type de résultat obtenu avec ces approches classiques. Dans le but de montrer le degré de robustesse de chacune des méthodes de correspondance testées, nous avons fixé les seuils de détections de manière à obtenir le même nombre d'appariements corrects entre l'image requête et la photographie à sa gauche (soit une soixantaine environ). On obtient ainsi les seuils de détection suivants pour l'ensemble des appariements d'images : $\varepsilon = 10^{-2}$ pour AC, $r = 0.8$ pour NN-DR, et $t = 0.45$ pour NN-DT).

On constate que notre méthode d'estimation automatique des seuils de validations sur la mesure de dissimilarité réduit le nombre de fausses détections (que l'objet d'intérêt soit présent ou non), tout en autorisant les mises en correspondances multiples pour les objets apparaissant plusieurs fois. De plus, nous pouvons constater une fois de plus que la limitation au plus proche voisin n'est pas nécessaire. En comparaison, le critère NN-DT utilisant un seuil fixe sur la mesure de dissimilarité entre descripteurs locaux donne des résultats très différents d'une image à une autre. Le nombre de fausses détections avec ce critère est plus élevé, malgré la restriction sur le nombre de correspondances par point d'intérêt de l'image requête. Le critère NN-DR, quant à lui, donne des résultats plus robustes sur la base par rapport au critère NN-DT. Le nombre de fausses détections est souvent plus réduit, mais lorsque l'objet d'intérêt apparaît plus d'une fois il y a alors très peu de correspondances correctes.

(a) Image requête au centre avec 8 images

(b) Mesure de dissimilarité D_{CEMD} et critère de mise en correspondance AC avec $\varepsilon = 10^{-2}$

FIG. 2.15 – ***Mise en correspondance de plusieurs images avec le même seuil de validation.*** *(*figure du haut*) L'image requête, au centre avec un cadre bleu, est mise en correspondance avec* 8 *autres images en utilisant le même critère et le même seuil de détection (voir le texte pour plus de détails). (*figure du bas*) Critère de mise en correspondance A Contrario, fondé sur la mesure de dissimilarité* D_{CEMD}.

(a) Distance euclidienne et critère de mise en correspondance NN-DR avec $r = 0.8$

(b) Distance euclidienne et critère de mise en correspondance NN-DT avec $t = 0.45$

FIG. 2.16 – *Mise en correspondance de plusieurs images avec le même seuil de validation.* (figure du haut) *Critère de mise en correspondance* NN-DR*, utilisant la distance euclidienne.* (figure du bas) *Critère de mise en correspondance* NN-DT*, utilisant la distance euclidienne.*

Chapitre 3

Groupement de mises en correspondance : problématique et état de l'art.

Nous avons pris le parti dans cette thèse, ce qui est classique, de réaliser la reconnaissance d'objets à partir d'une représentation locale des images. Dans les deux précédents chapitres, une approche pour extraire puis mettre en correspondance des points d'intérêt entre différentes images a été présentée. La dernière étape du processus de détection, que nous allons maintenant considérer, consiste à exploiter ces appariements de points pour détecter puis estimer la pose d'un objet entre plusieurs vues. Ce cadre de travail nécessite un large éventail d'invariances à divers phénomènes, tels que le changement de point de vue, les conditions d'éclairement, ou encore l'occultation partielle de l'objet recherché. De nombreuses autres contraintes sont à prendre en compte pour définir une méthode robuste de détection et d'estimation.

Nous allons tout d'abord introduire dans la section 3.1 la problématique liée à la détection et à l'estimation de la pose d'un objet. Après en avoir étudié les contraintes, nous définirons les objectifs auxquels nous souhaitons répondre.

Dans les sections suivantes (§ 3.2 et 3.3) seront étudiées les méthodes existantes pour détecter différents objets à partir de correspondances de points d'intérêt, ainsi que pour estimer leur pose respective de manière robuste. En particulier, nous analyserons leurs avantages respectifs ainsi que leurs limitations.

3.1 Problématique

3.1.1 Exemple de reconnaissance d'un objet

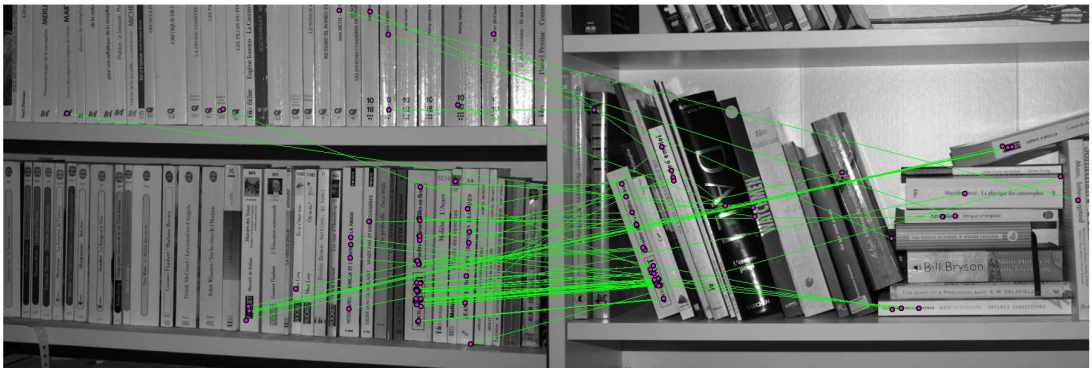
FIG. 3.1 – *Y a t-il un livre commun à ces deux bibliothèques ?*

Nous introduisons notre problématique avec l'exemple de la figure 3.1, représentant deux biblio-

thèques différentes. Étant donné une paire d'images de contenu quelconque, la reconnaissance d'objets consiste à répondre à la question suivante : existe-il un objet en commun entre ces deux photographies ? Dans l'affirmative, la seconde question à laquelle on souhaite répondre est : où se trouve-t-il ?

À l'aide du processus de mise en correspondance de points d'intérêt étudié au chapitre précédent, nous sommes capables d'apporter un début de réponse à ces questions. La figure 3.2(a) nous montre quels sont les points d'intérêt qui sont détectés et mis en correspondance avec la procédure décrite au deux précédents chapitres. De tels appariements sont représentés par des lignes vertes[1]. En raison du type de motif qui est ici répété (lettre de l'alphabet), de nombreux appariements ne correspondent pas physiquement au même objet entre les deux scènes.

On observe qu'une partie des points d'intérêt appariés correspondent à un même livre. La détection de cet objet repose alors sur l'identification d'un groupe de mises en correspondance *cohérentes avec une transformation géométrique* entre les deux images, ainsi que l'illustre la figure 3.2(b). Ceci permet de répondre à la question « existe-t-il un objet en commun entre ces deux photographies ? ». Au contraire, tout groupe de fausses correspondances doit être rejeté, de manière à pouvoir affirmer que les deux photographies n'ont pas d'autres objets en commun et à ne pas provoquer de fausses reconnaissances d'objets.



(a) Mise en correspondance de points d'intérêt.

(b) Détection d'un groupe de correspondances correspondant à un même objet.

FIG. 3.2 – ***Illustration de la reconnaissance d'objets par groupement de mises en correspondance de points d'intérêt.***

L'ensemble des correspondances du groupe détecté ont la particularité de suivre approximativement une même *transformation géométrique*. À partir d'un modèle géométrique, il est alors possible d'estimer directement la position de l'objet dans chacune des images, en utilisant les mises en correspondance sélectionnées. Dans notre exemple, le recalage des deux photographies ainsi obtenu est illustré en figure 3.3, pour une similitude. La qualité de cette estimation est grandement conditionnée par le nombre de correspondances correctes sélectionnées, ainsi que par la précision avec laquelle les points d'intérêt

[1] voir la version couleur de ce manuscrit

sont définis. Lorsque le modèle géométrique est inconnu, on doit être capable de pouvoir choisir quel est le modèle le plus approprié : est-ce une simple translation, une rotation de la caméra, ou une rotation 3D de l'objet ?

FIG. 3.3 – *Recalage de l'objet en commun dans les deux images (vue superposée des deux images).*

3.1.2 Objectifs

Cet exemple simple nous a permis de mettre en lumière deux aspects cruciaux de la reconnaissance d'objets :

1. **La détection :** Il s'agit d'être capable de décider si un objet est présent ou non entre plusieurs images, en identifiant un sous-ensemble de correspondances correctes.
2. **L'estimation :** La définition de la pose de l'objet détecté entres différentes scènes repose sur une estimation robuste de la transformation géométrique entre les points d'intérêt appariés.

Nous avons vu que ces deux tâches était réalisées simultanément et soumises à plusieurs contraintes, dont voici à présent une liste plus détaillée :

- **Robustesse aux fausses correspondances :** il s'agit non seulement de ne pas valider de groupes de fausses mises en correspondance, mais également, lorsque un même objet est présent dans les deux images, de ne pas valider un groupe hétérogène constitué à la fois de bonnes et de fausses correspondances ;
- **Robustesse à l'erreur sur la position des points :** les points d'intérêt qui sont appariés correspondent à des structures saillantes de l'objet dont l'aspect dépend de sa nature et des conditions d'éclairement. En outre, ces points sont obtenus au terme d'une chaîne de traitement, depuis l'acquisition de l'image jusqu'à leur détection. Leur position est donc entachée d'erreurs qui vont rendre d'autant plus difficile l'estimation de la pose de l'objet ;
- **Détection d'objets multiples :** Nous avons jusqu'à présent seulement considéré la détection d'un unique objet entre plusieurs images. Dans un cadre plus général, on souhaite pouvoir identifier plusieurs objets, ce qui nous amènera à considérer les méthodes de groupement multiple. Un cas particulier que nous envisageons également est la situation où un même objet peut être présent plusieurs fois dans une image.
- **Sélection du modèle géométrique de la pose des objets :** La pose de l'objet dépend de sa position tridimensionnelle vis-à-vis de la caméra, ainsi que du calibrage de celle-ci, pour chacune des prises de vue [2]. Cette différence de pose entre chaque image peut être modélisée par une transformation géométrique, mais dont la nature est *a priori* inconnue. Pour obtenir la meilleure interprétation

[2]Lorsque l'objet n'est pas rigide, sa pose dépend également de la déformation qu'il a subie.

possible, on doit alors à la fois choisir le modèle géométrique le plus adapté et estimer quels en sont les paramètres. On parle alors de *sélection de modèles*.

– **Connaissances *a priori* :** Nous verrons que la plupart des méthodes proposées dans la littérature reposent sur certains *a priori* (par exemple la loi de distribution des erreurs, le taux de fausses correspondances, ou encore le nombre d'objets à détecter). Afin d'être le plus générique possible, un système de reconnaissance d'objets ne doit reposer sur aucune connaissance *a priori*.

Nous proposons dans cette thèse une méthode visant à répondre à l'ensemble de ces objectifs. Cette méthode sera présentée au chapitre 4. Auparavant, nous allons formaliser la problématique de la reconnaissance d'objets en termes de groupement de points et de sélection de modèles. Ensuite, nous rappellerons les méthodes existantes pour traiter certains des objectifs qui viennent d'être présentés.

3.1.3 Notations

3.1.3.1 Cadre de travail

Nous travaillons avec des descripteurs locaux de type SIFT. Ces descripteurs offrent un certain nombre d'invariances (en particulier au zoom et à la rotation), et une robustesse au bruit et au changement de contraste affine (voir l'annexe B). On suppose en outre que les objets recherchés sont rigides, ce qui nous permet de restreindre le nombre de modèles géométriques envisageables pour la détection et l'estimation de la pose des objets.

Aucune autre connaissance n'est requise *a priori* (calibration et mouvement de la caméra, présence ou non d'un objet, nombre d'objets à détecter, pose de l'objet, caractéristiques de l'erreur sur la position des points d'intérêt *etc*.).

Sans perdre en généralité, nous considérons dorénavant la reconnaissance d'objets pour une paire d'images notées I et I'. Un critère de mise en correspondance (voir le chapitre 2) est utilisé pour sélectionner des appariements de points entre ces deux images, I étant l'image requête et I' l'image considérée comme la base de recherche. On note $\mathcal{C} = \{(m_i, m'_i),\ i = 1, \ldots, N\}$ l'ensemble des N mises en correspondance obtenues avec ce critère, où $\{m_i\}$ et $\{m'_i\}$ sont respectivement les points d'intérêt de l'image I et I'. Nous avons vu au chapitre précédent qu'un même point d'intérêt dans une image pouvait être mis en correspondance avec plusieurs points de l'autre image, ce que l'on qualifie de *mises en correspondance multiples*. Par conséquent, des points d'intérêt d'indices différents dans l'image I ou I' peuvent désigner un même point. Les descripteurs SIFT ayant potentiellement plusieurs orientations, il est également possible que deux correspondances d'indices différents représentent le même appariement : $(m_i, m'_i) = (m_j, m'_j)$ avec $i \neq j$.

3.1.3.2 Transformations considérées

L'objet présent dans les deux images I et I' étant supposé rigide, deux classes de transformations peuvent être utilisées pour modéliser le changement de son apparence : les transformations planes (isométrie, similitude, transformation affine, et homographie), et la géométrie épipolaire. On écarte ici les autres types de transformations liées aux défauts de la caméra : transformation radiale (distorsion en barillet ou en coussinet) et dispersion chromatique liées à la lentille, transformation liée à la disposition des capteurs sur une grille non régulière *etc*.

Soient m et m' des points de I et I' respectivement, représentant un même point de l'objet observé selon deux vues différentes. La relation entre les points m et m' dépend de la nature de l'objet (plan ou tridimensionnel), du mouvement relatif de l'objet vis-à-vis de la caméra entre les deux vues, ainsi que des paramètres internes de la caméra (en particulier, de la distance focale). L'expression de cette relation selon ces différents cas est rappelée en annexe C.

Rappelons toutefois que, lorsque l'objet est plan, son changement d'apparence entre les deux vues est décrit par une transformation plane, dont la forme la plus générale est l'homographie (ou géométrie projective). La relation entre les points m et m' est de la forme

$$m' = \mathcal{T} m\,, \tag{3.1}$$

où chacun des points est exprimé en coordonnées homogènes, c'est-à-dire $m = [x_m, y_m, 1]^T$ avec (x_m, y_m) les coordonnées du point m dans l'image I. $\mathcal{T}$ est une matrice 3×3 de p paramètres indépendants, avec $p = 4$ pour une similitude, $p = 6$ pour une transformation affine, et enfin $p = 8$ pour une homographie.

Le calcul des p paramètres de la transformation $\mathcal{T}$ requiert un groupe de correspondances que l'on désigne par la notation S'. Le cardinal de ce groupe est noté n. Ainsi, $n = p/2$ correspondances de points (m_i, m'_i) différentes sont requises pour estimer une unique transformation plane. Dans le cas de la transformation affine, on doit vérifier que les $n = 3$ points dans chaque image ne sont pas alignés ; avec l'homographie, c'est chacune des 4 combinaisons de triplet parmi les $n = 4$ points qui doivent être non-colinéaires.

Lorsque l'objet n'est pas plan, il existe certaines configurations particulières où une transformation plane peut décrire la transformation subie par l'objet. Dans le cas général cependant, la relation entre les points d'intérêt est décrite par la matrice fondamentale que nous noterons F. En utilisant les coordonnées homogènes, elle s'exprime ainsi

$$m'^T F m = 0\,. \tag{3.2}$$

Contrairement aux transformations planes, l'image du point $m \in I$ dans l'image I' est une droite paramétrée par Fm, que l'on appelle *ligne épipolaire*. L'ensemble des lignes épipolaires décrit un faisceau qui passe par un unique point, appelé épipole. Cet épipole représente l'image du centre de la première caméra (ayant capturé la vue I), par la seconde caméra. La matrice F étant définie à un facteur d'échelle près, en raison de la projection d'un point 3D sur le plan focal de la caméra, il faut un groupe S' de $n = 8$ points pour la définir de manière unique (sauf configuration dégénérée des points d'intérêt).

Le problème de la sélection de modèles sera étudié en détail en section 3.3. Il consiste dans notre cadre de travail à sélectionner parmi la similitude, la transformation affine, la géométrie projective et épipolaire, le modèle géométrique qui est le plus approprié pour expliquer la scène.

3.1.3.3 Évaluation de la qualité d'une transformation

Considérons pour l'instant le cas particulier des transformations planes $\mathcal{T}$. On note S l'ensemble des correspondances $\{(m_i, m'_i)\}$ entre des points d'un même objet. En raison des erreurs entachant la position estimée des points d'intérêt, il n'existe pas de transformation exacte permettant de vérifier l'expression (3.1) pour l'ensemble des correspondances S.

Autrement dit, pour n'importe quelle transformation $\mathcal{T}$ donnée, il existe un écart entre l'image $\mathcal{T}m$ du point m dans l'image I' et le point m' qui lui correspond. Réciproquement, dans l'image I, les points $\mathcal{T}^{-1}m'$ et m ne coïncident pas. On appelle cet écart illustré par la figure 3.4 *erreur résiduelle* ou résidu. Tous les estimateurs que nous étudierons dans la section 3.2 se basent sur cette notion de résidu. Une transformation sera jugée d'autant meilleure qu'elle minimise l'ensemble des erreurs résiduelles.

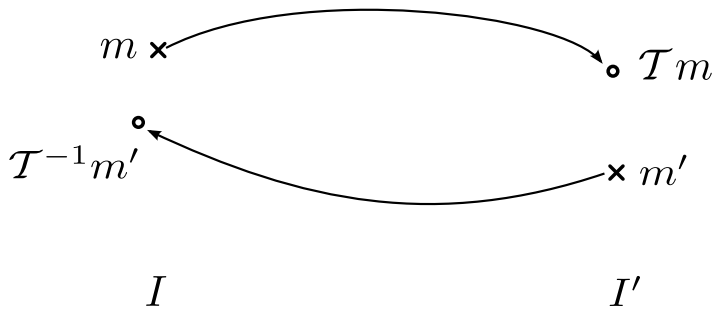


FIG. 3.4 – *Illustration de l'erreur résiduelle.*

Il existe plusieurs façon de mesurer les erreurs résiduelles de la transformation $\mathcal{T}$ sur les correspondances S. L'approche la plus intuitive consiste à regarder l'erreur résiduelle *géométrique* – encore

appelée erreur de transfert – qui est définie par la distance euclidienne entre les couples de points $\mathcal{T}m$ et m'. Plus généralement, on appelle *erreur de transfert symétrique* la mesure de l'écart résiduel qui dépend des erreurs de transferts calculées dans chacune des images (en coordonnées non homogènes) :

$$r_i = \left(||\mathcal{T}m_i - m'_i||_2^2 + ||\mathcal{T}^{-1}m'_i - m_i||_2^2 \right)^{\frac{1}{2}} ,$$

où $||.||$ désigne la norme euclidienne.

Dans le cas de la géométrie épipolaire, l'erreur résiduelle de transfert s'exprime comme la distance euclidienne d'un point à la ligne épipolaire qui lui est associée, soit :

$$r_i = \left(d(Fm_i, m'_i)^2 + d(F^T m'_i, m_i)^2 \right)^{\frac{1}{2}} ,$$

où $d(Fm, m')$ est la distance euclidienne entre m' et son projeté $m'_\perp$ sur la ligne épipolaire définie par Fm, de telle sorte que $m'^T_\perp Fm = 0$.

Il existe d'autres définitions de l'erreur résiduelle : erreur algébrique, erreur de rétro-projection et erreur de Sampson[3] par exemple. Leur principal intérêt est la simplification de la mise en œuvre des algorithmes dans lesquels elles sont utilisées (par exemple, et de manière non exhaustive, les méthodes des moindres carrés, l'algorithme DLT (*Direct Linear Transformation*), ou encore le *Gold Standard algorithm* [HZ04]).

3.1.3.4 Groupement de correspondances

Dans le cadre spécifique de la reconnaissance d'objets, il faut tenir compte (en plus de l'erreur sur les données) de la présence de fausses mises en correspondance. Au contraire des correspondances correctes de descripteurs locaux qui décrivent le même objet, les fausses mises en correspondance sont des données pour lesquelles il n'existe pas d'interprétation géométrique réelle. On parle alors d'échantillons « aberrants », ou encore d'« outliers » en anglais. Ces données sont parfois modélisées comme des réalisations d'un processus aléatoire, indépendantes et identiquement distribuées selon une loi connue *a priori* (la loi uniforme est le plus souvent utilisée). Les données régulières qui, au contraire, suivent un modèle géométrique déterminé mais inconnu, sont par opposition désignées par le terme d'*inliers*, terme dont il n'existe pas de véritable pendant en français.

Nous avons vu en introduction de cette section qu'en raison de la présence de tels *outliers* dans l'ensemble $\mathcal{C}$, la reconnaissance d'un objet revenait à isoler un groupe $S \subset \mathcal{C}$ d'*inliers*. Contrairement aux *outliers*, les correspondances de S partagent le fait d'être expliquées avec précision par une même transformation $\mathcal{T}$. Or, chaque correspondance de points appartenant à un espace de 4 dimensions, cela signifie que les éléments du groupe S sont localisés au voisinage d'une variété définie par la transformation $\mathcal{T}$. Les outliers occupent quant à eux un hypercube de $\mathbb{R}^4$ de manière aléatoire. Il est donc nécessaire d'utiliser des outils d'estimation robuste afin d'isoler le groupe S du reste des correspondances. Les méthodes classiques utilisées pour ce faire sont rappelées aux sections 3.2.1, 3.2.2 et 3.2.3.

De manière plus générale, lorsqu'il existe plusieurs objets ayant des transformations différentes, une difficulté supplémentaire consiste alors à d'identifier simultanément plusieurs groupes disjoints. La reconnaissance d'objets se ramène alors à une procédure de **groupement multiple** de correspondances, dont nous effectuons un état de l'art aux sections 3.2.2 et 3.2.4.

Après avoir analysé différents types d'estimateurs robuste fondés sur le groupement de correspondances, nous présenterons en section 3.3 quelques approches classiques pour la sélection de modèles.

[3] l'erreur de Sampson représente la distance approchée d'un point m à la projection orthogonale sur la variété définie par la transformation $\mathcal{T}$ considérée [HZ04].

3.2 État de l'art sur le groupement de correspondances de points

Dans un premier temps, nous revenons sur la notion d'estimateur robuste utilisant les méthodes des moindres carrés. Deux principales approches permettant de faire de la reconnaissance d'objets par groupement de mises en correspondance sont ensuite présentées : la transformée de Hough et l'algorithme RANSAC.

3.2.1 Estimateurs robustes des moindres carrés

Considérons le problème de l'estimation *directe* des paramètres de la transformation d'un objet à partir d'un ensemble de correspondances $\mathcal{C}$ de taille N. Nous avons vu qu'il suffisait de choisir n couples de points d'intérêt pour définir de manière unique une transformation $\mathcal{T}$. En raison de l'erreur sur la position des points, il est alors nécessaire d'estimer la transformation d'un objet à partir d'un plus grand nombre de correspondances $N > n$. Dans ce cas, on obtient un système d'équations surdéterminé.

L'approche la plus simple pour estimer directement la transformation d'un ensemble de correspondances est la méthode des moindres carrés (notée LS, pour *Least Square*). Elle consiste à définir la transformation optimale $\mathcal{T}_\mathcal{C}$ pour l'ensemble $\mathcal{C}$ comme la transformation minimisant l'erreur, définie comme la somme des résidus au carré de chaque couple de points :

$$e = \sum_{i=1}^{N} r_i^2 \ .$$

Différentes définitions pour les résidus r_i peuvent être utilisées (§ 3.1.3.3), un choix usuel étant l'erreur algébrique.

> **Remarque 1 :**
> Si le modèle géométrique choisi est la similitude ou la transformation affine, la solution des moindres carrés peut être exprimée analytiquement. Par contre, dans le cas de l'homographie et de la géométrie épipolaire, les matrices sont définies à un paramètre d'échelle près (équations (3.1) et (3.2)), ce qui se traduit par un système d'équations linéaires surdéterminé, à second membre nul. Une solution consiste à utiliser l'algorithme DLT (*Direct Linear Transformation*), qui repose sur une décomposition en valeur singulière de la matrice des coefficients du système.

Notons qu'une définition alternative, *les moindres carrés robustes*, a été proposée par [GL97] pour prendre en compte l'erreur résiduelle maximale pour une classe de perturbation. Cependant, ce type d'approche n'a pas été – à notre connaissance – utilisé pour le groupement de correspondances, où les données sont généralement largement suffisantes pour permettre la définition d'une transformation robuste.

À titre d'exemple dans [Low04], pour une application de reconnaissance d'objets, la transformation affine est définie comme la solution des moindres carrés à partir de correspondances de points d'intérêt.

Estimateurs robustes Le principal défaut des méthodes aux moindres carrés est la définition de l'erreur quadratique e qui la rend très sensible aux données aberrantes (outliers). Pour qualifier ce phénomène, on parle de « point de rupture » (*breakdown point* en anglais) : il s'agit du taux maximal de données aberrantes auquel un estimateur est robuste. Avec les moindres carrés, il suffit théoriquement d'une seule correspondance incorrecte pour perturber l'estimation de la transformation. Autrement dit, le taux d'outliers toléré est nul.

Diverses alternatives ont été proposées pour obtenir un point de rupture plus élevé, dont voici les plus connues. Rousseeuw a tout d'abord introduit la méthode des Moindres Carrés Médians [Rou84] (notée LMS, pour *Least Median of Squares*), où l'erreur est définie de la manière suivante :

$$e = \text{median}\{r_i^2 \ , \ i \in 1, \ldots, N\} \ .$$

Avec l'utilisation du médian, le point de rupture est de 50%, ce qui signifie que la moitié des données peut être contaminée par des outliers. La contrepartie de ce gain énorme en termes de robustesse est que l'estimation du modèle est seulement optimale pour la moitié des données, ce qui est une grandeur arbitraire. Pour remédier à cela, Rousseeuw a ensuite proposé dans [Rou85] une méthode des moindres carrés dits « tronqués » (notée LTS, pour *Least Trimmed Square*), qui consiste à fixer le nombre d'échantillons k utilisés dans le calcul de l'erreur :

$$e = \sum_{i=1}^{k} r_i^2 \, , \lceil \tfrac{N}{2} \rceil + 1 \leq k \leq N \, .$$

La constante k permet d'ajuster le point de rupture de la méthode : plus k tend vers $N/2$, plus l'estimateur est robuste, et moins la précision est grande (à supposer qu'il existe plus de k inliers). L'inconvénient de cette approche est sa mise en œuvre, beaucoup plus lente et complexe que les simples moindres carrés.

Plus généralement, les méthodes appelées *M-estimateurs* [Hub81] (M faisant référence au maximum de vraisemblance), consistent à utiliser une fonction de coût ρ qui varie selon l'amplitude du résidu. L'expression de l'erreur devient :

$$e = \sum_{i=1}^{N} \rho(r_i)$$

avec ρ une fonction de coût, telle que la fenêtre de Huber ou de Tukey. Cette fonction revient généralement à définir e comme une moyenne pondérée des résidus aux carrés. Les poids dépendent d'un seuil sur la valeur des résidus qui permet grossièrement de définir un échantillon en tant qu'inlier ou outlier. Les résidus plus petits que le seuil ont alors un poids proche de l'unité, tandis que les résidus plus grands que ce seuil ont un poids proche de 0. Le seuil est choisi en fonction de l'écart-type des erreurs sur les données. Bien que le point de rupture théorique des M-estimateurs est de 0%, ce sont des estimateurs connus pour être beaucoup plus robustes que les moindres carrés.

Considérations pratique pour la reconnaissance d'objets À ce stade nous devons rappeler que dans notre cadre d'étude, la proportion d'outliers est très variable selon le type de scénario envisagé. Lorsque l'on compare deux images n'ayant aucun objet en commun, toutes les correspondances validées par le processus de mise en correspondance seront donc fausses. La proportion d'outliers est dans ce cas de 100%. Il nous faut donc un **critère de décision robuste** nous permettant d'affirmer qu'il n'y a pas d'objet à détecter dans un tel cas de figure, ce que ne permettent pas les différents estimateurs présentés ici. Un autre cas de figure auquel on s'intéresse est la reconnaissance d'objets multiples, c'est-à-dire de plusieurs objets ayant chacun une transformation qui lui est propre. Nous avons vu dans la section d'introduction à ce chapitre que le problème de leur détection se ramenait à celui du groupement multiple de correspondances. Or, il est important de noter qu'en cherchant à estimer la transformation d'un seul objet, *les correspondances liées aux autres objets se comportent comme des outliers*. Ceci signifie, dans le cadre de la détection multiple, que le taux d'inliers correspondant à un unique objet peut ainsi être très faible. Pour ces deux raisons essentielles, il n'est donc pas envisageable d'utiliser les différentes méthodes d'estimations présentées dans ce paragraphe.

Pour permettre la détection et l'estimation sur des données présentant de telles caractéristiques, une autre famille de méthodes a été proposée dans la littérature : la transformée de Hough [Hou59] et algorithme RANSAC [FB81]. Ces deux approches, très différentes dans leur mise en œuvre, reposent sur la même notion de *consensus* (ou de « vote ») : en échantillonnant aléatoirement les données, on cherche à détecter un groupe de points cohérents selon une même transformation géométrique. Dans les deux paragraphes suivants, nous présentons en détail ces deux méthodes qui sont très largement utilisées dans le domaine de la vision par ordinateur, et dont diverses extensions ont été proposées pour la reconnaissance d'objets.

3.2.2 Transformée de Hough

Hough a introduit sa transformée éponyme dans [Hou59] pour la détection de lignes et de courbes dans des photographies de chambre de détection de particules du CERN. Cette transformation s'étend naturellement à la détection de n'importe quelle forme géométrique paramétrée (cercle, ellipse, ...). Par la suite, avec l'avènement des applications de vision robotique, le principe de la transformée de Hough a été généralisée par Ballard [Bal81] pour la détection de courbes définies arbitrairement.

Principe La transformée de Hough est une méthode de groupement de données qui est réalisée dans l'**espace des paramètres**. Cet espace de dimension p est quantifié selon une grille régulière, définissant ainsi des accumulateurs initialement vides. Le remplissage des accumulateurs est réalisé de la manière suivante : pour chaque groupe S' de $n' \leq n$ échantillons parmi l'ensemble $\mathcal{C}$, les paramètres de la forme recherchée sont estimés, puis un vote est ajouté à tous les accumulateurs qui correspondent à ces paramètres. Si $n' = n$, le groupe S' défini une unique transformation et vote pour un seul accumulateur. Si par contre $n' < n$, le groupe S' vote pour tous les accumulateurs compatibles. Une fois que toutes les combinaisons possibles de n'-uplets ont été testées, on obtient une distribution de votes dans l'espace des paramètres. Un maximum local de cette distribution représente un groupe de données ayant voté pour une même transformation, et dont les paramètres sont donnés par la position de l'accumulateur. Une recherche des maxima locaux permet donc à la fois d'identifier les groupes de données correspondant à une même transformation, et d'estimer quels en sont les paramètres. La validation de ces différents groupes dépend d'un seuil sur le nombre de votes accumulés.

Lorsque le nombre de paramètres devient trop important ($p \geq 3$), le test de toutes les combinaisons $\binom{N}{n'}$ de groupe S' est trop coûteux en termes de calcul. La *transformée de Hough aléatoire* (voir par exemple [XOK]) consiste alors à échantillonner uniformément des groupes S' de taille n.

Lorsque la forme recherchée n'est pas définie de manière paramétrique, c'est la *transformée généralisée de Hough* qui est utilisée. C'est ce qui se produit si l'on souhaite détecter dans une image une forme arbitraire, telle qu'un logo, selon certaines invariances géométriques (transformation affine par exemple). En substance, cette variante de la transformée de Hough consiste à construire au préalable pour la forme recherchée une table de correspondance (*Look-Up Table*) à partir d'un prototype (image binaire de la forme par exemple). Cette table permet de déterminer numériquement les configurations entre les différents points de ce prototype (gradient, position du centre de la forme, *etc.*). Pendant la phase de détection, où sont échantillonnés les groupes S', cette table de correspondance est utilisée pour estimer les transformations de S' et définir les accumulateurs qui doivent être remplis. Remarquons une forte analogie avec la méthode de *Geometric Hashing* [LW88], qui a fait l'objet d'une analyse comparative détaillée avec la transformée généralisée de Hough dans [HB94].

La transformée de Hough est un outil puissant très utilisé en traitement d'images (mais également dans d'autres domaines, tels que la chimie moléculaire), et en particulier en vision par ordinateur. En effet, la transposition de la transformée de Hough aux problèmes du groupement de correspondances est triviale. Des échantillons de n appariements sont sélectionnés puis, les paramètres de la transformation qui leur correspond ayant été calculés, les votes sont comptabilisés dans l'espace des p paramètres. De nombreuses applications utilisent cette transformée pour estimer la pose d'un objet (voir à titre d'exemple [Low04]).

Avantages et Limitations La transformée de Hough procure deux avantages considérables en vue de la reconnaissance d'objets multiples à partir de correspondances de points. En effet, la détection et l'estimation d'un groupe sont simultanément réalisées par une analyse de la distribution quantifiée par les accumulateurs. De plus, de part le principe même d'échantillonnage stochastique sur lequel cette méthode repose, la robustesse aux données aberrantes dépasse de loin les approches décrites au paragraphe précédent. Enfin, cette approche propose un cadre de travail propice à la détection multiples de groupes.

Néanmoins, la définition des accumulateurs de dimension p (taille, forme) est en pratique délicate. Elle a fait l'objet de nombreuses études [Maî86, Bro83, IK88, EGH90], et diverses solutions ont été pro-

posées notamment dans le but d'obtenir une distribution uniforme pour les données aléatoires. De plus, la précision de l'estimation des paramètres de la transformation d'un groupe est fortement dépendante de la définition des accumulateurs, donnant ainsi lieu à un compromis classique entre précision et détectabilité. Afin de rendre robuste la détection de groupes et pallier aux limitations du simple seuil sur le nombre de votes de la méthode originale, des critères de validation plus élaborés ont été proposés, fondés sur le maximum de vraisemblance [Ste91] ou la méthodologie *a contrario* [CDD+07].

> **Remarque 2 :**
> Bien que cette variante ne mentionne pas explicitement la transformée de Hough, l'algorithme de détection de mouvements multiples introduit par Meer et Subbarao [SM06] présente de nombreuses similarités avec cette approche. Des n-uplets de données sont aléatoirement échantillonnés de manière à générer un grand nombre d'hypothèses sur les différentes transformations présentes dans les données. Tout comme la transformée de Hough, l'ensemble de ces hypothèses sont représentées dans l'espace des paramètres de la classe de transformation considérée, à la différence près cependant qu'il n'existe aucune quantification de cet espace. On obtient ainsi à l'issue de cette première étape un nuage de points. L'originalité de l'approche de Meer et Subbarao réside dans la seconde étape, où les différents modes ne sont pas identifiés par le nombre de votes dans les accumulateurs, mais à l'aide de la méthode du « Mean shift » [FH75], qui est généralisée aux cas des variétés analytiques. Cette dernière est une méthode d'estimation non paramétrique des modes d'une densité, dont une étude détaillée est réalisée dans [Che95]. Cette alternative intéressante permet de s'affranchir de la définition des accumulateurs.

Deux principales limitations de la transformée de Hough sont ainsi fréquemment évoquées dans la littérature. La première concerne la précision de l'estimation, mais l'on pourrait objecter que la transformée de Hough peut être utilisé comme initialisation à un estimateur de type moindres carrés, comme cela est fait dans [Low04]. Le second et principal inconvénient de cette approche a trait à sa complexité algorithmique lorsque la dimension de l'espace des paramètres excède $p = 4$. En effet, même en utilisant la transformée de Hough aléatoire, l'échantillonnage de n-uplets requiert un nombre important de tirages pour voir émerger des groupes de consensus, et ce d'autant plus qu'il y a de groupes à détecter. De plus, la robustesse de la transformée de Hough dans le cas d'analyse de scènes complexes avec un fort taux d'outliers est très limitée [EGH90].

Pour cette dernière raison de complexité, l'algorithme RANSAC que nous présentons au prochain paragraphe – également fondé sur la recherche de consensus par vote – est beaucoup plus utilisé pour le groupement de correspondances lorsque l'on considère des transformations plus complexes comme l'homographie ou les transformations 3D (géométrie épipolaire).

3.2.3 RANSAC

L'acronyme RANSAC (RANdom SAmple Consensus) signifie littéralement « consensus à partir d'échantillons aléatoires ». Cet algorithme a été proposé par Fischler et Bolles [FB81] dans un tout autre objectif que la transformée de Hough. Pour un ensemble de données issues d'une *unique* transformation inconnue, son rôle est d'estimer les paramètres de la transformation tout en éliminant les données aberrantes.

Principe Le principe de l'algorithme RANSAC est très proche de celui de la transformée de Hough : on cherche à déterminer un consensus à partir de l'échantillonnage stochastique de n-uplets de données générant ainsi des hypothèses. La différence majeure avec la transformée de Hough se situe dans la manière dont sont exploitées ces hypothèses : au lieu de les agréger dans des accumulateurs de l'espace des paramètres, l'algorithme RANSAC consiste à *les tester successivement en mesurant pour chacune le consensus qu'elle génère*.

Dans [FB81], l'algorithme RANSAC est défini par Fischler et Bolles de la manière suivante : pour chaque itération i, jusqu'à l'arrêt de l'algorithme, un sous-ensemble S' de n échantillons est tiré aléatoirement parmi l'ensemble $\mathcal{C}$ de cardinal N, une transformation exacte $\mathcal{T}_{S'}$ est estimée à partir de S', puis l'ensemble des résidus r_i pour les $N - n$ échantillons restants sont estimés. L'ensemble S des échantillons ayant un résidu plus petit que le seuil r_{max} est construit. Si son cardinal $\#S$ est plus grand qu'une

quantité N_C, il est alors un candidat viable. À l'issue de l'examen des i_{max} hypothèses testées (les sous-ensembles S'), la transformation optimale $\mathcal{T}_{\text{opt}}$ est celle ayant le plus large consensus, c'est-à-dire la transformation donnant le groupe S avec le cardinal le plus important.

Une vue d'ensemble de RANSAC sous sa forme originale est donnée en table 3.1 pour le cas du groupement de correspondances.

TAB. 3.1 – *Vue d'ensemble de l'algorithme RANSAC.*

Algorithme 3.1 RANSAC

Entrées : Ensemble $\mathcal{C}$ de N correspondances $\{(m_i, m'_i)\}$.

Initialisation : groupe $S_{\text{opt}} := \emptyset$, compteur $i := 0$.

Paramètres définis par l'utilisateur : i_{max}, r_{max} et N_C.

1) **Échantillonnage aléatoire :** Tirage d'un jeu de n correspondances S' parmi $\mathcal{C}$.

Estimation de la transformation $\mathcal{T}_{S'}$.

2) **Sélection des inliers :** $S = \{(m_i, m'_i) \in \mathcal{C};\ \mathrm{r}_i < r_{max}\}$.

Si $\#S < N_S$, le groupe S est rejeté $\Rightarrow$ Étape 4).

3) **Consensus optimal :** Si $\#S > \#S_{\text{opt}}$, $S_{\text{opt}} := S$ et $\mathcal{T}_{\text{opt}} = \mathcal{T}_{S'}$.

4) **Critère d'arrêt :** tant que $i < i_{max}$, $i := i + 1$. Retour à l'étape 1).

Sorties : Sous-ensemble de correspondances S_{opt} et transformation $\mathcal{T}_{\text{opt}}$.

On peut voir que cet algorithme s'articule autour de quatre points : une méthode de **génération des hypothèses** qui sont successivement examinées, une **mesure de consensus** pour le test d'une hypothèse, un **critère de validation** d'une hypothèse et un **critère d'arrêt** de la boucle itérative. Nous allons dans les paragraphes suivants détailler les diverses extensions proposées dans la littérature pour chacun de ces quatre points. Notons auparavant que l'algorithme RANSAC à fait l'objet d'un nombre important de publications (applications, extensions, analyses et études comparatives) qu'il serait impossible de mentionner dans leur totalité. Nous nous focalisons donc par la suite sur les contributions proposées dans le cadre de la reconnaissance d'objets.

Génération des hypothèses (échantillonnage) L'échantillonnage de groupes S' peut être vu comme un procédé de génération d'hypothèses sur le modèle suivi par les données. En l'absence de connaissances *a priori*, un échantillonnage aléatoire uniforme de n-uplets S' est utilisé, chaque nouvelle hypothèse étant alors générée indépendamment des précédentes. Deux types d'heuristiques ont été proposées afin de faire converger l'algorithme plus rapidement.

Une première alternative consiste à échantillonner les n-uplets en exploitant le résultat des précédentes hypothèses. Dans [TD03], Torr et Davidson proposent une stratégie de recherche « coarse to fine ». L'algorithme RANSAC est d'abord appliqué localement, sur des correspondances de points appartenant à un voisinage spatial limité, de façon à définir pour chaque correspondance une probabilité d'être un inlier. L'algorithme est ensuite utilisé à des échelles de plus en plus grandes, en échantillonnant suivant ces probabilités qui sont successivement mises à jours. Plus récemment dans [CL09], une approche analogue est proposée – sans notion d'échelle cependant – de manière à accélérer l'élimination des points aberrants. Dans [CM08], Chum et Matas introduisent une procédure de test préliminaire des hypothèses qui permet (sous certaines conditions) d'accélérer le rejet des outliers. Moisan et Stival introduisent dans [MS04] l'algorithme ORSA (Optimal Random SAmpling) dont le principe très simple permet d'accélérer grandement la convergence de RANSAC, avec un facteur de l'ordre de 10. L'idée principale en est la suivante : lorsqu'une hypothèse est jugée satisfaisante, les échantillons sont tirés

parmi le meilleur sous-ensemble S_{opt} trouvé. Nous présenterons plus en détail cet algorithme dans le prochain chapitre (§ 4.1.2).

Une seconde approche de génération d'hypothèses consiste à exploiter des connaissances *a priori* sur les correspondances. Dans [Tor95], les correspondances sont échantillonnées de manière à ce que les points d'intérêt appartiennent à un voisinage restreint dans chacune des images. Pour cela, une première correspondance (m_i, m'_i) est tirée aléatoirement parmi $\mathcal{C}$, puis les $(n-1)$ correspondances restantes (m_j, m'_j) sont sélectionnées de telle sorte que : $||m_i - m_j|| \leq \delta$ et $||m'_i - m'_j|| \leq \delta'$, où δ et δ' sont les distances définissant la taille du voisinage. De manière analogue dans [ZKM05, TF08], la probabilité de tirer une correspondance (m_j, m'_j) est conditionnée par le premier échantillon (m_i, m'_i) : $\mathbb{P}(m_j|m_i) \propto \exp(-||m_i - m_j||^2/2\delta^2)$. Cette contrainte de voisinage permet d'augmenter les chances de tirer un n-uplet constitué de n correspondances correctes. Cependant, les distances δ et δ' sont des paramètres supplémentaires que l'utilisateur doit ajuster. Les auteurs de [Sch06] utilisent un procédé assez similaire où les groupes de correspondances sont échantillonnées de telle sorte que les points d'intérêt appartiennent à une même sous-partie dans chacune des images, découpées selon une grille régulière. Une autre possibilité de génération de n-uplet (complémentaire de la précédente), consiste à tirer les correspondances, non plus selon une probabilité uniforme, mais en tenant compte de la mesure de qualité qui leur est associée lors du processus de mise en correspondance. Cette stratégie correspond à l'intuition qu'il vaut mieux tester en premier lieu des configurations de correspondances en lesquelles on a le plus confiance. L'algorithme PROSAC [CM05] (PROgressive SAmple Consensus) et l'approche utilisée dans [NSB07] utilisent ce principe.

Notons que d'autres méthodes d'échantillonnage ont été proposées dans la littérature pour des applications très spécifiques qui sortent du cadre de ce chapitre, à l'image de *Preemptive RANSAC* [Nis05] où une gestion des priorités sur les hypothèses est définie en vue d'une application temps-réel de reconstruction 3D.

Mesure de qualité et validation d'un groupe et de sa transformation L'évaluation d'un groupe de correspondances S et celle de la transformation $\mathcal{T}_S$ qui lui correspond sont indissociables. Nous avons vu que l'approche originale de RANSAC consiste à sélectionner les inliers à l'aide d'un seuil r_{max} sur les résidus, et à rechercher la transformation donnant le plus d'inliers. Ceci revient [TZ00] à optimiser, par échantillonnage aléatoire, l'erreur e suivante :

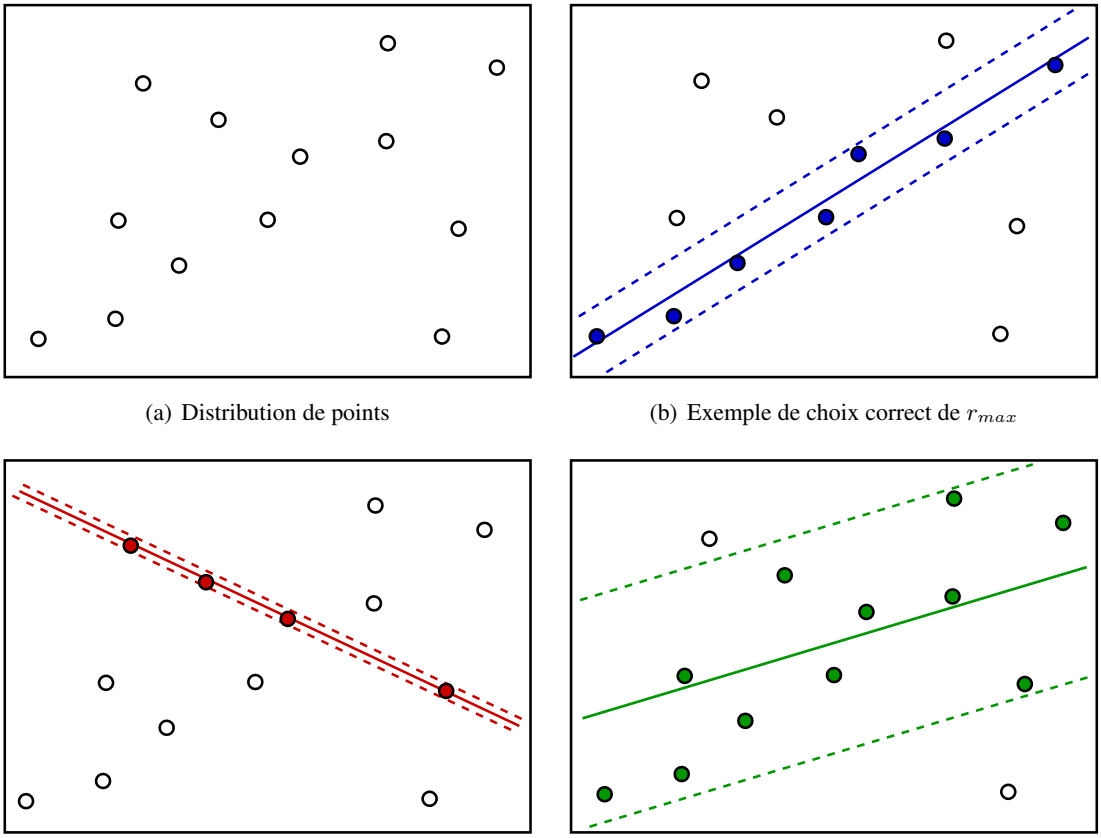
$$e = \sum_{i=1}^{N} \rho(r_i)\,, \text{ avec } \rho(r_i) = \begin{cases} 0 & \text{si } r_i < r_{max} \\ 1 & \text{si } r_i \geq r_{max} \end{cases}, \tag{3.3}$$

où r_i correspond à l'erreur résiduelle associée à la correspondance d'indice i, et dont la valeur dépend de la transformation testée $\mathcal{T}_S$. Plusieurs définitions de l'erreur résiduelle ont été rappelées au paragraphe 3.1.3. À l'issue de ce processus d'optimisation, la transformation optimale $\mathcal{T}_{\text{opt}}$ minimisant e est jugée fiable si le groupe d'inliers qui lui est associé représente plus de N_C correspondances.

La sélection et la validation du groupe d'inliers dépend ainsi de deux paramètres : r_{max}, le seuil sur les résidus et N_C, le seuil sur le cardinal. C'est le seuil r_{max} qui fait de RANSAC un estimateur robuste, pouvant tolérer plus de 50% d'outliers. Le choix pratique du seuil r_{max} est cependant très critique pour la sélection du groupe optimal et de sa transformation. Ainsi que l'illustre la figure 3.5, lorsque ce seuil est trop élevé, quelques outliers sélectionnés suffisent (comme pour les moindres carrés) pour obtenir une estimation très imprécise. Si le seuil est trop faible, il est difficile d'estimer une transformation fiable à partir d'un nombre restreint d'échantillons.

Le seuil N_C sur le cardinal est lui aussi très important car il concerne la validation du groupe optimal S_{opt} : il permet de décider *in fine* s'il existe un objet en commun entre deux images. Pourtant, peu d'études ont été menées afin de régler automatiquement ce paramètre. À notre connaissance, seuls Stewart [Ste95] ainsi que Moisan et Stival [MS04] ont proposé des critères de détection automatique.

Le problème du choix du seuil r_{max} a par contre suscité plus d'intérêt. Il a été notamment suggéré [TZ00] de définir le seuil à partir de l'estimation de la variance de l'erreur résiduelle des inliers, qui



(a) Distribution de points

(b) Exemple de choix correct de r_{max}

(c) Exemple de choix trop petit de r_{max}

(d) Exemple de choix trop grand de r_{max}

FIG. 3.5 – *Importance du seuil de sélection des inliers pour l'estimation de la transformation.*

dépend à la fois de l'erreur sur la position des points, et de la définition de l'erreur résiduelle. En supposant que le bruit sur la position des points d'intérêt appariés est gaussien, de moyenne nulle (détecteur sans biais) et de variance σ^2 connue, il est alors possible d'estimer la probabilité de rejet des inliers ayant un résidu supérieur à r_{max}. Fixer un seuil sur cette probabilité revient à définir le seuil r_{max} [HZ04].

> **Remarque 3 :**
> Notons que cette définition relativement naïve du bruit de mesure ne prend ni en compte le modèle géométrique considéré, ni même les paramètres de la transformation estimée, et suppose en outre une distribution normale. Comme l'indique plusieurs études (voir notamment [CL09]), l'estimation pratique de la variance σ^2 des erreurs résiduelles est très difficile. Elle dépend par ailleurs des conditions d'aquisition par la caméra (quantité de flou ou niveau de bruit par exemple).

Dans [TZ00], une nouvelle fonction de coût est définie en modifiant la fonction de pondération ρ de l'équation (3.3). Le principe est d'obtenir une fonction de coût similaire à celles utilisées par les M-estimateurs qui pénalisent les faibles résidus, contrairement à la fonction de coût de RANSAC. L'algorithme MSAC [TZ00] (M-estimator SAmple Consensus), utilise la fonction de coût suivante :

$$e = \sum_{i=1}^{N} \rho(r_i)\,, \text{ avec } \rho(r_i) = \begin{cases} r_i^2 & \text{si } r_i < r_{max} \\ r_{max}^2 & \text{si } r_i \geq r_{max} \end{cases}, \tag{3.4}$$

où le seuil r_{max} dépend de la variance σ^2 des résidus.

D'autres mesures de qualité ont été proposées pour la sélection du groupe S_{opt} et l'estimation de $\mathcal{T}_{opt}$. Elles reposent sur des hypothèses concernant la distribution des résidus, que l'on peut classer en différentes catégories :

- **Hypothèse : distribution normale des inliers et distribution uniforme des outliers.** Dans [TZ00], Zisserman et Torr introduisent une nouvelle mesure de qualité reposant sur la vraisemblance de cette hypothèse, la variance des erreurs résiduelles σ^2 étant supposée connue. La distribution des résidus r est donc un mélange d'une loi normale et d'une distribution uniforme, qui peut s'écrire :

$$\mathbb{P}(r) = \frac{\gamma}{\sqrt{2\pi}\sigma} e^{-\frac{r^2}{2\sigma^2}} + \frac{1-\gamma}{R} ,$$

où γ est le paramètre du mélange qui représente le *taux d'inliers*, et R est la plage des valeurs des résidus correspondant aux outliers. En pratique cependant, γ est inconnu. On note $\{\xi_j\}$ l'ensemble des N variables cachées prenant leur valeur dans $\{0, 1\}$ suivant que la correspondance d'indice j est considérée comme inlier ($\xi_j = 1$) ou outlier ($\xi_j = 0$). La « log-vraisemblance » s'exprime alors comme :

$$\log(\mathcal{L}) = \sum_{j=1}^{N} \log \left(\frac{\xi_j}{\sqrt{2\pi}\sigma} e^{-\frac{r_j^2}{2\sigma^2}} + \frac{1-\xi_j}{R} \right) .$$

Pour estimer le maximum de cette fonction, l'algorithme MLESAC [TZ00] consiste à utiliser itérativement le processus de génération d'hypothèses de RANSAC, alterné avec l'algorithme de maximisation de l'espérance (noté EM, pour *Expectation-Maximisation* en anglais) qui est utilisé pour estimer les variables cachées ξ_j. Pour chaque n-uplet S' tiré, l'algorithme EM est tout d'abord initialisé avec $\gamma = 1/2$. Ensuite, de manière alternée jusqu'à convergence de l'algorithme EM, on estime la valeur des ξ_j à l'aide des probabilités conditionnelles suivantes :

$$\forall j \in \{1, \ldots, N\} \;\; \mathbb{P}(\xi_j = 1 \mid \gamma) = \frac{p_j}{p_j + p_0} \quad , \quad \text{où} \quad p_0 = \frac{1-\gamma}{R} \quad \text{et} \quad p_j = \frac{\gamma}{\sqrt{2\pi}\sigma} e^{-\frac{r_j^2}{2\sigma^2}} ,$$

puis celle de γ avec : $\gamma = \frac{1}{N} \sum_{j=1}^{N} \xi_j$.
Le groupe optimal S_{opt} est défini par les $\lfloor N\gamma \rfloor$ correspondances de $\mathcal{C}$ ayant le plus faible résidu, selon la transformation $\mathcal{T}$ ayant maximisé la log-vraisemblance.

- **Hypothèse : distributions uniforme des outliers.** La mesure de qualité utilisée par l'algorithme MINPRAN [Ste95] (MINimize the Probability of RANdomness) offre une alternative intéressante à MLESAC. Le principe de MINPRAN est de valider les données qui rejettent l'hypothèse $\mathcal{H}$, selon laquelle les résidus des outliers sont supposés être uniformément distribués.
Soit une transformation $\mathcal{T}$ dont on cherche à déterminer le groupe d'inliers correspondant. Les résidus des N correspondances sont d'abord calculés, puis ordonnés de manière croissante. Considérons maintenant le groupe S constitué des k plus petits résidus, obtenu avec le seuil r. La probabilité d'observer un groupe de taille supérieure ou égale à S sous l'hypothèse $\mathcal{H}$, est la probabilité d'observer au moins k résidus plus petits que r parmi N résidus *i.i.d.* selon une loi uniforme, soit :

$$\mathbb{P}(S|\mathcal{H}) = \sum_{i=k}^{N} \binom{N}{i} \left(\frac{r}{R}\right)^i \left(1 - \frac{r}{R}\right)^{N-i} . \tag{3.5}$$

Un groupe S est considéré comme correct s'il rejette suffisamment l'hypothèse $\mathcal{H}$, c'est-à-dire lorque la probabilité $\mathbb{P}(S|\mathcal{H})$ est petite.
Afin trouver le groupe S et la transformation $\mathcal{T}$ qui minimise cette probabilité, la stratégie de génération d'hypothèses de l'algorithme RANSAC est utilisée. Pour chaque n-uplet échantillonné, la transformation et les résidus des points restants sont calculés, puis le groupe S qui minimise la probabilité est trouvé. À l'issue de cette recherche (i_{max} itérations), le groupe optimal S_{opt} est celui pour lequel l'expression (3.5) est la plus petite.

Stewart propose également une méthode de validation du groupe optimal. Il suggère de calculer la probabilité d'obtenir une valeur plus petite que $\mathbb{P}(S_{opt}|\mathcal{H})$ lorsque les N données vérifient le modèle $\mathcal{H}$, en testant i_{max} transformations différentes. Selon les termes de Stewart, cette probabilité exprime le fait que l'algorithme « hallucine » des groupes lorsqu'il y en a pas. Cette probabilité est ensuite seuillée par l'utilisateur pour valider ou non un groupe optimal, ce qui permet de n'avoir en pratique qu'un seuil de détection à régler. Cependant, comme l'indique Stewart :

> *To make the analysis feasible, we assume the [i_{max}] fits and their residuals are independant. Strictly speacking, this assumption is not correct [...] it makes MINPRAN more conservative in accepting fits.*

Ajoutons par ailleurs que le calcul de la probabilité est en pratique très complexe à mettre en œuvre en comparaison de la méthode suivante, proposée par les auteurs de [MS04].

- **Hypothèse : points d'intérêt indépendants et uniformément distribués dans les images.** Moisan et Stival [MS04] ont introduit, en vue de l'estimation de la transformation épipolaire, une nouvelle mesure de qualité d'un groupe de correspondances dans le cadre de la théorie de la détection *a contrario*. Nous nous référons dorénavant à cette méthode par l'acronyme *AC-RANSAC* (A Contrario RANdom SAmple Consensus), qui est présentée en détail en section 4.1.
  Cette approche présente de nombreuses similarités avec l'algorithme MINPRAN qui vient d'être présenté, notamment en ce qui concerne le test d'une *hypothèse nulle*. Dans le cas de AC-RANSAC, cette hypothèse suppose l'indépendance mutuelle et la distribution uniforme des points d'intérêt mis en correspondance dans chacune des images. Cependant, ces deux méthodes diffèrent sur quelques points. En particulier, les seuils de détections sont *automatiquement* définis avec AC-RANSAC. Une analyse comparative de ces deux approches est proposée au paragraphe 4.1.2.
  Notons enfin que dans [NSB07], Sur et Noury proposent une extension de cette mesure de qualité au cas où l'on possède un taux de confiance sur les mises en correspondance validées.

Nous terminons cet état de l'art sur l'algorithme RANSAC par une présentation du critère d'arrêt.

Critère d'arrêt L'avantage considérable de RANSAC est son temps de calcul extrêmement faible en comparaison de la transformée de Hough. Il suffit, en principe, de tirer un n-uplet correct parmi le groupe d'*inliers* pour le détecter, au lieu d'accumuler les « preuves » de son existence. Cependant, ce gain de temps dépend essentiellement du critère de sortie de l'algorithme, qui correspond au nombre d'itérations maximum i_{max}.

Une approche classique [HZ04] pour définir i_{max} consiste à exploiter la proportion d'inliers minimum, supposée pour l'instant connue, que l'on note q. La probabilité de tirer un groupe de n correspondances correctes dès la *première* tentative est :

$$p = \frac{\binom{qN}{n}}{\binom{N}{n}} = \prod_{i=0}^{n} \frac{qN-i}{N-i} \simeq q^n \text{ en supposant}^4\ n \ll qN\ .$$

La probabilité de tirer une mauvaise configuration (contenant au plus $n-1$ inliers) à la première tentative est alors $1-p$. Le tirage de chaque n-uplet se fait *avec remise*, *indépendamment* des tirages précédents, car il suffit d'un seul outlier sur n correspondances pour obtenir une mauvaise configuration. Par conséquent, la probabilité de tirer successivement m mauvaises configurations est alors :

$$P = (1-p)^m \simeq (1-q^n)^m\ .$$

Autrement dit, connaissant q, on peut exprimer le nombre d'itérations minimum à réaliser i_{min} de manière à ce que cette probabilité P soit plus petite qu'une valeur maximum P_{max} choisie par l'utilisateur :

$$i_{min}(q) = \left\lceil \frac{\log(P_{max})}{\log(1-q^n)} \right\rceil\ . \tag{3.6}$$

[4] Ce qui est le cas en pratique, même pour de faible fraction d'inliers q, sachant que l'on a typiquement $n \leq 7$ et $N \sim 10^3$.

Il est rare en pratique de *connaître a priori la quantitée* q, en particulier dans le contexte de la multi-détection, où les inliers d'un groupe sont des outliers pour les autres transformations [ZKM05], ou bien encore dans le cas général où il faut analyser des ensembles potentiellement sans inliers (la fouille dans une base d'images par exemple).

Pour ces différentes raisons, on utilise simultanément deux types de seuil (le minimum des deux) :

- Un premier seuil i_{max} correspondant au nombre maximal d'itérations autorisé. Ce seuil correspond implicitement au temps de calcul maximal toléré.
- Un second seuil $i_{min}(q)$ qui est évalué dès que l'on détecte un groupe d'inliers S, de sorte que $q = \#S/N$. Ce seuil est alors mis à jour pour chaque nouveau groupe détecté ayant une meilleure mesure de qualité et ayant un nombre d'inliers plus grand. Ce second seuil permet stopper plus rapidement l'algorithme dans le cas où sa valeur est plus petite que i_{max}.

Avantages et limitations de RANSAC L'intérêt premier de RANSAC est, en comparaison de la transformée de Hough, sa rapidité d'exécution (voir par exemple [TKLG07]) ainsi que sa simplicité de mise en œuvre. Il permet ainsi l'utilisation de modèles plus complexes (dans notre cas, l'homographie et la géométrie épipolaire) avec une augmentation raisonnable du temps de calcul. Par ailleurs, la transformation estimée est en pratique bien plus précise que celle obtenue avec la transformée de Hough.

Toutefois, l'algorithme RANSAC tel qu'il a été présenté jusqu'à présent souffre d'un inconvénient majeur : contrairement à la transformée de Hough, il ne permet pas la détection de structures multiples. Nous allons voir dans la section suivante quelques approches qui ont été proposées afin de bénéficier des avantages de RANSAC dans le cas plus général de la détection de plusieurs groupes distincts.

3.2.4 RANSAC et détection multiple

L'algorithme RANSAC, à l'instar des méthodes dérivées des moindres carrés (M-estimateurs et moindres carrés médian notamment), a été proposé pour l'identification et l'estimation simultanée du modèle d'un unique groupe parmi un ensemble de données. Lorsque les données sont le résultat de plusieurs transformations, l'approche originale de RANSAC, ainsi que les différentes variantes qui ont été présentées à la section précédente, ne sont pas adaptées à la détection de multiples groupes. L'adaptation de RANSAC pour la détection multiple est encore aujourd'hui l'objet de nombreuses études.

Dans cette section, nous souhaitons rappeler les différentes approches de détection multiple se basant sur le principe de recherche de consensus par échantillonnage aléatoire de RANSAC. Tout d'abord, nous présentons la méthode la plus simple et la plus largement utilisée, qui consiste à itérer RANSAC de manière à détecter successivement différentes transformations.

RANSAC séquentiel Afin de détecter plusieurs groupes avec RANSAC, de nombreuses études (par exemple, et de manière non exhaustive, [Ste95, VL01, TKLG07, Bar07]) ont suggéré d'en itérer l'algorithme selon le principe suivant : à chaque itération, l'algorithme RANSAC est utilisé sur les données restantes ; si un groupe optimal est validé, les données de ce groupe sont enlevées et l'on réitère le processus. L'algorithme stoppe lorsqu'il n'y a plus aucun nouveau groupe détecté. Cette stratégie d'utilisation itérée de RANSAC sur des données hétérogènes, dont on donne l'algorithme en table 3.2, est souvent appelée dans la littérature « RANSAC séquentiel ».

L'avantage de cette utilisation de RANSAC est que le nombre de transformations à détecter n'est pas requis *a priori*, tout comme avec la transformée de Hough. Un autre avantage que procure cette approche est la possibilité de détecter des groupes d'inliers de tailles très faibles vis-à-vis d'un groupe dominant. Pour illustrer ce dernier point, prenons l'exemple de deux prises de vues d'une scène fixe avec un objet en mouvement. Si l'objet est de taille très réduite dans chacune des images, il ne sera représenté que par une faible proportion de l'ensemble des correspondances ($< 1\%$). Il faut donc *a priori* un très grand nombre d'itérations pour le détecter directement avec RANSAC. Or, en utilisant RANSAC séquentiellement, c'est la transformation principale correspondant au fond qui est d'abord détectée : une fois les appariements de points correspondants éliminés, il est alors beaucoup plus aisé

TAB. 3.2 – *Utilisation séquentielle de RANSAC.*

Algorithme 3.2 RANSAC séquentiel

Entrées : Ensemble $\mathcal{C}$ de $N = \#\mathcal{C}$ correspondances, et $\hat{\mathcal{C}} := \mathcal{C}$.

1) **RANSAC** sur les correspondances restantes $\hat{\mathcal{C}}$.

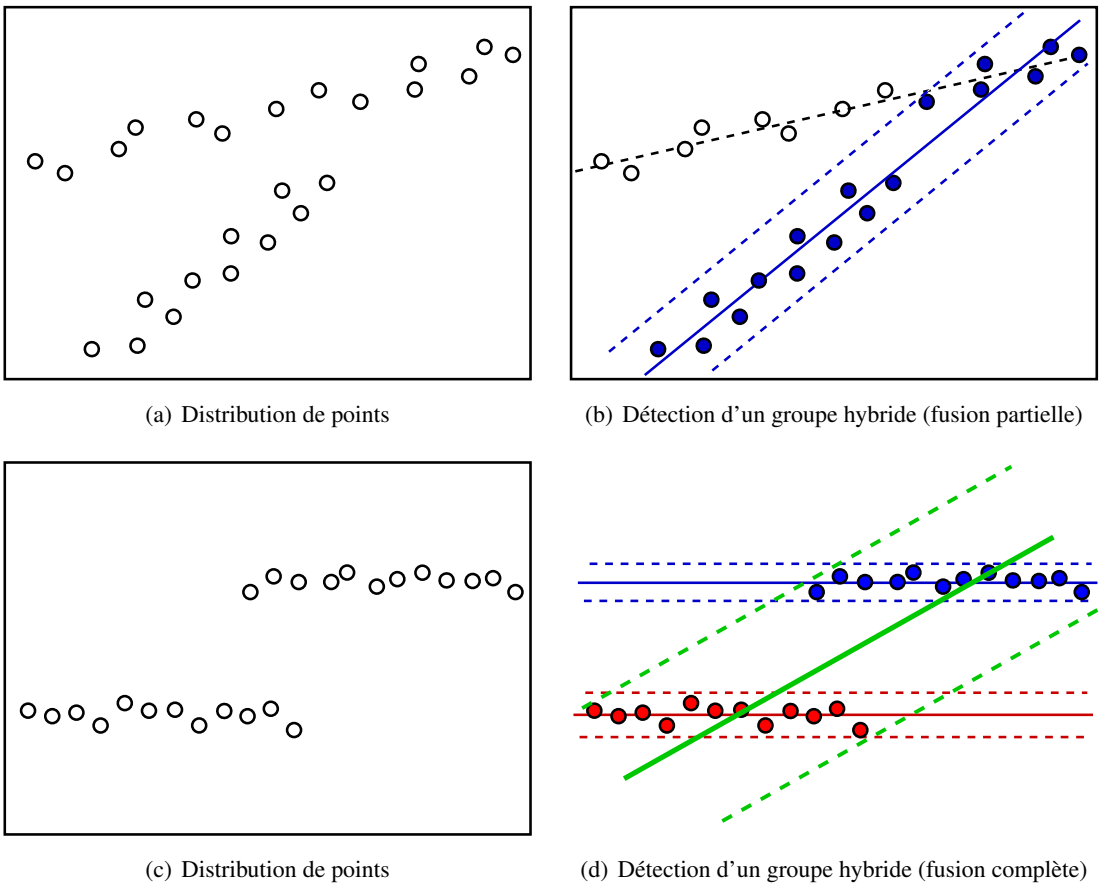
   Détection du groupe S_{opt} et de la transformation $\mathcal{T}_{\text{opt}}$.

2) **Filtrage des inliers :** $\hat{\mathcal{C}} = \hat{\mathcal{C}} \backslash S_{\text{opt}}$.

3) **Critère d'arrêt :** tant que $S_{\text{opt}} \neq \varnothing$, retour à l'étape 1).

Sorties : Listes des sous-ensemble de correspondances $\{S_{\text{opt}}\}$ et de leurs transformations $\{\mathcal{T}_{\text{opt}}\}$.

de détecter le groupe correspondant à l'objet. Nous verrons que ce n'est pas le cas des autres approches utilisant RANSAC pour la détection multiple.



(a) Distribution de points

(b) Détection d'un groupe hybride (fusion partielle)

(c) Distribution de points

(d) Détection d'un groupe hybride (fusion complète)

FIG. 3.6 – ***Illustration du problème de la fusion de groupes avec RANSAC en présence de transformations multiples***. *Dans le premier cas (figure 3.6(a)) les deux groupes corrects ont respectivement* 11 *et* 14 *points, mais le groupe optimal détecté (figure 3.6(b)) est un groupe hybride de* 17 *points. Le second cas (figure 3.6(c)) illustre le cas de la fusion de groupes décrit par Stewart [Ste95] (figure 3.6(d)).*

Cependant, différents phénomènes peuvent limitent les performances d'une utilisation séquentielle de RANSAC, suivant le réglage des différents seuils (r_{max}, N_C, i_{max}). L'une de ces limitations est la ***fusion de groupes***, qui est illustrée par la figure 3.6. Considérons le cas de la figure 3.6(a) avec deux groupes distincts. Avec RANSAC séquentiel, le premier groupe détecté est le groupe ayant le plus grand

nombre d'inliers. Malheureusement pour cette configuration, ce groupe ne correspond à aucun des deux groupes corrects (figure 3.6(b)) : il s'agit d'un groupe « hybride », qui résulte d'une fusion partielle des deux groupes. Dans [Ste95], Stewart montre que l'utilisation séquentielle de MINPRAN conduit au même résultat en raison de l'optimisation de la mesure de qualité qui conduit à ce phénomène de fusion. Il s'intéresse plus particulièrement à la fusion complète de deux groupes et à l'estimation de sa transformation, laquelle est décrite en tant que transformation « bridge ». Ce problème est illustré par les figures 3.6(c) et 3.6(d) : au lieu de détecter chaque groupe de manière distincte avec une grande précision (en rouge et bleu), c'est un unique groupe résultant de l'union des deux précédents qui est sélectionné, donnant une transformation très approximative. Ce cas de figure se présente quelle que soit la mesure de qualité utilisée (RANSAC, MSAC, MLESAC et MINPRAN). Nous aurons l'occasion de revenir sur ce phénomène dans le cas de AC-RANSAC (§ 4.4.4).

D'autres phénomènes que celui que nous venons de présenter peuvent se produire. Une analyse exhaustive des différentes limitations liées à l'utilisation séquentielle de RANSAC sera réalisée dans une prochaine section (§ 4.4.1). De manière générale, il est important de noter que le réglage des différents seuils de RANSAC est encore plus critique en vue de son utilisation séquentielle. À titre d'exemple, un seuil de détection trop permissif va conduire RANSAC séquentiel à détecter de multiples faux groupes.

D'autres méthodes fondées sur l'utilisation de RANSAC ont été récemment proposées.

RANSAC parallèle Zuliani *et al.* [ZKM05] ont proposé une utilisation de RANSAC « parallèle », où plusieurs groupes sont simultanément optimisés par échantillonnage aléatoire de n-uplets. Pour cela, le nombre de groupes est initialement donné par l'utilisateur, ce qui est une limitation pratique majeure. Les expériences présentées dans [ZKM05] montrent que cette utilisation parallèle de l'algorithme RANSAC améliore la précision de l'estimation des différents groupes, notamment parce que le fait de connaître le nombre de groupes à sélectionner permet d'éviter les phénomènes de fusion de groupes ou de fausses détections multiples. Par ailleurs dans [TF08], plusieurs expériences comparatives suggèrent que cette méthode ne présente pas un grand intérêt pratique vis-à-vis de RANSAC-séquentiel.

Regroupement d'hypothèses générées par RANSAC Une autre famille de détection de consensus multiples repose sur une stratégie d'agglomération des hypothèses générées par RANSAC, à la manière de la transformée de Hough (voir par exemple [SM06]). L'idée principale, reprise dans les trois méthodes qui vont être maintenant présentées, est de définir un grand nombre de groupes avec RANSAC, pour ensuite regrouper ceux ayant une transformation similaire.

- **Segmentation du mouvement** Dans [TM94], Torr et Murray utilisent RANSAC pour générer des hypothèses sur les différents mouvements entre deux séquences tirées d'une vidéo. Ces hypothèses se traduisent par de nombreux groupes de correspondances qui sont itérativement fusionnés deux à deux ou éliminés à l'aide de tests statistiques. À l'issue de ce processus, plusieurs groupes distincts sont alors identifiés, correspondant approximativement aux différents mouvements entre les deux images. Une étape d'optimisation est ensuite mise en œuvre, reposant sur l'optimisation de la vraisemblance associée à l'ensemble de ces groupes. Sans rentrer dans les détails, cette approche peut être considérée comme une généralisation de MLESAC dans le cadre de la détection multiple. La maximisation de la vraisemblance est cependant beaucoup plus complexe. Elle est réalisée à l'aide d'un algorithme de type « séparation et évaluation » (*branch and bound* en anglais).
- **Residual Histogram Analysis (RHA)** Plus récemment, Zhang et Kosecka [ZK06b] ont proposé une méthode pour la détermination du nombre de groupes distincts dans les données. Elle se base sur l'analyse de chaque histogramme de résidus obtenu par le tirage aléatoire d'un n-uplets. Cette analyse consiste à estimer le nombre de modes principaux de l'histogramme, supposé correspondre aux différents groupes ayant des transformations distinctes. C'est finalement le nombre médian de modes détectés dans l'ensemble des histogrammes de résidus qui est utilisé pour évaluer le nombre de groupes à valider. Une fois ce nombre estimé, les hypothèses sont ensuite agglomérées afin de sélectionner les correspondances et de les attribuer à un groupe.

Cette approche permet d'illustrer le fait que, lorsqu'il y a plus d'un groupe à détecter, l'hypothèse selon laquelle les outliers sont distribués uniformément est fausse, ce qui pose problème pour des méthodes comme MINPRAN ou MLESAC.

- **J-linkage** Toldo et Fusiello ont proposé récemment une approche similaire dans [TF08]. Comme les méthodes précédentes, de nombreuses hypothèses sont générées à l'aide de RANSAC. Pour chaque hypothèse, ils définissent un groupe à l'aide du seuil r_{max} sur les résidus. L'originalité de leur approche est ensuite de cesser de considérer les résidus des groupes, et de seulement s'intéresser aux étiquettes sur les données. Une matrice de consensus $C := [c_{ij}]$ est alors construite, où l'élément c_{ij} est égal à 1 si l'échantillon i des données appartient au groupe d'inliers de l'hypothèse j, et 0 sinon. Cette matrice donne ainsi, selon les auteurs, une représentation de l'espace des consensus, dans lequel ils prétendent que la prise de décision (fusion ou suppression de groupes) est rendue plus facile. La distance de Jacard est utilisée pour mesurer le degré de similitude entre deux groupes d'inliers S_i et S_j de différentes hypothèses ($i \neq j$) :

$$d_J(S_i, S_j) = 1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|} .$$

Un processus de fusion itératif est alors employé : les deux groupes ayant la distance la plus petite sont fusionnés jusqu'à ce que tous les groupes obtenus soient distincts, c'est-à-dire que chaque correspondances n'appartient qu'à un seul groupe.
Une des difficultés de cette approche est d'éliminer les groupes ne correspondant qu'à des outliers. Les auteurs suggèrent en pratique d'utiliser un seuil sur le cardinal des groupes (comme l'approche originale de RANSAC) pour identifier les outliers, ou bien d'éliminer tous les plus petits groupes jusqu'à ce que le taux d'outliers (alors supposé connu) soit atteint.

> **Remarque 4 :**
> Medioni *et al.* [TTM04] utilisent une stratégie inverse aux méthodes précédemment exposées. Dans le but de détecter des mouvements multiples sous la contrainte épipolaire, les données sont tout d'abord segmentées selon différents groupes dans l'espace des correspondances. Ils s'appuient sur une méthode de vote intitulée « tensor voting » [TML01], qui intègre les contraintes propres à la géométrie épipolaire entre les deux vues d'une scène (variétés de dimension 3 dans $\mathbb{R}^4$). Une fois ces groupes identifiés, l'estimation robuste des paramètres des transformations correspondantes, ainsi que l'élimination des outliers, sont réalisées en utilisant l'algorithme RANSAC sur chacun de ces groupes.

L'ensemble de ces approches, fondées sur le regroupement d'hypothèses, utilisent un seuil r_{max} défini par l'utilisateur pour définir les groupes de données qui sont ensuite examinées.

Une autre limite de ces approches est que la détection des différents groupes dépend beaucoup plus de leur tailles relatives qu'avec RANSAC séquentiel. En effet, rappelons qu'avec RANSAC il suffit théoriquement de tirer un n-uplet correct S' pour détecter le groupe entier S de K inliers qui correspond à la transformation donnée par S'. Plus la taille de S est petite relativement à l'ensemble $\mathcal{C}$ de N correspondances, plus il sera difficile de détecter ce groupe. Or, nous avons vu au paragraphe consacré à RANSAC séquentiel que les groupes de petites tailles peuvent être facilement détectés après avoir éliminé les correspondances des groupes dominants, qui sont facilement identifiables. Ce n'est pas le cas lorsque l'on cherche à détecter simultanément tous les groupes. Afin d'illustrer cela, reprenons l'exemple d'une paire d'images représentant une scène fixe dans laquelle se meut un objet. Supposons qu'il n'y a pas d'outliers, de telle sorte que les deux groupes de correspondances à détecter sont : l'objet (K correspondances) et le fond ($N - K$ correspondances). La proportion de n-uplets permettant de détecter l'objet est environ de $\binom{K}{n}/\binom{N}{n} \simeq (\frac{K}{N})^n$, en supposant $n \ll K$. Si l'objet ne représente que 1% des points, et si $n = 4$ (homographie), cela donne une proportion d'hypothèses correctes pour ce groupe de seulement 10^{-8}. La détection de l'objet va donc nécessiter un nombre d'itérations trop important. Avec RANSAC séquentiel, le groupe dominant va être détecté très facilement (environ 96% de n-uplets correct), ce qui signifie que 100% des correspondances restantes seront celles de l'objet après élimination du premier groupe. En conséquence, l'estimation du nombre d'itérations nécessaires à la détection de tous les groupes devient

encore plus critique que pour l'algorithme RANSAC séquentiel. Une autre raison pour laquelle les petits groupes sont difficilement identifiés est que ces méthodes de groupement sont fondées sur l'élimination des petits groupes (élagage, ou *pruning* en anglais). Cette procédure est très importante pour éviter la détection de multiples faux groupes. L'utilisation d'un seuil sur le cardinal pour éliminer les outliers peut ainsi conduire à la suppression des petits objets.

Enfin, une limitation des approches proposées dans [TF08, ZK06b] est que les auteurs considèrent que le modèle de la transformation recherchée est connu par l'utilisateur, et identique pour tous les objets recherchés, ce qui n'est pas forcément le cas en pratique.

> **Remarque 5 :**
> De nombreuses autres approches ont été proposées, plus spécifiques à la détection de mouvement entre des images temporellement proches issues de séquences vidéo, et qui n'utilisent pas de mises en correspondance entre des points d'intérêt. À titre d'exemple, voir [VCB06, BA96, BP09].

Dans la section suivante qui vient clore ce chapitre, nous allons maintenant nous intéresser au problème de la sélection de modèles en vue de l'estimation de la pose d'un objet.

3.3 État de l'art sur la sélection de modèles géométriques

Dans l'ensemble des méthodes d'estimation robustes précédemment décrites, la classe de transformation considérée (que l'on apelle ici modèle) était fixée. Cela suppose en pratique que le « vrai » modèle de génération des données est connu *a priori*, ou bien que l'utilisateur impose un choix de modèle qu'il sait être adéquat pour les données. Or, dans le cas général, le modèle n'est pas connu et l'on cherche alors à identifier, parmi un ensemble de modèles possibles donnés, quel est le plus satisfaisant pour estimer la pose de l'objet. Toute la difficulté est alors de définir un *critère de sélection de modèles*, qui formalise la comparaison de modèles.

Le premier critère qui nous vient sans aucun doute à l'esprit est de sélectionner le modèle qui minimise l'erreur sur les données. Si l'on choisit, de manière standard, la somme des résidus au carrés $e = \sum_{i=1}^{N} r_i^2$, on obtient l'estimateur des moindres carrés. D'un point de théorique, ce type de critère n'est cependant pas satisfaisant pour deux raisons :

- l'erreur résiduelle e peut être aussi faible que l'on veut en choisissant un modèle avec suffisamment de degré de liberté. Cela signifie en pratique que le modèle ayant le plus de paramètres sera systématiquement privilégié, phénomène que l'on qualifie de « sur-apprentissage » (*overfitting*).
- en plus du degré de liberté du modèle testé, le calcul de l'erreur résiduelle ne prend pas en compte les contraintes géométriques imposé par le modèle. Afin de comprendre ce dernier point, nous reprenons l'exemple donné dans [Tor98] pour la comparaison du calcul de l'erreur résiduelle pour les transformations planes et la matrice fondamentale (géométrie épipolaire). Comme l'illustre la figure 3.7, Torr rapelle que les modèles dont le résidu est défini comme la distance entre un point et une droite (matrice fondamentale F) sont privilégiés en comparaison de modèles où l'erreur est définie comme la distance entre deux points (homographie). Ainsi, en pratique, la géométrie épipolaire permet d'obtenir une erreur résiduelle toujours plus petite que l'homographie, et ce alors même que la géométrie épipolaire offre moins de degrés de liberté (7) que l'homographie (8).

Nous allons dans les paragraphes suivants rappeler brièvement quels sont les critères de sélection de modèles qui ont été proposés afin de prendre en compte ces deux phénomènes.

3.3.1 Critères usuels pour la sélection de modèles

La question de la sélection de modèles est un problème très général, dont nous souhaitons ici donner un léger aperçu avant de nous intéresser au cas particulier de la sélection de modèles géométriques pour

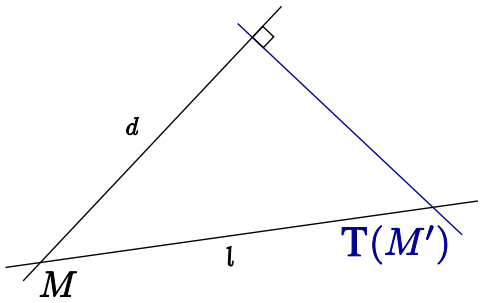


FIG. 3.7 – *Illustration de l'erreur de transfert définie comme la distance l entre le point m et le point $\mathcal{T}m'$ pour une transformation planaire $\mathcal{T}$, et la distance d entre le point m et la droite $F^T m'$ pour la géométrie épipolaire.*

des correspondances de points. On se réfère dorénavant au symbole $\mathcal{M}$ pour désigner un modèle, et à $\mathcal{T}$ pour les paramètres qui sont estimés pour ce modèle.

Examinons tout d'abord le problème du sur-apprentissage. Une alternative au critère des moindres carrés est donné par le principe du Rasoir d'Occam, qui peut être formulé ainsi :

> *« il ne faut pas multiplier les explications et les causes sans qu'on en ait une stricte nécessité. »*

Pour la sélection de modèles, ce principe de parcimonie peut se traduire par le fait de rejeter les modèles les plus complexes qui n'apportent pas un gain significatif sur la précision de modélisation des données. Cela revient en pratique à définir un critère permettant de réaliser un compromis entre précision et la complexité du modèle utilisé, ce que l'on désigne usuellement par l'expression « compromis biais-variance ».

De nombreux critères de sélection de modèles ont ainsi été proposés dans la littérature, dans des cadres théoriques très divers mais dont les expressions sont cependant très similaires. En effet, ces différents critères peuvent généralement s'écrire sous la forme (à une constante additive près) :

$$Q(\mathcal{C}, \mathcal{M}, \mathcal{T}) = -2\log(\mathcal{L}(\mathcal{T}, \mathcal{C})) + P(N, k) \ ,$$

où $\mathcal{L}(\mathcal{T}, \mathcal{C})$ désigne la vraisemblance des paramètres $\mathcal{T}$ du modèle considéré $\mathcal{M}$ en fonction des données $\mathcal{C}$ (N correspondances de points dans notre cas). La fonction P est un terme de pénalisation qui dépend du nombre de données N et qui prend en compte le nombre de paramètres k utilisé par le modèle. Pour simplifier les expressions suivantes, nous désignons désormais par $\mathcal{L}$ la vraisemblance du modèle. Déterminer le modèle optimal selon ce critère requiert alors l'estimation du maximum de vraisemblance de chacun des modèles en compétition.

On modélise généralement les N échantillons de données comme des variables aléatoires *iid*. Dans le cas où les données sont supposées suivre une loi normale, de moyenne nulle et de variance σ^2 **connue**, la log-vraisemblance peut s'exprimer en fonction de la somme des résidus aux carrés :

$$-\log(\mathcal{L}) = -\log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{r_i^2}{2\sigma^2}} = N/2\log(2\pi\sigma^2) + \frac{\sum_{i=1}^{N} r_i^2}{2\sigma^2} \ .$$

> **Remarque 1 :**
> Lorsque σ est inconnu, le nombre de paramètres est $(k+1)$ et l'estimateur de variance $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} r_i^2$ est utilisé.

Dans les paragraphes suivants sont donnés les premiers critères qui ont été proposés dans la littérature (et les plus utilisés aujourd'hui), qui ont par la suite fait l'objet de diverses extensions selon les applications et les hypothèses considérées.

Critère d'information AIC Akaike [Aka74] fut le premier (1974) à proposer un critère de sélection de modèles qui tient compte à la fois de la complexité du modèle et de sa précision. Ce critère intitulé par la suite AIC (pour *Akaike Information Criterion*) est un critère défini dans le cadre de la théorie de l'information. Le critère AIC est un estimateur asymptotiquement[5] non biaisé de l'espérance de l'information de Kullback-Leibler (aussi appelée divergence de Kullback-Leibler) qui peut s'écrire :

$$\mathrm{AIC} = -2\log(\mathcal{L}) + 2k \ ,$$

où l'on rappelle que k désigne le nombre de paramètres du modèle.

Dans le cas où le nombre d'échantillons N n'est pas suffisamment important ($N < 40k$), le critère AIC_c doit être utilisé [Sug78] pour éviter de privilégier des modèles trop complexes :

$$\mathrm{AIC}_c = AIC + \frac{2k(k+1)}{N-(k+1)} \ ,$$

AIC_c convergeant vers AIC lorsque $N \to \infty$.

BIC Schwarz [Sch78] proposa ensuite (1978) un critère alternatif à AIC en se plaçant dans le cadre de l'inférence bayésienne, de manière à pouvoir intégrer un *a priori* sur le modèle $\mathcal{M}$ et sur les paramètres de $\mathcal{T}$ conditionnellement au modèle testé. Le critère BIC, pour *Bayesian Information Criterion*, s'écrit alors de la manière suivante :

$$\mathrm{BIC} = -2\log(\mathcal{L}) + k\log(N) \ ,$$

où l'on rappelle que N désigne le nombre d'échantillons.

Parmi les nombreuses analyses comparatives des deux critères, il existe un consensus pour dire que le critère BIC privilégie le modèle « prédictif », tandis que le critère AIC privilégie le modèle « explicatif ». Autrement dit, le critère AIC tend à surestimer la complexité du modèle, tandis qu'au contraire, le critère BIC privilégie les modèles plus simples [HTF03].

> **Remarque 2 :**
> Il a été par ailleurs montré [BA04] que le critère AIC peut être vu comme un cas particulier de BIC, avec un *a priori* dépendant du nombre de paramètres des différents modèles testés.

MDL Dans un autre registre, Rissanen [Ris78] a proposé un autre critère de sélection de modèles fondé sur la *longueur minimale de description*, ou Minimum Description Length (MDL) en anglais. C'est un concept très simple qui exprime directement du principe du rasoir d'Occam en termes de complexité algorithmique, ou complexité de Kolmogorov. Autrement dit, le critère MDL consiste à déterminer, pour chacun des modèles en compétition, la longueur de code nécessaire à un programme pour représenter les données, le modèle optimal étant défini comme celui minimisant cette quantité. De manière analogue aux critères précédents, la longueur de code L d'un modèle est représenté par deux termes (on parle alors de « codage en deux parties ») :

$$L = L_E + L_{\mathcal{M}} \ ,$$

où L_E désigne le coût de représentation des écarts aux modèles (fonction des erreurs résiduelles), et $L_{\mathcal{M}}$ le coût de représentation des paramètres du modèle choisi. Or, selon la théorie de l'information, considérer la longueur d'un code permettant de représenter des données est équivalent à considérer la distribution de probabilité de ces données. Ainsi le terme L_E peut être exprimé comme l'opposé de la log-vraisemblance ($-\log(\mathcal{L})$) et la complexité du modèle $L_{\mathcal{M}}$ en fonction du nombre de paramètres k et du nombre d'échantillons N, soit finalement :

$$2L = -2\log(\mathcal{L}) + k\log(\frac{N}{2\pi}) \ .$$

Une fois encore, il est remarquable de constater la forte similarité de cette expression avec les critères précédents.

[5] c'est-à-dire lorsque $N \to \infty$.

3.3.2 Critères de sélection de modèles géométriques

Alors que l'estimation des paramètres de transformation en vision par ordinateur a fait l'objet d'un nombre important d'études, la sélection de modèles a suscité très peu d'intérêt dans ce domaine. La majorité des travaux concernent l'ajustement de courbe (*curve fitting* [Kan98]), le recalage d'image (voir par exemple [GBH08]), et la reconstruction de cartes de disparité (*range images* [BS98]). À notre connaissance, seul Torr [Tor97, Tor98, Tor02] étudie le problème de la sélection de modèles dans le cadre d'étude qui nous intéresse : le groupement de correspondances entre paires d'images.

Les différents critères du paragraphe précédent permettent de prendre en compte le nombre de paramètres du modèle considéré. Nous avons vu qu'il était également important de prendre en compte la façon dont est calculée l'erreur résiduelle entre les différents modèles.

GIC Pour y parvenir, Kanatani [Kan98, Kan04] montre qu'il faut prendre en compte la dimension de la variété décrite par le modèle utilisé en plus du nombre de paramètres k. Il définit le critère GIC, pour *Geometric Information Criterion*, fondé sur le critère AIC :

$$\text{GIC} = -2\log(\mathcal{L}) + 2(k + Nd) \ ,$$

où d désigne la dimension de la variété du modèle utilisé. Kanatani illustre l'intérêt de ce critère pour la comparaison de différents modèles géométriques de courbes : ligne, cercle et ellipse dans le plan, ainsi que plan et droite en 3D.

GRIC Torr montre l'intérêt du critère GIC proposé par Kanatani dans le cas de groupes d'appariements de points d'intérêt [Tor97]. Rappelons que dans ce cas, les échantillons de données sont des correspondances appartenant à un espace de 4 dimensions, que l'on note $d' = 4$. La dimension de la variété est inférieure ou égale à 3 ($d = 2$ pour une transformation plane et $d = 3$ pour la géométrie épipolaire). Il montre dans un premier temps le gain de cette approche en comparaison du critère AIC sur des données synthétiques et réelles.

Dans [Tor99, Tor02], Torr propose ensuite à partir du critère BIC un nouveau critère intitulé GBIC (Geométric Bayesian Information Criterion). Ce critère, à la manière de GIC, est défini de manière à prendre en compte la dimension d de la variété :

$$\text{GBIC} = -2\log\mathcal{L} + \lambda_1 Nd + \lambda_2 k \quad \text{avec} \quad \lambda_1 = \log 4 \ \text{ et } \ \lambda_2 = \log 4N \ .$$

En raison du problème de la présence de données aberrantes (et ce, potentiellement en grande proportion), il montre que le critère GIC n'est pas robuste. En s'inspirant des travaux de Ronchetti [HRRS05], qui défini un critère AIC robuste (AICR) en utilisant une pondération sur l'erreur résiduelle (fenêtre de Hubert), il propose de manière analogue dans [Tor97, Tor98] le critère GRIC pour *Geometric Robust Information Criterion*, une extension robuste du critère GIC. Pour cela, il utilise une estimation robuste de l'erreur résiduelle, à la façon des M-estimateurs (procédé qu'il a par ailleurs utilisé pour MSAC et MLESAC [TZ00]). Les données sont entachées d'erreurs indépendantes et normalement distribuées, de moyenne nulle et de variance σ^2 **connue**. Rappelons tout d'abord l'expression des critères AIC et GIC dans le cadre des hypothèses utilisées par Torr :

$$\text{AIC} = -2\log(\mathcal{L}) + 2k = \frac{1}{\sigma^2}\sum_{i=1}^{N} r_i^2 + 2k \quad \text{et} \quad \text{GIC} = \frac{1}{\sigma^2}\sum_{i=1}^{N} r_i^2 + 2(k + Nd) \ .$$

En présence d'outliers, Torr préconise l'utilisation du critère suivant :

$$\text{GRIC} = \sum_{i=1}^{N} \rho(r_i) + \lambda_1 Nd + \lambda_2 k \quad \text{avec} \quad \rho(r) = \min\{\frac{r^2}{\sigma^2} \ , \ \lambda_3(d' - d)\} \ ,$$

où le terme $(d' - d)$ est appelé « co-dimension », représentant la dimension de la contrainte utilisée pour le calcul de l'erreur résiduelle : $d' - d = 2$ pour les transformations planes et $d' - d = 1$ pour la géométrie épipolaire. La constante λ_3 représente l'opposé de la log-vraisemblance de la distribution uniforme des outliers. Dans [Tor98], Torr fixe $\lambda_1 = \lambda_2 = \lambda_3 = 2$ pour ses expériences.

Dans [Tor02], Torr examine en détail les hypothèses sur les distributions des inliers et des outliers. Il redéfinit alors le critère GRIC selon ces distributions et le taux d'outliers :

$$\text{GRIC} = \sum_{i=1}^{N} \rho(r_i) + \lambda_1 N d + \lambda_2 k \text{ avec } \lambda_1 = \log\left(\frac{L^2}{2\pi\sigma^2}\right) \text{ et } \lambda_2 = \log N \ ,$$

où L^2 est l'aire de l'image (supposée carrée, de côté L) et où la fonction de pondération ρ est définie comme :

$$\rho(r) = \begin{cases} \dfrac{r^2}{\sigma^2} & \text{en absence d'outliers} \\ \min\{\dfrac{r^2}{\sigma^2}\ ,\ \lambda_3\} & \text{en présence d'outliers} \end{cases} \ ,$$

avec

$$\lambda_3 = 2\log\left(\frac{p}{1-p}\right) + (d' - d)\lambda_1 \ ,$$

où p est un *a priori* sur le taux d'inliers.

Ce critère est utilisé par Pollefeys *et al.* [RP05] pour détecter, dans une séquence vidéo, les paires d'images où la transformation épipolaire l'emporte sur l'homographie. Ceci permet de sélectionner les vues principales servant à la reconstruction 3D de la scène. Schindler *et al.* [Sch06] utilise également ce critère pour sélectionner les modèles géométriques de chaque objet détecté entre plusieurs vues d'une scène.

Comme nous avons pu le voir dans ce chapitre consacré à l'état de l'art des méthodes de groupement, la majorité des approches proposées reposent en pratique sur un certain nombre de paramètres de détection. Le réglage de ces différents paramètres suppose toujours une certaine connaissance *a priori* sur les données qui sont examinées. Dans le prochain chapitre, nous proposons une approche automatique de groupement multiple et de sélection de modèles géométriques qui ne requiert aucun réglage de paramètres. Cette approche s'inspire de la méthode de groupement *a contrario* proposée dans [MS04].

Chapitre 4

MAC-RANSAC : groupement multiple et sélection de modèles

Dans le chapitre précédent ont été présentées les méthodes robustes de groupement multiple et de sélection de modèles, utilisées dans la littérature pour la reconnaissance d'objets. Nous avons vu que, pour l'essentiel, ces méthodes reposent sur un certain nombre d'hypothèses (en particulier sur la distribution des inliers et des outliers) qui nécessitent d'être vérifiées par les données, et qui requièrent en outre quelques connaissances *a priori* pour la prise de décision (typiquement la variance σ^2 sur l'erreur résiduelle des points d'intérêt).

Dans ce chapitre, nous introduisons un algorithme de détection multiple qui ne requiert aucun réglage de paramètres de décision. Pour cela, nous nous sommes inspirés des travaux de Moisan et Stival [MS04], qui ont défini une mesure de qualité du groupement de correspondances sous contrainte épipolaire, dans le cadre de la théorie de la détection *a contrario*. Nous allons dans un premier temps rappeler son principe dans le cadre d'une utilisation usuelle[1] de RANSAC, que l'on désignera par la suite comme l'algorithme AC-RANSAC. Nous étudierons ensuite son application pour le groupement de correspondances obtenues à l'aide d'un critère de mise en correspondance automatique de descripteurs locaux (section 4.2). Une nouvelle mesure de qualité du groupement de correspondances sera présentée pour les transformations géométriques du plan dans la section 4.3, en vue de la sélection de modèles. En section 4.4, nous proposerons de nouveaux critères de groupement en vue de la reconnaissance d'objets multiples.

Le nouvel algorithme obtenu, MAC-RANSAC, est évalué expérimentalement sur de nombreux exemples en dernière section de ce chapitre.

Ces travaux ont fait l'objet d'une publication dans [RDGM10].

4.1 Rappel sur AC-RANSAC

Dans [MS04], deux améliorations importantes de l'algorithme RANSAC sont présentées. La première consiste en une nouvelle mesure de qualité des groupes testés lors du processus d'échantillonnage de RANSAC. Cette mesure, appelée *rigidité*, est introduite pour l'évaluation et l'optimisation de la transformation entre deux vues stéréoscopiques (géométrie épipolaire). Le seuil de détection de la rigidité est estimé de manière automatique, sans nécessiter de réglage de la part de l'utilisateur ou de connaissance *a priori*. La seconde contribution est une nouvelle stratégie d'échantillonnage, ORSA (Optimal Random SAmpling), qui s'appuie sur la rigidité des groupes testés. Nous nous référons à l'ensemble de l'algorithme par l'acronyme AC-RANSAC.

[1] C'est-à-dire tel qu'il a été proposé dans [MS04], pour la détection d'une unique transformation épipolaire entre une paire d'images.

4.1.1 Critère de validation des groupes

Dans le but de s'affranchir des limitations de RANSAC – et des diverses variantes étudiées précédemment – Moisan et Stival ont utilisé le cadre théorique de décision *a contrario*, dans le but de sélectionner automatiquement les différents paramètres de sélection et de validation des groupes de correspondances pour la géométrie épipolaire.

Hypothèse nulle Rappelons que $C : \{(m_i, m_i')\,,\, i = 1, \ldots, N\}$ désigne l'ensemble des points d'intérêt appariés entre deux images I et I'. On désigne par $\mathcal{P}$ et $\mathcal{P}'$ les plans images correspondant aux images I et I'. On souhaite déterminer un sous-groupe de ces correspondances qui peut être expliqué par une unique transformation. Pour estimer une telle transformation dans le cadre de la méthodologie *a contrario*, on définit tout d'abord l'hypothèse nulle $\mathcal{H}_0$ qui décrit une distribution « générique » de correspondances aléatoires $(\mathbf{m}_i, \mathbf{m}_i')$, $i = 1, \ldots N$ pour lesquelles aucun groupement ne doit être validé. Ensuite, un groupe de correspondances est considéré comme significatif s'il est très improbable d'observer un tel groupe sous l'hypothèse nulle. Cette approche permet de limiter les *erreurs de type I* (ou fausses alarmes), c'est-à-dire les groupes validés de correspondances qui vérifient l'hypothèse nulle $\mathcal{H}_0$. Nous avons déjà fait usage de cette approche de test d'hypothèse au chapitre 2, consacré à la mise en correspondance automatique de descripteurs locaux.

Dans [MS04], l'hypothèse nulle $\mathcal{H}_0$ est définie de la manière suivante, $N \in \mathbb{N}^+$ étant fixé :

Définition 8 (Hypothèse Nulle) *Un ensemble* $\mathbf{C}$ *de* N *correspondances aléatoires* $\{(\mathbf{m}_i, \mathbf{m}_i')\}$ *suit l'hypothèse nulle* $\mathcal{H}_0$ *lorsque*

- *les points* $\mathbf{m}_i$ *et* $\mathbf{m}_j'$, $i, j = 1, \ldots, N$ *sont des variables aléatoires mutuellement indépendantes ;*
- *les points* $\mathbf{m}_i$, $i = 1, \ldots, N$ *sont uniformément distribués sur l'image* I *et les points* $\mathbf{m}_j'$, $j = 1, \ldots, N$ *sont uniformément distribués sur l'image* I'.

> **Remarque 1 :**
> Ce sont les seules hypothèses qui seront utilisées pour définir la mesure de qualité d'un groupe, ce qui distingue AC-RANSAC des différentes approches présentées au paragraphe § 3.2.3 (RANSAC, MSAC, MLESAC, MINPRAN notamment).

Nous allons maintenant définir la probabilité qu'un groupe de correspondances suive l'hypothèse nulle, et la mesure de qualité qui lui est associée dans le cadre de la géométrie épipolaire considéré dans [MS04].

Rappelons auparavant que n correspond au nombre d'appariements de points utilisés pour estimer une matrice fondamentale. Pour une matrice fondamentale F, les contraintes épipolaires impliquent que l'erreur résiduelle pour un couple de points (m, m') est déterminée par la distance des points m' et m aux lignes épipolaires Fm dans $\mathcal{P}'$ et $F^T m'$ dans $\mathcal{P}$ respectivement.

Probabilité sous l'hypothèse nulle dans le cas de la géométrie épipolaire Soit $\mathbf{C}$ un ensemble de N correspondances aléatoires, et $\mathbf{S}'$ un sous-ensemble de $\mathbf{C}$, tel que $\#\mathbf{S}' = n$. On note $F_{\mathbf{S}'}$ la matrice fondamentale déterminée à partir du sous-sensemble $\mathbf{S}'$. Théoriquement (pour plus de détails, voir en annexe C), le nombre de correspondances nécessaires à l'estimation d'une matrice fondamentale définie de manière unique – et exacte – est de $n = 8$ (algorithme à 8-points). Toutefois, il est également possible d'utiliser l'algorithme des 7-points qui, comme son nom l'indique, permet de définir à partir de $n = 7$ points, une ou trois matrices fondamentales différentes. Dans ce cas, $F_{\mathbf{S}'}$ désigne l'une des trois matrices potentiellement obtenues à partir de $\mathbf{S}'$. Nous supposons donc désormais que $n \in \{7, 8\}$, et nous reviendrons ultérieurement sur la conséquence du choix de n pour la définition de la mesure de qualité.

Supposons maintenant que $\mathbf{C}$ suit le modèle de fond $\mathcal{H}_0$, et que l'on a estimé une matrice $F_{\mathbf{S}'}$ à partir d'un sous-ensemble $\mathbf{S}' \subset \mathbf{C}$. Alors, pour n'importe quelle correspondance aléatoire $(\mathbf{m}, \mathbf{m}')$ de $\mathbf{C}$, la probabilité que la distance entre $\mathbf{m}'$ et la ligne épipolaire $F_{\mathbf{S}'}\mathbf{m}$ soit plus petite que α peut être *bornée supérieurement*. Cette borne est le rapport entre l'aire maximale d'une bande de largeur 2α dans

$\mathcal{P}'$, et l'aire de l'image I'. En notant $d(\mathbf{m}', F_{\mathbf{S}'}\mathbf{m})$ la distance euclidienne entre le point $\mathbf{m}'$ et la ligne épipolaire $F_{\mathbf{S}'}\mathbf{m}$, on peut alors écrire :

$$\forall \alpha > 0, \quad \mathbb{P}_{\mathcal{H}_0}[d(\mathbf{m}', F_{\mathbf{S}'}\mathbf{m}) \leq \alpha] \leq \frac{2D' \cdot \alpha}{A'}, \tag{4.1}$$

où D' et A' désignent respectivement la longueur de la diagonale et l'aire de l'image I'. On peut donc écrire que

$$\mathbb{P}_{\mathcal{H}_0}\left[\frac{2D'}{A'} d(\mathbf{m}', F_{\mathbf{S}'}\mathbf{m}) \leq \alpha\right] \leq \alpha \ .$$

De même, en considérant le point $\mathbf{m}$ et la ligne épipolaire $F_{\mathbf{S}'}^T\mathbf{m}'$, on a : $\mathbb{P}_{\mathcal{H}_0}\left[\frac{2D}{A} d(\mathbf{m}, F_{\mathbf{S}'}^T\mathbf{m}') \leq \alpha\right] \leq \alpha$, où D et A désignent respectivement la diagonale et l'aire de l'image I.

Moisan et Stival définissent l'erreur *symétrique* de transfert comme :

$$\max\left\{\frac{2D'}{A'} d(\mathbf{m}', F_{\mathbf{S}'}\mathbf{m}) \ , \ \frac{2D}{A} d(\mathbf{m}, F_{\mathbf{S}'}^T\mathbf{m}')\right\} \in [0, 1] \ .$$

Pour un couple de point aléatoire $(\mathbf{m}, \mathbf{m}') \subset \mathbf{C} \setminus \mathbf{S}'$, sachant que $\mathbf{m}$ et $\mathbf{m}'$ sont indépendants, la probabilité que cette quantité soit plus petite que α est bornée par α :

$$\mathbb{P}_{\mathcal{H}_0}\left[\max\left\{\frac{2D'}{A'} d(\mathbf{m}', F_{\mathbf{S}'}\mathbf{m}) \ , \ \frac{2D}{A} d(\mathbf{m}, F_{\mathbf{S}'}^T\mathbf{m}')\right\} \leq \alpha\right] \leq \alpha^2 \leq \alpha \ . \tag{4.2}$$

Considérons maintenant un sous-sensemble $\mathbf{S}$ de $\mathbf{C}$ tel que $\mathbf{S} \cap \mathbf{S}' = \emptyset$. Pour mesurer le degré de précision de la matrice fondamentale $F_{\mathbf{S}'}$ pour les correspondances de $\mathbf{S}$, on défini la $F_{\mathbf{S}'}$**-rigidité** de $\mathbf{S}$ comme l'erreur de transfert symétrique normalisée maximale sur tous les points de $\mathbf{S}$:

$$\alpha(\mathbf{S}, F_{\mathbf{S}'}) := \max_{(\mathbf{m},\mathbf{m}') \in \mathbf{S}} \max\left(\frac{2D'}{A'} d(\mathbf{m}', F_{\mathbf{S}'}\mathbf{m}), \frac{2D}{A} d(\mathbf{m}, F_{\mathbf{S}'}^T\mathbf{m}')\right) . \tag{4.3}$$

La F-rigidité $\alpha(S, F)$ d'un groupe de correspondances S mesure ainsi la cohérence entre l'ensemble S et la transformation F. Les correspondances aléatoires étant supposées indépendantes (hypothèse nulle $\mathcal{H}_0$), la probabilité d'observer une rigidité $\alpha(\mathbf{S}, F_{\mathbf{S}'})$ plus petite que α est donc bornée par $\alpha^{\#\mathbf{S}}$, où $\#\mathbf{S}$ désigne le cardinal de $\mathbf{S}$:

$$\mathbb{P}_{\mathcal{H}_0}\left[\alpha(\mathbf{S}, F_{\mathbf{S}'}) \leq \alpha\right] \leq \alpha^{\#\mathbf{S}} \ . \tag{4.4}$$

On peut ainsi mesurer de manière équivalente la cohérence d'un ensemble S de correspondances *réelles* selon une transformation F, en considérant la probabilité que la rigidité aléatoire $\alpha(\mathbf{S}, F_{\mathbf{S}'})$ soit plus petite que la *rigidité observée* $\alpha(S, F)$ sous l'hypothèse nulle $\mathcal{H}_0$. Ainsi, la quantité $\alpha(S, F)^{\#S}$ mesure à quel point on s'étonne d'observer un groupe de taille $\#S$ et de rigidité $\alpha(S, F)$ en supposant que ce groupe est généré aléatoirement. Le but étant de ne pas détecter de groupes dans du bruit, seuls les groupes pour lesquels cette probabilité est suffisamment faible seront validés.

Afin de définir un seuil de détection automatique sur la quantité $\mathbb{P}_{\mathcal{H}_0}\left[\alpha(\mathbf{S}, F_{\mathbf{S}'}) \leq \alpha\right]$, de manière analogue au critère de mise en correspondance *a contrario* présenté au chapitre 2, une mesure de qualité est introduite dans le but de contrôler l'espérance du nombre de fausses alarmes.

Critère de validation Nous avons vu que dans l'approche originale de RANSAC, deux paramètres étaient nécessaires à la sélection (résidu maximum r_{max}) et la validation de groupes de correspondances (cardinal minimum N_c). La majorité des méthodes proposées par la suite reposent également sur ces deux seuils : le seuil de sélection r_{max} sur les résidus, qui est supposé connu et fixé, et un seuil de validation sur le cardinal ou sur la vraisemblance d'un modèle (à l'image de l'algorithme MLESAC [TZ00]). À notre connaissance, il n'existe pas de méthode générique qui permette de *fixer automatiquement* ces deux paramètres. Seul MINPRAN propose un critère permettant de définir r_{max} en seuillant la probabilité

d'« halluciner » la détection de groupes dans du bruit, ce qui limite le choix des paramètres à un unique seuil.

En utilisant le cadre de travail des méthodes *a contrario*, l'algorithme AC-RANSAC permet de s'affranchir de ces limitations en définissant une nouvelle mesure de qualité correspondant à une borne supérieure sur l'espérance du nombre de fausses alarmes, *i.e.* le nombre de groupes validés qui suivent le modèle de fond. Dans [MS04], le critère de sélection et de validation ainsi proposé est spécifiquement défini pour l'algorithme des 7-points, qui estime une ou trois matrices fondamentales $F_{S'}$ à partir d'un sous ensemble S' de $n = 7$ échantillons. La mesure de qualité, notée « NFA », est alors définie de la manière suivante :

Définition 9 *Soit $\mathcal{C} = \{(m_i, m'_i) \mid i = 1, \dots N\}$ un ensemble de N correspondances entre les images I et I'. Soit S un sous-sensemble de $\mathcal{C}$, constitué de $\#S = K$ correspondances, avec $K \leq N - 7$. Pour $\varepsilon > 0$ donné, l'ensemble S est dit « ε-significatif » s'il existe un sous-ensemble S' de $\mathcal{C}$, tel que $\#S' = 7$, $S' \cap S = \emptyset$ et*

$$\mathrm{NFA}(S, S') := 3(N-7)\binom{N}{K}\binom{N-K}{7}\left(\min_{F_{S'}} \alpha(S, F_{S'})\right)^K \leq \varepsilon \ . \tag{4.5}$$

Avec cette mesure de qualité, un groupe S dont la rigidité est mesurée selon la (ou les) matrice(s) fondamentale(s) estimée(s) à partir de $\mathbf{S}'$ est d'autant plus significatif que la quantité $\mathrm{NFA}(\mathbf{S}, \mathbf{S}')$ est faible. Le NFA est défini de manière classique à partir de deux termes : une probabilité pondérée par un nombre de tests. La probabilité $\mathbb{P}_{\mathcal{H}_0}\left[\alpha(\mathbf{S}, F_{\mathbf{S}'}) \leq \alpha(S, F_{S'})\right]$ ne pouvant être directement estimée, c'est la borne supérieure définie au paragraphe précédent qui est employée. En raison de l'emploi de $\#S' = 7$ correspondances, c'est le minimum de la rigidité sur les 3 matrices $F_{S'}$ qui est utilisé – la transformation $F_{S'}$ retenue étant bien entendue celle correspondant à ce minimum. Le nombre de tests est directement inspiré du processus d'échantillonnage de RANSAC :

1. **Tirage aléatoire de S' :** le terme $3 \times \binom{N-K}{7}$ correspond au nombre de transformations F qu'il est possible d'estimer, *i.e.* le nombre de groupes S' de taille 7 parmi les $(N - K)$ correspondances restantes pondéré par le nombre maximum de transformations estimées par groupe (3) ;
2. **Analyse des résidus :** le terme $(N - 7)$ correspond aux données restantes dont on calcule les erreurs résiduelles, qui sont ensuite ordonnées en ordre croissant. Il existe alors $(N - 7)$ groupes de tailles différentes qui peuvent être sélectionnés ;
3. **Test d'un groupe S de taille K :** le terme $\binom{N}{K}$ correspond au nombre de groupes de taille $K \leq N - 7$ parmi un ensemble de taille N.

> **Remarque 2 :**
> Dans le cas de l'utilisation d'un algorithme à 8-points, la mesure de qualité devient alors :
>
> $$\mathrm{NFA}(S, S') := (N-8)\binom{N}{K}\binom{N-K}{8}\alpha(S, F_{S'})^K \ . \tag{4.6}$$

Conformément au principe des méthodes *a contrario*, la quantité NFA permet à la fois de définir une mesure de qualité pour la sélection de groupe optimal (*i.e.* le groupe le plus significatif minimisant le NFA), mais également un critère de validation par le choix du seuil ε. En effet, la définition 9 assure que l'espérance du nombre de groupes ε-significatifs parmi l'ensemble $\mathbf{C}$ de N correspondances aléatoires vérifiant l'hypothèse nulle $\mathcal{H}_0$ est plus petite que ε.

Proposition 3 *Si $\mathbf{C}$ est un ensemble de N correspondances aléatoires suivant l'hypothèse nulle $\mathcal{H}_0$, l'espérance du nombre de sous-ensembles ε-significatifs de $\mathbf{C}$ est plus petite que ε.*

Preuve L'espérance définie dans la proposition 3 s'écrit, en désignant par $\#\{F_{\mathbf{S}'}\}$ le nombre de matrices

fondamentales définies par un groupe $\mathbf{S}'$ (soit 1 ou 3) de taille $n = 7$:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{H}_0}\left[\#\left\{\mathbf{S} \subset \mathbf{C}\ ,\ \ \#\mathbf{S} \leq N-7 \ \text{ et } \mathbf{S} \text{ est } \varepsilon - \text{significatif}\right\}\right] \qquad (4.7)\\
&\quad = \sum_{K=1}^{N-7} \sum_{\#\mathbf{S}=K} \mathbb{P}_{\mathcal{H}_0}\left[\mathbf{S} \text{ est } \varepsilon - \text{significatif}\right]\\
&\quad = \sum_{K=1}^{N-7} \sum_{\#\mathbf{S}=K} \mathbb{P}_{\mathcal{H}_0}\left[\exists\, \mathbf{S}' \subset \mathbf{C} \setminus \mathbf{S}\ ,\ \#\mathbf{S}' = 7 \ \text{ et } \mathrm{NFA}(\mathbf{S}, \mathbf{S}') \leq \varepsilon\right]\\
&\quad \leq \sum_{K=1}^{N-7} \sum_{\#\mathbf{S}=K} \sum_{\mathbf{S}' \subset \mathbf{C}\setminus\mathbf{S}\ ,\ \#\mathbf{S}'=7} \sum_{\#\{F_{\mathbf{S}'}\}} \mathbb{P}_{\mathcal{H}_0}\left[\alpha(\mathbf{S}, F_{\mathbf{S}'}) \leq \left(\frac{\varepsilon}{3(N-7)\binom{N}{K}\binom{N-K}{7}}\right)^{\frac{1}{K}}\right]\\
&\quad \leq \sum_{K=1}^{N-7} \sum_{\#\mathbf{S}=K} \sum_{\mathbf{S}' \subset \mathbf{C}\setminus\mathbf{S}\ ,\ \#\mathbf{S}'=7} 3\frac{\varepsilon}{3(N-7)\binom{N}{K}\binom{N-K}{7}}\\
&\quad = \sum_{K=1}^{N-7} \sum_{\#\mathbf{S}=K} \frac{\varepsilon}{(N-7)\binom{N}{K}} = \sum_{K=1}^{N-7} \frac{\varepsilon}{(N-7)} = \varepsilon\ .
\end{aligned}
$$

□

Le seuil ε étant fixé (il sera toujours choisi égal à 1 dans les expériences), la définition 4.5 du NFA permet d'estimer automatiquement les seuils de détection adaptatifs sur les rigidités $\alpha(S, F_{S'})$.

Remarques 3 :

- Notons également que l'on peut définir de manière équivalente le nombre de tests comme $\mathcal{N}_T = \gamma(N-n)\binom{N}{n}\binom{N-n}{K}$ ou $\mathcal{N}_T = \gamma(N-n)\binom{N}{K}\binom{N-K}{n}$ (avec un facteur γ égal à 3 si $n = 7$ et égal à 1 si $n = 8$) ; ceci illustre le fait que l'ordre dans lequel sont tirés les groupes S et S' est indifférent (*i.e.* tirer n correspondances parmi N puis K parmi $N-K$, ou bien tirer K correspondances parmi N puis n parmi $N-K$).
- Contrairement à la mise en correspondance de descripteurs locaux qui correspond à un cas particulier où les groupes sont de tailles fixées (deux éléments), le nombre de tests dépend ici de la taille du groupe S considéré. Il aurait cependant été possible de définir un nombre de tests identique pour tous les groupes S, indépendamment de leur taille. Dans ce cas, un calcul analogue à (4.7) nous conduirait à définir le nombre de tests suivant :

$$
\mathcal{N}'_T = \sum_{K=1}^{N-n} \sum_{\mathbf{S}\subset\mathbf{C}\ ,\ \#\mathbf{S}=K} \sum_{\mathbf{S}'\subset\mathbf{C}\setminus\mathbf{S}\ ,\ \#\mathbf{S}'=n} \gamma = \gamma\binom{N}{n}\sum_{K=1}^{N-n}\binom{N}{K} = \gamma\binom{N}{n}(2^{N-n}-1)\ .
$$

La différence entre les nombre de tests $\mathcal{N}_T$ et $\mathcal{N}'_T$ est illustrée en figure 4.1. On constate, pour la plupart des groupes de taille K, que le nombre de tests $\mathcal{N}_T(K)$ est plus faible que la constante $\mathcal{N}'_T$. Ces deux définitions assurant le contrôle de l'espérance du nombre de fausses alarmes, cela signifie que la détection est rendue plus difficile avec un nombre de tests constant.

4.1.2 Algorithme

Le processus de l'algorithme AC-RANSAC est détaillé en table 4.1. À chaque itération i, un n-uplet S' de taille $n = 7$ est tiré parmi les N correspondances de $\mathcal{C}$. Une ou trois matrices fondamentales $F_{S'}$ sont alors estimées. Pour chacune de ces matrices, toutes les correspondances restantes $(m_i, m'_i) \in C \setminus S'$ sont ensuite ordonnées selon leur erreurs symétriques de transfert normalisées :

$$
\alpha_i = \max\left(\frac{2D'}{A'}d(\mathbf{m}', F_{S'}\mathbf{m}), \frac{2D}{A}d(\mathbf{m}, F_{S'}^T\mathbf{m}')\right) \qquad (4.8)
$$

Le NFA étant une fonction strictement croissante de α_i, la sélection du meilleur groupe pour chaque transformation estimée est simple : il suffit d'évaluer le NFA de chaque groupe S constitué des $K \leq$

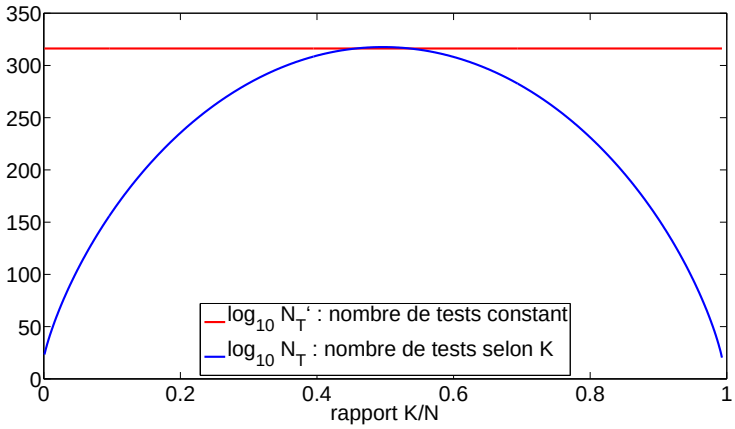


FIG. 4.1 – *Illustration du nombre de tests $\mathcal{N}_T(K)$ en fonction de la taille du groupe considéré K. Ce nombre de tests est comparé à la constante $\mathcal{N}'_T$ (voir la remarque 3) qui est une définition alternative du nombre de tests. Ces deux courbes sont tracées pour $N = 1000$ et $n = 7$ (logarithme de base 10).*

$N - 7$ plus petits résidus, ce qui représente jusqu'à $3(N - 7)$ groupes, puis on identifie le meilleur groupe comme celui minimisant le NFA : $\min_{F_{S'}} \{\text{NFA}(S, S')\}$. Cette étape se répète jusqu'à ce que le nombre d'itérations maximum $i_{\max}$ soit atteint, ou bien jusqu'à ce que l'on détecte un groupe tel que $\text{NFA}(S, S') < 1$. Dans ce cas, la phase de détection s'achève et une phase d'optimisation commence (processus d'échantillonnage ORSA décrit ci-après) à partir du groupe ainsi identifié.

Phase d'optimisation (ORSA) La quantité NFA recèle en pratique un autre avantage non négligeable. Nous avons vu au paragraphe 3.2.3 que l'échantillonnage aléatoire de transformations ne prenait pas en compte le résultat de l'analyse des résidus des tirages précédents. Si par chance le groupe d'inliers est identifié très tôt, les autres tirages devraient être choisis de manière à optimiser ce groupe, plutôt que de poursuivre l'échantillonnage de n-uplet parmi l'ensemble des correspondances $\mathcal{C}$. Le problème est cependant d'identifier l'instant où un groupe d'inliers est suffisamment correct pour décider de l'optimiser. L'idée de Moisan et Stival est de cesser la phase de détection dès que l'on identifie un groupe ayant un NFA plus petit que $\varepsilon = 1$. On définit alors ce groupe comme le groupe optimal S_{opt}. Bien souvent, la transformation estimée pour un tel groupe est peu précise et nécessite une optimisation qui est obtenue en *tirant des groupes S' parmi le groupe d'inliers S_{opt}*. Pour chaque transformation $\mathcal{F}_{S'}$ ainsi générée, un groupe S est sélectionné en sélectionnant des correspondances parmi $\mathcal{C} \setminus S'$. Si ce groupe est tel que son NFA est plus petit que S_{opt}, alors S est le nouveau groupe optimal. Ce principe très simple, appelé ORSA (Optimal Random SAmpling), permet en pratique de gagner un temps de calcul important (les tests menés dans [MS04] suggèrent un gain d'un ordre de grandeur).

Nous avons vu au paragraphe 3.2.3 que d'autres méthodes d'échantillonnage ont été proposées dans la littérature, qui reposent sur le tirage de correspondances dont les points d'intérêt appartiennent à un voisinage spatial restreint. ORSA peut être vu comme un moyen de définir automatiquement un tel voisinage, sans connaissance *a priori* sur les caractéristiques de l'objet recherché (taille, forme, et transformation entre les deux images).

Comparaison avec MINPRAN L'algorithme AC-RANSAC présente de nombreuses similarités avec MINPRAN [Ste95], mais également quelques différences fondamentales, ainsi que le note Desolneux *et al.* [DMM03b] au sujet de la comparaison des méthodes *a contrario* et de l'approche de Stewart :

> *We see that Stewart's method starts exactly as we propose. Stewart actually addresses but does not solve the two problems we intended to overcome. One is the generation of the set of samples, which generates in Stewart's method at least three user parameters, and the second one is the severe restriction about the independence of samples. We actually solved both*

TAB. 4.1 – *Algorithme AC-RANSAC.*

Algorithme 4.1 AC-RANSAC

Entrées : Ensemble $\mathcal{C}$ de N correspondances, nombre d'itérations maximum i_{max}, initialisation de $S_{\text{opt}} := \varnothing$, et de $i := 0$.

1) **Échantillonnage aléatoire :** Tirage d'un jeu de n correspondances S' parmi $\mathcal{C}$.
Estimation des trois matrices $\mathcal{F}_{S'}$.

2) **Sélection des inliers :** Tri des correspondances (m_i, m'_i) selon leur erreur symétrique de transfert normalisé α_i.
Sélection du groupe candidat S minimisant $NFA(S, S')$.

3) **Validation :**
Si $NFA(S, S') < 1$, passage à l'étape 4
Sinon, si $i < i_{max}$, $i := I + 1$ et retour à l'étape 1.
Sinon arrêt de l'algorithme.

4) **Optimisation (ORSA) :** Optimisation de $F_{S'}$ par échantillonnage aléatoire dans S.

Sorties : Sous-ensemble de correspondances S_{opt} et la matrice $\mathcal{F}_{\text{opt}}$.

difficulties simultaneously by introducing the number of samples as an implicit parameter of the method (computed from the image size and Shannon's principles) and by replacing in all calculations the "probability of hallucinating a wrong event" by the "expectation of the number of such hallucinations", namely what we call the number of false alarms.

Considérons tout d'abord leurs similitudes : ce sont deux méthodes qui utilisent un cadre probabiliste pour définir un critère de sélection de groupes de données à partir des erreurs résiduelles. Une mesure de qualité est définie en fonction d'un test d'hypothèse nulle sur les données. Pour une transformation échantillonnée aléatoirement, les résidus sont ordonnés en ordre croissant et le groupe minimisant cette mesure de qualité est sélectionné.

Ces deux approches se distinguent cependant sur différents aspects. D'une part, la théorie de la détection *a contrario* présente l'intérêt considérable de reposer sur un unique paramètre de détection ε qui est très intuitif. Il représente une borne sur l'espérance du nombre de fausses alarmes, fixée à 1 pour toutes les expériences. En comparaison, le paramètre de validation de MINPRAN, qui est plus difficile à régler, est un seuil sur une probabilité de sélectionner un faux groupe. L'estimation de cette probabilité est réalisée de manière empirique, par une méthode d'inférence plus complexe que celle de Moisan et Stival, et repose sur une hypothèse – fausse – d'indépendance des différentes transformations testées et de leurs résidus. Le seuillage de cette probabilité définit ensuite des seuils de détections sur la mesure de qualité des groupes testés. Comme le soulignent Sur *et al.* dans [MSM03], AC-RANSAC définit également des seuils de validation de manière automatique, mais à partir d'un calcul d'espérance qui ne requiert aucune approximation. D'autre part, le seuil de détection de MINPRAN dépend du nombre de groupes réellement testés (nombre d'itérations i_{max}). Au contraire, le critère de détection *a contrario* prend en compte l'ensemble des tests possibles, et est indépendant du nombre de tirages aléatoires.

Notons enfin que les applications visées par ces deux méthodes sont différentes, puisque MINPRAN est utilisé pour la reconstruction de plans à partir d'une image de disparité (points dans $\mathbb{R}^3$), tandis que AC-RANSAC est utilisé pour l'estimation de la transformation entre deux images à partir de correspondances (couples de points d'intérêt de $\mathbb{R}^4$). Par conséquent, les définitions des résidus sont différentes pour ces deux méthodes.

L'algorithme AC-RANSAC tel qu'il été introduit dans [MS04] a été défini afin de trouver une transformation épipolaire unique entre deux paires d'images, à partir de points de contrôle définis manuellement. Nous allons dans les paragraphes suivants présenter l'algorithme de détection multiple MAC-RANSAC. Il consiste en la modification de AC-RANSAC en vue de son utilisation séquentielle à partir de mises en correspondance automatique de descripteurs locaux, pour différents modèles géométriques.

4.2 Hypothèse nulle et mise en correspondance de descripteurs locaux

L'algorithme AC-RANSAC que nous venons de présenter utilise un modèle de fond permettant théoriquement le rejet de fausses détections. Ce modèle de fond repose sur une hypothèse nulle, qui suppose que les fausses correspondances sont des couples de points indépendants et distribués uniformément dans chaque image. L'objet de cette section est l'étude de la validité de cette hypothèse nulle dans notre cadre d'étude, où des appariements de points d'intérêt sont obtenus par mise en correspondance de descripteurs locaux.

En figure 4.2, nous proposons une expérience simple où l'on compare deux images différentes. Suivant la procédure décrite au chapitre 2, des points d'intérêt sont extraits de chacune des images et des descripteurs locaux sont construits puis mis en correspondance. Pour obtenir de nombreuses fausses correspondances, nous avons utilisé un seuil de détection très élevé. Un faux groupe est détecté par l'algorithme AC-RANSAC parmi ces fausses correspondances, illustrant le non respect du modèle de fond par les correspondances de descripteurs locaux. Nous en expliquons les raisons dans les paragraphes suivants, où sont proposés un principe de filtrage et une méthode de normalisation pour assurer la validité de l'hypothèse nulle.

4.2.1 Indépendance des correspondances

Le modèle de fond pour les correspondances aléatoires $\mathbf{C}$ repose sur l'hypothèse nulle $\mathcal{H}_0$, selon laquelle les points appariés $\mathbf{m}_i \in I$ et $\mathbf{m}'_i \in I'$ sont des variables aléatoires mutuellement indépendantes. Or, les faux appariements de points obtenus par un critère de mise en correspondance ne respectent pas nécessairement cette hypothèse d'indépendance, et ce pour deux raisons qui vont être ici détaillées.

Correspondances multiples entre points d'intérêt Nous avons vu que le critère de mise en correspondance *a contrario* introduit au chapitre précédent autorisait les appariements multiples entre un descripteur requête de l'image I et les descripteurs candidats de l'image I'. De manière plus générale, l'utilisation de n'importe quel critère – même restreint au plus proche voisin – peut conduire à l'appariement de plusieurs descripteurs requêtes de l'image I avec un même descripteur de l'image I'. Ces correspondances multiples ne peuvent être évitées qu'en utilisant un critère au plus proche voisin symétrique. Nous avons vu cependant que cela réduisait considérablement le nombre de correspondances dans le cas des occurrences multiples (objets apparaissant plusieurs fois, structures répétitives, ...).

En raison des correspondances multiples qui ne suivent pas le modèle de fond, il est possible de détecter des groupes de correspondances avec l'algorithme AC-RANSAC puisque l'hypothèse d'indépendance est de toute évidence fausse dans un tel cas. Pour éviter ce phénomène, tout en préservant les correspondances multiples, nous proposons le principe de maximalité suivant lors de l'examen des erreurs résiduelles de RANSAC.

Définition 10 (principe de maximalité) *Pour une transformation donnée par S', une seule mise en correspondance par point d'intérêt $\{\mathbf{m}_i\}_{i \leq N_Q}$ et $\{\mathbf{m}'_j\}_{j \leq N_C}$ peut être sélectionnée dans un groupe. Cette correspondance est celle qui minimise l'erreur résiduelle α_i (Éq. (4.8)).*

La mise en œuvre d'un tel principe est très simple car les correspondances sont ordonnées en ordre croissant de leur erreur résiduelle α lors de l'étape de sélection de groupe. Il suffit alors de parcourir la

FIG. 4.2 – *Fausse détection de groupe avec les mises en correspondance de descripteurs locaux.* *À gauche : De manière à obtenir de nombreuses fausses correspondances sur deux images différentes, le critère de mise en correspondance AC présenté au chapitre 2 est utilisé avec un seuil de validation très élevé ($\varepsilon = 1000$ au lieu de $\varepsilon \leq 1$). 6334 fausses correspondances sont ainsi obtenues, illustrées par des points violets dans chacune des images. Au centre : Un faux groupe très significatif de 5730 correspondances est détecté avec AC-RANSAC, avec une mesure de qualité de NFA $= 10^{-305}$. Cette fausse détection est liée au fait que les correspondances aléatoires obtenues ne suivent pas le modèle de fond, en raison de correspondances multiples et redondantes d'une part, et de la concentration des points d'intérêt d'autre part. À droite : En utilisant notre procédure de filtrage et de normalisation, présentée dans cette section, plus aucun faux groupe n'est détecté (*i.e. *le meilleur groupe trouvé a un NFA supérieur à 1).*

liste pour écarter temporairement les correspondances de points ayant déjà été sélectionnés. Ce principe permet effectivement d'éliminer les fausses détections de groupes liées aux correspondances multiples incorrectes.

Néanmoins, il existe une autre raison pour laquelle les fausses correspondances de points d'intérêt ne sont pas indépendantes. Elle concerne l'étape de détection des points d'intérêt.

Redondance des points d'intérêt Certaines structures d'intérêt sont détectées de manière *redondante*, quelque soit le détecteur de points (ou de régions) d'intérêt utilisé (MSER, Hessien, Laplacien, Harris) (voir la section B.1.4 en annexe). Le terme « redondant » signifie ici qu'une même structure peut être représentée par plusieurs points d'intérêt qui diffèrent très légèrement en position et en échelle. Typiquement, les coins sont détectés de manière redondantes en raison de la non-localisation d'un coin en espace-échelle linéaire. Soit une structure détectée de manière redondante par plusieurs points d'intérêt $\{m_i\}$ dans l'image I : si l'un de ces points d'intérêt est apparié avec m' dans l'image I', alors les autres points d'intérêt redondants ont de fortes chances d'être également mise en correspondance avec m'. De plus, en prenant en considération les mises en correspondances multiples, ces points d'intérêt redondants peuvent être appariés de multiples fois. En effet imaginons que deux points aléatoires m et m' sont appariés. Si chacun de ces points est détecté de manière redondante n fois, il y a donc potentiellement n^2 appariements incorrects qui **contredisent l'hypothèse nulle en terme d'indépendance**. Le principe de maximalité permet heureusement de réduire ce nombre à n appariements redondants incorrects, mais cela peut être insuffisant pour détecter de faux groupes. Il est donc primordial d'éliminer ces appariements redondants, tout en préservant les appariements multiples correspondant à des objets répétés.

Pour cela, nous proposons tout d'abord d'en donner une définition plus précise à partir du voisinage

des points d'intérêt.

Définition 11 (Correspondances redondantes) *Deux correspondances $c_i = (m_i, m'_i)$ et $c_j = (m_j, m'_j)$ de points d'intérêt sont considérées comme redondantes si* ***l'une des deux assertions est vraie*** *:*

- $m_i = m_j$ *et* $\|m'_i - m'_j\|_2 < \min\{\Delta_i, \Delta_j\}$
  ou
- $m'_i = m'_j$ *et* $\|m_i - m_j\|_2 < \min\{\Delta'_i, \Delta'_j\}$

où $\|.\|_2$ *désigne la norme euclidienne, et* Δ_k *représente l'échelle caractéristique du point d'intérêt* m_k*.*

Le voisinage d'un point d'intérêt est ici défini comme un disque dont le rayon correspond à son échelle caractéristique Δ, qui dépend de l'échelle σ à laquelle a été détecté le point d'intérêt en espace-échelle linéaire. Dans le cas des descripteurs locaux de type SIFT, qui sont rappelés en annexe B, cette échelle caractéristique a été fixée arbitrairement à $\Delta = 3\sigma$, et correspond au disque central utilisé pour l'extraction des histogrammes locaux (voir la figure B.8).

> **Remarque 1 :**
> D'autres définitions de voisinage peuvent être choisies selon le type de détecteur utilisé : MSER [MCUP02], détecteurs de point invariant affine [MS02], *etc*.

La définition 11 permet d'identifier les appariements redondants et un critère de sélection doit être appliqué afin de ne conserver qu'une seule mise en correspondance redondante par point d'intérêt. Nous proposons pour cela d'utiliser la mesure de qualité de la correspondance. Dans le cas du critère de correspondance *a contrario* présenté au chapitre 2, il s'agit du NFA (Formule 2.3). Le *principe d'exclusion* consiste alors à identifier pour chaque point d'intérêt les éventuelles correspondances redondantes, pour ne conserver que la correspondance redondante ayant le NFA le plus faible. Afin de traiter le plus efficacement l'ensemble des points d'intérêt appariés, les correspondances sont analysées par ordre croissant de NFA.

Le principe d'exclusion, dont l'algorithme est détaillé en table 4.2, est illustré par la figure 4.3.

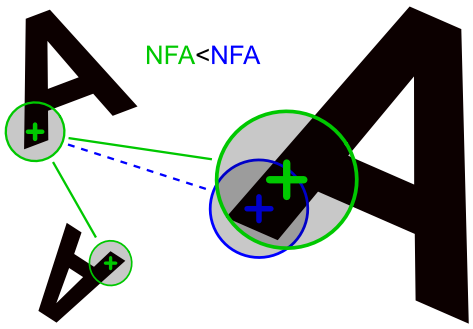


FIG. 4.3 – ***Illustration du principe d'exclusion des appariement redondants.*** *Le meilleur des appariements redondants (de NFA minimum) est sélectionné tandis que les autres sont éliminés. Les appariements multiples non redondants sont conservés par le principe d'exclusion.*

Si l'on reprend l'exemple de fausse détection introduit en début de cette section (figure 4.2), l'utilisation combinée du principe de maximalité et du principe d'exclusion permet de ne plus détecter aucun faux groupe. Nous verrons que le filtrage des correspondances redondantes présente un intérêt supplémentaire pour la détection de groupes multiples.

4.2.2 Normalisation

Nous venons de traiter la question de l'indépendance des mises en correspondance suivant le modèle de fond. Une autre hypothèse sur laquelle repose ce modèle est la distribution uniforme des points

TAB. 4.2 – *Élimination des appariements redondants.*

Algorithme 4.2 Principe d'exclusion des redondances

Entrée : Ensemble $\mathcal{C}$ de correspondances et de leurs mesures de qualité respectives.

1) **Ordonnancement :** Tri des correspondances c_i en ordre croissant de leur mesure de qualité q_i.

2) Pour chaque c_i, à partir de $i = 1$:

   Détection : Calcul de l'ensemble des correspondances redondantes

   $\mathcal{S} = \{c_j : j > i \text{ et } c_j \text{ est redondant avec } c_i\}$.

   Sélection : Élimination de ces correspondances $\mathcal{C} := \mathcal{C} \setminus \mathcal{S}$.

   $i := i + 1$.

Sortie : Liste de correspondances non redondantes $\mathcal{C}$.

d'intérêt appariés sur l'ensemble des domaines des images I et I'. La borne supérieure sur la probabilité d'observer un groupe de correspondance conditionnellement à $\mathcal{H}_0$, est ainsi définie (Éq. (4.2)) en considérant le résidu normalisé selon les dimensions des images (aire A et diagonale D).

Or, en utilisant un processus automatique de détection et d'appariement de points d'intérêt, il arrive fréquemment que ces points ne soient détectés que dans une sous-partie des images. C'est le cas par exemple des images contenant de larges régions uniformes, comme le ciel dans les photographies de la figure 4.2. Cela signifie, dans de tels cas, que l'on va s'étonner artificiellement d'observer des groupes de correspondances concentrés dans des sous-parties de I et I'. Du point de vue de l'estimation de la significativité, cela revient à *sous-estimer* le NFA. Concrètement, il existe donc une concentration critique des points d'intérêt pour laquelle de faux groupes vont être validés, ainsi que l'illustre la figure 4.4.

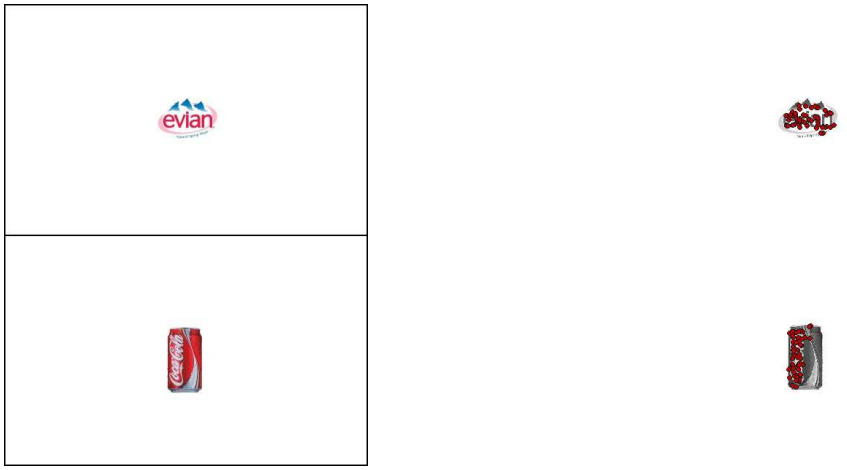


FIG. 4.4 – ***Illustration de la notion de taille critique pour la fausse détection.*** À gauche *: Deux images différentes où les objets d'intérêt occupent une sous partie de l'image. Des fausses correspondances sont obtenues en utilisant le critère de mise en correspondance AC avec seuil de détection très élevé* ($\varepsilon = 1000$ *au lieu de* $\varepsilon \leq 1$*).* À droite *: Lorsque les points d'intérêt appariés sont concentrés dans chacune des images, l'algorithme AC-RANSAC peut détecter un faux groupe. Dans cet exemple, le NFA du faux groupe validé est de* $10^{-12.1}$*. Avec le procédé de normalisation présenté dans ce paragraphe, ce faux groupe n'est plus validé (NFA* $= 10^{14.4}$*).*

Pour que la mesure de qualité (NFA) soit robuste à la distribution réelle des points dans chacune

des images, nous proposons d'estimer les paramètres de normalisation de l'erreur résiduelle (aire A et diagonale D) directement à partir des coordonnées des points appariés. Comme l'illustre le schéma de la figure 4.5, nous considérons simplement dans chaque image une ellipse qui contient une certaine proportion des points d'intérêt mis en correspondance.

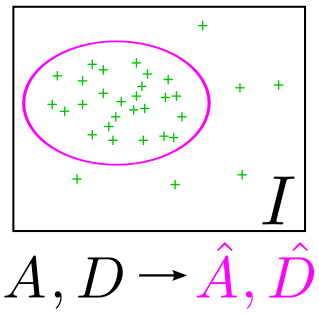


FIG. 4.5 – *La normalisation de la distribution des points est une étape simple mais nécessaire pour la robustesse de la définition du NFA (vérification du modèle de fond).*

Les paramètres d'une ellipse sont estimés selon les moments de second ordre de la distribution des points d'intérêt dans une image. On note δ_1 et $\delta_2 \leq \delta_1$ les écarts-types selon les orientations principales de la distribution de points dans l'image I. Nous avons choisi d'utiliser une ellipse de demi-grand axe égal à $2\delta_1$ et de demi-petit axe égal à $2\delta_2$. Dans le cas d'une distribution gaussienne, ce choix correspond à une ellipse contenant environ 90% des points. Les nouveaux paramètres de normalisation pour l'image I sont alors définis de la manière suivante :

$$\begin{cases} \hat{A} &= 4\pi\delta_1\delta_2 \\ \hat{D} &= 4\delta_1 \end{cases} .$$

Les paramètres de normalisation pour l'image I' sont définis de manière analogue.

Cette simple normalisation de l'aire et de la longueur caractéristique permet d'éviter le problème de la taille critique de fausse détection rencontré dans l'exemple présenté en début de ce paragraphe (figure 4.4).

D'un point de vue théorique, cette analyse exclusivement fondée sur les deux premiers moments de la distribution de points est naturellement insuffisante pour décrire des distributions plus complexes qu'une distribution uniforme ou normale. Par exemple, une scène avec plusieurs objets en mouvement indépendants sera mieux décrite par un mélange (par exemple, un mélange de gaussiennes). Au problème de l'estimation robuste de ces paramètres, s'ajoute alors celui de la redéfinition de A et D pour un tel modèle. Malgré tout, nous avons constaté expérimentalement que la normalisation proposée est suffisamment robuste pour éviter le phénomène de taille critique pour la détection.

> **Remarque 2 :**
> La normalisation des données selon la distribution réelle des points appariés présente un autre avantage. Hartley a ainsi montré qu'il est préférable – et nécessaire, dans le cas de la géométrie épipolaire [Har97] – de normaliser les coordonnées des couples de points d'intérêt dans le but d'estimer de façon robuste les paramètres de la transformation considérée. Il propose ainsi de normaliser les coordonnées des points de correspondances selon deux méthodes au choix, donnant des résultats similaires :
> – annuler le moment d'ordre 1 du nuage de points, et normaliser à l'unité les moments de second ordre (selon les orientations horizontales et verticales de l'image),
> – centrer le nuage puis normaliser avec un seul facteur d'échelle, de manière à ce que la distance moyenne d'un point par rapport à l'origine soit égal à $\frac{1}{\sqrt{2}}$.
> Nous utilisons également une telle normalisation des coordonnées des points d'intérêt pour le groupement de mises en correspondance.

4.3 Sélection de modèles *a contrario*

Nous proposons dans cette section un algorithme de sélection de modèles *a contrario*. Il repose sur le critère de validation initialement défini pour la géométrie épipolaire, qui est ici étendu à d'autres modèles géométriques.

4.3.1 Problématique

Dans le paragraphe 3.3 consacré à l'état de l'art sur la sélection de modèles, nous avons vu que les méthodes des moindres carrés conduisaient systématiquement à choisir le modèle ayant le plus grand degré de liberté. Les critères d'« information » (AIC, BIC, MDL, *etc*.) sont pour cette raison plus appropriés, permettant un compromis entre biais et variance (précision du modèle et complexité).

Lorsqu'il s'agit de la comparaison de modèles géométriques pour groupement de mises en correspondances (dans $\mathbb{R}^4$), il est nécessaire de prendre en compte la dimension des contraintes pour ne pas favoriser les modèles de dimension plus élevée. La comparaison de la géométrie épipolaire et de la géométrie projective pose ainsi problème car ces deux modèles présentent à la fois un degré de liberté différent (nombre de paramètres indépendants), mais également des contraintes différentes. Concrètement, l'erreur résiduelle est calculée de manière différente pour les deux transformations : distance entre des couples de points pour l'homographie, et distance de projection orthogonale d'un point sur une droite pour la matrice fondamentale.

En pratique, la définition d'un critère de sélection entre ces deux modèles est très importante, particulièrement en reconstruction 3D, afin de pouvoir estimer quel est le modèle géométrique le plus adapté pour décrire les données. Pour cela, Torr a proposé diverses extensions de son critère GRIC [Tor97, Tor98, Tor99, Tor02], dérivées des critères AIC et BIC, pour permettre la comparaison de l'homographie et de la géométrie épipolaire. Le critère GRIC, dont nous avons rappelé l'expression au paragraphe 3.3.2, repose sur le choix de nombreux paramètres qui dépendent de la distribution réelle des inliers et des outliers. En particulier, l'écart-type σ des résidus pour les inliers qui est supposé connu.

Comme alternative à ce critère de sélection de modèles, nous proposons d'utiliser la méthodologie *a contrario*, qui ne requiert aucune information *a priori* sur les données. Pour cela, le critère de validation présenté au paragraphe 4.1.1 doit être étendu à d'autres modèles géométriques.

4.3.2 Critère d'évaluation pour les transformations géométriques du plan

Le modèle de fond proposé pour la géométrie épipolaire (indépendance mutuelle et distribution uniforme des points d'intérêt) est très générique et peut donc être utilisé pour d'autres modèles géométriques. Nous nous intéressons dorénavant aux transformations géométriques du plan (similitude, transformation affine et homographie) qui sont souvent utilisées pour la reconnaissance d'objets. D'autres modèles géométriques que nous n'avons pas considérés pourraient également être étudiés (distorsion radiale, déformation polynomiale, *etc*.).

Soient $\mathbf{C}$ un ensemble de N correspondances aléatoires, et $\mathbf{S}'$ un sous-ensemble de $\mathbf{C}$, tel que $\#\mathbf{S}' = n$. On suppose que $\mathbf{C}$ suit l'hypothèse nulle $\mathcal{H}_0$ (définition 8). Dans le cas des transformations géométriques du plan, n est égal à 2, 3 ou 4 pour déterminer de manière exacte une unique[2] transformation d'ordre n, notée $\mathcal{T}_{\mathbf{S}'}$ entre les plans $\mathcal{P}$ et $\mathcal{P}'$ des images I et I' respectivement. L'erreur résiduelle de transfert dans le plan $\mathcal{P}'$ s'exprime maintenant comme la distance euclidienne $d(\mathcal{T}_{\mathbf{S}'}\mathbf{m}, \mathbf{m}') = ||\mathcal{T}_{\mathbf{S}'}\mathbf{m} - \mathbf{m}'||_2$ entre le point m' de l'image I' et le point $\mathcal{T}_{\mathbf{S}'}\mathbf{m}$.

Pour n'importe quelle correspondance $(\mathbf{m}, \mathbf{m}')$ de $\mathbf{C} \setminus \mathbf{S}'$, la probabilité conditionnellement à $\mathcal{H}_0$ que la distance $d(\mathcal{T}_{\mathbf{S}'}\mathbf{m}, \mathbf{m}')$ soit plus petite que α est bornée supérieurement par le rapport de l'aire du disque de rayon α divisé par l'aire A' de l'image I' : $\mathbb{P}_{\mathcal{H}_0}[d(\mathcal{T}_{\mathbf{S}'}\mathbf{m}, \mathbf{m}') \leq \alpha] \leq \pi\alpha^2/A' \quad \forall \alpha > 0$. Autrement écrit,

$$\mathbb{P}_{\mathcal{H}_0}\left[d(\mathcal{T}_{\mathbf{S}'}\mathbf{m}, \mathbf{m}')^2 \frac{\pi}{A'} \leq \alpha\right] \leq \alpha \ . \tag{4.9}$$

[2] à l'exclusion de cas dégénérés (alignement de plus de 2 points) qui sont rejetés par une procédure de test dans RANSAC.

Une expression similaire est obtenue en considérant l'erreur de transfert dans l'image I' en fonction de la transformation inverse $T_{\mathbf{S}'}^{-1}$. En considérant l'erreur maximale normalisée entre les deux images, nous avons finalement l'inégalité suivante :

$$\mathbb{P}_{\mathcal{H}_0}\left[\max\left(d(\mathcal{T}_{\mathbf{S}'}\mathbf{m},\mathbf{m}')^2\frac{\pi}{A'}, d(\mathbf{m},\mathcal{T}_{\mathbf{S}'}^{-1}\mathbf{m}')^2\frac{\pi}{A}\right)\leq\alpha\right]\leq\alpha\ . \tag{4.10}$$

Observons que cette probabilité dépend du *carré* de la distance entre deux points. Cette dernière expression suggère une nouvelle définition de la rigidité dans le cas des transformations géométriques du plan.

Soit un groupe de correspondances $\mathbf{S}$, sous-ensemble de $\mathbf{C}$ tel que $\mathbf{S}\cap\mathbf{S}'=\emptyset$. La $\mathcal{T}_{\mathbf{S}'}$-rigidité de $\mathbf{S}$ est définie comme :

$$\alpha(\mathbf{S},\mathcal{T}_{\mathbf{S}'}):=\max_{(\mathbf{m},\mathbf{m}')\in\mathbf{S}}\max\left(d(\mathcal{T}_{\mathbf{S}'}\mathbf{m},\mathbf{m}')^2\frac{\pi}{A'}, d(\mathbf{m},\mathcal{T}_{\mathbf{S}'}^{-1}\mathbf{m}')^2\frac{\pi}{A}\right). \tag{4.11}$$

La probabilité d'observer une $\mathcal{T}_{\mathbf{S}'}$-rigidité plus petite que α, pour un groupe $\mathbf{S}$ qui suit l'hypothèse nulle, est donc bornée par $\alpha^{\#\mathbf{S}}$:

$$\forall\alpha>0,\quad \mathbb{P}_{\mathcal{H}_0}\left[\alpha(\mathbf{S},\mathcal{T}_{\mathbf{S}'})\leq\alpha\right]\leq\alpha^{\#\mathbf{S}}.$$

Considérons maintenant l'ensemble $\mathcal{C}$ de correspondances de points d'intérêt entre deux images I et I'. De manière analogue au critère de validation proposé dans [MS04], nous définissons une mesure de qualité (NFA) pour les transformations géométriques du plan.

Définition 12 *Soit $\mathcal{C}=\{(m_i,m_i')\ ,\ i=1,\dots N\}$ un ensemble de N correspondances entre les images I et I'. Soit S un sous-ensemble de $\mathcal{C}$ de $\#S=K$ correspondances, tel que $K\leq N-n$. Pour $\varepsilon>0$ fixé, S est qualifié de groupe ε-significatif s'il existe un n-uplet $S'\subset C\setminus S$ tel que*

$$NFA(S,S'):=(N-n)\binom{N}{K}\binom{N-K}{n}\left(\alpha(S,T_{S'})\right)^K\leq\varepsilon\ . \tag{4.12}$$

La proposition 3 est également vraie pour cette définition du NFA (la démonstration est identique) : l'espérance du nombre de fausses alarmes (c'est-à-dire les sous-ensembles ε-significatifs qui suivent l'hypothèse nulle) est inférieure à ε.

> **Remarque 1 :**
> Les aires A et A' sont estimées à partir des distributions des points appariés dans chacune des images, par le procédé décrit au § 4.2.2.

4.3.3 Comparaison de modèles

La définition du NFA nous permet désormais de comparer différents modèles géométriques pour un ensemble de correspondances $\mathcal{C}$. Pour chaque modèle testé $\mathcal{M}$, l'algorithme AC-RANSAC est utilisé de manière à obtenir un sous-ensemble optimal S_{opt} dont la transformation est définie par un n-uplet S'_{opt}, et de significativité $\text{NFA}_{\mathcal{M}}(S_{\text{opt}},S'_{\text{opt}})$. Le modèle sélectionné est alors celui donnant le groupe le plus significatif :

$$\hat{\mathcal{M}}=\underset{\mathcal{M}}{\text{argmin}}\{\text{NFA}_{\mathcal{M}}(S_{\text{opt}},S'_{\text{opt}})\}\ .$$

Cette méthode de sélection de modèles géométriques est validée expérimentalement en section 4.5.3. Nous montrons sur des données synthétiques et réelles que la mesure de qualité NFA permet effectivement la sélection de modèles géométriques, constituant une alternative intéressante aux critères usuels.

Il est intéressant de comparer notre critère de sélection de modèles à celui proposé par Torr. Rappelons tout d'abord les expressions génériques du NFA et de GRIC [Tor97, Tor98] :

$$\begin{cases}\text{GRIC}=\sum_{i=1}^{N}\rho(r_i)+\lambda_1 Nd+\lambda_2 k\\ \log\text{NFA}_{\mathcal{M}}(S,S')=K\cdot\log\alpha(S,\mathcal{M}_{S'})+\log\mathcal{N}_T(N,K,n)\end{cases},$$

où, pour chacun des critères, les différents facteurs sont ainsi définis :

– GRIC : r_i désigne l'erreur résiduelle pour la correspondance d'indice i ; ρ est une fonction de pondération qui dépend de la variance σ^2 de l'erreur résiduelle des inliers, et d'un facteur λ_3 qui dépend de la proportion et de la distribution des outliers ; le coefficient λ_1 dépend de la distribution des outliers ; enfin, le coefficient λ_2 est une constante qui peut dépendre de N, le nombre de correspondances ;
– NFA : K est le cardinal du groupe S, $\mathcal{N}_T(N, K, n)$ désigne le nombre de tests qui dépend de K, n et de N.

Avec le critère GRIC, le premier terme correspond à la somme des N erreurs résiduelles pondérées. Le second terme permet de favoriser les modèles ayant une dimension plus faible ($d = 2$ pour les transformation planes, et $d = 3$ pour la matrice fondamentale). Le dernier terme permet de pénaliser les modèles ayant le plus de degrés de liberté (c'est-à-dire le nombre de paramètres $k = 4$, 6 et 8 pour les transformations géométriques du plan, et $k = 7$ pour la matrice fondamentale).

Dans notre cas, le premier terme du critère dépend de la rigidité $\alpha(S, \mathcal{M}_{S'})$, correspondant à l'erreur résiduelle normalisée maximale du groupe S selon la transformation $\mathcal{M}_{S'}$, et de K, le cardinal de S. Le second terme du NFA est le nombre de tests $\mathcal{N}_T(N, K, n)$, qui dépend en particulier de n qui varie donc selon le modèle choisi (c'est-à-dire $n = 2$, 3 et 4 pour les transformations géométriques du plan, et $n = 7$ pour la matrice fondamentale). Le nombre de tests tend ainsi à favoriser les modèles ayant besoin d'un faible nombre échantillons pour être définis. En reprenant les notations de GRIC, nous avons la relation suivante :

$$n = k \times (d' - d) \ ,$$

où $d' = 4$ est la dimension des correspondances, d la dimension de la variété du modèle, et k le nombre de paramètres du modèle. Ceci montre que le NFA est un critère de sélection qui traduit également un compromis entre la précision du modèle et sa complexité géométrique, en tenant compte à la fois de la dimension d du modèle et du nombre de paramètres k.

Le grand intérêt de notre critère en comparaison du critère GRIC est qu'**il ne requiert aucun réglage de paramètre**. De plus, la définition du NFA ne prend en compte que les K correspondances du groupe S parmi l'ensemble des N correspondances $\mathcal{C}$. Cela illustre le fait que notre critère permet à la fois de sélectionner le modèle géométrique $\mathcal{M}$ et le groupe d'inliers S qui lui correspond.

4.4 Détection multiple avec l'algorithme MAC-RANSAC

L'algorithme AC-RANSAC a été défini de manière à détecter une transformation unique pour un ensemble de données entachées d'erreurs. Diverses applications nécessitent la possibilité d'identifier plusieurs groupes de correspondances entre des images ; telles que

– Détection de plusieurs objets dans des photographies ;
– Détection multiple du même objet dans des photographies ;
– Segmentation du mouvement dans le cas d'objets avec des mouvements distincts en plus d'un éventuel mouvement du fond.

Nous avons présenté à la section 3.2.4 différentes approches utilisant l'algorithme RANSAC en vue de la détection de groupes multiples. Ces approches se déclinent en deux catégories : les méthodes utilisant RANSAC pour segmenter les données en plusieurs groupes qui sont ensuite séquentiellement agrégés, et les méthodes utilisant RANSAC de manière séquentielle de manière à détecter un seul groupe à la fois.

Nous introduisons dans cette section un nouvel algorithme, appelé MAC-RANSAC (Multiple A Contrario RANdom SAmple Consensus) qui adopte le second type d'approche, en utilisant *séquentiellement* l'algorithme AC-RANSAC. Ce choix est justifié par sa simplicité de mise en œuvre et par les nombreux avantages qu'il procure, notamment en raison du critère d'évaluation *a contrario*, qui permet à la fois de mesurer la qualité des groupes et de décider de leur validation. Toutefois, l'utilisation séquentielle de RANSAC n'est pas sans poser quelques problèmes et nous proposons deux critères pour s'en affranchir.

Dans la section suivante, nous allons étudier les avantages et les limitations de l'utilisation séquentielle de AC-RANSAC. Une vue d'ensemble de l'algorithme MAC-RANSAC est ensuite présentée dans la section 4.4.2, avant d'introduire dans les sections suivantes deux nouveaux critères.

4.4.1 Utilisation séquentielle de AC-RANSAC

Mise en œuvre L'utilisation séquentielle de AC-RANSAC est calquée sur l'algorithme séquentiel de RANSAC précédemment détaillé en table 3.2. À chaque fois qu'un groupe S_{opt} est détecté, puis optimisé, on élimine de l'ensemble de départ $\mathcal{C}$ les correspondances de ce groupe. Ensuite, AC-RANSAC est de nouveau utilisé sur les correspondances restantes, jusqu'à ce que plus aucun groupe significatif ne soit détecté.

À mesure que les groupes plus importants sont détectés, la taille relative des petits groupes augmente et facilite leur détection, ce qui requiert un nombre d'itérations moins élevé en comparaison des autres approches de détection multiple fondées sur RANSAC. Nous verrons dans la partie expérimentale que des objets représentant moins de 1% des correspondances peuvent ainsi être aisément détectés.

Afin de préserver la cohérence de la mesure de qualité (NFA), le nombre de correspondances initiales $N = \#\mathcal{C}$ n'est pas réévalué à chaque fois qu'un groupe est détecté. Le nombre de tests est ainsi identique pour des groupes de même taille pendant l'ensemble de la procédure. Cela permet de s'assurer que les NFA des différents groupes ne dépendent pas de l'ordre dans lequel ils sont détectés.

Un autre avantage procuré par une utilisation itérée de RANSAC est la sélection de modèles. Des groupes peuvent être ainsi évalués avec différents modèles, contrairement à des approches comme J-linkage [TF08] où le modèle est fixé pour tous les groupes.

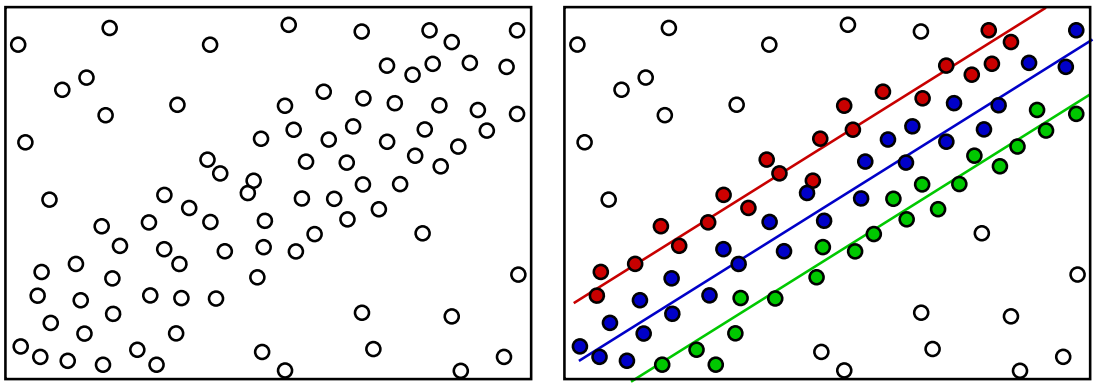
L'algorithme RANSAC étant défini pour estimer de manière robuste une unique transformation d'un ensemble de données, son utilisation séquentielle entraîne irrémédiablement quelques limitations que nous proposons de détailler dans le paragraphe suivant. Nous verrons ensuite ce que cela implique dans le cas de l'utilisation du critère de détection *a contrario*.

Limitations de l'utilisation séquentielle de RANSAC

1. **Détections de fausses transformations.** Lorsque tous les objets ont été détectés, RANSAC séquentiel requiert une dernière itération sur les données restantes afin de s'assurer qu'il n'existe plus de groupe à détecter. Le risque est donc de voir l'algorithme détecter plusieurs faux groupes avant de s'arrêter.
2. **Fusion de transformations proches.** La question de la fusion est par exemple évoquée dans [ZKM05, TF08] pour la détection de plans, et dans [Ste95] pour la reconstruction de surfaces à partir de données stéréoscopiques. Ce problème de fusion peut être interprété comme le résultat du comportement « glouton » de RANSAC, qui considère le meilleur groupe comme étant le plus grand sous-ensemble. Ce phénomène est d'autant plus important que le seuil de sélection r_{max} sur les résidus est permissif, ainsi que cela a été illustré en figure 3.6.
3. **Sur-segmentation en de multiples groupes.** Ce phénomène, opposé au précédent, se traduit par le fait de détecter pour un unique objet de multiples groupes au lieu d'un unique ensemble cohérent. Cela ce produit lorsque la tolérance sur l'erreur résiduelle r_{max} est sous-estimée. Ce phénomène est illustré par la figure 4.6.
4. **Détections d'« échos » liés à l'autosimilitude.** Ce terme désigne la détection de plusieurs transformations artificielles qui viennent faire écho à une unique transformation réelle. Cela se produit lorsqu'un objet est autosimilaire : un mur de brique, la façade d'un bâtiment, *etc*.

L'utilisation du critère *a contrario* permet de nous affranchir de plusieurs des limitations qui viennent d'être présentées.

L'un des intérêts majeurs de l'utilisation du critère *a contrario*, au lieu du critère classique de détection de RANSAC, est qu'il définit un *critère d'arrêt robuste* au processus séquentiel de recherche.



(a) Distribution de points (b) Exemple de sur-segmentation

FIG. 4.6 – *Illustration du problème de sur-segmentation.*

Le contrôle de l'espérance du nombre de fausses alarmes du critère *a contrario* permet de s'assurer en pratique qu'aucune fausse détection ne sera détectée parmi les données aberrantes restantes. Soulignons l'absence de paramètres à fixer *a priori*, à l'exception du seuil sur la significativité qui est toujours fixé à $\varepsilon = 1$, et du nombre d'itérations i_{max} pour la recherche d'un groupe. Nous verrons dans la partie expérimentale que ce critère d'arrêt est particulièrement efficace.

Par ailleurs, lorsque l'on considère la détection de plusieurs groupes indépendants, les éléments d'un groupe sont considérés comme des données aberrantes pour la transformation d'un autres groupes. Ces données aberrantes ne suivent cependant pas le modèle de fond (hypothèse d'indépendance mutuelle en particulier), ce qui pourrait poser problème. De ce point de vue, la mesure de rigidité proposée par Moisan et Stival est alors très intéressante car elle ne considère que les éléments d'un groupe pour évaluer sa significativité, et non les résidus des autres données à l'image de l'algorithme MLESAC [TZ00] ou du critère GRIC [Tor02]. *Il n'est donc pas nécessaire de modifier le modèle de fond utilisé pour la détection multiple.*

Un autre avantage dont tire profit l'utilisation séquentielle de AC-RANSAC est le principe d'échantillonnage ORSA. Avec l'approche classique de RANSAC et des approches par agrégation des hypothèses de RANSAC, l'échantillonnage n'est pas conditionné à la réussite des tirages effectués précédemment. Il faut alors attendre de générer un consensus suffisamment important, ou bien tirer un nombre suffisant d'échantillons afin de s'assurer que le groupe considéré est optimal. Avec AC-RANSAC, dès qu'un groupe de NFA plus petit que 1 est détecté, le processus de recherche est immédiatement stoppé ce qui représente un gain en temps de calcul dans le cadre de la multi-détection.

L'un des problèmes principaux de RANSAC est le réglage optimal du seuil r_{max} pour la sélection d'un groupe. Certaines approches (par exemple MLESAC) proposent d'estimer la valeur du seuil r_{max} à partir de la variance σ^2 de l'erreur résiduelle des inliers. Cependant, ce paramètre est généralement inconnu, notamment en raison de certains phénomènes qui ne sont pas pris en compte par le modèle géométrique utilisé (déformations en coussinet et barillet par exemple). Dans le cas d'une utilisation séquentielle de RANSAC, le rôle de ce seuil est encore plus critique car un mauvais choix se traduit par une fusion de plusieurs groupes ou bien la sur-segmentation d'un unique groupe. Le critère d'évaluation *a contrario* de AC-RANSAC est donc particulièrement intéressant car les seuils de validation sur l'erreur résiduelle (rigidité α) sont définis automatiquement, sans nécessiter de connaissance *a priori* sur les données utilisées. On observe ainsi que **ce critère permet d'éviter en pratique le problème de sur-segmentation**, lié au choix d'une tolérance spatiale trop faible.

Il reste néanmoins deux limitations à l'utilisation de AC-RANSAC séquentiel. La première concerne la détection des autosimilitudes. Ce type de détection est due à des correspondances artificielles liées à la répétition régulière d'un motif. Le problème de ces « échos » est qu'ils ne correspondent pas à de vraies transformations. Ils augmentent artificiellement le nombre d'objets détectés et introduisent une incerti-

tude sur la pose réelle de l'objet. Nous allons présenter plus en détail ce phénomène au paragraphe 4.4.3 et proposer un nouveau critère d'exclusion pour s'en affranchir. Notons qu'un cas particulier de la détection d'« échos » concerne la détection redondante du même objet en raison des correspondances redondantes. Cependant, le principe d'exclusion défini en section 4.2.1 permet d'éviter ce problème, comme nous le verrons en section expérimentale.

Le second problème est celui de la fusion de transformations similaires. Étant donné l'analogie de notre approche avec l'algorithme MINPRAN [Ste95], concernant le test d'une hypothèse nulle, il n'est pas surprenant d'observer le même phénomène avec le critère *a contrario*. Stewart réalise une analyse spécifique à ce problème dans [Ste97]. Il montre ainsi que dans différentes configurations impliquant deux plans (marche d'escalier, plans parallèles, ou en « triangle »), les différents critères des moindres carrés ainsi que MINPRAN détectent un groupe résultant de la fusion des deux plans. Pour éviter ce type de détection moyenne (voir l'illustration en figure 3.6), Stewart propose une extension de la mesure de qualité de MINPRAN pour examiner deux groupes simultanément, en utilisant une loi trinomiale. Cette nouvelle mesure de qualité lui permet d'améliorer la distinction entre les plans mais au prix d'un effort algorithmique beaucoup plus élevé. Sans rentrer dans les détails, on peut d'ores et déjà y voir une analogie avec la méthode de sélection de groupes, proposée par Sur *et al.* dans [CDD+07], où ce problème de fusion est également rencontré. Cependant, contrairement à cette approche, Stewart ne précise pas comment générer efficacement deux hypothèses de sous-groupes :

> *Unfortunately, the search of [sub-groups] is computationnaly expensive, and so the present implementation of MINPRAN2 uses a simple search heuristic that yields more biased result than the optimum shown here.*

De plus, il ne montre pas comment résoudre le problème de la fusion dans le cas général où plus de deux groupes sont impliqués. Cette limitation concernant MINPRAN conduisent Stewart et Miller à remarquer dans [MS96] :

> *Stewart's MINPRAN operator tolerates the large number of outliers in range images and identifies regions composed completely of outliers ; but MINPRAN's assumptions about the outlier distribution are not sufficient for extracting multiple surfaces.*

Nous nous intéresserons de plus près à ce problème de fusion de groupes au paragraphe § 4.4.4. Nous y introduirons un nouvel algorithme permettant à la fois de détecter les fusions de plusieurs transformations et d'identifier les différents groupes qui leurs correspondent.

4.4.2 Vue d'ensemble de l'algorithme MAC-RANSAC

L'ensemble de l'algorithme MAC-RANSAC est décrit en table 4.3. Il repose sur une utilisation séquentielle de AC-RANSAC, alternée avec différents critères définis dans les sections suivantes. Les seuils de détection pour l'ensemble des critères sont toujours fixés à $\varepsilon = 1$. Le seul paramètre défini par l'utilisateur est le nombre d'itérations maximum i_{max} que peut utiliser l'algorithme MAC-RANSAC pour détecter un groupe.

4.4.3 Filtrage des transformations d'autosimilitudes

Dans sa thèse sur le recalage d'image, P. Monasse décrit le problème de la détection de transformations multiples pour un même objet avec la transformée de Hough (voir la section 7.1 intitulée : « Quelle heure est-il ? » [Mon00]). Ce phénomène se produit lorsque l'objet détecté présente un fort degré d'**autosimilitude**, c'est-à-dire lorsqu'il est composé d'un motif qui se répète (voir par exemple la figure 4.7). Nous avons observé que ce phénomène d'« écho » se produit également avec une utilisation séquentielle de RANSAC. Ce problème se rencontre très fréquemment en raison de la forte autosimilitude des objets manufacturés qui nous entourent. Par exemple, Schaffalitzky et Zisserman exploitent ce principe d'autosimilitude dans les images afin d'en détecter les points et les lignes de fuites [SZ00].

TAB. 4.3 – *Algorithme de groupement multiple*

MAC-RANSAC

Entrées : Ensemble $\mathcal{C}$ de N correspondances non redondantes (critère 11) et nombre d'itérations i_{max}.

Initialisation du compteur $i := 0$ et de la liste de groupes $\mathcal{S} := \{\emptyset\}$.

1) **Détection :** Tant que $i < i_{max}$, tirage aléatoire d'un n-uplet $S' \subset \mathcal{C}$ puis

recherche du groupe $S \subset \mathcal{C} \setminus S'$ minimisant NFA(S, S').

– Si NFA$(S, S') < 1$, $(S_{\text{opt}}, S'_{\text{opt}}) := (S, S')$ puis passage à l'étape 2).

– Sinon $i := i + 1$. Si $i = i_{max}$, **arrêt de l'algorithme**.

2) **Optimisation (ORSA) :** Répéter $i_{max}/10$ fois le tirage de n-uplets $S' \subset S_{\text{opt}}$

et la recherche d'un groupe $S \subset \mathcal{C} \setminus S'$ minimisant NFA(S, S').

Si NFA$(S, S') <$ NFA$(S_{\text{opt}}, S'_{\text{opt}})$, $(S_{\text{opt}}, S'_{\text{opt}}) := (S, S')$.

3) **Détection de fusion de groupes :** Recherche d'un sous-groupe optimal de S_{opt} (algorithme 4.4).

– Si détection de fusion, obtention de 2 couples 1-significatifs (S_1, S'_1) et (S_2, S'_2).

– Sinon, $S_1 := S_{\text{opt}}$ et $S_2 := \emptyset$.

4) **Filtrage :** Élimination des correspondances autosimilaires avec S_1 dans $\mathcal{C}$ (critère 13),

puis élimination des correspondances de S_1 : $\mathcal{C} := \mathcal{C} \setminus S_1$.

5) **Itération :** Ajout de S_1 à la liste $\mathcal{S}$, initialisation du compteur $i := 0$.

– Si $S_2 = \emptyset$, retour à l'étape 1)

– Sinon, $(S_{\text{opt}}, S'_{\text{opt}}) := (S_2, S'_2)$ puis passage à l'étape 2).

Sortie : Liste de groupes disjoints $\mathcal{S}$.

Pourtant, à notre connaissance, seules quelques publications décrivent ce problème de la détection d'échos. Dans le cadre de la reconstruction 3D en environnement urbain, les auteurs de [ZK06a] en expliquent la cause : « ambiguities due to repetitive scene structures » . Dans un contexte similaire, les auteurs de [VL01] rapportent le même problème. Notons également que Lindenbaum propose dans [Lin97] une analyse des performances de reconnaissance d'objets en fonction des caractéristiques des données utilisées. Cette analyse prend notamment en compte le degré d'autosimilitude de l'objet recherché.

La figure 4.8 décrit le principe général de la détection d'autosimilitude. Une majorité des appariements entre les deux images correspondent à des points d'intérêt représentant physiquement le même objet. Des mises en correspondances multiples peuvent également être validées entre les structures répétitives de l'objet (dans notre exemple, il s'agit de la lettre "A" du mot "RANSAC"). Les correspondances entre les structures répétées sont alors groupées successivement en plusieurs consensus, dont un seul correspond à la vraie transformation. Le groupe principal ayant le plus de correspondances correspond à la transformation réelle de la pose de l'objet. Les autres groupes sont des transformations artificielles.

Comme nous l'avons déjà remarqué, les correspondances à l'origine de ces transformations artificielles sont dues à l'autosimilitude de l'objet, et contrairement aux correspondances redondantes, il n'est pas possible de les identifier comme telles *a priori*. Il est en effet nécessaire pour cela de détecter préliminairement la transformation principale. Il est important de noter que la transformation principale est théoriquement trouvée en premier en raison de son meilleur score. En effet, seule la « vraie » transformation peut expliquer globalement la nouvelle position de l'objet, et ce avec une précision au moins égale

FIG. 4.7 – *Exemple d'autosimilitude sur deux vues d'un bâtiment du château de Versailles. Le recalage correct des deux images nécessite un examen minutieux.*

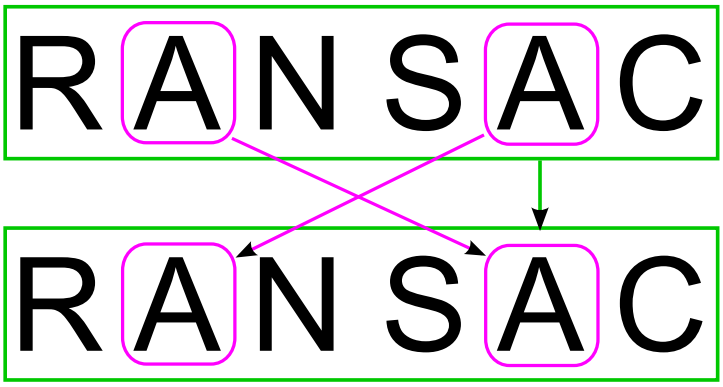


FIG. 4.8 – ***Principe des détections multiples pour un unique objet en raison de son autosimilitude.*** *Une unique transformation (en vert) permet d'expliquer globalement la relation entre les deux objets (ici le mot "RANSAC"). Cependant, en raison du phénomène d'autosimilitude (ici la lettre "A" qui est répétée), d'autres transformations artificielles sont identifiées (et symbolisées en magenta).*

aux transformations qui lui font écho. De fait, avec un plus grand nombre de correspondances, la vraie transformation est celle qui sera validée en premier par le processus d'optimisation.

Il nous faut donc un critère qui nous permette à ce stade (la transformation principale étant identifiée) de distinguer les correspondances multiples liées à une nouvelle occurrence de l'objet, de celles liées à son autosimilitude. Nous proposons pour cela de définir la région de l'objet détecté comme l'union des voisinages des points d'intérêt sélectionnés. La notion de voisinage a déjà été définie et utilisée pour le principe d'exclusion des redondances (définition 11). Nous avons fait le choix de définir le voisinage d'un point d'intérêt m_k comme un disque, dont le rayon dépend de l'échelle caractéristique Δ_k du point m_k. Les correspondances que l'on qualifie d'*autosimilaires* sont alors identifiées à l'aide du critère suivant :

Définition 13 (Correspondances autosimilaires) *Soit $\mathcal{G}$ un groupe validé, et $\mathcal{C}$ l'ensemble des correspondances restantes, de telle sorte que $\mathcal{C} \cap \mathcal{G} = \emptyset$. Soit $c_i = (m_i, m'_i)$ une correspondance entre les points $m_i \in I$ et $m'_i \in I'$. Elle est définie comme autosimilaire vis-à-vis du groupe $\mathcal{G}$ si* ***les deux conditions suivantes sont simultanément vérifiées*** *:*

- $\exists\, m \in \mathcal{G}$ *tel que* $\|m - m_i\|_2 < \min\{\Delta, \Delta_i\}$,
  et
- $\exists\, m' \in \mathcal{G}$ *tel que* $\|m' - m'_i\|_2 < \min\{\Delta', \Delta'_i\}$.

où $\|.\|_2$ désigne la norme euclidienne, et Δ_k représente l'échelle caractéristique du point d'intérêt m_k.

En pratique, à chaque fois qu'un nouveau groupe est validé, les correspondances autosimilaires de ce groupe sont éliminées des correspondances restantes $\mathcal{C}$ avant d'itérer AC-RANSAC. Ce nouveau principe d'exclusion spatiale est schématisé en figure 4.9, où l'on peut voir que les correspondances multiples liées à la répétition d'un objet sont effectivement préservées. Dans la partie expérimentale de ce chapitre, nous montrerons l'intérêt de ce simple critère pour l'élimination des transformations artificielles d'autosimilitude.

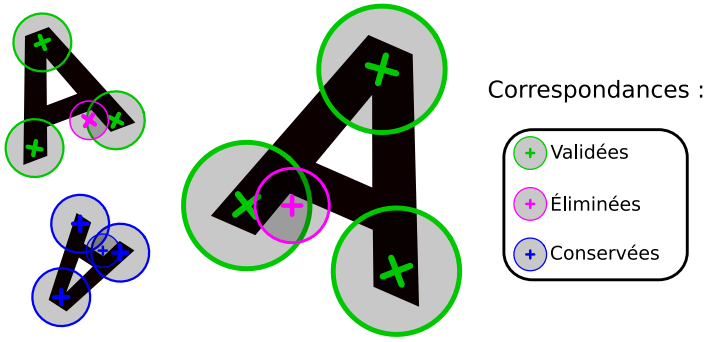


FIG. 4.9 – *Illutration du filtrage des correspondances autosimilaires après validation d'un groupe de correspondances (en vert), préservant les correspondances multiples (en bleu).*

> **Remarque 1 :**
> Une première possibilité que nous avons envisagée est de définir l'objet identifié en fonction de la région délimitée par l'enveloppe convexe des points d'intérêt regroupés. Les deux régions étant ainsi identifiées dans chacune des images, il suffit alors de d'éliminer les correspondances de points entre les deux régions. Le choix de l'enveloppe convexe ne convient malheureusement pas dans les cas où différents objets sont superposés. En effet, prenons l'exemple d'une scène statique avec un objet en mouvement. Le mouvement dominant qui correspond à l'arrière plan est d'abord détecté, puis toutes les correspondances restantes sont éliminés : de fait, l'objet en mouvement ne peut plus être détecté. Par ailleurs, une telle définition n'est pas robuste car il suffit d'un seul outlier sélectionné pour que la région délimitant l'objet soit fortement perturbée.

Bien que ce ne soit pas clairement identifié dans ce but, les auteurs de [VL01] proposent une approche qui pourrait également permettre d'éviter le phénomène d'écho. Ils utilisent en effet un filtre spatial pour éliminer les mises en correspondances non sélectionnées entre les deux régions identifiées comme similaires par la première transformation trouvée. Pour définir ces régions dans chacune des images, un masque binaire est construit à l'aide de différentes opérations. Elles consistent tout d'abord à recaler les deux images, puis à calculer la différence de niveau de gris, pour enfin utiliser un seuil afin de sélectionner les zones ayant approximativement le même niveau de gris. Le masque binaire obtenu étant approximatif, des opérations morphologiques sont ensuite utilisées pour améliorer le résultat. Le problème de cette approche est qu'elle repose sur la similitude des niveaux de gris, ce qui n'est pas robuste au changement de contraste de l'objet. Par ailleurs, elle nécessite le réglage d'un nombre important de paramètres. Une approche analogue est utilisée par A. Bartoli dans [Bar07], où une corrélation entre les parties des images correspondant aux plans détectés est utilisée pour vérifier l'estimation de la transformation.

4.4.4 *Détection de la fusion de plusieurs groupes*

Nous verrons dans la partie expérimentale, lorsque plus d'une transformation est en jeu, qu'il existe des situations où le critère de validation *a contrario* entraîne la détection d'un groupe résultant de la fusion de plusieurs transformations distinctes.

Ce problème de fusion est également présent dans [CDD+07], où Sur *et al.* proposent une méthode de groupement *a contrario* de points avec la transformée de Hough. Dans leur approche, les échantillons dans l'espace des paramètres sont initialement regroupés deux à deux de manière à construire une struc-

ture hiérarchique dyadique (c'est-à-dire un arbre dont le nombre de feuilles double à chaque nouveau nœud). Les auteurs définissent un critère de découpage, afin de pouvoir décider de la fusion ou de la séparation de deux groupes. Leur approche présente certaines analogies avec celle de Stewart [Ste95]. Cependant, cette comparaison se fait cette fois sans le moindre surcoût algorithmique, en raison de la structure dyadique de représentation des groupes qu'ils exploitent. Par ailleurs, un autre avantage de leur approche est que cela leur permet de détecter la fusion de plus de deux groupes, ce que MINPRAN est dans l'incapacité de réaliser.

Pour détecter la fusion de plusieurs groupes, nous proposons dans cette section un *critère de découpage* et un *algorithme de recherche récursive de découpage dyadique en sous-groupes* inspirés de ces travaux.

Critère de détection de fusion de deux groupes Pour simplifier l'étude du problème, nous nous intéressons dans un premier temps au cas de la fusion de deux transformations. On note S_0 un groupe de correspondances que l'on définit comme la réunion de deux sous-groupes S_1 et S_2, tels que $S_1 \cap S_2 = \emptyset$. On parle de fusion de groupes lorsque les transformations de S_1 et S_2, respectivement notée T_1 et T_2 sont *suffisamment* différentes[3]. La transformation T_0 du groupe S_0 est alors une transformation « moyenne » de T_1 et T_2.

La raison pour laquelle le groupe S_0 a été validé est que sa significativité (NFA) est plus faible que celles de chacun des deux sous-groupes S_1 et S_2, bien qu'ils aient chacun une plus faible rigidité :

$$\begin{cases} \text{NFA}(S_0 = S_1 \cup S_2, S'_0) & < & \min\left\{\text{NFA}(S_1, S'_1)\,,\ \text{NFA}(S_2, S'_2)\right\} \\ \alpha\big(S_0 = S_1 \cup S_2,\ \mathcal{T}_{S'_0}\big) & \geq & \max\left\{\alpha\big(S_1, \mathcal{T}_{S'_1}\big)\,,\ \alpha\big(S_2, \mathcal{T}_{S'_2}\big)\right\} \end{cases} . \tag{4.13}$$

Ce résultat est lié au comportement glouton de AC-RANSAC qui cherche le groupe minimisant le NFA. Le problème est que la relation (4.13) est également vraie lorsque S_0 décrit une unique transformation.

Afin de pouvoir détecter la fusion de deux groupes, nous devons donc être en mesure de comparer le groupe $S_0 = S_1 \cup S_2$ avec les deux groupes S_1 et S_2 simultanément. Nous proposons de réaliser cette comparaison à l'aide du critère naïf suivant, qui suppose que l'on peut estimer la mesure de qualité de l'union de deux groupes indépendants par le produit de leurs NFA respectifs.

Définition 14 (Critère de découpage) *Soit S_0 un groupe ε-significatif de $\mathcal{C}$, de transformation $\mathcal{T}_{S'_0}$. S'il existe S_1 et S_2, deux sous-ensembles disjoints de S_0, ainsi que deux n-uplets disjoints S'_1 et S'_2 de $S_0 \setminus \{S_1 \cup S_2\}$, tels que les conditions suivantes sont vérifiées :*
· $NFA(S_1, S'_1) \leq \varepsilon$ et $NFA(S_2, S'_2) \leq \varepsilon$
· $NFA(S_1, S'_1) \times NFA(S_2, S'_2) < NFA(S_0, S'_0)$,
alors, les 2 groupes S_1 et S_2 sont validés en tant que deux objets distincts à la place de S_0.

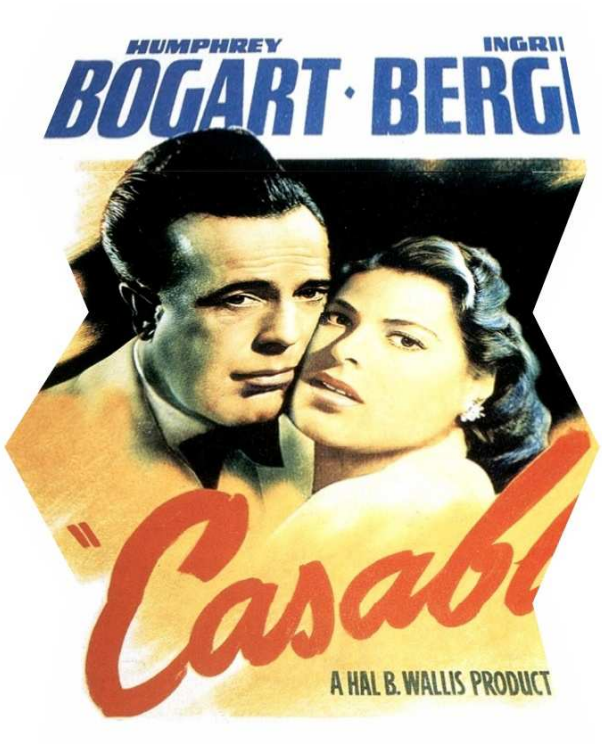
On peut montrer cependant que, contrairement au NFA, la quantité correspondant au produit des NFA ne permet pas de contrôler l'espérance du nombre de fausses alarmes. Toutefois, en raison de sa simplicité et de son efficacité qui sera démontrée expérimentalement, nous utilisons ce critère de découpage en pratique pour décider si deux groupes correspondent à une même transformation.

Afin de pouvoir exploiter le critère de découpage 14, nous allons maintenant définir une stratégie de découpage récursif en sous-groupes.

Recherche récursive de découpage dyadique en sous-groupes La figure 4.10 montre un exemple de situation où l'utilisation du critère *a contrario* conduit à la fusion de plusieurs groupes. L'image originale (poster du film « Casablanca », en figure 4.10(a)) est découpée en cinq bandes horizontales auxquelles sont appliquées des homographies différentes (figure 4.10(b)), de manière à ce que l'affiche obtenue corresponde à un pliage en « accordéon ». L'algorithme MAC-RANSAC détecte un groupe principal

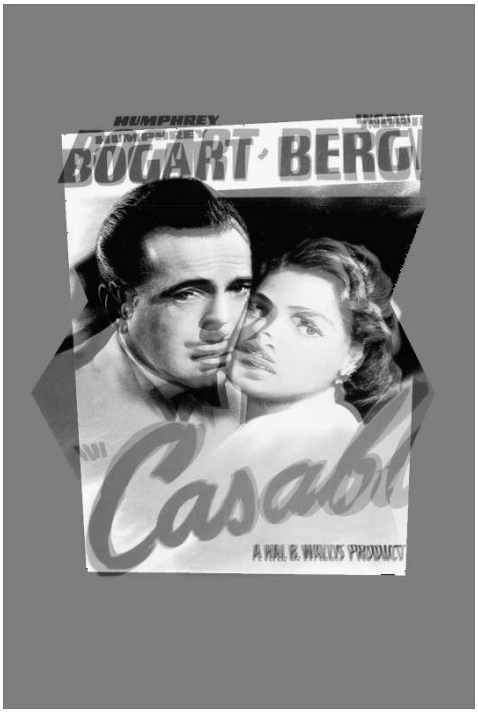
[3] En pratique, en raison des erreurs qui entachent les données, il est impossible que les deux transformations T_1 et T_2 soient rigoureusement identiques, même lorsque il n'y a pas de fusion. C'est la raison pour laquelle la définition de la fusion est quelque peu ambiguë.

(a) Image originale.



(b) Transformation synthétique en 5 homographies (pliage).

(c) Utilisation de AC-RANSAC.



(d) Recalage avec la transformation principale.

FIG. 4.10 – *Illustration de la fusion de plusieurs groupes.* *La figure 4.10(b) est obtenue en appliquant une transformation synthétique à l'image 4.10(a). L'algorithme AC-RANSAC détecte un groupe prépondérant, résultant de la fusion de cinq groupes dont les transformations sont distinctes (figure 4.10(c)). La transformation obtenue est peu précise, comme en atteste le recalage illustré en figure 4.10(d).*

qui résulte de la fusion approximative des cinq groupes (figure 4.10(c)), bien que le modèle géométrique utilisé soit correct (homographie). Quelques points restants sont groupés selon deux autres transformations incorrectes. Dans cet exemple difficile, la détection de la fusion suppose de pouvoir identifier puis comparer ces cinq sous-groupes. Pour contourner la difficulté de l'identification directe de tous les sous-groupes impliqués dans la fusion, nous proposons un algorithme dont le but est *d'identifier un seul de ces sous-groupes*. Pour cela, nous définissons une stratégie de découpage en deux sous-groupes qui est appliquée récursivement sur le groupe S_0 :

Recherche de deux sous-groupes Nous définissons S_1 comme le plus petit des deux sous-groupes S_1 et S_2 appartenant à S_0. Comme le remarque Stewart dans [Ste95], on vérifie alors la relation suivante : $\#S_1 \leq \#S_0/2 \leq \#S_2$. Le sous-groupe S_1 est obtenu en cherchant (par échantillonnage aléatoire) un n-uplet $S'_1 \subset S_0$ qui *minimise la quantité NFA*(S_1, S'_1) sous la contrainte que $\#S_1 \leq \frac{\#S_0}{2}$. Ce principe de recherche est illustré par la figure 4.11. Le sous-groupe S_2 est ensuite obtenu en cherchant un n-uplet S'_2 parmi les correspondances restantes de S_0 qui minimise NFA(S_2, S'_2). Le critère 14 nous permet alors de savoir si ce découpage est viable.

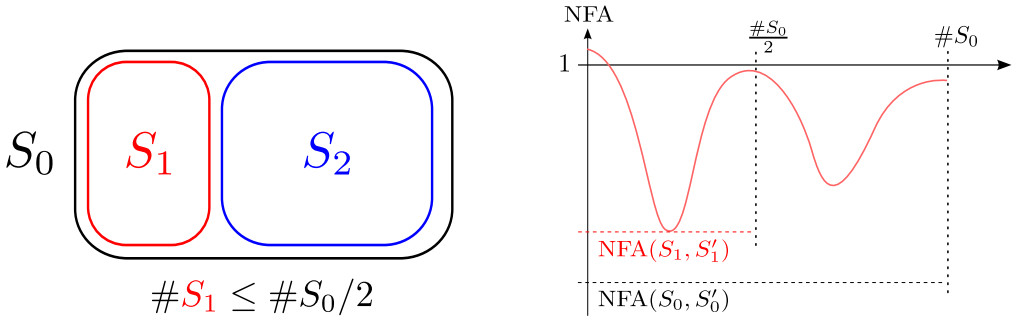


FIG. 4.11 – *À Gauche : lorsque l'on cherche à découper de manière optimale un groupe S_0 en deux sous-groupes, on recherche d'abord un premier sous-groupe S_1 dont le cardinal n'excède pas la moitié de celui de S_0. Le second sous-groupe est constitué initialement des points restants. À Droite : le groupe S_1 est détecté en tant que minimum local du NFA à la condition que son cardinal n'excède pas la moitié de celui de S_0.*

Découpage dyadique Tant que le critère valide le découpage, on applique **récursivement** cette stratégie de découpage en deux sous-groupes au groupe $S_1^{(k)}$ qui vient d'être identifié à l'itération k. Ce principe est illustré en figure 4.12 dans le cas de la fusion de 5 groupes.

Le processus s'arrête lorsque le dernier sous-ensemble $S_1^{(k)}$ trouvé correspond à une transformation unique : soit lorsque l'on n'a pas trouvé 2 nouveaux sous-groupes ε-significatifs, soit lorsque le critère de découpage rejette le découpage de $S_1^{(k)}$ en $S_1^{(k+1)}$ et $S_2^{(k+1)}$.

Si au terme de cette recherche le groupe S_0 est découpé en $S_1^{(k)}$ et $S_2^{(k)}$, seul le groupe $S_1^{(k)}$ est validé. Le groupe $S_2^{(k)}$ étant ε-significatif, il est quant à lui considéré comme une détection : l'algorithme MAC-RANSAC passe donc en phase d'optimisation (étape 2 de l'algorithme 4.3).

Cet algorithme de découpage que nous venons de décrire est résumé en table 4.4.

> **Remarque 2 :**
> Nous avons vu que le groupe S_0, s'il résulte d'une fusion de plusieurs groupes, est estimé à partir d'une transformation très approximative $\mathcal{T}_{S'_0}$. Le groupe S_0 ne capture donc pas nécessairement l'ensemble des correspondances des différents objets regroupés. Le groupe $S_1^{(k)}$ validé risque alors d'être incomplet (risque de sur-segmentation). Pour éviter cela, on lui ajoute toutes les correspondances $\tilde{S}_1$ de l'ensemble $\mathcal{C} \setminus S_0$ dont la rigidité est telle que $\alpha(\tilde{S}_1, S'_1) \leq \alpha(S, S'_1)$.

La figure 4.13 montre le résultat de l'application de l'algorithme 4.4 de découpage sur l'exemple synthétique des images 4.10(a) et 4.10(b). L'utilisation de notre procédure récursive de recherche de

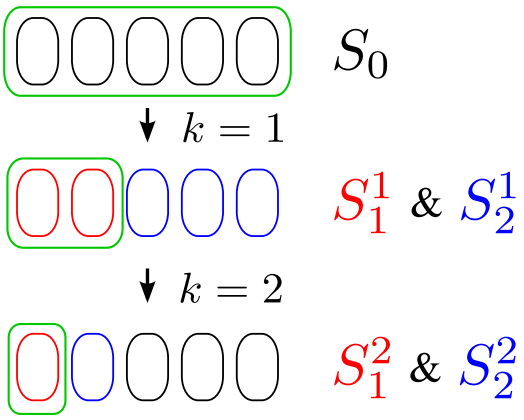


FIG. 4.12 – Détermination d'un sous-groupe optimal par découpage récursif en 2 sous-groupes.

TAB. 4.4 – *Algorithme de recherche récursive du plus petit sous-ensemble optimal d'un groupe.*

Algorithme de découpage 4.4

Entrées : Groupes $S_0, S'_0 \subset \mathcal{C}$ tels que $\text{NFA}(S_0, S'_0) < 1$ et $S'_0 \cap S_0 = \emptyset$. Nombre d'itérations i_{max}.

Initialisation du sous-groupe $S_1^{opt} := \{\emptyset\}$.

1) **Recherche d'un sous-groupe optimal :** Tirage de i_{max} n-uplets $S'_1 \subset S_0$,
avec recherche de $S_1 \subset S_0 \setminus S'_1$ minimisant $\text{NFA}(S_1, S'_1)$ tel que $\#S_1 < \frac{\#S_0}{2}$.
Si le groupe optimal est tel que $\text{NFA}(S_1, S'_1) \geq 1$: arrêt de l'algorithme.

2) **Définition du complémentaire :** $S_2 := S_0 \setminus S_1$ et $S'_2 = S'_0$.
Tirage de i_{max} n-uplets $S'_2 \subset S_0 \setminus S_1$, avec recherche de S_2 minimisant $\text{NFA}(S_2, S'_2)$.
Si le groupe optimal est tel que $\text{NFA}(S_2, S'_2) \geq 1$, arrêt de l'algorithme.

3) **Validation de la segmentation :** Si $\text{NFA}(S_1, S'_1) \times \text{NFA}(S_2, S'_2) < \text{NFA}(S_0, S'_0)$,
définition de $S_1^{opt} := S_1$ et $S_2^{opt} := S_2$. Retour à l'étape 1) avec $S_0 := S_1$ et $S'_0 := S'_1$.
Sinon, arrêt de l'algorithme.

Sortie : Sous-ensemble optimal S_1^{opt}.

– Si $S_1^{opt} \neq \emptyset$, S_1^{opt} est validé à la place de S_0,
et S_2^{opt} est utilisé pour initialiser l'étape 2 d'optimisation de MAC-RANSAC.

– Sinon S_0 est validé et retour à l'étape 1 de détection de MAC-RANSAC.

sous-groupes, illustrée par le schéma 4.12, permet de retrouver chacun des groupes (figure 4.13(a)). La précision de l'estimation de la pose sur chacune des parties de l'affiche est alors améliorée, comme que le montre les 5 recalages obtenus en la figure 4.13(b).

Pour des images naturelles, où le modèle géométrique utilisé n'est qu'une approximation de la transformation réelle de l'objet, on pourrait s'attendre à ce que notre algorithme détecte plusieurs petits groupes lorsqu'il n'y a qu'un seul objet à détecter. Nous allons voir dans la partie expérimentale (§ 4.5.2.1) que cela n'est pas le cas en pratique, et ce, toujours en raison de la propension du critère *a contrario* à fusionner les groupes. C'est une procédure robuste que l'on peut donc systématiquement appliquer en pratique.

(a) Résultat du groupement par MAC-RANSAC avec utilisation de l'algorithme de découpage récursif. Les cinq parties de l'affiche sont correctement identifiées.

(b) Recalage avec les 5 transformations estimées.

FIG. 4.13 – *La figure (4.13(a)) montre l'application de MAC-RANSAC avec l'algorithme 4.4 sur les images de la figure 4.10(c). Les 5 objets sont correctement détectés, ce qui permet d'accroître la précision de leurs transformations respectives (voir les 5 différents recalages en figure 4.13(b)).*

4.5 Validation expérimentale

Cette section est dédiée à l'évaluation des différents principes introduits à la section précédente. La section 4.5.1 est consacrée à l'étude des filtres proposés pour les redondances et les autosimilitudes. Dans la section 4.5.2, nous démontrons l'intérêt pratique de notre approche pour la reconnaissance d'objets multiples. Nous traitons le cas de la détection de différents objets, ainsi que de la répétition d'un même objet. La robustesse du découpage en sous-groupes y est également illustrée. Nous nous intéressons à la sélection de modèles géométriques dans la section 4.5.3, où des expériences sont à la fois réalisées sur des images synthétiques et sur des images réelles.

Dans toute cette partie expérimentale, nous utilisons le critère de mise en correspondance *a contrario* défini au chapitre 2, I étant l'image requête et I' l'image considérée comme la base de données. Le critère de validation de mise en correspondance est fixé à $\varepsilon = 1$ pour toutes les expériences. De même, le critère de validation de MAC-RANSAC est également fixé à $\varepsilon = 1$. Rappelons que le seul paramètre de la méthode est le nombre de tirages maximum choisi par l'utilisateur, que nous avons fixé à $i_{max} = 10000$. Si un objet est détecté (NFA$<$1), l'optimisation ORSA est réalisée sur $i_{max}/10$ itérations. Sauf mention contraire, l'ensemble de la chaîne de traitement de MAC-RANSAC est systématiquement appliquée, incluant les différents filtres ainsi que les procédures de normalisation et l'algorithme de découpage en sous-groupes.

4.5.1 Évaluation des différents filtres proposés

4.5.1.1 Élimination des détections redondantes

Nous avons présenté au paragraphe 4.2.1 une procédure de filtrage préliminaire des correspondances redondantes, afin de s'assurer de la validité du modèle de fond pour les fausses détections.

Nous avons vu par ailleurs que ce filtre permet d'éviter les détections redondantes du même objet, tout en préservant les correspondances liées à des occurrences multiples de l'objet recherché. Dans la figure 4.14, nous présentons une expérience de reconnaissance d'un objet apparaissant plusieurs fois dans une seconde image (figure 4.14(a)). Sans utilisation de ce filtre, on détecte chaque d'objet de manière redondante (figure 4.14(b)). Avec le principe d'exclusion, chacune des occurrences de l'objet de la première image est bien détectée une seule fois dans la seconde image.

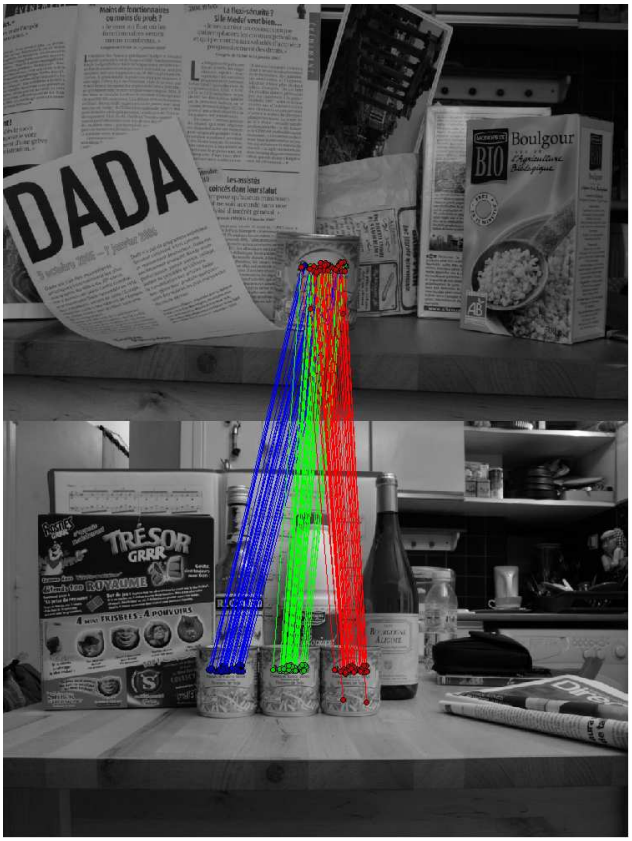
4.5.1.2 Élimination des détections d'échos

Au paragraphe 4.4.3, nous nous sommes intéressés à la détection des objets présentant une autosimilitude. De tels objets ont la particularité d'avoir des sous-parties identiques, ce qui a pour effet de créer des « échos » lors de la procédure de détection. Un exemple a été donné en figure 4.7 avec deux photographies d'un bâtiment sous différents angles de vue. Lorsque l'on utilise MAC-RANSAC sans utiliser de filtre spécifique, de nombreux groupes sont détectés (figure 4.15(b)) : le premier groupe (le plus significatif) correspond à la « vraie » transformation, tandis que les groupes suivants sont des détections d'autosimilitude entre les deux objets qui viennent d'être détectés. Un exemple de ce type de transformation est donné en figure 4.15(c). L'utilisation du principe d'exclusion des correspondances autosimilaires après la détection du premier groupe permet de supprimer tout phénomène d'écho, comme le montre la figure 4.15(d). Le premier groupe détecté après optimisation est celui correspondant à la transformation recherchée, comme en atteste le recalage (homographie) réalisé en figure 4.15(e).

Afin d'illustrer le fait que ce filtrage des autosimilitudes préserve à la fois les correspondances multiples et les correspondances des autres objets non détectés, nous proposons une deuxième expérience. La figure 4.16(a) montre la paire d'images utilisée : dans la première image, les deux objets recherchés sont superposés de manière à démontrer l'intérêt d'utiliser l'union des régions d'intérêt plutôt que l'enveloppe convexe des points d'intérêt. Le résultat est bien celui souhaité (figure 4.16(b)) : après reconnaissance du groupe principal (la boîte de céréale), les deux occurrences de la boîte circulaire sont bien détectées. Sans préservation des correspondances multiples, seule la première transformation est détectée.

(a) Paire d'images analysée.

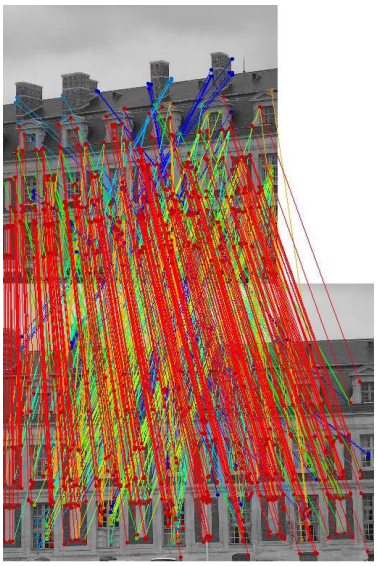
(b) Sans principe d'exclusion, six groupes sont détectés au lieu de trois.



(c) Avec principe d'exclusion, les trois groupes sont correctement identifiés.

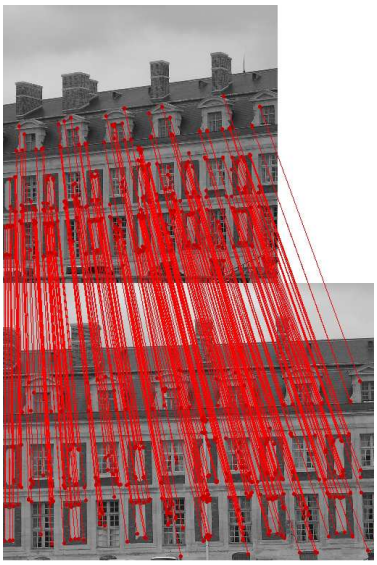
FIG. 4.14 – *Illustration de l'intérêt du filtre d'exclusion des redondances pour le problème de la détection répétée de la même transformation.*

(a) Paire d'images d'une façade de bâtiment avec une forte autosimilitude



(b) Détection de 7 groupes sans filtrage des autosimilitudes

(c) Exemple de transformation liée à l'autosimilitude

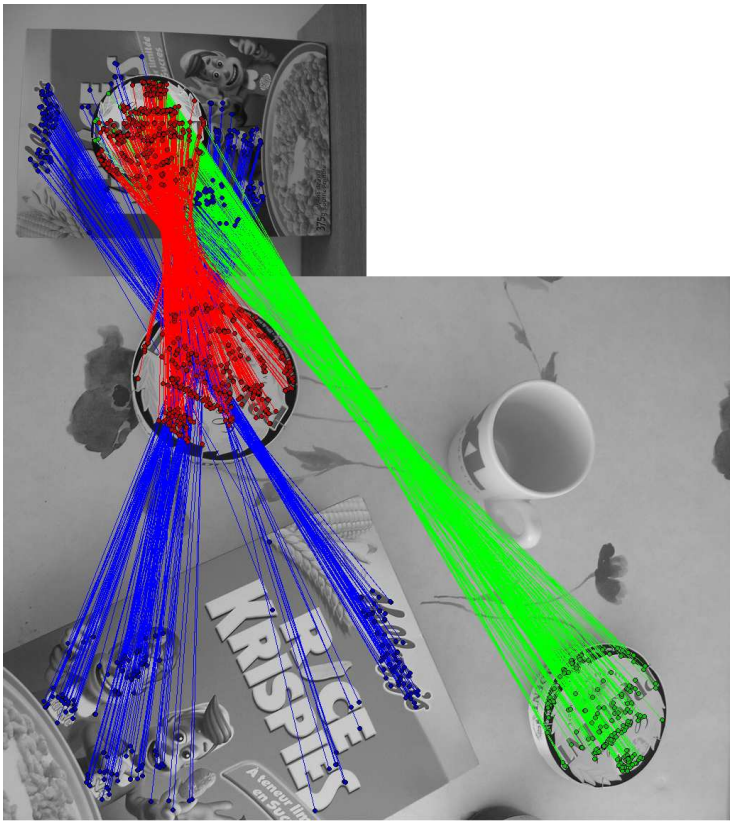


(d) Détection d'un unique groupe avec le filtrage des autosimilitudes

(e) L'unique transformation estimée est correcte

FIG. 4.15 – *La paire d'image 4.15(a) présente un fort degré d'autosimilitude. L'utilisation de MAC-RANSAC sans filtrage des correspondances autosimilaires conduit à la détection de plusieurs groupes dont la transformation est artificielle (figures 4.15(b) et 4.15(c)). Le filtrage des correspondances autosimilaires permet lever l'ambiguïté sur la position de l'objet : une unique transformation correcte est estimée à partir du groupe le plus significatif (figures 4.15(d) et 4.15(e)).*

(a) Paire d'images analysée.



(b) Reconnaissance de chacun des objets superposés.

FIG. 4.16 – *Résultat de MAC-RANSAC pour la détection de deux objets superposés dans la première image. Le filtrage par le critère d'union des masques préserve à la fois les objets qui apparaissent plusieurs fois, mais également les objets superposés.*

4.5.2 Évaluation de la détection multiple

Cette section se compose de trois parties. Dans la première est tout d'abord démontré l'intérêt pratique de l'algorithme 4.4 de découpage récursif en sous-groupes. Ensuite nous présentons deux types d'expériences illustrant la faculté de MAC-RANSAC à détecter séquentiellement, d'une part, plusieurs objets (§ 4.5.2.2), et d'autre part, les occurrences multiples d'un même objet (§ 4.5.2.3).

4.5.2.1 Évaluation de l'algorithme de découpage récursif

Nous illustrons ici l'intérêt du critère de découpage en sous-groupes par deux expériences.

La première expérience reproduit l'expérience synthétique étudiée au paragraphe 4.4.4 (affiche "Casablanca", figure 4.10). Ici nous avons plié un morceau de carton en trois morceaux, et photographié l'objet avant et après pliage 4.17. Comme nous considérons exclusivement les transformations rigides,

la seule interprétation possible de cette scène est que trois objets distincts ont chacun une transformation différente (une homographie). Une fois encore, sans utilisation du critère de sélection de sous-groupes (figure 4.17(c)), l'interprétation de la scène est mauvaise : un groupe dominant est obtenu (en rouge), correspondant à la moyenne des différentes transformations ; le second groupe détecté (en bleu) correspond à une transformation correcte mais qui ne recouvre pas toutes les correspondances concernées. L'utilisation de notre procédure de découpage récursif en sous-groupes permet au contraire de grouper avec précision les correspondances selon trois ensembles corrects, ainsi que leur transformation (figure 4.17(b)).

La seconde expérience démontre l'intérêt de notre approche dans le cas où le modèle géométrique utilisé n'est pas approprié. Il s'agit de la paire d'images de la tour de Pise (figure 4.18(a)) précédemment utilisée au chapitre 2. Il s'agit de deux vues d'une scène 3D, mais le mouvement de la caméra entre les deux prises de vue est faible en comparaison de la distance à l'objet. On peut donc considérer en première approximation que la transformation peut être approchée par une homographie. Sans procédure de découpage en sous-groupes, un groupe dominant est trouvé (voir la figure 4.18(b)) mais le recalage obtenu de la tour est très imprécis (figure 4.18(c)). Avec l'exploration des sous-groupes, deux groupes sont détectés (figures 4.18(d) et 4.18(f)), permettant un recalage beaucoup plus précis de la tour (figure 4.18(d)).

Dans toutes les expériences présentées par la suite dans cette section expérimentale, le critère de découpage est systématiquement utilisé afin de démontrer sa robustesse, même dans des situations où il n'est pas nécessaire. Ceci permet d'illustrer le fait que le critère de détection ne conduit pas à une sur-segmentation des objets détectés.

4.5.2.2 Détection de plusieurs objets

Nous expérimentons trois types de scénario dans cette section. Dans le premier, une caméra en mouvement filme une scène fixe. C'est le cas de la séquence d'image [Pol] qui est utilisée dans [PVGV+04] pour faire de la reconstruction 3D. Nous utilisons deux images de cette séquence (figure 4.19(a)), pour en extraire puis mettre en correspondance des points d'intérêt. Cette scène représente un bâtiment en forme de 'L', qui présente certaines autosimilitudes. Nous utilisons MAC-RANSAC avec la transformation épipolaire (figure 4.19(b)) et avec l'homographie (figure 4.19(c)). Comme on peut s'y attendre, un seul groupe est détecté avec la géométrie épipolaire. Cependant, avec le modèle de la géométrie projective, la meilleure interprétation de la scène consiste à segmenter les correspondances de points en plusieurs plans : on retrouve alors trois groupes différents correspondant aux différents plans du bâtiment, ainsi qu'un groupe supplémentaire lié aux arbres en arrière-plan (groupe en bleu).

> **Remarque 1 :**
> Notons que d'autres plans pourraient être détectés : le sol et les pans du toit de chacune des parties du bâtiment. Cependant, en l'absence de structures contrastées sur ces plans, il n'y a pas de points d'intérêt détectés.

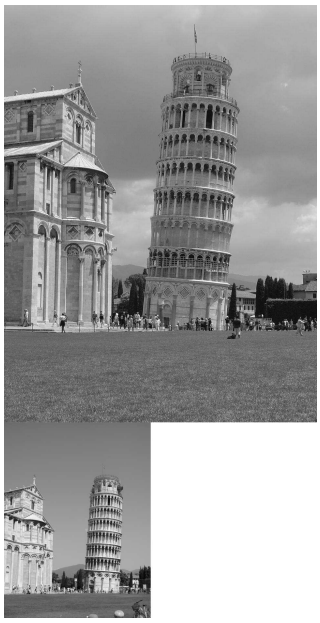
Dans le second scénario examiné, la caméra est encore en mouvement, mais les objets de la scène ne sont plus statiques. Dans la figure 4.20(a), deux photographies sont prises d'une scène où un objet a été déplacé entre les deux prises de vue. Avec la géométrie épipolaire (figure 4.20(b)) , deux groupes sont alors identifiés : un groupe dominant (en rouge) correspondant à la partie statique de la scène, et un groupe (en bleu) correspondant au téléphone qui a été déplacé entre les deux prises de vues. Si l'on recherche des homographies (figure 4.20(b)), on retrouve le même groupe pour le téléphone (en jaune), tandis que le reste de la scène est fragmenté en différents plans : le plan du livre (en vert), la souris (en bleu foncé), la télécommande (en rouge) et un dernier plan (en cyan) qui regroupe la table, un boîtier de CD et le dessus de l'ordinateur portable. Il est intéressant de voir que le critère de découpage a préféré ici regrouper les trois derniers objets plutôt que de les sursegmenter. La raison en est la suivante : les correspondances de ces trois objets représentent trois groupes de points distincts, répartis sur des plans parallèles très proches ; pour le critère de découpage, il est inutile de distinguer ces plans les uns des autres.

(a) Paire d'images analysée

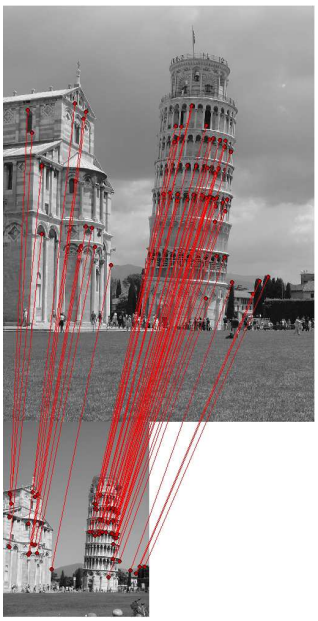
(b) Groupement de correspondances avec le critère de découpage

(c) Groupement de correspondances sans le critère de découpage

FIG. 4.17 – *Dans cette expérience, une affiche est pliée de manière à constituer trois plans (figure 4.17(a)). La mesure de qualité de AC-RANSAC tend à fusionner les transformations d'objets pourtant distincts (figure 4.17(c)). Le critère de découpage que nous avons défini pour comparer deux groupes, combiné à la recherche récursive de sous-groupes, permet de distinguer les trois plans (figure 4.17(b)).*

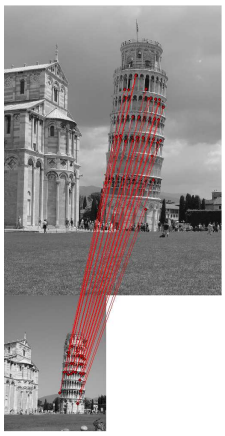


(a) Deux vues d'une scène 3D

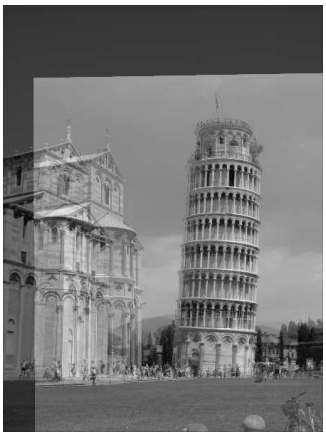


(b) Groupement de correspondances sans le critère de découpage en sous-groupes

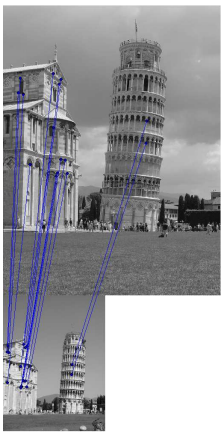
(c) Le recalage de la tour est très imprécis.



(d) Premier sous-groupe.



(e) Recalage obtenu avec le premier sous-groupe.



(f) Second sous-groupe.

(g) Recalage obtenu avec le second sous-groupe.

FIG. 4.18 – *Recherche d'objets en commun d'une scène 3D en vue du recalage par une homographie. Figure 4.18(b) : sans le critère de découpage, un unique groupe est obtenu. Le recalage est très imprécis (figure 4.18(c)). Le découpage en sous groupe est validé par le critère (figure 4.18(d) et 4.18(f)). Ceci permet d'améliorer le recalage de la tour (figure 4.18(e)).*

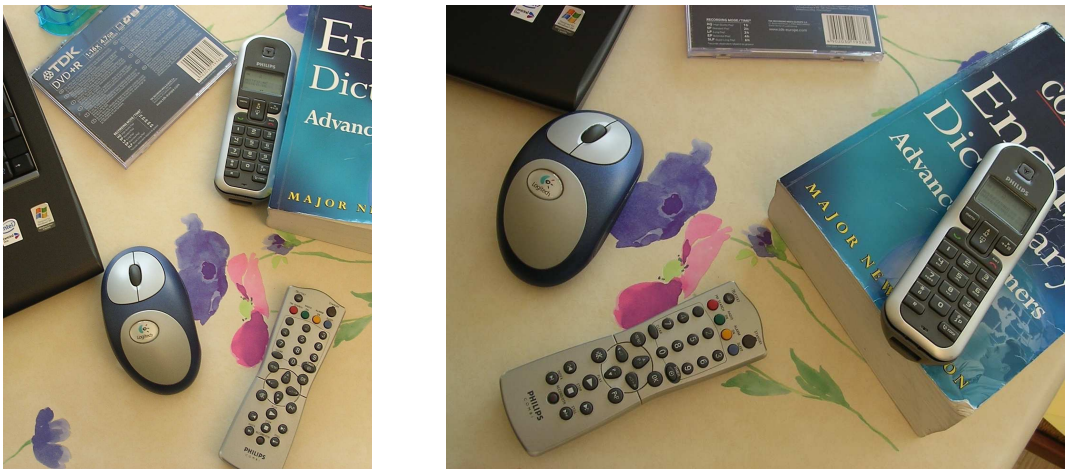
Dans le dernier exemple, nous illustrons le cas de la reconnaissance d'objets dans des contextes différents. Nous avons utilisé deux images de la base [PLRS04] (voir la figure 4.21(a), où deux objets sont communs à chacune des images. L'homographie (figure 4.21(b)) tout comme la géométrie épipolaire (figure 4.21(c)) permettent à MAC-RANSAC de détecter deux groupes. Dans le cas de l'homographie, les deux groupes permettent de localiser précisément les deux objets, tandis que dans le cas de la géométrie épipolaire, un objet est segmenté en deux groupes. Ceci est due à l'ambiguïté sur la position 3D de l'objet cylindrique vis à vis de la peluche (pour plus de détail sur la géométrie épipolaire, voir l'annexe C). Nous reviendrons sur ce phénomène au dernier paragraphe (§ 4.5.4).

(a) Paire d'images utilisée (frames 8 et 13 d'une séquence de 27 images [Pol])

(b) Unique groupe obtenu avec la géométrie épipolaire

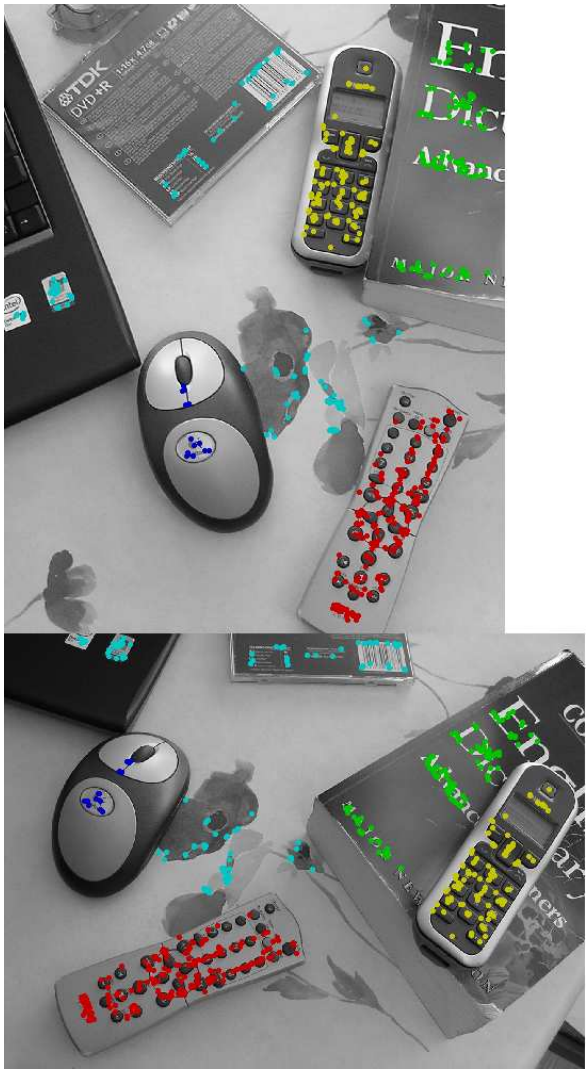
(c) 4 groupes obtenus avec l'homographie

FIG. 4.19 – *La figure (4.19(a)) montre deux images tirées d'une séquence vidéo [Pol]. Le groupement avec la transformation épipolaire donne un seul groupe (4.19(b)). Le groupement avec l'homographie donne 4 groupes (4.19(c)) : 3 plans correspondants aux murs du bâtiment, et un groupe supplémentaire pour les arbres en arrière plan.*



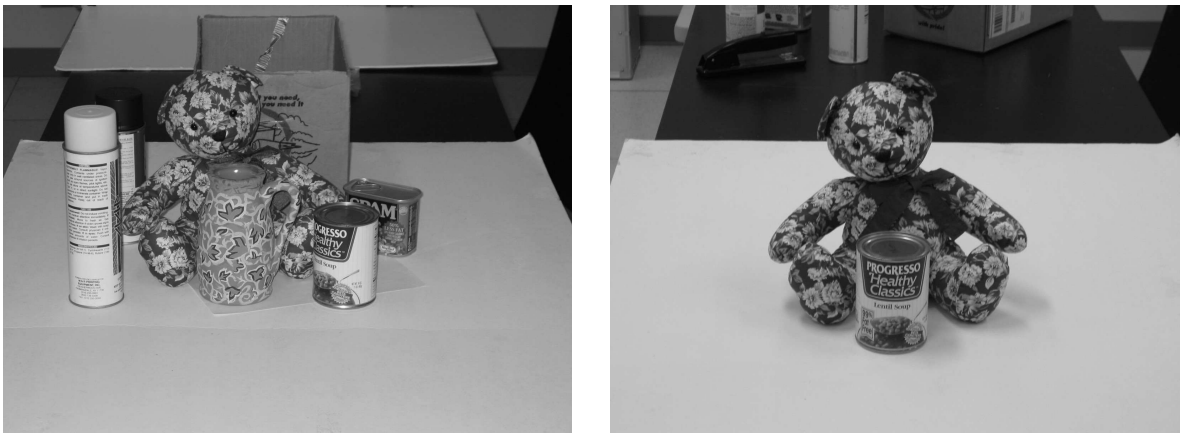
(a) Paire d'images utilisée

(b) Groupement sous contrainte épipolaire

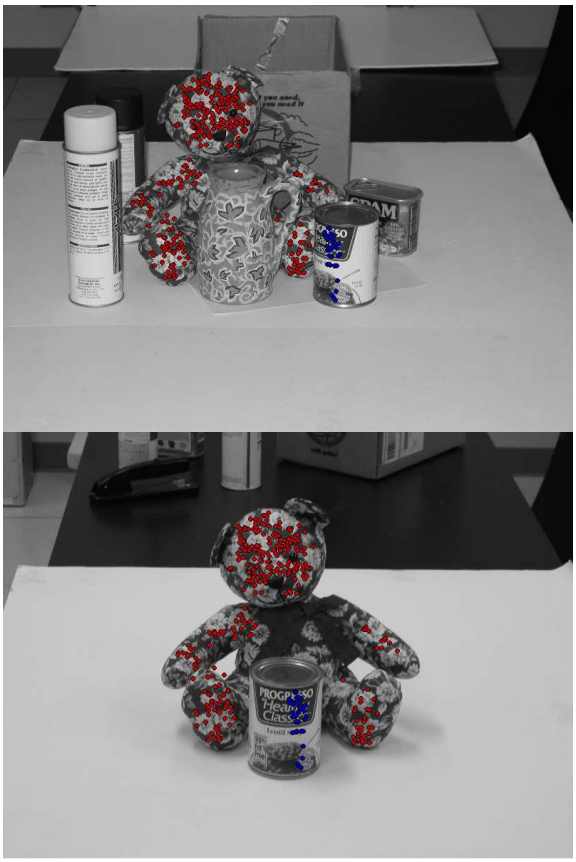


(c) Groupement sous contrainte projective

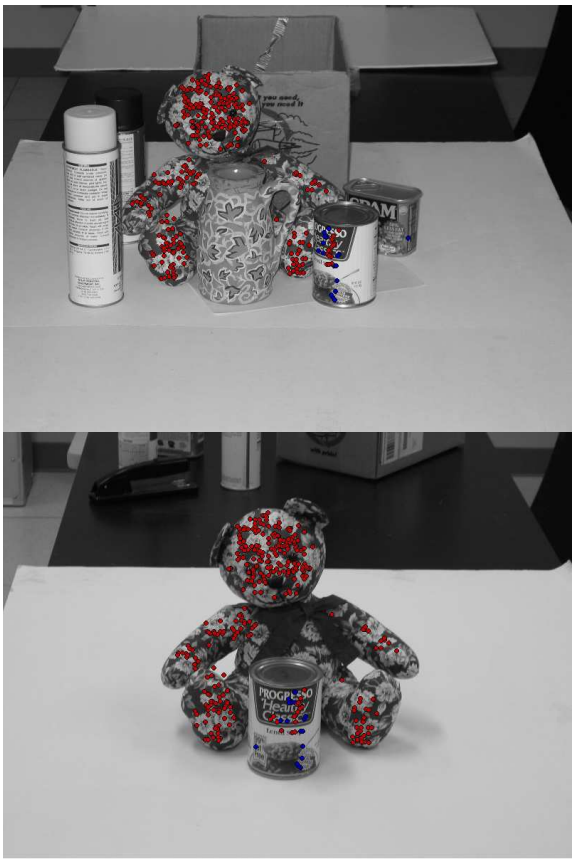
FIG. 4.20 – *La figure 4.20(a) montre deux photographies d'une scène où un objet a été déplacé entre les deux prises de vue. Deux groupes sont alors identifiés avec la géométrie épipolaire (figure 4.20(b)). La scène est découpée en différents plans lorsque l'on recherche une transformation projective (figure 4.20(c)).*



(a) Paire d'images utilisée



(b) Groupement obtenu avec l'homographie.



(c) Groupement obtenu avec la géométrie épipolaire.

FIG. 4.21 – *La figure 4.21(a) montre deux images de la base [PLRS04]. Il existe deux objets commun à ces deux images, en présence de fouillis. Le groupement avec l'homographie identifie correctement les 2 objets (4.21(b)). Le groupement avec la transformation épipolaire donne quasiment le même résultat (4.21(c)), mais il existe une ambiguïté sur la position tri-dimensionnelle de la boîte cylindrique par rapport à la peluche.*

4.5.2.3 Détection multiple du même objet

Nous avons défini l'algorithme MAC-RANSAC de manière à prendre en compte les correspondances liées à des occurrences multiples d'un objet dans chacune des images I et I'. Nous allons montrer, avec quelques exemples, que MAC-RANSAC permet de détecter si un objet apparaît plusieurs fois dans une paire d'images.

La figure 4.22(a) montre une paire de photographies, dont la première représente une canette de soda. La seconde image représente une scène avec 28 canettes ayant le même logo, dans des positions différentes. Précisons que la première image est obtenue à partir d'un point de vue différent de la seconde. Le résultat du groupement de correspondances est donné en figure 4.22(b). On vérifie que l'algorithme permet de détecter les 28 occurrences de la canettes. Nous aimerions en particulier insister sur le fait que chacun des groupes détectés correspond à une faible proportion de l'ensemble des correspondances. En particulier, les deux groupes correspondant à deux canettes fortement occultées (en bleu foncé) ne représentent chacun que 1% du total des correspondances. Ceci montre l'intérêt de rechercher itérativement chaque objet, de manière à augmenter progressivement la proportion relative des groupes correspondant à des objets de petites tailles.

Dans le second exemple sont utilisées deux photographies contenant certains objets en commun dans des poses différentes (figure 4.23(a)) : une boîte de céréales, ainsi que trois canettes identiques. Rien ne distingue les trois canettes, si bien que 9 transformations sont théoriquement possibles entre les deux vues pour ces 3 objets. En utilisant MAC-RANSAC, un unique groupe correspondant à la boîte de céréales est correctement détecté, ainsi que 9 groupes de correspondances entre les canettes (figure 4.23(b)). Il est intéressant de voir que si le critère de découpage récursif en sous-groupe n'est pas utilisé (figure 4.23(c)), l'interprétation de cette paire d'image est incomplète. En effet, deux des trois canettes sont considérées comme étant un unique objet rigide, selon une transformation très imprécise. Une seule transformation pour ce groupe est détectée, ce qui donne un total de 5 groupes au lieu de 10.

(a) Paire d'images avec l'objet recherché à gauche, et une scène contenant de nombreuses occurrences de l'objet à droite.

(b) Résultat du groupement avec MAC-RANSAC : les 28 occurences de l'objet recherché sont reconnues.

FIG. 4.22 – *Détection d'un objet apparaissant plusieurs fois dans une paire d'images (4.22(a)). Les différents filtres utilisés (principe de maximalité, élimination des correspondances redondantes et auto-similaires) préservent les correspondances liées à la détection multiples d'un objet. L'algorithme MAC-RANSAC permet dans le cas présent de détecter les 28 canettes de soda présentes dans la scène (4.22(b)).*

(a) Deux scènes différentes partageant des objets en commun.

(b) L'algorithme MAC-RANSAC détecte les 10 groupes recherchés.

(c) Sans la procédure de découpage récursif en sous-groupes, seulement 5 groupes sont détectés.

FIG. 4.23 – *Dans cet exemple, deux photographies contiennent plusieurs objets en commun : une boîte de céréales, et trois canettes identiques (4.23(a)). L'algorithme MAC-RANSAC détecte 10 groupes (figure 4.23(b)) : un groupe correspond à la boîte de céréales, et les* 9 *restants sont les détections des* 3×3 *transformations des objets identiques. La figure 4.23(c) montre ce qu'il se produit sans la procédure de détection de fusion. Des paires de canettes ayant des transformations grossièrement similaires sont regroupées et sont considérées comme un même objet rigide. Il n'y a alors plus que* 2×2 *groupes correspondant aux canettes.*

4.5.3 Évaluation expérimentale de la sélection de modèles

Dans cette section dédiée à la sélection de modèles, le critère présenté en section 4.3 est utilisé pour sélectionner le meilleur modèles parmi les transformations planes (similitude, transformation affine ou projective) et la géométrie épipolaire. Nous analysons dans un premier temps ses performances sur des données synthétiques (§ 4.5.3.1). Ensuite, nous illustrons son intérêt pour la reconnaissance d'objets dans des images réelles (§ 4.5.3.2).

4.5.3.1 Expérimentations sur des données synthétiques

Nous utilisons dans ce paragraphe des correspondances synthétiques dont on connaît le vrai modèle. Pour cela, un objet est défini à partir d'un nuage de points 3D, qui sont projetés sur le plan focal de chacune des caméras (une illustration est donnée en figure 4.24). Les deux caméras sont considérées comme parfaites, modélisées par un sténopé dont le principe est rappelé en annexe (voir la section C.1).

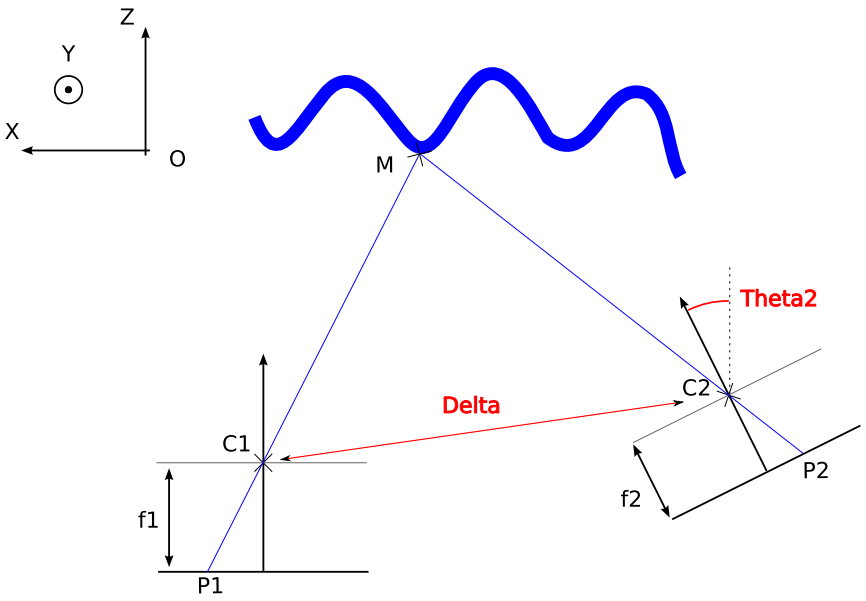


FIG. 4.24 – *Illustration de l'obtention des paires d'images synthétiques à partir de deux sténopés.*

Nous obtenons par ce procédé des correspondances de points parfaites entre les deux images synthétiques. À titre d'exemple, la figure 4.26 montre les deux vues obtenues lorsque la caméra fait un mouvement 3D autour d'un objet plan (carré). Pour simuler la quantification liée au capteur, les coordonnées des points d'intérêt synthétiques sont définies sur une grille de 1000×1000 pixels. Un bruit blanc de moyenne nulle est ensuite ajouté sur les coordonnées 2D des points pour simuler l'erreur en position des points d'intérêt appariés. Sauf indications contraires, l'écart-type σ a été fixé à 0.1.

Afin de simuler différents modèles géométriques, nous utilisons deux types d'objets : un objet plan pour commencer, puis une paraboloïde. Rappelons que le modèle géométrique de la relation entre les points vus par chaque caméra dépend à la fois de la nature de l'objet (plan ou non) et des caractéristiques de la caméra (distance focale, orientation de l'axe optique et position du centre de la caméra). Le modèle géométrique correspondant aux différents cas de figure rencontrés est rappelé en annexe C.

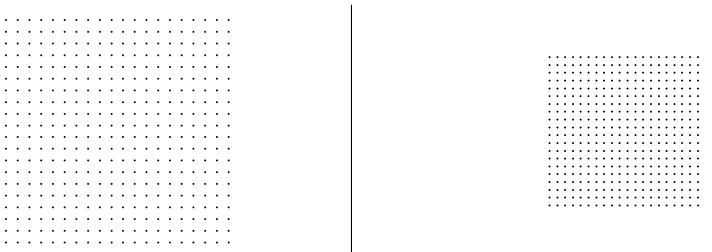
À partir de ces ensembles de correspondances synthétiques, nous utilisons l'algorithme MAC - RANSAC pour estimer la transformation optimale de chacun des modèles considérés. Par analogie avec la log-vraisemblance, et pour simplifier la lecture des différentes valeurs, les scores pour chaque modèle sont exprimés par la suite comme la quantité $-log\text{NFA}$ (logarithme en base 10). Le meilleur modèle est

celui maximisant cette quantité. Dans le tableau 4.5 sont répertoriés les scores de chacun des modèles, ainsi que les figures correspondant à chacune des expériences réalisées que nous allons maintenant présenter. Les cases à fond bleu indiquent le vrai modèle, et le score en gras montre quel est le modèle choisi par notre critère de sélection).

TAB. 4.5 – *Comparaison de la mesure de qualité du modèle ($-\log NFA$), avec entre parenthèses la mesure de précision (rigidité). Les modèles comparés sont : Similitude (S), Affine (A), Homographie (H), matrice Fondamentale (F). Les mouvements de la caméra utilisés sont : Translation (T), Zoom (Z) et Rotation (R) (voir illustration 4.24).*

Expériences		Modèles testés			
Objet	Mouvement (et Figure)	S	A	H	F
Plan	T (4.25)	**2433** (0.46)	2418 (0.49)	2172 (0.9)	905 (1.6)
	T_∞, Z_∞, R (4.27)	542 (116)	**2345** (0.6)	1957 (1.5)	944 (1.4)
	T+R (4.26)	570 (110)	504 (100)	**2190** (0.9)	947 (1.5)
Paraboloïde	T+R (4.28)	525 (112)	506 (104)	531 (110)	**1088** (0.57)
	R (4.30)	1088 (23)	1047 (25)	**2277** (0.6)	932 (1.25)
	Z (4.29)	**2312** (0.7)	2267 (0.72)	2246 (0.75)	916 (1.0)
	$T \to 0$ (4.31)	**1200** (13.7)	1190 (15)	1191 (14.8)	900 (1.6)
	T, Z_∞, R (4.32)	-	-	**923** (17)	835 (1.1)

Objet Plan Lorsque l'objet est plan, on vérifie qu'une transformation plane est sélectionnée quel que soit le mouvement de la caméra (voir les trois premières lignes du tableau 4.5). Lorsque la caméra se translate de manière à ce que son axe optique soit toujours perpendiculaire au plan de l'objet, quel que soit le changement de focale ou de la rotation autour de son axe optique, il s'agit d'une similitude.

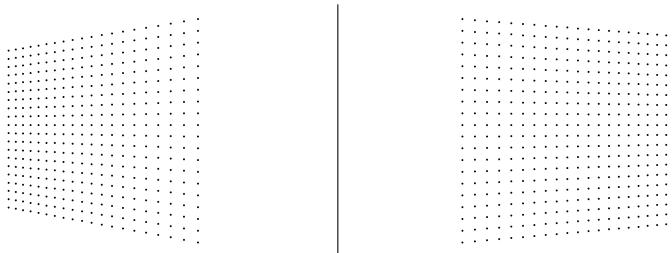


(a) Caméra 1 (b) Caméra 2

FIG. 4.25 – *Objet plan (carré), avec une translation de la caméra entre les deux vues de telle sorte que son axe optique reste perpendiculaire au plan de l'objet. La transformation entre ces deux ensembles de points est une similitude.*

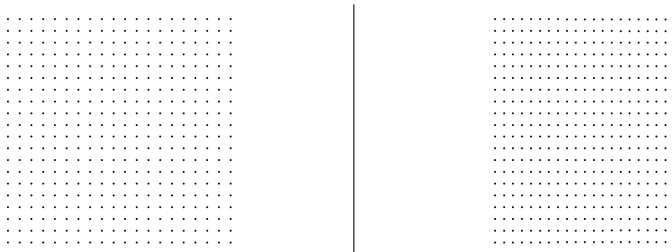
Dans le cas de la figure 4.25, nous avons simplement translaté la caméra. Si par contre la caméra effectue une rotation de telle sorte que son axe optique n'est plus perpendiculaire au plan de l'objet, alors il s'agit d'une homographie (figure 4.26). Nous vérifions que ce sont les modèles effectivement sélectionnés pour ces deux expériences.

Le cas de la transformation affine est un peu plus particulier : il s'agit d'un cas limite où la perspective préserve le parralélisme. Cela correspond au cas où l'objet est vu avec un recul infini de la caméra et un zoom infini, ce que nous simulons en figure 4.27. Il est intéressant de voir que, même dans ce cas limite, c'est une fois encore le modèle correct qui est sélectionné par notre approche.



(a) Caméra 1 (b) Caméra 2

FIG. 4.26 – *Objet plan (carré), avec un mouvement 3D de la caméra entre les deux vues. Dès que l'axe de la caméra est non orthogonal au plan de l'objet, les effets de la perspectives sont observables, d'autant plus que la caméra est proche de l'objet. La transformation entre ces deux ensembles de points est une homographie.*



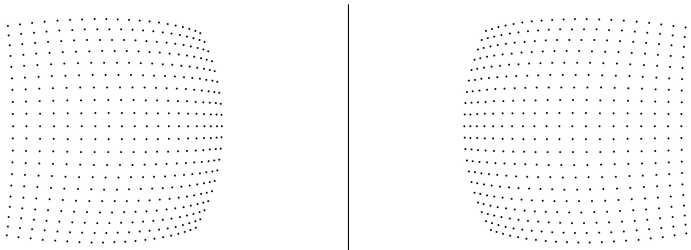
(a) Caméra 1 (b) Caméra 2

FIG. 4.27 – *Figure 4.27(a) : objet plan (carré) orienté face à la caméra. Figure 4.27(b) : autre vue de l'objet avec une rotation de la caméra de $\pi/6$, combinée avec un recul à l'infini (ici obtenu par un éloignement d'une distance de* 1000 *la taille de l'objet) et zoom infini (ici par un changement de focale d'un facteur* 100*). Dans ce cas limite, la transformation est affine : le carré devient un rectangle quasiment parfait.*

> **Remarque 2 :**
> Nous indiquons également dans le tableau de résultat 4.5, la mesure de rigidité associée à chacune des transformations estimées. Ceci permet d'avoir une idée de la précision obtenue. Rappelons néanmoins qu'il s'agit de l'erreur de transfert résiduelle *maximum* du groupe sélectionné. On pourrait s'étonner que la rigidité est plus élevée avec des modèles ayant un plus grand degré de liberté. Ceci est le résultat du procédé d'optimisation par échantillonnage aléatoire. D'une part, les transformations (et donc la rigidité) sont estimées à partir de n-uplets de correspondances « bruitées ». Lorsque l'on utilise un modèle avec un degré de liberté plus que nécessaire, la transformation estimée à partir du n-uplet va excessivement tenir compte du bruit sur la position des points d'intérêt (un sur-apprentissage en quelque sorte). Autrement dit, l'estimateur d'un modèle trop complexe est *moins robuste au bruit* que celui d'un modèle plus simple. D'autre part, l'optimisation (ORSA) est réalisée par échantillonnage aléatoire de 1000 n-uplets. Cela signifie que l'on ne teste pas toutes les configurations possibles. Néanmoins, afin de s'assurer de la robustesse de la solution obtenue, nous avons pour chaque expérience utilisé 10 fois l'algorithme MAC-RANSAC, et choisi la solution donnant le meilleur score.

Objet 3D Lorsque l'objet n'est pas plan (ici, une paraboloïde de révolution), la transformation observée est *a priori* 3D et donc décrite par la géométrie épipolaire. C'est ce que nous vérifions avec la figure 4.28 où la caméra effectue un mouvement 3D autour de l'objet : le modèle épipolaire est alors le seul pouvant expliquer la scène avec une excellente précision (voir le tableau 4.5).

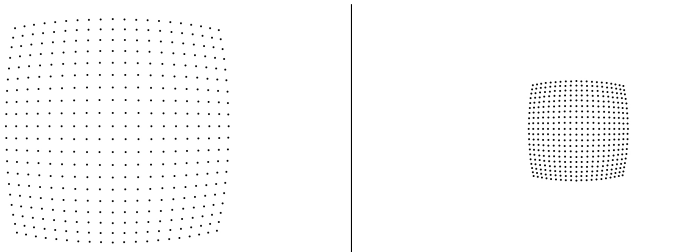
Toutefois, il existe des configurations particulières où le modèle peut être réduit à une simple transformation plane. Lorsque la caméra est fixe et que l'on effectue un changement de focale (figure 4.29), c'est effectivement la similitude qui est sélectionnée. Lorsque le centre de la caméra est fixé mais que son axe optique change d'orientation, la transformation est une homographie (figure 4.30), ce que le critère



(a) Caméra 1 (b) Caméra 2

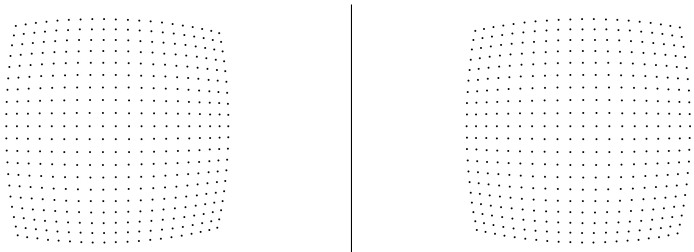
FIG. 4.28 – *Objet 3D (paraboloïde) avec un mouvement 3D entre les deux vues.*

de sélection permet une fois encore de détecter.



(a) Caméra 1 (b) Caméra 2

FIG. 4.29 – *Objet 3D (paraboloïde) avec un zoom (changement de focale) entre les deux vues. Dans ce cas particulier, la transformation entre les deux images est une similitude.*



(a) Caméra 1 (b) Caméra 2

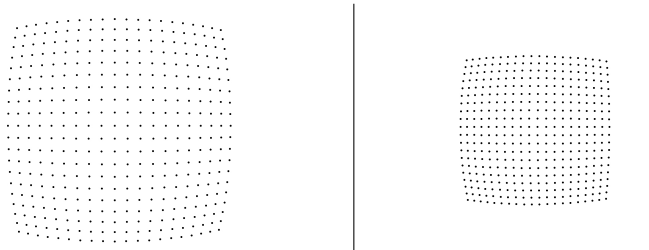
FIG. 4.30 – *Objet 3D (paraboloïde) avec une rotation de la caméra autour de son centre optique (fixe), entre les deux vues. Dans ce cas particulier, la transformation entre les deux images est une homographie.*

Détection de faibles mouvements 3D Dans le domaine de la reconstruction 3D, il est nécessaire d'avoir une grande base pour que les effets du mouvement 3D soient suffisamment importants. Ainsi dans [RP05], les auteurs sélectionnent les images d'une séquence vidéo pour lesquelles la ligne de base est suffisamment grande. Pour cela ils utilisent le critère GRIC [Tor98] afin de choisir les images pour lesquelles le modèle de la géométrie épipolaire l'emporte sur l'homographie.

Nous rencontrons exactement le même phénomène avec notre critère : dans certaines situations, le critère de sélection de modèles *tend à privilégier les transformations planes* au détriment du modèle épipolaire. En fait, de manière analogue au cas limite de la transformation affine qui est sélectionnée à la place de l'homographie, on observe que le critère de sélection de modèles tend à choisir un modèle plus simple si celui-ci offre une approximation suffisante. Nous allons voir avec trois exemples différents dans quels cas ceci se produit.

En premier lieu, lorsque la caméra se translate, on observe les effets combinés de la perspective et de la profondeur de l'objet et c'est le modèle épipolaire qui l'emporte. Cependant, lorsque ces effets de

perspectives et de profondeur sont trop faibles, le critère de sélection alors privilégie une transformation plane. C'est le cas de la paire d'image synthétique de la figure 4.31 où le mouvement de la caméra est tel que les effets de perspectives sont très limités. C'est, dans ce cas, la similitude qui obtient le meilleur score, au lieu de la matrice fondamentale.



(a) Caméra 1 (b) Caméra 2

FIG. 4.31 – *Objet 3D (paraboloïde), avec une translation entre les deux vues. Il s'agit d'une transformation 3D, mais la caméra étant éloignée de l'objet et son déplacement étant faible, le modèle sélectionné est une simple similitude.*

Lorsque le déplacement de la caméra est important entre les deux vues (grande base, ou *wide baseline* en anglais), le changement de perspective est considérable. Si la caméra s'éloigne fortement de l'objet, la perception de la profondeur diminue. Les objets éloignés nous paraissent plans : on parle d'« écrasement de perspective ». C'est par exemple le cas dans le domaine des images aériennes (cartographie, imagerie satellitaire). Pour de telles images où seul l'effet de perspective apparaît, le recalage des images est généralement réalisé avec une homographie. Nous illustrons ce phénomène par la figure 4.32, où le déplacement de la caméra est très important, avec un éloignement suffisant pour diminuer fortement la notion de profondeur sur l'objet. Dans cet exemple, c'est l'homographie qui est sélectionnée au lieu de la géométrie épipolaire.

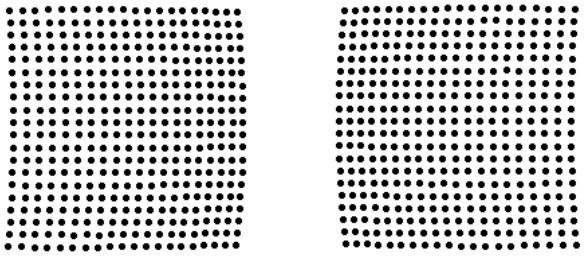


FIG. 4.32 – *Objet 3D (paraboloïde), en vue lointaine (Zoom et Recul) avec mouvement 3D de la caméra. Du fait de ce recul important, les perspectives sont "écrasées". Une fois encore, le modèle homographique est sélectionnée ($-\log NFA = 923$) au lieu de l'épipolaire ($-\log NFA = 835$) car, du point de vue du modèle de fond, c'est une approximation convenable.*

De manière plus générale, il est couramment admis que lorsqu'un objet possède un plan dominant, il est difficile de distinguer la géométrie de l'homographie (voir par exemple [Chu05]). En réalisant des expériences (non présentées ici) avec d'autres objets 3D synthétiques, nous avons également observé que plus l'objet considéré possède un plan dominant, plus le modèle de la géométrie projective est privilégiée. Nous reviendrons sur ce phénomène dans la section suivante, sur des paires d'images réelles.

Un dernier phénomène affecte la sélection du modèle de la géométrie épipolaire : le bruit sur la position des points d'intérêt. Dans le tableau 4.6, les scores de chaque modèle sont indiqués pour une même transformation synthétique 3D, illustrée en figure 4.33, en fonction de l'écart-type σ de l'erreur ajoutée à la position des points d'intérêt. En augmentant σ, la mesure de qualité de la transformation épipolaire diminue, alors qu'elle reste inchangée pour l'homographie (dont l'erreur résiduelle est déjà

très grande). Lorsque l'on atteint un certain niveau de bruit ($\sigma = 2.5$), l'homographie est sélectionnée à la place du modèle 3D. La figure 4.33 illustre la paire d'images synthétiques obtenue avec un tel bruit.

> **Remarque 3 :**
> Comme pour les expériences précédentes, l'algorithme MAC-RANSAC est utilisé pour estimer une transformation optimale en $i_{max} = 10000$ itérations. En raison du bruit blanc gaussien sur la position des points appariés, le groupe optimal ne contient généralement qu'une partie de l'ensemble des correspondances. C'est la raison pour laquelle à la fois la rigidité α et la mesure de qualité $-\log \text{NFA}$ peuvent grandement varier en fonction de l'écart-type σ du bruit.

TAB. 4.6 – *Comparaison de la mesure de qualité du modèle ($-\log NFA$), avec entre parenthèse la rigidité, en fonction de l'écart-type σ du bruit blanc gaussien ajouté à la position des points d'intérêt. Les modèles comparés sont : Homographie (H) et matrice Fondamentale (F). Les deux images synthétiques sont obtenues pour un mouvement 3D de caméra autour de la paraboloïde (figure 4.33).*

Ecart-type σ	.1	.2	.5	1.	1.5	2.	2.5
H	602 (99)	595 (87)	588 (110)	593 (99)	605 (86)	590 (109)	**601** (105)
F	**910** (1.6)	**919** (1.7)	**831** (2.9)	**745** (4.9)	**737** (4.4)	**622** (9.3)	588 (9.9)

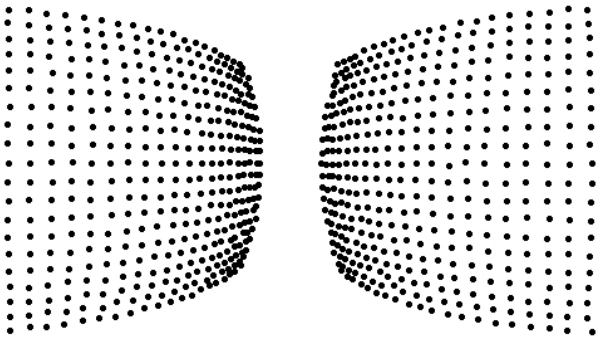
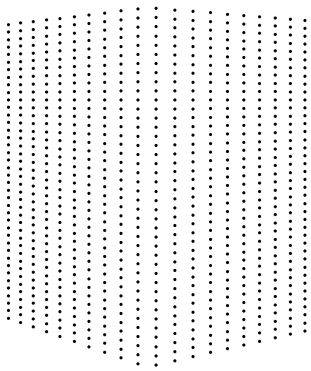
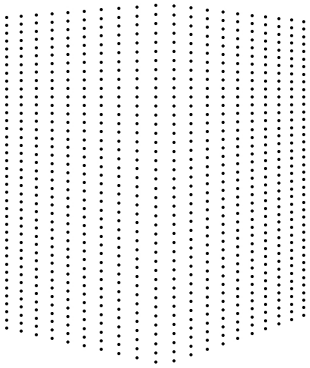


FIG. 4.33 – *Objet 3D (paraboloïde), avec un mouvement 3D de la caméra entre les deux vues et ajout d'un bruit blanc gaussien sur la position des points, d'écart-type $\sigma = 2.5$. En raison du manque de précision de la transformation épipolaire, le modèle homographique est sélectionné (voir tableau 4.6).*

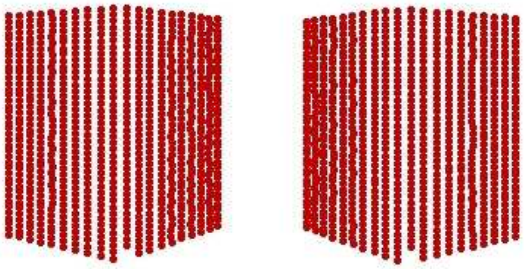
Illustration de l'intérêt du critère de découpage Avant d'illustrer le comportement du critère de sélection de modèles sur des images réelles, nous aimerions brièvement présenter l'intérêt du critère de découpage en sous-groupe sur des données synthétiques. En figures 4.34(a) et 4.34(b), sont montrées deux vues d'un objet composé de deux plans formant un angle droit. En utilisant le groupement avec le modèle projectif, pourtant correct, un unique groupe est obtenu (figure 4.34(c)) au lieu de deux. Ce groupe est en effet plus significatif ($-\log \text{NFA} = 1180$, $\alpha = 7$) que le plus significatif des deux plans ($-\log \text{NFA}_1 = 911$, $\alpha_1 = .49$ et $-\log \text{NFA}_2 = 880$, $\alpha_2 = .46$). En utilisant le critère de découpage (figure 4.34(d)), on obtient deux groupes correspondant à chaque plan, avec une précision nettement supérieure.



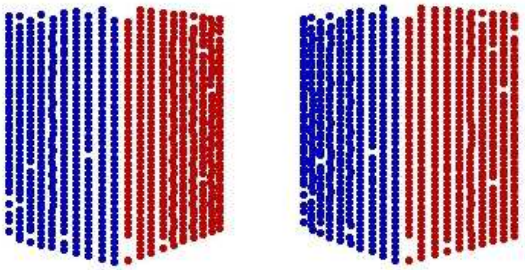
(a) Caméra 1



(b) Caméra 2



(c) Groupement sans découpage : un seul groupe imprécis est détecté ($-\log \text{NFA} = 1180$, $\alpha = 7$)



(d) Groupement avec découpage : les deux groupes obtenus sont corrects ($-\log \text{NFA}_1 = 911$, $\alpha_1 = .49$ et $-\log \text{NFA}_2 = 880$, $\alpha_2 = .46$)

FIG. 4.34 – *Figures 4.34(a) et 4.34(b) : Expérience synthétique avec un objet possédant 2 plans, et un mouvement 3D entre les deux vues. Sans la procédure de découpage récursif, le groupement avec l'homographie donne un seul groupe (figure 4.34(c)). Avec cette procédure, on obtient deux groupes correspondant à chacun des plans (figure 4.34(d)).*

4.5.3.2 Expérimentations sur des images réelles

Nous présentons dans les deux paragraphes suivants le résultat de notre critère de sélection de modèles sur des paires de photographies d'objets plans ou 3D. L'ensemble des scores obtenus pour chaque modèle et pour chaque expérience est donné en table 4.7.

TAB. 4.7 – *Comparaison de la mesure de qualité du modèle ($-\log NFA$). Les modèles comparés sont : Similitude (S), Affine (A), Homographie (H), matrice Fondamentale (F). Le vrai modèle est mis en valeur sur fond bleu, et le modèle sélectionné est indiqué en fonte grasse.*

Expériences		Modèles testés			
Objet	Figure	S	A	H	F
Plan	BD (4.35)	**726**	718	704	284
	Dali (4.36)	64	**178**	172	83
	Portrait (4.37)	260	273	**340**	151
3D	Jouet (4.38(a) & 4.38(b))	525	535	545	**996**
	Teddy (4.38(c) & 4.38(d))	334	441	432	**510**
3D à plan dominant	Chapelle (4.39)	350	400	**430**	205
	Étagère (4.40)	215	310	**370**	242

Transformations planes Nous considérons dans un premier temps le cas des transformations planes. Dans les trois exemples suivants, le modèle épipolaire est systématiquement rejeté au profit d'un modèle plan.

Dans le premier exemple de la figure 4.35, deux photographies d'un livre sont prises avec une translation et un changement de la distance focale de la caméra, l'axe optique restant perpendiculaire au plan de l'objet. C'est la similitude qui est correctement sélectionnée pour cet exemple, dont on donne le recalage en figure 4.35(b).

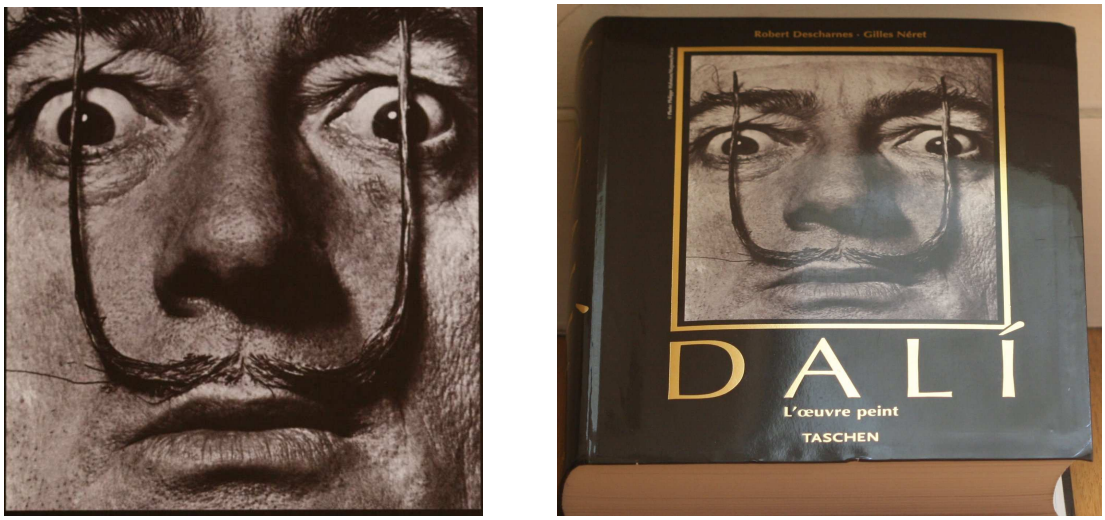
Dans le second exemple (figure 4.36(a)), nous nous sommes placés dans des conditions de prise de vue avec un fort recul avec un zoom, de telle sorte que la transformation obtenue soit une transformation affine. C'est effectivement le modèle choisi par notre méthode (voir 4.7), et le recalage obtenu confirme que le modèle affine suffit pour décrire le changement de point de vue.

Nous considérons dans le dernier exemple la transformation projective pour un objet plan. En figure 4.37(a) sont mises en correspondance deux photographies d'un tableau, selon deux points de vues différents. Du fait de l'effet de perspective prépondérant, c'est l'homographie qui est sélectionnée par notre modèle. En effet, seul le recalage avec l'homographie donne un résultat visuellement satisfaisant (figure 4.37(c)).

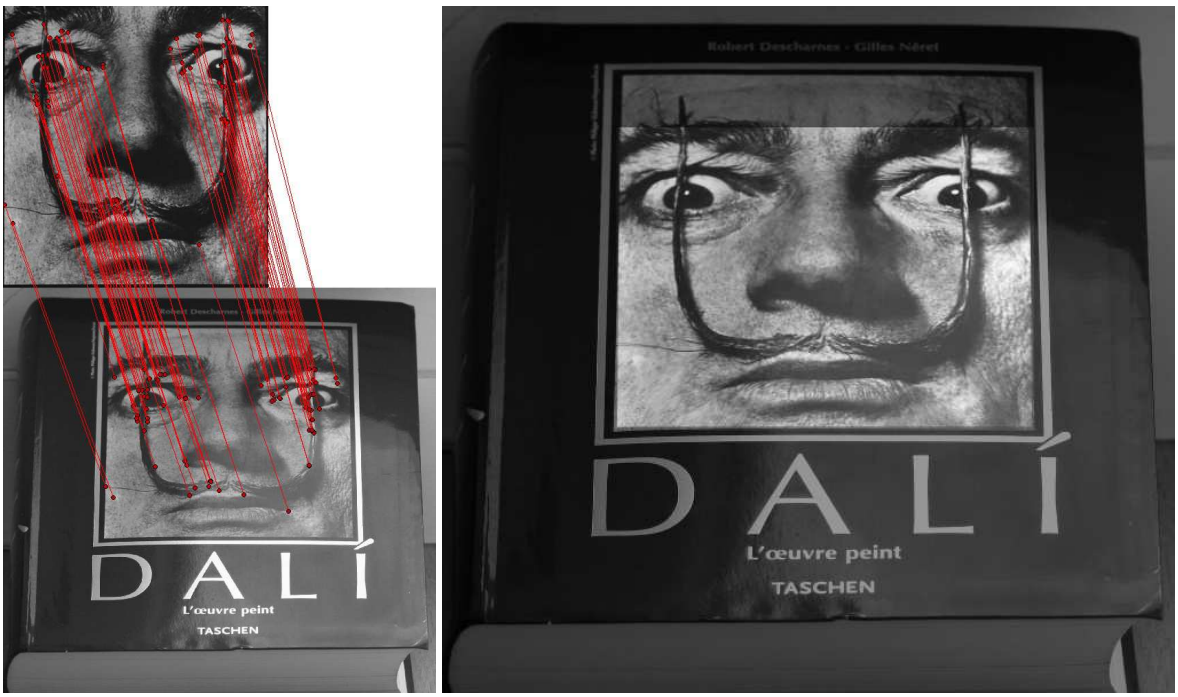
(a) À gauche : Zoom sur un objet plan. À droite : groupe de correspondances de points d'intérêt sélectionnées (la similitude obtient le score le plus élevé, voir le tableau 4.7).

(b) Superposition des deux images selon les paramètres de la similitude sélectionnée.

FIG. 4.35 – *Figure 4.35(a) : exemple de sélection de la similitude pour un objet plan, avec un mouvement de la caméra tel que son axe optique reste perpendiculaire au plan de l'objet. Le recalage montre que le modèle choisi est satisfaisant (figure 4.35(b)).*



(a) Photo en vue frontale et en vue oblique d'un objet plan.



(b) Groupement de correspondances et superposition des deux images selon la transformation affine estimée.

FIG. 4.36 – *Figure 4.36(a) : 2 vues éloignées avec zoom d'un objet plan. Le groupe de correspondances sélectionné est le mieux expliqué par une transformation affine (voir le tableau 4.7). Figure 4.36(b) : superposition des deux photographies selon la transformation estimée. On constate que le modèle affine permet de décrire la transformation entre les deux images.*

(a) À gauche : paire d'images d'un objet plan sous deux angles de vues différents. À droite : le groupe de correspondances sélectionné obtient le meilleur score avec la transformation projective (voir le tableau 4.7).

(b) Superposition des deux images selon la transformation affine estimée.

(c) Superposition des deux images selon l'homographie, qui est identifié comme le meilleur modèle.

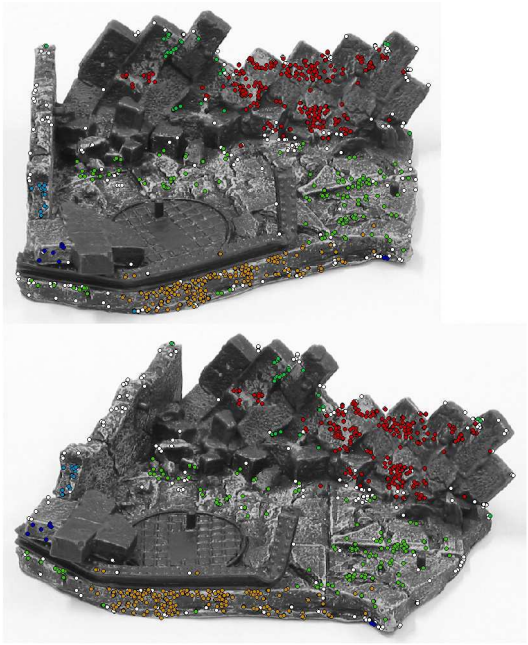
FIG. 4.37 – *Figure 4.37(a) : deux photographies d'un tableau, selon deux points de vues différents, sont mises en correspondance. L'homographie, qui est sélectionnée par notre critère, donne le meilleur recalage (figure 4.37(c)).*

Sélection du modèle de la géométrie épipolaire pour des objets 3D En expérimentant sur des données synthétiques, nous avons observé que le modèle épipolaire était sélectionné pour des objets 3D si le changement de point de vue était suffisamment important. Par contre, lorsque les objets considérés possèdent un plan dominant, une transformation plane est généralement privilégiée. Nous allons montrer avec les expériences suivantes que l'on observe le même phénomène sur des paires d'images.

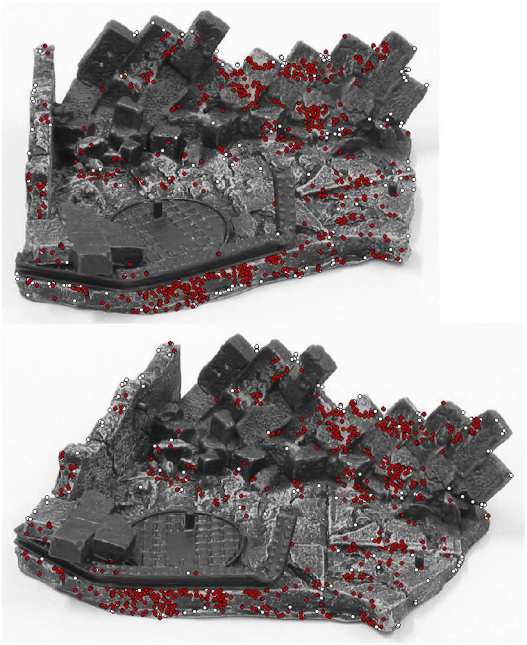
Les deux exemples suivants sont tirés de la base de données [PLRS04], où différentes vues d'objets 3D sont photographiées. Pour les paires d'images utilisées en figure 4.38 le changement de point de vue est tel que la géométrie épipolaire obtient le meilleur score pour chacun de ces exemples (voir le tableau 4.7).

Nous présentons ensuite deux exemples où un objet possède un plan dominant en figures 4.39(a) et 4.40(a). Dans ces deux cas, le modèle épipolaire est rejeté au profit d'une transformation plane (ici l'homographie). Ceci se produit lorsqu'une des dimensions de l'objet considéré est négligeable devant les deux autres (par exemple, les tranches de livres grossièrement alignées sur une étagère). Nous avons vu que ce phénomène était renforcé par l'écrasement de perspective, c'est-à-dire lorsque le mouvement de la caméra ne suffit pas à mettre en évidence la profondeur de l'objet.

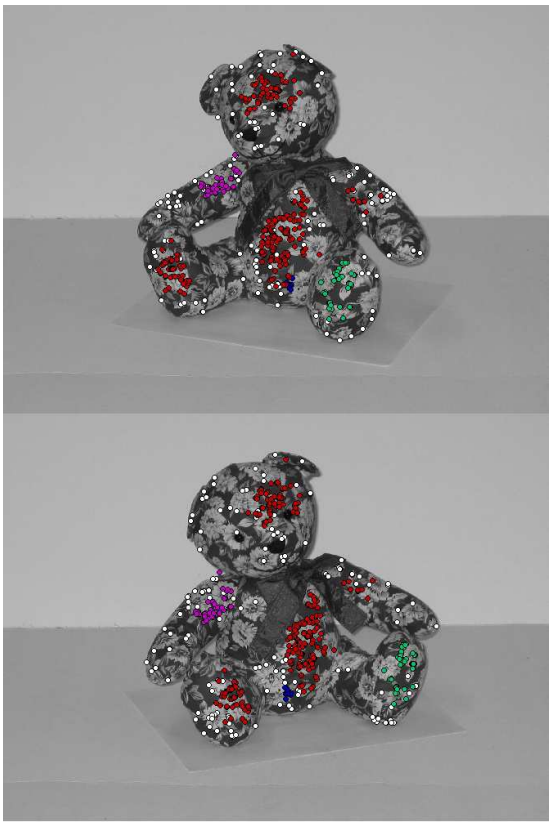
Dans ces deux exemples, seule la géométrie épipolaire permet d'expliquer l'ensemble des correspondances entre les deux vues, et avec précision. Pourtant, l'homographie est sélectionnée pour ces deux exemples où une large majorité de points reposent approximativement sur un même plan. Le recalage des deux paires d'images permet de vérifier ce constat (figures 4.39(d) et 4.40(b)).



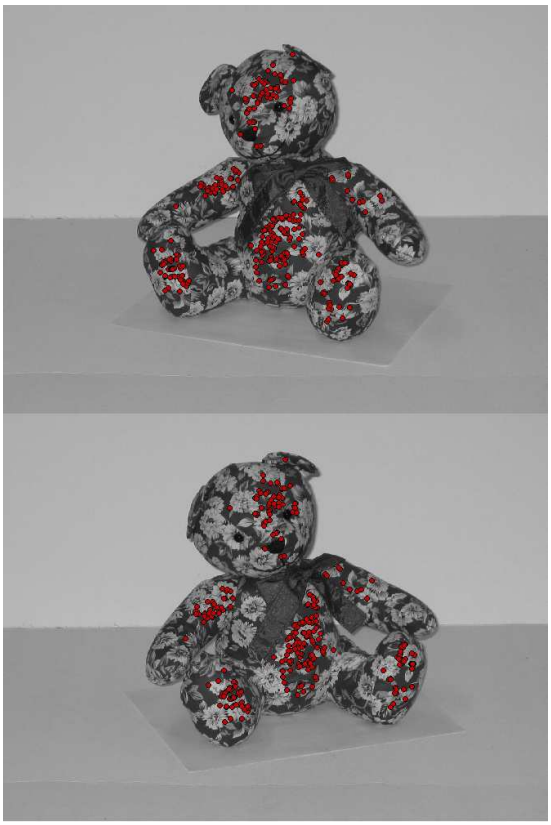
(a) Avec une transformation plane (ici l'homographie), l'algorithme MAC-RANSAC détecte plusieurs groupes correspondant aux plans principaux de l'objets.



(b) Un seul groupe est détecté avec la géométrie épipolaire, qui est le modèle sélectionné par notre critère de sélection.



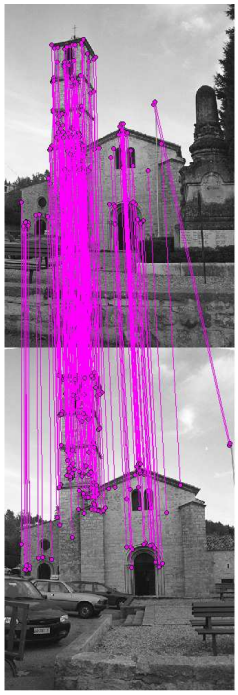
(c) Avec une transformation plane (ici l'homographie), l'algorithme MAC-RANSAC détecte plusieurs groupes correspondant aux plans principaux de l'objets.



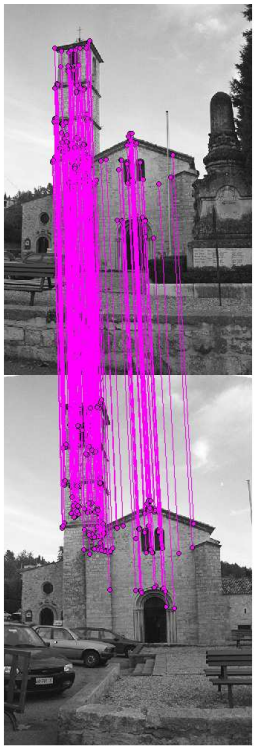
(d) Un seul groupe est détecté avec la géométrie épipolaire, qui est le modèle sélectionné par notre critère de sélection.

FIG. 4.38 – *Sélection de modèles sur un objet 3D. Les transformations planes segmentent les correspondances en plusieurs groupes, selon les plans principaux qui composent l'objet (figures 4.38(a) et 4.38(c) avec la géométrie projective). Un unique groupe est détecté avec la géométrie épipolaire (figure 4.38(b) et 4.38(d)), ce qui lui permet d'obtenir le meilleur score (tableau 4.7).*

(a) Deux vues d'une scène avec un plan dominant



(b) Groupement selon la géométrie épipolaire



(c) Groupement selon l'homographie

(d) Superposition des deux vues selon l'homographie estimée.

FIG. 4.39 – *Église de Valbonne* *Figure 4.39(b) : seule la géométrie épipolaire permet d'expliquer l'ensemble des correspondances entre les deux vues de la figure 4.39(a). Pourtant, les transformations planes obtiennent toutes un meilleur score (tableau 4.7) car une majorité de points reposent approximativement sur un même plan (le frontispice de l'église). En effet, le groupement obtenu selon l'homographie ne sélectionne que les points appartenant à ce plan (figure 4.39(c)). Le recalage du modèle sélectionné (homographie) est en effet suffisamment précis pour ce plan (figure 4.39(d)).*

(a) Deux vues d'une scène avec un plan dominant

(b) Superposition des deux images avec le modèle homographique sélectionné : le résultat est peu précis mais la transformation obtenue est la plus significative au regard de l'hypothèse nulle utilisée.

FIG. 4.40 – *Seule la géométrie épipolaire permet d'expliquer l'ensemble des correspondances entre les deux vues de la figure 4.40(a) avec précision. Cependant, l'homographie obtient le meilleur score (tableau 4.7) car une majorité de points reposent approximativement sur un même plan (l'alignement des tranches de livres). Le recalage du modèle sélectionné (homographie) correspond effectivement à ce plan approximatif (figure 4.40(b)).*

4.5.4 Limitations et analyse de configurations d'échec

Dans certaines situations particulières, le résultat obtenu avec l'algorithme MAC-RANSAC n'est pas pleinement satisfaisant. Il s'agit principalement de deux cas de figure que nous allons maintenant présenter.

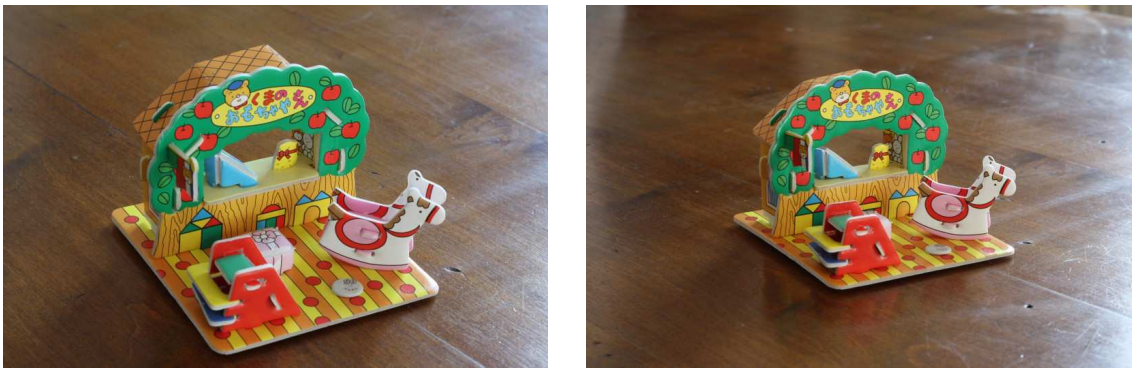
4.5.4.1 Limitation de la géométrie épipolaire pour la détection multiple

Nous avons vu avec l'expérience de la figure 4.21(c) que la détection de plusieurs objets avec la matrice fondamentale peut conduire à certaines ambiguïtés. En effet, en ajoutant une notion de profondeur, plusieurs solutions sont parfois possibles, et la géométrie épipolaire peut conduire à regrouper plusieurs objets ayant des mouvements différents entre les deux prises de vues. En pratique, cette ambiguïté se traduit au niveau de la sélection des inliers qui repose sur le calcul de l'erreur résiduelle entre des points et les lignes épipolaires. Dans l'exemple de la figure 4.41, nous avons tracé les lignes épipolaires correspondant à la matrice fondamentale du groupe sélectionné (il s'agit de la même expérience que celle réalisée en figure 4.14(a) avec une transformation plane). On peut voir que les lignes épipolaires sont parallèles et horizontales, et les trois canettes sont regroupées au lieu d'être séparées. Du point de vue géométrique, tout se passe comme si les trois canettes de la seconde image étaient alignées selon l'axe optique de la caméra dans la première image.

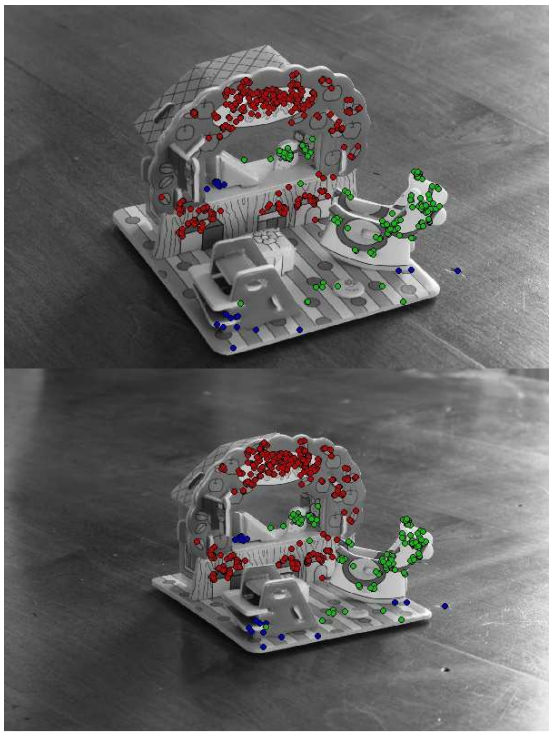
FIG. 4.41 – *Retour sur l'expérience de la figure 4.14(a) dans le cas de la géométrie épipolaire. Illustration de l'ambiguïté de la géométrie épipolaire, où les trois objets sont regroupés. Tout se passe comme si les trois objets étaient alignés dans la première image selon l'axe optique de la caméra.*

4.5.4.2 Limitation du découpage en sous-groupes

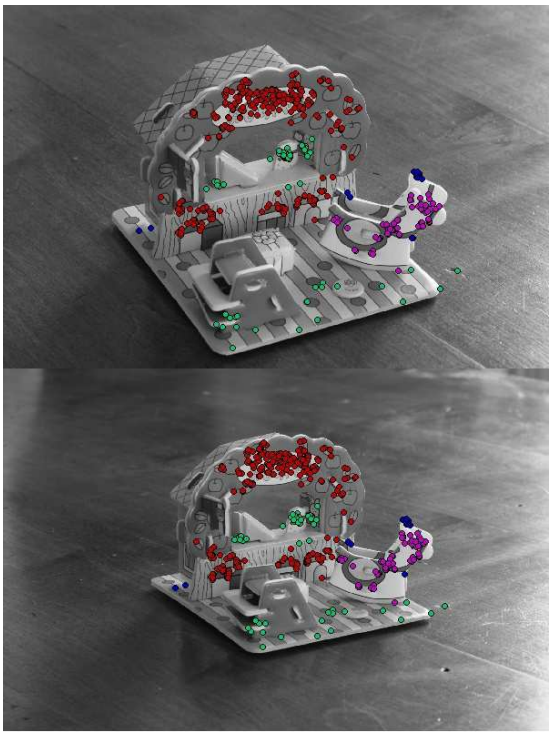
L'exemple de fusion de groupes de la figure 4.42 illustre un autre limitation de notre approche. L'utilisation itérative de MAC-RANSAC, sans utilisation du critère de découpage, nous donne pour cet exemple trois groupes. Un seul de ces groupes correspond à l'un des plans principaux de l'objet, les deux autres groupes résultant de la fusion de plusieurs plans. La procédure de découpage récursif en sous-groupes (algorithme 4.4) permet de détecter un plan supplémentaire (celui du cheval) mais le plan horizontal n'est pas identifié comme tel. Deux autres groupes incorrects sont à la place détectés.



(a) Paires d'images représentant un objet avec plusieurs plans



(b) Résultat de la recherche de plans avec MAC-RANSAC **sans** découpage en sous-groupe



(c) Résultat de la recherche de plans avec MAC-RANSAC **avec** découpage en sous-groupe

FIG. 4.42 – *Dans la paire d'images de la figure 4.42(a)), l'objet possède plusieurs plans principaux que l'on souhaite détecter. Sans l'algorithme de découpage récursif, MAC-RANSAC identifie correctement un seul plan (figure 4.42(b)). Avec notre procédure de détection de fusion, on récupère un plan supplémentaire (celui du cheval, en figure 4.42(c)) mais on ne parvient pas à segmenter le reste des points selon des plans qui ont du sens.*

Deuxième partie

Transport entre histogrammes

Chapitre 5

Problématique

Cette deuxième partie du manuscrit est consacrée à l'étude de différentes applications du transport optimal. Nous allons dans ce chapitre présenter la théorie du transport de Monge-Kantorovich, puis rappeler quelques unes de ses applications en vision par ordinateur et traitement des images.

5.1 Présentation de la théorie du transport de Monge-Kantorovich

La théorie du transport optimal est formalisée pour la première fois en 1781 par Monge dans son mémoire [Mon81] sur la « théorie des déblais et des remblais ». En 1942, Kantorovich [Kan42] propose une linéarisation de cette formulation, que nous décrivons dans ce qui suit. Soient f et g deux distributions de probabilité sur $\mathbb{R}^n$, c'est-à-dire deux mesures positives et de somme 1. La distribution f peut être vue comme un déblai que l'on va déplacer pour remplir le remblai $(-g)$. Soit $c(.,.)$ une fonction de coût sur $\mathbb{R}^n \times \mathbb{R}^n$. La quantité $c(x,y)$ représente le coût du transport d'une masse élémentaire depuis x vers y. On définit $\Pi(f,g)$ comme l'ensemble des mesures de probabilité π sur $\mathbb{R}^n \times \mathbb{R}^n$ ayant pour marginales f et g, soit

$$\Pi(f,g) := \{ \quad \pi \text{ mesure de probabilité sur } \mathbb{R}^n \times \mathbb{R}^n \,; \\ \forall A, B \in \mathbb{R}^n \;\; \pi(A \times \mathbb{R}^n) \; = f(A) \text{ et } \pi(\mathbb{R}^n \times B) \; = g(B) \;\; \} \quad .$$

Pour toute mesure π dans $\Pi(f,g)$, on peut définir le coût de transport de la mesure f vers la mesure g par le plan π comme

$$C_\pi(f,g) = \iint_{x,y} c(x,y)\, d\pi(x,y) \;\; . \tag{5.1}$$

On appelle alors transport optimal la mesure π qui minimise le coût $C_\pi(f,g)$, quand elle existe. La quantité $\inf_{\pi \in \Pi} C_\pi(f,g)$ est appelée le coût de transport optimal entre f et g.

Remarquons que le coût de transport optimal dépend fortement du choix de la fonction de coût c. Celle-ci est souvent désignée par le terme « distance au sol » (*ground distance*). Dans le cas où $c(x,y) = ||x-y||^p$, $||.||$ désignant la norme euclidienne, on peut montrer que $(\inf_{\pi \in \Pi} C_\pi(.,.))^{\frac{1}{p}}$ est une distance sur l'ensemble des mesures de probabilité sur $\mathbb{R}^n$ lorsque $p \geq 1$. Cette distance est appelée distance de Monge-Kantorovich, ou encore « p-Kantorovich norm » [ACB+03]. On note alors :

$$\mathrm{MK}_p(f,g) = \inf_{\pi \in \Pi(f,g)} \left(\iint_{x,y} ||x-y||^p \, d\pi(x,y) \right)^{\frac{1}{p}} \;\; . \tag{5.2}$$

On choisit généralement $p > 1$ pour avoir un coût strictement convexe ($p = 2$ le plus souvent). Ceci permet de s'assurer de certaines propriétés, telles que l'unicité de la solution et la préservation de l'ordre notamment[1].

[1] pour plus de détails, le lecteur est invité à consulter le livre de C. Villani [Vil03]

La distance de Monge-Kantorovich est très intéressante en pratique car elle permet de définir une mesure de dissimilarité entre deux distributions dont l'interprétation est très intuitive. Pour cette raison, la distance de Monge-Kantorovich est parfois appelée « distance L^p minimale » (*Minimal L^p metric*).

> **Remarque 1 :**
> La problématique du transport optimal a été plusieurs fois redécouverte dans la littérature, avec parfois des formulations différentes. On trouve ainsi de nombreuses autres appellations pour désigner le fait de définir une distance comme un coût optimal de transport : la distance du « cantonnier » (*Earth Mover Distance*) la distance de Kantorovich (ou *Kantorovich metric*), la distance de Wasserstein, ou encore la distance de Mallows.

Transport optimal entre des mesures discrètes Bien que ce lien ne soit pas toujours explicitement mentionné, de nombreux travaux en analyse d'images s'inscrivent dans le cadre du transport optimal. Ce sont généralement des histogrammes qui sont manipulés, et la distance de Monge-Kantorovich (5.1) est alors exprimée entre deux mesures discrètes sur une grille régulière en dimension n. Par exemple dans [SW83, WPR85], le transport est utilisé pour comparer des histogrammes de caractéristiques (texture, forme, *etc.*).

En dehors des histogrammes, d'autres types de structures de données se prêtent également au problème du transport. Par exemple dans [WPMK86], Werman *et al.* s'intéressent au problème de l'assignement entre deux ensembles finis de points (*Bipartite graph matching*). Par la suite, Rubner *et al.* [RTG00] généralisent ce principe en introduisant le concept de « signature », un ensemble d'éléments pondérés (*weighted bipartite graph*). La signature d'une image $\{(x_i, p_i)\}_{i=1,\dots,N}$ se compose d'une liste d'éléments $\{x_i\}$ caractéristiques (couleur, texture, *etc.*), et d'une liste de poids $\{p_i\}$ qui en mesurent la composition. Un histogramme $h[i]$ peut ainsi être vu comme une signature particulière où les éléments $\{x_i\}$ correspondent à une grille régulière, et où la liste de poids est telle que $\{p_i = h[i]\}$. Il est important de souligner que, contrairement au calcul d'une distribution empirique, le poids total d'une signature n'est pas normalisé. Le poids peut donc varier d'une image à une autre.

Rubner *et al.* définissent dans ce cadre de travail une distance entre signatures qu'ils nomment **EMD** pour « Earth Mover Distance ». Si les deux signatures ont le même poids, cette mesure n'est autre que le transport optimal entre les deux signatures, vues comme des mesures discrètes. Si les deux signatures sont de poids différent, ils proposent de calculer le coût de transport de la signature de masse totale la plus faible vers la signature de masse la plus élevée. Si les deux signatures s'écrivent $f : \{(x_f[i], p_f[i])\,,\ \forall i = 1, \dots, N_f\}$ et $g : \{(x_g[j], p_g[j])\,,\ \forall j = 1, \dots, N_g\}$ et ont pour poids totaux respectifs $P_f = \sum_i p_f[i]$ et $P_g = \sum_j p_g[j]$, alors la distance $\text{EMD}(f, g)$ s'écrit :

$$\text{EMD}(f, g) := \min_{(\alpha_{i,j}) \in \mathcal{M}} \frac{\sum_{i=1}^{N_f} \sum_{j=1}^{N_g} \alpha_{i,j}\, c(x_f[i], x_g[j])}{\sum_{i=1}^{N_f} \sum_{j=1}^{N_g} \alpha_{i,j}}\,, \tag{5.3}$$

avec

$$\mathcal{M} = \left\{ (\alpha_{i,j});\ \alpha_{i,j} \geq 0,\ \sum_{j=1}^{N_g} \alpha_{i,j} \leq p_f[i],\ \sum_{i=1}^{N_f} \alpha_{i,j} \leq p_g[j],\ \sum_{i \leq N_f, j \leq N_g} \alpha_{i,j} = \min\left\{P_f, P_g\right\} \right\}$$

où $c(.,.)$ est toujours le coût de transport d'une masse unitaire entre deux positions. Dans les travaux de Rubner, ce coût est généralement la distance euclidienne.

> **Remarque 2 :**
> L'intérêt de l'utilisation de signatures à la place d'histogrammes est discuté dans [LB01]. La différence entre ces deux approches est que l'EMD correspond à un transport partiel dans le premier cas. Levina et Beckel montrent que si le temps de calcul est ainsi réduit, cela se traduit également par des performances moindres en termes de taux de reconnaissance.

Les travaux de Rubner *et al.* [RTG00] ont contribué à populariser ce type d'approche en traitement d'images. À un tel point qu'il est devenu courant dans ce domaine d'utiliser le terme « EMD » pour désigner le calcul du transport optimal entre des histogrammes.

Nous rappelons dans la section suivante quelques applications en analyse d'images pour lesquelles le transport optimal a été employé. Ces applications peuvent être classées selon deux catégories. La première, logiquement, concerne l'utilisation de la distance de Monge-Kantorovich comme mesure de dissimilarité entre des histogrammes. La seconde catégorie englobe les applications où c'est le transport optimal en tant que tel qui est exploité.

5.2 Applications du transport optimal et précédents travaux

Nous nous intéressons dans cette section à différentes applications du transport pour les histogrammes.

5.2.1 Comparaison d'histogrammes

Il existe de nombreuses applications qui nécessitent la comparaison d'histogrammes de caractéristiques, notamment en reconnaissance d'objets (voir la première partie de ce manuscrit). On appelle « bin », une cellule de quantification d'un histogramme. Pour comparer des histogrammes, on peut distinguer deux catégories de mesure de dissimilarité :

- les distances « bin-à-bin » (*bin to bin*), qui limitent la comparaison de deux histogrammes à des opérations sur les valeurs en des positions (ou indices) identiques. C'est le cas par exemple des normes L^p, ou encore de la divergence de Kullback-Leibler ;
- les distances « inter-bins » (*cross bin*), qui au contraire permettent de comparer des valeurs de positions différentes, à l'image de la distance de Monge-Kantorovich, ou de l'EMD.

Comme l'illustre le schéma de la figure 5.1, une mesure de dissimilarité définie dans le cadre du transport optimal est beaucoup plus robuste que les distances usuelles bin-à-bin à certaines classes de perturbations, comme l'erreur de quantification ou les translations des modes principaux des histogrammes.

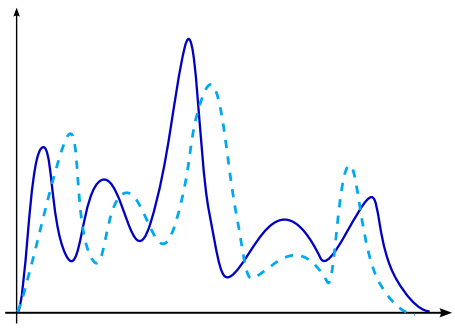


FIG. 5.1 – *Illustration des différents types de perturbations affectant un histogramme.*

Pour cette raison, l'utilisation du transport optimal en tant que mesure de dissimilarité a connu un essor spectaculaire, en particulier à la suite des travaux de [RTG00]. Voici quelques exemples d'applications :

- Reconnaissance de formes à partir de graphes bipartites [WPR85], et d'histogrammes [LO07] ;
- Recherche d'images par comparaison de signatures de couleurs ou de textures [RTG00] ;
- Détection de contours et de jonctions [RT01] ;
- Recherche d'images par analyse de l'organisation spatiale des couleurs [HGS08] ;
- Reconnaissance d'objets par mise en correspondance de descripteurs locaux [LO07] ;
- Détection d'objets par sac de mots (*Bag of features*) [ZMLS07].

Une des limitations pratiques du transport est son estimation numérique lorsque la dimension de l'espace est supérieure ou égale à 2. Dans la majorité des travaux précédemment cités, le coût du transport optimal est estimé en utilisant un algorithme du simplexe, qui permet de définir la solution exacte dans le cas discret. En utilisant une distance au sol euclidienne, la complexité de cet algorithme est –de manière

empirique– entre $\mathcal{O}(N^3)$ et $\mathcal{O}(N^4)$, où N désigne le nombre de bins des histogrammes. Kaijser [Kai98] montre que l'on peut réduire la complexité de cet algorithme à $\mathcal{O}(N^2)$ en utilisant une distance au sol $c(x,y) = \|x-y\|_1$ à la place de la distance euclidienne. Ce principe est utilisé par Ling et Okada [LO07] qui proposent un algorithme pour le calcul de l'EMD avec cette distance au sol. Toutefois, le temps de calcul reste prohibitif lorsque l'on souhaite utiliser ce type de distance sur une grande base de données ou comparer des histogrammes de grandes tailles.

Approximation de l'EMD Pour remédier à ce problème de complexité, plusieurs approximations ont été proposées.

Une première catégorie d'approximation concerne l'approche dite de « plongement métrique » (*metric embedding*). Son principe consiste à estimer de manière approchée une distance à partir d'un autre espace métrique, en contrôlant la distorsion de l'estimation. Dans [IT03], la mesure de dissimilarité EMD est calculée par plongement dans L^1. Cette estimation est obtenue par des grilles de quantification de l'espace à différentes échelles qui sont translatées aléatoirement. Les auteurs de [IT03] utilisent cette approximation pour accélérer de deux ordres de grandeur la recherche d'image dans une base avec les descripteurs définis par [RTG00]. Une approche analogue est proposée dans [LCL04]. Ce principe a ensuite été repris dans [GD04, GD05] pour la reconnaissance de formes et d'objets respectivement.

Très récemment, deux autres approches ont été suggérées pour calculer l'EMD de manière approchée en complexité linéaire. Shirdhonkar et Jacobs [SJ08] utilisent une formalisation duale de la distance de Monge-Kantorovich. Ils montrent qu'elle peut être approximée par un calcul de coefficients d'ondelettes. Dans [PW09], Pele et Werman utilisent une distance au sol tronquée, c'est-à-dire constante au-delà d'une certaine distance c_{max} : $c(x,y) = \min\{|x-y|, c_{max}\}$, ce qui leur permet de diminuer le temps de calcul.

Une dernière possibilité consiste à considérer le transport en une dimension. En effet, dans le cas particulier $1D$, le coût du transport s'exprime dans certains cas de manière analytique, ce qui réduit considérablement le temps de calcul. Pour se ramener à des comparaisons $1D$, il est possible de modifier la définition de la mesure de dissimilarité, comme nous le proposons dans le chapitre 7 pour la comparaison de descripteurs SIFT. Une autre solution, indépendamment proposée par Pitié et Kokaram [PKD07] et par Marc Bernot, repose sur des projections aléatoires $1D$. La mise en œuvre de ce principe est détaillée au chapitre 8.

Ceci nous amène au cas particulier du transport entre histogrammes unidimensionnels.

Transport sur la droite Soient f et g deux distributions continues unidimensionnelles, dont on note respectivement F et G les fonctions de répartition. Si le coût $c(x,y)$ est la distance $|x-y|$, un résultat bien connu est que le coût du transport optimal $\mathrm{MK}_1(f,g)$ s'exprime simplement comme $||F-G||_1$, la norme L^1 de la différence entre les fonctions de répartition F et G. Ce résultat, désigné par le terme *match distance* dans [RTG00, LB01], est lié au fait que le transport de Monge-Kantorovich préserve l'ordre des points sur $\mathbb{R}$ [Vil03] lorsque le coût c est une fonction convexe de $|x-y|$ (par exemple lorsque $p \geq 1$ dans l'équation (5.2)).

Transport sur le cercle Nous nous intéressons dans le chapitre 6 au cas particulier du transport entre des histogrammes *circulaires*, c'est-à-dire des distributions empiriques de variables périodiques 1D. Il existe plusieurs applications où l'on rencontre des histogrammes périodiques (non nécessairement 1D) :

- comparaison de descripteurs globaux : par exemple, les histogrammes d'orientation du gradient (reconnaissance de caractères avec [CS02]), et les histogrammes en espace HSV [LCL04] où la teinte est définie sur le cercle ;
- comparaison de descripteurs locaux : par exemple les SIFTs [Low04], composés d'histogrammes d'orientation du gradient, ou encore les descripteurs "Shape Context" [BMP02], histogrammes de contours en coordonnées polaires.

Dans le cas d'histogrammes circulaires, la distance au sol $c(.,.)$ est définie comme le plus court chemin sur le cercle. La propriété de préservation de l'ordre valable sur $\mathbb{R}$ n'a donc plus de sens. Dans le

prochain chapitre, nous allons montrer que la distance de Monge-Kantorovich peut cependant s'exprimer de manière très simple à partir des fonctions de répartition des distributions considérées. La distance obtenue, que l'on appelle CEMD, sera utilisée pour comparer des descripteurs d'images globaux et circulaires dans le cas unidimensionnel (histogrammes de teinte, d'orientation du gradient). Une analyse générale de l'intérêt du transport pour la comparaison d'histogrammes sera ensuite réalisée. Nous verrons au chapitre 7 comment exploiter CEMD pour la comparaison de descripteurs locaux dans le cadre de la reconnaissance d'objets, que nous avons présentée dans la première partie de ce manuscrit.

5.2.2 Transformation d'histogrammes

Nous avons considéré jusqu'à présent le coût du transport comme le moyen de mesurer la dissimilarité entre deux distributions. Une autre application consiste à exploiter le transport optimal en tant que tel, c'est-à-dire comme un moyen d'envoyer une mesure sur une autre de la manière la moins coûteuse possible.

L'une des applications les plus courantes est la spécification d'histogramme, qui permet d'ajuster le contraste d'une image. Plus récemment, ce principe a été étendu aux couleurs pour modifier la palette d'une image. En particulier, on notera les travaux de J. Delon [Del04] sur la définition d'histogrammes mi-chemin (*midway histograms*) pour égaliser les palettes couleurs de deux images. L'ensemble de ces applications sont étudiées au chapitre 8.

Une autre application du transport consiste à faire du morphage (ou *morphing*) d'images [ZYHT07] ou de texture (*texture mapping*) [DT09].

Une des limitations du flot du transport optimal pour la transformation d'histogrammes est qu'il produit des artefacts dans les images obtenues, dont les causes sont analysées au chapitre 6. En effet, le transfert défini par cette méthode ne tient pas compte des dépendances entre les valeurs des pixels dans l'image d'origine. Nous proposons dans le chapitre 8 une nouvelle méthode de régularisation du transport qui permet de préserver la géométrie de l'image d'origine dans l'image obtenue par transfert.

Chapitre 6

Comparaison d'histogrammes circulaires

Dans ce chapitre, nous étudions le transport optimal sur le cercle. Cette étude nous permettera de définir une distance appelée CEMD. L'intérêt du transport en tant que mesure de dissimilarité est ensuite analysé à partir de quelques exemples.

6.1 Étude du transport pour les histogrammes unidimensionnels et circulaires (CEMD)

Dans cette section, nous nous intéressons à l'étude du transport pour les histogrammes unidimensionnels et circulaires (*i.e.* représentant des données périodiques, telle qu'une orientation). Nous étudions la distance de Monge-Kantorovich MK_λ définie au chapitre précédent (Formule (5.2)), où cette fois λ est un réel positif. Nous allons ici montrer que, dans le cas discret, cette distance peut être exprimée à partir des fonctions de répartition sur le cercle des deux distributions qui sont comparées.

Une preuve alternative de l'expression de la distance de Monge-Kantorovich dans le cas $p = 1$ (voir formule (5.2)) à été proposée par Werman *et al* [WPMK86] pour des distributions discrètes de points deux à deux distincts sur le cercle.

6.1.1 Notations

Nous considérons dans cette section les deux ensembles discrets de points $\{x_1, \dots x_P\}$ et $\{y_1, \dots y_P\}$ sur le cercle unité (noté S^1), et leurs distributions discrètes respectives

$$f = \frac{1}{P}\sum_{k=1}^{P} \delta_{x_k}, \text{ et } g = \frac{1}{P}\sum_{k=1}^{P} \delta_{y_k},$$

où les notations x_k, y_k représentent indifféremment la position des points sur le cercle ou leurs coordonnées dans $[0, 1[$. Par ailleurs, les indices sont également définis de manière circulaire, *i.e.* $k - 1 \equiv P$ lorsque $k = 1$ et $k + 1 \equiv 1$ lorsque $k = P$. Soit $c(.,.)$ la distance périodique L^1 sur $[0, 1[$:

$$c(x, y) = \min(|x - y|, 1 - |x - y|). \tag{6.1}$$

Pour tout $\lambda > 0$, on rappelle que l'on peut définir une distance entre deux distributions f et g de la manière suivante (voir le chapitre 5)

$$\text{MK}_\lambda(f, g) := \min_{(\alpha_{i,j})\in\mathcal{M}} \left(\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_{i,j} c(x_i, y_j)^\lambda\right)^{\frac{1}{\lambda}}, \text{ où} \tag{6.2}$$

$$\mathcal{M} = \{(\alpha_{i,j});\ \alpha_{i,j} \geq 0,\ \sum_j \alpha_{i,j} = \frac{1}{P},\ \sum_i \alpha_{i,j} = \frac{1}{P}\}\ . \tag{6.3}$$

Ces distances (voir [Vil03] pour la preuve que ce sont bien des distances) sont appelées distances de Monge-Kantorovich. Dans le cas particulier où $\lambda = 1$, et où les ensembles $\{x_k\}$ et $\{y_k\}$ ont le même cardinal, cette définition est équivalente à la distance du « cantonnier », ou Earth Mover's Distance (EMD) [RTG00].

La distance de Monge-Kantorovich MK_λ entre deux distributions discrètes f et g revient à calculer (voir par exemple l'introduction du livre de Villani [Vil03])

$$\mathrm{MK}_\lambda(f,g) := \min_{\sigma \in \Sigma_P} W_\sigma^\lambda(f,g) \tag{6.4}$$

où Σ_P est l'ensemble des permutations de $\{1,\dots,P\}$ et où

$$W_\sigma^\lambda(f,g) := \frac{1}{P}\left(\sum_k c(x_k, y_{\sigma(k)})^\lambda\right)^{\frac{1}{\lambda}} \tag{6.5}$$

est le coût de transport de f vers g selon la permutation σ. Autrement dit, trouver le transport optimal des masses de f vers celles de g revient à trouver la permutation optimale σ entre les points $\{x_k\}$ et $\{y_k\}$.

Définition du chemin Si x et y sont deux points distincts de S^1, on considère $\gamma(x,y)$ **la géodésique reliant** x **à** y **sur** S^1. $\gamma(x,y)$ est défini comme un chemin ouvert : il ne contient ni x, ni y. Dans le cas particulier où x et y sont en des positions opposées sur le cercle (soit $y = x + 1/2\,[1]$), $\gamma(x,y)$ est défini comme le chemin allant de x vers y dans le sens trigonométrique sur S^1. Ainsi, le chemin $\gamma(x,y)$ est toujours défini de manière unique. On qualifie la géodésique $\gamma(x,y)$ de **positive** lorsque le chemin de x vers y est dans le sens trigonométrique. Sinon, elle est qualifiée de **négative**.

Fonctions de répartition sur le cercle La fonction de répartition de $f = \frac{1}{P}\sum_{k=1}^P \delta_{x_k}$ sur le segment $[0,1[$ est généralement définie à partir de l'origine 0 :

$$\forall y \in [0,1[,\quad F(y) = \frac{1}{P}\sum_{k=1}^{P} \mathbb{1}_{\{x_k \in [0,y[\}}. \tag{6.6}$$

Or, dans notre cas, le segment $[0,1[$ peut être vu comme le cercle unité S^1, si bien qu'il n'y a pas d'ordre privilégié entre les différents points. Autrement dit, il est possible de définir une fonction de répartition à partir de n'importe quel point de référence sur le cercle. Nous définissons ainsi F_x, la fonction de répartition de f à partir du point $x \in [0,1[$ sur le cercle unité dans le sens trigonométrique. F_x peut être exprimée analytiquement en fonction de f et de x de la manière suivante :

$$\forall y \in [0,1[,\quad F_x(y) = \begin{cases} \frac{1}{P}\sum_{k=1}^P \mathbb{1}_{\{x_k \in [x, y+x[\}} & \text{si } y < 1-x, \\ \frac{1}{P}\sum_{k=1}^P \mathbb{1}_{\{x_k \in [x,1[\cup[0,y-1+x[\}} & \text{si } y \geq 1-x \end{cases}. \tag{6.7}$$

La figure 6.1 donne un exemple d'une fonction de répartition F à partir de l'origine 0, et de la fonction F_x à partir de la coordonnée x qui lui correspond. On remarque ainsi que l'on peut définir la fonction F_x à partir de la fonction F :

$$\forall (x,y) \in [0,1[^2,\quad F_x(y) = \begin{cases} F(y+x) - F(x) & \text{si } y < 1-x\ . \\ F(y+x-1) + 1 - F(x) & \text{si } y \geq 1-x \end{cases}, \tag{6.8}$$

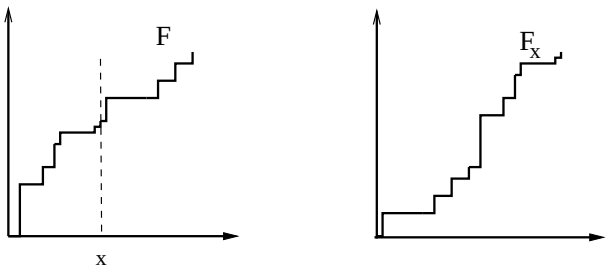


FIG. 6.1 – *(*Figure de gauche*) Fonction de répartition F depuis l'origine* 0*. (*Figure de droite*) Fonction de répartition F_x depuis la coordonnée x.*

6.1.2 Calcul de la distance de Monge-Kantorovich sur le cercle

Dans ce qui suit, nous allons montrer que la distance de Monge-Kantorovich sur le cercle peut être exprimée analytiquement et calculée à moindre coût, notamment dans le cas $\lambda = 1$ qui correspond à la distance EMD entre deux ensembles de points. Pour cela, nous montrons que si σ est une permutation optimale au sens de l'équation (6.4), alors il existe toujours un point sur le cercle qui n'appartient à aucun des chemins optimaux de σ. Ce résultat est obtenu dans un premier temps avec $\lambda > 1$ pour n'importe quelle permutation optimale σ, et ensuite avec $\lambda = 1$ pour une configuration bien choisie de permutation optimale. Cela signifie que le calcul de la distance de Monge-Kantorovich sur le cercle revient au calcul de la même distance sur un intervalle de $\mathbb{R}$.

Proposition 4 *Supposons que $\lambda > 1$. Soient $x_1, \dots x_P$ et $y_1, \dots y_P$, P points de $[0, 1[$. Supposons que tous ces points sont deux à deux différents. Alors pour toute permutation σ de Σ_P qui minimise (6.4) avec le coût (6.1), il existe $k \in \{1, \dots P\}$ tel que pour tout $l \neq k$, $x_k \notin \gamma(x_l, y_{\sigma(l)})$.*

La preuve de la proposition 4 repose sur le lemme suivant, qui donne quelques propriétés des géodésiques $\gamma(x_l, y_{\sigma(l)})$ obtenues lorsque σ est un minimiseur de (6.4) avec $\lambda > 1$.

Lemme 1 *Supposons que $\lambda > 1$. Soit σ un minimiseur de (6.4), et les chemins $\gamma_l = \gamma(x_l, y_{\sigma(l)})$ et $\gamma_k = \gamma(x_k, y_{\sigma(k)})$ (avec $l \neq k$) deux géodésiques selon l'assignement défini par σ. Supposons également que $x_l \neq x_k$ et $y_{\sigma(l)} \neq y_{\sigma(k)}$. Alors, l'une des deux hypothèses est vérifiée :*

- *$\gamma_l \cap \gamma_k = \emptyset$;*
- *$\gamma_l \cap \gamma_k \neq \emptyset$ et dans ce cas γ_l et γ_k ont la même orientation (tous deux positifs ou négatifs) et aucun n'est contenu dans l'autre.*

Preuve du lemme 1 Supposons que $\gamma_l \cap \gamma_k \neq \emptyset$. Si $\gamma_l \cap \gamma_k$ est égal à $\gamma(x_l, x_k)$ ou $\gamma(y_{\sigma(l)}, y_{\sigma(k)})$, alors

$$c(x_l, y_{\sigma(l)})^\lambda + c(x_k, y_{\sigma(k)})^\lambda > c(x_l, y_{\sigma(k)})^\lambda + c(x_k, y_{\sigma(l)})^\lambda,$$

ce qui contredit l'hypothèse d'optimalité de σ. De plus, λ étant strictement supérieur à 1, la fonction $x \mapsto |x|^\lambda$ est strictement convexe. Si, par exemple, le chemin γ_l est inclus dans γ_k, alors

$$c(x_l, y_{\sigma(l)})^\lambda + c(x_k, y_{\sigma(k)})^\lambda > c(x_l, y_{\sigma(k)})^\lambda + c(x_k, y_{\sigma(l)})^\lambda,$$

ce qui contredit également l'hypothèse d'optimalité de σ. La seule alternative possible est donc que $\gamma_l \cap \gamma_k$ est égal à $\gamma(x_l, y_{\sigma(k)})$ ou bien à $\gamma(x_k, y_{\sigma(l)})$. Il s'ensuit que γ_k et γ_l sont tous deux de même orientation sur le cercle (soit tous les deux positifs, soit tous les deux négatifs). □

Preuve de la proposition 4 Soit σ un minimiseur de (6.4). Par souci de clarté, on note dans ce qui suit γ_l le chemin $\gamma(x_l, y_{\sigma(l)})$. Sans perdre en généralité, on peut supposer que les points $\{x_1, \dots, x_P\}$ sont donnés dans l'ordre trigonométrique sur le cercle.

Supposons que la proposition 4 est fausse. Dans ce cas, pour chaque $l \in \{1, \dots P\}$, il existe $q(l) \neq l$ tel que x_l appartient au chemin ouvert $\gamma_{q(l)}$. Alors, pour chaque l, on obtient $\gamma_{q(l)} \cap \gamma_l \neq \emptyset$, ce qui signifie d'après le lemme 1 que les chemins $\gamma_{q(l)}$ et γ_l ont la même orientation. Supposons par exemple qu'ils sont tous les deux positifs, et montrons que dans ce cas $x_l \in \gamma_{l-1}$.

Si $q(l) = l - 1$, le résultat est immédiat. Si $q(l) \neq l - 1$, cela signifie que $x_{q(l)}$, x_{l-1}, x_l sont dans l'ordre trigonométrique sur le cercle. Comme $\gamma_{q(l)}$ est un chemin positif partant du point $x_{q(l)}$ et contenant x_l, il contient également le point x_{l-1} (car les points étant supposés deux à deux distincts, on a $x_{l-1} \neq x_{q(l)}$). Ainsi $\gamma_{l-1} \cap \gamma_{q(l)} \neq \emptyset$, ce qui implique que le chemin γ_{l-1} est également positif.

On en déduit alors que x_l est nécessairement inclus dans γ_{l-1}, car d'après le lemme 1, le chemin γ_{l-1} ne peut être inclus dans $\gamma_{q(l)}$. En conclusion, s'il existe pour chaque $l \in \{1, \dots P\}$, $q(l) \neq l$ tel que $x_l \in \gamma_{q(l)}$, alors $x_l \in \gamma_{l-1}$ lorsque γ_l est positif, et on montre de la même manière que lorsque γ_l est négatif, alors $x_l \in \gamma_{l+1}$.

Supposons maintenant que pour un $k \in \{1, \dots P\}$ donné, le chemin γ_k soit positif, si bien que d'après le raisonnement précédent $x_k \in \gamma_{k-1}$. D'après le lemme 1, le chemin γ_{k-1} est également positif, ce qui signifie que x_{k-1} doit appartenir à γ_{k-2}. De manière récursive, on en déduit que pour chaque $l \in \{1, \dots P\}$, $x_l \in \gamma_{l-1}$. Finalement, envoyer x_l sur $y_{\sigma(l-1)}$ est strictement moins coûteux que de l'envoyer sur $y_{\sigma(l)}$: $c(x_l, y_{\sigma(l-1)}) < c(x_l, y_{\sigma(l)})$. En comparant le coût de l'assignement σ avec celui de l'assignement correspondant à l'élément précédent, on obtient

$$\sum_{l=1}^{P} c(x_l, y_{\sigma(l)})^\lambda > \sum_{l=1}^{P} c(x_l, y_{\sigma(l-1)})^\lambda, \tag{6.9}$$

ce qui vient contredire le fait que σ est un minimiseur de (6.4). On arrive à la même conclusion en supposant que pour un $k \in \{1, \dots P\}$ donné, le chemin γ_k est négatif. □

Le même résultat peut être obtenu lorsque $\lambda = 1$, mais il n'est cependant vérifié que pour un bon choix de permutation σ minimiseur de (6.4), et non pour toutes les permutations qui sont solutions de (6.4) comme dans le cas $\lambda > 1$. Ce résultat peut être vu comme le cas limite de la proposition 4 où $\lambda \to 1$.

Corollaire 1 *Supposons que $\lambda = 1$. Soient $\{x_1, \dots, x_P\}$ et $\{y_1, \dots, y_P\}$ deux ensembles de P points dans $[0, 1[$. On suppose que tous ces points sont deux à deux distincts. Alors, il existe une permutation σ de Σ_P qui minimise (6.4) et un point $x_k \in \{x_1, \dots x_P\}$ tel que pour tout $l \neq k$, $x_k \notin \gamma(x_l, y_{\sigma(l)})$.*

Preuve du corollaire 1 Nous savons d'après la proposition 4 que pour tout $\lambda > 1$, si σ_λ minimise le coût $\sigma \mapsto W_\sigma^\lambda(f, g)$, il existe $k \in \{1, \dots P\}$ tel que pour tout $l \neq k$, $x_k \notin \gamma_l = \gamma(x_l, y_{\sigma_\lambda(l)})$.

Si σ et les points $x_1, \dots, x_P$, et $y_1, \dots, y_P$ sont *fixés*, $W_\sigma^\lambda(f, g)$ est alors une fonction continue de λ. Ainsi, pour tout $\varepsilon > 0$, il existe $\beta > 1$ tel que pour tout $\lambda \in [1, \beta]$, $|W_\sigma^\lambda(f, g) - W_\sigma^1(f, g)| \leq \varepsilon$. Σ_P étant un ensemble fini, on peut choisir β suffisamment proche de 1 pour que cette propriété soit satisfaite pour tout σ de Σ_P. On peut également choisir β de manière à vérifier $|\min_\sigma W_\sigma^1(f, g) - \min_\sigma W_\sigma^\lambda(f, g)| \leq \varepsilon$ (le minimum d'un ensemble fini de fonctions continues étant une fonction continue). Il s'ensuit que pour $\lambda \in [1, \beta]$, en notant $\sigma_\lambda = \underset{\sigma \in \Sigma_P}{\operatorname{argmin}} W_\sigma^\lambda(f, g)$

$$\begin{aligned} |\min_\sigma W_\sigma^1(f, g) - W_{\sigma_\lambda}^1(f, g)| &= |\min_\sigma W_\sigma^1(f, g) - W_{\sigma_\lambda}^\lambda(f, g) + W_{\sigma_\lambda}^\lambda(f, g) - W_{\sigma_\lambda}^1(f, g)| \\ &\leq |\min_\sigma W_\sigma^1(f, g) - \min_\sigma W_\sigma^\lambda(f, g)| + |W_{\sigma_\lambda}^\lambda(f, g) - W_{\sigma_\lambda}^1(f, g)| \\ &\leq 2\varepsilon. \end{aligned}$$

Ce qui montre que lorsque λ est suffisamment proche de 1, un minimiseur σ_λ de $W_\sigma^\lambda(f, g)$ est également un minimiseur de $W_\sigma^1(f, g)$. Cela prouve qu'il existe alors *au moins un minimiseur* σ de $\sigma \mapsto W_\sigma^1(f, g)$ tel que $x_k \notin \gamma(x_l, y_{\sigma_\lambda(l)})$ pour $k \in \{1, \dots, P\}$, et ceci quel que soit $l \neq k$. □

Nous allons maintenant montrer que la conséquence directe de ce résultat est que le calcul du transport optimal sur le cercle revient à chercher une coupure de S^1, de telle sorte que ce transport soit optimal sur la droite réelle. Cela nous permet d'écrire la distance de Monge-Kantorovich d'indice λ dans le cas du cercle sous une nouvelle forme à partir de la proposition 4 et du corollaire 1, lorsque tous les points sont distincts.

Corollaire 2 *Supposons que $x_1, \dots, x_P$ et $y_1, \dots, y_P$ sont des points deux à deux différents. Alors,*

$$\forall \lambda \geq 1, \quad \mathrm{MK}_\lambda(f,g) = \left(\inf_{x \in S^1} \int |F_x^{-1} - G_x^{-1}|^\lambda \right)^{1/\lambda}, \tag{6.10}$$

où F_x^{-1} et G_x^{-1} sont respectivement les fonctions pseudo-inverses des fonctions croissantes F_x et G_x, définies comme : $F_x^{-1}(y) = \inf\{t;\ F_x(t) > y\}$ et $G_x^{-1}(y) = \inf\{t;\ G_x(t) > y\}$.

Preuve du corollaire 2 Nous avons vu que pour tout $\lambda \geq 1$, si $x_1, \dots, x_P$ et $y_1, \dots, y_P$ sont des points distincts, une permutation optimale σ_λ (minimisant 6.4) peut être choisie de manière à ce qu'il existe un point x_k qui n'est contenu dans aucun des chemins $\{\gamma(x_l, y_{\sigma_\lambda(l)})\}_{1 \leq l \leq P}$ définis par σ_λ (pour rappel, les chemins $\gamma(x_l, y_{\sigma_\lambda(l)})$ ont été définis comme des ouverts, si bien qu'il ne contiennent pas leurs frontières x_l et $y_{\sigma_\lambda(l)}$). Puisque les points sont supposés être deux à deux différents, le seul chemin qui inclut une partie du voisinage de x_k est γ_k. Il existe donc un ensemble ouvert ne contenant pas x_k (situé d'un côté ou de l'autre de x_k selon l'orientation de γ_k), qui n'est inclus dans aucun des chemins définis par la permutation optimale σ_λ. En particulier, le milieu x de cet ensemble ouvert n'appartient à aucun de ces chemins. **On peut donc « couper » le cercle S^1 en ce point x de façon à réduire le problème du transport sur le cercle à celui du transport sur la ligne réelle**, ainsi que l'illustre la figure 6.2. Or, dans le cas de la droite réelle, la distance Monge-Kantorovich d'indice λ entre deux distributions f et g sur $\mathbb{R}$ peut être exprimée à l'aide des pseudo-inverses des fonctions de répartition F et G (voir la remarque 2 du théorème 13 dans [Vil03]) :

$$\forall \lambda \geq 1, \quad \mathrm{MK}_\lambda(f,g) = \left(\int |F^{-1} - G^{-1}|^\lambda \right)^{1/\lambda},$$

où F^{-1} et G^{-1} sont respectivement les fonctions pseudo-inverses des fonctions de répartition F et G depuis $-\infty$, définies ainsi : $F^{-1}(y) = \inf\{t \in \mathbb{R};\ F(t) > y\}$ et $G^{-1}(y) = \inf\{t \in \mathbb{R};\ G(t) > y\}$. Ainsi, en choisissant x comme point de référence sur le cercle, on peut exprimer de manière similaire la distance Monge-Kantorovich d'indice λ entre deux distributions f et g sur S^1 à l'aide des pseudo-inverses de F_x et G_x, les fonctions de répartition de f et de g à partir de x :

$$\forall \lambda \geq 1, \quad \mathrm{MK}_\lambda(f,g) = \left(\inf_{x \in S^1} \int |F_x^{-1} - G_x^{-1}|^\lambda \right)^{1/\lambda} = \inf_{x \in S^1} \|F_x^{-1} - G_x^{-1}\|_\lambda . \tag{6.11}$$

□

Nous venons de montrer que le transport optimal entre deux distributions de points sur le cercle, minimisant la distance Monge-Kantorovich d'indice λ, permet de couper S^1 en un point par lequel ne passe aucun chemin. Dans le cas $\lambda = 1$, sachant que les fonctions F_x et G_x sont des applications de $[0,1[$ dans $[0,1]$, l'expression (6.11) peut être réécrite sous la forme (voir la formule (72) [Vil03]) :

$$\mathrm{MK}_1(f,g) = \inf_{x \in S^1} \|F_x - G_x\|_1. \tag{6.12}$$

Afin de généraliser l'expression (6.12) au cas où les points $x_1, \dots, x_P$ et $y_1, \dots, y_P$ peuvent coïncider, nous allons tout d'abord montrer que $\mathrm{MK}_1(f,g)$ peut s'écrire comme $\min_{l \in \{-P, \dots P\}} \|F - G - \frac{l}{P}\|_1$,

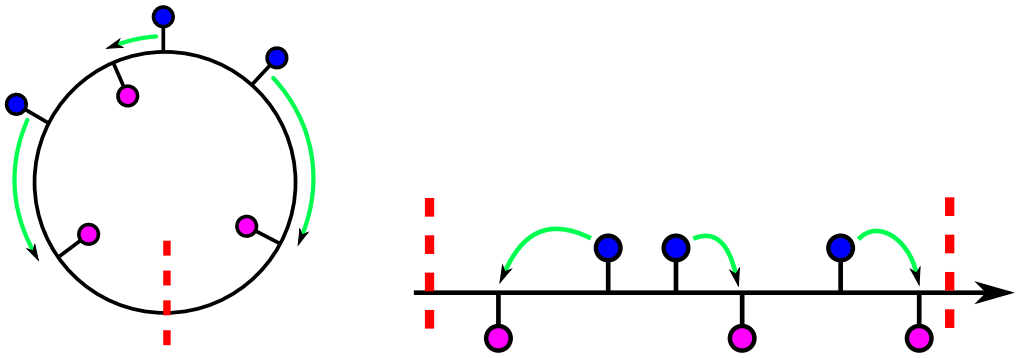


FIG. 6.2 – *Avec une distance au sol strictement convexe ($\lambda > 1$) et dans le cas limite où $\lambda = 1$, le transport optimal sur le cercle S^1 revient au transport sur la droite réelle si l'on connaît le lieu du cercle où l'on peut couper.*

expression qu'il sera plus facile de manipuler par la suite. Nous avons vu en préliminaire de cette section (équation (6.8)), que la fonction F_x peut s'exprimer exclusivement à l'aide de F, ce qui nous donne ici :

$$\begin{aligned}
\|F_x - G_x\|_1 &= \int_0^{1-x} |F(t+x) - F(x) - G(t+x) + G(x)|dt \\
&\quad + \int_{1-x}^{1} |F(t+x-1) - F(x) - G(t+x-1) + G(x)|dt \\
&= \int_x^1 |F(t) - F(x) - G(t) + G(x)|dt + \int_0^x |F(t) - F(x) - G(t) + G(x)|dt \\
&= \|F - F(x) - G + G(x)\|_1 \ .
\end{aligned}$$

Or, bien que l'ensemble des x solutions de (6.12) soit un ensemble ouvert de $[0,1[$, la quantité $F(x) - G(x)$ est quantifiée et prend l'une des valeurs à $\{\frac{l}{P};\ l = -P, \dots P\}$. Cela signifie donc que la quantité $\|F - G - \frac{l}{P}\|_1$ atteint un minimum pour l'une de ces valeurs (solution non nécessairement unique). On peut donc finalement écrire

$$\mathrm{MK}_1(f,g) = \min_{l \in \{-P,\dots P\}} \|F - G - \frac{l}{P}\|_1.$$

Nous allons montrer, à la suite du corollaire suivant, que ce résultat peut être généralisé au cas où les points $x_1, \dots, x_P$ et $y_1, \dots, y_P$ ne sont pas nécessairement distincts.

Corollaire 3 *Soient $f = \frac{1}{P}\sum_{k=1}^P \delta_{x_k}$ et $g = \frac{1}{P}\sum_{k=1}^P \delta_{y_k}$ deux distributions discrètes sur S^1, les points $x_1, \dots, x_P$ et $y_1, \dots, y_P$ pouvant coïncider. La distance Monge Kantorovich d'indice $\lambda = 1$ entre f et g peut être exprimée comme*

$$\mathrm{MK}_1(f,g) = \min_{l \in \{-P,\dots P\}} \|F - G - \frac{l}{P}\|_1. \tag{6.13}$$

Preuve du corollaire 3 Soient $2P$ points $x_1, \dots, x_P, y_1, \dots, y_P$ sur S^1, non nécessairement différents. On peut construire pour chaque $\varepsilon > 0$, les points $x_1^\varepsilon, \dots x_P^\varepsilon, y_1^\varepsilon, \dots y_P^\varepsilon$, deux à deux différents, tels que $\forall k \in \{1, \dots P\},\ c(x_k^\varepsilon, x_k) \leq \varepsilon$ et $c(y_k^\varepsilon, y_k) \leq \varepsilon$. Soient $f^\varepsilon = \frac{1}{P}\sum_{k=1}^P \delta_{x_k^\varepsilon}$ et $g^\varepsilon = \frac{1}{P}\sum_{k=1}^P \delta_{y_k^\varepsilon}$ les distributions qui leur sont respectivement associées. Alors, σ étant fixé, $\forall \sigma \in \Sigma_P$

$$|W_\sigma^1(f^\varepsilon, g^\varepsilon) - W_\sigma^1(f,g)| \leq \sum_{k=1}^P |c(x_k^\varepsilon, y_{\sigma(k)}^\varepsilon) - c(x_k, y_{\sigma(k)})| \leq 2P\varepsilon \ . \tag{6.14}$$

Puisque MK_1 est un minimum sur un ensemble fini, il s'ensuit que

$$\mathrm{MK}_1(f^\varepsilon, g^\varepsilon) \xrightarrow[\varepsilon\to 0]{} \mathrm{MK}_1(f, g) \ .$$

Or, si l'on note F^ε et G^ε les fonctions de répartition de f^ε et g^ε, on a :

$$\forall l \in \{-P, \dots P\},\ \|F^\varepsilon - G^\varepsilon - \frac{l}{P}\|_1 \xrightarrow[\varepsilon\to 0]{} \|F - G - \frac{l}{P}\|_1.$$

Puisque l prend un nombre fini de valeurs, on obtient la convergence suivante :

$$\min_{l\in\{-P,\dots P\}} \|F^\varepsilon - G^\varepsilon - \frac{l}{P}\|_1 \xrightarrow[\varepsilon\to 0]{} \min_{l\in\{-P,\dots P\}} \|F - G - \frac{l}{P}\|_1.$$

Nous pouvons ainsi finalement écrire

$$\mathrm{MK}_1(f, g) = \min_{l\in\{-P,\dots P\}} \|F - G - \frac{l}{P}\|_1 \ .$$

□

En remarquant que la quantité $\|F - G - \frac{l}{P}\|_1$ atteint son minimum pour l'une des valeurs prises par $\|F - G - F(x) + G(x)\|_1$, on peut donc conclure que la formule (6.12) est valide dans le cas général où les points peuvent coïncider.

6.1.3 Calcul de l'EMD pour des histogrammes discrets sur le cercle

Considérons maintenant deux histogrammes discrets $f = (f[i])_{i=0\dots N-1}$ et $g = (g[i])_{i=0\dots N-1}$, échantillonnés sur N cellules de quantification de tailles égales (c'est-à-dire avec un pas de quantification uniforme). On suppose que f et g sont normalisés, *i.e.* $\sum_{i=0}^{N-1} f[i] = \sum_{i=0}^{N-1} g[i] = 1$, et qu'ils sont « circulaires », c'est-à-dire que la première cellule (d'indice 0) est voisine de la dernière (d'indice $N-1$). Ces deux histogrammes peuvent alors être considérés comme des distributions de probabilité sur le cercle unité S^1, ou de manière équivalente, comme des distributions périodiques de période 1 sur $\mathbb{R}$. Ces distributions s'expriment comme

$$f = \sum_{i=0}^{N-1} f[i]\delta_{i/N} \ , \text{ et } g = \sum_{i=0}^{N-1} g[i]\delta_{i/N} \ ,$$

où i/N représente la coordonnée de la i-ième cellule de quantification sur le cercle S^1 ou sur l'ouvert $[0, 1[$. Les poids $f[i]$ et $g[i]$ sont rationnels dans le cas général (histogrammes obtenus numériquement), il est donc toujours possible de dupliquer les points qui leurs sont associés autant de fois que nécessaire de manière à pouvoir écrire

$$f = \frac{1}{P}\sum_{k=1}^{P} \delta_{x_k} \ , \text{ et } g = \frac{1}{P}\sum_{k=1}^{P} \delta_{y_k},$$

où $\{x_1, \dots x_P\}$ et $\{y_1, \dots y_P\}$ sont deux ensembles de points sur S^1 qui peuvent coïncider, leurs positions étant quantifiées selon $\{\frac{i}{N}\}_{0\le i\le N}$. À titre d'exemple, si $N = 3$ et $f = (3/4, 1/4, 0)$, on peut définir de manière équivalente $f = \frac{1}{4}(\delta_0 + \delta_0 + \delta_0 + \delta_{1/3})$.

La comparaison de deux histogrammes définis sur le cercle peut ainsi être effectuée avec la distance de Monge-Kantorovich calculée selon le coût défini à l'équation (6.1) et $\lambda = 1$. On note CEMD, pour « Circular Earth Mover's Distance », la distance ainsi définie entre deux histogrammes f et g

$$\mathrm{CEMD}(f, g) = \mathrm{MK}_1(f, g) = \inf_{x\in[0,1[} \|F - G - F(x) + G(x)\|_1 = \min_{l\in\{-P,\dots P\}} \|F - G - \frac{l}{P}\|_1 \ , \quad (6.15)$$

où F et G sont les fonctions de répartition de f et de g. Remarquons que l'on peut alors écrire de manière équivalente $\text{CEMD}(f,g) = \min_{l\in\{-P,\dots P\}} \sum_{k=-P}^{P} \left|\frac{k-l}{P}\right| \omega_k$, avec $\omega_k = mesure\{z \in [0,1[;\ F(z) - G(z) = \frac{k}{P}\}$. Autrement dit, la valeur de l qui minimise cette quantité peut être calculée comme le médian pondéré des valeurs de $F(z) - G(z)$.

Or, lorsque les histogrammes sont discrets, les points x_k et y_k sont localisés sur une grille régulière de N niveaux de quantification, la fonction de coût $c(.,.)$ est alors définie comme :

$$\forall i,j \in \{0,\dots,N-1\}\,,\ c(i,j) = \frac{1}{N}\min(|i-j|, N-|i-j|)\,, \tag{6.16}$$

et les *histogrammes cumulés discrets* F et G de f et g s'écrivent ainsi :

$$\forall j \in \{0,\dots,N-1\}\,,\ F[j] = \sum_{i=0}^{j} f[i]\,,\ \text{et } G[j] = \sum_{i=0}^{j} g[i]\,.$$

Cela revient à exprimer $\text{CEMD}(f,g)$ dans le cas discret sous la forme :

$$\text{CEMD}(f,g) = \frac{1}{N}\min_{j\in\{0,\dots,N-1\}} \sum_{i=0}^{N-1} |F[i] - G[i] - F[j] + G[j]| = \frac{1}{N}\sum_{i=0}^{N-1} |F[i] - G[i] - \mu|\,, \tag{6.17}$$

où μ est le médian des valeurs $\{F[i] - G[i]\}_{0\le i\le N-1}$.

Une définition équivalente de CEMD peut être obtenue de manière analogue à la distance de Monge-Kantorovich définie avec la formule (6.12).

$$\text{CEMD}(f,g) = \frac{1}{N}\min_{k\in\{0,\dots,N-1\}} \|F_k - G_k\|_1\,, \tag{6.18}$$

où F_k et G_k sont les histogrammes cumulés discrets de f et g depuis la k-ième cellule de quantification :

$$\forall k,i \in \{0,\dots,N-1\}\,,\ F_k[i] = \begin{cases} \displaystyle\sum_{j=k}^{k+i} f[j] & \text{if} \quad 0 \le i \le N-k-1 \\ \displaystyle\sum_{j=k}^{N-1} f[j] + \sum_{j=0}^{k+i-N} f[j] & \text{if} \quad N-k \le i \le N-1 \end{cases}.$$

La définition est similaire pour G_k en remplaçant f par g. Avec cette définition, on a

$$\forall i,k \in \{0,\dots,N-1\}\,,\ F_k[i] = \begin{cases} F[k+i] - F[k-1] & \text{si } i \le N-k-1 \\ F[k+i-N] + 1 - F[k-1] & \text{si } i \ge N-k \end{cases},$$

avec la convention $F[-1] = 0$. Ce qui nous donne l'équivalence $\forall k \in \{0,\dots,N-1\}$,

$$\begin{aligned} \|F_k - G_k\|_1 &= \sum_{i=0}^{N-1} |F_k[i] - G_k[i]| \\ &= \sum_{i=0}^{N-1-k} |F[k+i] - F[k-1] - G[k+i] + G[k-1]| \\ &\quad + \sum_{i=N-k}^{N-1} |F[k+i-N] - F[k-1] - G[k+i-N] + G[k-1]| \\ &= \sum_{j=k}^{N-1} |F[j] - G[j] - F[k-1] + G[k-1]| + \sum_{j=0}^{k-1} |F[j] - G[j] - F[k-1] + G[k-1]| \\ &= \|F - G - F[k-1] + G[k-1]\|_1\,. \end{aligned}$$

Ainsi, la distance CEMD(f,g) revient à calculer le minimum sur k de la distance L^1 entre F_k et G_k, les histogrammes cumulés de f et g depuis la k-ième cellule de quantification. Encore une fois, cela montre que la distance EMD définie avec coût circulaire revient à chercher sur le cercle la coupure qui minimise le coût de transport total de l'EMD sur la droite (voir la figure 6.3).

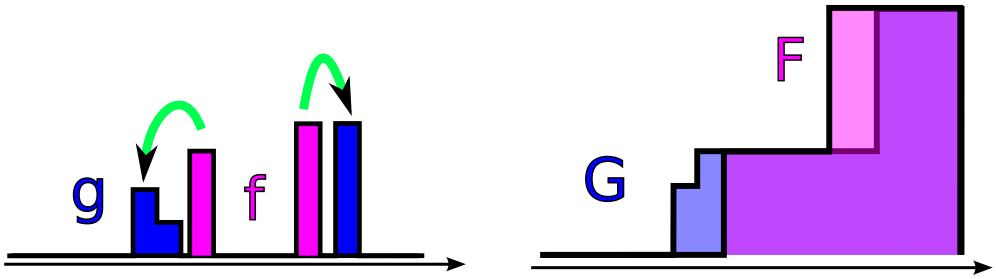


FIG. 6.3 – *Lorsque ($\lambda = 1$), le coût du transport optimal entre deux histogrammes discrets sur le cercle est obtenu en cherchant la coupure de S^1 minimisant le transport sur la droite obtenue. Cela revient à minimiser la norme L^1 de la différence des histogrammes cumulés (ici définis à l'équation 6.19) selon le premier bin choisi pour le cumul.*

Remarquons pour finir que la formule 6.18 reste valable pour n'importe quelle version translatée de F_k et G_k. En particulier, en définissant

$$\forall k, i \in \{0,\dots,N-1\}\,,\ \tilde{F}_k[i] = \begin{cases} \displaystyle\sum_{j=k}^{i} f[j] & \text{si} \quad i \geq k \\ \displaystyle\sum_{j=k}^{N-1} f[j] + \sum_{j=0}^{i} f[j] & \text{si} \quad i \leq k-1 \end{cases}, \tag{6.19}$$

la distance CEMD(f,g) peut s'exprimer comme

$$\text{CEMD}(f,g) = \frac{1}{N} \min_{k\in\{0,\dots,N-1\}} \|\tilde{F}_k - \tilde{G}_k\|_1\ . \tag{6.20}$$

En effet, on peut vérifier une nouvelle fois que $\tilde{F}_k$ s'écrit en fonction de F

$$\forall i, k \in \{0,\dots,N-1\}\,,\ \tilde{F}_k[i] = \begin{cases} F[i] - F[k-1] & \text{si } i \geq k \\ F[i] + 1 - F[k-1] & \text{si } i \leq k-1 \end{cases},$$

toujours avec la convention $F[-1] = 0$. Ce qui nous donne l'équivalence $\forall k \in \{0,\dots,N-1\}$,

$$\begin{aligned} \|\tilde{F}_k - \tilde{F}_k\|_1 &= \sum_{i=0}^{N-1} |\tilde{F}_k[i] - \tilde{F}_k[i]| \\ &= \sum_{i=0}^{k-1} |F[i] - F[k-1] - G[i] + G[k-1]| + \sum_{i=k}^{N-1} |F[i] - F[k-1] - G[i] + G[k-1]| \\ &= \|F - G - F[k-1] + G[k-1]\|_1\ . \end{aligned}$$

Complexité de CEMD Dans le cas non circulaire et unidimensionnel, le calcul de l'EMD (tout comme les distances bin-à-bin L^1 et L^2) représente une complexité de $\mathcal{O}(N)$ (exprimé en nombre de multiplications). En exploitant la structure circulaire des histogrammes, le calcul du transport avec CEMD en

utilisant la formule (6.17) nécessite le calcul du médian μ de la liste de N valeurs $\{F[i]-G[i]\}_{0\leq i\leq N-1}$, soit une complexité totale de $\mathcal{O}(N)$.

> **Remarque 1 :**
> La formule alternative (6.20), utilisant les histogrammes cumulés depuis chaque cellule de quantification, est de complexité plus grande en raison de la minimisation qui requiert le calcul suplémentaire de $N-1$ normes : $\mathcal{O}(N^2)$.

6.2 Analyse de l'intérêt du transport pour la comparaison d'histogrammes globaux

Nous avons présenté dans la section précédente une nouvelle expression pour calculer efficacement le transport optimal entre deux histogrammes unidimensionnels circulaires. Dans cette thèse, nous nous intéressons à deux domaines dans lesquels des histogrammes circulaires apparaissent :

- la reconnaissance d'objets, domaine dans lequel les histogrammes utilisés sont *locaux* et construits à partir de peu d'échantillons ; l'étude de la distance CEMD dans ce cadre fait l'objet du chapitre 7 ;
- la recherche d'images par le contenu, application pour laquelle les histogrammes sont *globaux* et représentent une statistique sur la totalité de l'image d'intérêt, avec un nombre d'échantillons *a priori* plus important[1]. L'étude de l'intérêt de la distance CEMD pour cette application est justement l'objet de cette section 6.2.

La recherche d'images par le contenu (*Content Based Image Retrieval*) est très différente de la mise en correspondance d'images. Plutôt que de rechercher l'élément le plus proche dans une base, il s'agit de retrouver toutes les images appartenant à une même classe, ce qui requiert une mesure de similarité particulièrement robuste à la variabilité « intra-classe », c'est-à-dire aux perturbations subies par les différents éléments au sein d'une même classe, par opposition à la variabilité « inter-classe ». De nombreuses études ont montré la supériorité de la distance EMD en comparaison des distances bin-à-bin pour ce type d'application [RTG00, GDR00, RT01, Dvi02, LCL04, LZLM05, ZWG06, HGS08, PW09].

Nous allons dans un premier temps appliquer CEMD à deux types de représentations pour lesquelles l'utilisation du transport circulaire peut s'avérer utile : les histogrammes d'orientation du gradient pour la reconnaissance de caractères, et les histogrammes de teinte pour l'indexation d'images couleurs. Une analyse comparative de la robustesse des distances fondées sur le transport et sur les comparaisons bin-à-bin est ensuite réalisée pour différents types de perturbations. Enfin, des exemples sur des données réelles sont étudiés afin de conclure sur l'intérêt du transport pour la comparaison d'histogrammes globaux.

6.2.1 *Applications de* CEMD *pour la comparaison d'histogrammes globaux*

Le principe des expériences de recherche d'image (*image retrieval*) que nous présentons est le suivant : étant donnée une base de N images, divisée en plusieurs catégories (ou classes), on cherche pour chaque image requête de la base à retrouver toutes les images de sa classe. Chaque image est décrite par un histogramme global d'un attribut (orientation du gradient, teinte, etc.). Pour une distance donnée entre histogrammes (L^1 ou CEMD par exemple), et pour une image requête donnée, on peut ordonner les $(N-1)$ images restantes par ordre croissant de dissimilarité. Les performances des distances utilisées seront fonction de leur capacité à bien ordonner les images (images de la même classe en premier).

Pour illustrer ces performances, nous traçons des courbes de *performance moyenne*, qui sont utilisées traditionnellement en recherche d'images et légèrement différentes de celles présentées au chapitre 2 pour la mise en correspondance. Nous rappelons ci-dessous leur définition.

Courbes de performance moyenne Etant donnée une image requête et une distance entre histogrammes, on commence par ordonner les $(N-1)$ images restantes. Chaque image se voit attribuer un rang, le rang 1 correspondant à l'image requête et le rang N à l'image qui en est la plus éloignée. Supposons que l'on sélectionne toutes les images jusqu'au rang r. Si l'on reprend la terminologie empruntée au domaine de la classification (voir le tableau 1.1 dans le chapitre 1), *le taux de rappel* est alors défini comme la proportion d'images correctes ainsi retrouvées parmi l'ensemble des images de la même classe. Dit autrement, le taux de rappel est le nombre $\#\{vp(r)\}$ de vrais-positifs parmi les r images sélectionnées, divisé par la somme $\#\{vp(N)+fn(N)\}$ des nombres de vrais-positifs et de faux-négatifs. Le *taux de précision* désigne la proportion d'images correctes retrouvées parmi l'ensemble des images sélectionnées (à la fois les vrais-positifs et les faux-positifs, soit $\#\{vp(r)+fp(r)\}$). Les taux de rappel

[1] ce qui a un effet non négligeable sur les performances d'une distance bin-à-bin (voir la section 6.2.2).

et de précision sont donc :

$$\begin{cases} \text{taux de rappel}(r) = & \dfrac{\#\{\,\text{vp}(r)\,\}}{\#\{\,\text{vp}(N)+\text{fn}(N)\,\}}\,, \\ \text{taux de précision}(r) = & \dfrac{\#\{\text{vp}(r)\,\}}{\#\{\,\text{vp}(r)+\text{fp}(r)\,\}}\,. \end{cases}$$

Une courbe de performance est tracée en faisant varier r, le nombre d'images sélectionnées. Les courbes de performance *moyenne* sont ensuite obtenues en utilisant chaque image de la base comme image requête et en calculant les moyennes en fonction du paramètre r. Dans la suite, nous traçons deux types de courbes de performance moyenne : le taux de rappel moyen en fonction du nombre d'images sélectionnées (r), ainsi que le taux de précision moyen en fonction du taux de rappel.

Un premier exemple : histogrammes d'orientation D'autres applications, en dehors de celles utilisant les descripteurs SIFT [Low04], se basent sur des histogrammes d'orientation du gradient en tant que représentation globale d'une image. Par exemple, dans [CS02], les auteurs proposent d'utiliser ce type d'histogramme pour la reconnaissance de caractères. N'ayant pas une telle base à notre disposition, nous en avons créé une de petite taille avec le logiciel GIMP. Cette base est constituée d'imagettes en niveau de gris représentant les 10 premières lettres de l'alphabet, avec pour chacune des classes 10 occurrences dans des polices et des styles variés (gras, italique, *etc.*). La figure 6.4 montre quelques unes des ces imagettes.

La méthode de construction des descripteurs est inspirée des SIFTs (dont le principe est rappelé en annexe B). Les images sont légèrement lissées par convolution avec un noyau gaussien afin d'éviter le phénomène de crénelage (*aliasing*). La norme puis la phase du gradient sont calculées sur l'image, pour les pixels ayant un gradient de norme suffisamment élevée (supérieur à 1) pour être robuste au bruit (ou aux artefacts) de l'image. L'histogramme de l'orientation du gradient est ensuite construit empiriquement sur q bins (la valeur sera ultérieurement précisée), en pondérant le vote de chaque pixel par la norme du gradient. Cela permet d'être robuste au lissage effectué sur l'image, qui fait apparaître de nouvelles orientations. L'histogramme de l'image est ensuite normalisé, de telle manière que sa masse totale (norme L^1) soit égale à l'unité.

Afin de mesurer l'intérêt du transport circulaire, nous traçons les courbes de performance pour les distances CEMD, EMD (non circulaire) et L^1 en figure 6.5 pour diverses situations. On constate sur cet exemple simple l'intérêt de tirer parti de la circularité des histogrammes, la distance CEMD donnant systématiquement de meilleures performances moyennes en termes de recherche d'images que la distance EMD. Les figures 6.5(a) et 6.5(b) montrent les performances moyennes obtenues avec une quantification des histogrammes de $q = 360$ bins. Globalement, on observe que la distance CEMD donne de meilleures performances que la distance L^1. Si l'on applique une transformation affine aléatoire sur chacune des imagettes pour perturber les histogrammes globaux d'orientation (figures 6.5(e) et 6.5(f)), on observe que la distance CEMD est bien plus robuste que la distance L^1. Cependant, avec une quantification de seulement $q = 36$ bins et sans perturbation affine (figures 6.5(c) et 6.5(d)), l'écart de performances diminue significativement. Nous étudierons plus en détail en section 6.2.2 les raisons pour lesquelles on observe de telles variations de performances.

Un second exemple : histogrammes de teinte Pour les applications de recherche d'images par le contenu (*Content Based Image Retrieval*), il est courant d'utiliser – entre autres – un histogramme de la distribution des couleurs de l'image [HGS08, RTG00]. En nous inspirant de ce type d'application, nous présentons ici à nouveau une expérience d'indexation sur une petite base d'images couleurs en comparant cette fois leurs histogrammes de teinte. La teinte étant définie de manière circulaire, la distance CEMD est alors toute indiquée.

La base est présentée en figure 6.6, divisée suivant 14 classes de 9 objets. Les images ont été obtenues en photographiant sur un fond identique différents objets *selon la même pose* mais sous des éclairages

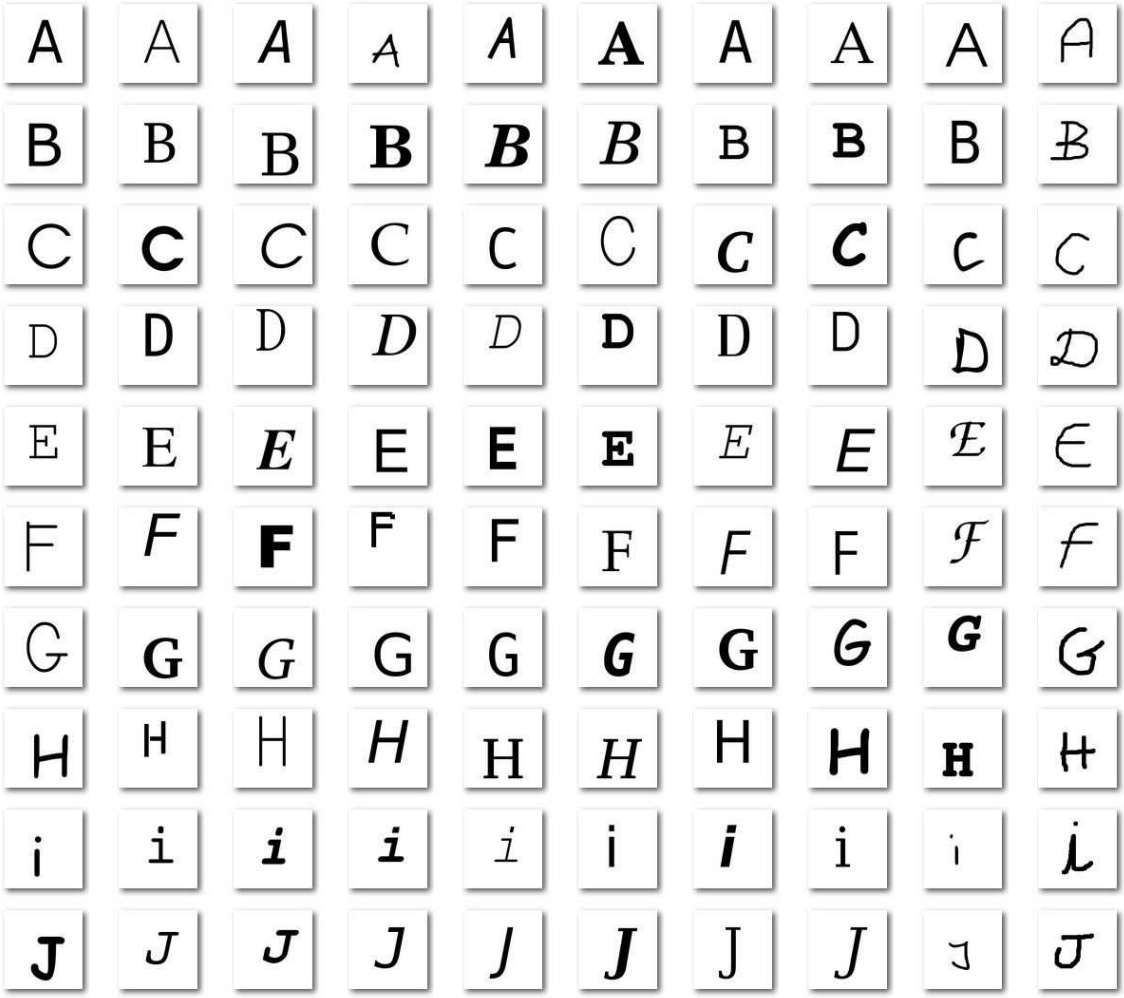


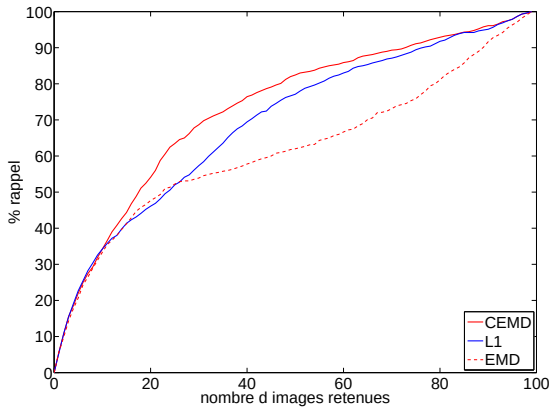
FIG. 6.4 – *Base de caractères de l'alphabet.*

et des conditions d'acquisition différents (avec et sans flash, différents réglages de l'appareil pour le temps de pose, la sensibilité et l'ouverture, *etc.*). Pour extraire leur histogramme de teinte, les images sont d'abord représentées dans l'espace colorimétrique HSV (*Hue Saturation Value*), dont on extrait la teinte (hue) et la saturation. La teinte et la saturation, notées respectivement H et S, sont définies de la manière suivante à partir de la représentation usuelle RVB (Rouge-Vert-Bleu) :

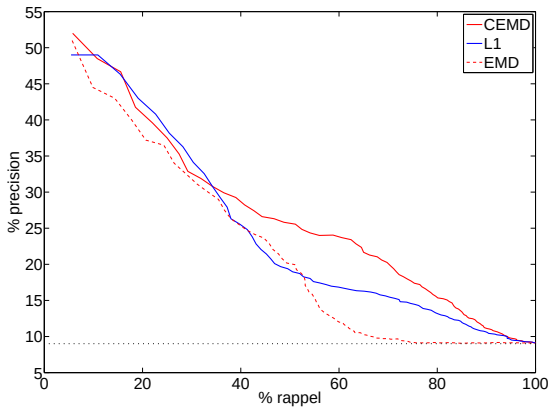
$$M = \max\{R,V,B\} \text{ et } m = \min\{R,V,B\}\,,\ M,m \in \{0,\dots,255\}$$

$$H = \begin{cases} 0, & \text{si } M = m \\ 60 \cdot \frac{V-B}{M-m}, & \text{si } M = R \\ 60 \cdot \frac{B-R}{M-m} + 120, & \text{si } M = V \\ 60 \cdot \frac{R-V}{M-m} + 240, & \text{si } M = B \end{cases} \in [0,360] \quad \text{et} \quad S = \begin{cases} 0, & \text{if } M = 0 \\ 1 - \frac{m}{M}, & \text{si } M \neq 0 \end{cases} \in [0,1]\,.$$

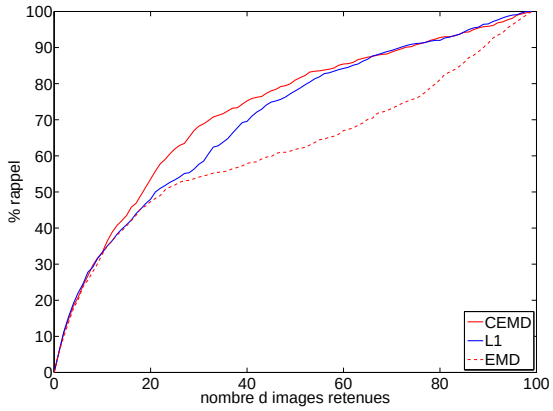
L'histogramme de chaque image est construit sur q bins à partir des pixels dont la saturation est plus grande qu'un certain seuil que l'on a fixé expérimentalement à $S_{\min} = 0.2$. Cela permet d'éliminer les pixels dont la couleur est très peu saturée (et dont l'apparence est proche des niveaux de gris), et qui peuvent considérablement perturber l'histogramme. La valeur de $S_{\min} = 0.2$ est celle donnant sur cette base des résultat optimaux à la fois pour L^1 et CEMD. L'histogramme circulaire de teinte est finalement normalisé à l'unité.



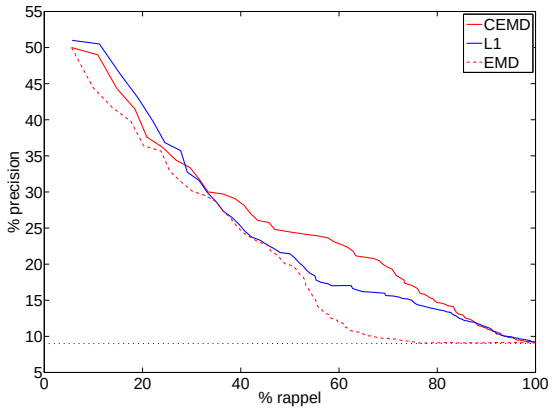
(a) Courbe de rappel moyen ($q = 360$)



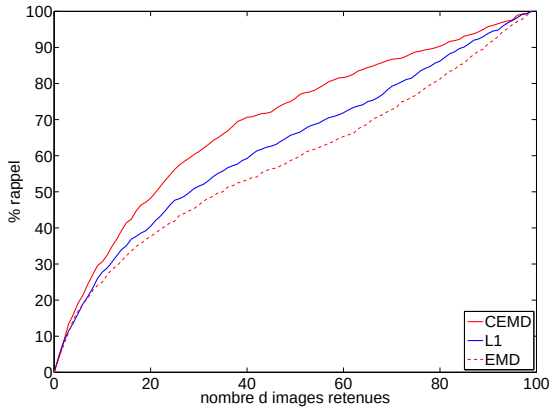
(b) Courbe de précision-rappel moyen ($q = 360$)



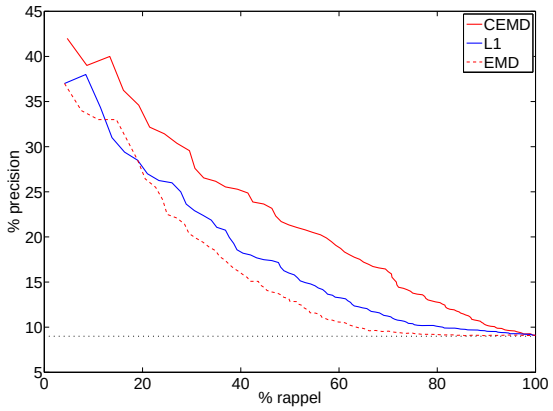
(c) Courbe de rappel moyen ($q = 36$)



(d) Courbe de précision-rappel moyen ($q = 36$)



(e) Courbe de rappel moyen ($q = 360$ + transformation affine)



(f) Courbe de précision-rappel moyen ($q = 360$ + transformation affine)

FIG. 6.5 – *Courbes de précision-rappel moyen de l'indexation d'une base de caractères. Des histogrammes d'orientation du gradient sont comparés avec différentes distances :* CEMD *en trait rouge continu, EMD en trait rouge interrompu, et* L^1 *en trait bleu continu. La première rangée de figures correspond à des histogrammes de* $q = 360$ *bins. La seconde rangée de figures correspond à des histogrammes de* $q = 36$ *bins. La troisième rangée de figures correspond à des histogrammes de* $q = 360$ *bins, calculés à partir d'imagettes perturbées par une transformation affine.*

FIG. 6.6 – *Base d'objets en prise de vue fixe selon différents réglages de la caméra.*

Les courbes de performance moyenne pour les distances CEMD et L^1 sont données en figure 6.7, pour différentes valeurs de quantification q. Une fois encore, on observe que la distance CEMD donne de meilleurs résultats que la distance bin-à-bin L^1, et ce d'autant plus que le nombre de bins est élévé.

Au travers de ces deux exemples de recherche d'images, nous avons illustré l'intérêt potentiel de la distance CEMD vis à vis d'une distance bin-à-bin comme L^1 pour des histogrammes circulaires. D'autres applications de l'EMD dans le cas non circulaire confirment un tel intérêt : comparaison de signatures [RTG00] pour les images couleurs et les textures, comparaison d'histogrammes de sac de mots [ZMLS07] (*bag of features*, voir le paragraphe 1.3.5) pour la classification d'images, comparaison d'histogrammes mélangeant position spatiale et bi-couleurs dominantes [HGS08] pour la recherche d'images par l'organisation spatiale de la couleur. Notons également une application récente de la distance CEMD pour la comparaison d'histogrammes circulaires en imagerie médicale.

Cependant, certains résultats obtenus dans la première expérience semblent suggérer certaines limitations du transport, phénomène que nous allons étudier dans la section suivante.

6.2.2 Analyse du transport pour des histogrammes de mélange de gaussiennes

Avec les deux exemples précédents, la distance CEMD semble être plus robuste que la distance L^1 à certaines classes de perturbations, comme la quantification des histogrammes, le changement d'éclairage pour la teinte de l'objet, ou les transformations géométriques dans le cas des caractères. Nous allons cependant montrer dans cette section qu'une distance définie comme un coût optimal de transport présente une limitation importante en comparaison d'une distance bin-à-bin. Cette limitation n'étant pas liée à la circularité des histogrammes, nous allons par la suite faire une analyse comparative de l'EMD avec L^1 exclusivement.

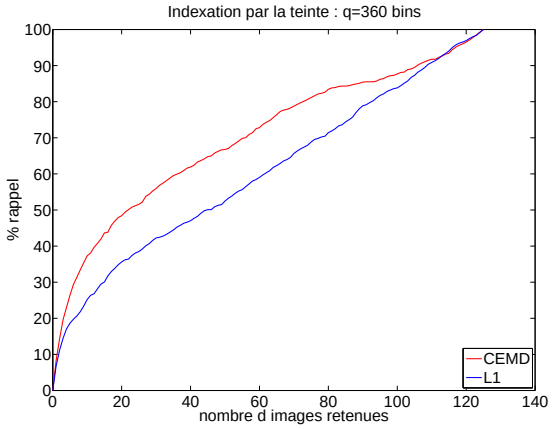
Dans les précédents exemples, les histogrammes globaux présentent des modes principaux correspondant aux teintes dominantes (histogrammes couleurs) ou aux contours dominants (histogrammes d'orientation). Ils peuvent en pratique être bien modélisés comme des mélanges de gaussiennes. En nous inspirant de ces expériences, nous allons analyser les performances de chaque distance en considérant le problème d'indexation suivant, avec une base divisée en deux classes d'objets, où chacune est représentée par des histogrammes de mélange de gaussiennes.

Principe On suppose que les objets à comparer sont divisés en deux classes, A et B. Comme dans les deux exemples précédents, chacun de ces objets est représenté par un histogramme. Cet histogramme est obtenu comme la distribution empirique de n réalisations d'une variable aléatoire de densité f_A, si l'objet appartient à la classe A, ou de densité f_B si l'objet appartient à la classe B, avec

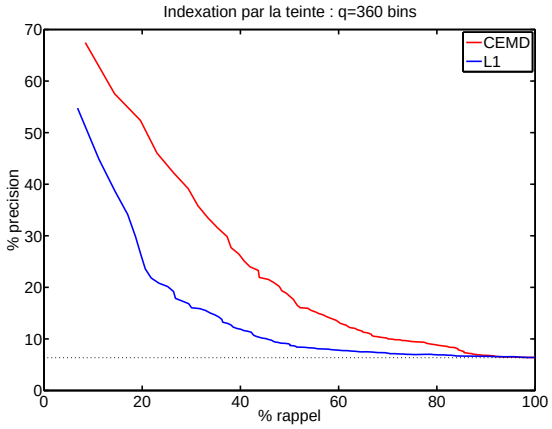
$$\begin{cases} f_A : x \mapsto p^A \times G_{\sigma_1^A}(x-\mu_1^A) + (1-p^A) \times G_{\sigma_2^A}(x-\mu_2^A) \\ f_B : x \mapsto p^B \times G_{\sigma_1^B}(x-\mu_1^B) + (1-p^B) \times G_{\sigma_2^B}(x-\mu_2^B) \end{cases}, \text{ où } G_\sigma(x-\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

avec μ et σ^2 représentant respectivement la moyenne et la variance d'une distribution normale $G_\sigma(x-\mu)$. Les deux distributions f_A et f_B sont illustrées en figure 6.8. Le terme p^A (respectivement p^B) correspond à la probabilité qu'une réalisation soit tirée suivant la loi normale de densité $G_{\sigma_1^A}(x-\mu_1^A)$ (respectivement $G_{\sigma_1^B}(x-\mu_1^B)$). De même, $1-p^A$ correspond à la probabilité qu'une réalisation soit tirée suivant la seconde loi normale, notée $G_{\sigma_2^A}(x-\mu_2^A)$.

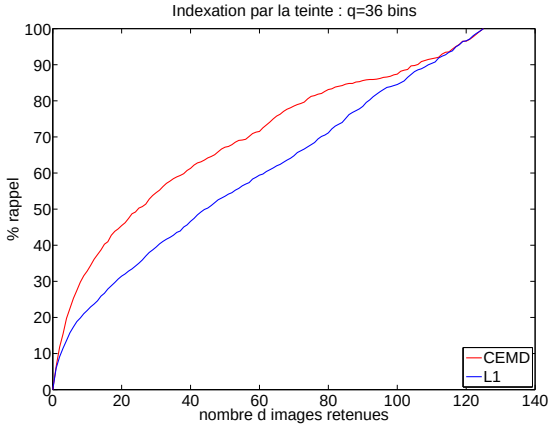
On suppose que l'on a N objets différents pour chaque classe, dont on construit les $2N$ histogrammes correspondants. Une indexation est ensuite réalisée en calculant pour chacun des histogrammes la distance avec les $2N-1$ autres histogrammes. Idéalement, les $N-1$ histogrammes les plus proches sont ceux correspondant aux objets appartenant à la même classe que l'objet requête. En comptabilisant le nombre d'objets corrects retrouvés parmi les $N-1$ plus proches, on peut donc déterminer quelle est la distance la plus robuste.



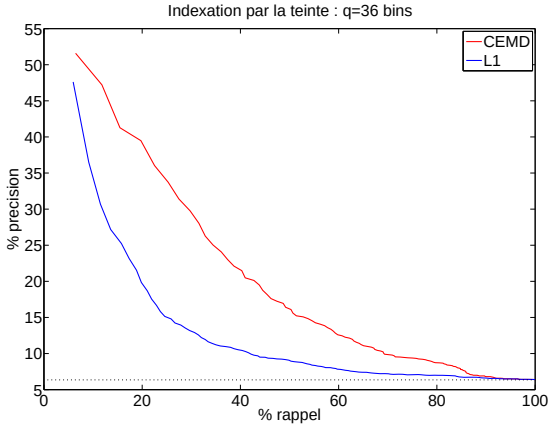
(a) Courbe de rappel moyen ($q = 360$)



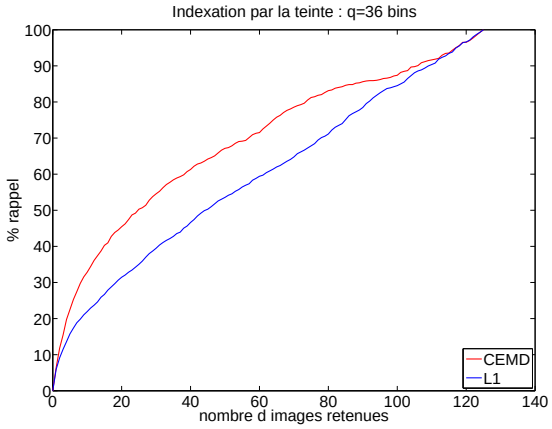
(b) Courbe de précision-rappel moyen ($q = 360$)



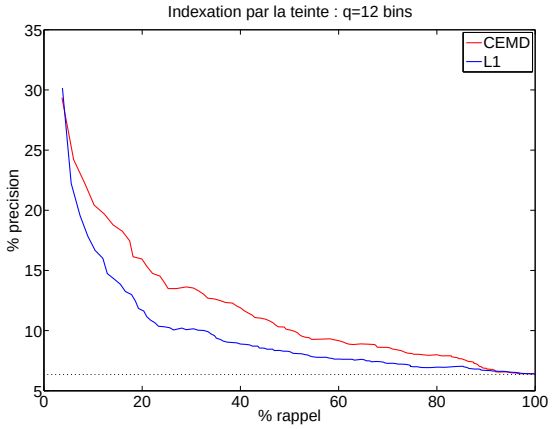
(c) Courbe de rappel moyen ($q = 36$)



(d) Courbe de précision-rappel moyen ($q = 36$)



(e) Courbe de rappel moyen ($q = 12$)



(f) Courbe de précision-rappel moyen ($q = 12$)

FIG. 6.7 – *Indexation d'une base d'objet en prise de vue fixe selon différents réglages de la caméra.* *Les courbes de précision-rappel moyen sont tracées selon différentes quantifications, avc un nombre de bins égal* $q = 360, 36,$ *et* 12. *Le trait rouge continu représente la distance* CEMD, *le trait rouge interrompu la distance EMD (transport sans tenir compte de la circularité), et le trait bleu continu la distance* L^1.

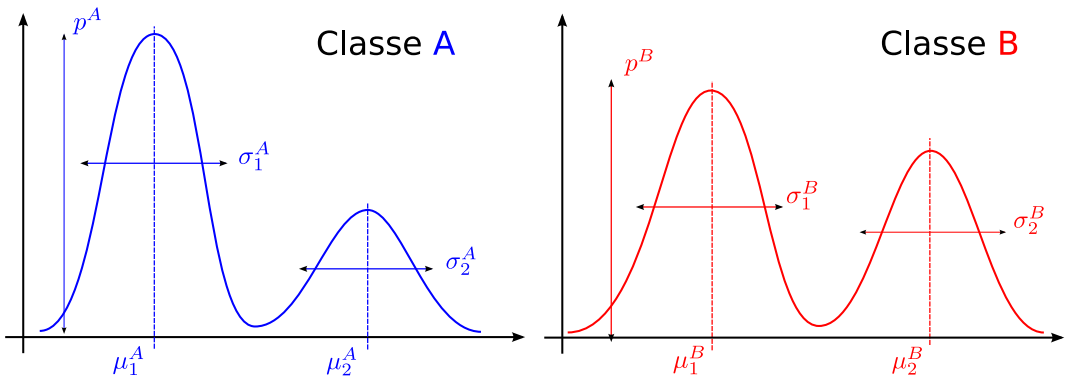


FIG. 6.8 – *Les deux classes sont définies comme un mélange de deux gaussiennes. En modélisant des perturbations par des changements de moyenne μ, de poids p et de variance σ^2, nous allons tester la robustesse de la distance de transport EMD et de la distance bin-à-bin L^1.*

Supposons que le nombre d'échantillons n tende vers l'infini, de sorte que les histogrammes tendent vers les distributions f_A ou f_B, selon la classe de l'objet. Dans ce cas limite, les distances EMD et L^1 entre ces deux distributions (appelées *distances inter-classes*) s'expriment ainsi :

$$D_{L^1}(f_A, f_B) = ||f_A - f_B||_1 = \int_{-\infty}^{\infty} \left| p^A\, G_{\sigma_1^A}(x-\mu_1^A) + (1-p^A)\, G_{\sigma_2^A}(x-\mu_2^A) - p^B\, G_{\sigma_1^B}(x-\mu_1^B) - (1-p^B)\, G_{\sigma_2^B}(x-\mu_2^B) \right| dx\ , \tag{6.21}$$

et

$$\mathrm{EMD}(f_A, f_B) = ||F_A - F_B||_1 = \frac{1}{2} \times \int_{-\infty}^{\infty} \left| p^A \,\mathrm{erf}\left(\frac{x-\mu_1^A}{\sqrt{2}\sigma_1^A}\right) + (1-p^A)\,\mathrm{erf}\left(\frac{x-\mu_2^A}{\sqrt{2}\sigma_2^A}\right) - p^B\,\mathrm{erf}\left(\frac{x-\mu_1^B}{\sqrt{2}\sigma_1^B}\right) - (1-p^B)\,\mathrm{erf}\left(\frac{x-\mu_2^B}{\sqrt{2}\sigma_2^B}\right) \right| dx, \tag{6.22}$$

où F_A et F_B sont les fonctions de répartition des classes A et B, et où la fonction d'erreur « erf » est définie comme :

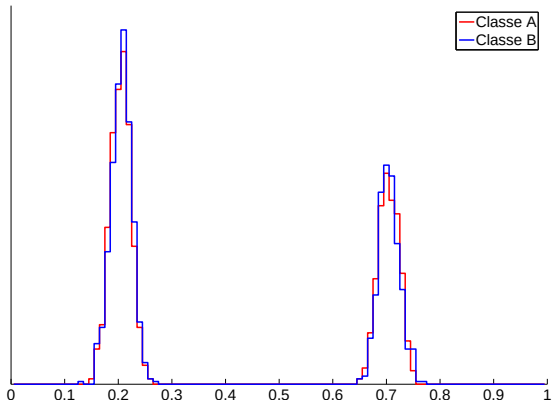
$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt = \frac{2}{\sqrt{\pi}} \int_{-\infty}^x e^{-t^2} dt - 1\ .$$

La distance entre deux histogrammes de la même classe est bien entendu nulle si les réalisations sont identiquement distribuées. Cependant, il existe en réalité des variations **intra-classe** (changement de la police du symbole et changement d'éclairage de l'objet dans les exemples précédents) qui se traduisent par des variations des caractéristiques principales de la distribution (moyenne, variance, *etc.*). Par exemple, il est connu que les distances bin-à-bin – contrairement à l'EMD – sont très peu robustes à de petites translations des histogrammes, ce qui peut être modélisé par des variations de la moyenne des couples de gaussiennes de chaque classe. Nous proposons de réaliser cette analyse à partir de simulations d'histogrammes discrets, en traçant des courbes de performance moyenne d'indexation de manière analogue aux exemples précédents. Différents types de perturbation seront testés en variant les différents paramètres intervenant dans ce type de problème : quantification q, nombre d'échantillons n, et paramètres des distributions de chaque classe (poids du premier mode p, de moyenne μ et variance σ^2).

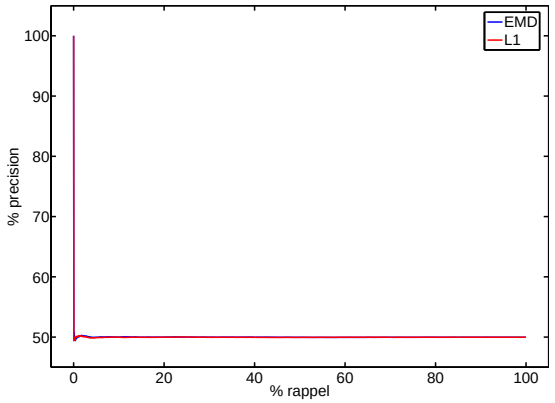
Protocole Dans les différentes expériences suivantes, sauf indications contraires, $N = 10^3$ histogrammes empiriques sont construits pour chacune des deux classes A et B, à partir de $n = 10^3$ échantillons dont les valeurs sont uniformément quantifiées sur $q = 10^2$ bins sur le domaine $[0, 1]$. Les histogrammes sont ensuite normalisés de telle sorte que leur norme L^1 soit égale à l'unité.

Afin d'illustrer chaque perturbation étudiée par la suite, nous traçons pour chaque expérience une courbe de précision-rappel moyen, ainsi qu'un exemple d'histogramme issu de chaque classe. Pour donner un ordre de grandeur de la robustesse de la distance pour la perturbation testée, nous donnerons également le taux de précision moyen sur les $N-1$ plus proches voisins (parmi $2N-1$). À titre d'exemple, un classifieur aléatoire donne une précision moyenne de 50% sur les $N-1$ plus proches voisins.

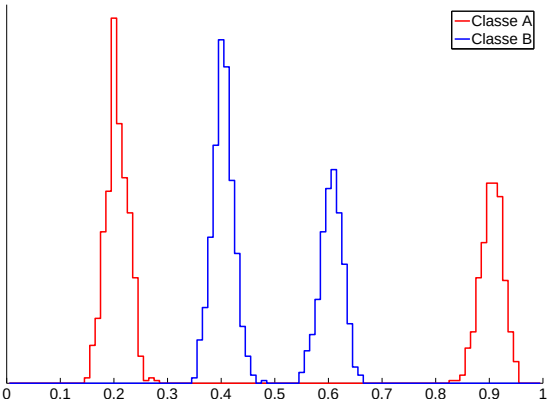
Pour vérifier la pertinence du nombre N de réalisations effectuées, on étudie deux cas particuliers : lorsque les deux classes sont identiques ($p^a = p^b = 0.6$, $\mu_1^a = \mu_1^b = 0.2$, $\mu_2^a = \mu_2^b = 0.7$, $\sigma_1^a = \sigma_1^b = \sigma_2^a = \sigma_2^b = 0.02$), et lorsque les deux classes sont parfaitement distinctes ($p^a = p^b = 0.6$, $\mu_1^a = 0.2$ et $\mu_1^b = 0.4$, $\mu_2^a = 0.9$ et $\mu_2^b = 0.6$, $\sigma_1^a = \sigma_1^b = \sigma_2^a = \sigma_2^b = 0.02$). Des exemples de réalisations d'histogrammes sont donnés en figure 6.9(a) et 6.9(c), pour chacun des cas. Dans le cas où les deux classes sont indistinctes, les courbes ROC obtenues en figure 6.9(b) correspondent avec une bonne approximation à la droite de 50% de précision moyenne (une chance sur deux d'attribuer la bonne classe), ce qui montre que le nombre de réalisations est suffisant ($N = 10^3$ histogrammes construits à partir de $n = 10^3$ échantillons). Lorsque les classes A et B sont totalement distinctes, on vérifie cette fois que les deux courbes de performance en figure 6.9(d) correspondent parfaitement à la droite idéale à 100% de précision moyenne pour les $N-1$ premiers histogrammes.



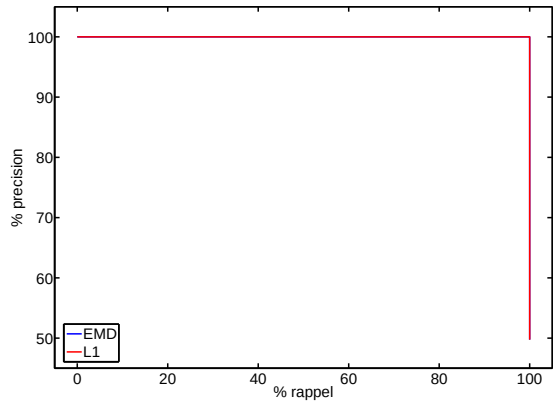
(a) Les deux classes ne sont pas distinctes



(b) Courbes précision-rappel moyen dans le cas de classes identiques



(c) Les deux classes sont totalement distinctes



(d) Courbes précision-rappel moyen dans le cas de classes parfaitement distinctes

FIG. 6.9 – *Illustration du procédé d'évaluation des distances EMD et L^1. Les figures de la première colonne montrent des exemples de réalisations d'histogrammes pour chaque classe. La seconde colonne illustre la courbe de ROC (taux de précision moyen en fonction du taux de rappel) obtenue en comparant $N = 10^3$ histogrammes générés aléatoirement suivant deux classes A et B. Dans le premier cas, les deux classes sont indistinctes (figure 6.9(a)) et la précision moyenne obtenue est alors effectivement de 50% environ (figure 6.9(b)). Dans le second cas, les deux classes sont parfaitement distinctes (figure 6.9(c)) et la précision moyenne est alors de 100% (figure 6.9(d)).*

Expérience 1 : effets de l'échantillonnage Dans un premier temps, nous allons vérifier qu'une distance fondée sur le transport est plus robuste qu'une distance bin-à-bin aux effets de l'échantillonnage : phénomène de quantification et nombre d'échantillons.

Nous fixons les paramètres de la manière suivante : $p^a = 0.6$ et $p^b = 0.8$, $\mu_1^a = \mu_1^b = 0.2$, $\mu_2^a = \mu_2^b = 0.7$, $\sigma_1^a = \sigma_1^b = \sigma_2^a = \sigma_2^b = 0.05$ ($N = 10^3$ et $n = 10^3$). Un exemple de réalisation pour chacune des classes est donné en figure 6.10(a) avec $q = 10^2$ bins, et en figure 6.10(c) avec $q = 10^3$ bins. Avec $q = 10^2$ bins, on obtient une courbe ROC parfaite avec les deux distances (figure 6.10(b)) : la différence de poids de chacune des gaussiennes permet de les distinguer. Lorsque l'on augmente le nombre de bins à $q = 10^3$, les performances de la distance EMD restent inchangées (figure 6.10(d)), tandis que les performances de la distance L^1 sont affectées : sur les $N - 1$ premiers histogrammes, seuls 93.5% sont en moyenne corrects.

> **Remarque 1 :**
> Si l'on fixe cette fois la quantification à $q = 10^2$ bins et que l'on diminue le nombre d'échantillons à $n = 4.10^2$ (figure 6.10(e)), on observe un résultat similaire (figure 6.10(f)) : les performances de la distance EMD restent inchangées, alors que la précision moyenne de la distance L^1 est de 95.4% sur les $N - 1$ plus proches voisins.

Nous allons dans les deux expériences suivantes analyser la robustesse de la distance bin-à-bin L^1 et de la distance de transport EMD aux perturbations intra-classe.

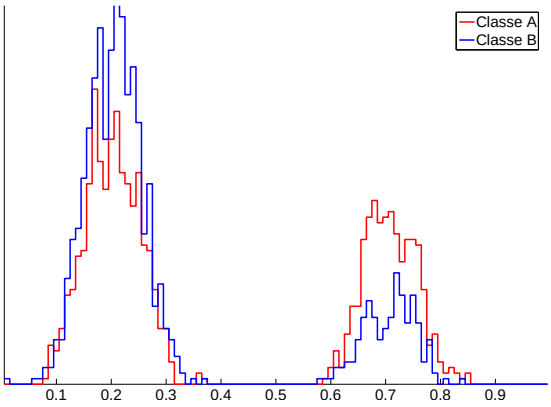
Expérience 2 : variabilité intra-classe avec perturbation sur les moyennes Nous nous intéressons tout d'abord au phénomène de translations des modes principaux.

Nous fixons maintenant les paramètres de la manière suivante : $p^a = 0.6$ et $p^b = 0.8$, $\mu_2^a = \mu_2^b = 0.7$, $\sigma_1^a = \sigma_1^b = \sigma_2^a = \sigma_2^b = 0.05$ (toujours avec $N = 10^3$, $n = 10^3$ et $q = 10^2$). Cette fois, les moyennes du mode principal des deux classes (μ_1^a et μ_1^b) sont des variables aléatoires *iid*, tirées suivant la loi uniforme, centrée en 0.2 de largeur ϵ_μ. Cela permet de modéliser une variabilité intra-classe qui se traduit par des petites translations du mode principal avec $\mu_1^a, \mu_1^b \in [0.2 - \frac{\epsilon_\mu}{2}, 0.2 + \frac{\epsilon_\mu}{2}]$.

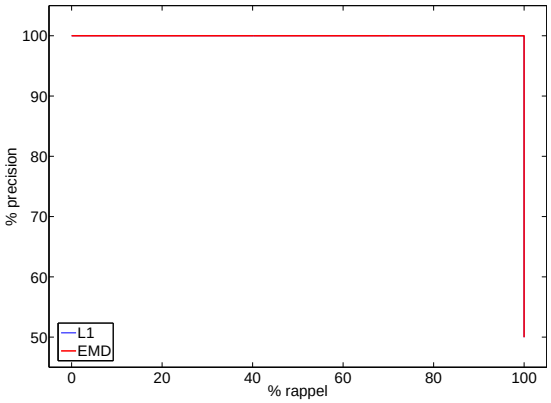
Afin de visualiser ces translations intra-classe, nous avons représenté plusieurs histogrammes par classe en dégradé de couleur sur la figure 6.11(c) lorsque $\epsilon_\mu = 0.05$. Les courbes de précision-rappel moyen sont tracées en figure 6.11(d). Comme l'on pouvait s'y attendre, la distance de transport est invariante à ce type de perturbation tandis que les performances de la distance L^1 sont fortement affectées par ce phénomène, avec 83.9% de précision moyenne sur les $N - 1$ plus proches voisins. À titre de comparaison, lorsque les moyennes des modes principaux des deux classes sont fixées (soit $\epsilon_\mu = 0$, voir la figure 6.11(a) avec plusieurs histogrammes représentés par classe), la distance L^1 obtient 100% de précision à l'instar de l'EMD (figure 6.11(b)). Lorsque l'on diminue les variances des modes ($\sigma_1^a = \sigma_1^b = \sigma_2^a = \sigma_2^b = 0.02$), tout en gardant $\epsilon_\mu = 0.05$, les performances se dégradent d'autant plus pour la distance L^1 (61.1% de précision moyenne sur les $N - 1$ plus proches voisins), la distance EMD étant toujours invariante à ces perturbations.

Expérience 3 : variabilité intra-classe avec perturbation sur les poids Nous allons maintenant mettre en évidence une autre type de perturbation à laquelle l'EMD se montre cette fois très sensible. En effet, en examinant les différents cas de figure où l'EMD montrait des performances moindres que la distance L^1, nous avons pu constater que cela se produisait lorsque les poids relatifs des modes principaux variaient au sein de la même classe. Afin de vérifier cette hypothèse, nous allons introduire dans l'expérience suivante une perturbation sur les termes de pondération p^a et p^b dans chacune des classes, de manière analogue aux translations intra-classe opérées à l'expérience précédente.

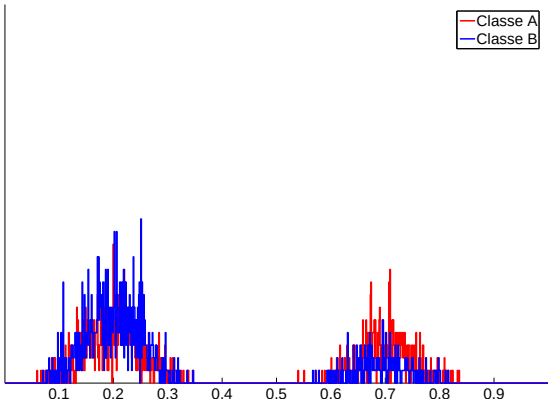
Les paramètres sont les suivants : $\mu_1^a = 0.2$ et $\mu_1^b = 0.25$, $\mu_2^a = 0.7$ et $\mu_2^b = 0.75$, $\sigma_1^a = \sigma_1^b = \sigma_2^a = \sigma_2^b = 0.02$ (toujours avec $N = 10^3$, $n = 10^3$ et $q = 10^2$). Cette fois, les poids des modes principaux des deux classes (p^a et p^b) sont des variables aléatoires *iid*, tirées suivant la loi uniforme de largeur ϵ_p, centrée en 0.6 pour la classe A, et centrée en 0.8 pour la classe B. Cela permet de modéliser une variabilité intra-classe qui se traduit par des variations du poids du mode principal avec $p^a \in [0.6 - \frac{\epsilon_p}{2}, 0.6 + \frac{\epsilon_p}{2}]$ et $p^b \in$



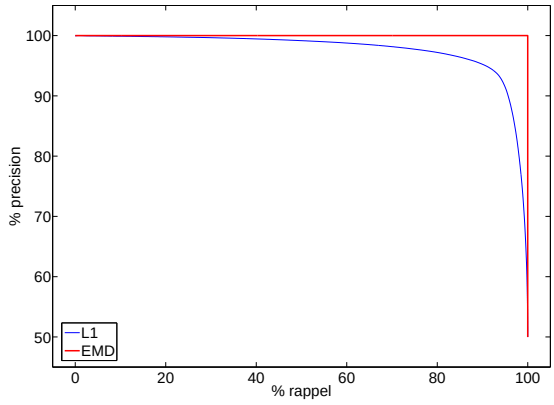
(a) $q = 10^2$ et $n = 10^3$



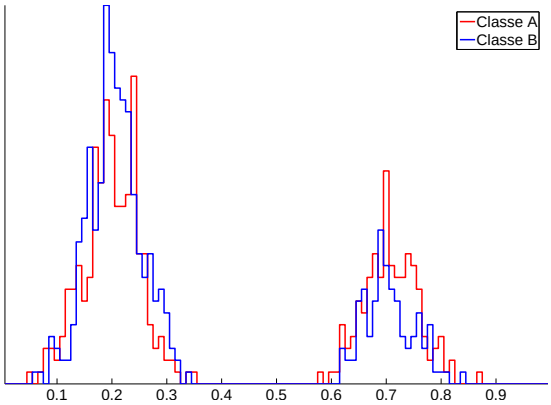
(b) Courbes de précision-rappel moyen



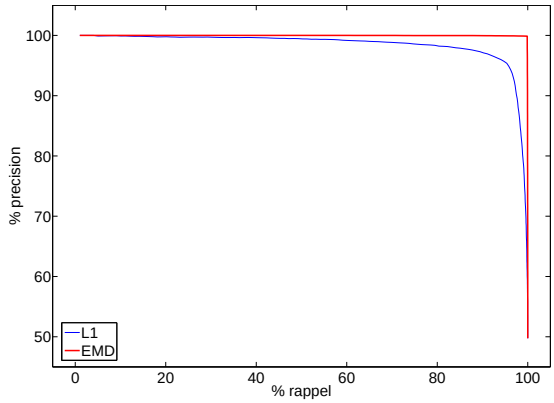
(c) $q = 10^3$ et $n = 10^3$



(d) Courbes de précision-rappel moyen



(e) $q = 10^3$ et $n = 4.10^2$

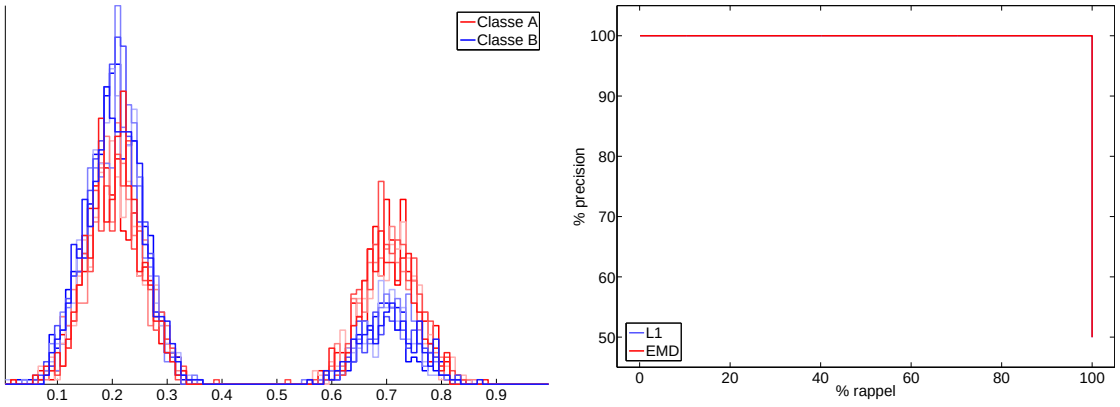


(f) Courbes de précision-rappel moyen

FIG. 6.10 – *Analyse de la robustesse des distances EMD et L^1 aux effets de l'échantillonnage (nombre de bins q et d'échantillons n).*

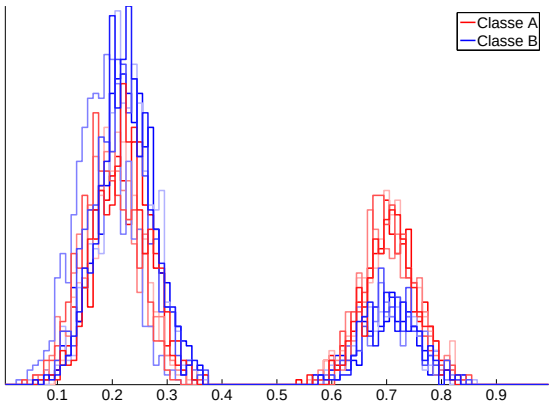
$[0.8 - \frac{\epsilon_p}{2}, 0.8 + \frac{\epsilon_p}{2}]$. Les gaussiennes étant en des positions (moyennes) fixes – différentes selon la classe – et les poids des modes principaux étant de domaines différents selon la classe, les classes devraient *a priori* être facilement distinguées. Pour tester cela, nous traçons les courbes de performance pour différentes valeurs de ϵ_p : $\epsilon_p = 0.1$ en figure 6.12(b), $\epsilon_p = 0.2$ en figure 6.12(d), et $\epsilon_p = 0.4$ en figure 6.12(f). Pour chacune de ces valeurs de ϵ_p sont respectivement représentés plusieurs histogrammes par classe dans les figures 6.12(a), 6.12(c) et 6.12(e). Pour $\epsilon_p = 0$ (poids fixés), on obtient pour chacune des distances une courbe ROC parfaite (courbes avec 100% de précision, non représentées ici).

Tout d'abord, on constate que la distance L^1 est effectivement très robuste à ce type de perturbation,

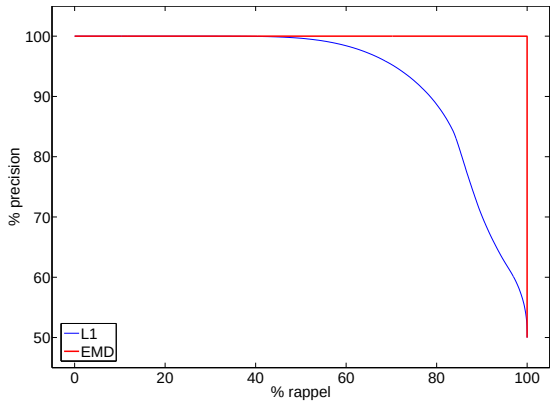


(a) $\epsilon_\mu = 0$ et $\sigma_1^A = \sigma_1^B = 0.05$

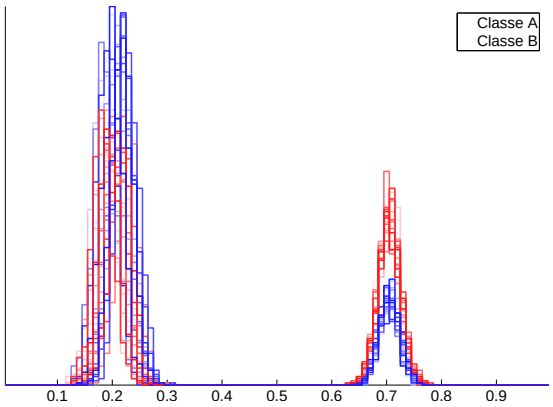
(b) Courbes de précision-rappel moyen



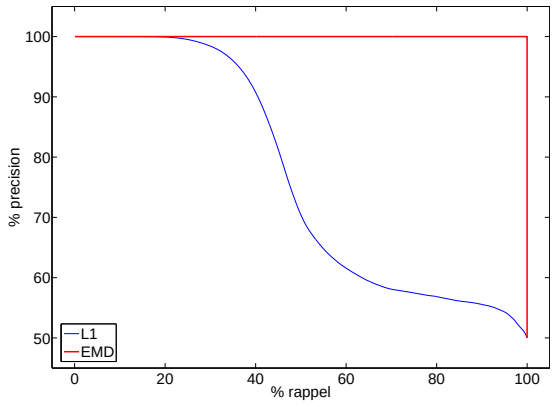
(c) $\epsilon_\mu = 0.05$ et $\sigma_1^A = \sigma_1^B = 0.05$



(d) Courbes de précision-rappel moyen



(e) $\epsilon_\mu = 0.05$ et $\sigma_1^A = \sigma_1^B = 0.02$



(f) Courbes de précision-rappel moyen

FIG. 6.11 – *Analyse des distances selon la variabilité intra-classe de la position du mode principal.*

puisque l'on obtient 100% de précision, quelle que soit la plage de variation intra-classe ϵ_p. La raison pour laquelle la distance bin-à-bin L^1 est robuste à un tel phénomène est évidente : puisque seuls les bins de même indice sont comparés, il suffit que les modes principaux des histogrammes soient suffisamment distincts pour des classes différentes pour que L^1 donne de bons résultats. Lorsque cela n'est pas vérifié, la distance L^1 n'est plus aussi discriminante, comme l'illustrent les courbes de la figure 6.12(h) correspondant aux cas où les modes des deux classes coïncident.

Par contre, on observe que les performances de la distance EMD dépendent fortement de la valeur de ϵ_p. La figure 6.12(b) correspond au cas limite pour lequel l'EMD permet de distinguer parfaitement les deux classes ($\epsilon_p \leq 0.1$). Dès que la plage de variabilité intra-classe ϵ_p des variables aléatoires p^a et p^b

excède cette valeur, on observe que l'EMD commence à confondre les histogrammes des deux classes. On obtient ainsi 94.042% de précision moyenne sur les $N-1$ plus proches voisins lorsque $\epsilon_p = 0.2$ (figure 6.12(d)) et 70.2% lorsque $\epsilon_p = 0.4$ (figure 6.12(f)).

Ces expériences confirment donc la limitation de l'EMD dans le cas d'une variabilité intra-classe des poids relatifs aux différents modes principaux. Afin d'en comprendre la raison, et de définir la valeur critique de ϵ_p pour laquelle la confusion entre les classes apparaît, nous allons étudier le coût du transport correspondant à ce phénomène. Toutefois, pour en simplifier l'analyse, nous nous intéressons au cas particulier où les variances des distributions normales tendent vers 0. Dans ce cas, les pseudo distributions de probabilité des deux classes s'écrivent :

$$\begin{cases} f_A : x \mapsto p^A\,\delta(x-\mu_1^A) + (1-p^A)\,\delta(x-\mu_2^A) \\ f_B : x \mapsto p^B\,\delta(x-\mu_1^B) + (1-p^B)\,\delta(x-\mu_2^B) \end{cases},$$

où $\delta(.)$ désigne la distribution de Dirac.

La distance L^1 s'écrit

$$D_{L^1}(f_A, f_B) = ||f_A - f_B||_1 = 2 + \Big(|p^a - p^b| - 1\Big)\,\delta_{\mu_1^A,\mu_1^B} + \Big(|p^a - p^b| - 1\Big)\,\delta_{\mu_2^A,\mu_2^B}\ ,$$

où $\delta\{.,.\}$ est le symbole de Kronecker. La distance de transport entre ces deux fonctions f_A et f_B est

$$\begin{aligned} \text{EMD}(f_A, f_B) = ||F_A - F_B||_1 = \quad & |\mu_1^A - \mu_1^B| \cdot \begin{cases} p^a & \text{si } \mu_1^A < \mu_1^B \\ p^b & \text{si } \mu_1^A > \mu_1^B \\ 0 & \text{si } \mu_1^A = \mu_1^B \end{cases} \\ & + |\mu_2^A - \mu_2^B| \cdot \begin{cases} 1-p^a & \text{si } \mu_2^A < \mu_2^B \\ 1-p^b & \text{si } \mu_2^A > \mu_2^B \\ 0 & \text{si } \mu_2^A = \mu_2^B \end{cases} \\ & + |p^a - p^b| \cdot \big(\min\{\mu_2^A, \mu_2^B\} - \max\{\mu_1^A, \mu_1^B\}\big)\ . \end{aligned}$$

Afin de simplifier encore ces expressions, on note $\Delta_p = |p^A - p^B|$ la différence de poids des premiers modes de chacune des classes, $L = \min\{\mu_2^A, \mu_2^B\} - \max\{\mu_1^A, \mu_1^B\}$ la plus petite distance séparant les deux modes, et $\Delta_\mu = |\mu_1^A - \mu_1^B| = |\mu_2^A - \mu_2^B|$ la différence entre les positions des modes principaux respectifs des deux classes, de telle sorte que si $\Delta_p = 0$, f_B est la translation de $\pm\Delta_\mu$ de f_A. Pour une meilleure compréhension des notations, voir la figure 6.13. Les deux distances s'expriment alors :

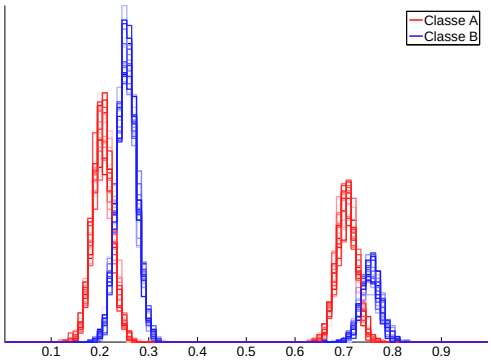
$$\begin{cases} D_{L^1}(f_A, f_B) &= 2\left(1 - \delta_{0,\Delta_\mu}(1-\Delta_p)\right) \\ \text{EMD}(f_A, f_B) &= \Delta_\mu + \Delta_p \cdot L \end{cases}.$$

Le coût de transport entre les distributions f_A et f_B est donc fonction de la translation Δ_μ entre les deux histogrammes, mais également de la différence de poids inter-classe Δ_p qui est transportée sur une distance au sol de L.

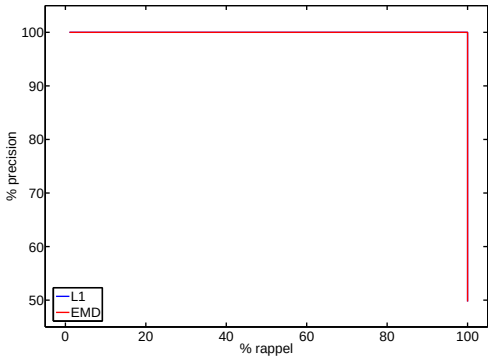
Considérons maintenant la distribution $f_A{}'$, ayant les mêmes caractéristiques que la distribution f_A (les deux masses sont aux mêmes positions μ_1^A et μ_2^A) à la différence près que les poids sont différents : $p^{A'} = p^A + \epsilon_p$. On définit la longueur $L' = |\mu_2^A - \mu_1^A| = |\mu_2^B - \mu_1^B| = L + \Delta_\mu$ correspondant à la distance entre les deux modes principaux au sein d'une même classe.

$$\begin{cases} D_{L^1}(f_A, f_A{}') &= 2\epsilon_p \\ \text{EMD}(f_A, f_A{}') &= \epsilon_p \cdot L' = \epsilon_p \cdot (L + \Delta_\mu) \end{cases}$$

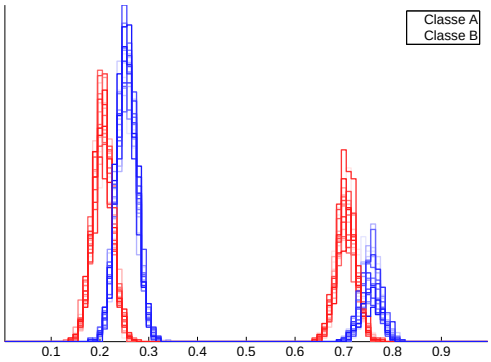
Cette fois, le coût de transport entre les deux histogrammes de la même classe A et A' correspond au transport de la différence de poids intra-classe ϵ_p sur une distance au sol de L'. Cela signifie que, du fait de la définition du transport, **la distance intra-classe dépend de la position relative des deux modes principaux**.



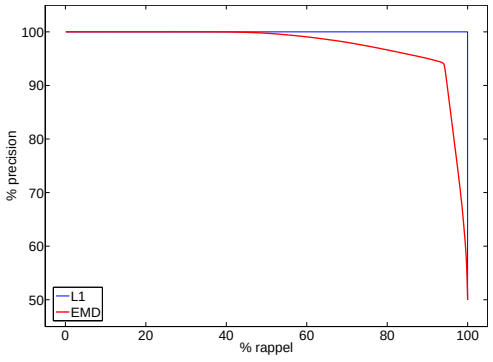
(a) $\epsilon_p = 0.1$ avec $\mu_1^B - \mu_1^A = \mu_2^B - \mu_2^A = 0.05$



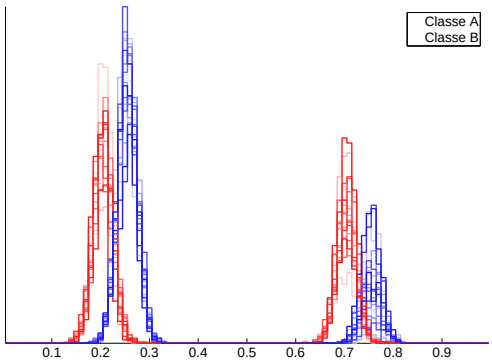
(b) Courbes de précision-rappel moyen



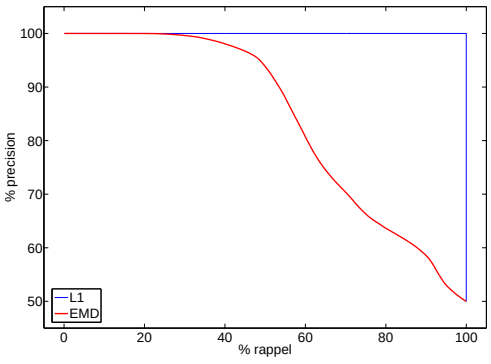
(c) $\epsilon_p = 0.2$ avec $\mu_1^B - \mu_1^A = \mu_2^B - \mu_2^A = 0.05$



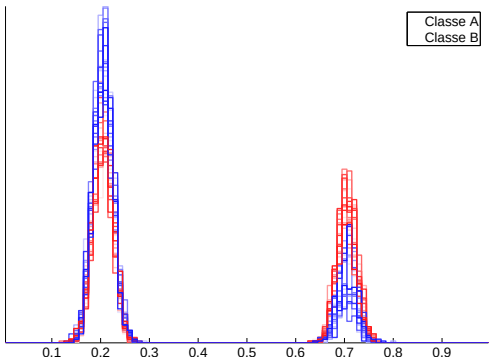
(d) Courbes de précision-rappel moyen



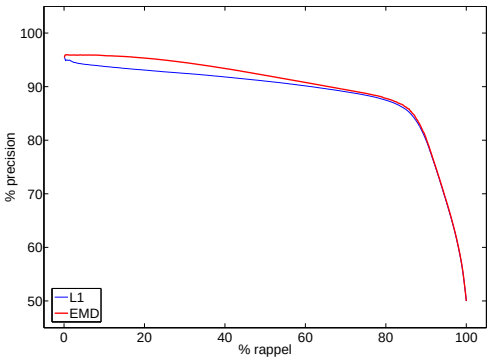
(e) $\epsilon_p = 0.4$ avec $\mu_1^B - \mu_1^A = \mu_2^B - \mu_2^A = 0.05$



(f) Courbes de précision-rappel moyen



(g) $\epsilon_p = 0.2$ avec $\mu_1^A = \mu_1^B$, $\mu_2^A = \mu_2^B$



(h) Courbes de précision-rappel moyen

FIG. 6.12 – *Analyse des distances selon la variabilité intra-classe du poids du mode principal.*

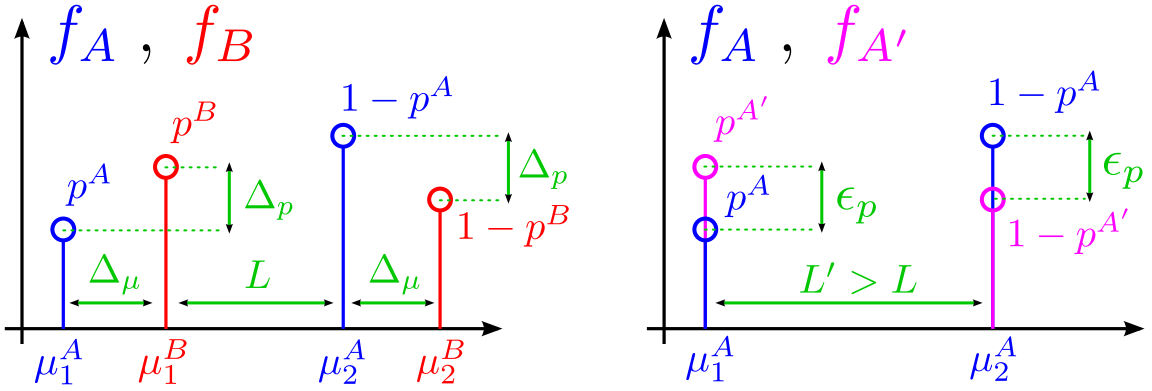


FIG. 6.13 – *En faisant tendre les variances vers zéro, les deux classes sont définies comme un mélange de deux distributions de Dirac.*

Avec ces expressions, on peut aisément retrouver les résultats obtenus avec l'expérience sur la variabilité intra-classe de la pondération. Dans le cas de L^1, on obtient sachant que $\epsilon_p < 1$:

$$D_{L^1}(f_A, f_A{}') \geq D_{L^1}(f_A, f_B) \Leftrightarrow \epsilon_p \geq 1 + \delta_{0,\Delta_\mu}(1-\Delta_p) = \begin{cases} \Delta_p & \text{si } \Delta_\mu = 0 \\ 1 & \text{si } \Delta_\mu \neq 0 \end{cases} .$$

On retrouve ainsi le résultat trivial suivant : lorsque les positions des modes principaux des deux classes sont distinctes ($\Delta_\mu \neq 0$), il est impossible de confondre les deux classes avec la distance L^1 (figures 6.12(b), 6.12(d), et 6.12(f)) ; par contre, lorsque les modes sont confondus ($\Delta_\mu = 0$), il suffit que la différence de poids intra-classe ϵ_p soit plus grande que la différence de poids inter-classe Δ_p pour qu'il y ait confusion entre les deux classes (figure 6.12(h)).

Dans le cas de la distance EMD, le résultat dépend cette fois des positions relatives des modes principaux :

$$\text{EMD}(f_A, f_A{}') \geq \text{EMD}(f_A, f_B) \Leftrightarrow \epsilon_p \geq \Delta_p \frac{L}{L'} + \frac{\Delta_\mu}{L'} = \begin{cases} \Delta_p & \text{si } \Delta_\mu = 0 \\ \dfrac{\Delta_p L + \Delta_\mu}{L + \Delta_\mu} \geq \Delta_p & \text{si } \Delta_\mu \neq 0 \end{cases} .$$

Examinons tout d'abord le cas le plus simple où les modes des deux classes sont confondus ($\Delta_\mu = 0$) : il suffit alors que la différence de poids intra-classe ϵ_p soit plus grande que la différence de poids inter-classe Δ_p à l'instar de la distance L^1, ce qui corrobore le résultat des courbes de la figure 6.12(h). Par contre, dans le cas général où les positions des modes principaux des deux classes sont distinctes ($\Delta_\mu \neq 0$), la condition sur ϵ_p pour que la distance intra-classe $\text{EMD}(f_A, f_A{}')$ soit plus grande que la distance inter-classe $\text{EMD}(f_A, f_B)$ dépend à la fois de la différence de masse inter-classe Δ_p mais également de la position relative des modes (les distances L et Δ_μ). Cependant, cette inégalité montre que la sensibilité de la distance EMD au phénomène de variabilité du poids dépend surtout du rapport entre Δ_μ / L : si la position relative des modes est suffisament grande ($\Delta_\mu \gg L$), alors la valeur critique de ϵ_p tend vers 1 et la distance EMD devient robuste à ce phénomène ; au contraire, lorsque $\Delta_\mu \ll L$, cette valeur critique tend vers Δ_p, valeur pour laquelle il commence à y avoir confusion entre les classes.

Les expériences que nous avons réalisées sur les mélanges de gaussiennes corroborent ces résultats (existence d'un seuil critique et sa dépendance en fonction de Δ_p, Δ_μ et L). Par exemple, nous avons vu en figure 6.12(b) (avec $\Delta_\mu = 0.05$) qu'il existait un valeur maximale de ϵ_p égale à 0.1 pour laquelle l'EMD est robuste aux variations de poids intra-classe. Si ϵ_p dépasse cette valeur critique, il y a confusion entre les classes. Bien que nous n'ayons pas établi l'expression de cette valeur critique pour le mélange de gaussienne, nous pouvons vérifier expérimentalement sa dépendance en fonction de Δ_μ : en augmentant la distance Δ_μ (par exemple 0.1 au lieu de 0.05), la valeur critique de ϵ_p augmente également ($\epsilon_p \geq 0.2$ au lieu de 0.1).

En conclusion, nous avons étudié dans cette section la robustesse de la distance bin-à-bin L^1 et de la distance de transport EMD en fonction de l'échantillonnage et pour deux types de perturbation intra-classe : les translations et les changement de poids, illustrés sur la figure 6.14. Nous avons en particulier montré que, dans le cas du mélange de gaussiennes, comme dans le cas limite du mélange de distributions de Dirac, le coût de transport dépend à la fois de la distance au sol entre les différents modes principaux et de la différence de poids des modes inter et intra-classe. Du fait de cette sensibilité du transport à la variabilité du poids des modes principaux, **les performances de la distance CEMD peuvent donc être altérées** lorsque ce type de perturbation intervient, voire même être en deçà d'une distance bin-à-bin – nonobstant la sensibilité de celle-ci aux perturbations liées à la quantification ou aux translations. Par conséquent, suivant le type de perturbation à laquelle on souhaite être invariant, l'une ou l'autre des distances (bin-à-bin ou transport) doit être *a priori* privilégiée. Il reste cependant à déterminer dans quel cadre ces perturbations se produisent en pratique, ce que nous étudions dans la section suivante.

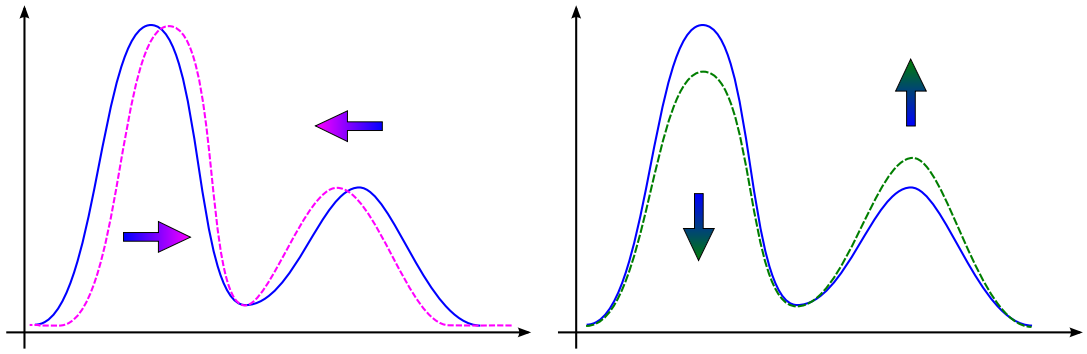


FIG. 6.14 – *Illustration de deux classes de transformations sur les histogrammes : perturbations intra-classe sur la position des modes (*à gauche*) et sur leur poids (*à droite*).*

6.2.3 Expériences sur la robustesse du transport aux perturbations intra-classe

Dans la section 6.2.1, nous avons illustré l'intérêt de la distance de transport circulaire CEMD en comparaison de la distance L^1. Les deux expériences qui ont été réalisées font intervenir des perturbations (liées à l'échantillonnage et au phénomène de translation des modes) pour lesquelles nous avons montré qu'une distance de transport est plus robuste qu'une distance bin-à-bin.

Nous allons maintenant montrer des exemples d'indexation d'images où le phénomène du changement de pondération est prépondérant. Nous avons constaté qu'il intervient en pratique dans deux cas de figures : dans le cas d'un changement de pose et d'un changement de la balance des blancs.

Expérience avec changement de balance des blancs La perception de la teinte d'un objet dépend du spectre de la source lumineuse qui l'éclaire. Par exemple, il est très difficile de distinguer certaines couleurs avec un éclairage public classique de couleur orangée (lampe à décharge au sodium). Pour obtenir des couleurs plus naturelles, les appareils photographiques numériques permettent de prendre en compte le type de source lumineuse par une correction dite de « balance des blancs ». On parle alors de « température de couleur » pour désigner le type d'éclairage utilisé (théorie du corps noir). Typiquement, on considère que la température d'une lumière incandescente (en degré Kelvin) est de $3200°K$, de $5200°K$ pour un ciel ensoleillé, ou encore de $8000°K$ pour un ciel nuageux. Ainsi, selon la température de référence sélectionnée par l'utilisateur, les teintes de l'image obtenue sont très différentes.

Pour étudier la robustesse des distances CEMD et L^1 à ce type de correction, nous avons utilisé 22 photographies en format brut (RAW) que nous avons ensuite déclinées selon 10 balances de blancs différentes, en utilisant des températures de référence entre $4400°K$ et $6200°K$ avec un logiciel d'édition CANON. Les images obtenues sont données en figure 6.15. Pour chaque image, un histogramme de teinte

est ensuite construit, de manière analogue à l'expérience de la section 6.2.1, avec une quantification sur 360 bins.

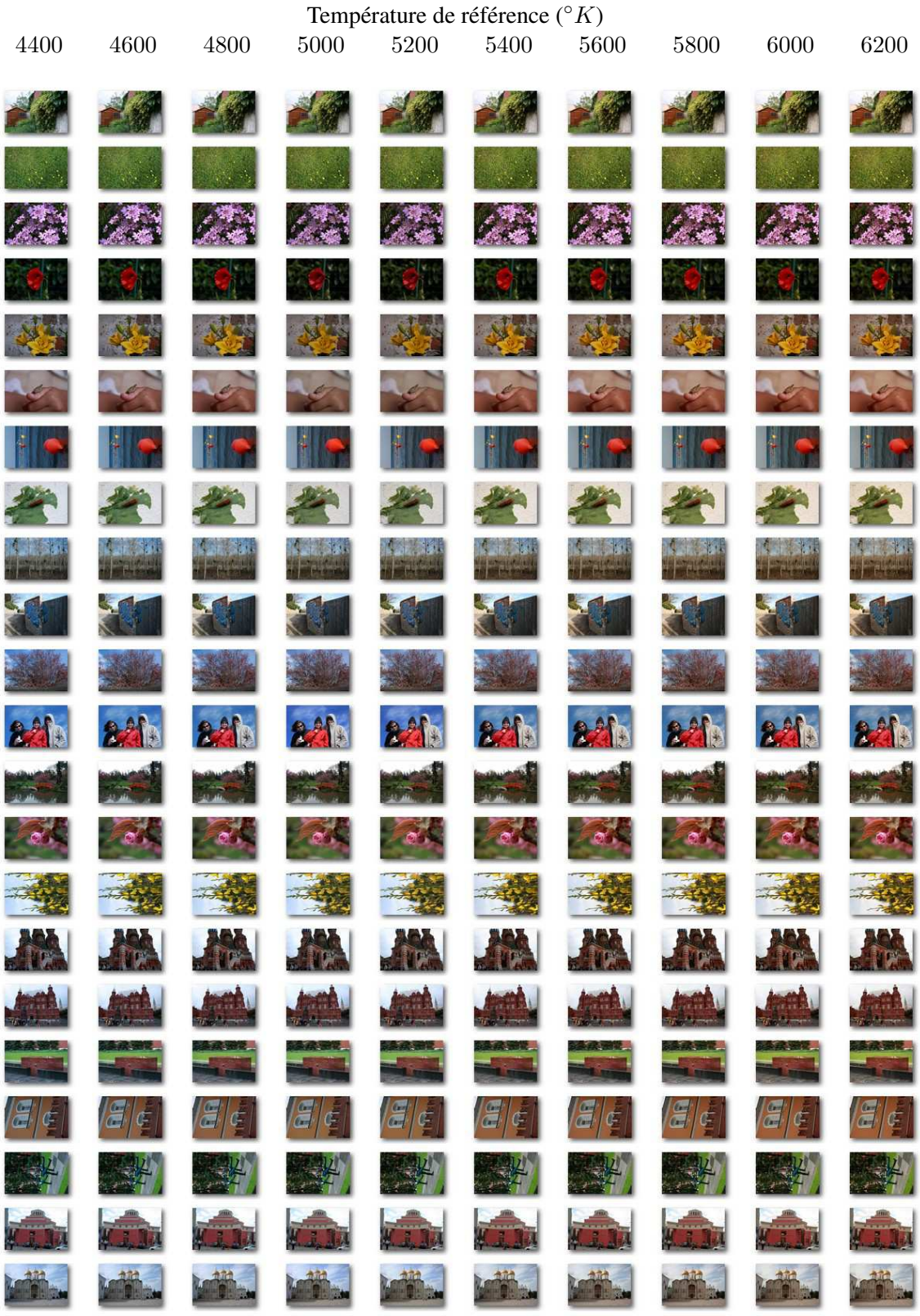
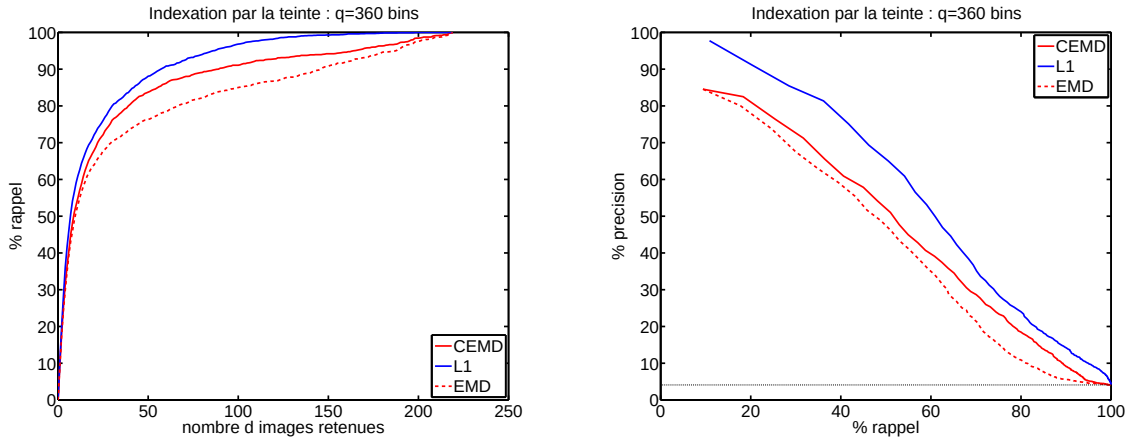


FIG. 6.15 – *Base de photographies avec modification de la balance des blancs.*

Les courbes de performance moyenne de l'indexation de cette base pour les distances CEMD, EMD et L^1 sont données en figure 6.16. Cette fois, on observe que la distance bin-à-bin L^1 est en moyenne plus performante que la distance de transport CEMD. Ceci s'explique en considérant la figure 6.17 où sont montrés les 10 histogrammes de teinte d'une même classe et la courbe de précision-rappel moyen pour cette classe. On peut constater que le changement de température de référence se traduit à la fois par une translation des modes mais également par un changement de poids des modes. Il est intéressant de voir que, suivant l'importance de la perturbation en poids ou en translation, c'est l'une ou l'autre des distances L^1 ou CEMD qui est la meilleure. Ainsi, dans le cas de la figure 6.17(a), c'est principalement un changement de poids qui se produit, ce qui conduit aux mauvaises performances de CEMD (figure 6.17(b)). Dans le cas de la figure 6.17(e), les deux phénomènes se produisent simultanément mais la variabilité de poids est prépondérante, ce qui réduit considérablement les performances de la distance CEMD (figure 6.17(f)). Par contre, en figure 6.17(c), il n'y a qu'un seul mode par histogramme, ce qui réduit le problème du changement de poids pour la distance CEMD qui prend alors l'avantage vis-à-vis de la distance L^1 (figure 6.17(d)). Dans le cas de la figure 6.17(g), c'est le phénomène de translation qui réduit cette fois les performances de la distance L^1 (figure 6.17(h)). En moyenne cependant, le changement de poids est souvent très important ce qui explique pourquoi la distance L^1 prend l'avantage sur la distance CEMD en considérant l'ensemble de la base (figure 6.16).



(a) Courbe de rappel moyen

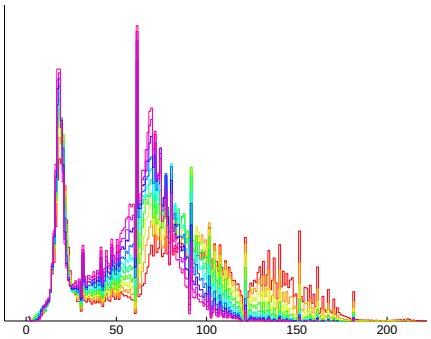
(b) Courbe de précision-rappel moyen

FIG. 6.16 – *Courbes de précision-rappel moyen de l'indexation d'une base d'images avec changement de la balance des blancs. La distance de transport* CEMD *est représentée en trait rouge continu, l'EMD (correspondant à un transport sans prise en compte de la circularité de la teinte) en trait rouge interrompu, et la distance L^1 en trait bleu continu.*

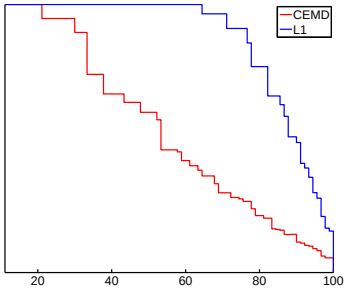
Expérience avec changement de pose Dans les expériences d'indexation d'images couleurs précédentes, une classe représentait un objet selon un même point de vue. Nous allons voir que le changement de point de vue (ou de pose) génère également une modification intra-classe du poids relatif des modes principaux dans l'histogramme de teinte.

Nous avons utilisé la base d'images de Nistér et Stewénius décrite dans [NS06][2], composées de 10200 images de résolution 640×480 pixels. Cette base est divisée en 2550 classes de 4 vues selon des poses différentes de scènes variées. En figure 6.18 sont donnés quelques exemples de ces photographies. Comme pour l'expérience précédente, nous réalisons une indexation de la base à partir d'histogrammes de teinte pondérés par la saturation et quantifiés sur 360 bins. En raison de limitation de mémoire, nous nous sommes restreints à l'indexation des 1000 premières classes.

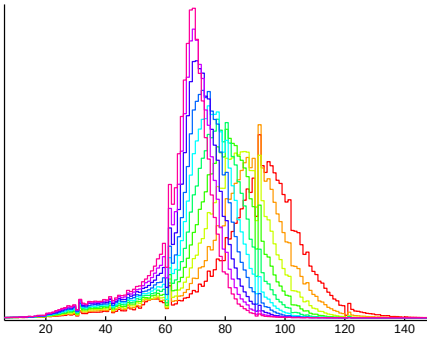
[2]Cette base est disponible à l'adresse : http://vis.uky.edu/~stewe/ukbench/



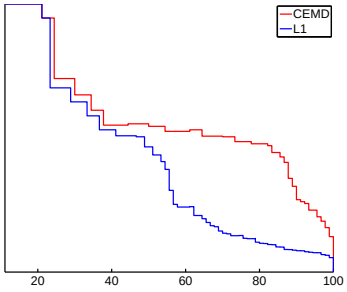
(a) Histogrammes de teinte des images de la classe 1



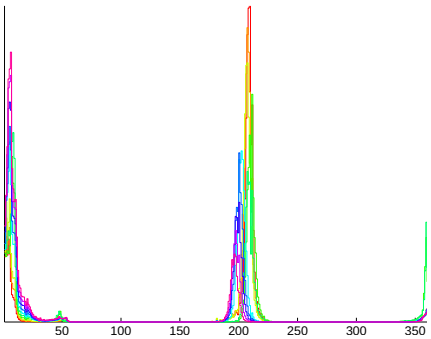
(b) Courbe de précision-rappel moyen sur la classe 1



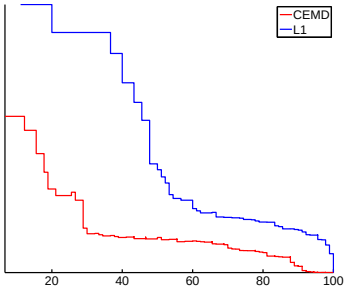
(c) Histogrammes de teinte des images de la classe 2



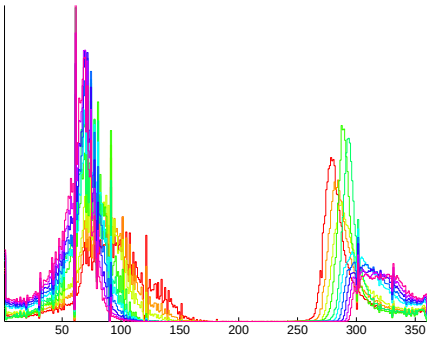
(d) Courbe de précision-rappel moyen sur la classe 2



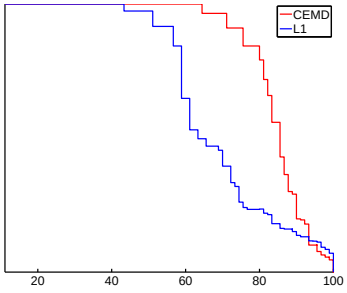
(e) Histogrammes de teinte des images de la classe 7



(f) Courbe de précision-rappel moyen sur la classe 7



(g) Histogrammes de teinte des images de la classe 3



(h) Courbe de précision-rappel moyen sur la classe 3

FIG. 6.17 – *Performances pour quelques classes (l'indice de la classe correspond au numéro de ligne dans la figure 6.15). Les figures de la colonne de gauche représentent les histogrammes au sein d'une même classe. Les courbes de performance pour chacune des classes sont données dans la colonne de droite, en fonction des distances L^1 (en bleu) et* CEMD *(en rouge).*

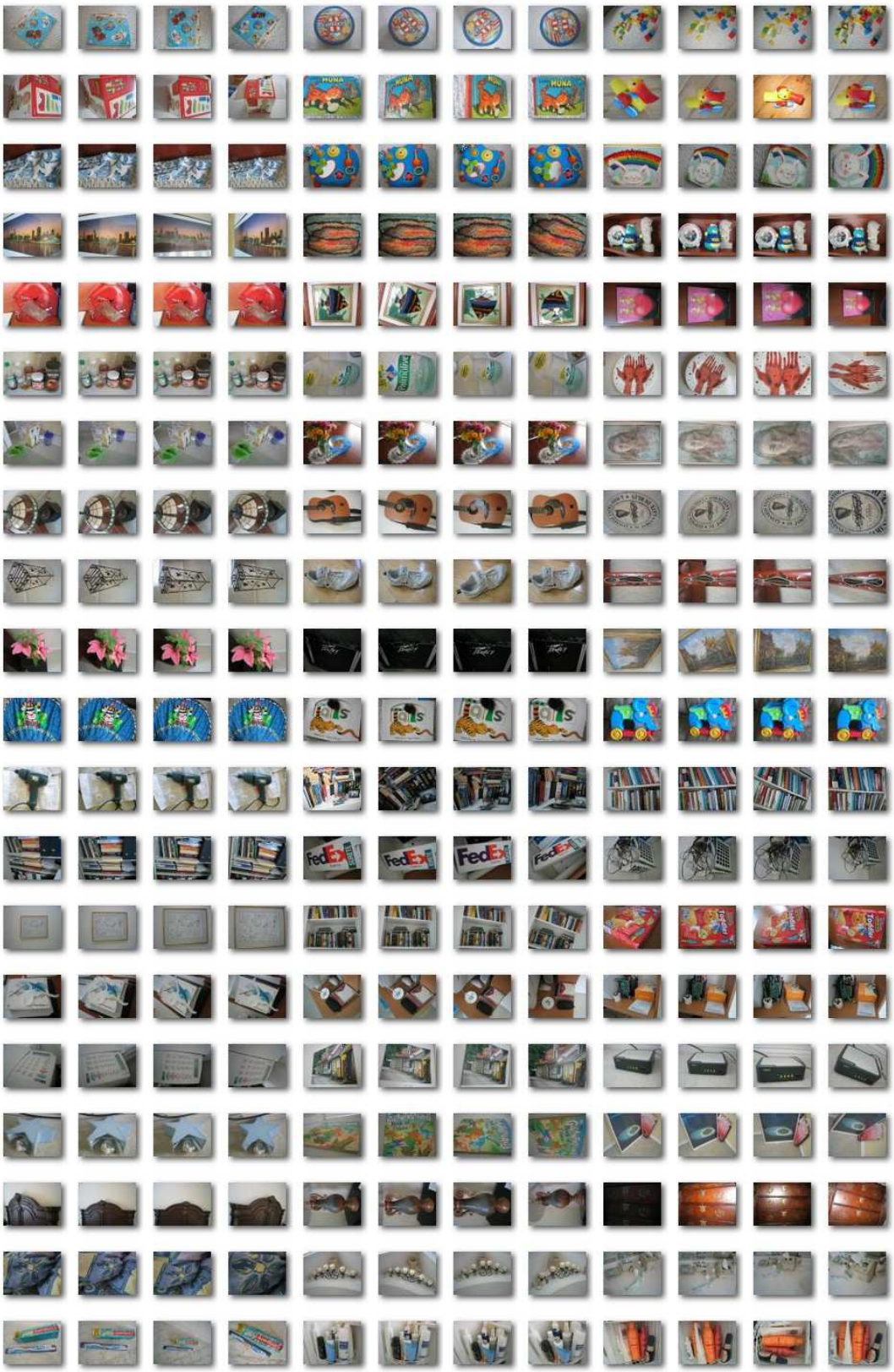


FIG. 6.18 – *Quelques images de la base de photographies avec prises de vue variables (4 vues par classe) [NS06].*

Les courbes de performance moyenne pour les distances CEMD et L^1 sont données en figure 6.5. On observe une fois encore que la distance L^1 donne de meilleurs résultats en moyenne que la distance de transport CEMD.

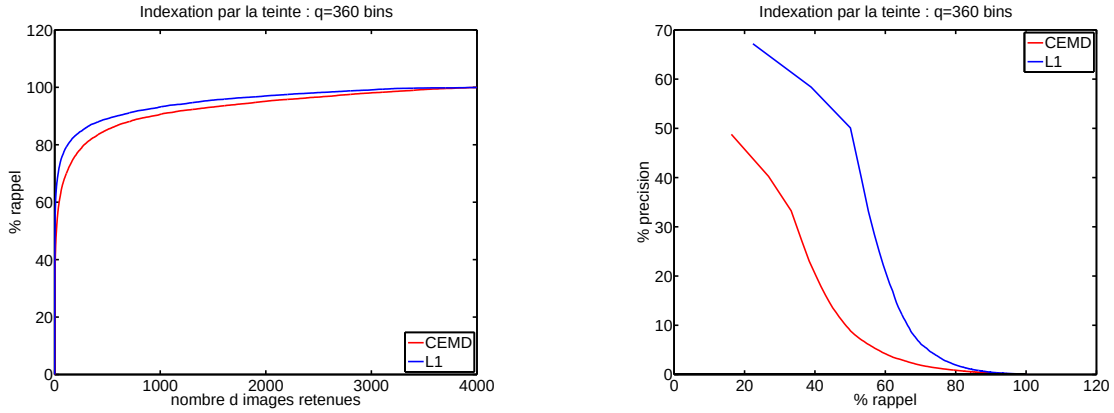


FIG. 6.19 – *Courbes de précision-rappel moyen de l'indexation avec des variations de pose intra-classe de* 4000 *images de la base [NS06]. La distance de transport* CEMD *est représentée en trait rouge continu, et la distance* L^1 *en trait bleu continu.*

Tout comme dans les cas précédents, les changements de balance de blancs éventuels entre les différentes vues se traduisent à la fois par des perturbations au sol et sur le poids des modes. Cependant, la principale différence vient du fait que cette base d'images a été conçue pour la mesure de performance d'une méthode d'indexation robuste au changement de point de vue. Les photographies d'une même classe sont donc des transformations géométriques de la même scène, ce qui se traduit du point de vue de la teinte par des perturbations intra-classe très importantes au niveau du poids des modes principaux. Par exemple, lorsque l'on photographie un objet sur un fond uniforme, l'histogramme possède un mode pour la teinte du fond et un autre correspondant à celle de l'objet. Si l'on effectue un zoom sur l'objet, sa taille relativement au fond augmente et par conséquent, le poids du mode relatif à la teinte de l'objet augmente également. La figure 6.20 illustre ce phénomène pour les quatre histogrammes de l'une des classes de la base. Ce phénomène vient une nouvelle fois limiter les performances de la distance CEMD.

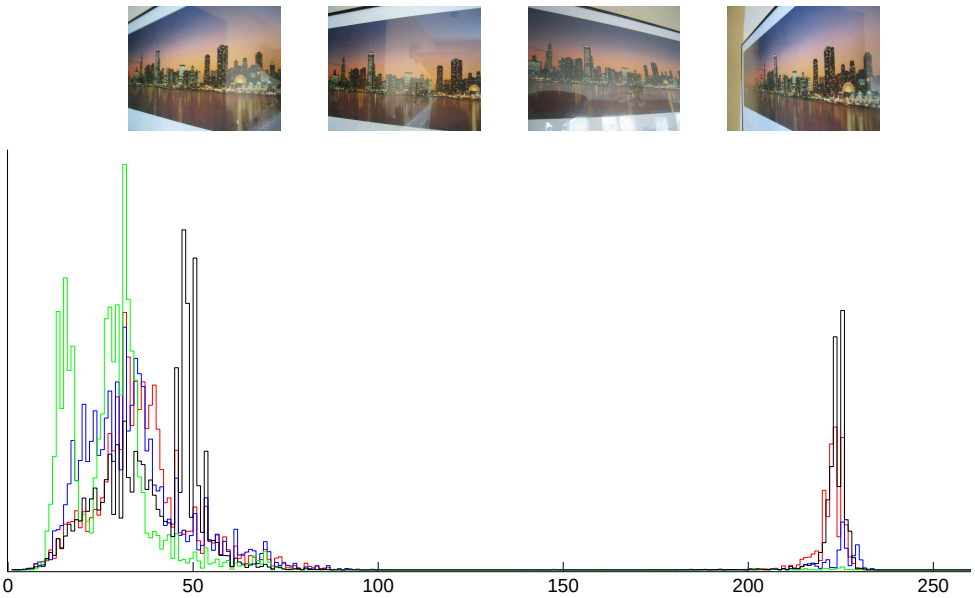


FIG. 6.20 – *Exemple de l'effet du changement de point de vue sur l'histogramme de teinte.*

Dans le même temps, nous avons pu remarquer que les perturbations liées à l'échantillonnage – auxquelles la distance L^1 est très peu robuste – étaient très limitées. D'une part, comme pour les expériences précédentes, un grand nombre d'échantillons sont utilisés pour la construction des histogrammes de teinte. D'autre part, en raison de la compression JPEG, le phénomène de quantification de la teinte est très limité. En effet, comme l'illustrent les 4 histogrammes de la figure 6.21 provenant de la même classe, les mêmes bins sont systématiquement remplis. Ce dernier point est la raison pour laquelle la distance L^1 surpasse la distance de transport dans un tel cas.

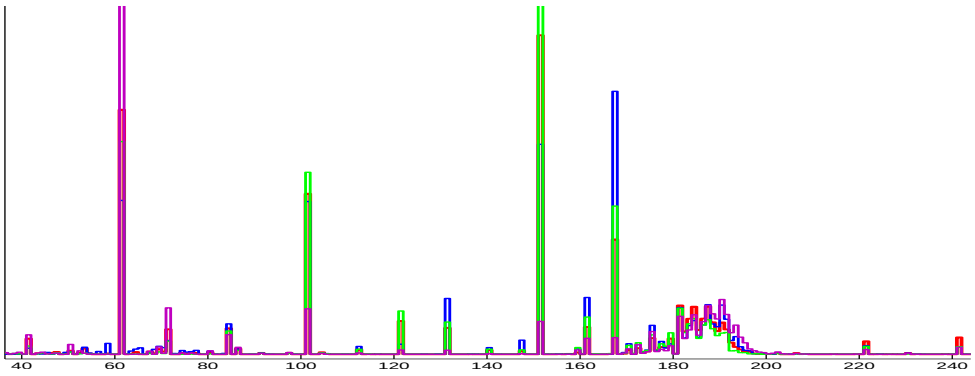


FIG. 6.21 – *Exemple de la quantification des teintes sur la base d'images.* *Les quatres histogrammes de teinte représentés correspondent aux différentes images d'une même classe.*

> **Remarque 2 :**
> Dans l'étude [RBSW00], l'utilisation de l'EMD est comparée à des distances de type bin-à-bin pour indexer des images de la base COREL, représentées par des histogrammes de couleur 3D (dans l'espace HSV). Ainsi que le font remarquer les auteurs, la distance de transport EMD est – théoriquement – d'autant plus intéressante vis-à-vis des distance bin-à-bin que, dans le cas des histogrammes multidimentionnels, le phénomène de quantification y est plus important. Ils s'étonnent pourtant de constater que la distance EMD donne des résultats presque similaires à la distance L^1, sans toutefois avancer d'explications. On serait néanmoins tenté d'y voir, une fois encore, la manifestation du compromis entre la robustesse à la quantification et aux translations de l'EMD d'une part, et la robustesse au changement de poids de la distance L^1 d'autre part.

Discussion sur la pertinence de la distance au sol Nous avons vu que l'utilisation d'une distance de transport permet d'être plus robuste aux effets de l'échantillonnage (quantité d'échantillons et quantification) et d'une certaine classe de perturbation (qui se manifeste par des translations) qui peuvent affecter les histogrammes. Nous avons également vu qu'il était possible de tirer parti de la structure circulaire d'un histogramme, en utilisant une distance au sol c sur le cercle qui lui est mieux adaptée : $c(x,y) = \min\{d(x,y), 1-d(x,y)\}$ avec $d(x,y) = |x-y|$. Cependant, un aspect dont nous n'avons pas discuté jusqu'à présent est le choix de la distance $d(x,y)$ en lui-même.

La problématique de la définition d'une distance analogue à la perception cognitive a été abordée dans la litérature. Par exemple, le modèle de représentation des couleurs CIE-Lab a été spécialement conçu dans le but que la distance euclidienne corresponde *localement* aux nuances perçues par un individu moyen. Shepard [She87] réalise des expériences montrant que la réaction d'un sujet (mesurée en terme de probabilité de confusion) à deux stimuli différents (tels que la couleur) décroit exponentiellement en fonction de la dissimilarité de ces stimuli. En s'appuyant sur ce résultat, Tomasi *et al.* [RT01] proposent d'utiliser une distance au sol de type *exponentielle* pour la comparaison d'histogrammes couleurs par l'EMD :

$$\hat{c}(x,y) = 1 - e^{-\frac{d(x,y)}{\tau}},$$

où $d(.,.)$ est une distance dans l'espace couleur. Ce choix a ensuite été repris dans plusieurs travaux,

dont Hurtut *et. al.* [HGS08] pour la comparaison de pixels avec des attributs couleurs et spatiaux. Une alternative consiste à utiliser une distance au sol tronquée, comme le suggère Lv *et. al.* dans [LCL04] :

$$\hat{c}(x, y) = \min\{c(x, y), c_{\max}\} \ ,$$

où $c_{\max}$ est la pénalité maximale de transport pour une masse unité. Très récemment, les auteurs de [PW09] ont montré que ce choix de distance au sol tronqué permet de diminuer le temps de calcul de l'EMD pour les histogrammes multidimensionnels. Ils utilisent cette distance pour l'indexation d'images couleurs.

Avec ce type de distance au sol, on peut s'attendre à augmenter la robustesse de la distance de transport EMD au phénomène de changement de pondération. En effet, nous avons vu que le seuil critique ϵ_p dans le cas d'un mélange de deux gaussiennes dépendait de la distance entre les deux modes. Cependant, l'utilisation d'une telle distance au sol ne permet pas d'éviter complètement ce phénomène : comme le remarquent Peleg et Werman, sur certaines classes « faciles » (voir la figure (4.f) de [PW09]), la distance bin-à-bin L^1 fait légèrement mieux que la distance de transport tronquée. Notons par ailleurs que ces distances requièrent le réglage d'un paramètre supplémentaire τ ou $c_{\max}$.

Le problème de la différence du poids relatif des modes qui vient d'être mis en évidence dans ce chapitre a des conséquences pour le transfert de caractéristiques par le transport optimal. Nous verrons au chapitre 8 une méthode de régularisation permettant d'en limiter les effets.

Dans le chapitre suivant, nous allons étudier l'application de la distance CEMD pour la comparaison de descripteurs locaux de type SIFT. La différence principale avec l'indexation d'images que nous avons étudiée dans ce chapitre, est que les histogrammes locaux sont obtenus à partir d'un nombre plus faible d'échantillons et qu'ils sont soumis à de plus grandes perturbations géométriques.

Chapitre 7

Application de CEMD aux descripteurs locaux de type SIFT

Ce chapitre est dédié à la mesure de dissimilarité utilisée pour la comparaison de descripteurs locaux de type SIFT. Nous avons vu dans la première partie de cette thèse l'importance jouée par la mesure de dissimilarité dans le processus de mise en correspondance (chapitres 1 et 2). Ce processus consiste, pour un descripteur requête, à ordonner les descripteurs candidats de la base de données par degré croissant de dissimilarité, puis à valider les plus similaires d'entre eux à l'aide d'un critère de décision.

Dans le chapitre précédent, nous avons étudié la distance de Monge-Kantorovich dans le cas unidimensionnel et circulaire. Cette distance de transport, appelée CEMD, a ensuite été utilisée pour la comparaison d'histogrammes globaux dans le cadre de l'indexation de bases d'images. Nous proposons ici d'utiliser la distance CEMD pour la comparaison de descripteurs SIFT, constitués d'histogrammes circulaires d'orientation du gradient.

Dans un premier temps, nous allons brièvement rappeler quelles sont les distances utilisées dans la littérature pour comparer de tels descripteurs (§ 7.1), pour ensuite étudier la manière d'employer la distance CEMD dans le cas des SIFTs (§ 7.2). Les performances de la mesure de dissimilarité obtenue sont par la suite analysées à l'aide d'une large base d'images (§ 7.3).

Ces travaux ont fait l'objet d'une publication dans [RDG08b, RDG09].

7.1 Précédents travaux sur la comparaison de descripteurs SIFT

Rappelons qu'un descripteur local a de type SIFT peut à la fois être considéré comme un histogramme 3D, ou bien comme une collection de M histogrammes 1D circulaires d'orientation du gradient a_m. Chaque histogramme a_m est quantifié sur N bins, et l'ensemble du descripteur est normalisé au poids unité, de telle sorte que si l'on choisit la norme L^1 :

$$\sum_{m=1}^{M} \sum_{n=1}^{N} a_m[n] = 1 \ .$$

Pour de plus amples détails, voir l'annexe B qui est consacrée à la construction des descripteurs SIFT.

La comparaison de deux descripteurs SIFT a et b revient donc à examiner les paires d'histogrammes d'orientation du gradient (a_m, b_m). Deux catégories de distances sont envisageables pour réaliser cette comparaison : les distances que l'on appelle « bin-à-bin », qui consistent à ne comparer que les valeurs des bins ayant la même position, et les distances « inter-bins » qui au contraire permettent de comparer des bins ayant des positions différentes. Dans les deux prochains paragraphes, nous allons faire l'état de l'art des différentes métriques utilisées pour comparer des descripteurs SIFT selon cette classification.

7.1.1 Distances bin-à-bin

Les distances les plus utilisées pour la comparaison de descripteurs locaux sont les distances bin-à-bin. Dans les premiers travaux de Lowe [Low04], les SIFTs sont initialement normalisés avec la norme L^2 puis comparés avec la distance euclidienne :

$$\mathrm{D}_{L^2}(a,b) := \sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N}\left(\,a_m[n]-b_m[n]\,\right)^2}\,,\ \text{avec}\ \sum_{m=1}^{M}\sum_{n=1}^{N}a_m[n]^2=\sum_{m=1}^{M}\sum_{n=1}^{N}b_m[n]^2=1\,. \tag{7.1}$$

On peut généraliser ce principe à la norme L^p. En particulier, avec la norme L^1, on obtient :

$$\mathrm{D}_{L^1}(a,b) := \sum_{m=1}^{M}\sum_{n=1}^{N}\left|\,a_m[n]-b_m[n]\,\right|\,,\ \text{avec}\ \sum_{m=1}^{M}\sum_{n=1}^{N}a_m[n]=\sum_{m=1}^{M}\sum_{n=1}^{N}b_m[n]=1\,. \tag{7.2}$$

Une solution alternative proposée par Belongie *et al.* [BMP02] pour les descripteurs locaux *Shape Context*, est d'utiliser la distance du χ^2 définie ainsi :

$$\mathrm{D}_{\chi^2}(a,b) := \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N}\frac{(a_m[n]-b_m[n])^2}{a_m[n]+b_m[n]}\,,\ \text{avec}\ ||a||_1=||b||_1=1\,. \tag{7.3}$$

Dans l'article de Zhang *et al.* [ZMLS07], cette distance est utilisée pour la comparaison d'histogrammes d'occurrences de SIFTs (*bag of features*). Cette distance est également reprise par Forssén et Lowe dans [FL07] pour la comparaison de descripteurs de type SIFT.

Cependant, d'autres distances classiques utilisées pour la comparaison d'histogrammes pourraient également l'être pour des descripteurs locaux. Par exemple, la divergence de Jeffrey qui est l'expression symétrisée de la divergence de Kullback-Leibler $\mathrm{D}_J(a,b)=\mathrm{KL}(a,\frac{a+b}{2})+\mathrm{KL}(b,\frac{a+b}{2})$, soit

$$\mathrm{D}_J(a,b) := \sum_{m=1}^{M}\sum_{n=1}^{N}a_m[n]\log\left(\frac{2\,a_m[n]}{a_m[n]+b_m[n]}\right)+b_m[n]\log\left(\frac{2\,b_m[n]}{a_m[n]+b_m[n]}\right)\,,\ \text{avec}\ ||a||_1=||b||_1=1\,. \tag{7.4}$$

Nous avons montré dans le chapitre précédent (section 6.2) que les distances bin-à-bin ne sont pas robustes à deux types de phénomènes :

- **Échantillonnage** Les histogrammes locaux sont construits à partir d'un nombre réduit d'échantillons. Contrairement aux histogrammes globaux, on obtient des distributions très irrégulières. Or, les distances bin-à-bin sont peu robustes à ce type de variabilité intra-classe. Par ailleurs, plus le nombre de bins est élevé, plus le pouvoir discriminant du descripteur est théoriquement grand. C'est un aspect très important pour les applications de reconnaissance d'objets où l'on souhaite retrouver exactement l'objet requête, contrairement à ce que l'on cherche à faire dans le cas de l'indexation d'images. Cependant, en raison de la non robustesse des distances bin-à-bin à la quantification (et pour limiter le temps de calcul), le nombre de bins est en pratique limité, ce qui réduit les performances de la mise en correspondance.
- **Perturbations sur les positions des modes** En cas de perturbations géométriques, la position des modes principaux des histogrammes peut changer. Typiquement, une erreur de la définition de l'orientation principale dans le cas des SIFTs se traduit par une translation de l'ensemble de l'histogramme, ce à quoi les distances bin-à-bin sont très sensibles.

Pour éviter ces différentes limitations, plusieurs études ont suggéré l'utilisation d'une distance inter-bins pour la comparaison de descripteurs locaux.

7.1.2 Distances inter-bins

On trouve dans la littérature quelques exemples d'utilisation de distances inter-bins en tant que mesure de dissimilarité entre descripteurs SIFT.

Tout d'abord, dans l'étude comparative de descripteurs locaux réalisée dans [MS05], Mikolajczyk et Schmid proposent d'utiliser la distance de Mahalanobis pour des descripteurs locaux obtenus par des filtres tels que les *Steerable Filters* [FA91] ou par le calcul de moments invariants. En s'inspirant de cette étude, Moreels et Perona comparent dans [MP05] la distance euclidienne à la distance de Mahalanobis pour plusieurs descripteurs construits à partir d'histogrammes locaux (SIFT [Low04], PCA-SIFT [KS04], et Shape Context [BMP02]).

La distance de Mahalanobis entre deux descripteurs SIFT a et b est définie ainsi dans [MP05] :

$$\mathrm{D}_M(a,b) := \sqrt{(a-b)^t C^{-1}(a-b)} = \left(\sum_{m,m'=1}^{M} \sum_{n,n'=1}^{N} \left(a_m[n] - b_m[n]\right) \omega_{m,n,m',n'} \left(a_{m'}[n'] - b_{m'}[n']\right) \right)^{\frac{1}{2}} \tag{7.5}$$

avec $||a||_2 = ||b||_2 = 1$, et où C est une matrice de covariance de SIFTs, qui est en pratique estimée empiriquement sur une base d'images. Par ce procédé, la comparaison de deux histogrammes n'est plus seulement limitée à la comparaison des bins de même indice, mais potentiellement étendue à l'ensemble des autres bins. Afin que la distance ait du sens, la matrice de covariance inverse C^{-1} est utilisée pour pondérer chacune des comparaisons (donnant les poids $\omega_{m,n,m',n'}$). On retrouve ainsi la distance L^2 en remplaçant cette matrice par la matrice identité. Si la distance obtenue est bien « inter-bins », elle diffère grandement d'une distance de transport car les poids $\omega_{m,n,m',n'}$, au lieu de dépendre des histogrammes a et b, sont appris sur une base de descripteurs et sont ensuite fixés pour toutes les comparaisons de descripteurs, quels que soient a et b. C'est la raison pour laquelle cette distance en pratique ne présente pas de véritable intérêt par rapport à la distance L^2. C'est d'ailleurs la conclusion à laquelle aboutissent les auteurs de [MP05] : comme on peut le constater sur la figure 7.1, les performances sont pour la plupart des descripteurs inchangées, et dans le cas des SIFTs, les performances sont même très en deçà de celles obtenues avec la distance euclidienne.

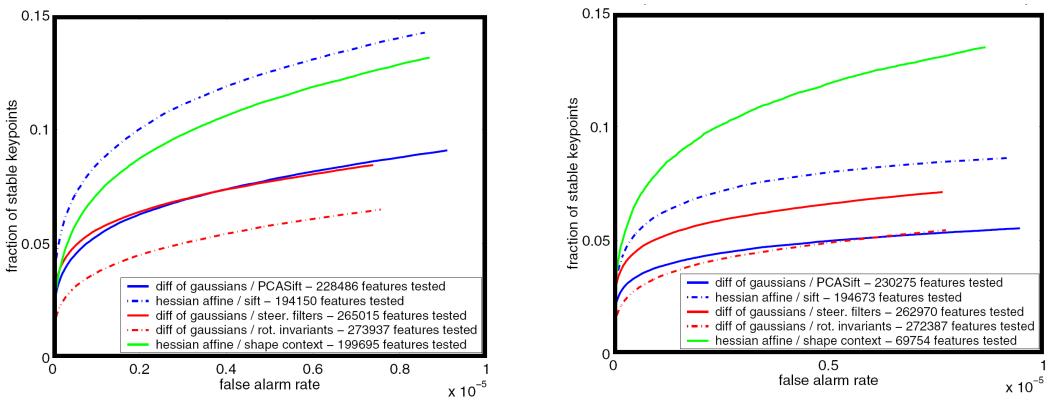


FIG. 7.1 – *Comparaison de la distance euclidienne (D_{L^2}, figure de gauche) et de la distance de Mahalanobis (D_M, figure de droite) pour différents descripteurs locaux pour la mise en correspondance d'images (graphiques extraits des figures 9 et 10 de [MP05]). Dans le cas des SIFTs (courbes de performance en trait bleu interrompu), la distance euclidienne donne de meilleurs résultats que la distance de Mahalanobis.*

Notons qu'une distance appelée *quadratic distance* à été proposée dans [NBE+93], dont l'expression est très proche de la distance de Mahalanobis. La seule différence réside dans le fait que la matrice de pondération C^{-1} est directement définie par l'utilisateur.

Distance de transport Une alternative plus intéressante, afin de répondre au problème de la quantification, est de considérer une mesure de dissimilarité fondée sur le transport, telle que celle formalisée par l'EMD (Earth Mover Distance) [RTG00]. Dans [LO07], Ling et Okada proposent de comparer les descripteurs SIFT en utilisant la distance EMD. Rappelons que les descripteurs SIFT peuvent être vus

comme des histogrammes tridimensionnels, normalisés, représentant à la fois une orientation de gradient ainsi que la position spatiale des régions desquelles ils sont extraits. Dans leurs expériences, les performances obtenues avec l'EMD sont alors bien meilleures qu'avec la distance euclidienne, ce qui montre l'intérêt de l'utilisation du transport en terme de robustesse.

Ling et Okada présentent un nouvel algorithme de calcul de l'EMD dans le cas où la distance au sol est la distance L^1 (appelé EMD-L^1), ce qui permet d'en réduire la complexité et de gagner deux ordres de grandeur en temps de calcul pour les SIFTs. Néanmoins, le temps de calcul nécessaire à la comparaison de descripteurs SIFT par l'algorithme EMD-L^1 est en pratique trop prohibitif pour envisager son utilisation pratique. En effet, d'après le tableau VII dans [LO07], la distance EMD-L^1 est empiriquement 720 fois plus longue à calculer que la distance euclidienne.

De plus, les auteurs de cette étude ne mentionnent pas deux aspects critiques du transport en vue de la comparaison des descripteurs SIFT :

- **Circularité** L'une des trois dimensions de l'histogramme 3D est circulaire (orientation du gradient), et par conséquent la distance au sol utilisée doit théoriquement prendre en compte cet aspect afin que le transport ait du sens ; or, la validité de l'algorithme proposé dans [LO07] n'a pas été montrée dans le cas où l'une des dimensions est circulaire.
- **Normalisation** Les descripteurs SIFT construits à l'aide du code original de Lowe [Low] sont normalisés avec la norme L^2. Les auteurs ne précisent pas si les descripteurs sont renormalisés selon la norme L^1.

Une autre distance inter-bins a été également proposée par Ling et Okada dans [LO06]. Cette distance de diffusion (*diffusion distance*) est fondée sur la convolution de la différence des deux histogrammes à comparer. Cette convolution est réalisée à différentes échelles, et la distance mesure à quelle vitesse la différence des histogrammes tend vers le vecteur nul 0. Dans les expériences réalisées sur des SIFTs, les performances sont similaires à celle de l'EMD, mais le temps de calcul est beaucoup plus rapide (un facteur d'environ 6 dans le tableau 3 donné dans [LO06]) Cependant, la distance de diffusion nécessite la définition de plusieurs paramètres contrairement à l'EMD : le choix du noyau de convolution (et en particulier les différentes échelles utilisées), ainsi que le nombre d'itérations. Par ailleurs, la question de la circularité des SIFTs n'est pas abordée.

Nous allons dans la section suivante proposer une mesure de dissimilarité fondée sur la distance CEMD, présentée à la section 6.1, qui permet d'exploiter la structure circulaire des histogrammes d'orientation du gradient, tout en en limitant la complexité par la comparaison d'histogrammes unidimensionnels. Notons que des travaux similaires ont été récemment publiés dans [PW08], dont nous ferons par la suite une étude plus approfondie.

7.2 Utilisation de la distance CEMD pour la comparaison de descripteurs locaux

Nous considérons dorénavant un descripteur a de type SIFT comme une collection de M histogrammes circulaires et uni-dimensionnels : $\{a_m\}_{1\leq m\leq M}$. On se place ainsi dans le cadre d'étude du chapitre précédent, à la différence près que les descripteurs SIFT sont construits de telle sorte que les histogrammes sont normalisés *globalement* : chacun des histogrammes a_m possède alors une norme différente. Par la suite, nous supposons que c'est la norme L^1 qui est employée, ce qui nous donne pour un descripteur a :

$$\sum_{m=1}^{M} ||a_m||_1 = \sum_{m=1}^{M}\sum_{n=1}^{N} a_m[n] = 1 \ .$$

Dans le but d'utiliser la distance CEMD pour définir une mesure de dissimilarité entre deux descripteurs a et b, il nous faut répondre à deux questions :

- comment utiliser la distance CEMD entre des histogrammes non normalisés a_m et b_m ?
- comment combiner ces distances pour définir une mesure de dissimilarité globale entre a et b ?

7.2.1 Normalisation des histogrammes

Plusieurs alternatives sont possibles pour comparer les histogrammes 1D a_m et b_m. Une première possibilité consiste à utiliser la définition originale de l'EMD, donnée au chapitre 5, entre des ensembles de poids total différents. Avec cette définition, seul le minimum entre les poids des deux histogrammes, soit ici $\min\{||a_m||_1, ||b_m||_1\}$, est transporté. Le problème avec ce choix de distance est que la mesure de dissimilarité est artificiellement faible pour les descripteurs ayant plusieurs histogrammes vides (ou de normes L^1 très petites). En effet, imaginons que l'on détecte un point d'intérêt sur la frontière d'un objet *photographié sur un fond plus clair et uniforme*, et prenons pour simplifier un descripteur avec $M = 2$ régions : une pour l'intérieur de l'objet, et une autre pour l'extérieur. Le descripteur correspondant à ce point se compose d'un histogramme a_1 vide (*i.e.* de norme nulle), le gradient étant nul dans la région du masque coïncidant avec le fond, et un histogramme a_2 de norme 1. Supposons qu'il existe un autre point d'intérêt (par exemple sur la frontière d'un objet avec un fond sombre) pour lequel on trouve la situation inverse : un histogramme b_1 de norme 1 et un deuxième histogramme vide b_2. Quelle que soit la combinaison entre les distances utilisées par la suite, étant donné que $EMD(a_1, b_1) = EMD(a_2, b_2) = 0$, la distance totale entre a et b sera nulle, quand bien même les deux descripteurs sont très différents. Cet exemple simple montre que cette définition n'a pas de sens dans le cas des SIFTs. En outre, il n'existe pas d'expression analytique permettant de calculer directement cette distance, qui nécessite par conséquent l'utilisation d'un algorithme de type « simplexe » [Str89] de complexité élevée.

Une seconde possibilité est de normaliser individuellement chacun des histogrammes du descripteur a avec la norme L^1 : $\forall m = 1, \ldots, M \;\; ||a_m||_1 = 1$. Cette normalisation *locale* plutôt que globale des descripteurs SIFT, signifie que le descripteur est désormais invariant localement –c'est-à-dire pour chaque région du masque– à un changement de contraste affine. La distance entre deux histogrammes a_m et b_m peut alors s'exprimer comme la distance CEMD définie par la formule (6.18), qui utilise la distance au sol L^1 sur le cercle, soit $c(i,j) = \frac{1}{N}\min(|i-j|, N-|i-j|),\ \forall (i,j) \in \{1,\ldots N\}^2$:

$$\text{CEMD}(a_m, b_m) = \frac{1}{N} \min_{k \in \{1,\ldots,N\}} \|A_m^k - B_m^k\|_1 \ , \text{ avec } ||a_m||_1 = ||b_m||_1 = 1 \tag{7.6}$$

où A_m^k et B_m^k sont les histogrammes cumulés depuis le k-ième bin des histogrammes a_m et b_m respectivement, ce qui donne pour A_m^k (la définition de B_m^k est la même en remplaçant A par B)

$$\forall i, k \in \{1, \ldots, N\} \ , \ A_m^k[i] = \begin{cases} A_m[i] - A_m[k-1] & \text{si } i \geq k \\ A_m[i] - A_m[k-1] + A_m[N] & \text{si } i \leq k-1 \end{cases} \ , \tag{7.7}$$

avec la convention $A[0] = 0$, et où $A_m[N] = 1$ est égal à la norme L^1 de a_m .

> **Remarque 1 :**
> Cette distance peut également s'exprimer selon la formule (6.17), les histogrammes étant renormalisés à l'unité :
> $$\text{CEMD}(a_m, b_m) = \frac{1}{N}||A_m - B_m - \mu||_1 \ \text{ avec } ||a_m||_1 = ||b_m||_1 = 1 \ , \tag{7.8}$$
> où A_m et B_m sont les histogrammes cumulés depuis le premier bin des histogrammes a_m et b_m respectivement, et où μ est le médian de l'ensemble des valeurs $\{(A_m[i] - B_m[i])\}_{1 \leq i \leq N}$.

Cependant, ainsi que cela a été remarqué par Lowe dans [Low04], les performances de la distance euclidienne sont meilleures lorsque l'on utilise une normalisation globale, au lieu de locale. Nous verrons dans la partie expérimentale que cela est également vrai pour toutes les distances bin-à-bin que nous avons comparées. L'explication la plus plausible à ce phénomène est le compromis classique entre invariance et pouvoir discriminant du descripteur : avec une normalisation locale, le descripteur gagne en invariance au changement de contraste, mais il perd l'information de « contraste moyen » sur l'ensemble du masque.

Pour cette raison, nous définissons une mesure de dissimilarité alternative à la formule 7.6, en utilisant les histogrammes a_m et b_m normalisés globalement, afin de conserver cette information importante

des descripteurs SIFT. Ce qui nous donne ici :

$$\text{CEMD}(a_m, b_m) = \frac{1}{N} \min_{k \in \{1,\dots,N\}} \|A_m^k - B_m^k\|_1 \ , \text{ avec } \sum_{m=1}^{M} ||a_m||_1 = \sum_{m=1}^{M} ||b_m||_1 = 1 \ , \tag{7.9}$$

où A_m^k et B_m^k sont toujours définis selon l'équation (7.7), mais cette fois $A_m[N]$ et $B_m[N]$ ne sont plus nécessairement égaux à 1. Nous verrons dans la partie expérimentale l'intérêt de cette définition alternative de CEMD avec une normalisation globale.

> **Remarques 2 :**
> - Cette formulation alternative de CEMD n'est plus une distance en raison de la non normalisation des deux histogrammes qui sont comparés. En effet, les propriétés de symétrie ($\forall a, b$ $\text{CEMD}(a,b) = \text{CEMD}(b,a)$), et de séparation ($\forall a, b$ $\text{CEMD}(a,b) = 0 \Leftrightarrow a = b$) sont toujours vraies, mais l'inégalité triangulaire n'est plus vérifiée : $\forall a, b, c$ $\text{CEMD}(a,b) \nleq \text{CEMD}(a,c) + \text{CEMD}(b,c)$. Par exemple, soient c un histogramme vide de taille N ($c[n] = 0 \, \forall n$), a un histogramme valant 1 en un bin i et nul partout ailleurs ($a[n] = \delta\{n-i\}$) et b valant 1 en un bin $j \neq i$ et nul partout ailleurs ($b[n] = \delta\{n-j\}$). Si les bins i et j ne sont pas voisins, alors on obtient l'inégalité $\text{CEMD}(a,b) = |i-j|/N \geq \text{CEMD}(a,c) + \text{CEMD}(c,b) = 2/N$.
> - Il n'est plus possible de décliner l'expression (7.9) comme une fonction du médian, comme c'est le cas lorsque les histogrammes sont de poids égaux (équation (7.8)). La complexité de cette formulation est donc plus grande. Nous reviendrons ultérieurement sur son expression au paragraphe 7.2.3.

Il est difficile de donner une interprétation simple de l'utilisation de CEMD avec des histogrammes non normalisés. La formulation (7.9) permet néanmoins d'en comprendre le principe. En minimisant selon k la norme L^1 de $A_m^k - B_m^k$, cela revient comme dans le cas normalisé à chercher sur le cercle une coupure par laquelle ne passe aucun chemin afin de trouver le transport optimal. Cette fois par contre, le coût de transport total prend en compte le transport de la différence de poids $\big|\, ||a||_1 - ||b||_1 \,\big|$ jusqu'au bin le plus éloigné sur le cercle, c'est-à-dire $k + \lceil N/2 \rceil [N]$. La mesure de dissimilarité CEMD employée avec la normalisation globale des SIFTs pénalise donc également la différence de poids entre les histogrammes.

Notons qu'une autre solution intéressante à été récemment proposée dans [PW08], qui ont également observé que la normalisation globale des descripteurs SIFT donnait de meilleurs résultats. Werman et Pele proposent une nouvelle définition de distance fondée sur l'EMD pour des histogrammes non normalisés. En reprenant leur notation, on définit :

$$\widehat{\text{EMD}}(a_m, b_m) = EMD(a_m, b_m) + \frac{1}{2}\big|\, ||a_m||_1 - ||b_m||_1 \,\big| \ , \text{ avec } \sum_{m=1}^{M} ||a_m||_1 = \sum_{m=1}^{M} ||b_m||_1 = 1 \ , \tag{7.10}$$

où $EMD(a_m, b_m)$ désigne le coût de transport de la masse minimum $\min\{||a_m||_1, ||b_m||_1\}$ et le second terme représente le coût de transport associé à la masse restante $\big|\, ||a_m||_1 - ||b_m||_1 \,\big|$. L'intérêt de cette formulation est que la mesure $\widehat{\text{EMD}}$ est bien une distance. Cependant, ainsi que nous l'évoquions au début de paragraphe, il n'existe pas d'expression analytique du coût de transport $EMD(a_m, b_m)$ dans le cas non normalisé, ce qui signifie que sa mise en œuvre est très coûteuse en comparaison d'une distance bin-à-bin.

Pour limiter le temps de calcul de $\widehat{\text{EMD}}$, Werman et Pele utilisent une distance au sol tronquée :

$$\forall (i,j) \in \{1,\dots N\}^2 \ \ \tilde{c}(i,j) = \begin{cases} \min(|i-j|, N - |i-j|) & \text{si } \min(|i-j|, N-|i-j|) \leq 2 \\ 2 & \text{sinon} \end{cases} .$$

Le coût de transport circulaire est donc arbitrairement constant au delà de 2 bins. Cette approximation leur permet de gagner un ordre de grandeur en temps de calcul, mais il est cependant important de noter que cette définition n'est pas indépendante de la quantification des histogrammes.

7.2.2 Combinaison des distances entre histogrammes

Une mesure de dissimilarité D_{CEMD} entre des descripteurs SIFT a et b peut être définie en combinant les distances CEMD entre les différentes paires d'histogrammes de même indice a_m et b_m. Nous avons choisi de définir D_{CEMD} comme la somme de ces distances :

$$D_{\text{CEMD}}(a,b) := \sum_{m=1}^{M} \text{CEMD}(a_m, b_m). \tag{7.11}$$

Nous verrons dans la partie expérimentale que l'on obtient de meilleurs résultats si l'on calcule cette distance entre des histogrammes non normalisés. Notons que d'autres combinaisons sont possibles, notamment le maximum des distances ($\max \text{CEMD}(a_m, b_m)$), à l'image de ce qui est fait dans [MSC+06] pour la comparaison de formes. Cependant, nous avons observé que la distance (7.11) était plus robuste aux occultations. Remarquons que c'est également la solution retenue par Pele et Werman dans [PW08] pour comparer des descripteurs SIFT.

7.2.3 Mise en œuvre et complexité

Nous avons vu que la complexité de la distance CEMD pour deux histogrammes normalisés est en $\mathcal{O}(N \log N)$, où N est le nombre de bins, avec une formulation utilisant le médian. Dans le cas non normalisé (expression (7.9)), la complexité est plus grande car une telle formulation n'est plus possible.

On définit $X_m^k = A_m^k - B_m^k$ la différence des histogrammes cumulés depuis le bin d'indice k, ce qui donne l'expression suivante : $\text{CEMD}(a_m, b_m) = \frac{1}{N} \min_{k\in\{1,\dots,N\}} \|X_m^k\|_1$. X_m^k peut être écrit en fonction de X_m^1, avec la convention $X_m^1[0] = 0$:

$$X_m^k[i] = \begin{cases} X_m^1[i] - X_m^1[k-1] & \text{si } i \geq k \\ X_m^1[i] - X_m^1[k-1] + X_m^1[N] & \text{si } i \leq k-1 \end{cases},$$

où $X_m^1[N]$ représente la différence entre les poids des histogrammes a_m and b_m. Ainsi, dans le cas non normalisé, le calcul de CEMD ne nécessite pas l'estimation des k différents histogrammes cumulés A_m^k et B_m^k.

En comparaison des distances bin-à-bin L^1 et L^2 de complexité $\mathcal{O}(MN)$, le calcul supplémentaire requi par D_{CEMD} entre deux descripteurs SIFT de taille $M \times N$ correspond au calcul :

- de M médians dans le cas des histogrammes normalisés individuellement. La complexité en terme de multiplication est de $\mathcal{O}(MN \log N)$.
- du minimum selon k de $\|X_k\|_1$, la norme L^1 de X_k dans le cas non normalisé. La complexité est alors de $\mathcal{O}(MN^2)$.

Pour avoir un ordre d'idée des temps de calcul des différentes distances utilisées dans la partie expérimentale, nous donnons dans le tableau 7.1 le temps moyen de comparaison de deux images de 1000 descripteurs SIFT, composés de $M = 9$ histogrammes quantifiés sur $N = 12$ bins. La distance intitulée « EMD » correspond à l'utilisation de l'EMD pour comparer des descripteurs SIFT en tant qu'histogrammes 3D.

7.3 Résultats expérimentaux

Afin d'illustrer les performances de la mesure de dissimilarité que nous avons proposée, nous allons procéder à différentes comparaisons et analyses sur une large base d'images.

7.3.1 Protocole expérimental

À l'image de l'étude faite par Rubner *et al.* (voir par exemple [RTG00]), il est nécessaire d'utiliser une base d'images variées et suffisamment large afin d'évaluer les performances moyennes d'une mesure

TAB. 7.1 – *Temps d'exécution moyen de calcul de la mesure de dissimilarité entre* 10^6 *paires de descripteurs SIFT de tailles* $M = 9$ *et* $N = 12$.

Mesure de dissimilarité	Temps moyen d'exécution (en secondes)
L^2 ou $L1$	5
χ^2	8
Jeffrey	28
D_{CEMD} (cas normalisé)	8
D_{CEMD} (cas non normalisé)	13
EMD *	$15.\,10^3$

(* Les détails de mise en œuvre sont donnés en section 7.3.4)

de dissimilarité. Nous avons pour cela collecté 732 photographies de scènes diverses, de 800×600 pixels[1]. Cette base d'images est également utilisée pour l'évaluation de la mise en correspondance de descripteurs locaux au chapitre 2, où quelques photographies sont visibles en figure 2.4. Nous avons extrait au total plus de 3.10^6 descripteurs de type SIFT, soit environ 4000 points d'intérêt en moyenne par image.

Le procédé d'évaluation que nous proposons d'utiliser est identique au protocole intitulé $A \rightarrow A'$ dans la partie expérimentale (section 2.2.2.2) du chapitre 2. Ce protocole consiste à mettre en correspondance les descripteurs SIFT entre une image A de la base et une image A', dégradation synthétique de l'image A. Cette dégradation permet à la fois de perturber les histogrammes d'orientation des descripteurs SIFT entre les deux images, et d'obtenir une vérité terrain. L'image A' est en pratique obtenue en appliquant une transformation affine à l'image A , et par ajout d'un bruit blanc gaussien. Le paramètre de « tilt » de la transformation affine est fixé à 2.5, Yu et Morel ayant montré dans [MY09] que les SIFTs sont grossièrement invariants pour des transformations affines dont le tilt n'excède pas 2. Une analyse de l'influence de ce paramètre sur la robustesse de la distance CEMD est par la suite proposée. Le bruit additif est gaussien, de moyenne nulle et d'écart-type $\sigma = 5$ pour des images quantifiées sur 256 niveaux par canal.

Pour évaluer la performance d'une distance, chacun des N_Q descripteur SIFT de l'image A est mis en correspondance avec son plus proche voisin dans l'image A' selon la mesure de dissimilarité évaluée. Un appariement de deux descripteurs SIFT est considéré comme faux (*i.e.* faux-positif[2]) ou correct (*i.e.* vrai-positif), selon un critère de tolérance spatiale. Suivant le protocole utilisé par [MS05], une correspondance entre deux points d'intérêt est considérée comme correcte si l'erreur de superposition est plus petite que 50%. Cette erreur de superposition est définie par le rapport entre l'aire d'intersection et l'aire d'union des supports des descripteurs SIFT (masque circulaire) dans l'image A. Ainsi, en ordonnant la liste de l'ensemble des N_Q mises en correspondance obtenues pour l'image A par ordre croissant de dissimilarité, on peut tracer une courbe de performance (appelée courbe ROC). Cette courbe représente, pour un seuil donné sur la mesure de dissimilarité utilisée, le taux de bonnes correspondances (ou taux de rappel) en fonction du taux de fausses alarmes.

> **Remarque 1 :**
> Cette expérience correspond au protocole $A \rightarrow A'$ combiné avec le critère NN-DT présenté au chapitre 2.

Les courbes ROC obtenues selon ce procédé pour 6 images de la base de 732 photographies sont données en figure 7.2. Elles ont été obtenues en appliquant les distances CEMD (en rouge), L^1 (en bleu) et L^2 (en vert) à des descripteurs SIFT normalisés *globalement* (chacun des $M = 9$ histogrammes étant par conséquent de poids différents), et quantifiés sur $N = 12$ bins. Comme nous avons déjà pu le constater pour la comparaison de critères de mise en correspondance, les performances des différentes distances varient selon les images utilisées. En particulier, il existe des exemples où la distance L^2 est meilleure que la distance L^1, et réciproquement. C'est la raison pour laquelle il est préférable d'analyser

[1] Cette base est disponible à l'adresse : http://www.tsi.enst.fr/~rabin/matching/
[2] selon la taxonomie donnée en introduction du chapitre 2, dans le tableau 1.1

une mesure de dissimilarité sur une large base de données.

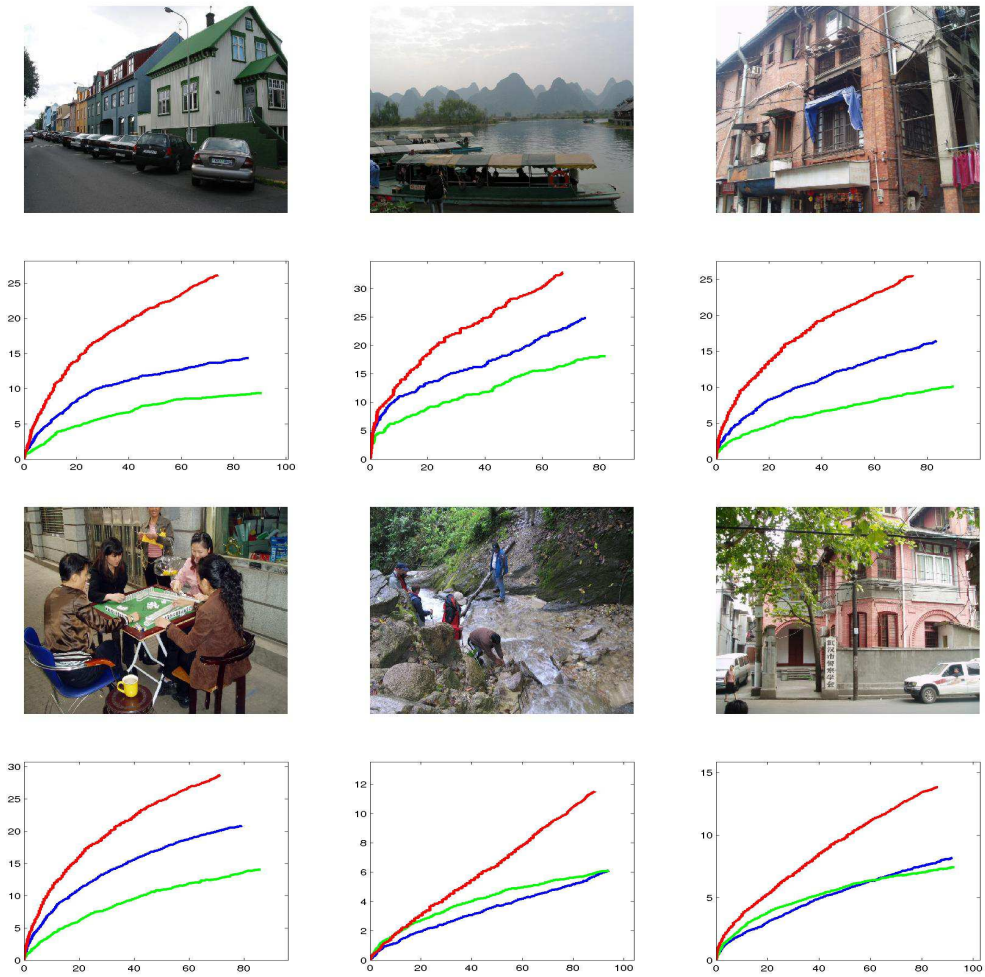


FIG. 7.2 – *Six échantillons d'images provenant de la base de données et les courbes ROC obtenues pour trois mesures de dissimilarité, calculées sur des descripteurs SIFT normalisés globalement. La courbe rouge correspond à la distance* CEMD*, la bleue à la distance L^1 et la verte à la distance euclidienne. Cette comparaison des différentes courbes sur quelques images illustre la variabilité des performances de chaque mesure de dissimilarité.*

Dans le but d'évaluer comparativement les performances moyennes obtenues par différentes mesures de dissimilarité sur notre base d'images, nous nous sommes inspirés des méthodes d'analyses en indexation d'images (voir par exemple [RTG00, LO07]) reposant sur des **courbes ROC moyennes**. En utilisant le protocole $A \rightarrow A'$, ces courbes permettent d'évaluer la capacité d'une mesure de dissimilarité à classer en premier le bon candidat parmi la base, ce qui est primordial pour la mise en correspondance. Nous définissons ici le taux *moyen* de correspondances correctes (équation (7.3.1)) comme la moyenne du taux de rappel pour un même taux de fausses détections dans chaque image A_i, pondérée par le nombre de descripteurs $N_{Q,i}$ de l'image A_i. Rappelons que le taux de rappel est défini comme le rapport du nombre de correspondances correctes sur le nombre total de correspondances correctes possibles, ce qui nous donne l'expression du taux moyen :

$$\text{Taux moyen de rappel} = \frac{1}{\sum_{i=1}^{732} N_{Q,i}} \sum_{i=1}^{732} \left(N_{Q,i} \frac{\#\text{correspondances correctes}(A_i)}{\#\text{correspondances correctes possibles}(A_i)} \right) ,$$

où les quantités $\#\text{correspondances correctes}(A_i)$ sont évaluées pour un même taux de fausses détections (*i.e.* une même proportion du nombre de correspondances incorrectes parmi celles sélectionnées). Le fait

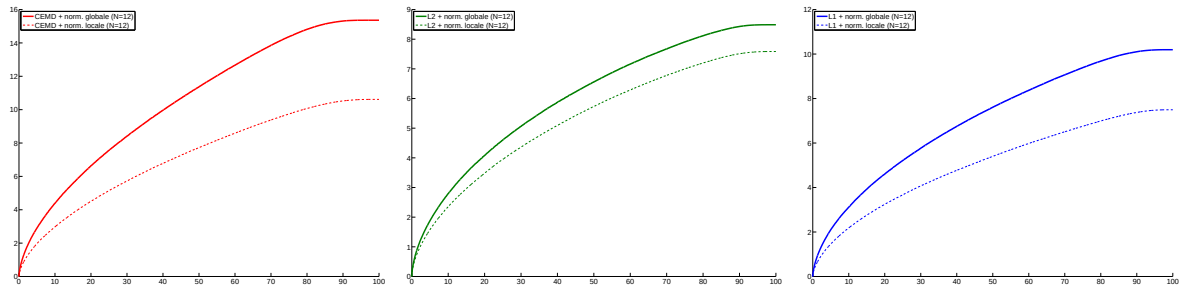
de calculer une moyenne pondérée plutôt qu'un moyenne simple permet d'éviter théoriquement qu'une image avec très peu de points d'intérêt vienne fausser les résultats (par analogie avec le nombre d'images par classe en indexation). En pratique cependant, il n'y a pas de différence significative sur notre base. Une courbe ROC moyenne est obtenue en traçant le taux moyen de rappel en fonction du taux de fausses détections.

Pour l'ensemble des expériences qui vont suivre, les descripteurs SIFT sont construits avec $M = 9$ histogrammes quantifiés sur $N = 8$ ou 12 bins.

7.3.2 Performances selon la normalisation

Cette première expérience a pour but d'illustrer le fait que le choix de la normalisation pour les SIFTs est très important pour la distance CEMD, comme pour les distances bin-à-bin. Rappelons qu'un descripteur a est normalisé *globalement* lorsque que $\sum_{m=1}^{M} ||a_m||_1 = 1$, et qu'il est normalisé *localement* lorsque $||a_m||_1 = 1 \, \forall m = 1, \ldots, M$.

La figure 7.3 montre les courbes de performance moyenne obtenues pour les distances L^2, L^1 et la mesure de dissimilarité D_{CEMD}, chacune pour les deux types de normalisation, avec $N = 12$ bins par histogramme.



(a) D_{CEMD} (b) L^2 (c) L^1

FIG. 7.3 – *Courbes ROC moyennes sur* 732 *images pour les mesures de dissimilarité* D_{CEMD} *en rouge,* L^2 *en vert,* L^1 *en bleu, avec une quantification de* $N = 12$*. Deux normalisations différentes sont utilisées :* locale *pour les courbes en trait interrompu, et* globale *pour les courbes en trait continu. Quelle que soit la mesure de dissimilarité utilisée, la normalisation globale donne systématiquement de meilleurs résultats.*

En premier lieu, on retrouve bien le fait que la distance euclidienne (figure 7.3(b)) donne de meilleurs résultats avec la normalisation globale, confirmant ainsi l'analyse faite par Lowe dans [Low04]. La mesure de dissimilarité L^1 (figure 7.3(c)) donne des résultats analogues. Cela montre qu'en choisissant une normalisation locale, on perd en robustesse ce que l'on gagne en invariance : la normalisation locale autorise des changements locaux affines de contraste pour chacune des M régions de la grille de localisation, tandis que la normalisation globale limite l'invariance au changement de contraste affine pour l'ensemble des régions.

Les courbes de performance moyenne en figure 7.3(a) montrent pour la mesure de dissimilarité D_{CEMD} le même phénomène : les performances sont meilleures avec la normalisation globale. Pour l'ensemble des expériences suivantes, la distance D_{CEMD}, tout comme les autres mesures de dissimilarité utilisées pour les analyses comparatives, sera calculée en utilisant la normalisation globale des SIFTs.

Remarques 2 :
- Les courbes de performance obtenues par Werman et Pele dans [PW08] suggèrent le même résultat (voir les figures 4 à 7). La distance de transport $\hat{EMD}$ correspondant à la normalisation locale (notée EMD_{MOD}) donne systématiquement de moins bons résultats que la distance obtenue avec une normalisation globale (appelée SIFT_{DIST}).
- Par souci de clarté nous n'avons montré que les distances L^2 ou L^1, mais nous avons observé exactement le même phénomène pour l'ensemble des distances bin-à-bin qui sont utilisés au paragraphe suivant.

7.3.3 Comparaison de D_{CEMD} avec les distances bin-à-bin

Dans ce paragraphe, nous comparons les performances de la mesure de dissimilarité D_{CEMD} avec celles des différentes distances bin-à-bin introduites à la section 7.1.1 : distance L^1 (Manhattan), distance L^2 (euclidienne), divergence de Jeffrey, et distance du χ^2. La normalisation globale sur les SIFTs est employée pour toutes ces mesures. En plus des perturbations apportées aux histogrammes d'orientation du gradient des SIFTs par la déformation géométrique (transformation affine) et le bruit entre les images A et A', nous utilisons deux niveaux de quantification pour la comparaison des distances : $N = 8$ et $N = 12$ bins. La figure 7.4 montre l'ensemble de ces 10 courbes ROC moyennes obtenues à partir des 732 images de la base de données. L'aspect très « lisse » des courbes s'explique par le nombre de comparaisons de descripteurs réalisées (approximativement 25.10^9 paires de descripteurs).

Remarque 3 :
L'ensemble des distances utilisent une normalisation globale selon L^1 pour la comparaison des descripteurs SIFT, à l'exception de la distance euclidienne qui utilise une normalisation globale selon L^2.

La figure 7.4 montre clairement la supériorité de D_{CEMD} par rapport aux distances bin-à-bin usuelles, et ce, pour tous les choix de quantification. Comme on pouvait s'y attendre, cette mesure de dissimilarité « inter-bins » est beaucoup plus robuste que les distances bin-à-bin aux déformations géométriques appliquées aux images (transformation affine simulant le changement de point de vue), qui se traduisent localement par des petites translations dans les histogrammes d'orientation du gradient des descripteurs SIFT. Par ailleurs, il est intéressant de constater que lorsque le nombre de bins est plus important, les performances moyennes sont améliorées avec la mesure de dissimilarité D_{CEMD}, car le pouvoir discriminant des descripteurs SIFT s'en trouve accru. Ce n'est pas le cas des distances bin-à-bin qui sont par contre très sensibles à la quantification. En effet dans [Low04], Lowe montre que la distance euclidienne donne une performance optimale pour $N = 8$ bins, ce que l'on corrobore une nouvelle fois ici. L'utilisation de la distance CEMD permet donc de s'affranchir de ce compromis entre pouvoir discriminant et robustesse à la quantification, le choix du nombre de bins n'étant alors régi que par des considérations de temps de calcul.

Remarques 4 :
- Les courbes ROC moyennes pour la divergence de Jeffrey, la distance du χ^2 et la distance L^1 sont quasiment confondues pour $N = 8$ bins.
- Les performances de la distance L^2 sont clairement en retrait par rapport aux autres distances bin-à-bin, notamment vis-à-vis de la distance du χ^2, ce qui corrobore la remarque de Lowe dans [FL07].
- Par analogie avec la partie expérimentale du chapitre 2, nous avons ici utilisé un protocole identique à celui intitulé $A \rightarrow A'$. Cependant, d'autres protocoles auraient pu tout aussi bien être utilisés, mais les courbes de performance sont similaires car ces protocoles conçus pour l'analyse des performances de la mise en correspondance n'apportent rien dans un cadre de comparaison de mesure de dissimilarité. Toutefois, à titre de comparaison, nous donnons la figure 7.5 correspondant au protocole $A \rightarrow \{^{A'}_{B}$: dans ce protocole, l'image A est à la fois comparée à A' et à une image B différente.

7.3.4 Comparaison de D_{CEMD} avec l'EMD

Nous avons vu au paragraphe 7.1.2 que la distance EMD pouvait être utilisée pour comparer directement les descripteurs SIFT en tant qu'histogrammes 3D, combinant à la fois l'information d'orientation du gradient mais également une information spatiale sur la position relative du gradient selon une grille

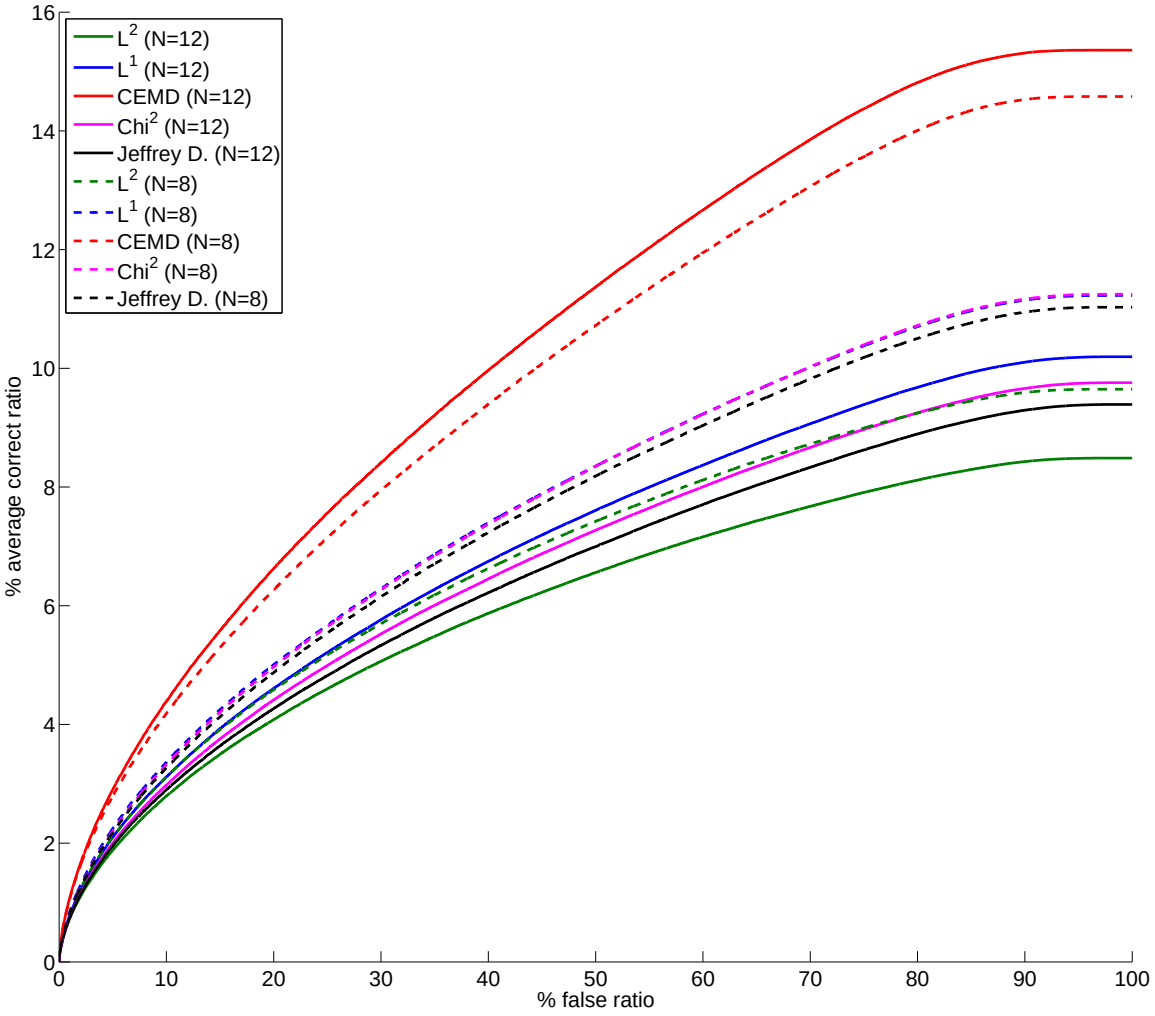
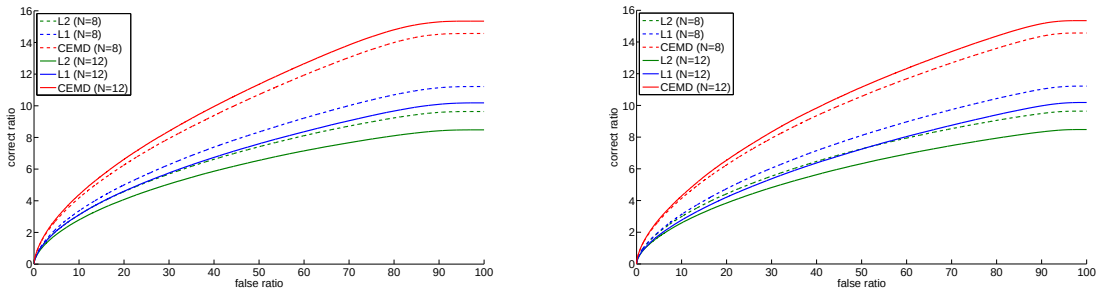


FIG. 7.4 – *Courbes ROC moyennes sur* 732 *images et* 3.1 *millions de descripteurs pour les mesures de dissimilarité D_{CEMD} en rouge, L^1 en bleu, L^2 en vert, χ^2 en magenta et la divergence de Jeffrey en noir. Deux quantifications différentes sont utilisées : $N = 8$ bins pour les courbes en trait interrompu, et $N = 12$ pour les courbes en trait continu.*



(a) protocole $A \rightarrow A'$ (courbes provenant de la figure 7.4)

(b) protocole $A \rightarrow \{^{A'}_{B}$

FIG. 7.5 – *Comparaison des courbes ROC moyennes avec les protocoles $A \rightarrow A'$ (figure 7.5(a)) et $A \rightarrow \{^{A'}_{B}$ (figure 7.5(b)) pour les mesures de dissimilarité D_{CEMD} en rouge, L^1 en bleu, et L^2 en vert. Deux quantifications différentes sont utilisées : $N = 8$ bins pour les courbes en trait interrompu, et $N = 12$ pour les courbes en trait continu.*

de localisation. Une telle comparaison est néanmoins coûteuse en temps de calcul (voir le tableau 7.1). Cependant, il est légitime de se demander si le fait de comparer des histogrammes 1D avec CEMD, et donc de perdre l'information spatiale, limite ses performances vis à vis de la comparaison tridimensionnelle proposée par Ling *et al.* dans [LO07]. Nous reprenons leur notation « EMD-L^1 » pour désigner l'utilisation de l'EMD sur des SIFTs avec une distance au sol L^1. Cette notation est néanmoins abusive car l'une des dimensions de l'histogramme 3D est circulaire (orientation du gradient) ; par conséquent la distance au sol doit prendre théoriquement en compte cette circularité pour que le transport ait du sens. Par ailleurs, la définition de la distance au sol pour les SIFTs qui combinent des données hétérogènes (deux dimensions spatiales dans le plan, et une dimension circulaire) n'est pas explicité dans [LO07]. On suppose ici que la distance au sol est définie comme la somme de la distance normalisée sur le cercle (notée d_θ) pour la dimension circulaire, et de la distance L^1 normalisée (notée $d_\mathcal{G}$) pour les dimensions spatiales.

Soient $x = (p_x, \theta_x)$ et $y = (p_y, \theta_y)$ les vecteurs de position de deux bins dans un descripteur SIFT considéré comme un histogramme 3D, où p_x et p_y désignent la position spatiale de l'histogramme d'orientation du gradient calculé sur la grille de localisation (rappelée en figure 7.6(a)), et θ_x et θ_y représentent l'orientation quantifiée dans l'histogramme d'orientation. La distance normalisée sur le cercle est définie comme :

$$\forall \theta_x, \theta_y \in [0; 2\pi[\ , d_\theta(\theta_x, \theta_y) := \frac{1}{\pi} \min(|\theta_x - \theta_y|, 2\pi - |\theta_x - \theta_y|) \in [0; 1]$$

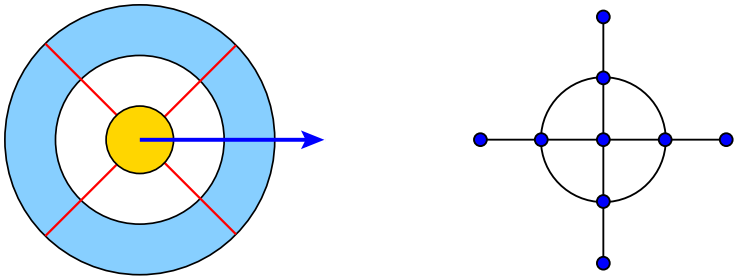
Les positions spatiales p_x et p_y sont définies selon la grille de localisation, dont on donne le graphe en figure 7.6(b) (les disques bleus symbolisent les différentes positions possibles). Nous avons choisi de définir $d_\mathcal{G}$ comme la distance sur ce graphe, c'est-à-dire le nombre d'arcs entre deux positions. Pour obtenir une distance normalisée (*i.e.* dans $[0; 1]$), on divise cette distance par la distance maximale sur le graphe, soit ici 4. Ainsi, la fonction de coût de transport (ou distance au sol) d'une masse unité de la position x à la position y est définie de la manière suivante pour EMD-L^1 :

$$c_{\text{EMD-}L^1}(x, y) := \frac{1}{\pi} \min(|\theta_x - \theta_y|, 2\pi - |\theta_x - \theta_y|) + \frac{1}{4} d_\mathcal{G}(p_x, p_y) \tag{7.12}$$

Nous avons utilisé le code source proposé par Rubner [Rub] (fondé sur l'algorithme du simplexe) pour calculer l'EMD entre les descripteurs SIFT avec le coût défini par la formule (7.12). Étant donné la complexité de calcul (plus de 1000 fois plus lent que D_{CEMD} avec une quantification de $N = 12$ bins), nous avons utilisé un échantillon de 10 images de la base seulement pour réaliser les courbes ROC moyennes.

> **Remarque 5 :**
> D'autres combinaisons et d'autres définitions de d_θ et $d_\mathcal{G}$ auraient pu être choisies. Nous avons cherché à définir avec la formule (7.12) la distance au sol qui se rapproche le plus d'une distance définie comme « L^1 » dans [LO07].

En figure 7.7 sont tracées les courbes ROC moyennes pour les mesures de similarité D_{CEMD} en rouge, L^1 en bleu et EMD-L^1 en cyan. Ces courbes sont obtenues pour deux quantifications ($N = 8$ bins en trait interrompu, et $N = 12$ en trait continu) avec une normalisation globale des SIFTs (norme L^1). Tout comme la mesure D_{CEMD} fondée sur le transport, EMD-L^1 donne de meilleures performances lorsque la quantification est plus fine, ce à quoi l'on pouvait s'attendre. Néanmoins, les performances de EMD-L^1 sont nettement en deçà de celles de D_{CEMD}, et assez proches de la distance L^1. Cela semble ainsi suggérer que, pour le coût défini en formule (7.12), l'information spatiale n'est finalement pas importante pour la comparaison des SIFTs, voire même qu'elle diminue les performances de cette mesure de dissimilarité. En effet, rappelons que les distances bin-à-bin n'exploitent pas cette information, et pourtant avec une quantification de $N = 8$ bins, la distance L^1 est plus robuste que la distance EMD-L^1. Werman et Pele font le même constat : EMD-L^1 est moins performante que toutes les autres distances utilisées (voir les figures 4 à 7 dans [PW08]). Cela signifie donc qu'utiliser une distance au sol mélangeant à la fois la position spatiale des histogrammes et les orientations du gradient n'a pas vraiment de sens pour la comparaison de SIFTs.



(a) Masque circulaire découpé en 9 régions.

(b) Graphe utilisé pour le calcul de la distance entre les régions.

FIG. 7.6 – *Illustration du masque et du graphe associé d'un descripteur SIFT. (Figure de gauche) Masque circulaire découpé en* 9 *régions distinctes, utilisé pour l'extraction des histogrammes d'orientation du gradient. (Figure de droite) Graphe de relation entre les régions.*

7.3.5 Performances selon la perturbation

Dans les expériences précédentes, la transformation affine permettant d'obtenir l'image A' était fixée avec un paramètre de tilt égal à 2.5. Afin d'illustrer l'influence de la transformation affine sur les performances de la mesure de dissimilarité, le protocole $A \rightarrow A'$ est réalisé pour différentes valeurs de tilt. En effet, les auteurs de [MY09] ont montré que c'est le paramètre le plus critique concernant la robustesse de représentation locale des descripteurs SIFT. Nous estimons cette fois les courbes ROC moyennes sur plusieurs images pour différentes valeurs de tilt, les autres paramètres de la transformation affine étant fixés (à 0 pour les deux paramètres de rotation, et à 1 pour le paramètre d'échelle). On obtient ainsi plusieurs courbes ROC moyennes à partir de 65 images de la base pour 12 valeurs de tilt différentes, les descripteurs SIFT étant quantifiés sur $N = 12$ bins.

Afin de comparer ces 12 courbes –similaires à celles en figure 7.4–, on mesure la valeur de l'aire sous la courbe ROC en fonction du tilt, pour les mesures de dissimilarité D_{CEMD} et L^1. L'aire sous une courbe ROC est en effet un moyen d'évaluer la performance globale d'un classifieur : plus elle est proche de 1, plus la classification obtenue est parfaite. Nous traçons en figure 7.8 la valeur de l'aire sous la courbe ROC moyenne en fonction du tilt. Comme l'on pouvait s'y attendre, plus le facteur de tilt est grand, plus les performances moyennes diminuent pour les deux distances. Cependant, les performances moyennes de la mesure de dissimilarité fondée sur D_{CEMD} décroissent moins vite qu'avec la distance L^1, ce qui illustre une nouvelle fois la plus grande robustesse de D_{CEMD} aux perturbations géométriques.

Pour une faible valeur de tilt (tilt $= 1.3$), l'aire sous la courbe ROC moyenne est environ de 0.995 pour la distance L^1, alors que pour la mesure de dissimilarité fondée sur CEMD, l'aire est plus petite : environ 0.975. Si en pratique cette différence ne correspond qu'à quelques fausses détections supplémentaires et est par conséquent assez négligeable, ce résultat n'en est pas moins étonnant. En effet, pour toutes les autres valeurs de tilt, les performances de D_{CEMD} sont supérieures à L^1.

Pour expliquer ce résultat, considérons le résultat de l'analyse réalisée au chapitre précédent. Nous avons montré que la distance de transport EMD est plus robuste que les distances bin-à-bin pour des perturbations liées à la quantification ou des perturbations sur les positions des modes principaux. Néanmoins, nous avons vu que les distances bin-à-bin sont plus robustes que l'EMD aux variations de poids des modes principaux des histogrammes comparés. Suivant le type de perturbation prépondérant, il a été montré pour une application d'indexation d'images que l'une ou l'autre de ces distances pouvait donner de meilleures performances. Dans le cas des descripteurs SIFT, une transformation affine se traduit par deux perturbations sur les structures géométriques détectées (typiquement une jonction entre deux bords) : un changement d'angle qui correspond à une translation dans les histogrammes d'orientation, et un changement de longueur des bords qui correspond à un changement de poids. Il est possible que pour de faibles valeurs de tilt le phénomène de variation de poids soit prépondérant, ce qui expliquerait pourquoi la distance L^1 est meilleure que D_{CEMD} dans ce cas.

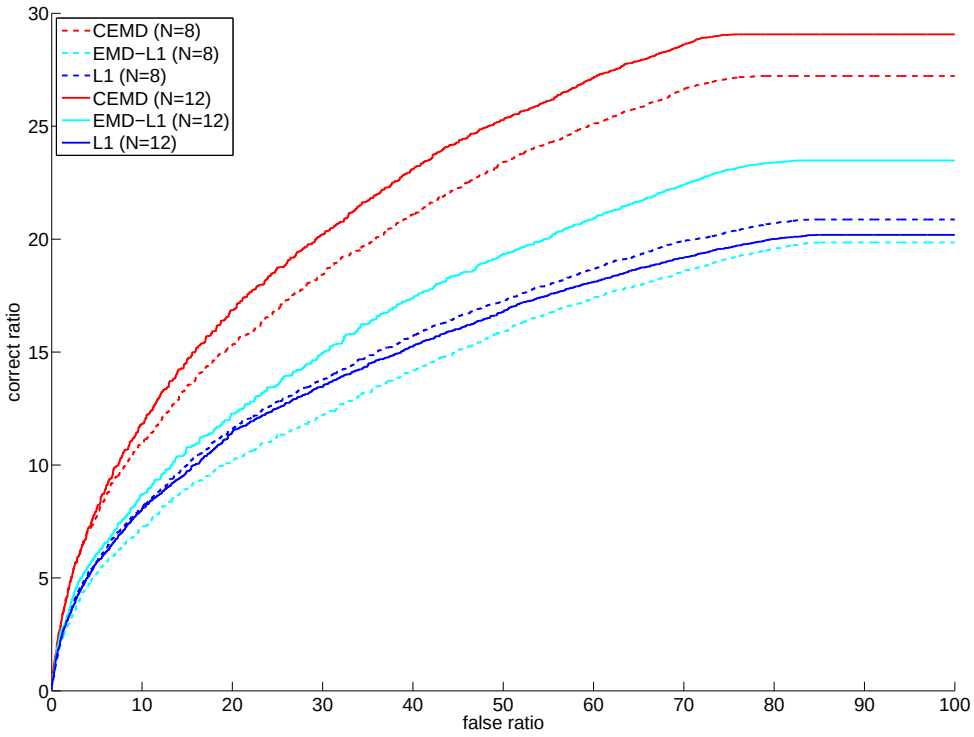


FIG. 7.7 – *Courbes ROC moyennes (sur* 10 *images) pour les distances* D_{CEMD} *(en rouge) et EMD tridimensionnel (en cyan), et* L^1 *(en bleu) avec deux quantifications différentes (*$N = 8$ *en trait interrompu et* $N = 12$ *en trait continu).*

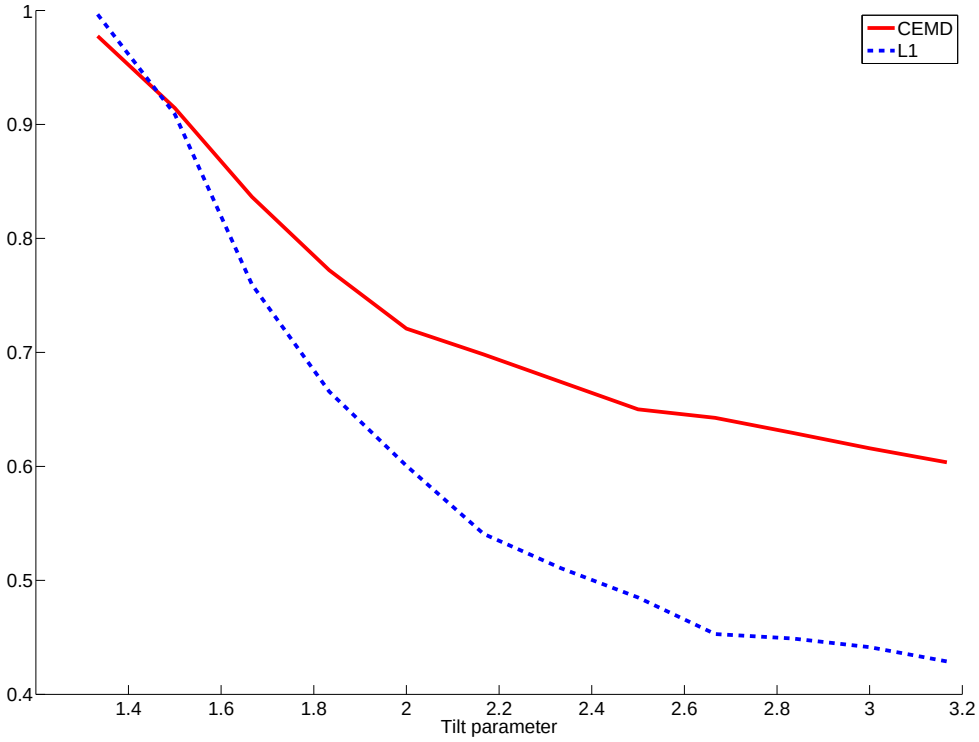


FIG. 7.8 – *Aire sous la courbe ROC (moyenne sur* 65 *images) pour* D_{CEMD} *(courbe rouge en trait continu) et* L^1 *(en trait bleu interrompu), avec* $N = 12$ *bins, selon différentes valeurs de tilt.*

Chapitre 8

Régularisation du transport pour le transfert de caractéristiques

Nous avons vu, dans les deux précédents chapitres, que le transport pouvait être utilisé pour la comparaison d'histogrammes. Dans ce dernier chapitre, nous nous intéressons cette fois aux applications du transport pour le transfert de caractéristiques. Par le terme de « transfert », nous entendons le fait de modifier une image dans le but de lui imposer certaines caractéristiques (contraste, domaine de couleur, *etc*.). Nous considérons ici les applications de transfert qui peuvent être vues comme un *transport* entre les distributions de caractéristiques de l'image à traiter et de l'image que l'on souhaite obtenir. Dans ce cadre de travail, nous verrons que les différentes approches qui ont été proposées par le passé produisent certains effets visuels indésirables. Nous proposons dans ce chapitre une méthode de régularisation du transport, appelée régularisation non locale de la carte de transport (*non local map regularization*), qui vise à améliorer le rendu de l'image obtenue (un exemple de régularisation du transport pour le transfert de couleurs est donné en figure 8.1).

(a) Image originale (Auguste Renoir, *Le déjeuner des Canotiers*, 1881).

(b) Image de style (Paul Gauguin, *Mahana no atua – le jour de dieu* , 1894).

(c) Transfert de la couleur, puis régularisation par notre méthode.

FIG. 8.1 – ***Illustration du transfert de palette.*** *Les deux premières figures (8.1(a) et 8.1(b)) représentent des tableaux ayant des palettes de couleurs différentes. Le principe du transfert de palette est de reproduire la première image en utilisant la palette de la seconde image. Ce transfert est défini par le transport optimal entre les histogrammes couleurs de ces deux images. La figure 8.1(c) est obtenue en utilisant une nouvelle méthode de régularisation du transport introduite dans ce chapitre.*

Les applications du transport pour le transfert de caractéristiques sont tout d'abord présentées en détails dans la section 8.1. Nous y décrivons également leurs principales limitations (diminution du rapport signal sur bruit, artefacts de compression, suppression de structures, incohérences spatiales ...). Une nouvelle méthode de régularisation du transport est ensuite présentée dans la section 8.2. Notre approche s'inspire du principe de filtrage par un opérateur de moyenne pondérée dans l'espace couleur des images, qui a été introduit dans le cadre du débruitage (*Bilateral filter* [TM98] et *Non Local Means* [BCM05] notamment). Nous illustrons dans une dernière section l'intérêt de notre approche pour les applications de transfert de palette et de réhaussement de contraste.

8.1 Présentation du problème

Nous nous intéressons dans cette section aux applications de transfert de caractéristiques dans le cadre du transport optimal. Deux méthodes en traitement d'images correspondent à ce cadre de travail : l'ajustement de contraste et le transfert de couleurs.

8.1.1 Spécification d'histogramme

Les méthodes d'ajustement automatique de contraste visent à modifier la dynamique du signal afin d'améliorer la visualisation des détails d'une photographie, ou encore d'harmoniser le contraste d'une série d'images. Les applications qui en découlent sont multiples : impression de photographies, visualisation en imagerie médicale, construction de panorama à partir de plusieurs images, restauration de films anciens, réduction de dynamique d'image (restriction des niveaux de gris), *etc*.

Soit l'image en niveau de gris $u : x \in \Omega \mapsto u(x) \in \mathbb{R}^+$, où Ω désigne le domaine de l'image u dont on souhaite modifier le contraste. Cette opération revient à définir une fonction f de $\mathbb{R}^+$ à valeur dans $\mathbb{R}^+$ que l'on doit appliquer à l'image u. Afin de ne pas modifier la géométrie de l'image u (suppression de structures, inversion de contraste), il est nécessaire que f soit une fonction strictement croissante. On note $h_u : \mathbb{R}^+ \mapsto \mathbb{R}^+$ la distribution des niveaux de gris de l'image u, et $H_u : \mathbb{R}^+ \mapsto [0, 1]$ la fonction de répartition correspondante.

Un moyen de définir l'opérateur f de changement de contraste global est de considérer le transport optimal. Soient h_T une distribution cible et H_T la fonction de répartition qui lui correspond. On cherche l'opérateur f tel que la distribution de l'image $f(u)$ soit égale à la distribution cible : $h_{f(u)} = h_T$. L'application f représente le transport optimal qui permet de transformer la distribution h_u en h_t, définie de la manière suivante (voir par exemple [Vil03]) :

$$f = H_T^{-1} \circ H_u$$

où $H_T^{-1}(t) = \inf\{y \,|\, H_T(y) \geq t\}$ désigne la fonction pseudo-inverse de H_T, qui permet de définir le nouveau niveau de gris pour un pixel de rang t.

En pratique, cette approche est très simple à mettre en œuvre pour des images discrètes $u : x \in \Omega \subset \mathbb{Z}^2 \mapsto u(x) \in \{0, \ldots, 255\}$. Le transport f est alors défini de manière approchée à partir d'histogrammes, et l'on parle de *spécification d'histogramme*. L'histogramme cible h_T est spécifié par l'utilisateur, soit de manière explicite, soit à partir d'une autre image.

> **Remarque 1 :**
> Dans le cas continu, le transport optimal d'une distribution h vers g correspond à l'application de $f = G^{-1} \circ H$, à la condition que h et g soient sans atomes (absence de masses de Dirac). Dans le cas discret cependant, la définition équivalente de f ne correspond plus tout à fait au transport optimal : tous les pixels de même niveau de gris auront la même affectation, ce qui n'est pas nécessairement le cas pour le transport optimal. Cependant, le résultat obtenu est en pratique une excellente approximation du transport optimal.

De nombreuses extensions ont été proposées à partir de ce cadre de travail. Les différents traitements reposant sur la spécification d'histogramme peuvent être distingués en deux catégories, selon la manière dont est défini l'histogramme h_T.

Une application très courante consiste à spécifier explicitement la distribution h_T que l'on souhaite obtenir. En particulier, on parle d'*égalisation d'histogramme* lorsque h_T est un histogramme plat (distribution uniforme sur les 256 niveaux de gris). Dans ce cas, l'expression du transport devient $f = 255 \cdot H_u$. Dans [PAA+87] les auteurs proposent d'appliquer le principe d'égalisation d'histogramme localement, dans le cadre de l'imagerie médicale (rayons X, IRM, *etc*.) afin d'en faciliter l'interprétation. Pour cela, l'image est découpée en imagettes qui sont traitées indépendamment. Cependant, pour limiter le rehaussement de contraste excessif dans les zones homogènes, ils proposent par ailleurs de limiter le gradient du transport (*i.e.* la pente de l'histogramme cumulé de chaque imagette). Dans [CCM99, CLMS99], le contraste est modifié localement tout en préservant la carte topographique de l'image.

Une autre utilisation du transport consiste à définir l'histogramme cible h_T à partir d'une autre image. À titre d'exemple, ceci permet d'obtenir un rendu visuel plus naturel pour une séquence d'images (tirées d'un film ou bien d'une série de photographies destinées à la construction d'une mosaïque) représentant une même scène selon des éclairements et des conditions de prises de vue différentes. Dans le cas spécifique où l'on souhaite ajuster le contraste de deux images $u1$ et $u2$ d'une même scène, J. Delon montre dans [Del04] qu'il est préférable de définir h_T comme l'histogramme mi-chemin (*midway*) des histogrammes des images originales : $H_T^{-1} = \frac{1}{2}\left(H_{u1}^{-1} + H_{u2}^{-1}\right)$. Ce principe est également utilisé pour la restauration de films anciens (réduction des effets de papillonnement, ou *flicker*) dans [Del06, DD09].

Nous avons jusqu'à présent considéré le cas des images en niveaux de gris. Dans le cas des images couleurs (trois canaux RVB), les images sont représentées dans un autre espace (Yuv, Lab, HSV, HSI, ...). Seul le canal représentant l'intensité lumineuse est concerné par le changement de contraste, en utilisant le même principe que les précédentes méthodes exposées. Toutefois, quelques ajustements supplémentaires sur les deux autres canaux peuvent être requis pour préserver les couleurs (par exemple, voir [CLMS99]). Divers exemples d'égalisation d'histogramme pour des images (couleurs ou en niveau de gris) sont donnés en figure 8.2.

Dans le prochain paragraphe, nous nous intéressons cette fois au transport des trois canaux à la fois. Il ne s'agit alors plus de changement de contraste, et l'on parle de transfert de couleurs.

8.1.2 Le transfert de couleurs

On trouve dans la littérature de nombreuses méthodes de transfert de couleurs. Le principe de ces méthodes est de modifier une image u – appelée image *source* – de manière à lui transférer les caractéristiques couleurs d'une image v, appelée image de *style*.

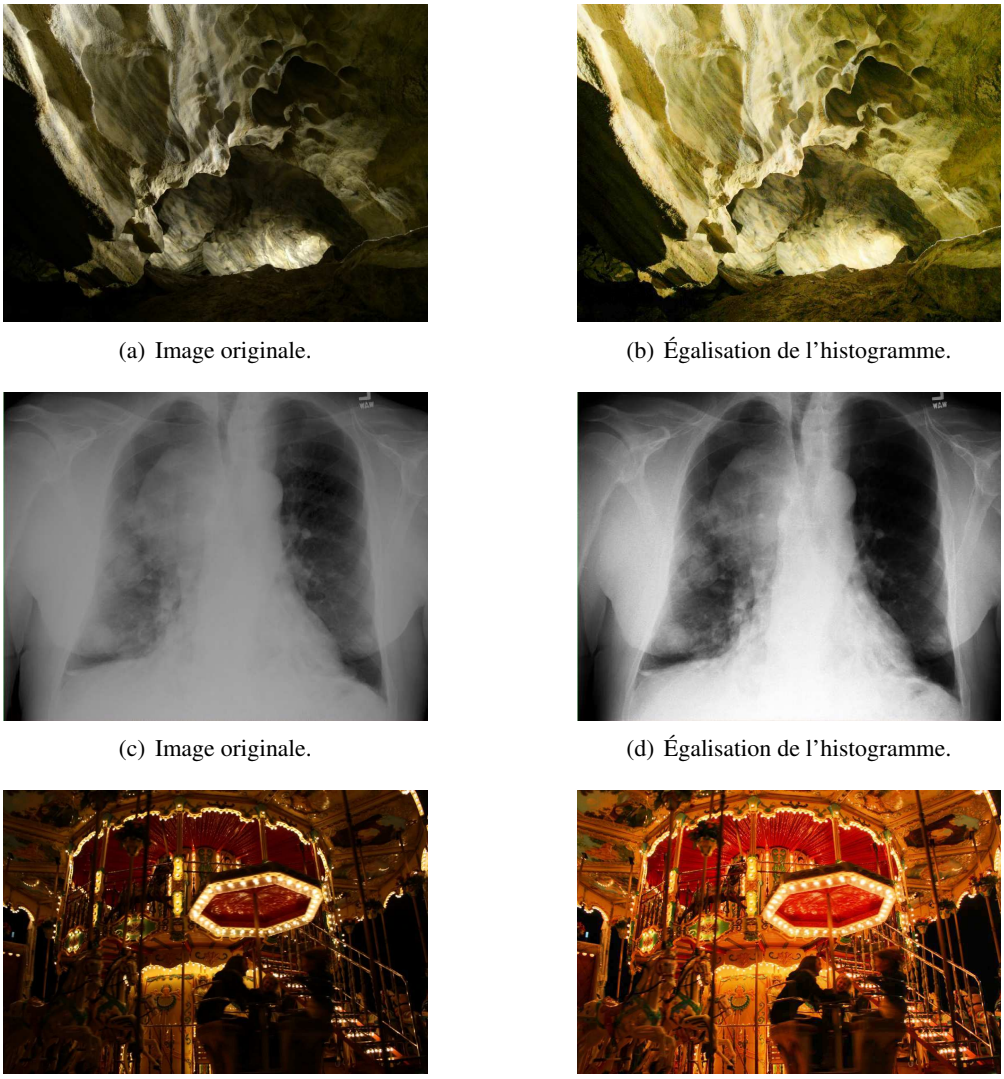
Les premiers à s'être intéressés à la problématique du transfert des caractéristiques couleurs sont, à notre connaissance, les auteurs de [RAGS01]. Leur idée est très simple : il s'agit d'appliquer une transformation affine de telle manière que la moyenne et la variance de l'image u coïncident avec celles de l'image v. Pour cela, il est tout d'abord nécessaire d'appliquer un changement de système de coordonnées : les images u et v, initialement en coordonnées RVB, sont alors exprimées dans un système (Lab, $l\alpha\beta$, HSV, HSI, *etc.*) où la luminosité est exprimée suivant une dimension indépendante des deux autres dimensions, lesquelles représentent alors des caractéristiques de la couleur (teinte et saturation par exemple). Les histogrammes empiriques h_u et h_v sont estimés dans cet espace, respectivement à partir de u et de v. On estime ensuite les moyennes $\{\mu_u^i, \mu_v^i\}$ et les écarts-types $\{\sigma_u^i, \sigma_v^i\}$ de chaque dimension i des distributions 3D h_u et h_v. En notant $u(x) = (u^1(x), u^2(x), u^3(x))^T$ les valeurs du pixel x de l'image u, on obtient la nouvelle image $f(u)$ par application de la transformation affine f suivante :

$$f(u)^i(x) = \frac{\sigma_v^i}{\sigma_u^i}\left(u^i(x) - \mu_u^i\right) + \mu_v^i \ .$$

Cette méthodologie est en fait une simple extension en 3D de la correction de contraste par étirement affine de l'histogramme en niveau de gris (*histogram stretching*). Le problème de ce type de transformation est qu'elle n'est efficace que pour des distributions relativement simples (loi normale par exemple). Lorsque les distributions sont plus complexes (comme c'est le cas en figure 8.3), ce type de transformation devient hasardeux et le résultat visuel est peu satisfaisant (saturation de luminosité, apparition de fausses couleurs, ou encore pas de changement notable).

Néanmoins, même si la méthode de [RAGS01] est naïve et demande en pratique de nombreuses interventions de la part de l'utilisateur, elle a contribué à populariser ce nouveau champ applicatif et de nombreuses approches ont été par la suite proposées. À titre d'exemple, les méthodes de transfert automatique utilisées dans [Kot05, GH03] s'inspirent de cette approche.

Une autre catégorie de méthodes de transfert a ensuite été proposée, formalisant le transfert de couleurs en un problème de transport optimal. Elle consiste à étendre aux cas des images couleurs le principe



(a) Image originale. (b) Égalisation de l'histogramme.

(c) Image originale. (d) Égalisation de l'histogramme.

(e) Image originale. (f) Égalisation de l'histogramme.

FIG. 8.2 – ***Illustration de l'égalisation d'histogramme.*** *La première colonne montre les images originales, et la seconde les images obtenues après égalisation de l'histogramme de luminosité (dans le cas présent, le canal 'V' de la représentation HSV).*

(a) Image source. (b) Image de style. (c) Transfert par transformation affine [RAGS01].

FIG. 8.3 – ***Illustration du transfert de couleurs par transformation affine.*** *L'image 8.3(c) est obtenue par la méthode de transfert de couleurs proposée dans [RAGS01], appliquée à l'image source 8.3(a) en considérant l'image 8.3(b) comme une image de style. Le résultat est peu satisfaisant, les couleurs obtenues ne correspondant pas à celle de l'image de style.*

présenté au paragraphe précédent, ce qui revient à considérer le transport f entre des histogrammes tridimensionnels. Cependant, sa mise en œuvre tout autant que l'effet produit sont radicalement différents. Intéressons nous dans un premier temps à cet effet.

Rendu visuel Tout comme la spécification d'histogramme, le transport optimal va dépendre de l'histogramme couleur h_u de l'image source u et de h_T, un histogramme cible explicitement défini par l'utilisateur, ou bien défini à partir d'une image de style v : $h_T = h_v$. Le fait de définir le transfert de couleurs comme un problème de transport impose que l'image obtenue $f(u)$ soit telle que son histogramme $h_{f(u)}$ soit identique à l'histogramme spécifié h_T. Ce résultat peut alors être interprété de la manière suivante : *étant donné une palette de couleurs définie par h_T, on souhaite « repeindre » une image u en utilisant exclusivement les couleurs de h_T, dans les mêmes proportions.*

Par analogie avec la spécification d'histogramme, deux solutions sont envisageables pour définir h_T. Une première solution consiste à étendre la notion d'égalisation d'histogramme aux cas des images couleurs. Étant donné h_T un histogramme 3D uniforme, on obtient une image $f(u)$ utilisant toutes les couleurs possibles (avec une dynamique plus élevée que pour les niveaux de gris : 24 bits au lieu de 8), ce qui se traduit par un rendu artificiel. Une mise en œuvre de ce principe est suggérée dans [PNS03], où l'égalisation couleur est utilisée pour permettre la visualisation d'images non naturelles (imagerie par microscopie électronique par exemple). Toutefois, en dehors de cette application très spécifique, cette approche de transfert de couleurs présente un intérêt limité.

La seconde approche, qui consiste à modifier une image couleur u en prenant exemple sur une autre image v, a en revanche suscité récemment un intérêt bien plus marqué. Nous allons dans le paragraphe suivant présenter les différentes solutions qui ont été proposées pour sa mise en œuvre.

Mise en œuvre Contrairement au cas 1D, il n'existe pas de solution analytique au problème du transport optimal en 3D. Une solution consiste à déterminer le transport optimal à l'aide d'un algorithme itératif (l'algorithme du simplexe par exemple), à l'image de [MS03]. Cependant, en raison de la taille des histogrammes couleurs (2^{24} bins), la recherche du transport optimal est très coûteuse en temps de calcul.

Dans [NN05], Neumann et Neumann proposent de définir le transport de *manière approchée*. Leur idée est d'appliquer le transport successivement sur chaque canal, dans un ordre précis, conditionnellement aux précédents canaux. On note $h_u(x_1, x_2, x_3)$ l'histogramme 3D de l'image à modifier u, et h_v l'histogramme de l'image v utilisée comme modèle. On suppose pour simplifier que h_u et h_v sont des distributions continues sur $\mathbb{R}^3$, les trois variables (x_1, x_2, x_3) représentant chacun des canaux de u. Les transports sur chacun des canaux sont définis de la manière suivante, en estimant à chaque fois les histogrammes empiriques correspondant pour chacune des images :

1. définition du transport $f_1(x_1)$ selon les valeurs de x_1 ;
2. définition du transport $f_2(x_2|x_1)$ selon les valeurs de x_2, conditionnellement à la valeur prise par le premier canal ;
3. définition du transport $f_3(x_3|x_1, x_2)$ selon les valeurs de x_3, conditionnellement à la valeur prise par les deux autres canaux.

La nouvelle image $f(u)$ est obtenue en appliquant le transport sur chaque canal :

$$f(u(x)) = \Big(f_1\big(u^1(x)\big)\ ,\ f_2\big(u^2(x)|u^1(x)\big)\ ,\ f_3\big(u^3(x)|u^1(x), u^2(x)\big) \Big)^T .$$

La solution obtenue dépend donc fortement de l'ordre dans lequel sont transportés chacun des histogrammes.

Contrairement à l'égalisation d'histogramme pour l'ajustement du contraste, cet algorithme repose sur le transport 1D d'un très grand nombre d'histogrammes. En effet, pour une image dont chaque canal est quantifié sur n bits, le transport f_3 du dernier canal requiert la définition de 2^{2n} histogrammes cumulés conditionnels, soit 65536 pour une image standard RVB codée sur 8 bits/canal. Pour être mis

en œuvre avec un temps de calcul raisonnable, Neumann et Neumann suggèrent en pratique d'utiliser une quantification plus grossière pour la construction des histogrammes. De plus, certaines fonctions $f_2(.|x_1)$ et $f_3(.|x_1, x_2)$ ne sont potentiellement pas définies en raison de la différence des distributions des images u et v. Les auteurs proposent d'utiliser une carte de transport par défaut (égalisation) pour ces cas de figure et, par ailleurs, de lisser les histogrammes de h_u et h_v pour réduire le nombre de pixels concernés par ce problème. Cet algorithme, défini de manière *ad-hoc*, ne garantit en rien que la solution obtenue soit proche du transport optimal, ni même que l'histogramme de l'image $f(u)$ soit identique à celui de v. Les différents exemples présentés dans [NN05] montrent en effet que les couleurs obtenues après transfert ne sont pas fidèles aux couleurs de l'image de style.

Pitié *et al.* ont proposé dans [PKD07] une solution alternative intéressante au problème du transport approché. Leur idée de départ est similaire à celle de [NN05] : plutôt que de définir explicitement le transport 3D, le transport est estimé successivement sur des histogrammes 1D à partir des histogrammes 3D des images u et v. L'originalité de leur approche réside dans le fait que les histogrammes 1D sont des projections orthogonales aléatoires des histogrammes h_u et h_v. L'algorithme est itéré pour que le transport converge vers une solution où $h_{f(u)} = h_v$, comme cela est décrit par le tableau 8.1. Cet algorithme est appelé IDT (*Iterative Distribution Transfer*).

TAB. 8.1 – *Vue d'ensemble de l'algorithme de transfert de couleurs [PKD07].*

Algorithme 8.1 IDT (*Iterative Distribution Transfer*)

Entrées : Image source u, image de style v, et calcul de l'histogramme discret h_v.

Initialisation : $k := 0$ et $u^{(k)} := u$.

1) **Rotation aléatoire :** Tirage de trois axes orthogonaux $R = [r_1, r_2, r_3]$ et calcul de l'histogramme h_u de l'image $u^{(k)}$; calcul de h_u^i et h_v^i, projections de h_u et h_v suivant chacun des r_i.

2) **Transfert 1D :** Pour chaque axe r_i, définir les histogrammes cumulés correspondants H_u^i et H_v^i puis les fonctions de transfert 1D $f_i = (H_v^i)^{(-1)} \circ H_u^i$, $\forall i = 1, 2, 3$.

3) **Transfert itératif 3D :** Définition de la fonction de transfert f 3D à partir de $\{f_i\}_{i=1,2,3}$.
Application à l'image $u^{(k)}$: $u^{(k+1)} = R\, f\left(R^T\, u^{(k)}\right)$.

4) **Critère d'arrêt :** tant que $k < k_{max}$, $k := k + 1$. Retour à l'étape 1).

Sortie : Nouvelle image $u^{k_{max}}$.

Pitié *et al.* montrent que cette procédure converge presque sûrement dans le cas où h_v est une distribution normale. En pratique, on observe expérimentalement que l'histogramme de l'image transformée correspond bien à l'histogramme de l'image v, au terme de plusieurs dizaines d'itérations. Par contre, aucune analyse n'a été faite pour mesurer la proximité de la solution obtenue vis-à-vis de la solution optimale au problème de transport de Monge-Kantorovitch.

Différents exemples d'application de l'algorithme IDT pour le transfert de couleurs sont donnés en figure 8.4. Au contraire des précédentes approches de transfert de couleurs, l'image obtenue possède exactement la même palette de couleurs que l'image de style. Cependant, cette méthode possède plusieurs limitations que nous étudions dans la section 8.1.3.

Pour finir, rappelons que le transfert de caractéristiques concerne aussi d'autres applications. Une première application consiste en la réduction de la dynamique des images couleurs (*gamut mapping*) pour la visualisation et l'impression de photographies (voir par exemple [BSL08, DD02]). Efros *et al.* ont introduit une méthode automatique de transfert de texture par copie de patchs [EF01]. Perez *et al.* ont proposé une méthode supervisée de transfert de texture et de couleur [PGB03]. On citera également les travaux sur le transfert générique de style présentés dans [HJO+01] qui ouvrent des perspectives intéressantes.

(a) Image originale. (b) Image de style couleur. (c) Transfert de la couleur par l'algorithme IDT.

(d) Image originale. (e) Image de style couleur. (f) Transfert de la couleur par l'algorithme IDT.

FIG. 8.4 – *Illustration du transfert de style de couleur.* *La première colonne montre les images originales, et la seconde les images de style utilisées. En troisième colonne sont placées les images de transfert de couleurs obtenues par la méthode IDT de [PKD07].*

Plus récemment, des travaux ont été réalisés sur le transfert de caractéristiques entre les photographies avec et sans flash [ED04, PSA+04], sur lesquels nous reviendrons ultérieurement. On notera enfin les applications de la synthèse par l'exemple, qui consistent à remplacer les données manquantes dans une image (*inpainting*) [CPT04, PGB04], ou encore à synthétiser une texture (voir par exemple [Pey09]).

> **Remarques 2 :**
> Il est important de noter que de nombreux outils ont été proposés pour la colorisation d'images en noir et blanc, application différente de celle que nous venons d'étudier, mais qui présente par certains aspects quelques similarités. Cette application vise à mettre en couleur des images ou des films en noir et blanc, le plus souvent de manière supervisée. Ainsi, en s'inspirant des travaux de [RAGS01], les auteurs de [WAM02] proposent une approche supervisée de colorisation d'image noir et blanc. Dans [YK03], une méthode de colorisation de film infra-rouge est décrite. Plus récemment, Charpiat *et al.* ont introduit une approche de colorisation automatique fondée sur la similarité de descripteurs locaux [CHS08]. Des approches de colorisation directes (sans utilisation d'image de style) ont également été proposées, en particulier [YS06]. Une approche de ce type a ensuite été reprise pour la colorisation d'autres styles d'images, tel que le dessin et la peinture [SBv04, QWH06].

8.1.3 Limitations du transport

Nous avons présenté avec les figures 8.2 et 8.4 des images pour lesquelles le traitement appliqué (égalisation d'histogramme et transfert de couleurs) donne un résultat satisfaisant. En pratique cependant, le rendu visuel est généralement de mauvaise qualité, principalement en raison de l'utilisation du transport optimal comme solution au problème du transfert de caractéristiques. En effet, le transport étant réalisé sans prendre en compte la cohérence spatiale des pixels de l'image source, l'aspect de l'image obtenue est « artificiel ». Nous allons décrire dans les deux paragraphes suivants les effets et les causes de ces limitations du transport.

Apparition d'artefacts et augmentation du bruit La principale conséquence malheureuse –et bien connue– de l'égalisation d'histogramme est l'augmentation du niveau de bruit (ou réduction du rapport signal à bruit, noté SNR en anglais pour *signal-to-noise ratio*). Cela se produit lorsque le transport implique localement un rehaussement du contraste, comme cela est illustré par les figures 8.5(b) et 8.5(d). Cet effet est conjugué en pratique à l'apparition d'artefacts en raison de la compression des images utilisées. Un exemple est donné en figure 8.5(f), où des artefacts (fausses couleurs, apparition de blocs) dus à la compression JPEG sont nettement amplifiés par l'égalisation d'histogramme.

Le même phénomène se produit lorsque l'on considère le transfert de couleurs, et est d'autant plus remarquable que la palette de couleurs de l'image de style est plus riche que celle de l'image source. Le problème du rehaussement du bruit est visible sur les exemples précédemment donnés en figure 8.4 pour l'illustration du transfert de couleurs par l'algorithme IDT [PKD07]. L'effet produit lorsque l'on utilise des images compressées est encore plus visible, ainsi que l'illustrent les exemples de la figure 8.6.

Ces phénomènes ne se limitent pas aux approches fondées sur le transport. Le problème du rehaussement du rapport signal à bruit concerne toutes les méthodes fondées sur la manipulation d'histogramme, dont notamment la méthode de [RAGS01]. De même, le problème de l'apparition d'artefacts de compression peut être rencontré pour d'autres applications du transfert de caractéristiques. Par exemple, avec l'approche de transfert de texture présentée dans [BPD06], les auteurs évoquent l'apparition d'artefacts dus à la compression.

Le problème de la proportion des couleurs de la palette Nous avons étudié au chapitre 6 l'intérêt du transport optimal pour la comparaison d'histogrammes dans le cas d'un mélange de deux gaussiennes. Nous avons conclu que la limitation intrinsèque du transport optimal était lié au problème de la différence relative des poids de chaque mode. La raison en est que l'excédent de masse relatif au premier mode est transporté sur le second mode, ce qui augmente fortement la dissimilarité intra-classe pour la comparaison d'histogrammes.

Dans le cadre du transport optimal, rappelons que nous avons interprété le transfert de couleurs comme le fait de « peindre » l'image source en utilisant les couleurs de l'image de style, *dans les mêmes quantités*. Considérons maintenant le cas où l'image source u et l'image de style v ont des compositions similaires, par exemple, un ciel et une forêt, dont la distribution peut être assimilée à un mélange de deux gaussiennes. Dès que les deux composantes sont de tailles relatives différentes dans les deux images, on retrouve le phénomène précédemment décrit : l'excédent de couleur de la composante principale de l'image de style est distribué sur la composante mineure de l'image source. Deux exemples illustrent ce point en figure 8.7 : dans l'image source 8.7(a) le ciel occupe ainsi une place plus réduite que dans l'image de style 8.7(b), et dans l'image source 8.7(d), les fleurs occupent une place plus importante que dans l'image de style 8.7(e). Le résultat expérimental du transfert par l'algorithme IDT confirme ce problème : le champ de blé devient en partie bleu (image 8.7(c)) et les pétales des fleurs et les feuilles ne sont plus de couleur homogène (image 8.7(f)).

Notons que ce problème de différence de composition a précédemment été évoqué dans la littérature. Dans [RAGS01], Reinhard *et al.* soulignent déjà ce type de phénomène, et évoquent la nécessité d'une intervention de l'utilisateur afin de segmenter l'image en différentes composantes sur lesquelles appliquer séparément le transfert de couleurs. Ce type de solution a par exemple été mis en œuvre dans [AK07]. Une modélisation des distributions de couleurs par un mélange de gaussiennes est réalisée dans [TJT05] pour traiter ce problème de manière automatique. Dans [NN05], les auteurs évoquent un problème similaire. Étonnamment, Pitié *et al.* n'évoquent pas ce problème dans [PK06, PKD07].

En raison de ces différentes limitations du transport (bruit, artefact, différence de composition), il est nécessaire de régulariser le résultat obtenu. Différentes stratégies peuvent être alors adoptées, suivant que l'on souhaite faire intervenir ou non l'utilisateur. Nous nous restreindrons dans la suite de ce chapitre aux méthodes de régularisation sans intervention de l'utilisateur.

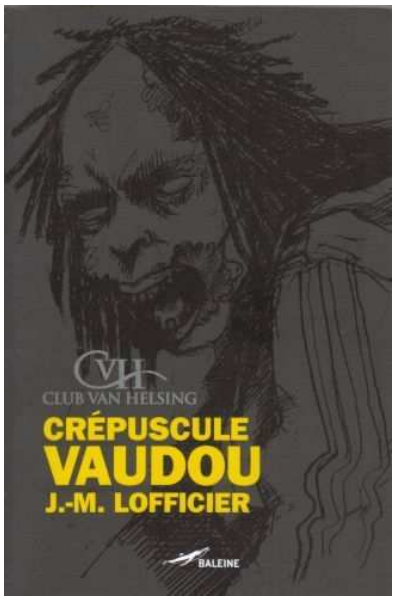
À notre connaissance, une unique méthode [PK06] a été proposée pour limiter le problème du rehaussement du niveau de bruit pour le transfert de couleurs, que nous allons présenter dans le paragraphe

(a) Image originale.

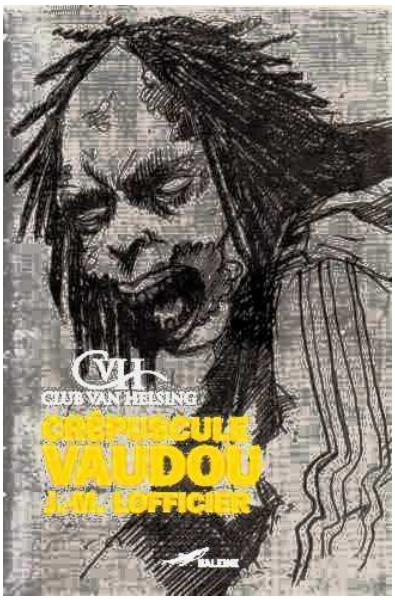
(b) Égalisation d'histogramme.

(c) Image originale.

(d) Égalisation d'histogramme.



(e) Image originale.



(f) Égalisation d'histogramme.

FIG. 8.5 – *Illustration du problème de l'égalisation d'histogramme.* *La première colonne montre les images originales, et la seconde colonne les images obtenues après égalisation de l'histogramme de luminosité (*i.e. *canal 'V' de la représentation HSV).*

(a) Image source. (b) Image de style couleur. (c) Transfert de la couleur par l'algorithme IDT.

(d) Image source. (e) Image de style couleur. (f) Transfert de la couleur par l'algorithme IDT.

FIG. 8.6 – *Illustration du problème de compression JPEG pour le transfert de couleurs.* *La première colonne montre les images originales, et la seconde les images de style utilisées. En troisième colonne sont placées les images de transfert de couleurs obtenues par la méthode IDT de [PKD07].*

(a) Image source. (b) Image de style couleur. (c) Transfert de la couleur par l'algorithme IDT.

(d) Image source. (e) Image de style. (f) Transfert de la couleur par l'algorithme IDT.

FIG. 8.7 – *Illustration du problème de proportion de couleur.* *La première colonne montre les images originales, et la seconde les images de style utilisées. En troisième colonne sont placées les images de transfert de couleurs obtenues par la méthode IDT de [PKD07].*

suivant. Nous montrerons son intérêt à l'aide de quelques exemples, puis nous en étudierons les limitations.

8.1.4 Restauration du grain par méthode variationnelle (regraining)

Dans le cadre du problème de transfert de couleurs et d'égalisation de contraste, Pitié *et al.* proposent [PKD07] de régulariser l'image après transport selon une approche variationnelle. En notant $t(u)$ l'image obtenue après application du transport t sur l'image source u, la fonctionnelle suivante est définie :

$$J_u(w) = \int_{x\in\Omega} \left\{ \phi(x) \cdot ||\nabla w(x) - \nabla u(x)||^2 + \psi(x) \cdot ||w(x) - t \circ u(x)||^2 \right\} dx \ , \tag{8.1}$$

où ∇ désigne l'opérateur de gradient et $||.||$ la norme euclidienne. Les termes ϕ et ψ sont des fonctions de pondération telles que ϕ privilégie le premier terme de J_u dans les régions où le gradient de u est faible, et ψ privilégie le second terme sur les fortes transitions de u :

$$\phi(x) = \frac{30}{1 + 10\,||\nabla u(x)||} \quad \text{et} \quad \psi(x) = \begin{cases} 1 & \text{si } ||\nabla u(x)|| > 5 \\ \frac{1}{5}||\nabla u(x)|| & \text{si } ||\nabla u(x)|| \leq 5 \end{cases} .$$

L'originalité de cette définition est qu'elle dépend à la fois de l'image à régulariser $t(u)$, mais également de l'image originale u. En minimisant cette fonctionnelle, on cherche une image w dont les couleurs sont proches de celles de $t(u)$ (attache aux données), mais dont le gradient est proche de celui de l'image originale u lorsque ce dernier est faible (restauration) : de cette manière, la solution conserve à la fois les zones plates de u et les petites structures. Pitié *et al.* nomment cette procédure de régularisation *regraining*, car elle permet de restaurer le grain d'une photographie qui a été traitée par manipulation de son histogramme. Ils proposent un schéma numérique itératif de cette approche dont le temps de calcul est de quelques dizaines de secondes pour une image d'un million de pixels.

Expérimentalement, cette méthode de régularisation permet effectivement de diminuer le bruit lorsqu'il y a rehaussement local du contraste, et de récupérer des détails lorsqu'il y a diminution de contraste. La régularisation des exemples de la figure 8.4 est illustrée en figure 8.8 (une évaluation des résultats requiert une visualisation électronique de ce document). On notera notamment le gain en détails les fleurs 8.8(d) et la réduction du bruit sur les poissons 8.8(b). Cette approche dépend du *compromis* entre réduction du bruit et rehaussement de détails : dans l'exemple des poissons-clown, la limitation du bruit sur les structures fines des tentacules de l'anémone se traduit par un léger flou.

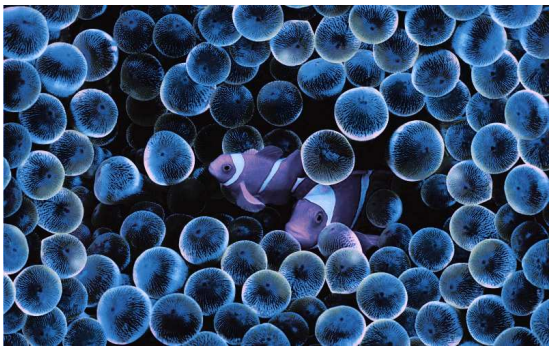
Néanmoins, certains problèmes demeurent. Nous verrons dans la section 8.3 que les artefacts liés à la compression de l'image (voir les exemples donnés en figures 8.19(d) et 8.18(d)) ne sont pas éliminés, et que le problème du changement de la proportion de couleur n'est pas traité. Par exemple, dans la figure 8.4(f), certaines feuilles originellement vertes dans l'image source (voir le détail donné en figure 8.8(e)) comportent des taches jaunes après le transport, et des structures vertes et marron sont apparues dans les pétales de fleurs (voir détail en figure 8.8(f)). Après le regraining, l'image restaurée en figure 8.8(d) présente toujours ces problèmes (voir le détail donné en figure 8.8(g)).

Nous introduisons dans la section suivante une nouvelle approche pour la régularisation des images obtenues par application d'un transport de couleurs.

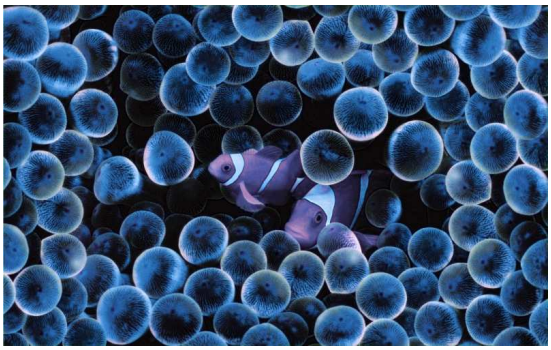
8.2 Régularisation du transport

Afin de régulariser l'ensemble des défauts liés au transfert tout en préservant les détails de l'image originale, nous proposons dans cette section une nouvelle méthode de régularisation du transport fondée sur le filtrage.

Dans un premier temps, nous formalisons les différents problèmes du transfert étudiés dans la section précédente en considérant la carte de transport. Nous rappelons dans le paragraphe suivant quelques unes des nombreuses méthodes de filtrage existantes et dont nous nous sommes inspirés, avant de présenter notre approche.



(a) Transfert de la couleur sans regraining.



(b) Après restauration par regraining.

(c) Transfert de la couleur sans regraining.

(d) Après restauration par regraining.

(e) Détail de l'image source (figure 8.7(d)).

(f) Détail avant regraining (figure 8.8(c)).

(g) Détail après regraining (figure 8.8(d)).

FIG. 8.8 – ***Illustration de la régularisation par regraining.*** *La première colonne montre les images après transfert de couleurs, et la seconde, l'image obtenue après régularisation par la méthode de [PKD07].*

8.2.1 Formalisation

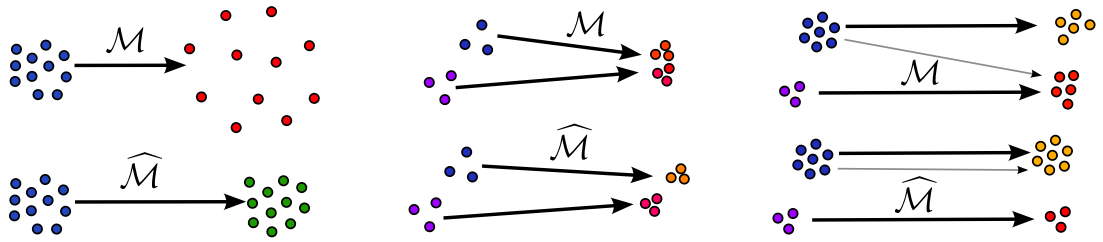
Rappelons que u désigne l'image source, et v l'image de style. On suppose que l'on obtient par un procédé quelconque une nouvelle image $t(u)$, où t est une application de transfert qui dépend de l'image de style. Dans la partie expérimentale, nous considérerons les cas de l'égalisation d'histogramme (sur un seul canal) et du transfert de couleurs, mais d'autres types de transfert sont envisageables.

L'originalité de notre approche réside dans le fait que nous proposons de régulariser l'image $t(u)$ en considérant la carte de transport. Cette carte de transport $\mathcal{M}$ est simplement définie par l'application $\mathcal{M} = t - id : u \in \mathbb{R}^3 \mapsto t(u) - u \in \mathbb{R}^3$, où id désigne l'identité. Voyons comment se caractérisent, du point de vue de cette carte de transport, les différents problèmes évoqués dans la section précédente :

– ***Rehaussement du bruit*** Considérons une image bruitée u dont le contraste est rehaussé. Les pixels bruités d'une région homogène de u correspondent ainsi à un nuage de point, dont on modifie la moyenne par application de $\mathcal{M}$, mais également la taille (augmentation de l'écart type). On souhaiterait, au contraire, définir une carte de transport $\widehat{\mathcal{M}}$ qui modifie la moyenne sans accroître

la variance du bruit (voir la figure 8.9(a)). Précisons que l'on ne cherche pas à *diminuer* la variance du nuage, comme c'est le cas avec les applications de débruitage d'images.

– ***Apparition d'artefacts*** Les artefacts se produisent lorsque des pixels de couleurs très proches (indistinguables à l'œil nu), se retrouvent affectés à des couleurs très différentes. Dans le cas de la compression JPEG, ces pixels sont répartis en blocs dans le domaine de l'image. C'est un problème tout à fait analogue au rehaussement de bruit.
– ***Perte des détails*** La perte de détails est due en partie à la réduction de la dynamique de l'image, comme le montre la figure 8.9(b). Deux groupes de points très distants dans l'image originale vont devenir très rapprochés. Il est nécessaire dans ce cas de maintenir une distance entre les groupes si l'on souhaite pouvoir préserver les détails de l'image source u.
– ***Proportion de couleur*** La figure 8.9(c) illustre le problème de la proportion de couleur présenté en § 8.1.3 (les proportions sont ici symbolisées par le nombre de points de chacun des groupes). Cet exemple illustre que tous les points d'un même groupe ont sensiblement le même type d'affectation, à l'exception des points supplémentaires du mode principal de la distribution de u qui se retrouvent affectés à un mode différent de la distribution de l'image de style v. Idéalement, la carte de transport $\widehat{\mathcal{M}}$ devrait être définie de telle sorte que tous les points d'un mode soit transportés sur un même mode.



(a) Augmentation du bruit (b) Disparition des détails (c) Problème du transport de la couleur

FIG. 8.9 – *Pour chaque illustration, on représente une carte de transport $\mathcal{M}$ qui présente une irrégularité. En dessous, on représente le transport $\widehat{\mathcal{M}}$ que l'on souhaite obtenir pour éviter ce problème.*

Cette taxonomie des différents problèmes du transfert de caractéristiques nous amène finalement au constat suivant : du point de vue du transport, tous les effets indésirables du transfert peuvent être interprétés comme des irrégularités spatiales de la carte de transport. Considérer la régularisation de la carte de transport nous permet de résoudre l'ensemble de ces problèmes avec *une unique méthode*, sans compromis entre préservation des détails et réduction du bruit. En outre, ces irrégularités sont facilement détectables en considérant la distribution des points de l'image source u. Cela signifie que la régularisation de l'application $t - id$ doit prendre en compte la distribution de l'image source u.

La solution que nous allons proposer, qui permet de régulariser la carte de transport $\mathcal{M} = t - id$ en fonction de u, s'inspire des filtres non-locaux, dont nous rappelons le principe dans le paragraphe suivant.

8.2.2 Filtres non-locaux

En régularisation d'images, une solution alternative à l'approche variationnelle est l'utilisation de filtres. Une approche standard consiste à utiliser des filtres *locaux* tels que la moyenne pondérée par un noyau gaussien. La convolution d'une image par un noyau gaussien reproduit le processus de diffusion de l'équation de la chaleur, ce qui se traduit par la perte des détails de l'image. D'autres approches, comme le filtre médian, permettent de préserver les structures contrastées, mais entraînent la perte des structures fines et peu contrastées.

Pour permettre la régularisation des images tout en préservant les structures telles que les bords ou les textures, différents filtres *semi-locaux* ou *non-locaux* ont été proposés. Ils connaissent un large succès

dans la littérature pour les applications de débruitage, mais leur usage s'est également étendu à d'autres applications.

Filtres semi-locaux Soit u une image continue en couleur, définie selon trois canaux, telle que $u : x \in \Omega \mapsto u(x) \in \mathbb{R}^3$, où Ω est un ensemble borné de $\mathbb{R}^2$. Soit $D_\rho(x) \in \mathbb{R}^2$, disque de rayon ρ centré sur x, le voisinage spatial de x sur lequel on régularise $u(x)$. Le filtre de Yaroslavsky [Yar85] consiste à calculer pour tout x la moyenne pondérée des valeurs de $u(y)$ sur le voisinage $y \in D_\rho(x)$, de la manière suivante :

$$Y(u(x)) = \frac{1}{C(x)} \int_{y \in D_\rho(x)} u(y)\, e^{-\frac{||u(x)-u(y)||^2}{2\sigma^2}} dy \ ,$$

où $||.||$ désigne la norme euclidienne d'un point de $\mathbb{R}^3$, et où $C(x)$ est une constante dépendant de x qui représente la somme des termes de pondération de la moyenne, soit

$$C(x) = \int_{y \in D_\rho(x)} e^{-\frac{||u(x)-u(y)||^2}{2\sigma^2}} dy \ .$$

La variable σ est un paramètre de réglage, définissant le voisinage dans l'espace couleur des valeurs de u que l'on va moyenner. Le filtre de Yaroslavsky consiste donc, *localement* sur le voisinage D_ρ, en un lissage de l'image selon la similarité des échantillons dans l'espace couleur, ce qui permet de préserver les structures comme les bords. En comparaison avec la convolution, on peut alors qualifier ce type de filtre de *semi-local*, puisque le voisinage spatial D_ρ n'intervient que pour la sélection des échantillons, et non pour leur comparaison. Dans [BCM05], les auteurs qualifient ce filtre de « *neighborhood filter* ». Ils en étudient le comportement dans le cas où le voisinage est négligeable devant le paramètre d'échelle σ ($\rho \ll \sigma$) et lorsque $\sigma \to 0$. Dans ce cas limite, ils montrent que le bruit éliminé par le filtre de Yaroslavsky dérive d'une équation de la chaleur, tout comme la convolution par un noyau gaussien, avec un terme de conductivité dépendant de la dérivée première de l'image u.

De multiples variantes sont possibles à partir du filtre de Yaroslavsky, parmi lesquelles les filtres bilatéraux (*bilateral filters*). Ils ont été indépendamment proposés par [SB97] et [TM98], et consistent à remplacer la définition du voisinage $D_\rho(x) \in \mathbb{R}^2$ par un nouveau terme de pondération de la moyenne qui dépend de la proximité spatiale des échantillons :

$$BF(u(x)) = \frac{1}{C(x)} \int_{y \in \Omega} u(y)\, e^{-\frac{|x-y|^2}{2\rho^2}}\, e^{-\frac{|u(x)-u(y)|^2}{2\sigma^2}} dy \ ,$$

avec cette fois $C(x) = \int_{y \in \Omega} e^{-\frac{|x-y|^2}{2\rho^2}}\, e^{-\frac{|u(x)-u(y)|^2}{2\sigma^2}} dy$.

Une étude intéressante [DD02] fait le lien entre ce type de filtre et les approches variationnelles de diffusion, et suggère entre autres l'utilisation d'autres noyaux de pondération. Les auteurs proposent également une approximation des filtres bilatéraux qui se base à la fois sur un sous-échantillonnage de l'image, et sur l'utilisation de la transformée de Fourier rapide.

Dans [TM98], les auteurs indiquent qu'une seule itération de ce type de filtre est en pratique suffisante pour régulariser l'image u, à la condition de définir correctement les paramètres σ et ρ. Nous verrons qu'il est toutefois possible d'en itérer l'application, mais le choix pratique de σ reste un point critique de la méthode.

Non-Local means Dans [BCM05], A. Buades *et al.* proposent une extension des précédents filtres qui reposent à la fois sur la notion de voisinage spatial et de voisinage en espace couleur. Leur principale motivation est la suivante : les précédents filtres préservent les structures principales de l'image (bords contrastés) mais éliminent les structures les plus fines (texture), qui se comportent à l'échelle du pixel comme du bruit. Pour conserver ces structures, il faut que la moyenne soit effectuée entre des pixels

similaires d'une même texture. Le problème est qu'il est impossible de distinguer ces pixels dans un voisinage en comparant seulement leur intensité.

Leur approche consiste à mesurer la similarité entre les pixels en observant les patchs centrés sur les pixels comparés. Puisque de tels pixels similaires ne sont pas nécessairement voisins spatialement, il est pertinent de les chercher dans tout le domaine Ω de l'image. Ce filtre effectuant des moyennes non-locales (ou *Non-Local means*) s'exprime ainsi sous la forme :

$$NL(u(x)) = \frac{1}{C(x)} \int_{y \in \Omega} u(y)\, e^{-\frac{||u(x-.) - u(y-.)||^2_{G_a}}{2\sigma^2}} dy \ ,$$

où $||.||^2_{G_a}$ désigne la mesure de dissimilarité définie pour comparer les patchs $u(x-.)$ et $u(y-.)$, centrés respectivement en x et en y. $C(x)$ est une constante dépendant de x qui représente la somme des termes de pondération de la moyenne pour le pixel x. Dans [BCM05], la mesure de dissimilarité est définie comme :

$$||u(x-.) - u(y-.)||^2_{G_a} = \int_{t \in \Omega} G_a(t) ||u(x-t) - u(y-t)||^2 dt \ ,$$

où G_a désigne un noyau gaussien de moyenne nulle et de variance a^2.

Pour des raisons pratiques de temps de calcul, la distance $||u(x-.) - u(y-.)||^2_{G_a}$ est calculée entre des patchs carrés de n pixels de côté (n valant typiquement $3, 5, 7$), et la comparaison des patchs est restreinte à un voisinage spatial de x, comme pour les filtres semi-locaux.

Les NL-means ont été utilisés avec succès pour de nombreuses applications : débruitage d'images et de vidéos [BCM08, KB08, BKB07], démosaïquage [BCMS07], et régularisation de cartes de disparité [BHS08] pour la reconstruction stéréoscopique.

Une étude récente des NL-means [SSN09] montre que ce filtre peut être interprété en terme de marche aléatoire sur un graphe. Une analyse de son utilisation itérée montre qu'elle résulte en une diffusion dans l'espace couleur, et non une diffusion spatiale comme c'est le cas avec les filtres locaux.

De manière similaire aux filtres bilatéraux, le filtre des NL-means est en pratique utilisé une seule fois pour restaurer l'image. Pour la restauration de texture [BC08], T. Brox et D. Cremers montrent néanmoins qu'il n'existe pas de "bon choix" du paramètre σ. Idéalement, σ devrait être défini localement et non globalement pour éviter ce problème (voir par exemple [KB08, BKB07]). Comme alternative, ils proposent de contourner la difficulté de l'estimation de σ en calculant la moyenne sur les n-plus proches voisins, et ce, de manière itérée. Il est ainsi possible de définir trois schémas itératifs différents [SSN09], selon que l'on mette à jour les coefficients de pondération et/ou les échantillons utilisés par le filtre. Afin de simplifier les expressions suivantes, on note $\mathrm{NL}_b(a)$ l'opérateur de moyennes non locales appliqué à l'image a, en utilisant les coefficients de pondération définis à partir de l'image b :

1. Mise à jour des échantillons seulement : $u_{k+1} = \mathrm{NL}_{u_0}(u_k)$. Cette solution est étudiée par Singer *et al.* dans [SSN09].
2. Mise à jour des échantillons et des coefficients : $u_{k+1} = \mathrm{NL}_{u_k}(u_k)$. Ce schéma est utilisé par Boulanger et Kervrann dans [KB08].
3. Mise à jour des coefficients seulement : $u_{k+1} = \mathrm{NL}_{u_k}(u_0)$. C'est le schéma proposé par Cremers et Brox [BC08].

Nous allons maintenant introduire une méthode de régularisation pour le transfert de caractéristiques, qui consiste à utiliser un filtre de moyennes non-locales sur la carte de transport.

8.2.3 Régularisation de carte de transport par filtrage non-local itératif

Non-Local Map Regularization Dans le but d'exploiter les informations de l'image source et d'avoir une approche adaptative, nous proposons d'utiliser les filtres non-locaux. D'une part, ils permettent de définir *localement* un noyau de pondération dont la forme s'adapte aux données comparées. D'autre part, à l'image des NL-means, ces approches permettent par des opérations simples sur les pixels des images

(moyennes pondérées) de faire de la diffusion dans l'espace des patchs, sans nécessiter de travailler directement dans cet espace.

Le filtrage direct de l'image $t(u)$ se traduirait par une perte des détails, ce que l'on veut justement éviter. Or, nous avons vu au paragraphe 8.2.1 que le filtrage de la carte de transport permet de traiter l'ensemble des problèmes rencontrés avec le transport, en particulier le rehaussement des détails. Nous proposons donc d'appliquer à la carte de transport $\mathcal{M}(u) = t(u) - u$ l'opérateur des moyennes non locales NL_u dont les poids sont calculés à partir des patchs de l'image u, soit :

$$\mathrm{NL}_u\,\mathcal{M}(u(x)) = \frac{1}{C(x)} \int_{y\in\mathcal{N}(x)} \mathcal{M}(u(y))\, e^{-\frac{||u(x-.)-u(y-.)||^2}{n^2\sigma^2}} dy\ , \tag{8.2}$$

où $\mathcal{N}(x)$ désigne un voisinage du point x inclus dans le domaine Ω de l'image u. Le terme $||u(x-.)-u(y-.)||^2$ désigne la norme entre des patchs carrés de taille $n \times n$ pixels. La constante $C(x)$ de normalisation est définie comme : $C(x) = \int_{y\in\mathcal{N}(x)} \exp(-||u(x-.)-u(y-.)||^2/n^2\sigma^2)dy$.

Le filtrage de la carte de transport $\mathcal{M}(u)$ nous permet ainsi d'introduire un nouvel opérateur de régularisation de l'image $t(u)$ dépendant de l'image source u. Nous notons cet opérateur **NLMR**, pour *Non-Local Map Regularization* ; il est défini de la façon suivante :

$$\mathrm{NLMR}_u(t(u)) := u + \mathrm{NL}_u\,\mathcal{M}(u) = \underbrace{\mathrm{NL}_u\big(t(u)\big)}_{\text{filtrage de l'image } t(u)} + \underbrace{u - \mathrm{NL}_u(u)}_{\text{détails de l'image } u}\ . \tag{8.3}$$

L'expression (8.3) se compose d'une somme de deux termes, signifiant que la régularisation de l'image $t(u)$ par le filtre NLMR est réalisée à partir de deux opérations distinctes. Le premier terme est l'application du filtrage de l'image $t(u)$ par l'opérateur de moyenne non-locale NL_u, en considérant les poids de similarité à partir de l'image source u. Cette opération permet de réduire les effets visuels liés à l'augmentation de la dynamique par le transfert (artefacts de compression, augmentation du bruit et proportion de couleur). Cependant, cette opération seule n'est pas suffisante : du fait du filtrage, on perd le détail des structures fines, de la texture et du grain de la photographie. Le rôle du second terme est ainsi de reconstituer les détails de l'image source, ce qui permet d'obtenir un rendu plus naturel. Les détails de l'image source sont obtenus en calculant $u - \mathrm{NL}_u(u)$, la différence entre u et le résultat du filtrage par les moyennes non-locales, ce qui est considéré dans les applications de débruitage comme du bruit.

Définition des paramètres Le réglage de l'opérateur NLMR repose sur trois paramètres : la taille du voisinage $\mathcal{N}$, la taille du patch extrait en chaque pixel (patch carré de n pixels de côté), et enfin le paramètre σ de comparaison des patchs.

Contrairement au débruitage d'image, l'intérêt de comparer des patchs ($n > 1$) plutôt que la valeur du pixel ($n = 1$) est moins évident pour la régularisation de transfert de caractéristiques. Nous examinerons dans la partie expérimentale le résultat de la régularisation dans ces deux situations.

La définition du voisinage spatial $\mathcal{N}$ est en principe l'ensemble du domaine Ω de l'image u. Cependant, pour des raisons de temps de calcul, nous serons contraints dans la partie expérimentale d'utiliser un voisinage local. On définit alors $\mathcal{N}(x)$ comme le disque $D_\rho(x)$ centré sur le pixel d'intérêt x, et de rayon ρ.

Enfin, nous verrons que le paramètre le plus important de notre méthode est le paramètre σ.

Propriétés du filtre NLMR Il est important de souligner que dans la définition (8.3) du filtre NLMR, c'est le même opérateur NL_u qui est à la fois appliqué à u ainsi qu'à $t(u)$, en utilisant les mêmes poids pour le calcul de la moyenne. On s'assure ainsi de la propriété suivante : NLMR est **invariant par translation des niveaux de gris de chaque canal**. En effet, pour tout transfert t tel que $t : u(x) \mapsto u(x) + T$ où T est un vecteur constant de $\mathbb{R}^3$ quel que soit le pixel $x \in \Omega$, l'application de NLMR ne change pas l'image $t(u)$: $\mathrm{NLMR}_u(t(u)) = t(u)$.

Bien entendu, ce n'est plus vrai pour un changement linéaire des couleurs de l'image source. Si t est tel que $t : u(x) \mapsto \alpha u(x)$ où α est une constante de $\mathbb{R}^+$ quel que soit le pixel $x \in \Omega$, l'application de NLMR sur $t(u)$ donne : $\text{NLMR}_u(t(u)) = t(u) + (1-\alpha) \times (u - \text{NL}_u(u))$. Si α est plus grand que 1, il y a augmentation de la dynamique de l'image source u, et donc rehaussement du bruit. L'opérateur NLMR permet alors de *diminuer* l'amplitude du bruit après régularisation. Au contraire, lorsque α est plus petit que 1, l'opérateur NLMR permet de *rehausser* le contraste des détails.

Une autre caractéristique du filtre NLMR est que le filtrage du transport par des moyennes non-locales assure que la distribution de l'image régularisée $\text{NLMR}_u(t(u))$ est contenue dans l'enveloppe convexe de la distribution de $t(u)$. Cette propriété est intéressante car elle prévient l'apparition de « fausses couleurs ».

Avant de vérifier ces propriétés sur des exemples simples, nous allons étudier l'utilisation itérée de ce filtre.

Itération du filtre NLMR Nous avons vu au paragraphe 8.2.2 pour le débruitage d'image que les filtres non-locaux étaient appliqués une seule fois lorsque la variance du bruit est connue, ce qui permet de définir σ. Cependant lorsque cela n'est pas le cas, plusieurs études [KB08, BC08, SSN09] ont montré que les filtres locaux pouvaient être appliqués successivement pour obtenir un bon résultat sans pour autant connaître σ *a priori*. Pour cela, différents schémas itératifs ont été proposés.

Le fait d'itérer la régularisation de la carte de transport nous permet en pratique d'améliorer grandement les résultats. Par ailleurs, il rend le choix de σ moins critique. Nous avons choisi un schéma itératif où les coefficients de pondération sont toujours les mêmes, calculés à partir de l'image source et de la constante σ. La carte de transport à l'itération k, notée $\text{NL}_u^k\mathcal{M}$, est ainsi définie par application récursive de l'opérateur NL_u sur la carte à l'itération précédente :

$$\text{NL}_u^k\,\mathcal{M}(u) = \underbrace{\text{NL}_u \circ \ldots \circ \text{NL}_u}_{k \text{ termes}}\big(t(u) - u\big)\ .$$

L'image finale au terme de k itérations est alors :

$$\text{NLMR}_u^k(t(u)) := \text{NL}_u^k\big(t(u)\big) + u - \text{NL}_u^k(u)\ .$$

Convergence Dans [SSN09], les auteurs étudient l'application itérative de l'opérateur NL_u sur une image u, selon le même schéma itératif que le nôtre, c'est-à-dire en ne mettant pas à jour les coefficients de pondération pour le calcul de la moyenne. Ils montrent que l'image filtrée obtenue **converge** vers une solution (minimum d'un potentiel) qui dépend du choix de σ.

Afin d'étudier empiriquement la convergence de l'opérateur NLMR, nous considérons une image u de bord de taille 100×100 px[1], avec un bruit iid normal sur chaque canal RVB (figure 8.10(a)). On ajuste ensuite le contraste de l'image source u par égalisation d'histogramme (figure 8.10(b)). On donne en table 8.2 la moyenne et la variance du niveau de gris de chaque palier de chaque image. L'égalisation d'histogramme de u permet d'augmenter l'écart entre les moyennes de chaque palier, mais elle augmente considérablement l'écart-type du bruit (près d'un facteur 4).

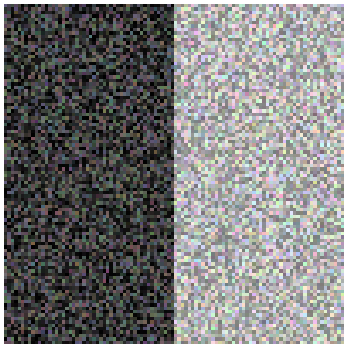
On applique ensuite l'opérateur NLMR itérativement jusqu'à ce qu'il y ait convergence. Au bout d'une dizaine d'itérations, on observe effectivement que la norme de variation de la carte de transport entre deux itérations $||\mathcal{M}^k(u(x)) - \mathcal{M}^{k-1}(u(x))||$ en chaque pixel x est plus petite que 1. On considérera par la suite qu'il y a convergence numérique lorsque l'on vérifie cette propriété. En figure 8.10(c) est donnée l'image régularisée en utilisant $\sigma = 10$, des patchs de 1 pixel, et un voisinage $\mathcal{N}$ étendu au domaine de toute l'image Ω. Le résultat obtenu est celui que l'on attendait : le bruit a été réduit tout en préservant le contraste moyen entre les deux paliers obtenu par le transfert t. La table 8.2 confirme ce constat : l'écart-type du bruit est effectivement redevenu proche de celui de l'image originale u, tandis que la différence des moyennes entre paliers (le contraste) est très proche de celui de l'image $t(u)$.

[1] l'abréviation *px* désigne une unité en nombre de pixels.

TAB. 8.2 – *Moyennes et écart-type des niveaux de gris des paliers dans chaque image.*

Image (et figure)	Caractéristiques (moyenne μ et écart-type δ)				
	μ_1	μ_2	$\mu_2 - \mu_1$	δ_1	δ_2
u (8.10(a))	100	150	50	7.6	7.7
$t(u)$ (8.10(b))	56.7	178	121.3	32	31.5
$\text{NLMR}_u^\infty\, t(u)$ (8.10(c)), $n = 1$	48	170	122	8.4	8.5
$\text{NLMR}_u^\infty\, t(u)$ (8.11(a)), $n = 3$	54.2	175.2	121	8.4	8.5

(a) Image source u (bord avec bruit additif *iid* normal sur chaque canal).



(b) Image $t(u)$ après égalisation d'histogramme de u.

(c) Régularisation par NLMR itéré avec $\sigma = 10$, $n = 1$, $\mathcal{N} = \Omega$.

FIG. 8.10 – *Utilisation du filtrage NLMR sur un bord bruité. Le contraste de l'image source u (figure 8.10(a)) est rehaussé par égalisation d'histogramme (figure 8.10(b)). On applique successivement le filtre NLMR jusqu'à convergence pour obtenir le résultat en figure 8.10(c). On peut observer que le changement de contraste obtenu par égalisation est préservé, tandis que le réhaussement du niveau de bruit a été fortement réduit (voir le tableau 8.2). Nous avons utilisé des patchs de $n = 1$ pixel, $\sigma = 10$ et un voisinage $\mathcal{N}$ étendu à tout le domaine Ω de l'image source.*

(a) Régularisation par NLMR itéré avec $\sigma = 10$, $n = 3$, $\mathcal{N} = \Omega$.

(b) Régularisation par NLMR itéré avec $\sigma = 10$, $n = 1$, et $\mathcal{N} = \mathcal{D}_\rho$ où $\rho = 10$ px.

(c) Régularisation par NLMR itéré avec $\sigma = 100$, $n = 1$, $\mathcal{N} = \Omega$.

FIG. 8.11 – *Régularisation du bord bruité (figure 8.10(b)) par itération de NLMR jusqu'à convergence. En utilisant des patchs ($n = 3$ en figure 8.11(a)), ou un voisinage restreint ($\mathcal{N} = \mathcal{D}_\rho$ où $\rho = 10$ px, figure 8.11(b)) le résultat obtenu après convergence est très similaire au résultat de la figure 8.10(c). Par contre, en utilisant un paramètre de régularisation beaucoup plus grand ($\sigma = 100$, figure 8.11(c)), on perd le changement de contraste apporté par l'égalisation d'histogramme.*

Afin d'illustrer l'importance des différents paramètres de l'approche, nous reproduisons cette expérience en modifiant les trois paramètres du filtre NLMR.

Nous avons utilisé des patchs de 1 pixel dans le premier exemple de régularisation (figure 8.10(c)). En figure 8.11(a) est donné le résultat de la régularisation NLMR en comparant des patchs 3×3. L'image régularisée est assez proche de celle obtenue précédemment, sauf au niveau du bord où le nombre de patchs utilisés pour le calcul de la moyenne est moins important. Comme on peut le voir dans le tableau 8.2, l'écart type du bruit est identique à celui obtenu pour la comparaison de patchs de 1 pixel.

Si l'on se restreint à un voisinage $\mathcal{N}$ de rayon $\rho = 10$ pixels au lieu du domaine Ω, la convergence requiert le double d'itérations pour la même valeur $\sigma = 10$, mais le résultat obtenu est très proche (figure 8.11(b)) du précédent résultat (figure 8.10(c)). Ceci illustre le phénomène de diffusion qui se produit dans les régions homogènes de l'image source. Cela signifie également que la restriction à un voisinage spatial $\mathcal{N}$ limite la diffusion aux régions homogènes connexes de l'image source. Nous illustrerons ce principe dans la partie expérimentale.

C'est finalement le choix de σ qui s'avère le plus critique, car ce paramètre définit quels sont les pixels considérés comme « homogènes ». Dans notre exemple, $\sigma = 10$ est un choix convenable, car l'écart moyen entre deux pixels de chacun des paliers dans l'image source est de 50 en niveau de gris (voir le tableau 8.2). La figure 8.11(c) montre le résultat de la régularisation lorsque $\sigma = 100$: il y a diffusion entre les pixels des deux plateaux, et l'on perd le contraste obtenu par le transfert. Si, au contraire, σ est petit devant l'écart-type du bruit δ dans l'image source, la diffusion devient excessivement lente et la régularisation entre deux itérations est imperceptible.

8.2.4 *Discussion*

Nous avons présenté au paragraphe 8.2.2 les filtres bilatéraux comme un outil de débruitage. Toutefois, ces filtres ont également été utilisés dans quelques travaux en rapport avec le transfert de caractéristiques, notamment dans le but de préserver les détails. Bien que les filtres bilatéraux ne soient pas utilisés dans ce contexte pour régulariser le transfert obtenu, ces travaux que nous allons maintenant rappeler présentent quelques analogies avec notre approche. Nous illustrerons ensuite par un cas pratique l'intérêt du filtre NLMR.

Le filtre bilatéral a été utilisé dans d'autres applications que le débruitage, en particulier [DD02] pour l'affichage des images de grande dynamique (*i.e.* de plus de 24 bits) sur des moniteurs ayant une dynamique plus faible. Un simple ajustement de l'histogramme d'intensité fait généralement perdre les détails. Pour éviter cela, Durand et Dorsey utilisent les filtres bilatéraux afin de séparer les basses et hautes fréquences de l'image à traiter, l'ajustement de dynamique étant uniquement réalisé sur l'image de base (basses fréquences). Les détails (hautes fréquences) sont ensuite simplement ajoutés à l'image de base. Notons que dans [BSL08], Bonnier *et al.* utilisent une approche similaire pour l'impression de photographies (*gamut mapping*). Contrairement aux applications précédentes de débruitage, le filtrage n'est pas utilisé pour régulariser les images, son utilisation n'est donc pas itérée. Ce principe de décomposition des images par les filtres bilatéraux a ensuite été repris et modifié pour deux autres applications de transfert de caractéristiques :

- **Utilisation de paires de photographies avec et sans flash.** Deux approches très analogues ont été simultanément proposées dans [ED04, PSA+04], dont le but est d'obtenir une photographie de haute qualité d'une scène sans flash, tout en bénéficiant des détails donnés par la même scène prise avec flash. Pour parvenir à ce résultat, ces deux travaux se basent sur l'utilisation de la décomposition qui vient d'être introduite, afin d'obtenir une nouvelle image en recombinant la couleur et les détails de l'image avec flash, avec l'image de base extraite de l'image sans flash. Cependant, l'extraction de cette image de base est souvent de mauvaise qualité en raison du très faible rapport signal sur bruit. Pour contourner cette difficulté, le filtre bilatéral est mis en œuvre sur l'image *sans flash* en calculant les poids de la moyenne à partir de l'image *avec flash*. Cette variante est appelée *cross bilateral filter* (ou *joint bilateral filter*).
- **Transfert de texture.** Dans [BPD06], S. Bae *et al.* s'intéressent au transfert de caractéristiques

(contraste et texture) entre deux photographies. Le même cadre de travail que pour l'utilisation de paires de photographies avec et sans flash est utilisé. Le *cross bilateral filter* permet alors de transférer la texture d'une image à une autre.

Une fois encore, le *cross bilateral filter* est utilisé dans ces dernières applications comme un moyen de réaliser le transfert de caractéristiques et non comme une régularisation du transfert. Cependant, cette approche illustre l'intérêt pour certaines applications d'utiliser une mesure de similarité à partir d'une autre image.

Comparaison avec notre méthode Nous avons vu, avec la méthode décrite dans [DD02] pour la réduction de dynamique, que les filtres bilatéraux sont utilisés pour décomposer une image source u en une image de base u_B et une image de détail Δu. Le transfert t_B est alors appliqué sur l'image de base u_B, puis l'image de détail Δu est ajoutée à l'image obtenue. Ce principe de la décomposition peut être mis en équation de la façon suivante avec les filtres de moyennes non-locale NL_u :

$$D(u) = t_B\big(u_B\big) + \Delta u = t_B\big(\mathrm{NL}_u(u)\big) + u - \mathrm{NL}_u(u)\,,$$

où $D(u)$ désigne l'image obtenue par ce principe de décomposition.

Ce résultat présente une forte analogie avec notre approche lorsque l'on applique une seule itération du filtre NLMR à l'image $t(u)$:

$$\mathrm{NLMR}(t(u)) = \mathrm{NL}_u\big(t(u)\big) + u - \mathrm{NL}_u(u)\,.$$

Dans ce cas particulier où l'on ne réalise qu'une itération, la différence majeure entre ces deux méthodes réside dans le fait de régulariser l'image *avant* le transfert de caractéristiques ou *après*.

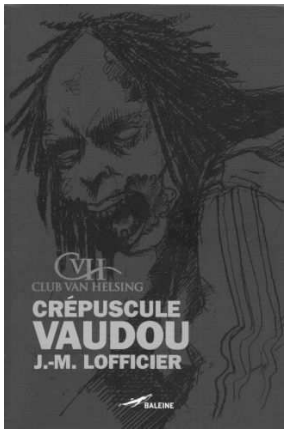
Pour illustrer cette différence, nous étudions dans la figure 8.12 le cas de l'égalisation d'histogramme pour une photographie en niveau de gris (image 8.12(a)). Afin de procéder selon la méthode décrite dans [DD02], le filtre NL_u est utilisé avec des patchs de 1 pixel ($n = 1$), pour un voisinage spatial défini par un disque de rayon $\rho = 10$ px. Nous avons sélectionné le paramètre $\sigma = 15$ donnant le meilleur rendu visuel. Par application du filtre NL_u, on obtient l'image de base u_B (figure 8.12(b)) et l'image de détail Δu (figure 8.12(c)). Dans [DD02], une réduction de contraste est utilisée pour diminuer la dynamique de l'intensité des images. Ici, le transfert t_B qui est appliqué à l'image de base est une égalisation d'histogramme, ce qui a pour effet d'augmenter le contraste (figure 8.12(d)). Il est intéressant de constater que le transfert t_B produit beaucoup moins d'artefacts et de bruit que le transfert t qui est obtenu par égalisation de l'image u (figure 8.12(g)). L'image finale est obtenue par ajout des détails (figure 8.12(e)), et gagne ainsi en netteté. Cependant, il reste de nombreux artefacts et l'image obtenue est plus « floue » que l'image originale.

La figure 8.12(h) montre le résultat pour une seule itération du filtre NLMR sur l'image $t(u)$, défini à partir du même filtre NL_u que la méthode de décomposition précédente. Avec notre approche, l'image obtenue est beaucoup plus proche de l'image originale. Ceci illustre le fait que la régularisation de la carte de transport permet de conserver les détails de l'image, contrairement à une approche classique de filtrage d'image.

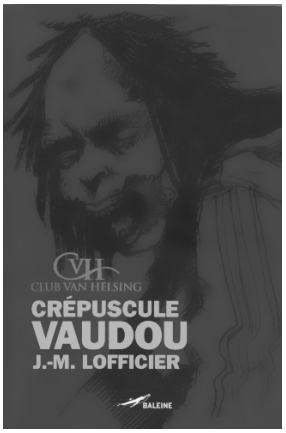
8.2.5 Considérations pratiques

Dans ce paragraphe, nous précisons quelques points essentiels de la mise en œuvre de notre approche.

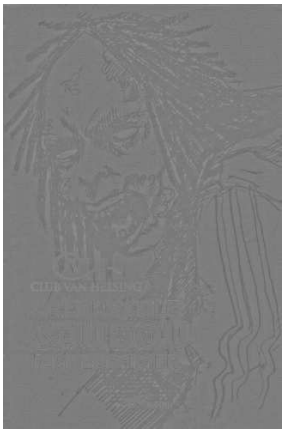
Paramètre de comparaison de patchs σ Le paramètre σ est le seul paramètre critique de notre approche, le choix de la taille du voisinage ainsi que le nombre d'itérations étant seulement motivés par des considérations computationnelles. Idéalement, σ doit être choisi de manière à s'adapter aux données. Dans les résultats expérimentaux qui seront donnés par la suite, la valeur est toujours fixée à $\sigma = 10$, sauf indication contraire. Nous reviendrons dans la section 8.4 sur la définition de ce paramètre.



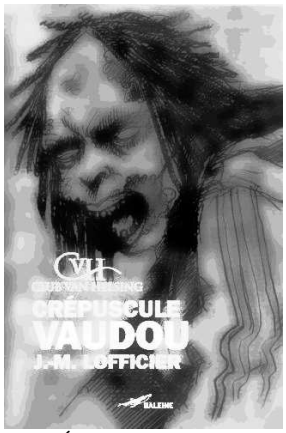
(a) Image originale u.



(b) Composante de base $u_B = \mathrm{NL}_u(u)$.

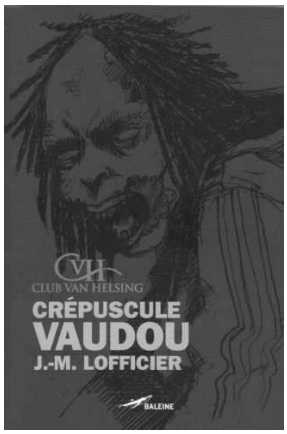


(c) Composante de détail $\Delta u = u - u_B$ (ici affichée avec une moyenne de 125)

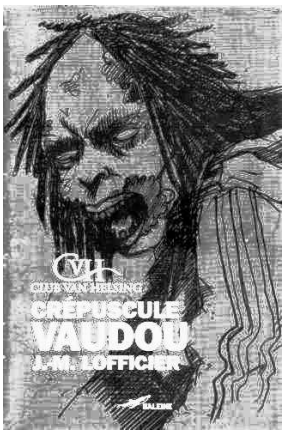


(d) Égalisation d'histogramme sur u_B : $t_B(u_B)$.

(e) Image finale obtenue par ajout des détails : $t_B(u_B) + \Delta u$.



(f) Image originale u.



(g) Image après transfert $t(u)$, obtenue par égalisation de l'image u.

(h) Notre approche avec une seule itération : $\mathrm{NL}_u(t(u)) + \Delta u$.

FIG. 8.12 – *Comparaison de la méthode de transfert par décomposition en image de base et de détail [DD02] avec notre approche de régularisation du transport.* *Les deux approches utilisent le même principe de filtrage, mais de manière différente. Le résultat visuel est complètement différent : notre approche vise à régulariser le transport, tandis que la méthode de [DD02] permet d'appliquer un transfert sans modifier les détails.*

Espace couleur Le choix de l'espace couleur intervient pour le calcul de la distance entre les patchs ainsi que pour le calcul de la moyenne de la carte de transport. Nous utilisons dans la partie expérimentale l'espace RVB (Rouge Vert Bleu), mais d'autres espaces sont envisageables afin d'améliorer la notion de similarité, à l'image de la transformation CIE Lab. Toutefois, il convient de noter que l'utilisation d'espaces ayant une dimension circulaire (comme le canal de teinte 'H' des espaces HSV et HSI) entraîne une modification de la définition de la moyenne pondérée dans la formule (8.3), qui requiert l'analyse de la congruence des différents termes de la moyenne.

Taille des patchs Nous avons pour l'instant montré un exemple jouet où la taille des patchs était ramenée à un pixel ($n = 1$). Il peut sembler *a priori* plus intéressant[2] de considérer des patchs de plus grande taille, afin de préserver la texture et les structures de l'image source. Afin de comprendre le rôle de la taille des patchs pour la régularisation, nous utiliserons dans la section expérimentale des patchs de taille $n = 1$ et $n = 3$ sur divers exemples. Dans ces deux cas, nous utiliserons le paramètre $\sigma = 10$.

Taille du voisinage spatial En raison de l'itération du filtre NLMR, la taille du voisinage spatial $\mathcal{N}$ n'a pas une grande influence sur le résultat final, mais son choix reste néanmoins crucial du point de vue du temps de calcul. À titre d'exemple, l'algorithme des NL-means sur une image 8 bits de taille 512×512 requiert une demi-heure de calcul avec un voisinage étendu à l'ensemble du domaine de l'image et $n = 3$ (pour une mise en œuvre en langage C). C'est la raison pour laquelle on se restreint en pratique à un voisinage de quelques dizaines de pixels, ce qui ramène le temps de calcul à plusieurs dizaines de secondes. Sauf mention contraire, on utilise un disque de rayon $\rho = 10$ pixels.

Carte de convergence Nous avons vu qu'il était *a priori* inutile de définir un nombre d'itérations au préalable, l'utilisation itérée de l'opérateur convergeant vers une carte $\mathcal{M}_\infty$. Nous n'avons pas à l'heure actuelle de preuve théorique de cette convergence, mais toutes les expériences que nous avons menées vont dans ce sens. Une analyse similaire à celle proposée par [BCM05], en faisant tendre σ et ρ vers 0, suggère une telle convergence, mais son intérêt reste limité en raison des conditions dans lesquelles elle est valide.

On considère en pratique qu'il y a convergence lorsque la norme de variation de la carte de transport entre deux itérations est plus petite qu'un seuil : $|\{\mathcal{M}_{k+1} - \mathcal{M}_k\}u(x)| < \epsilon \; \forall x \in \Omega$. En pratique, tous les pixels ne convergent pas en même temps. Pour cette raison, afin d'éviter d'appliquer l'opérateur NLMR en des pixels où le transport est stable, on ne traite que les points x pour lesquels $|\{\mathcal{M}_{k+1} - \mathcal{M}_k\}u(x)| \geq \epsilon$. Cependant, il est nécessaire de prendre en compte les pixels voisins de ces points ; en effet, en raison de la restriction au voisinage spatial $\mathcal{N}$, la diffusion se fait de proche en proche. Afin de préserver ce phénomène de diffusion, nous utilisons une carte de convergence qui est définie de la manière suivante :

1. Définition de l'image binaire $M(x) = 1$ si $|\{\mathcal{M}_{k+1} - \mathcal{M}_k\}u(x)| \geq 1$, et $M(x) = 0$ sinon.
2. Dilatation du masque binaire M par un disque de rayon ρ.

En itérant le filtre NLMR, seuls sont traités les pixels x pour lesquels $M(x) = 1$. Ceci permet de gagner un temps considérable si l'on souhaite itérer le filtre NLMR jusqu'à convergence. Dans l'exemple étudié en figure 8.10(a) où les points convergent à peu près à la même vitesse, l'utilisation de cette carte permet pourtant de gagner un facteur 3.

Comparaison rapide de patch Malgré l'utilisation de la carte de convergence, le filtre NLMR requiert un temps de calcul important : une seule itération sur tous les pixels d'une image couleur de taille 512×512 requiert en effet 20 secondes sur un processeur 2 GHz.

Nous n'avons pas cherché à diminuer ce temps de calcul, mais il est intéressant de noter que plusieurs approximations ont été proposées dans la littérature pour accélérer le calcul des moyennes non-locales.

[2] sauf en terme de complexité

Une première approximation consiste à calculer les moyennes entre patchs à partir d'une image sous-échantillonnée [KB08, DD02]. Dans [DCC+08], les comparaisons de patchs sont calculées en utilisation des « images intégrales » (*integral sum squared image*). Dans [PD09], les auteurs proposent une approximation des filtres bilatéraux fondée sur une convolution en coordonnées homogènes. Récemment, une méthode de recherche approchée des patchs par *kd-tree* a été proposée dans [AGDL09] pour les filtres non-locaux. D'après les gains en temps de calcul annoncés dans ces différentes études, toutes ces approximations permettent de diminuer le temps de calcul de deux ordres de grandeur. Dans notre cas, il serait également intéressant de tirer parti du fait que les poids sont toujours calculés à partir de l'image source u pour chaque itération du filtre NLMR. Il n'est donc pas nécessaire en théorie de recalculer la distance entre les patchs pour chaque nouvelle itération.

8.3 Validation expérimentale

Nous réalisons dans cette section plusieurs expériences visant à démontrer l'intérêt de notre approche pour la régularisation de deux types de transfert. Nous étudions dans un premier temps la régularisation par le filtre NLMR (Non Local Map Regularization) du transfert par égalisation d'histogramme. Dans la section suivante, nous nous intéressons à la comparaison de notre méthode avec celle de *regraining* [PKD07], pour le transfert de couleurs.

8.3.1 Égalisation d'histogrammes

Dans les expériences suivantes, nous appliquons un ajustement de contraste par égalisation d'histogramme du canal 'V' de la représentation HSV de l'image source u. Le filtre NLMR présenté dans la section précédente est ensuite appliqué avec un voisinage spatial de rayon 10 pixels, avec un paramètre σ égal à 10, pour des patchs de 1 ou 3 pixels de coté (paramètre n). Ce filtre est utilisé de manière itérée jusqu'à convergence numérique, tel que cela a été décrit au §8.2.5.

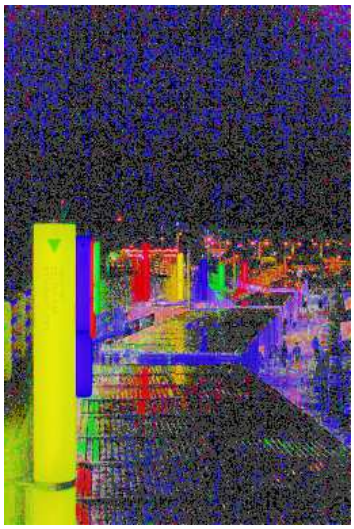
Le premier exemple que nous examinons est un cas extrême de réhaussement de contraste, permettant d'illustrer le phénomène de diffusion lié à l'itération du filtre NLMR. L'image 8.13(a) est une photographie de luminaires prise de nuit. L'égalisation d'histogramme permet de réhausser le contraste du fond (figure 8.13(b)), mais de nombreux artefacts bleus apparaissent dans l'image. Nous comparons ici l'utilisation du filtre NLMR avec des patchs carrés de $n = 1$ et 3 pixels de côté. Dès la première itération, la majorité des artefacts ont disparu (figure 8.13(c) pour $n = 1$, 8.13(e) pour $n = 3$). En raison du voisinage spatial utilisé, la diffusion est limitée à de petites régions : des structures apparaissent encore dans le ciel nocturne par exemple. En itérant le filtre NLMR jusqu'à convergence, on obtient les figures 8.13(d) ($n = 1$) et 8.13(f) ($n = 3$). Cette itération permet de propager la diffusion dans les zones homogènes de l'image source, tout en préservant les structures saillantes. On observe ainsi que les régions délimitant les luminaires sont préservées. Dans le même temps, les fausses couleurs apparues dans le ciel sont mélangées, ce qui donne un aspect plus homogène et naturel. Par contre, en raison du contraste extrêmement faible de l'image source, les arrêtes du batîment (en bas à droite) ne sont pas préservées. En effet le choix de $\sigma = 10$ ne permet pas de restreindre la diffusion à chacune des régions.

Un autre point important de cette expérience est l'importance de la taille du patch (n). Nous avons vu que pour le débruitage d'image, l'utilisation de patch permet d'améliorer la comparaison entre des pixels bruités. Dans notre cas, cela semble moins évident puisque l'on dispose de l'image originale u. Cependant, il est *a priori* plus intéressant d'utiliser des patchs afin de préserver la texture et les structures de l'image source. De manière assez surprenante, on observe en pratique que ce n'est pas nécessairement le cas. Les deux images obtenues pour $n = 1$ et $n = 3$ sont très similaires (figures 8.13(d) et 8.13(f) respectivement). Nous donnerons d'autres exemples dans la suite de cette section afin de comprendre le rôle des patchs dans la régularisation.

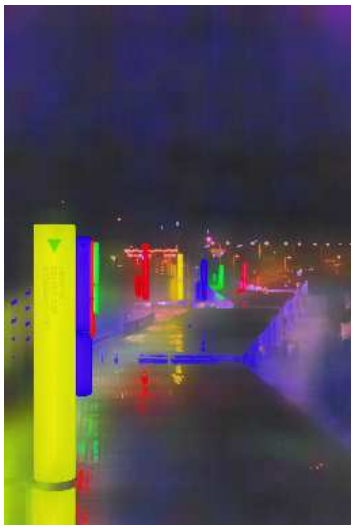
Nous revenons sur l'exemple de la couverture de livre (§8.2.4) avec la figure 8.14 . Cette fois nous illustrons le résultat obtenu par application du filtre NLMR jusqu'à convergence, après égalisation d'histogramme sur l'image source (figure 8.14(b)). Nous donnons en figure 8.14(c) le résultat pour $n = 1$, et en figure 8.14(d) pour $n = 3$. Il est intéressant de voir que l'itération du filtre NLMR permet effectivement d'éliminer tous les artefacts générés par l'égalisation d'histogramme. En particulier, on observe que l'utilisation de patch de 1 pixel permet de récupérer de nombreux détails (hachure du dessin). Le résultat de la régularisation pour des patchs de 3×3 pixels est légèrement différent : les structures fines sont plus contrastées mais également moins bien localisées.

Nous observons les mêmes résultats sur l'exemple de la figure 8.15. L'égalisation de l'histogramme de l'image originale 8.15(a) permet de réhausser le contraste du tableau, mais fait apparaître de nombreux détails (craquelures notamment) qui étaient peu visibles auparavant. La figure 8.15(c) montre l'image obtenue par régularisation de la carte de transport, avec $n = 1$. Une fois encore, l'utilisation de notre filtre NLMR permet de préserver les fines structures qui étaient suffisamment contrastées dans l'image source, comme les cheveux ou les plumes, mais également les nouvelles structures à grande échelle,

(a) Image originale.

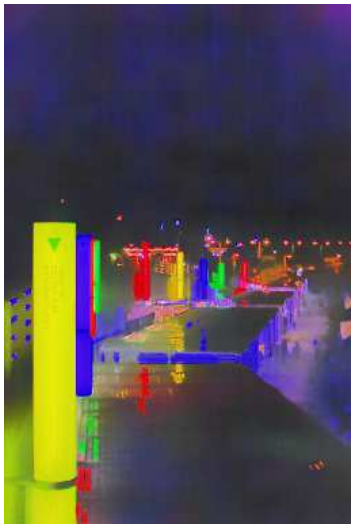


(b) Égalisation d'histogramme.



(c) Régularisation avec 1 itération ($n = 1$).

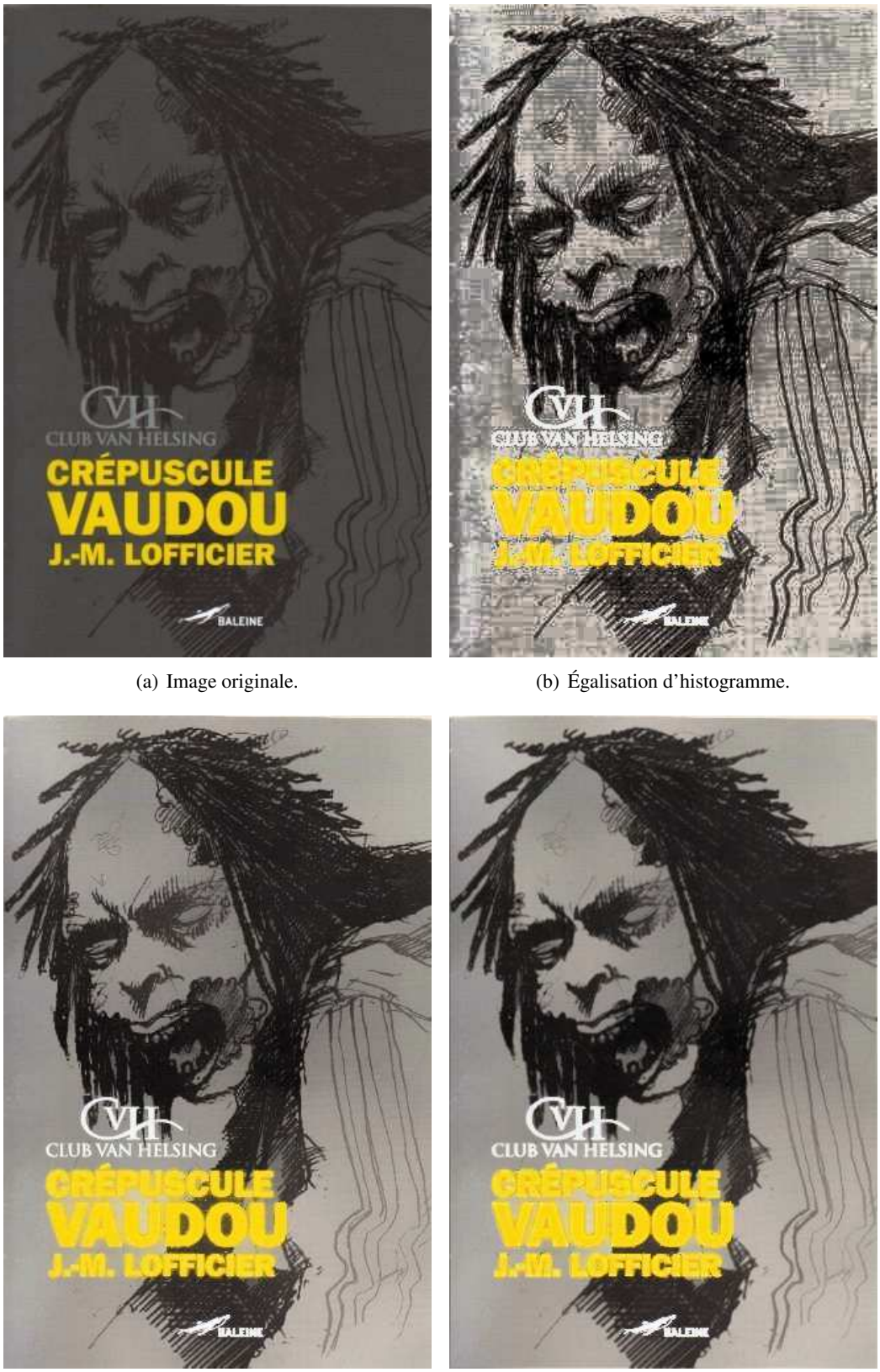
(d) Régularisation jusqu'à convergence ($n = 1$).



(e) Régularisation avec 1 itération ($n = 3$).

(f) Régularisation jusqu'à convergence ($n = 3$).

FIG. 8.13 – *Illustration de la régularisation pour l'égalisation d'histogramme.* *Il s'agit de l'égalisation de l'histogramme de luminosité (canal 'V' de la représentation HSV). Cet exemple extrême de transfert permet d'illustrer le phénomène de diffusion opéré par le filtre NLRM lorsqu'il est itéré.*



(a) Image originale.

(b) Égalisation d'histogramme.

(c) Régularisation par NLRM avec $n = 1$.

(d) Régularisation par NLRM avec $n = 3$.

FIG. 8.14 – *La figure 8.14(a) montre l'image source à laquelle on applique un transfert par égalisation d'histogramme (figure 8.14(b)). Les figures 8.14(c) et 8.14(d) sont le résultat de l'application itérée du filtre NLRM jusqu'à convergence en utilisant respectivement des patchs de 1 ou 9 pixels.*

(a) Image originale.

(b) Égalisation d'histogramme.

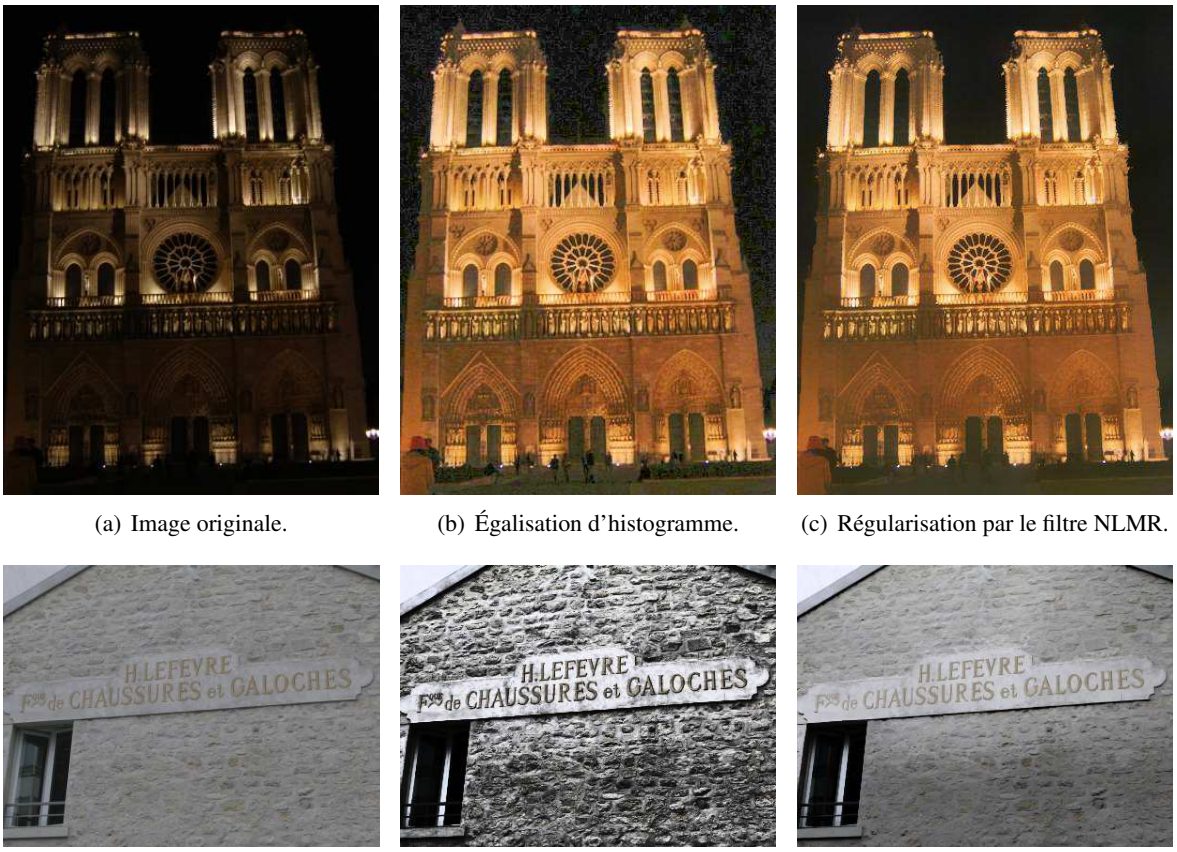
(c) Régularisation par NLMR, avec patch de 1px.

(d) Régularisation par NLMR, avec patch de 9px.

FIG. 8.15 – *La figure 8.15(a) montre l'image source à laquelle on applique un transfert par égalisation d'histogramme (figure 8.15(b)). Les figures 8.15(c) et 8.15(d) sont le résultat de l'application itérée du filtre NLRM jusqu'à convergence en utilisant respectivement des patchs de 1 ou 9 pixels.*

comme la branche d'arbre (en haut, à gauche). Les nouvelles petites structures qui sont apparues sur la tableau sont considérées comme du bruit et donc atténuées. Rappelons que le paramètre σ a été fixé arbitrairement à 10 comme pour les autres expériences. Le fait de changer ce paramètre, ou le nombre d'itérations, permet d'ajuster le degré de la régularisation. La figure 8.15(d) est obtenue avec des patchs de 3 px de côté. Une fois encore, le résultat est très similaire. Seul un examen attentif permet de constater que les structures fines dues aux craquelures sont encore plus atténuées. Étant donné le faible intérêt pratique apporté par l'utilisation des patchs, nous allons dans la suite seulement présenter les résultats obtenus avec $n = 1$.

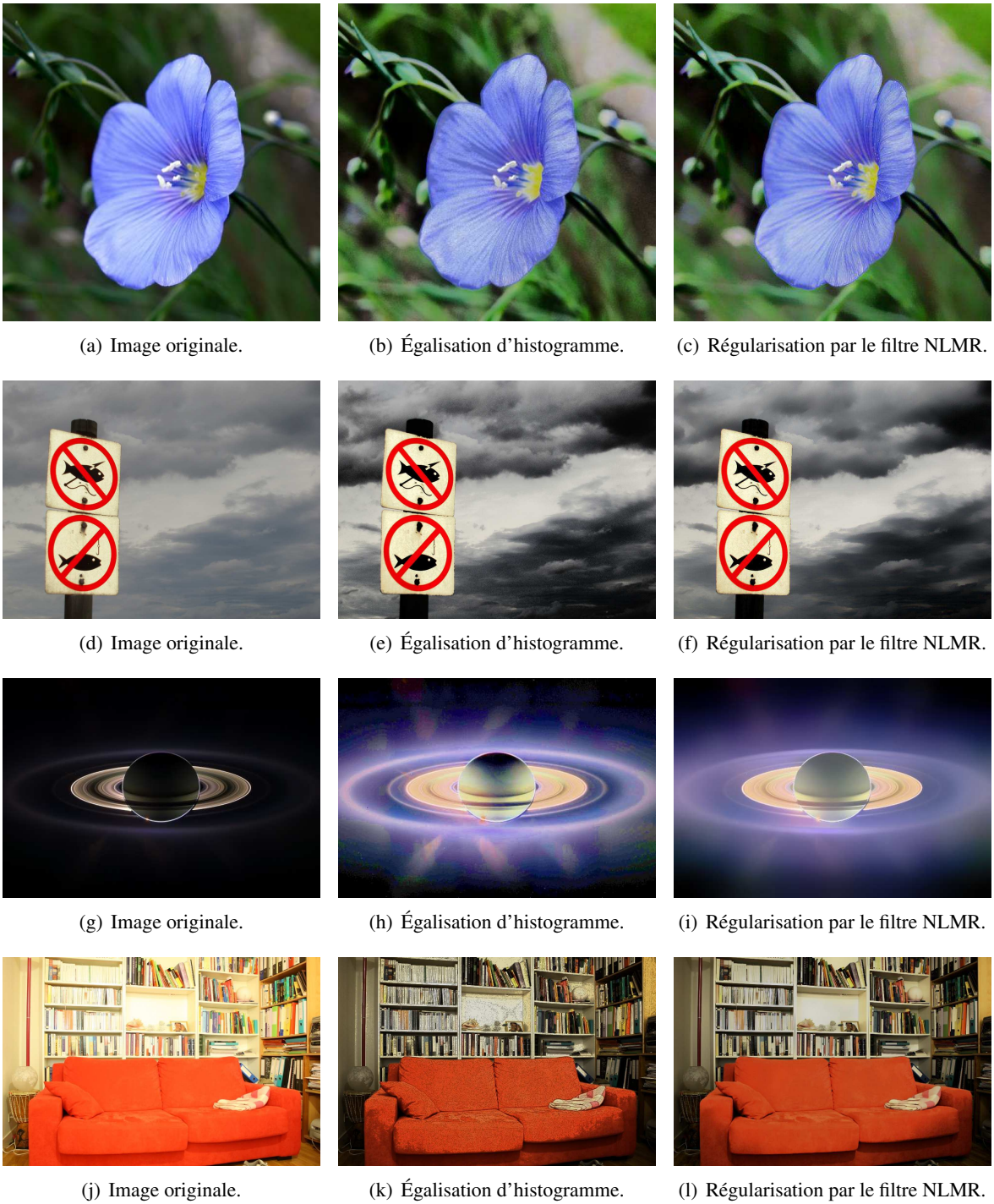
Nous concluons cette partie expérimentale en présentant les résultats de la régularisation par le filtre NLMR sur les autres images présentées en section 8.1.3, où l'égalisation d'histogramme seule donne un résultat peu satisfaisant. En figure 8.16(c), on constate que le filtre NLMR permet d'éliminer le bruit fortement rehaussé dans le ciel. En figure 8.16(f), il est intéressant de noter que le filtre NLMR restaure la texture du mur qui avait été complètement dégradée par le transfert. D'autres exemples de régularisation sont également donnés en figure 8.17.



(a) Image originale. (b) Égalisation d'histogramme. (c) Régularisation par le filtre NLMR.

(d) Image originale. (e) Égalisation d'histogramme. (f) Régularisation par le filtre NLMR.

FIG. 8.16 – *Dans la colonne de gauche figurent les images sources. Les images obtenues après égalisation d'histogramme sont au centre. La colonne de droite montre le résultat de l'application itérée du filtre NLRM jusqu'à convergence en utilisant des patchs de 1 pixel.*



(a) Image originale. (b) Égalisation d'histogramme. (c) Régularisation par le filtre NLMR.

(d) Image originale. (e) Égalisation d'histogramme. (f) Régularisation par le filtre NLMR.

(g) Image originale. (h) Égalisation d'histogramme. (i) Régularisation par le filtre NLMR.

(j) Image originale. (k) Égalisation d'histogramme. (l) Régularisation par le filtre NLMR.

FIG. 8.17 – *Dans la colonne de gauche figurent les images sources. Les images obtenues après égalisation d'histogramme sont au centre. La colonne de droite montre le résultat de l'application itérée du filtre NLRM jusqu'à convergence en utilisant des patchs de 1 pixel.*

8.3.2 Transfert de couleurs

Nous présentons dans ce paragraphe le résultat du filtrage NLMR pour les images obtenues par transfert de couleurs selon l'algorithme IDT [PKD07]. Contrairement au cas de l'égalisation d'histogramme étudié au paragraphe précédent, le transfert de couleurs concerne le transport sur les trois canaux de l'image source. Nous avons vu que les phénomènes de réhaussement de bruit, de perte de détails ou d'apparition d'artefact de compression étaient donc plus importants. Nous allons commencer par examiner ces différents aspects à l'aide de quelques exemples, afin de montrer l'intérêt de notre approche basée sur la régularisation de la carte de transport. Ensuite, nous nous intéresserons plus particulièrement au problème de la proportion des couleurs entre la palette de l'image source et celle de l'image de style.

Nous avons présenté au paragraphe 8.1.3 deux images sources pour lesquelles le transfert de couleurs fait apparaître des artefacts (fausses couleurs, structures en bloc), principalement en raison de la compression JPEG (voir les figures 8.18 et 8.19). C'est un phénomène qui se produit dès que l'image de style possède une palette de couleur distincte de celle de l'image source.

L'algorithme de *regraining* de Kokaram et Pitié [PKD07] ne prend pas en compte ce phénomène : l'image régularisée conserve donc tous les artefacts apparus avec le transfert (voir les figures 8.18(d) et 8.19(d)). Au contraire, notre approche basée sur la régularisation de la carte de transport permet d'atténuer les structures (bruit ou artefacts) qui étaient imperceptibles dans l'image d'origine. Dans les figures 8.18(e) et 8.19(e), l'image régularisée est obtenue en appliquant itérativement le filtre NLMR jusqu'à convergence, avec un voisinage de 10 px, des patchs de 1 px et $\sigma = 10$. Dans ces deux exemples, le ciel retrouve un dégradé plus naturel. Dans l'exemple du coucher de soleil (figure 8.18(e)), l'approche permet de reconstituer la brume autour des rochers.

À titre d'exemple, nous avons également utilisé le filtre NLMR avec des patchs de 9 pixels ($n = 3$), dont le résultat est donné en figure 8.18(f). Une fois encore, il y a peu de différences visibles en modifiant la taille du patch. Pour cette raison, nous allons dans la suite présenter les résultats de notre approche avec $n = 1$ seulement.

L'exemple suivant a pour but de démontrer la faculté de notre approche à réhausser les détails. Nous revenons sur l'exemple des poissons clown étudié en première section de ce chapitre (en haut de la figure 8.20 sont placées l'image source, l'image de style et l'image résultant du transfert). Le méthode de *regraining* permet de restaurer une partie des détails (image 8.20(d)), mais en comparaison, notre approche produit un résultat plus naturel (image 8.20(e)). En particulier, on notera que les structures fines sur les tentacules de l'anémone de mer sont correctement restituées, tandis que les détails perdus par la diminution de contraste au centre de l'image sont restaurés. Avec l'approche variationnelle de [PKD07], les petites structures sont systématiquement lissées en raison de la norme euclidienne qui est employée pour pénaliser le gradient. En effet, il est connu que la norme 2 favorise les transitions douces, contrairement à la norme 1 (à l'instar de la régularisation par variation totale [ROF92]).

Nous avons vu que l'opérateur NLMR (équation (8.3)) est obtenu par deux filtrages indépendants. La première action consiste à filtrer l'image $t(u)$ obtenue par transfert avec l'opérateur des moyennes non-locales NL_u, où les poids sont calculés à partir de l'image u. Avec k itération, ce filtrage s'exprime comme $\mathrm{NL}_u^k\big(t(u)\big)$. La seconde opération consiste à extraire les détails de l'image source u, en calculant $u - \mathrm{NL}_u^k(u)$. Pour illustrer chacune de ces opérations sur l'exemple des poissons clowns, nous donnons en figure 8.21 les deux images correspondant au résultat du filtrage de transfert (figure 8.21(b)) et à l'extraction de détails (figure 8.21(a)). Ces deux images sont simplement ajoutées pour former le résultat présenté en figure 8.20(e).

Un des problèmes majeurs du transfert de couleurs est la différence de proportion entre les palettes de couleurs de l'image source et de l'image de style. L'algorithme de *regraining* ne permet pas de traiter ce problème puisque le terme d'attache aux données utilisé dans la définition (8.1) est le résultat du transfert $t(u)$. Nous allons montrer aux travers des exemples suivants que notre approche permet d'atténuer ce problème lié à l'utilisation du transport.

Revenons tout d'abord sur l'exemple de la figure 8.22, que nous avons déjà étudié. Il s'agit *a priori* d'un exemple simple pour le transfert de couleurs puisque les deux images ont seulement deux compo-

santes de couleurs principales. Pourtant, nous avons vu que la différence subtile entre la taille des fleurs dans chacune des images suffit pour avoir un impact visuel conséquent. La régularisation par *regraining* réduit le bruit lié au transfert mais conserve tous les problèmes liés à cette différence de composition : apparition d'artefacts marron sur les pétales, excédents de couleur qui se traduit par des taches jaunes sur les feuilles et vertes sur les pétales (figure 8.22(d)). En appliquant plusieurs fois le filtre NLMR, les fausses couleurs sont fortement atténuées par le phénomène de diffusion, ce qui donne un rendu plus beaucoup naturel à l'image (figure 8.22(e)). Un résultat similaire est obtenu en figure 8.23, en utilisant une image de style différente pour la même image source.

> **Remarque 1 :**
> Contrairement aux cas précédents, nous n'avons pas itéré le filtre NLMR jusqu'à convergence. En effet, en raison de la diffusion liée au problème du transfert de la couleur, la convergence devient très lente. Ceci montre que quelques itérations du filtre NLMR suffisent à obtenir un résultat satisfaisant.

Nous avons présenté au tout début de ce chapitre un exemple de transfert de la palette couleur d'un tableau de Gauguin sur un tableau de Renoir. Le résultat de ce transfert par l'algorithme IDT est donné en figure 8.24(c). Contrairement au cas précédent où les images étaient composées de couleurs dominantes, ces deux peintures ont une palette très riche. Pour cette raison, c'est l'ensemble de l'image qui est cette fois affecté par le problème de la proportion des couleurs. En particulier, l'excédent de couleurs bleues et orangées de la palette de Gauguin se retrouve distribué sur l'ensemble du tableau de Renoir.

Une comparaison de la régularisation par notre approche et par celle de Kokaram et Pitié est donnée en figure 8.24. Une fois encore, la régularisation par regraining (figure 8.24(d)) ne traite pas ce problème d'excédent de couleurs. Beaucoup de régions ayant une teinte homogène dans l'image originale se voient ainsi affectées plusieurs couleurs à la fois. Avec notre approche (figure 8.24(e)), ce problème est fortement atténué. Une fois encore, les détails perdus lors du transfert sont également restaurés.

Pour clore cette section expérimentale, nous allons étudier deux exemples difficiles où les proportions de couleurs entre les deux images sont très importantes. Le premier exemple est donné en figure 8.25, où le but est de transférer les couleurs d'une photographie d'épis de blé ensoleillés sur une image représentant un champ de blé sous un ciel gris. La composition de ces deux photographies est très inégale, et par conséquent le transfert de couleurs ne produit pas l'effet désiré, une partie du champ devenant bleu (figure 8.25(c)). La régularisation par regraining (figure 8.25(d)) n'évite pas ce phénomène.

En appliquant 10 itérations de notre filtre NLMR (figure 8.26(a)), le rendu de la texture du blé est bien meilleur mais une grande partie du champ reste bleu. Il est alors nécessaire de choisir un voisinage plus grand afin de permettre une diffusion étendue à tout le champ. La figure 8.26(b) montre le résultat obtenu en seulement deux itérations, avec un voisinage de rayon $\rho = 100$ au lieu de 10 pixels. La teinte du champ de blé devient ainsi homogène, mais au prix d'une complexité plus élevée. Ceci montre la limite de la mise en œuvre de notre approche : pour des raisons de temps de calcul, nous sommes obligés de restreindre le voisinage de comparaison des patchs, ce qui limite la portée de la diffusion.

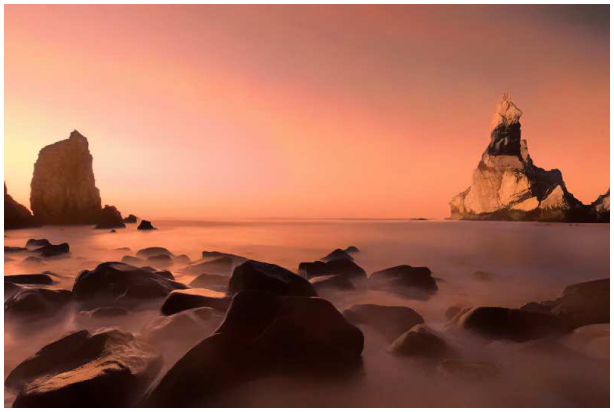
Pour illustrer spécifiquement ce dernier point, nous proposons une dernière expérience. Dans la figure 8.27, le transfert de la couleur de l'image de style sur l'image source est problématique car des régions homogènes entières se voient affecter une couleur incorrecte. Cette fois, l'utilisation de notre filtre sur un petit voisinage (figure 8.27(d)) comme sur un grand (figure 8.27(e)) ne permet pas de résoudre ce problème d'affectation de couleur. Ce type de problème requiert la mise au point d'une nouvelle approche de transfert de couleurs.

(a) Image originale.

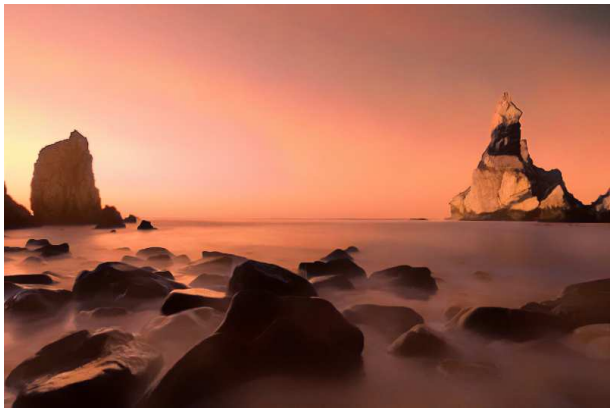
(b) Image de style couleur.

(c) Transfert de la couleur par IDT [PKD07].

(d) Régularisation par regraining [PKD07].



(e) Régularisation par itération du filtre NLMR ($n = 1$).



(f) Régularisation par itération du filtre NLMR ($n = 3$).

FIG. 8.18 – *Régularisation du transfert de couleurs. Cet exemple met en évidence l'intérêt de notre approche pour la suppression des artefacts de compression JPEG.*

(a) Image originale.

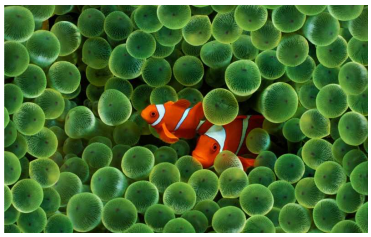
(b) Image de style couleur.

(c) Transfert de la couleur.

(d) Régularisation par regraining [PKD07].

(e) Régularisation par itération du filtre NLMR ($n = 1$).

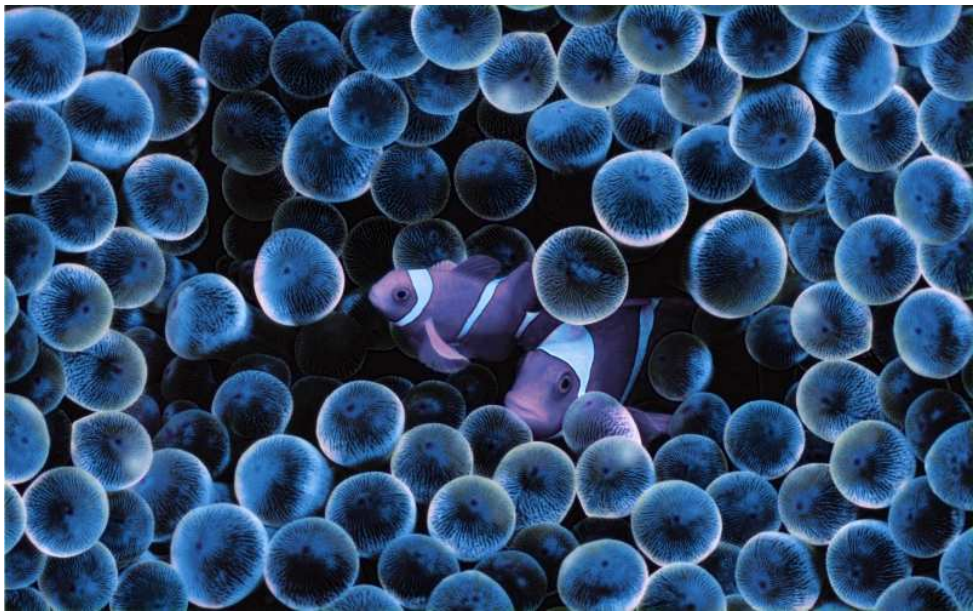
FIG. 8.19 – *Régularisation du transfert de couleurs. Cet exemple met en évidence l'intérêt de notre approche pour la suppression des artefacts de compression JPEG.*



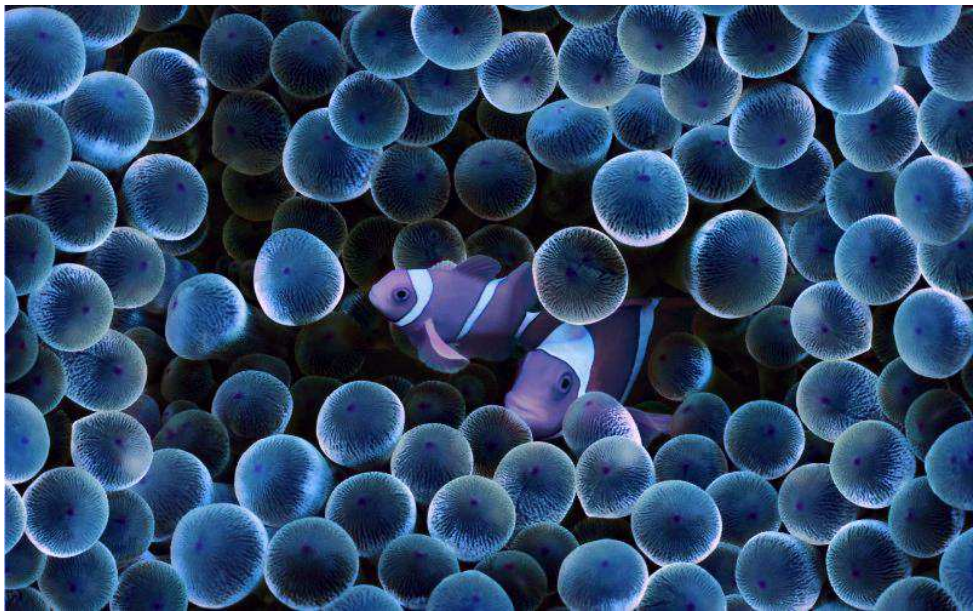
(a) Image originale.

(b) Image de style couleur.

(c) Transfert de la couleur par IDT.

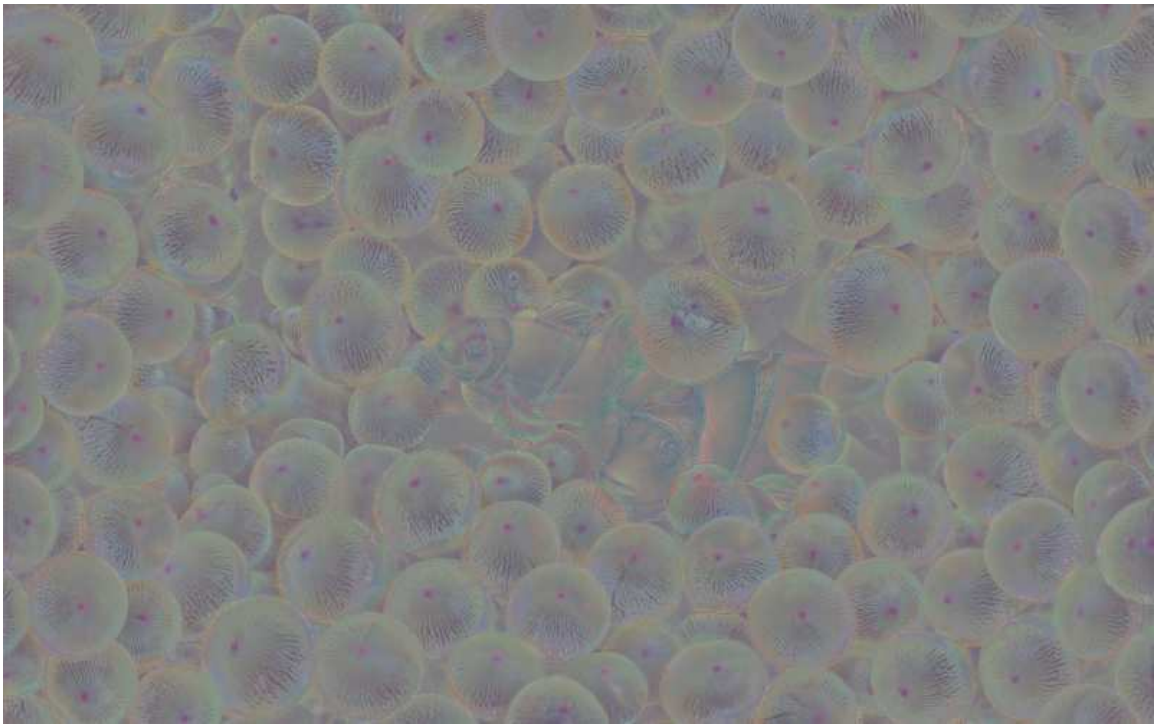


(d) Régularisation par regraining.

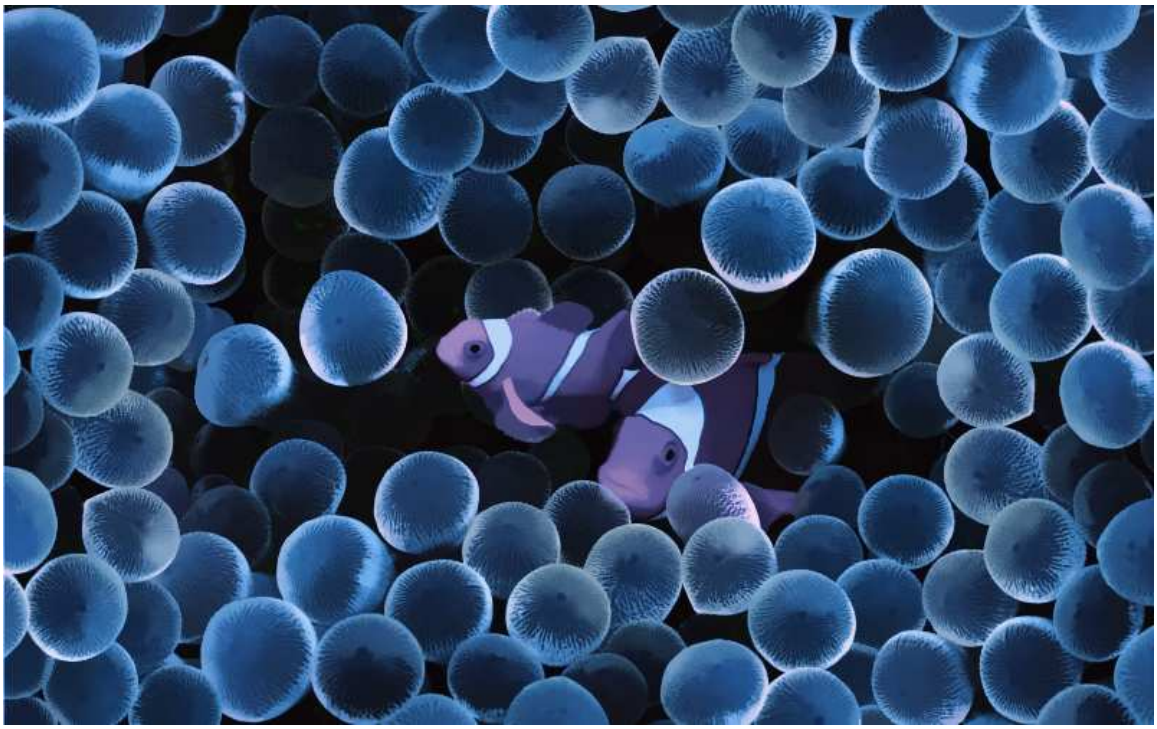


(e) Régularisation par itération du filtre NLMR.

FIG. 8.20 – *Régularisation du transfert de couleurs. Cet exemple met en évidence l'intérêt de notre approche pour le réhaussement des détails.*



(a) Extraction des détails de l'image source u (la moyenne a été rehaussée pour permettre la visualisation).



(b) Filtrage de l'image $t(u)$.

FIG. 8.21 – *Illustration des deux actions menées par le filtre NLMR. La figure 8.21(a) montre les détails extraits de l'image source. La figure 8.21(b) montre le résultat du filtrage de l'image de transfert $t(u)$. On obtient ainsi une décomposition de type « cartoon + texture ». Ces deux images sont simplement ajouté pour obtenir l'image finale.*

(a) Image originale.

(b) Image de style couleur.

(c) Transfert de la couleur par IDT.

(d) Régularisation avec regraining.

(e) Régularisation par itération du filtre NLMR.

FIG. 8.22 – *Régularisation du transfert de couleurs. Cet exemple met en évidence l'intérêt de notre approche pour le problème de la proportion de couleur.*

(a) Image originale.

(b) Image de style couleur.

(c) Transfert de la couleur par IDT.

(d) Régularisation avec regraining.

(e) Régularisation avec filtre NLMR.

FIG. 8.23 – *Régularisation du transfert de couleurs. Cet exemple met en évidence l'intérêt de notre approche pour le problème de la proportion de couleur.*

(a) Image originale.

(b) Image de style.

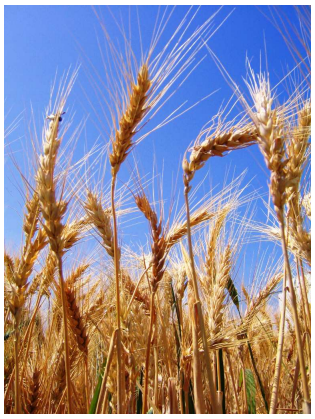
(c) Transfert de la couleur par IDT.

(d) Régularisation avec regraining.

(e) Régularisation avec filtre NLMR.

FIG. 8.24 – *Régularisation du transfert de couleurs. Cet exemple met en évidence l'intérêt de notre approche pour le problème de la proportion de couleur.*

(a) Image originale.



(b) Image de style couleur.

(c) Transfert sans régularisation.

(d) Régularisation par regraining [PKD07].

FIG. 8.25 – *Dans cet exemple, la différence de proportion entre l'image de style et l'image source se traduit par un excès de teinte bleu dans l'image obtenu par l'algorithme IDT (figure 8.25(c)). La restauration obtenue par regraining [PKD07] ne permet pas de traiter ce type de problème (figure 8.25(d)).*

(a) Itération de NLRM sur un disque de rayon $\rho = 10$ px (10 itérations).

(b) Itération de NLRM sur un disque de rayon $\rho = 100$ px (2 itérations).

FIG. 8.26 – *En utilisant le filtre NLMR introduit dans ce chapitre, le phénomène d'excès de couleur peut être atténué. La figure 8.26(a) présente le résultat de notre algorithme pour* 10 *itérations, avec les mêmes réglages que pour les expériences précédentes (*$n = 1$, $\sigma = 10$ *et* $\rho = 10$*). En raison de la restriction spatiale, une partie du champ de blé reste toujours teinté de bleu. En utilisant un voisinage spatial plus large (*$\rho = 100$*), la diffusion est plus rapide et permet en seulement deux itérations d'éliminer la composante bleu (figure 8.26(b)).*

(a) Image originale.

(b) Égalisation d'histogramme.

(c) Transfert couleur puis régularisation par regraining.

(d) Itération de NLRM sur un disque de rayon $\rho = 10$ px.

(e) Itération de NLRM sur un disque de rayon $\rho = 100$ px.

FIG. 8.27 – *Limitation de la régularisation du transport* *Dans cet exemple, la différence de proportion est trop importante entre les palettes de couleurs de l'image source et de l'image de style. L'image obtenue par l'algorithme IDT (figure 8.27(c)) possède des régions entièrement affectées à de fausses couleurs. Dans un tel cas de figure, notre approche ne permet pas de restaurer l'homogénéité de l'image source (figure 8.27(d)), même en utilisant un plus grand voisinage (figure 8.27(e)).*

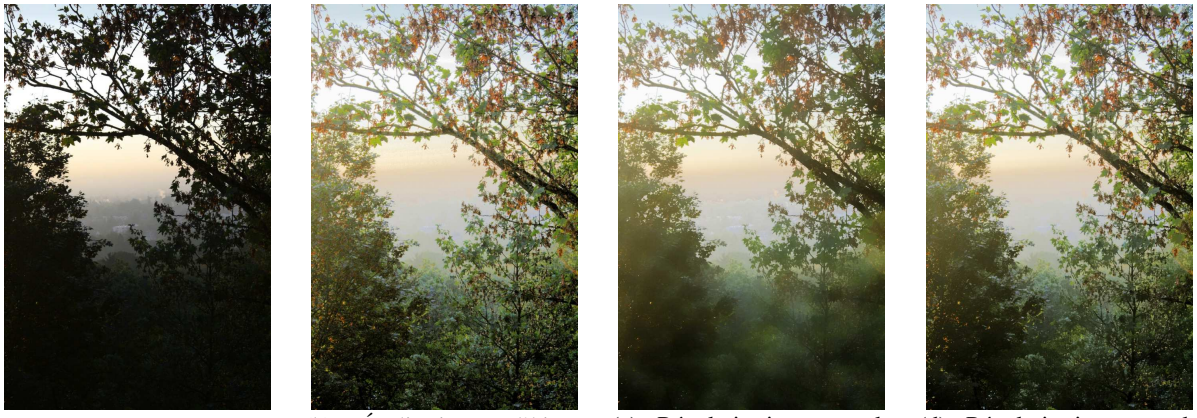
8.4 Perspectives

Nous avons mis en évidence, dans la précédente section, l'intérêt de notre approche pour la régularisation du transfert de caractéristiques. Certaines expériences que nous avons réalisées suggèrent quelques pistes de travail à explorer.

La première concerne l'étude de la convergence de schéma itératif. Nous avons vu sur de nombreux exemples que l'utilisation itérative du filtre NLMR converge vers une solution satisfaisante. Étant donné que le schéma que nous utilisons présente certaines similarités avec celui étudié dans [SSN09], il serait intéressant de faire le lien avec leur résultat.

Par ailleurs, cette approche est coûteuse en temps de calcul mais de nombreuses approximations ont été proposées pour accélérer le calcul des filtres bilatéraux ou des non-local means. Il serait intéressant de voir quelles dégradations entraînent ces différentes approximations. Nous avons également vu sur plusieurs exemples que quelques itérations suffisent parfois pour obtenir un résultat satisfaisant, ce qui permet de gagner du temps.

Une autre piste de travail concerne le choix de σ. Nous avons utilisé dans toutes les expériences $\sigma = 10$, sans nous soucier de choisir *a posteriori* la valeur donnant le meilleur résultat visuel. Nous avons observé que l'utilisation itérative du filtre NLMR rend le choix de ce paramètre moins critique qu'il ne l'est lorsqu'il s'agit de ne réaliser qu'une itération. Toutefois, la valeur de σ a un impact important sur le résultat final comme le montre a figure 8.28, ce qui rend la définition automatique de ce paramètre nécessaire.



(a) Image originale. (b) Égalisation d'histogramme. (c) Régularisation par le filtre NLMR avec $\sigma = 10$. (d) Régularisation par le filtre NLMR avec $\sigma = 2$.

FIG. 8.28 – *Illustration de la nécessité de définir automatiquement le paramètre σ.*

Il serait également intéressant de définir le paramètre σ de manière locale, à l'image de [KB08]. Ceci permettrait, d'une part, d'accélérer la convergence du processus, et par ailleurs, de préserver les structures peu contrastées. Pour illustrer ce point, nous donnons un exemple de transfert supervisé de couleur en figure 8.29. Dans cet exemple, l'image source est grossièrement colorisé à l'aide d'un logiciel d'édition d'image (figure 8.29(b)), puis l'image est régularisé avec notre approche. Le résultat en figure 8.29(c) montre que le choix de $\sigma = 10$ convient pour certaines régions où la diffusion reste confinée à chaque régions homogènes. Par contre, cette même valeur $\sigma = 10$ est trop élevée pour d'autres régions : certaines les bandes blanches sur le ballon se retrouvent ainsi mélangées aux régions voisines.

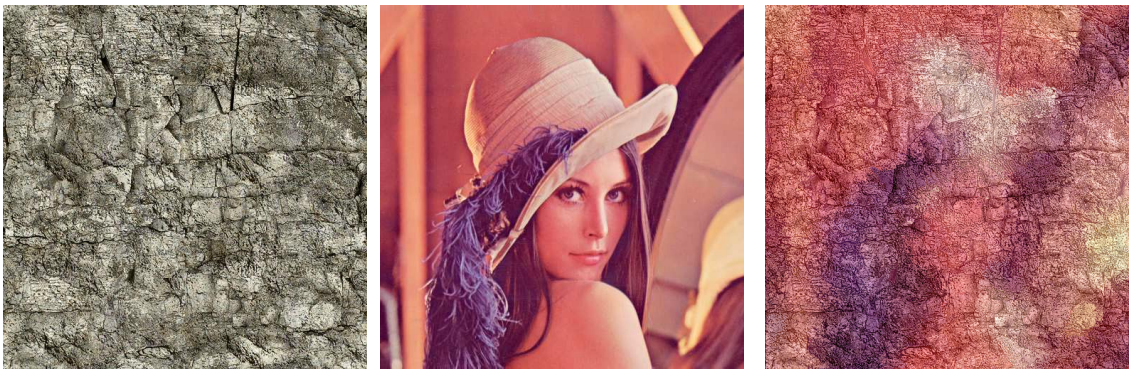
Enfin, il serait intéressant d'étendre l'utilisation du filtre NLMR à d'autres applications de transfert. Pour le transfert de style proposé dans [BPD06], les auteurs évoquent l'apparition d'artefacts dus à la compression JPEG. Leur approche pourrait bénéficier de notre méthode pour éviter ce problème. De nombreuses approches de transfert de textures ont été proposées, à l'image de [HJO+01, EF01]. Une variante se basant sur notre approche pourrait être définie, comme l'illustre la figure 8.30. Une dernière application envisageable est la réduction de la dynamique des images couleurs [BSL08, DD02] pour leur visualisation (*tone mapping*) et leur impression (*gamut mapping*).

(a) Image originale u.

(b) Image $t(u)$ (édition manuelle par logiciel).

(c) Régularisation avec le filtre NLMR ($\sigma = 10$).

FIG. 8.29 – *Illustration de la nécessité de définir localement le paramètre σ.*



(a) Image de texture considérée comme l'image source.

(b) Image de style.

(c) Transfert par filtre NL (patch 3×3 pixels, $\sigma = 100$).

FIG. 8.30 – *Exemple de transfert de texture avec le filtre NLMR.*

Conclusion et perspectives

Conclusion

Nous avons abordé dans cette thèse la problématique de la comparaison d'images selon des représentations locales et globales.

Nous avons présenté un système complet de reconnaissance d'objets à partir de descripteurs locaux, dont nous avons étudié les différentes étapes en détail. Une nouvelle mesure de dissimilarité pour la comparaison de descripteurs de type SIFT a été introduite, exploitant la structure circulaire des histogrammes d'orientation du gradient qui les composent. Nous avons montré, comparativement à plusieurs distances, l'intérêt de cette mesure sur une base de 3 millions de descripteurs. Un critère de mise en correspondance de descripteurs locaux autorisant les mises en correspondances multiples a été proposé. Ce critère de décision à été défini dans le cadre théorique de la détection *a contrario*, ce qui permet de contrôler l'espérance du nombre de fausses détections. L'intérêt de notre approche, vis-à-vis de celles existantes, a été démontré expérimentalement par la comparaison de plusieurs dizaines de milliers de paires d'images. Nous avons également proposé un algorithme de groupement de correspondances, MAC-RANSAC, rendant possible la reconnaissance d'objets multiples. Cet algorithme repose sur une mesure de qualité géométrique initialement proposée dans [MS04] pour la géométrie épipolaire. Nous avons tout d'abord adapté cette mesure aux transformations planes, qui sont souvent utilisées pour la reconnaissance d'objets. Ceci nous a également permi de traiter le problème de la sélection de modèles géométriques. Nous avons en outre introduit divers critères permettant d'améliorer la robustesse et la précision de la reconnaissance d'objets multiples. Les résultats expérimentaux de cet algorithme ont été présentés sur différents types de données, et pour différents cas de figure. Notons enfin que le sytème obtenu présente le grand intérêt, de par l'utilisation de critères de détection *a contrario*, de ne nécessiter aucun réglage de paramètre de détection, contrairement aux autres algorithmes existants.

Nous nous sommes également intéressés à la comparaison d'images *via* le transport optimal entre des histogrammes globaux. Une étude de la distance de transport de Monge-Kantorovich dans le cas des histogrammes unidimensionnels et circulaires nous a permis de définir la distance CEMD. Cette distance a été utilisée pour définir une mesure de dissimilarité entre descripteurs SIFT. Une analyse comparative de la distance de transport EMD et des distances bin-à-bin pour la comparaison d'histogrammes a mis en évidence leurs avantages et leurs inconvénients respectifs. Nous avons ensuite illustré ces résultats par l'indexation de plusieurs bases d'images. Enfin, nous avons étudié le transfert de caractéristiques entre des images par transport optimal. Faisant le lien avec la précédente analyse, nous avons mis en lumière plusieurs problèmes liés au transport qui dégradent fortement le rendu visuel. Une méthode de régularisation non locale de la carte de transport (NLMR : Non Local Map Regularization) a été proposée pour traiter l'ensemble de ces problèmes. L'intérêt du filtre NLMR a été démontré pour les applications d'ajustement de contraste et de transfert de palette de couleurs.

Perspectives

Nous avons d'ores et déjà évoqué au cours de ce manuscrit quelques pistes de travail pour les différents thèmes abordés dans ce manuscrit.

Représentation locales des images Nous avons vu dans la partie expérimentale du chapitre 4 que certains objets ne sont pas détectés en raison du manque de points d'intérêt détectés. Ceci principalement en raison de la non invariance au changement de contraste du détecteurs de points (voir l'annexe B). Une perspective intéressante est donnée par l'approche de détection *a contrario* proposée dans [Cao04] pour la détection de coins. Cependant, cette méthode fondée sur la carte topographique doit être adaptée pour la définition d'une échelle caractéristique.

Un autre problème auquelle nous avons été confronté concerne la reconnaissance d'objets dans le cas d'un fort changement de point de vue. Il s'agit dans ce cas d'une limitation des descripteurs locaux de type SIFT que nous employons. L'utilisation des descripteurs ASIFT [MY09], qui possèdent une très grande robustesse pour de telles transformations, permettrait de s'affranchir de ce problème.

Dans certaines situations, les descripteurs de type SIFT ne permettent pas la reconnaissance d'un objet, par exemple en raison d'une inversion de contraste ou d'un fort changement d'éclairage. Il serait intéressant d'étudier la combinaison de ce type de représentation avec des descripteurs comme les morceaux de lignes [MSC+06] qui autorisent de tels changement.

Nous avons également mis en évidence le problème de la redondance des points d'intérêt détectés. Ce phénomène est lié au principe de maximalité utilisé en espace échelle (voir l'annexe B). Une piste est proposée dans [Sur07], qui consiste à suivre la trajectoire des coins détectés par le détecteur de Harris dans un espace-échelle affine (Affine Morphological Scale-Space [AGLM93]).

Reconnaissance d'objets Dans cette thèse, nous n'avons pas pris en considération le temps de calcul lié à la comparaison exhaustive des descripteurs SIFT entre deux images. Nous avons brièvement mentionné dans le chapitre 1 quelques algorithmes rapides de recherche approchée (*Approximate Nearest Neighbor*) qui ont été proposé pour réduire la complexité de ce type de recherche. L'utilisation de ce type d'approche est envisageable pour notre système, mais elle nécessite d'adapter notre critère de mise en correspondance proposé au chapitre 2. Plus précisément, nous devons étudier s'il est possible d'estimer de manière approchée les distributions empiriques des distances entre histogrammes à partir de quelques échantillons seulement, au lieu d'utiliser toute la base.

Une autre piste que nous souhaiterions suivre est l'application à la reconstruction 3D de notre critère de sélection de modèles géométriques, proposé au chapitre 4. À l'image de [RP05], il pourrait être utilisé pour la sélection de paires d'images ayant une ligne de base suffisante.

Régularisation du transport Nous avons proposé au chapitre 8 un filtre itératif de régularisation du transport (NLMR). La limitation principale de cette approche est son temps d'exécution. Nous avons évoqué en section 8.4 de nombreuses pistes pour accélérer cette régularisation, que nous aimerions mettre en œuvre.

Un autre point que nous aimerions étudier est la définition automatique, de manière locale, du paramètre d'échelle du filtre NLMR.

D'autres applications de notre approche sont également envisagées, dont notamment le rehaussement de détails pour l'affichage des images de grande dynamique (*tone mapping*) et le transfert de texture.

Fonction de coût pour le transport optimal Nous avons souligné l'importance pour le transport optimal du poids des modes principaux dans les histogrammes comparés. Ce phénomène limite à la fois les performances de la comparaison d'histogrammes pour l'indexation des images, mais également pour le transfert de caractéristique.

Il serait intéressant de se pencher sur le rôle du coût exponentiel utilisé pour le calcul du transport en indexation d'images (voir le chapitre 6) vis-à-vis de ce phénomène. À notre connaissance, il n'existe pas d'analyse du transport dans un tel cas.

Dans le cadre du transfert de palettes de couleurs, nous envisagons d'étudier le moyen d'intégrer des contraintes propres à cette application pour l'estimation du transport optimal.

Annexes

Annexe A

Présentation de la méthodologie *a contrario*

Nous présentons dans cette annexe le principe de la détection *a contrario* qui est appliqué pour la mise en correspondance de descripteurs locaux dans le chapitre 2, pour le groupement de mises en correspondance dans le chapitre 4, et pour la détection des orientations principales des points d'intérêt en annexe B.

A.1 Motivation

Une grande variété d'applications en analyse d'images requièrent une étape de décision. Nous avons vu au travers de divers exemples que le critère de décision est souvent ramené à un simple seuil sur une mesure de qualité : les critères de mise en correspondance présentés au chapitre 1, l'algorithme RANSAC [FB81] (chapitre 3) et le détecteur de coins de Harris [HS88] (annexe B). Différentes stratégies sont possibles pour définir la valeur optimale d'un tel seuil de décision. Elles dépendent principalement des performances visées (rapidité d'exécution, robustesse aux perturbations liées au bruit, prise en compte des tests multiples effectués, taux d'acceptation ou de rejet, *etc*.). Afin de définir des seuils de détection adaptatifs, de nombreuses approches fondées sur la théorie statistique de la décision ou sur l'inférence bayésienne ont été proposées.

Ces approches reposent le plus souvent sur la définition d'un modèle générique pour décrire la classe d'objets que l'on souhaite valider. Par exemple, avec l'algorithme MLESAC [TZ00], un modèle gaussien est utilisé pour décrire les erreurs résiduelles correspondant aux données que l'on souhaite sélectionner, et un modèle uniforme pour les erreurs correspondant aux données aberrantes. Or, il existe beaucoup d'applications pour lesquelles il est difficile de définir de tels modèles. Soit parce qu'aucune information *a priori* n'est connue, soit parce que la classe d'objets recherchés présente une forte variabilité.

Dans le cadre de l'étude du groupement perceptuel par les lois de la Gestalt, Desolneux, Moisan et Morel [DMM03a] ont proposé une nouvelle approche pour définir des seuils adaptatifs de détection. Cette méthodologie, désormais connue sous le nom de « méthode *a contrario* », a pour principe de mesurer la qualité d'un groupe de caractéristiques en considérant le rejet d'un modèle de fond. Ce modèle aléatoire représente une situation générique pour laquelle on ne souhaite détecter aucun groupe. Cette approche, qui se fonde sur le principe de Helmholtz, a par la suite été reprise dans de nombreux champs d'applications [DMM08] (voir la section A.2.4).

Dans la section suivante, nous détaillons ce principe de décision, puis nous décrivons sa mise en œuvre générique.

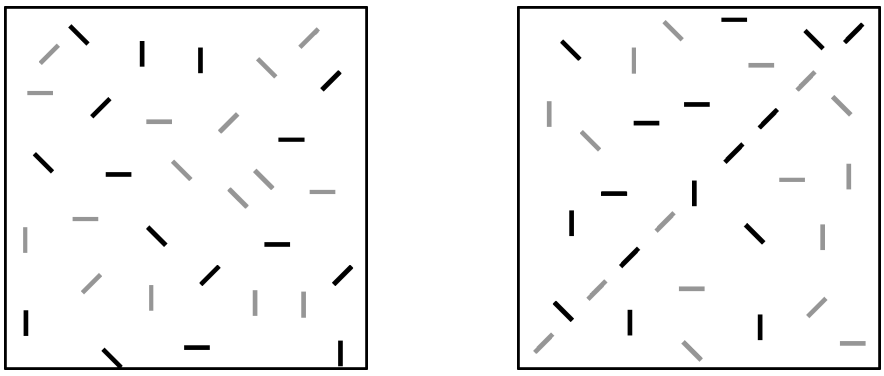
A.2 Présentation générale

Nous allons tout d'abord introduire le principe de Helmholtz sur lequel repose la théorie de la détection *a contrario*.

A.2.1 Principe de Helmholtz

Un principe général utilisé en analyse d'images postule que l'on ne perçoit une structure que lorsque celle-ci a très peu de chance de se produire par hasard. L'une des premières formalisations de ce principe, auquel les auteurs de [DMM08] se refèrent en tant que « principe de Helmholtz », a été proposée par Lowe [Low85].

Pour illustrer ce principe, nous donnons en figure A.1(a) une configuration de segments obtenue en tirant aléatoirement (de manière indépendante et uniforme) leur position, leur couleur, leur orientation. Dans une telle image, aucune structure ne se dégage car la configuration obtenue est générique. Au contraire, en observant la configuration de la figure A.1(b), on remarque immédiatement une structure correspondant à l'alignement de certains des segments. Une telle structure est très improbable sous l'hypothèse que les segments sont indépendants : on parle alors de structure « significative ». Une structure significative peut prendre de multiples formes, mais elle se définit toujours comme un groupe d'éléments qui possèdent une caractéristique similaire, dans le cas présent l'orientation des segments. Selon la caractéristique considérée, d'autres types de structures peuvent être perçues.



(a) (b)

FIG. A.1 – *Illustration du groupement perceptuel (principe de Helmholtz).* *Aucun structure ne se dégage de l'image A.1(a) où les segments sont tirés aléatoirement, de manière indépendante. Au contraire, dans le cas de l'image A.1(b), on ne peut s'empêcher de regrouper certains segments.*

A.2.2 Mise en œuvre

L'objet de la théorie de la détection *a contrario* [DMM08] est la mise en œuvre du principe de perception que nous venons de présenter dans un cadre statistique. Ce principe repose sur deux notions fondamentales : un **modèle de fond**, qui décrit un processus génératif pour lequel on ne perçoit pas de structure significative, et une mesure de similarité visuelle des caractéristiques qui composent un groupe. Desolneux, Moisan et Morel proposent d'exploiter ces deux notions pour mesurer la significativité d'un groupe.

Hypothèse nulle En pratique, un modèle de fond est tout d'abord défini. Généralement, il repose sur une hypothèse d'indépendance mutuelle des caractéristiques examinées. Il s'agit d'une *hypothèse nulle*, notée $\mathcal{H}_0$, qui est courante en théorie de la décision statistique. Dans notre exemple précédent avec les segments, le modèle de fond traduit le fait que les caractéristiques de couleur, de position et

d'orientation des segments sont indépendantes entres elles et d'un segment à l'autre. Suivant le type de données analysées, d'autres modèles de fond sont envisageables.

Mesure de similarité Une mesure de similarité doit ensuite être définie afin de quantifier l'adéquation des éléments d'un groupe testé vis-à-vis du type de structure que l'on souhaite détecter. Si l'on souhaite par exemple détecter des alignements de segments (exemple de la figure A.1(b)), cette mesure va dépendre de l'orientation et de la position relative des segments du groupe testé.

Mesure de significativité Le cœur des méthodes *a contrario* consiste à définir une mesure de *significativité* permettant le contrôle des fausses détections lors de l'examen d'un grand nombre de groupes. Pour cela, on considère tout d'abord la probabilité associée à un groupe testé. On désigne par G_i, le i-ième groupe testé. On note X la mesure de similarité des éléments d'un groupe, et x_i la mesure associée au groupe G_i. Il est possible d'exprimer la probabilité sous l'hypothèse nulle d'observer le groupe G_i avec la mesure de qualité x_i. Plus généralement, on calcule la « p-valeur » qui correspond à la probabilité d'observer un groupe de caractéristiques aléatoires dont la mesure de qualité est au moins aussi grande sous l'hypothèse nulle. On note cette probabilité : $\mathbb{P}_{\mathcal{H}_0}(X \geq x_i)$. Notons que cette probabilité dépend souvent du *cardinal* du groupe G_i, c'est-à-dire du nombre d'éléments de ce groupe.

Un groupe testé est d'autant plus significatif que cette probabilité est petite. Une approche classique consiste à seuiller cette probabilité pour décider si l'on doit valider un groupe. En pratique cependant, il est difficile de définir un tel seuil. Si le seuil est un peu trop permissif, le nombre de fausses détections (ou encore « faux-positifs ») risque d'exploser pour un nombre de tests important. En effet, pour revenir à notre cas d'étude, en supposant que n segments sont tirés, on a alors $N = 2^n - n - 1$ groupes de segments à tester.

L'originalité de la théorie de la détection *a contrario* est de prendre en compte ce nombre de tests de manière à contrôler l'espérance du nombre de fausses alarmes. Pour cela, une mesure de significativité appelée **NFA**, pour « Nombre de Fausses Alarmes », est alors définie pour un groupe en fonction du nombre de tests et de la probabilité qui lui est associée. Dans sa forme la plus simple, le NFA du groupe G_i s'exprime comme le produit du nombre total de groupes testés N, et de la probabilité $\mathbb{P}_{\mathcal{H}_0}(X \geq x_i)$:

$$\text{NFA}(G_i, x_i) := N \times \mathbb{P}_{\mathcal{H}_0}(X \geq x_i) \ . \tag{A.1}$$

Cette définition du NFA est par exemple utilisée pour la détection d'alignements dans [DMM00], et au chapitre 2 pour notre critère de mise en correspondance de descripteurs locaux (voir l'expression (2.3)).

On appelle alors « ε-significatif », un groupe dont le NFA est plus petit que ε. La définition (A.1) assure une propriété fondamentale du NFA :

> *l'espérance du nombre de groupes ε-significatifs vérifiant l'hypothèse nulle est plus petite que ε.*

On appelle « fausse alarme » un groupe validé qui vérifie le modèle de fond (c'est-à-dire qui ne correspond pas à une structure significative). Le critère de décision consiste alors à ne valider que les groupes ε-significatifs, ce qui permet de s'assurer, en moyenne, que le nombre de fausses alarmes est plus petit que ε. Ceci revient à définir un seuil sur la probabilité égal à $\frac{\varepsilon}{N}$.

Notons que dans sa forme la plus générale, le NFA est défini de la manière suivante :

$$\text{NFA}(G_i, x_i) := n_i \times \mathbb{P}_{\mathcal{H}_0}(X \geq x_i) \ , \tag{A.2}$$

où l'ensemble $\{\frac{\varepsilon}{n_i}\}_i \leq N$ représente une famille de seuils sur la probabilité, choisis de manière à vérifier la relation suivante :

$$\sum_{i \leq N} \frac{1}{n_i} \leq 1 \ . \tag{A.3}$$

Comme c'est souvent le cas des applications de la méthode *a contrario*, le critère de groupement de mises en correspondance MAC-RANSAC, introduit au chapitre 4, utilise une famille de tests qui dépend du cardinal du groupe testé (voir l'expression (4.12)).

Maximalité Un principe de maximalité est généralement nécessaire en pratique pour définir le meilleur groupe correspondant à la structure détectée. Ceci permet d'éviter les détections redondantes de la même structure (voir par exemple le détecteur de points d'intérêt en annexe B). Dans le cadre de la méthodologie *a contrario*, l'application de ce principe revient à sélectionner le groupe dont le NFA est le plus faible, de manière locale ou globale.

Dans ce manuscrit, nous avons eu plusieurs fois recours à ce principe de maximalité. Tout d'abord, dans le cadre de la mise en correspondance de descripteurs locaux (chapitre 2), nous avons vu que ce principe se traduisait par le fait de ne sélectionner que le plus proche voisin dans une base de descripteurs. Par la suite, nous avons défini au chapitre 4 un principe d'exclusion spatial (définition 11), afin d'éliminer les correspondances redondantes. Dans ce même chapitre, nous avons proposé un algorithme de groupement qui applique le principe de maximalité à deux reprises. Dans un premier temps, un groupe est identifié en tant que minimum global du NFA sur l'ensemble des données. Ensuite, un critère de découpage est utilisé pour tester si ce groupe résulte de la fusion de deux groupes.

A.2.3 Intérêts de la détection a contrario

La détection *a contrario* présente plusieurs avantages en comparaison d'autres méthodes existantes.

Tout d'abord, elle ne nécessite que la définition d'un modèle de fond, qui est un modèle naïf fondé sur une hypothèse d'indépendance mutuelle. D'autres approches, à l'image de l'approche ANOVA (*Analysis of variance*), requièrent au contraire la définition de plusieurs hypothèses : une hypothèse nulle $\mathcal{H}_0$, mais également une hypothèse alternative $\mathcal{H}_1$ qui décrit la structure que l'on souhaite détecter.

En outre, la méthode *a contrario* prend en compte le nombre de tests théoriquement effectués de manière à contrôler l'espérance du nombre de fausses alarmes. Nous avons vu que cela revenait à pondérer la probabilité par un nombre de tests (équation (A.2)), ce qui correspond à la correction dite de Bonferroni [Bon36]. Dans le cadre des tests multiples, cette approche consiste à définir un seuil α sur la probabilité $\mathbb{P}_{\mathcal{H}_0}$ qui est appelé niveau de détection . Afin de prendre en compte l'ensemble des N tests effectués, Bonferroni montre qu'il est nécessaire de normaliser le niveau de détection pour chacun des tests : $\mathbb{P}_{\mathcal{H}_0} \leq \frac{\alpha}{N}$.

Le seuil de détection ε est très souvent fixé à 1, ce qui conduit généralement à dire que la méthode *a contrario* est « sans paramètres ». L'autre avantage procuré par la définition du NFA est la définition de seuils de détection adaptatifs sur la mesure de qualité X. Pour chaque groupe i, le seuil de détection $\hat{x}_i$ est ainsi défini comme :

$$\hat{x}_i = \underset{x}{\operatorname{argmax}}\{\mathbb{P}_{\mathcal{H}_0}(X \geq x) \leq \frac{\varepsilon}{n_i}\} \ .$$

> **Remarque 1 :**
> Rappelons que la variable aléatoire X peut dépendre des caractéristiques du groupe d'indice i testé, telles que son cardinal.

Autres approches Il est intéressant de noter que d'autres approches ont été proposées dans le but de contrôler les fausses alarmes (généralement désignées comme des erreurs de première espèce, ou de type I). Afin de comprendre la différence entre ces approches, nous rappelons dans le tableau suivant les résultats des tests en fonction de la validité de l'hypothèse nulle $\mathcal{H}_0$ (rappelons qu'un test est positif lorsqu'il conduit à la validation d'un groupe, soit dans notre cas lorsque l'hypothèse nulle est rejetée).

On parle généralement de *Per Comparison Error Rate* (PCER) pour désigner la probabilité de valider une fausse alarme pour un unique test, et *Experimentwise Error Rate* (EWER) la probabilité de valider au moins une fausse alarme pour l'ensemble des N tests.

La méthode *a contrario* consiste à contrôler l'espérance du nombre de fausses alarmes, soit $\mathbb{E}(N_{fp})$. Une approche alternative, *False Discovery Rate* (FDR), considère l'espérance du taux de fausses alarmes validées, soit $\mathbb{E}(\frac{N_{fp}}{N_p})$.

TAB. A.1 – *Répartition des différents résultats sur l'ensemble de N tests*

Résultats sur N tests / Validité de $\mathcal{H}_0$	Positif (Validé)	Négatif (Rejeté)	Total
$\mathcal{H}_0$ invalide	Vrai-Positif (N_{vp})	Faux-Négatif (N_{fn})	$N_v = N_{vp} + N_{fn}$
$\mathcal{H}_0$ valide	Faux-Positif (N_{fp})	Vrai-Négatif (N_{vn})	$N_f = N_{fp} + N_{vn}$
Total	$N_p = N_{vp} + N_{fp}$	$N_n = N_{vn} + N_{fn}$	N

Il est important de souligner l'analogie des travaux présentés dans [BL01] avec la théorie de la détection *a contrario*. Une prévision des performances de l'algorithme, fondée sur l'estimation de la probabilité de fausses détections, y est utilisée pour définir automatiquement les seuils de détection. Lindenbaum *et al.* proposent dans [ELS04] une mesure de qualité pour le groupement de caractéristiques, inspirée de la théorie de l'information. La mesure de qualité d'un groupe est définie à partir d'un test d'hypothèse vis-à-vis d'un modèle de dégradation des données. Cette quantitée correspond alors à « surprise » d'observer une telle configuration sous l'hypothèse que le modèle est vrai.

Notons également qu'une approche de groupement perceptuel a récemment été proposée dans [YS09].

A.2.4 *Un aperçu des applications de la détection* a contrario

La détection *a contrario* a initialement été proposée pour mettre en application certains principes de groupement perceptuel de la Gestalt. Desolneux *et al.* ont ainsi appliqué ce principe pour la détection d'alignements [DMM00], de groupes [DMM03a], et de points de fuite [ADV03]. Ils ont également démontré la faculté de cette approche à combiner plusieurs critères de groupement [DMM03a].

Par la suite, le cadre théorique de la détection *a contrario* s'est révélé être générique et a été utilisé avec succès pour de nombreuses autres applications. Nous avons regroupé en différentes catégories les domaines qui ont fait l'objet d'une mise en application de la théorie de décision *a contrario*.

Analyse des alignements Détection d'alignements [DMM00, DMM03b], détection des points de fuite [ADV03], détection de segments [GJMR08] ;

Détection de points d'intérêt ou de contours Détection de contours et de coins [Cao04], de jonctions en T [Bél06], de lignes de niveau contrastées [DMM03a] et de traits [HGCS08] ;

Analyse des histogrammes Analyse des modes [DMM03a], segmentation [DDLP07b], estimation de la palette de couleurs [DDLP07a] ;

Groupement de caractéristiques Groupement selon plusieurs caractéristiques [DMM03a], groupement de correspondances de formes [CDD+07], groupement de correspondances de points d'intérêt et estimation du mouvement [MS04, NSB07] ;

Définition de seuils de mise en correspondance de formes [MSC+06], de points d'intérêt [CLM+08, RDG09], et d'images [HGS08] ;

Détection de mouvement [DPK05, VCB06] ;

Détection d'anomalies dans des textures [GM09] ;

Imagerie satellitaire Détection de changement [RMLHM], détection de routes agricoles [Gil07], et détection de zones urbaines [Jak06] ;

Construction de carte de disparité sub-pixellique [SAM08].

Annexe B

Une mise en œuvre des descripteurs de type SIFT

Dans cette annexe, nous rappelons le principe de la représentation locale des images par des descripteurs SIFT et nous donnons les détails de notre mise en oeuvre.

B.1 Détection de points d'intérêt

La première étape de la représentation locale des images par des descripteurs SIFT est la détection de points d'intérêt. Des points d'intérêt candidats sont d'abord obtenus à partir d'une représentation en espace-échelle de l'image analysée (§ B.1.1). Un critère de validation est ensuite utilisé pour sélectionner les structures géométriques les plus intéressantes (§ B.1.2).

B.1.1 Critère de sélection en espace-échelle

Soit $I : (x, y) \in \Omega \subset \mathbb{R}^2 \mapsto I(x, y) \in \mathbb{R}$ une image en niveaux de gris dont on souhaite extraire les points d'intérêt. Sa représentation en espace échelle linéaire, notée I_σ, est obtenue par convolution de l'image I avec un noyau gaussien de moyenne nulle et de taille σ :

$$I_\sigma(x, y) := G_\sigma * I(x, y) = \frac{1}{2\pi\sigma^2} \iint_{(u,v)\in\Omega} I(u, v)\, e^{-\frac{(x-u)^2+(y-v)^2}{2\sigma^2}}\, dudv\ .$$

Cette représentation reproduit le phénomène de diffusion de la chaleur.

Pour extraire de cette représentation tridimensionnelle de l'image I des points, de manière robuste et répétable, T. Lindeberg propose dans [Lin94] l'opérateur du *laplacien normalisé*. Les extremums du laplacien correspondant à des transitions dans l'image, son idée est de regarder l'évolution de ces structures en fonction de l'échelle pour en déterminer une taille caractéristique. Pour obtenir une invariance au changement d'échelle (changement de résolution, zoom, *etc.*), Lindeberg montre que la réponse du laplacien doit être normalisée en fonction de l'échelle analysée. Dans [MS01, Low04], l'opérateur est défini en utilisant une normalisation σ^2 :

$$\Delta_\sigma I(x, y) = \sigma^2 \cdot \Delta I_\sigma(x, y) := \left\{ \left(\frac{x^2 + y^2}{\sigma^2} - 2 \right) G_\sigma(x, y) \right\} * I(x, y)\ .$$

Les points d'intérêt candidats sont alors définis comme les extremums locaux de $\Delta_\sigma I(x, y)$ en espace (x, y) et en échelle d'analyse σ [Lin98]. On détecte typiquement des structures de bord, de jonction et de « blob » :

- pour un disque de rayon R, l'extremum est situé au centre du disque pour une échelle $\sigma = R/\sqrt{2}$;
- pour une gaussienne de moyenne μ et d'écart-type s, l'extremum est situé en μ pour une échelle $\sigma = s$;

– pour une bande de largeur L, les extremums en échelle sont situés sur la ligne médiane pour une échelle $\sigma = L/2$. Il n'y a théoriquement pas d'extremum local en espace, mais en présence de bruit, de nombreux extremums sont détectés à cette échelle ;
– pour une jonction, les extremums en espace sont situés sur la bissectrice. Il n'y a théoriquement pas d'extremum au sens strict en échelle, mais nous verrons qu'en pratique on peut détecter de nombreux extremums à différentes échelles.

En pratique, pour des images discrètes, le domaine des échelles est échantillonné suivant une progression géométrique : $\sigma_k = \sigma_0 \cdot r^k$, où k est l'indice de l'échelle, σ_0 l'échelle la plus faible analysée et r le rapport entre deux échelles successives. Nous avons choisi les valeurs suivantes : $\sigma_0 = 0.63$, $r = 2^{1/3}$ et $k \in \{1, \ldots, 13\}$. Un extremum local est recherché en chaque point (x, y, σ) de la représentation en espace échelle, en considérant un certain voisinage. Nous avons choisi la 26-connexité. Cela signifie que les échelles des points détectés sont entre 0.8 et 8.

Un exemple des points ainsi détectés à différentes échelles est présenté en figure B.1. On peut observer que le laplacien normalisé permet de détecter différents types de structures. Cependant, le nombre de points d'intérêt candidats est trop important. En particulier, de nombreux points correspondent à des structures de bord qui, comme le souligne D. Lowe, sont très peu informative pour la reconnaissance d'objets. Pour les éliminer, un critère de sélection présenté au prochain paragraphe est utilisé.

Notons que dans [Low04], deux approximations que nous n'avons pas mise en œuvre sont utilisées pour accélérer cette phase de détection. La première approximation proposée par D. Lowe est d'estimer le laplacien normalisé à partir de la différence de gaussiennes (*Difference of Gaussian*, ou DoG). En effet, la différence de deux images successives en espace-échelle donne :

$$I_{\sigma_{k+1}}(x,y) - I_{\sigma_k}(x,y) = \{G_{\sigma_{k+1}} - G_{\sigma_k}\} * I(x,y) \approx \sigma_k (r-1) \left.\frac{\partial G_\sigma(x,y)}{\partial \sigma}\right|_{\sigma_k} .$$

Or, d'après l'équation de la chaleur $\frac{\partial G_\sigma(x,y)}{\partial \sigma} = \sigma \Delta G_\sigma(x,y)$, ceci permet d'écrire que la différence de deux images successives en espace-échelle est une approximation de l'opérateur du laplacien normalisé (à une constante près) :

$$I_{\sigma_{k+1}}(x,y) - I_{\sigma_k}(x,y) \approx (r-1)\Delta_{\sigma_k} I_{\sigma_k}(x,y) .$$

La seconde approximation utilisée dans [Low04] consiste à construire une pyramide, en sous-échantillonnant d'un facteur 2 l'image analysée dès que l'échelle d'analyse σ_k est une puissance de 2. Le gain en complexité est alors très grand pour les grandes échelles où la convolution par un noyau gaussien est très lent. Pour compenser la perte en précision de l'estimation de la position d'un point, une interpolation est ensuite utilisée.

B.1.2 L'élimination des points de bord

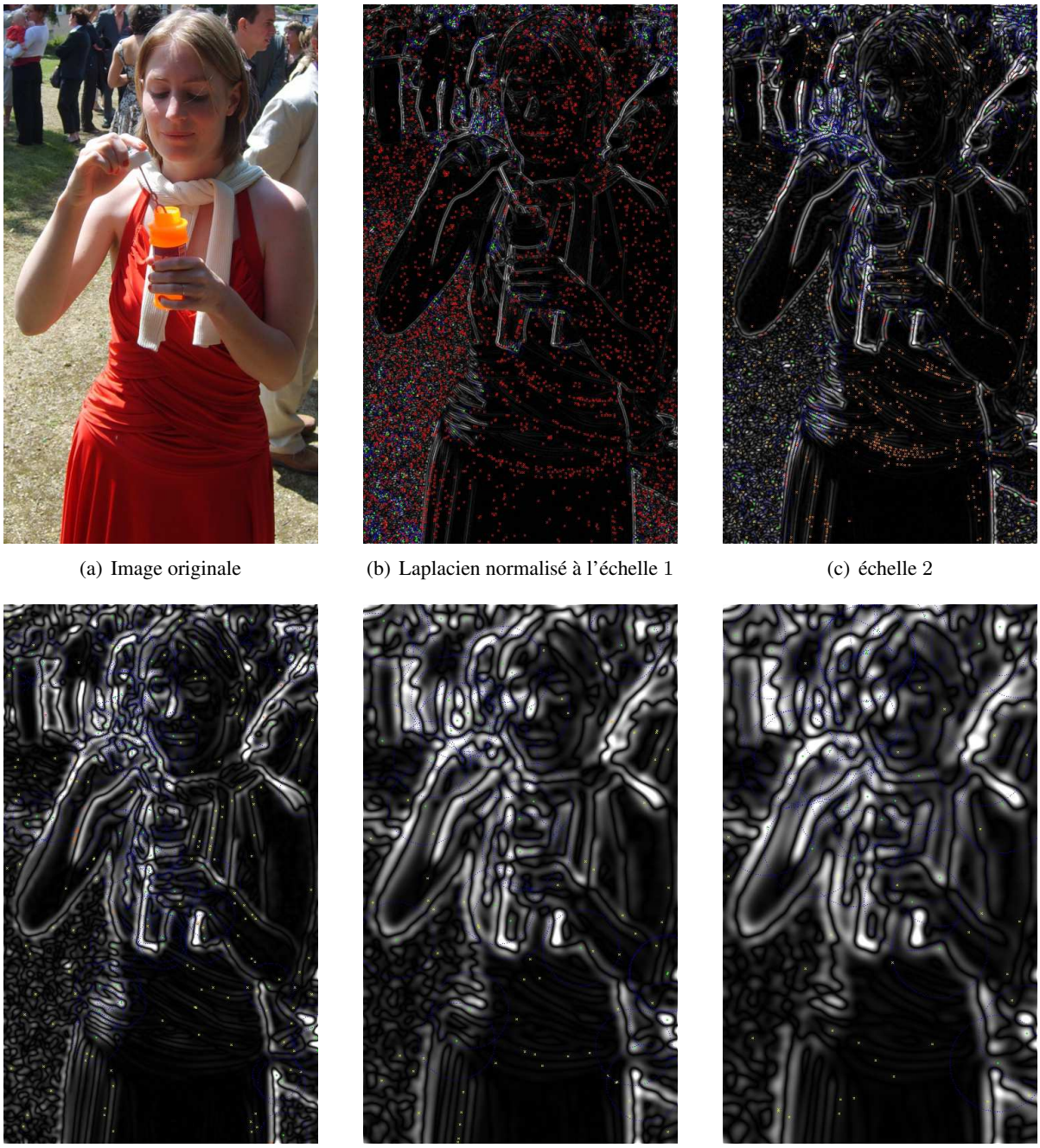
Dans [Low04], un critère de filtrage des points est dérivé de la matrice hessienne. Rappelons tout d'abord que cette matrice des dérivées secondes en un point de l'image $I_\sigma(x,y)$ s'écrit :

$$H_\sigma(x,y) = \begin{bmatrix} \frac{\partial^2 I_\sigma(x,y)}{\partial x^2} & \frac{\partial^2 I_\sigma(x,y)}{\partial y \partial x} \\ \frac{\partial^2 I_\sigma(x,y)}{\partial x \partial y} & \frac{\partial^2 I_\sigma(x,y)}{\partial y^2} \end{bmatrix} .$$

L'étude des valeurs propres de cette matrice permet d'analyser la courbure en un point (x,y) de la surface I_σ. Pour éliminer les points de bord, il faut sélectionner les points ayant deux valeurs propres simultanément élevées. Afin de limiter les calculs, Lowe propose – de manière analogue au détecteur de Harris [HS88] – d'utiliser le test suivant :

$$Det(H_\sigma(x,y)) > t \cdot Tr(H_\sigma(x,y))^2 ,$$

où Det et Tr désignent respectivement le déterminant et la trace de la matrice H_σ, et le paramètre t est un seuil de détection (fixé à 12.1).



(a) Image originale (b) Laplacien normalisé à l'échelle 1 (c) échelle 2

(d) échelle 4 (e) échelle 6.35 (f) échelle 8

FIG. B.1 – ***Illustration du laplacien normalisé en espace-échelle et de la sélection de points saillants.*** *L'image originale est donnée en figure B.1(a). La valeur absolue du laplacien normalisé est représentée en niveaux de gris, pour différentes échelles d'analyse. Les positions des extremums locaux détectés à la fois en espace et en échelle sont indiquées par des croix de couleur. Étant donné le nombre extrême de candidats détectés, un critère de sélection géométrique est nécessaire pour ne sélectionner que les structures les plus intéressantes (représentées par des croix vertes et par un cercle bleu).*

En pratique, cette procédure permet de réduire considérablement le nombre de points reposant sur des bords, sans toutefois les éliminer totalement. Nous avons observé que ces points génèrent de nombreuses fausses correspondances. Pour éviter ce problème, nous avons utilisé le critère de Harris multi-échelle proposé par Mikolajczyk et Schmid dans [MS01].

Le détecteur de coins de Harris [HS88], fondé sur l'approche proposée par Moravec [Mor80], consiste

à étudier les valeurs propres de la matrice $C(x,y)$ définie à partir des dérivées premières de $I(x,y)$:

$$C(x,y) = G_s * \begin{bmatrix} (\frac{\partial I}{\partial x})^2 & \frac{\partial I}{\partial x}\frac{\partial I}{\partial y} \\ \frac{\partial I}{\partial x}\frac{\partial I}{\partial y} & (\frac{\partial I}{\partial y})^2 \end{bmatrix} , \tag{B.1}$$

où G_s est un noyau de convolution gaussien d'écart type s permettant d'obtenir des grandeurs moyennes. Harris et Stephens montrent que l'analyse des valeurs propres de cette matrice répond à la problématique de la détection de structures de forte courbure. Afin de s'épargner les coûts de calcul des valeurs propres, ils proposent le test suivant :

$$\mathcal{C}(x,y) = Det\,(C(x,y)) - k \cdot Tr(C(x,y))^2 > t\ , \tag{B.2}$$

où t et k sont les deux paramètres de détection. La quantité $\mathcal{C}(x,y)$ est une grandeur mesurant la ressemblance à une structure de coin (*cornerness*) qui dépend des valeurs propres de la matrice $C(x,y)$ et du paramètre k. La figure B.2 montre l'allure de cette quantité en fonction des valeurs propres : seules les structures pour lesquelles les deux valeurs propres sont simultanément grandes sont validées.

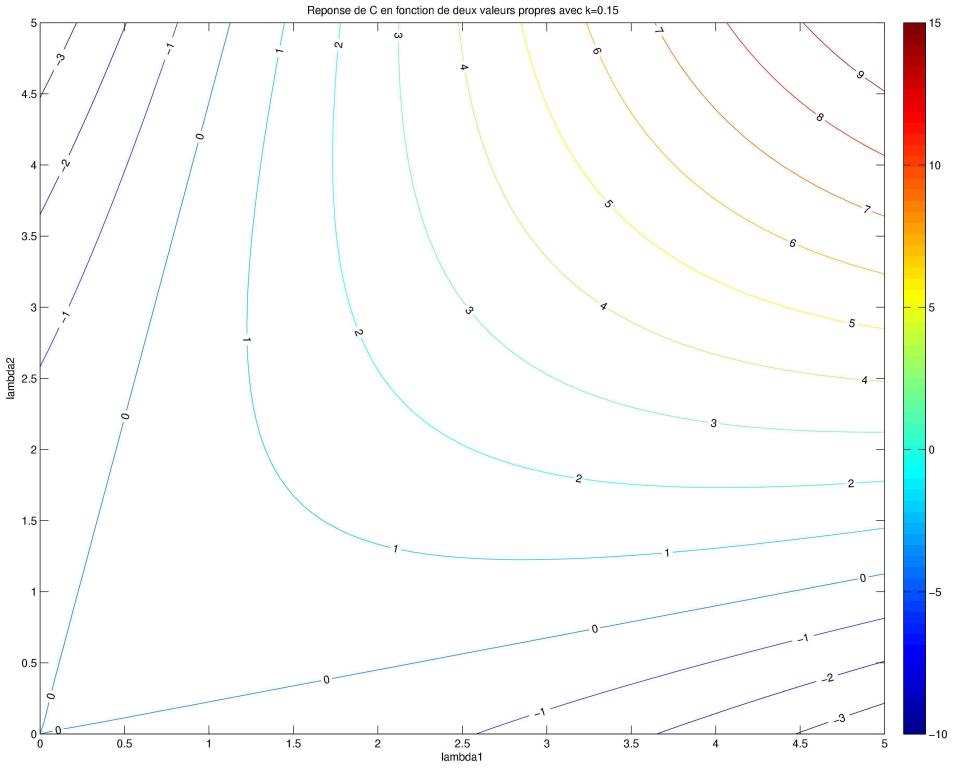


FIG. B.2 – *Réponse de la mesure de ressemblance $\mathcal{C}$ pour $k = 0.15$ en fonction des valeurs propres de la matrice $C_\sigma(x,y)$.*

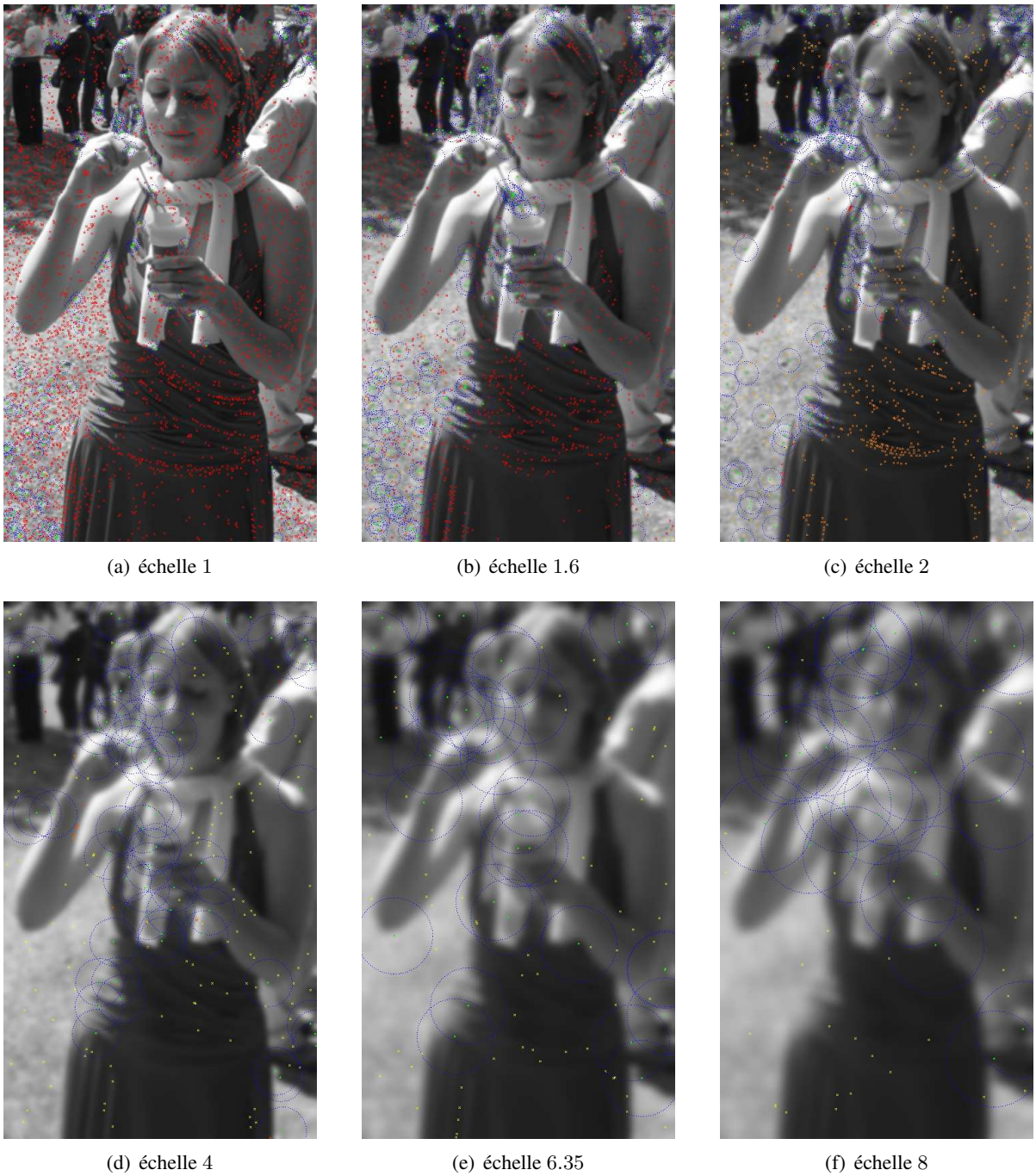
Dans [MS01], les auteurs proposent d'adapter ce critère pour l'analyse en espace-échelle. Pour cela, la matrice $C_\sigma(x,y)$ est définie en prenant en compte l'échelle d'analyse σ :

$$C_\sigma(x,y) = \sigma^2 \cdot G_s * \begin{bmatrix} (\frac{\partial I_\sigma}{\partial x})^2 & \frac{\partial I_\sigma}{\partial x}\frac{\partial I_\sigma}{\partial y} \\ \frac{\partial I_\sigma}{\partial x}\frac{\partial I_\sigma}{\partial y} & (\frac{\partial I_\sigma}{\partial y})^2 \end{bmatrix} . \tag{B.3}$$

Notons une nouvelle fois le paramètre de normalisation σ^2. Le test de ressemblance repose alors sur les trois paramètres de détection t, σ et k :

$$\mathcal{C}_\sigma(x,y) = Det\,(C_\sigma(x,y)) - k \cdot Tr(C_\sigma(x,y))^2 > t\ . \tag{B.4}$$

Nous utilisons cette mesure $\mathcal{C}_\sigma(x, y)$ pour sélectionner les points candidats détectés en tant qu'extremums du laplacien normalisé. Les paramètres utilisés sont : $t = 2000$, $k = 0.04$, et nous avons choisi de fixer le paramètre de moyennage s de manière à ce que $s = \sqrt{2}\sigma$. La figure B.2 montre quels sont les points candidats sélectionnés aux différentes échelles d'analyse de l'image présentée en figure B.1(a).



(a) échelle 1 (b) échelle 1.6 (c) échelle 2

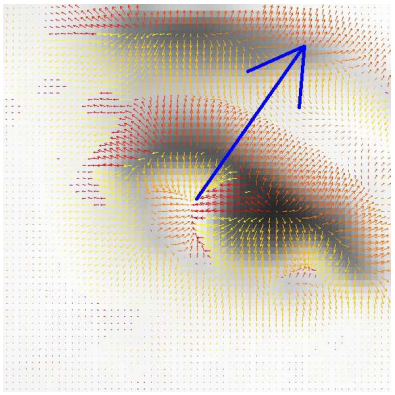
(d) échelle 4 (e) échelle 6.35 (f) échelle 8

FIG. B.3 – *Illustration de l'élimination des points de bord avec le critère de Harris multi-échelle. Les points rejetés sont en rouge, et les points validés sont en vert, entourés par un cercle bleu de rayon proportionnel à l'échelle de détection.*

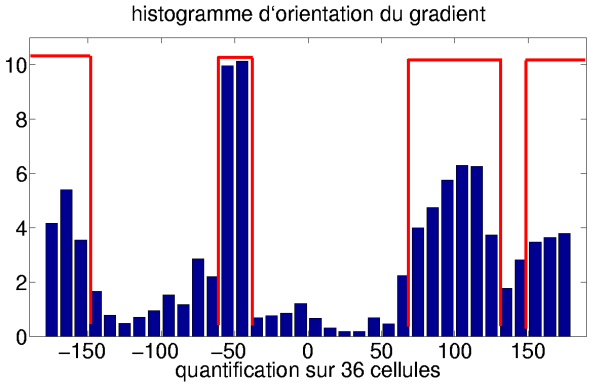
B.1.3 Sélection des orientations principales

À ce stade de la détection de points d'intérêt, des points ont été localisés en espace-échelle. Dans l'approche originale des SIFTs [Low04], une ou plusieurs orientations principales sont ensuite assignées à chaque point pour définir ultérieurement un descripteur invariant à la rotation. D. Lowe propose de construire pour chaque point (x, y, σ) un histogramme d'orientation du gradient (pondéré par la norme) dans un voisinage centré en (x, y) et proportionnel à σ. L'histogramme est quantifié sur 36 bins. Les orientations principales d'un point sont ensuite détectées en recherchant les maxima locaux de cet histogramme. Un seuil de détection est fixé à 80% du maximum de l'histogramme pour limiter le nombre d'orientations.

Pour définir les orientations de manière plus robuste, notamment en ce qui concerne la quantification, nous avons utilisé la méthode de segmentation d'histogramme de [DMM03a]. C'est une approche *a contrario* permettant d'identifier les modes principaux d'un histogramme : une orientation principale peut alors être définie comme le *barycentre circulaire* d'un mode. Une illustration de cette approche est donnée en figure B.4. Jusqu'à deux orientations par point d'intérêt sont ainsi définies, correspondant dans le cas d'un coin aux directions normales à chacun des bords. Une comparaison avec l'approche originale des SIFTs est montrée en figure B.5.

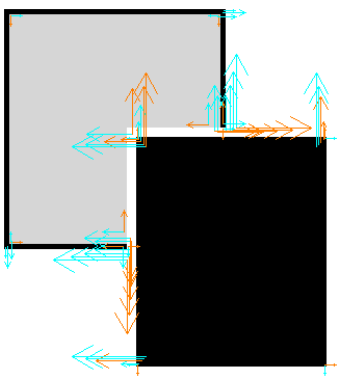


(a) Exemple d'orientation principale détectée

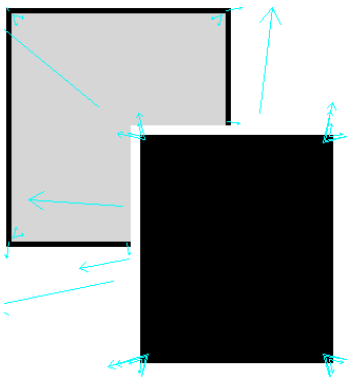


(b) Détection des modes principaux par la méthode *a contrario* de [DMM03a]

FIG. B.4 – *Illustration de l'estimation des orientations principales.*



(a) Orientations principales détectées avec notre approche



(b) Orientations principales détectées avec l'approche originale des SIFTs

FIG. B.5 – *Comparaison de l'estimation des orientations principales.*

B.1.4 Invariance et redondance

Les points d'intérêt détectés en combinant le laplacien normalisé et le critère de Harris (ou « Laplace-Harris ») sont donnés en figure B.6(b). À titre de comparaison, nous présentons le résultat de l'approche originale des SIFTs en figure B.6(c), et du détecteur de Harris en figure B.6(d). Comme on peut le voir sur cet exemple, les points de Laplace-Harris sont moins nombreux sur les bords et dans les zones bruitées très contrastées. Ceci permet une mise en correspondance plus robuste, car l'appariement de tels points d'intérêt est peu informative. Cependant, toutes ces méthodes sont très sensibles au bruit et au changement de contraste. En particulier, on observe qu'aucun point d'intérêt n'est détecté sur la robe dans l'ombre (figure B.6(a)), malgré la présence de nombreuses jonctions. Une perspective intéressante est donnée par les approches de détection *a contrario* qui ont été proposées pour les détections de jonctions en L [Cao04] (figure B.6(e)) ou en T [Bél06] (figure B.6(f)) : des points sont détectés dans les zones peu contrastées mais très structurées, tout en limitant le nombre de fausses détections dans les zones bruitées.

Cette question de l'invariance des détecteurs de points d'intérêt est primordiale car elle conditionne l'ensemble des performances d'un système de reconnaissance d'objets. La figure B.7(a) illustre ce problème dans le cas de la recherche d'un logo dans une publicité. Sur chaque bouteille, le logo apparaît deux fois : en grand sur le bas de la bouteille, et en haut sur le col. En raison de la qualité médiocre de l'image (faible résolution, artefacts de compression, faible contraste sur certains objets), très peu de points d'intérêt sont détectés dans la seconde image. Les mises en correspondance de ces points sont données en figure B.7(b). En conséquence, tous les logos à petite échelle ou peu contrastés ne peuvent pas être détectés (figure B.7(c)), en raison de l'absence ou d'un nombre insuffisant de points d'intérêt.

Nous avons également souligné au chapitre 4 le problème des répétitions de points d'intérêt à différentes échelles, comme on peut le voir sur les images B.6(b) et B.6(c). Ce phénomène se produit notamment pour les structures correspondant à des jonctions de bords (typiquement un coin), qui n'ont pas d'échelle caractéristique (voir la figure B.5). Une solution proposée dans [Sur07] consiste à étudier la trajectoire de ces points en espace-échelle non linéaire.

Notons également l'approche proposée dans [MS02] pour une détection des points d'intérêt robuste au changement affine.

(a) Image avec des variations fortes de contraste

(b) Détecteur de Laplace-Harris

(c) Détecteur de SIFT

(d) Détecteur de Harris

(e) Détecteur de coins *a contrario* (Cao)

(f) Détecteur de jonctions en T *a contrario* (Bélardi)

FIG. B.6 – *Illustration de la non invariance de la détection de points d'intérêt au changement de contraste. Points détectés avec plusieurs types de détecteurs.*

(a) Paires d'images avec le logo recherché et une scène où il apparaît de nombreuses fois

(b) Correspondances détectées avec le critère AC

(c) Résultat de MAC-RANSAC

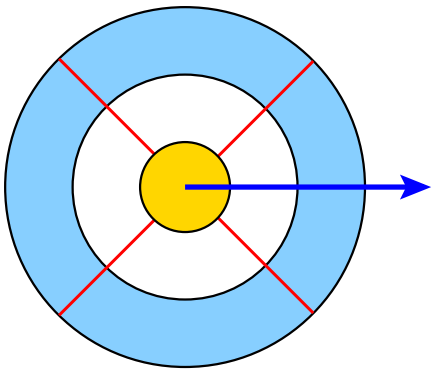
FIG. B.7 – *On recherche un logo dans une photographie où il apparaît de nombreuses fois (figure B.7(a)). MAC-RANSAC ne permet de détecter qu'une faible proportion d'occurrences (figure B.7(c)), en raison du nombre insuffisant de points d'intérêt détectés initialement (figure B.7(b)).*

B.2 Construction des descripteurs locaux

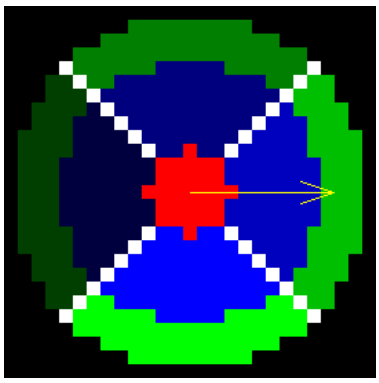
Pour chaque point d'intérêt (x, y, σ) détecté, un descripteur local SIFT est construit en fonction du voisinage centré en (x, y) et du paramètre d'échelle σ.

Taille et forme du descripteur Pour compléter l'invariance à la rotation du descripteur SIFT, le voisinage est défini à partir d'un masque circulaire divisé en plusieurs régions distinctes, de manière analogue au descripteur Shape Context [BMP02]. Ce masque est schématisé en figure B.8(a) avec $M = 9$ régions. L'orientation principale du point d'intérêt permet de définir l'origine angulaire à la fois du masque, mais également de la direction du gradient. Ainsi pour chaque orientation principale du point d'intérêt, un nouveau descripteur SIFT est construit.

La taille du masque est proportionnelle à l'échelle caractéristique σ du point détecté. Le principe du descripteur SIFT est d'être très discriminant, en codant à la fois la structure détectée et son contexte. Le rayon du masque circulaire est de 12σ pixels (figure B.8(b) avec $\sigma = 1$) avec le premier rayon intérieur de 3σ et le second à 8σ.



(a) Masque circulaire découpé en 9 régions.



(b) Masque circulaire de 12 pixels de rayon.

FIG. B.8 – *(Figure de gauche) Masque circulaire découpé en* 9 *régions distinctes, utilisé pour l'extraction des histogrammes d'orientation du gradient. (Figure de droite) Exemple du masque circulaire construit pour un rayon de* 12 *pixels.*

Construction des histogrammes Dans [Low04], un histogramme de l'orientation du gradient est construit pour chaque région du masque utilisé. Pour un point d'intérêt détecté à l'échelle σ, le gradient est directement calculé à partir de l'image I_σ. Nous avons observé que les histogrammes obtenus était souvent plat : en effet, l'échelle à laquelle un point d'intérêt est détecté correspond au « flou » nécessaire pour détecter la structure d'intérêt, ce qui élimine la texture et les détails à plus petite échelle. Or, tout l'intérêt des descripteurs SIFT réside dans le fait de représenter à la fois la géométrie de la structure d'intérêt, mais également de coder les différentes textures du voisinage. Pour conserver cette information, nous utilisons pour la construction des histogrammes une échelle $\sigma' = 0.4 \cdot \sigma$.

Chaque histogramme est ensuite construit en considérant l'orientation du gradient des pixels appartenant à une même région (figure B.9). Chaque vote est pondéré par la norme du gradient pour être plus robuste au bruit. Notons que dans [Low04], les SIFTs sont construits à l'aide d'une grille régulière carrée orientée selon l'orientation principale. En raison des problèmes de quantification générés par une telle grille, D. Lowe utilise une interpolation trilinéaire (en espace et en orientation) pour le vote de chaque bin. Dans notre cas, l'utilisation d'un masque circulaire limite ce problème, car les anneaux du masque sont invariants par rotation du masque. Toutefois, les pixels jouxtant plusieurs régions votent pour toutes ces régions afin de prendre en compte ce problème de quantification.

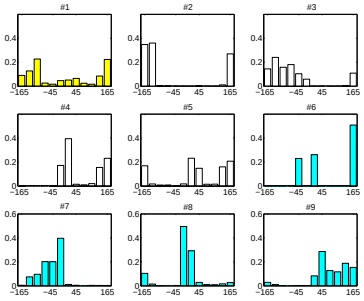


FIG. B.9 – *Exemple de* 9 *histogrammes d'orientation du gradient extraits de chacune des régions correspondantes.*

Finalement, l'ensemble des histogrammes sont normalisés de manière à ce que la norme du descripteur soit égale à 1. La norme utilisée dépend de la mesure de dissimilarité utilisée. Pour de plus amples détails sur la normalisation, voir la section 7.2.

Notons que de très nombreuses variations ont par la suite été proposées à partir de la méthode originale [Low04] de représentation des SIFTs. À titre d'exemple : réduction des dimensions (PCA SIFT [KS04] et MSIFT [MM07]), prise en compte de l'orientation du gradient des images couleurs (CSIFT [AHF06]), utilisation des images intégrales pour accélérer le calcul des descripteurs (SURF [BTG06]), utilisation préliminaire des MSER en tant que détecteur de régions d'intérêt [FL07], prise en compte du contexte des points d'intérêt sur un plus grand voisinage (SIFT with Shape Context [MDS05]), descripteur complètement invariant à la rotation (GLOH [MS05]) ou aux transformations affines (ASIFT [MY09]).

Annexe C

Géométrie de la caméra

Nous nous sommes intéressés, dans la première partie de cette thèse, à la mise en correspondance de deux images, où un objet est photographié depuis deux points de vue différents. Nous allons dans les sections suivantes rappeler les différents modèles géométriques pouvant être utilisés suivant le mouvement relatif de la caméra vis à vis de l'objet entre les deux prises de vue, et suivant la nature de l'objet en lui-même.

C.1 Modèle du sténopé

On rappelle ici le modèle caméra utilisé pour les expériences sur images synthétiques du chapitre 4. Il s'agit du sténopé schématisé en figure C.1, qui modélise la projection d'un point tridimensionnel sur le plan image d'une caméra, dans le cas particulier où son ouverture est réduite à un point (*pine-hole camera*). C'est un modèle simple où n'apparaissent pas les problèmes liés à l'utilisation d'une lentille (distorsion géométrique, chromatique, *etc.*).

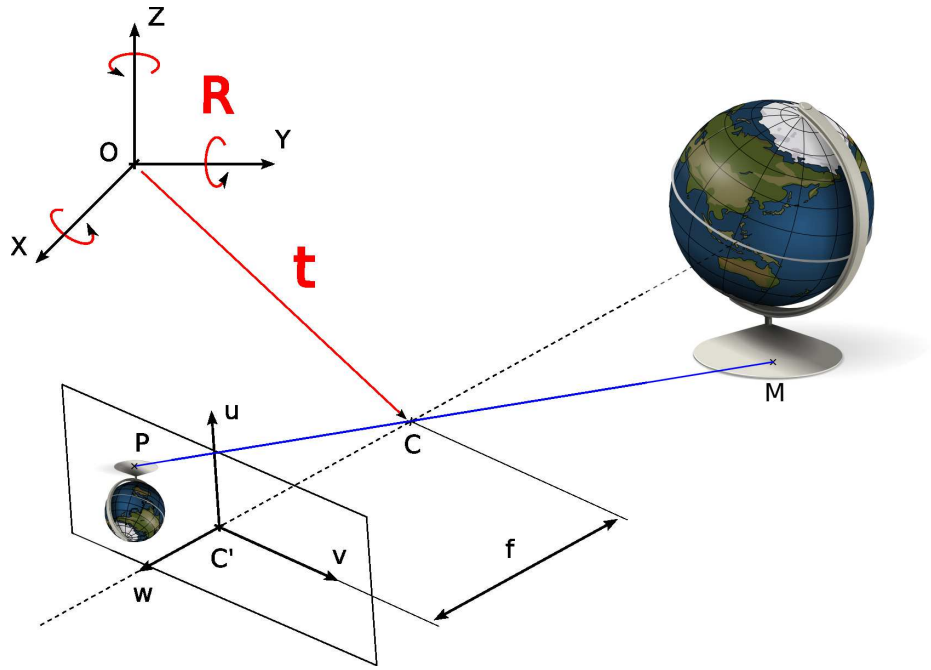


FIG. C.1 – *Modèle du sténopé.*

Soit $\mathcal{R}$ le repère de référence de centre O, et $\mathcal{R}_C : (C, \vec{u}, \vec{v}, \vec{w})$ le repère lié à la caméra. C désigne le centre optique, et l'axe optique (CC') est dirigé selon $\vec{w}$. Le plan focal image $\mathcal{F}$ est le plan $(C', \vec{u}, \vec{v})$. On appelle *distance focale* la distance séparant C de C'.

On désigne par $\mathbf{R}$ la matrice 3×3 exprimant le changement de repère de $\mathcal{R}$ à $\mathcal{R}_C$ lié à la rotation de la caméra, et par $\mathbf{t} = \overrightarrow{CO}$ la position du point O vis-à-vis du centre optique de la caméra, exprimée dans le repère $\mathcal{R}$. La pose de la caméra dépend de 6 degrés de liberté, exprimés par les *paramètres extrinsèques* de la caméra. Soit un point $M : (X, Y, Z)^T$ dont les coordonnées sont exprimées dans le repère $\mathcal{R}_C$. On note $M' : (X', Y', Z')^T$ les coordonnées du point M dans le repère de la caméra $\mathcal{R}_C$, soit :

$$M' = [\mathbf{R} \mid \mathbf{t}] \begin{pmatrix} M \\ 1 \end{pmatrix} .$$

La projection du point M' par le centre C sur le plan $\mathcal{F}$ est le point P ayant pour coordonnées $P : (u, v)$ dans le plan focal de la caméra. En coordonnées homogènes, on a la relation suivante :

$$P = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{f}M' = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} ,$$

où $u = \frac{x}{z}$ et $v = \frac{y}{z}$.

Afin de modéliser les problèmes liés à l'acquisition des images numériques *via* un capteur (transistors disposés sur une grille non régulière, non coïncidence du capteur et du plan focal image, *etc*.), on appelle $P' : (u', v')$ l'image du point M dans la photographie acquise par le capteur. Ses coordonnées homogènes dépendent des *paramètres intrinsèques* de la caméra regroupés (à l'exception de la distance focale f) dans la matrice notée $\mathbf{K}$:

$$P' = \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \mathbf{K}P = \begin{bmatrix} s_x & s_{xy} & o_x \\ 0 & s_y & o_y \\ 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} ,$$

où (o_x, o_y) désigne un décalage de l'origine (*offset*), s_x et s_y désignent des paramètres d'échelles et s_{xy} est un paramètre lié à la non-orthogonalité de la grille. Les coordonnées de P' sont alors définies selon : $u' = \frac{x'}{z'}$ et $v' = \frac{y'}{z'}$. Finalement, la relation entre le point M et le point P' est la suivante :

$$P' = \mathbf{K}\,\mathbf{f}\,[\mathbf{R} \mid \mathbf{t}] \begin{pmatrix} M \\ 1 \end{pmatrix} .$$

Par la suite, nous considérons que la caméra est calibrée, c'est à dire que la distorsion liée au capteur est connue et compensée, de telle sorte que $\mathbf{K}$ est égale à la matrice identité. Seuls le paramètre f et la matrice $[\mathbf{R}' \mid \mathbf{t}']$ sont inconnus.

Nous allons maintenant nous intéresser aux relations de points d'intérêts correspondant à deux prises de vue d'un même objet.

Transformations liées au mouvement de la caméra Supposons que le repère de référence soit lié à l'objet photographié et que la caméra se déplace entre deux prises de vue de l'objet en question. L'image du point M est notée m dans l'image I, et m' dans l'image I', de telle sorte que :

$$\begin{cases} m &= \mathbf{f}\,[\mathbf{R} \mid \mathbf{t}] \begin{pmatrix} M \\ 1 \end{pmatrix} \\ m' &= \mathbf{f}'\,[\mathbf{R}' \mid \mathbf{t}'] \begin{pmatrix} M \\ 1 \end{pmatrix} \end{cases}$$

Le déplacement dans le repère $\mathcal{R}_C$ du centre optique C de la caméra entre les deux vues correspond à $(\mathbf{R}'^{\mathbf{T}}\mathbf{t}' - \mathbf{R}^{\mathbf{T}}\mathbf{t})$, et la rotation 3D de la caméra correspond à la matrice : $\mathbf{R}'\mathbf{R}^{\mathbf{T}}$.

La transformation géométrique entre les points m et m' dépend de la nature de l'objet considéré sur lequel repose le point M et du mouvement de la caméra entre les deux vues.

Lorsque l'objet est plan, il est toujours possible de définir une transformation exacte entre les points conjugués m et m'. Ces transformations, dites *planes*, sont décrites dans la section suivante (§ C.2).

Lorsque l'objet n'est pas plan, le moindre mouvement de la caméra peut se traduire par des effets 3D comme l'auto-occultation : dans ce cas, le point M visible dans la première image peut être caché dans la seconde. On utilise la géométrie *épipolaire* pour décrire la relation entre les points m et m' dans de telles situations, qui est rappelée en section C.3.

C.2 Les transformations planes

Lorsque l'objet est plan, la transformation la plus générale pour le couple de points (m, m') est appelée *homographie*. Elle permet notamment de décrire les effets du changement de perspective entre deux points de vue. Nous verrons ensuite que, pour quelques mouvements particuliers de la caméra, la transformation peut se ramener à une similitude ou une transformation affine.

Homographie Dans le modèle du sténopé, la photographie est obtenue par projection centrale de points de l'espace sur le plan focal image dans la direction d'un point fixe, le centre optique. Si tous les points appartiennent à un même plan, alors les points obtenus dans le plan image conservent leurs alignements. L'homographie (ou transformation projective) désigne la classe de transformations qui conservent les alignements. Les relations entre les coordonnées des points $m : (x, y) \in I$ et $m' : (x', y') \in I'$ sont linéaires en fonction des paramètres de la transformation, mais pour exprimer la relation entre les points m et m' sous forme matricielle, il faut utiliser les notations en coordonnées homogènes :

$$m^{'} = \begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = H \cdot m = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}, \tag{C.1}$$

où les coordonnées du point m' dans l'image I' sont obtenues de la manière suivante :

$$\begin{cases} x' = \dfrac{X'}{Z'} = \dfrac{h_{11}\,x + h_{12}\,y + h_{13}}{h_{31}\,x + h_{32}\,y + h_{33}} \\ \\ y' = \dfrac{Y'}{Z'} = \dfrac{h_{21}\,x + h_{22}\,y + h_{23}}{h_{31}\,x + h_{32}\,y + h_{33}} \end{cases}. \tag{C.2}$$

La matrice H étant définie à un facteur d'échelle près, quatre couples de points sont nécessaires pour définir de manière unique (en dehors de situations dégénérées) les huit paramètres indépendants de l'homographie.

Dans le cas où l'objet n'est pas plan, il existe plusieurs cas de figure où l'homographie modélise la transformation entre un couple de points. En particulier, c'est le cas lorsque la caméra effectue une rotation de telle sorte que son centre optique reste fixe entre les deux prises de vue : il n'y a alors pas de phénomène d'auto-occultation. Cette situation correspond au fait de changer la direction de son regard sans bouger la tête. Ce cas particulier de mouvement de caméra est par exemple utilisé pour la construction de mosaïques d'images (ou panorama).

Dans le cas particulier où l'axe optique de la caméra est **normal** au plan de l'objet pour les deux vues, il n'y a pas d'effet de perspective et l'homographie se réduit alors à une simple similitude.

Similitude Une similitude est une transformation à quatre degrés de liberté, qui résulte de la composition d'une rotation plane de paramètre θ, d'un changement d'échelle s et d'une translation T :

$$m^{'} = \begin{pmatrix} x' \\ y' \end{pmatrix} = s\,R(\theta) \cdot m + T = s \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix}, \tag{C.3}$$

où $R(\theta)$ représente la matrice de rotation dans le plan de l'image I'. Le paramètre s dépend alors du changement de focale (facteur de zoom f'/f) et de la translation selon la direction de l'axe optique $|(\mathbf{t}' - \mathbf{t}.\vec{w})|$ de la caméra entre les deux vues. Les paramètres (t_x, t_y) représentent la composante du mouvement de la caméra parallèle au plan de l'objet. Enfin, le paramètre θ décrit la rotation de la caméra autour de son axe optique.

Dans le cas d'un objet non plan, la similitude permet de modéliser la transformation entre les deux prises de vue lorsque le mouvement de la caméra est limité à une rotation autour de son axe optique et à un changement de focale.

Deux couples de points sont nécessaires pour déterminer les paramètres de la similitude. En coordonnées homogènes, la relation entre le point m et m' peut être mise sous la forme :

$$m^{'} = Hm = \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{bmatrix} sR(\theta) & T \\ (0,0) & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} .$$

Transformation affine Une autre transformation plane est la transformation affine qui possède six degrés de liberté et qui a la propriété de conserver le parallélisme. Elle modélise le cas très particulier où l'on observe un objet plan avec zoom infini ($f, f' \rightarrow \infty$) et où la caméra se trouve infiniment loin de l'objet ($||t||, ||t'|| \rightarrow \infty$). Cette transformation est largement utilisée en vision par ordinateur (reconnaissance et détection d'objets) car elle est souvent une approximation convenable de la transformation entre deux photographies obtenues avec une grande focale (c'est typiquement le cas avec un appareil photographique compact).

La transformation affine est la composition d'une rotation θ selon l'axe optique, de deux facteurs d'échelle s_1 et s_2 selon deux directions orthogonales paramétrées par ϕ et d'une translation T :

$$m^{'} = A \cdot m + T = R(\theta) \cdot R(-\phi) \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \cdot R(\phi) \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} . \tag{C.4}$$

Lorsque $s_1 = s_2$, la transformation affine est réduite à une similitude. Ces paramètres correspondent à un changement d'échelle différent selon deux directions (effet de cisaillement ou *skew*), et permettent de compenser en partie les effets de perspective. Dans [MY09], la transformation affine est écrite sous la forme :

$$A = \lambda R(\phi) \cdot \begin{pmatrix} \tau & 0 \\ 0 & 1 \end{pmatrix} \cdot R(\psi)$$

où le paramètre $\tau = \dfrac{1}{\arccos(\theta)} > 1$ est appelé « tilt ». θ représente l'angle entre la normale au plan de l'objet et l'axe optique.

Il nous faut trois couples de points pour déterminer les paramètres de la transformation affine. En coordonnées homogènes, la relation entre le point m et m' peut être mise sous la forme :

$$m^{'} = Hm = \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{bmatrix} A & T \\ (0,0) & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} .$$

C.3 La géométrie épipolaire

Dans la section précédente, nous avons présenté le cas particulier des objets plans où il était possible de définir une application linéaire bijective (transformation plane) définissant la position d'un point m' dans une image I' à partir d'un point conjugé m dans une image I'. Nous avons vu que dans certains cas particuliers, ces transformations planes peuvent également être utilisées pour décrire la transformation entre deux prises de vue d'un objet 3D.

Dans un cas plus général où le mouvement de la caméra est quelconque et pour des objets non plans, il est impossible de définir une telle application. Néanmoins, Faugeras et Luong ont mis en évidence [FLP01] les contraintes dites « épipolaires » sur les positions relatives d'un couple de points conjugués (m, m'). Ces contraintes sont illustrées en figure C.2. L'image du point M par la caméra de centre optique C est le point $m \in I$. Dans la seconde image I', la position du point m' est définie de manière unique par le point M. En pratique cependant, on ne connaît pas la position tridimensionnelle du point M : on sait seulement qu'il appartient à la droite (Cm). L'image du point m dans l'image I' correspond alors à une droite appelée *ligne épipolaire*. Cette ligne correspond à l'ensemble des positions du point m' lorsque le point M se déplace sur la droite (Cm). Il s'agit de l'intersection du plan focal $\mathcal{F}'$ et du plan (CMC'). Ainsi, quelle que soit la position du point $m \in I$, la ligne épipolaire dans l'image I' passe par un point fixe e', qui est l'image du centre optique de la première caméra C par la seconde. Ce point est appelé épipole.

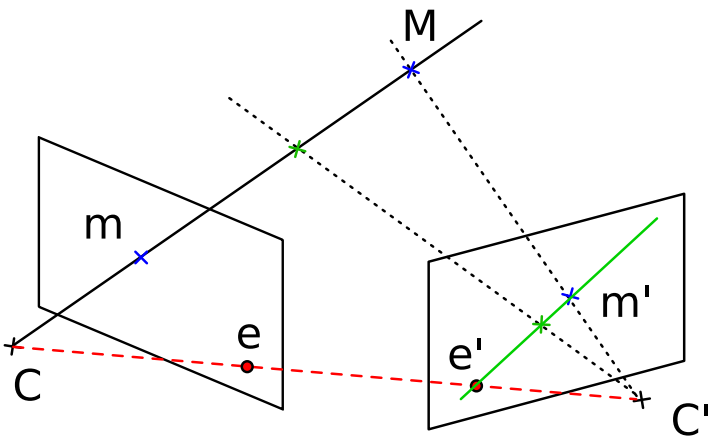


FIG. C.2 – *Illustration des contraintes épipolaires pour un couple de points* (m, m').

La relation entre les points m et m' (en coordonnées homogènes) est donnée par la *matrice fondamentale*

$$m'^T F\, m = 0\,, \tag{C.5}$$

où F est une matrice 3×3 de rang 2. Cette expression traduit la coplanarité des points m, C, m', C', et M. Pour F et m donnés, le point m' appartient à la droite épipolaire décrite en coordonnées homogènes par le vecteur Fm.

Comme dans le cas de l'homographie, où la matrice H est définie à un facteur d'échelle près, la matrice fondamentale F est décrite par huit paramètres indépendants. 8 couples de points sont donc nécessaires étant donné que la contrainte (C.5) est scalaire.

Géométrie épipolaire affine Notons que dans le cas particulier affine (objet à l'infini), la structure de la matrice fondamentale affine – notée F_A – est la suivante :

$$F_A = \begin{bmatrix} 0 & 0 & f_{13} \\ 0 & 0 & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix}\,.$$

Les lignes épipolaires sont, dans ce cas particulier, parallèles entre elles.

C.4 Transformations non rigides

Nous avons jusqu'à présent considéré dans cette annexe les déformations *rigides*. D'autres transformations non rigides peuvent être utilisées pour modéliser les déformations liées à la caméra (système optique), ou encore à l'objet en lui-même. Voici, à titre d'exemple, quelques modèles utilisés en vision par ordinateur :

- Déformations polynomiales (imagerie satellitaire) ;
- Distorsion radiale (déformations en coussinet ou en barillet, liées au système optique de la caméra) ;
- Modèle rigide articulé (recherche de la pose d'une personne).

Publications

Revue

A statistical approach to the matching of local features, J. Rabin, J. Delon et Y. Gousseau, *SIAM Journal on Imaging Science*, 2009.

Conférences

MAC-RANSAC : Multiple A Contrario Consensus for object recognition, J. Rabin, J. Delon, Y. Gousseau et L. Moisan, *3D'PVT*, 2010 (Présentation orale).

MAC-RANSAC : reconnaissance automatique d'objets multiples, J. Rabin, J. Delon, Y. Gousseau et L. Moisan, *RFIA*, 2010 (Présentation orale).

Circular Earth Mover's Distance for the comparison of local features, J. Rabin, J. Delon et Y. Gousseau, *ICPR*, 2008 (Présentation orale).

A contrario matching of SIFT-like descriptors, J. Rabin, J. Delon et Y. Gousseau, *ICPR*, 2008 (Présentation orale).

Mise en correspondance de descripteurs géométriques locaux par méthode a contrario, J. Rabin, J. Delon et Y. Gousseau, *GRETSI*, 2007 (Poster).

Prépublication

Transportation distances on the circle, J. Rabin, J. Delon et Y. Gousseau, Preprint HAL , 2009.

Article en préparation avancée

Non Local Map Regularization for color transfer, J. Rabin, J. Delon et Y. Gousseau.

Bibliographie

[ACB+03] L. Ambrosio, L.A. Caffarelli, Y. Brenier, G. Buttazzo, and C. Villani. *Optimal Transportation and Applications*, volume 1813 of *Lecture Notes in Mathematics*. Springer, Berlin / Heidelberg, mathematics and statistics edition, 2003. 131

[ADV03] A. Almansa, A. Desolneux, and S. Vamech. Vanishing points detection without any a priori information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4) :502–507, april 2003. 239

[AGDL09] Andrew Adams, Natasha Gelfand, Jennifer Dolson, and Marc Levoy. Gaussian KD-trees for fast high-dimensional filtering. *ACM Trans. Graph.*, 28(3) :1–12, 2009. 209

[AGLM93] L. Alvarez, F. Guichard, P.L. Lions, and J.M. Morel. Axioms and fundamental equations of image processing. *Archive for Rational Mechanics and Analysis*, 123(3) :199–257, 1993. 232

[AHF06] A. E. Abdel-Hakim and A. A. Farag. CSIFT : A SIFT Descriptor with Color Invariant Characteristics. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1978–1983, 2006. 251

[AK07] Arash Abadpour and Shohreh Kasaei. An efficient PCA-based color transfer method. *J. Vis. Comun. Image Represent.*, 18(1) :15–34, 2007. 194

[Aka74] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6) :716–723, 1974. 70

[AM97] Luis Alvarez and Freya Morales. Affine morphological multiscale analysis of corners and multiple junctions. *Int. J. Comput. Vision*, 25(2) :95–107, 1997. 10

[AMN+98] Sunil Arya, David M. Mount, Nathan S. Netanyahu, Ruth Silverman, and Angela Y. Wu. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM*, 45(6) :891–923, 1998. 17

[BA96] Michael J. Black and P. Anandan. The robust estimation of multiple motions : parametric and piecewise-smooth flow fields. *Comput. Vis. Image Underst.*, 63(1) :75–104, 1996. 68

[BA04] Burnham and Anderson. Multimodel Inference : Understanding AIC and BIC in Model Selection. *Sociological Methods Research*, 33 :261–304, 2004. 70

[Bal81] D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 13(2) :111–122, 1981. 57

[Bar07] Adrien Bartoli. A random sampling strategy for piecewise planar scene segmentation. *Comput. Vis. Image Underst.*, 105(1) :42–59, 2007. 7, 64, 93

[Bau00] A. Baumberg. Reliable feature matching across widely separated views. In *Proc. CVPR*, 2000. 10, 11, 13, 14

[BC08] Thomas Brox and Daniel Cremers. Iterated nonlocal means for texture restoration. *Scale Space and Variational Methods in Computer Vision*, pages 13–24, 2008. 201, 203

[BCM05] A. Buades, B. Coll, and J. M. Morel. A review of image denoising algorithms, with a new one. *Multiscale Modeling & Simulation*, 4(2) :490+, 2005. 187, 200, 201, 208

[BCM08] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Nonlocal image and movie denoising. *Int. J. Comput. Vision*, 76(2) :123–139, 2008. 201

[BCMS07] A. Buades, B. Coll, J.M Morel, and C. Sbert. Non local demosaicing. *Preprint CMLA 2007-15, disponible à : http://dmi.uib.es/~abuades/publicacions/CMLA2007-15.pdf*, 2007. 201

[BHS08] Philipp Jenke Benjamin Huhle, Timo Schairer and Wolfgang Straßer. Robust non-local denoising of colored depth data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Workshop on Time of Flight Camera based Computer Vision (TOF-CV)*, 2008. 201

[BKB07] Jerome Boulanger, Charles Kervrann, and Patrick Bouthemy. Space-time adaptation for patch-based image sequence restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6) :1096–1102, 2007. 201

[BL01] Alexander Berengolts and Michael Lindenbaum. On the performance of connected components grouping. *Int. J. Comput. Vision*, 41(3) :195–216, 2001. 239

[BL07] Matthew Brown and David G. Lowe. Automatic panoramic image stitching using invariant features. *Int. J. Comput. Vision*, 74(1) :59–73, 2007. 7, 14

[BMP02] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4) :509–522, 2002. 13, 20, 26, 134, 172, 173, 250

[Bon36] C. Bonferroni. Teoria statistica delle classi et calcolo delle probabilita. *Pubblicazioni del Instituto Superiore de Scienze Economiche e Commerciali di Firenze*, 8 :3–62, 1936. 238

[BP09] Aurélie Bugeau and Patrick Pérez. Detection and segmentation of moving objects in complex scenes. *Comput. Vis. Image Underst.*, 113(4) :459–476, 2009. 68

[BPD06] Soonmin Bae, Sylvain Paris, and Frédo Durand. Two-scale tone management for photographic look. *ACM Trans. Graph.*, 25(3) :637–645, 2006. 194, 205, 228

[Bro83] C.M. Brown. Inherent bias and noise in the Hough transform. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 5(5) :493–505, September 1983. 57

[BS98] Kishore Bubna and Charles V. Stewart. Model selection and surface merging in reconstruction algorithms. In *ICCV '98 : Proceedings of the Sixth International Conference on Computer Vision*, page 895, Washington, DC, USA, 1998. IEEE Computer Society. 71

[BSL08] Nicolas Bonnier, Francis Schmitt, and Christophe Leynadier. Improvements in spatial and color adaptive gamut mapping algorithms. In *Proc. of the 4th European Conference on Colour in Graphics, Imaging, and Vision*, 2008. 192, 205, 228

[BSW05] M. Brown, R. Szeliski, and S. Windner. Multi-image matching using multi-scale oriented patches. In *Proc. CVPR*, pages 510–517, 2005. 15

[BTG06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF : Speeded up robust features. In *ECCV*, pages 404–417, 2006. 251

[BZM08] Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4) :712–727, 2008. 8

[Bél06] Stéphane Bélardi. Détection de jonctions en T dans les images. 2006. 9, 239, 247

[Cao04] Frédéric Cao. Application of the gestalt principles to the detection of good continuations and corners in image level lines. *Comput. Vis. Sci.*, 7(1) :3–13, 2004. 9, 232, 239, 247

[CCM99] Vicent Caselles, Bartomeu Coll, and Jean-Michel Morel. Topographic maps and local contrast changes in natural images. *Int. J. Comput. Vision*, 33(1) :5–27, 1999. 188

[CDD+07] Frédéric Cao, Julie Delon, Agnès Desolneux, Pablo Musé, and Frédéric Sur. A unified framework for detecting groups and application to shape recognition. *J. of Mathematical Imaging and Vision*, 27(2), 2007. 58, 90, 93, 239

[Che95] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8) :790–799, 1995. 58

[CHS08] Guillaume Charpiat, Matthias Hofmann, and Bernhard Scholkopf. Automatic image colorization via multimodal predictions. In *10th European Conference on Computer Vision, ECCV 2008*, pages 126–139. Springer, 10 2008. Marseille. 193

[Chu05] O. Chum. *Two-View Geometry Estimation by Random Sample and Consensus*. PhD thesis, 2005. 116

[CJ07a] Gustavo Carneiro and Allan D. Jepson. Flexible spatial configuration of local image features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12) :2089–2104, 2007. 16

[CJ07b] C.D. Castillo and D.W. Jacobs. Using stereo matching for 2-D face recognition across pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007. 16

[CK08] A. Chariot and R. Keriven. GPU-boosted online image matching. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4, 2008. 17

[CL09] Chia-Ming Cheng and Shang-Hong Lai. A consensus sampling technique for fast and robust model fitting. *Pattern Recogn.*, 42(7) :1318–1329, 2009. 59, 61

[CLM+08] F. Cao, J.L. Lisani, J.-M. Morel, P. Musé, and F. Sur. *A theory of shape identification*, volume 1948 of *Lecture Notes in Mathematics*. Springer, 2008. 20, 21, 239

[CLMS99] V. Caselles, J.L. Lisani, J.M. Morel, and G. Sapiro. Shape preserving local histogram modification. *IEEE Transactions on Image Processing*, 8(2) :220–230, February 1999. 188, 189

[CM05] O. Chum and J. Matas. Matching with PROSAC - progressive sample consensus. In *Proc. CVPR*, pages 220–226, 2005. 13, 15, 60

[CM08] O. Chum and J. Matas. Optimal randomized RANSAC. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8) :1472–1482, August 2008. 59

[CPS+07] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall : Automatic query expansion with a generative feature model for object retrieval. In *Proceedings of the 11th International Conference on Computer Vision, Rio de Janeiro, Brazil*, 2007. 7, 17, 18

[CPT04] A. Criminisi, P. Pérez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *Image Processing, IEEE Transactions on*, 13(9) :1200–1212, 2004. 193

[CS02] Sung-Hyuk Cha and Sargur N. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6) :1355–1370, June 2002. 134, 148

[DCC+08] J. Darbon, A. Cunha, T.F. Chan, S. Osher, and G.J. Jensen. Fast nonlocal filtering applied to electron cryomicroscopy. In *In the proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI'08)*, 2008. 209

[DD02] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *SIGGRAPH '02 : Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 257–266, New York, NY, USA, 2002. ACM. 192, 200, 205, 206, 207, 209, 228

[DD09] J. Delon and A. Desolneux. Flicker stabilization in image sequences. *Preprint HAL, disponible à :* `http://hal.archives-ouvertes.fr/docs/00/40/77/96/PDF/flicker_hal_juillet2009.pdf`, 2009. 189

[DDLP07a] Julie Delon, Agnès Desolneux, Jose Luis Lisani, and Ana Belen Petro. Automatic color palette. *Inverse Problems and Imaging*, 1(2) :265–287, 2007. 239

[DDLP07b] Julie Delon, Agnès Desolneux, Jose Luis Lisani, and Ana Belen Petro. A nonparametric approach for histogram segmentation. *IEEE Transactions on Image Processing*, 16(1) :253–261, 2007. 239

[Del04] J. Delon. Midway image equalization. *JMIV*, 21(2) :119–134, September 2004. 2, 135, 189

[Del06] J. Delon. Movie and video scale-time equalization application to flicker reduction. *Image Processing*, 15(1) :241–248, January 2006. 189

[DIIM04] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *SCG '04 : Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262, New York, NY, USA, 2004. ACM. 17

[DMA07] Francois Destrempes, Max Mignotte, and Jean-Francois Angers. Localization of shapes using statistical models and stochastic optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9) :1603–1615, 2007. 15

[DMM00] Agnès Desolneux, Lionel Moisan, and Jean-Michel Morel. Meaningful alignments. *Int. J. Comput. Vision*, 40(1) :7–23, 2000. 20, 237, 239

[DMM03a] A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4) :508–513, 2003. 235, 239, 246

[DMM03b] A. Desolneux, L. Moisan, and J.-M. Morel. Maximal meaningful events and applications to image analysis. *Annals of Statistics*, 31(6) :1822–1851, 2003. 78, 239

[DMM08] A. Desolneux, L. Moisan, and J.-M. Morel. *From Gestalt Theory to Image Analysis : A Probabilistic Approach*. Springer Verlag, 2008. 2, 19, 20, 235, 236

[DPK05] F. Dibos, S. Pelletier, and G. Koepfler. Real-time segmentation of moving objects in a video sequence by a contrario detection. *IEEE International Conference on Image Processing*, 2005. 239

[DT09] Ayelet Dominitz and Allen Tannenbaum. Texture mapping via optimal mass transport. *IEEE Transactions on Visualization and Computer Graphics*, 99(2), 2009. 135

[Dvi02] Guy Dvir. Context-based image modelling. In *Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 4*, page 40162. IEEE Computer Society, 2002. 147

[DZLF94] R. Deriche, Z. Zhang, Q.-T. Luong, and O. Faugeras. Robust recovery of the epipolar geometry for an uncalibrated stereo rig. In *ECCV '94 : Proceedings of the third European conference on Computer vision (vol. 1)*, pages 567–576, Secaucus, NJ, USA, 1994. Springer-Verlag New York, Inc. 13, 14

[DZM+07] Hongli Deng, Wei Zhang, Eric Mortensen, Thomas Dietterich, and Linda Shapiro. Principal curvature-based region detector for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 30

[ED04] Elmar Eisemann and Frédo Durand. Flash photography enhancement via intrinsic relighting. *ACM Trans. Graph.*, 23(3) :673–678, 2004. 193, 205

[EF01] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. *Proceedings of SIGGRAPH 2001*, pages 341–346, August 2001. 192, 228

[EGH90] W. Eric, L. Grimson, and Daniel P. Huttenlocher. On the Sensitivity of the Hough Transform for Object Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12, 1990. 57, 58

[ELS04] E.A. Engbers, M. Lindenbaum, and A.W.M. Smeulders. An information-based measure for grouping quality. In *ECCV*, pages Vol III : 392–404, 2004. 239

[FA91] William T. Freeman and Edward H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13 :891–906, 1991. 173

[FB81] Martin A. Fischler and Robert C. Bolles. Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6) :381–395, 1981. 2, 11, 56, 58, 235

[FFJS08] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Groups of adjacent contour segments for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(1) :36–51, 2008. 10

[FH75] K. Fukunaga and L.D. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IT*, 21(1) :32–40, January 1975. 58

[FL07] P.E. Forssén and D.G. Lowe. Shape descriptors for maximally stable extremal regions. In *IEEE ICCV*, 2007. 10, 26, 172, 181, 251

[FLP01] Olivier Faugeras, Quang-Tuan Luong, and T. Papadopoulou. *The Geometry of Multiple Images : The Laws That Govern The Formation of Images of A Scene and Some of Their Applications*. MIT Press, Cambridge, MA, USA, 2001. 257

[FTG06] Vittorio Ferrari, Tinne Tuytelaars, and Luc Gool. Simultaneous object recognition and segmentation from single or multiple model views. *Int. J. Comput. Vision*, 67(2) :159–188, 2006. 13, 14, 16

[GBH08] Niloofar Gheissari and Alireza Bab-Hadiashar. A comparative study of model selection criteria for computer vision applications. *Image Vision Comput.*, 26(12) :1636–1649, 2008. 71

[GD04] K. Grauman and T.J. Darrell. Fast Contour Matching Using Approximate Earth Mover's Distance. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages I : 220–227, 2004. 134

[GD05] Kristen Grauman and Trevor Darrell. Efficient image matching with distributions of local invariant features. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–634, Washington, DC, USA, 2005. IEEE Computer Society. 134

[GDR00] H. Greenspan, G. Dvir, and Y. Rubner. Region correspondence for image matching via emd flow. In *CBAIVL '00 : Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'00)*, page 27, Washington, DC, USA, 2000. IEEE Computer Society. 147

[GH03] Gary R. Greenfield and Donald H. House. Image recoloring induced by palette color associations. *Journal of WSCG*, 11 :189–196, 2003. 189

[Gil07] Jérôme Gilles. *La recherche des alignements dans les images digitales et ses applications à l'imagerie satellitaire*. PhD thesis, ENS de Cachan, 2007. 239

[GJMR08] Raphael Grompone, Jeremie Jakubowicz, Jean-Michel Morel, and Gregory Randall. LSD : a fast line segment detector with a false detection control. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2008. 239

[GL97] Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18(4) :1035–1064, 1997. 55

[GM09] Bénédicte Grosjean and Lionel Moisan. A-contrario detectability of spots in textured backgrounds. *J. Math. Imaging Vis.*, 33(3) :313–337, 2009. 239

[Har97] Richard I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6) :580–593, 1997. 84

[HB94] Y. Hecker and R. Bolle. On geometric hashing and the generalized Hough transform. volume 24, pages 1328–1338, 1994. 57

[HGCS08] T. Hurtut, Y. Gousseau, F. Cheriet, and F. Schmitt. Pictorial analysis of line-drawings. In *Computational Aesthetics in Graphics (CAe'08), Eurographics, Lisbon*, Jun. 2008. 9, 239

[HGS08] T. Hurtut, Y. Gousseau, and F. Schmitt. Adaptive image retrieval based on the spatial organization of colors. *Computer Vision and Image Understanding, in Press*, 2008. 133, 147, 148, 152, 169, 239

[HJO+01] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. In *SIGGRAPH '01 : Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 327–340, New York, NY, USA, 2001. ACM. 192, 228

[Hou59] P.V. Hough. Machine analysis of bubble chamber pictures. In *International Conference on High Energy Accelerators and Instrumentation*, pages 554–556, 1959. 11, 56, 57

[HRRS05] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics : The Approach Based on Influence Functions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, New York, revised edition, April 1986, 2005. 71

[HS88] Chris Harris and Mike Stephens. A combined corner and edge detector. *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988. 9, 235, 242, 243

[HTF03] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, corrected edition, July 2003. 70

[HU90] D.P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *IJCV*, 5(2) :195–212, November 1990. 8

[Hub81] Peter. J Huber. *Robust Statistics*. Wiley, 1981. 56

[HZ04] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision – 2nd Edition*. Cambridge University Press, 2004. 54, 61, 63

[IK88] J. Illingworth and J. Kittler. A Survey of the Hough Transform. *Computer Vision, Graphics and Image Processing*, 44 :87–116, 1988. 57

[IT03] P. Indyk and N. Thaper. Fast image retrieval via embeddings. In *3rd International Workshop on Statistical and Computational Theories of Vision*, Nice, France, 2003. 134

[Jak06] Jérémie Jakubowicz. *Décomposition et détection de structures géométriques en imagerie*. PhD thesis, ENS de Cachan, 2006. 239

[JD01] Frédéric Jurie and M. Dhome. Real time template matching. In *International Conference on Computer Vision*, pages 544–549, Vancouver, Canada, July 2001. 8

[JT08] Jiaya Jia and Chi-Keung Tang. Image stitching using structure deformation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4) :617–631, 2008. 13, 14

[Kai98] Thomas Kaijser. Computing the Kantorovich distance for images. *J. Math. Imaging Vis.*, 9(2) :173–191, 1998. 134

[Kan42] L. Kantorovich. On the transfer of masses (en russe). *Translated in Management Science, Vol. 5, pp. 1–4, 1959*, 37(2) :227–229, 1942. 131

[Kan98] K. Kanatani. Geometric information criterion for model selection. *IJCV*, 26(3) :171–189, March 1998. 71

[Kan04] K. Kanatani. Uncertainty modeling and model selection for geometric inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(10) :1307–1319, October 2004. 71

[KB08] Ch. Kervrann and J. Boulanger. Local adaptivity to variable smoothness for exemplar-based image denoising and representation. *International Journal of Computer Vision*, 79(1) :45–69, August 2008. 201, 203, 209, 228

[Kot05] H. Kotera. A scene-referred color transfer for pleasant imaging on display. In *ICIP*, pages II : 5–8, 2005. 189

[KP06] A. Kushal and J. Ponce. Modeling 3D objects from stereo view and recognizing them in photographs. In *Proc. ECCV*, 2006. 13, 16

[KS04] Yan Ke and R. Sukthankar. PCA-SIFT : a more distinctive representation for local image descriptors. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).*, volume 2, pages II–506–II–513 Vol.2, 2004. 173, 251

[LB01] E. Levina and P. Bickel. The Earth Mover's distance is the Mallows distance : some insights from statistics. In *Proceedings of the Eighth IEEE International Conference onComputer Vision (ICCV))*, volume 2, pages 251–256 vol.2, 2001. 132, 134

[LCL04] Qin Lv, Moses Charikar, and Kai Li. Image similarity search with compact data structures. In *CIKM '04 : Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 208–217, New York, NY, USA, 2004. ACM. 134, 147, 169

[Lin94] T. Lindeberg. *Scale-Space Theory in Computer Vision*. Norwell, MA, USA. Kluwer Academic Publishers, 1994. 241

[Lin97] M. Lindenbaum. An integrated model for evaluating the amount of data required for reliable recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(11) :1251–1264, 1997. 91

[Lin98] T. Lindeberg. Feature Detection with Automatic Scale Selection. *International Journal of Computer Vision*, 30(2) :79–116, 1998. 9, 241

[LO06] Haibin Ling and Kazunori Okada. Diffusion distance for histogram comparison. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 246–253, Washington, DC, USA, 2006. IEEE Computer Society. 174

[LO07] Haibin Ling and Kazunori Okada. An efficient Earth Mover's distance algorithm for robust histogram comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5) :840–853, may 2007. 133, 134, 173, 174, 179, 183

[Low] David G. Lowe. Code exécutable SIFT : `http://www.cs.ubc.ca/~lowe/keypoints/`. 174

[Low85] David G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, USA, 1985. 236

[Low99] David G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99 : Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, Washington, DC, USA, 1999. IEEE Computer Society. 10

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2) :91–110, 2004. 1, 8, 9, 13, 14, 15, 17, 19, 20, 26, 31, 33, 55, 57, 58, 134, 148, 172, 173, 175, 180, 181, 241, 242, 246, 250, 251

[LPK07] P. Labatut, J.-P. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct 2007. 7

[LQ00] Maxime Lhuillier and Long Quan. Robust dense matching using local and global geometric constraints. *Pattern Recognition, International Conference on*, 1 :1968, 2000. 16

[LW88] Y. Lamdan and H. J. Wolfson. Geometric hashing : A general and efficient model-based recognition scheme. In *Computer Vision., Second International Conference on*, pages 238–249, 1988. 8, 57

[LZLM05] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. Region-based image retrieval with high-level semantic color names. In *MMM '05 : Proceedings of the 11th International Multimedia Modelling Conference*, pages 180–187, Washington, DC, USA, 2005. IEEE Computer Society. 147

[Maî86] H. Maître. Contribution to the prediction of performances of the Hough transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(5) :669–674, September 1986. 57

[MCUP02] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002. 9, 10, 13, 82

[MDS05] E. N. Mortensen, H. Deng, and L. Shapiro. A SIFT Descriptor with Global Context. In *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 184–190, 2005. 251

[MG98] Pascal Monasse and Frédéric Guichard. Fast computation of a contrast-invariant image representation. *IEEE Trans. on Image Proc*, 9 :860–872, 1998. 10

[ML09] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP'09 : Proceedings of the International Conference on Computer Vision Theory and Applications*, 2009. 17, 18

[MM07] Krystian Mikolajczyk and Jiri Matas. Improving SIFT for fast tree matching by optimal linear projection. page 8, Los Alamitos, CA, USA, 2007. IEEE Computer Society. 17, 251

[Mon81] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. Histoire de l'Académie Royale des Sciences, 1781. 2, 131

[Mon00] Pascal Monasse. *Représentation morphologique d'images numériques et application au recalage*. PhD thesis, CMLA, ENS de Cahan, 2000. 90

[Mor80] Hans Moravec. Obstacle avoidance and navigation in the real world by a seeing robot rover. In *tech. report CMU-RI-TR-80-03, Robotics Institute, Carnegie Mellon University & doctoral dissertation, Stanford University*, number CMU-RI-TR-80-03. September 1980. 243

[MP05] Pierre Moreels and Pietro Perona. Evaluation of features detectors and descriptors based on 3D objects. In *Int. Conf. Comput. Vision (ICCV)*, volume 73, Hingham, MA, USA, 2005. Kluwer Academic Publishers. 173

[MP07] Pierre Moreels and Pietro Perona. Evaluation of features detectors and descriptors based on 3D objects. *Int. J. Comput. Vision*, 73(3) :263–284, 2007. 11, 15, 31, 32, 33, 36, 38

[MS96] James Miller and Charles V. Stewart. Muse : Robust surface fitting using unbiased scale estimates. In *Proceedings of the 1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 300–306, 1996. 90

[MS01] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. *International Conference on Computer Vision*, pages 525–531, 2001. 9, 241, 243, 244

[MS02] Krystian Mikolajczyk and Cordelia Schmid. An affine invariant interest point detector. In *Proc. European Conf. Computer Vision*, pages 128–142. Springer Verlag, 2002. 9, 82, 247

[MS03] J. Morovic and P.L. Sun. Accurate 3D image colour histogram transformation. *PRL*, 24(11) :1725–1735, July 2003. 191

[MS04] L. Moisan and B. Stival. A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *International Journal of Computer Vision*, 57(3) :201–218, 2004. 2, 16, 59, 60, 63, 72, 73, 74, 76, 78, 80, 86, 231, 239

[MS05] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10) :1615–1630, 2005. 10, 15, 31, 33, 34, 173, 178, 251

[MSC+06] Pablo Musé, Frédéric Sur, Frédéric Cao, Yann Gousseau, and Jean-Michel Morel. An a contrario decision method for shape element recognition. *Int. J. Comput. Vision*, 69(3) :295–315, 2006. 9, 10, 20, 177, 232, 239

[MSM03] P. Musé, F. Sur, and J.-M. Morel. Sur les seuils de reconnaissance des formes. *Traitement du Signal*, 20(3) :279–294, 2003. 79

[MTS+05] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2) :43–72, 2005. 9

[MY08] J.M. Morel and G. Yu. On the consistency of the SIFT Method. Technical report, CMLA, 2008. 10

[MY09] J.M. Morel and G. Yu. ASIFT : A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences*, 2(2) :438–469, 2009. 10, 33, 178, 184, 232, 251, 256

[NBE+93] Carlton W. Niblack, Ron Barber, Will Equitz, Myron D. Flickner, Eduardo H. Glasman, Dragutin Petkovic, Peter Yanker, Christos Faloutsos, and Gabriel Taubin. Qbic project : querying images by content, using color, texture, and shape. volume 1908, pages 173–187. SPIE, 1993. 173

[Nis05] David Nistér. Preemptive RANSAC for live structure and motion estimation. *Mach. Vision Appl.*, 16(5) :321–329, 2005. 60

[NN05] Laszlo Neumann and Attila Neumann. Color style transfer techniques using hue, lightness and saturation histogram matching. In *Computational Aesthetics in Graphics, Visualization and Imaging 2005*, pages 111–122, May 2005. 191, 192, 194

[NS06] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168, June 2006. 164, 166, 167

[NSB07] N. Noury, F. Sur, and M.-O. Berger. Fundamental matrix estimation without prior match. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 513–516, San Antonio (Texas, USA), September 2007. 60, 63, 239

[PAA+87] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart Ter Haar Romeny, and John B. Zimmerman. Adaptive histogram equalization and its variations. *Comput. Vision Graph. Image Process.*, 39(3) :355–368, 1987. 188

[PD09] Sylvain Paris and Frédo Durand. A fast approximation of the bilateral filter using a signal processing approach. *Int. J. Comput. Vision*, 81(1) :24–52, 2009. 209

[Pey09] Gabriel Peyré. Texture synthesis with grouplets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2009. 193

[PGB03] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Trans. Graph.*, 22(3) :313–318, 2003. 192

[PGB04] P. Pérez, M. Gangnet, and A. Blake. Patchworks : Example-based region tiling for image editing. Technical report, Microsoft Research, 2004. 193

[Pie05] Matti Pietikäinen. Image analysis with local binary patterns. pages 115–118. 2005. 10

[PK06] F. Pitié and A. Kokaram. The Linear Monge-Kantorovitch Colour Mapping for Example-Based Colour Transfer. In *Proceedings of 3rd European Conference on Visual Media Production (CVMP'06)*, London, November 2006. 194

[PKD07] F. Pitié, A. Kokaram, and R. Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, February 2007. 2, 134, 192, 193, 194, 196, 197, 198, 210, 216, 218, 219, 225

[PLRS04] Jean Ponce, Svetlana Lazebnik, Fredrick Rothganger, and Cordelia Schmid. Toward true 3D object recognition. In *Reconnaisance de Formes et Intelligence Artificelle*, 2004. 105, 108, 123

[PNS03] E. Pichon, M. Niethammer, and G. Sapiro. Color histogram equalization through mesh deformation. In *Proceedings of the International Conference on Image Processing*, pages 117–120, 2003. 191

[Pol] M. Pollefeys. Séquence vidéo du chateau de leuven, disponible à : `http://www.cs.unc.edu/~marc/data/castlejpg.zip`. 103, 106

[PSA+04] Georg Petschnigg, Richard Szeliski, Maneesh Agrawala, Michael Cohen, Hugues Hoppe, and Kentaro Toyama. Digital photography with flash and no-flash image pairs. In *SIGGRAPH '04 : ACM SIGGRAPH 2004 Papers*, pages 664–672, New York, NY, USA, 2004. ACM. 193, 205

[PSZ08] J. Philbin, J. Sivic, and A. Zisserman. Geometric lda : A generative model for particular object discovery. In *Proceedings of the British Machine Vision Conference*, 2008. 7

[PVGV+04] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual modeling with a hand-held camera. *Int. J. Comput. Vision*, 59(3) :207–232, 2004. 103

[PW08] O. Pele and M. Werman. A Linear Time Histogram Metric for Improved SIFT Matching. In *ECCV08*, 2008. 174, 176, 177, 181, 183

[PW09] Ofir Pele and Michael Werman. Fast and Robust Earth Mover's Distances. In *ICCV*, 2009. 134, 147, 169

[QWH06] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng. Manga colorization. *ACM Trans. Graph.*, 25(3) :1214–1220, 2006. 193

[RAGS01] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Comput. Graph. Appl.*, 21(5) :34–41, 2001. 189, 190, 193, 194

[RBSW00] K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. A comparison of measures for visualising image similarity. In *Proceedings of The Challenge of Image Retrieval (CIR 2000)*, 2000. 168

[RDG08a] Julien Rabin, Julie Delon, and Yann Gousseau. A contrario matching of SIFT-like descriptors. In *Proc. ICPR*. IEEE Computer Society, 2008. 19

[RDG08b] Julien Rabin, Julie Delon, and Yann Gousseau. Circular Earth Mover's Distance for the comparison of local features. In *Proc. ICPR*. IEEE Computer Society, 2008. 171

[RDG09] J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. *SIAM Journal on Imaging Sciences*, 2(3) :931–958, 2009. 19, 171, 239

[RDGM10] J. Rabin, J. Delon, Y. Gousseau, and L. Moisan. MAC-RANSAC : reconnaissance automatique d'objets multiples. In *RFIA (à paraître)*, 2010. 73

[Ris78] J. Rissanen. Modeling by shortest data description. *Automatica*, 14 :465–471, 1978. 70

[RLSP07] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Segmenting, modeling, and matching video clips containing multiple moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3) :477–491, 2007. 13, 14

[RMLHM] Amandine Robin, Lionel Moisan, and Sylvie Le Hégarat-Mascle. An a-contrario approach for sub-pixel change detection in satellite imagery. Preprint MAP5 2009-15 , à paraître dans IEEE Transactions on Pattern Analysis and Machine Intelligence. 239

[ROF92] L.I. Rudin, S.J. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *PhysicaD*, 60 :259–268, 1992. 216

[Rou84] P. J Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79 :871–880, 1984. 55

[Rou85] P. J Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, B :283–297, 1985. 56

[RP05] Jason Repko and Marc Pollefeys. 3D models from extended uncalibrated video sequences : Addressing key-frame selection and projective drift. In *3DIM '05 : Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling*, pages 150–157, Washington, DC, USA, 2005. IEEE Computer Society. 72, 115, 232

[RT01] Mark A. Ruzon and Carlo Tomasi. Edge, junction, and corner detection using color distributions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11) :1281–1295, 2001. 133, 147, 168

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The Earth Mover's distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2) :99–121, 2000. 2, 132, 133, 134, 138, 147, 148, 152, 173, 177, 179

[Rub] Yossi Rubner. Code source EMD : http ://robotics.stanford.edu/ rubner/. 183

[SAH08] C. Silpa Anan and R.I. Hartley. Optimised KD-trees for fast image descriptor matching. In *CVPR*, pages 1–8, 2008. 17

[SAM08] Neus Sabater, Andres Almansa, and Jean-Michel Morel. Rejecting wrong matches in stereovision. *Preprint CMLA 2008-28*, 2008. 16, 20, 239

[SARK08] H. Sahbi, J.Y. Audibert, H. Rabarisoa, and R. Keriven. Context-dependent kernel design for object matching and recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, Jun 2008. 16

[SB97] Stephen M. Smith and J. Michael Brady. Susan—a new approach to low level image processing. *Int. J. Comput. Vision*, 23(1) :45–78, 1997. 200

[SBv04] Daniel Sýkora, Jan Buriánek, and Jiří Žára. Unsupervised colorization of black-and-white cartoons. In *NPAR '04 : Proceedings of the 3rd international symposium on Non-photorealistic animation and rendering*, pages 121–127, New York, NY, USA, 2004. ACM. 193

[Sch78] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 2(6) :461–464, 1978. 70

[Sch96] C. Schmid. *Appariement d'images par invariants locaux de niveaux de gris*. PhD thesis, Institut National Polytechnique de Grenoble, July 1996. 14, 16

[Sch06] Two-view multibody structure-and-motion with outliers through model selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(6) :983–995, 2006. Schindler, Konrad and Suter, David. 60, 72

[She87] Roger N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820) :1317–1323, 1987. 168

[SJ08] S. Shirdhonkar and D.W. Jacobs. Approximate Earth Mover's Distance in linear time. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 134

[SM97] Cordelia Schmid and Roger Mohr. Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 :530–535, 1997. 10

[SM06] R. Subbarao and P. Meer. Nonlinear mean shift for clustering over analytic manifolds. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 1168–1175, 2006. 58, 66

[SSN09] A. Singer, Y. Shkolnisky, and B. Nadler. Diffusion interpretation of nonlocal neighborhood filters for signal denoising. *SIAM, SIIMS*, 2(1) :118–139, 2009. 201, 203, 228

[SSS08] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *Int. J. Comput. Vision*, 2008. 15

[Ste91] R.S. Stephens. Probabilistic approach to the Hough transform. *IVC*, 9(1) :66–71, February 1991. 58

[Ste95] Charles V. Stewart. Minpran : A new robust estimator for computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(10) :925–938, 1995. 60, 62, 64, 65, 66, 78, 88, 90, 94, 96

[Ste97] Charles V. Stewart. Bias in robust estimation caused by discontinuities and multiple structures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(8) :818–833, 1997. 90

[Str89] James K. Strayer. Linear programming and its applications. In *Undergraduate Texts in Mathematics*. Springer, 1989. 175

[Sug78] Nariaki Sugiura. Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7(1) :13–26, 1978. 70

[Sur07] F. Sur. Invariant image descriptors and affine morphological scale-space. Research Report 6250, INRIA, July 2007. 10, 232, 247

[SW83] H.C. Shen and A.K.C. Wong. Generalized texture representation and metric. *CVGIP*, 23(2) :187–206, August 1983. 132

[SZ00] F. Schaffalitzky and A. Zisserman. Planar grouping for automatic detection of vanishing lines and points. *IVC*, 18(9) :647–658, June 2000. 90

[SZ06] J. Sivic and A. Zisserman. Video Google : Efficient visual search of videos. In *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 127–144. Springer, 2006. 7, 16, 18

[TD03] Philip H. S. Torr and Colin Davidson. Impsac : Synthesis of importance sampling and random sample consensus. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(3) :354–364, 2003. 59

[TF08] Roberto Toldo and Andrea Fusiello. Robust multiple structures estimation with J-linkage. In *ECCV (1)*, pages 537–547, 2008. 60, 66, 67, 68, 88

[TJT05] Yu-Wing Tai, Jiaya Jia, and Chi-Keung Tang. Local color transfer via probabilistic segmentation by expectation-maximization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Washington, DC, USA, 2005. IEEE Computer Society. 194

[TKLG07] F. Tarsha Kurdi, T. Landes, and P. Grussenmeyer. Hough-Transform and Extended RANSAC Algorithms for Automatic Detection of 3D Building Roof Planes from Lidar Data. In *Laser07*, page 407, 2007. 64

[TM94] Philip H. S. Torr and David W. Murray. Stochastic motion clustering. In *ECCV (2)*, pages 328–337, 1994. 66

[TM98] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV '98 : Proceedings of the Sixth International Conference on Computer Vision*, page 839, Washington, DC, USA, 1998. IEEE Computer Society. 187, 200

[TML01] Chi-Keung Tang, Gérard Medioni, and Mi-Suen Lee. N-dimensional tensor voting and application to epipolar geometry estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(8) :829–844, 2001. 67

[Tor95] P.H.S. Torr. *Outlier Detection and Motion Segmentation*. PhD thesis, University of Oxford, 1995. 60

[Tor97] P.H.S. Torr. An assessment of information criteria for motion model selection. *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 47–52, Jun 1997. 71, 85, 86

[Tor98] P. H. S. Torr. Geometric motion segmentation and model selection. *Phil. Trans. Royal Society of London A*, 356 :1321–1340, 1998. 68, 71, 72, 85, 86, 115

[Tor99] P. H. S. Torr. Model selection for structure and motion recovery from multiple images. Technical report, Microsoft Research, 1999. 71, 85

[Tor02] P. H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *Int. J. Comput. Vision*, 50(1) :35–61, 2002. 71, 72, 85, 89

[TTM04] Wai-Shun Tong, Chi-Keung Tang, and Gerard Medioni. Simultaneous two-view epipolar geometry estimation and motion segmentation by 4d tensor voting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9) :1167–1184, 2004. 67

[TZ00] P. H. S. Torr and A. Zisserman. MLESAC : a new robust estimator with application to estimating image geometry. *Comput. Vis. Image Underst.*, 78(1) :138–156, 2000. 60, 61, 62, 71, 75, 89, 235

[VCB06] Thomas Veit, Frédéric Cao, and Patrick Bouthemy. An a contrario decision framework for region-based motion detection. *Int. J. Comput. Vision*, 68(2) :163–178, 2006. 68, 239

[Ved08] A. Vedaldi. *Invariant Representations and Learning for Computer Vision*. PhD thesis, University of California at Los Angeles, 2008. 18

[Vil03] C. Villani. *Topics in optimal transportation*. American Math. Soc., 2003. 131, 134, 138, 141, 188

[VJ02] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2002. 8

[VL01] E. Vincent and R. Laganiere. Detecting planar homographies in an image pair. *Image and Signal Processing and Analysis, 2001. ISPA 2001. Proceedings of the 2nd International Symposium on*, pages 182–187, 2001. 64, 91, 93

[WAM02] Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. Transferring color to greyscale images. *ACM Trans. Graph.*, 21(3) :277–280, 2002. 193

[WPMK86] M. Werman, S. Peleg, R. Melter, and TY Kong. Bipartite graph matching for points on a line or a circle. *Journal of Algorithms*, 7(2) :277–284, 1986. 132, 137

[WPR85] M. Werman, S. Peleg, and A. Rosenfeld. A distance metric for multidimensional histograms. *CVGIP*, 32(3) :328–336, December 1985. 132, 133

[XOK] L. Xu, E. Oja, and P. Kultanen. A new curve detection method : randomized Hough transform (RHT), journal = Pattern Recogn. Lett., volume = 11, number = 5, year = 1990, issn = 0167-8655, pages = 331–338, doi = http ://dx.doi.org/10.1016/0167-8655(90)90042-Z, publisher = Elsevier Science Inc., address = New York, NY, USA. 57

[Yar85] L. Yaroslavsky. *Digital Picture Processing, An Introduction*. Springer-Verlag, Berlin, 1985. 200

[YK03] Wei-Qi Yan and M. S. Kankanhalli. Colorizing infrared home videos. In *ICME '03 : Proceedings of the 2003 International Conference on Multimedia and Expo*, pages 97–9100, Washington, DC, USA, 2003. IEEE Computer Society. 193

[YS06] L. Yatziv and G. Sapiro. Fast image and video colorization using chrominance blending. *Image Processing, IEEE Transactions on*, 15(5) :1120–1129, 2006. 193

[YS09] G. Yu and J.J. Slotine. Visual grouping by neural oscillators. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), Taipei*, 2009. 239

[YSST07] Gehua Yang, Charles V. Stewart, Michal Sofka, and Chia-Ling Tsai. Registration of challenging image pairs : Initialization, estimation, and decision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11) :1973–1989, 2007. 7

[ZK06a] Wei Zhang and Jana Kosecka. Generalized RANSAC Framework for Relaxed Correspondence Problems. In *3DPVT '06 : Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, pages 854–860, Washington, DC, USA, 2006. IEEE Computer Society. 13, 15, 16, 42, 91

[ZK06b] Wei Zhang and Jana Kosecká. Nonparametric estimation of multiple structures with outliers. In *WDV*, pages 60–74, 2006. 66, 68

[ZKM05] M. Zuliani, C. S. Kenney, and B. S. Manjunath. The MultiRANSAC Algorithm and its Application to Detect Planar Homographies. In *IEEE International Conference on Image Processing*, Sep 2005. 60, 64, 66, 88

[ZMLS07] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories : A comprehensive study. *Int. J. Comput. Vision*, 73(2) :213–238, 2007. 133, 152, 172

[ZWG06] Qing-Fang Zheng, Wei-Qiang Wang, and Wen Gao. Effective and efficient object-based image retrieval using visual phrases. In *MULTIMEDIA '06 : Proceedings of the 14th annual ACM international conference on Multimedia*, pages 77–80, New York, NY, USA, 2006. ACM. 147

[ZYHT07] L. Zhu, Y. Yang, S. Haker, and A. Tannenbaum. An image morphing technique based on optimal mass preserving mapping. *Image Processing*, 16(6) :1481–1495, June 2007. 135

[ZYS09] Huiyu Zhou, Yuan Yuan, and Chunmei Shi. Object tracking using SIFT features and mean shift. *Computer Vision and Image Understanding*, 113(3) :345 – 352, 2009. Special Issue on Video Analysis. 16