Représentation et reconnaissance d'objets par champs réceptifs

Vincent Colin de Verdière

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

THÈSE

pour obtenir le grade de

DOCTEUR DE L'INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

Discipline : Informatique — Option Imagerie Vision Robotique

présentée et soutenue publiquement
par

Vincent COLIN de VERDIÈRE

le 10 décembre 1999

REPRÉSENTATION ET RECONNAISSANCE D' OBJETS PAR CHAMPS RÉCEPTIFS

Directeur de thèse : M. James L. CROWLEY

JURY

M. Alain CHEHIKIAN, Président
M. Mike BRADY
M. Bernt SCHIELE
M. Jan Olof EKLUNDH, Rapporteur
M. Marc RICHETIN, Rapporteur

Thèse préparée dans le laboratoire GRAVIR – IMAG au sein du projet PRIMA
INRIA Rhône-Alpes, 655 av. de l'Europe, 38330 Montbonnot Saint Martin.

Résumé

Cette thèse se place dans le domaine de la modélisation et de la reconnaissance d'objets par leur apparence. Chaque objet est modélisé par une collection d'images et la reconnaissance est obtenue par l'appariement d'une nouvelle image avec une image modèle. Les images sont modélisées par des mesures sur des caractéristiques locales. Plusieurs bases de descripteurs locaux sont évaluées théoriquement et expérimentalement et la base des dérivées de Gaussiennes est sélectionnée pour ses propriétés de discriminabilité avec une description très concise et son paramétrage en orientation et en échelle. Une invariance à l'orientation de la caméra par rapport à l'objet est obtenue par un calage des dérivées sur la direction du gradient et une invariance à l'échelle est obtenue par une technique novatrice qui consiste à sélectionner en chaque point une échelle caractéristique pour décrire son voisinage. Cette échelle caractéristique correspond au maximum en échelle d'un opérateur Laplacien. Ces invariances sont validées par des expérimentations systématiques.

Dans notre système, une image est décomposée en une grille de fenêtres recouvrantes puis représentée par une grille de descripteurs locaux calculés sur ces fenêtres. Cette représentation très redondante nous a permis de définir deux stratégies de reconnaissance robustes : l'une fondée sur un vote et l'autre fondée sur une stratégie par prédiction–vérification qui consiste à découper la reconnaissance en une phase de génération d'hypothèses d'appariement pour une fenêtre suivi d'une phase de vérification de ces hypothèses sur les fenêtres voisines en incluant des contraintes de cohérence spatiale à cette vérification.

Mots-clés

reconnaissance d'objets, apparence, vision par ordinateur, modélisation par caractéristiques locales, invariance à l'échelle, cohérence spatiale, vote, prédiction–vérification.

Object representation and recognition using receptive fields
Abstract

This thesis belong to the field of modelisation and recognition of objects using their appearance. Each object is modeled by a collection of images and recognition is obtained by the matching between a new image and a model image. Images are modeled by measures on local features. Several local descriptor bases are theoretically and experimentally evaluated and a Gaussian derivative basis is selected for its properties which include: high discriminability for a concise description, scalability and steerability. Invariance to orientation is obtained by setting derivative directions according to the local gradient. Invariance to scale is obtained by a new technique which locally selects a characteristic scale for describing a neighborhood. This scale correspond to a maximum over scale of a Laplacian operator. These invariances are experimentally validated.

In our system, an image is decomposed in a grid of overlapping windows which is represented by a corresponding grid of local features computed on these windows. This highly redundant representation enables us to design two robust recognition techniques : the first one is based on a simple vote and the second one based on the prediction–verification method consists in splitting the recognition process in an hypothesis generation step for a single window followed by a verification step which checks the generated hypothesis on the neighboring windows with inclusion of geometric coherence constraints.

keywords

object recognition, appearance, computer vision, modelisation using local features, scale invariance, spatial coherence, vote, prediction–verification.

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au travail présenté dans cette thèse.

Tout d'abord, je voudrais remercier mon directeur de thèse James L. Crowley pour m'avoir fourni la motivation qui m'a souvent manquée, un environnement de travail fructueux et de nombreux conseils pour ce travail. Je remercie les personnes qui me font l'honneur de participer à mon jury : Marc Richetin et Jan Olof Eklundh comme rapporteurs, Alain Chehikian pour président ainsi que Mike Brady et Bernt Schiele. Je les remercie pour le temps et le travail fournis pour l'évaluation de ma thèse.

Je tiens aussi à remercier, plus particulierement, Bernt Schiele pour m'avoir introduit à ce domaine de recherche dans le cadre de mon DEA. Je remercie Jérôme Martin mon co-bureau pendant trois ans. Je remercie aussi Augustin Lux pour ses conseils précieux sur plusieurs aspects de ma thèse et pour son aide à la correction de mon manuscrit. Je remercie particulièrement Olivier Chomat pour la coopération très profitable que nous avons eu au cours de la dernière année et pour son aide et ses conseils pour la correction de mon manuscrit.

D'autre part, je tiens à remercier Augustin Lux, Bruno Zoppis et Claude Poizat pour m'avoir fourni le précieux outil *Ravi* et Christophe Le Gal pour tout le temps passé à la lourde tache de maintenance de notre système informatique.

Je remercie toute l'équipe PRIMA pour l'ambiance de travail particulièrement décontractée. Je remercie en particulier Alvaro, Bénédicte, Bill, Bruno, Christophe (l'ancien), Claude, Claus, Daniela, Erik, Fabrice, Frank, Guillaume, Jean-Baptiste, Karl, Nikla, Steve et Vincent.

Je remercie naturellement mes parents pour leur soutien en particulier pendant les périodes de découragement.

Vincent, le 10 décembre 1999

Sommaire

Chapitre 1

Introduction

1.1 Reconnaissance d'objets 3D

La reconnaissance d'objets est un problème fondamental de la vision artificielle qui vise à identifier les éléments pertinents dans une image d'une scène. Elle consiste à mettre en correspondance l'image d'un objet avec un modèle de celui-ci. En vision par ordinateur, la reconnaissance d'objets se décompose en deux phases :

- Une phase d'apprentissage ou modélisation permet d'associer un modèle à chaque objet. Le modèle d'un objet peut être construit manuellement ou automatiquement. La base d'apprentissage est constituée par des images de l'objet ou par des données extérieures. Cette phase permet de construire une base de modèles qui est l'unique représentation des objets pour la seconde phase.

- La phase de reconnaissance consiste à apparier automatiquement une image inconnue avec un élément de la base des modèles. Cet appariement est obtenu en mesurant le degré d'appartenance de l'image à ce modèle. Il donne l'identification de l'objet et, pour certaines stratégies d'appariement, la détection simultanée de la position du modèle dans l'image.

La grande variété des classes d'objets à reconnaître implique une grande variété dans les modèles de représentation possibles. Une approche classique de la reconnaissance se concentre sur des objets de type manufacturés. Ces objets sont caractérisés par une forme plutôt polyhèdrique qui permet de les représenter par un modèle géométrique de type CAO. Leurs modèles sont fondés sur des indices visuels d'images comme des contours ou des points caractéristiques. Ainsi, ROBERTS [Rob65] a proposé l'un des premiers systèmes de reconnaissance d'objets 3D qui consiste à modéliser la scène 3D qui a générée une image plutôt que l'image elle-même. Son système reconnaît des objets modélisés dans des images ainsi que leurs positions et orientations, mais il se limite à quelques objets polyhédriques simples représentés par les arêtes de leurs faces. Cette direction de recherche

suivie par de nombreux auteurs a présenté de grandes difficultés pour la reproduction des résultats obtenus en laboratoire Ceci s'explique par la faiblesse des indices visuels utilisés : les segments de droites sont théoriquement robustes aux variations du point de vue d'observation ou de l'éclairage mais, en pratique, leurs extrémités sont difficiles à détecter de façon précise et il arrive souvent qu'un segment ne soit pas retrouvé ou qu'il soit coupé pendant sa détection.

Pour remédier à ces inconvénients, une nouvelle approche est apparue récemment en lien avec les progrès techniques. Cette approche modélise les objets par leurs images elles-mêmes et refuse ainsi les modèles de type géométrique ou fondés sur des caractéristiques instables comme les segments. On parle, dans ce cas, d'une modélisation fondée sur l'apparence des objets qui permet une reconnaissance via l'appariement entre une image et une image modèle de l'objet. Nous nous plaçons dans cette approche. Le modèle d'un objet est obtenu automatiquement en parcourant l'ensemble de ce qui est observable sur cet objet. Pour cela, la sphère des vues possibles est échantillonnée et chaque élément de cet échantillonnage est une image qui montre l'une des apparences de l'objet. Ainsi, un objet est représenté par une *collection d'images* et la reconnaissance est fondée sur l'*appariement* d'une nouvelle image de l'objet avec une image de cette collection. Cet appariement donne simultanément une identification et un positionnement de l'objet.

Le problème de la reconnaissance d'objets est converti en un problème d'appariement d'une image avec une image modèle. Cet appariement nécessite une modélisation des images robuste à différentes perturbations comme le bruit de la chaîne d'acquisition, un changement d'éclairage, une variation faible du point de vue ou l'occultation partielle de l'objet dans l'image. Pour cela, nous proposons, dans cette thèse, une modélisation des images fondée sur une décomposition en sous-images ou *imagettes* recouvrantes de petites dimensions par rapport à la taille de l'image. Chacune des imagettes est représentée par un *vecteur de mesures*. L'appariement entre images est obtenu grâce à l'appariement des vecteurs de mesures. L'utilisation de contraintes de *cohérence spatiale* entre appariements permet d'obtenir une technique de reconnaissance d'objets robuste aux changements de points de vue et à l'occultation partielle.

Les mesures des caractéristiques locales sont les réponses de *champs réceptifs visuels*. Ce terme issu de la biologie permet de modéliser le cortex visuel humain comme un ensemble de champs réceptifs visuels. Ces champs sont des capteurs localisés sur la rétine qui réagissent à des stimuli issus des cellules photo-receptives de l'œil, les cônes et les bâtonnets. Chacun des champs réceptifs peut être modélisé par un réseau de neurones qui mesure la présence d'une caractéristique locale particulière. La reconnaissance par le système visuel humain est fondée sur l'analyse des réponses d'une large gamme de champs réceptifs localisés sur un maillage dense de la rétine. En vision par ordinateur, il est possible de définir des champs réceptifs sur des images. Un champs réceptif synthétique ou *descripteur local* est un *opérateur* qui mesure la présence d'une caractéristique locale particulière. La réponse du champs réceptif est une mesure de la quantité d'une caractéristique locale particulière présente en un point. L'utilisation d'une base de descripteurs locaux

permet d'évaluer en tout point d'une image un vecteur de mesures locales qui forme une représentation condensée du voisinage du point.

1.2 Une stratégie de reconnaissance d'objets par appariements de caractéristiques locales

Cette thèse présente une technique de modélisation d'objets par caractéristiques locales pour la reconnaissance suivant deux axes principaux : une étude de la stratégie de modélisation et de ses paramètres puis une étude de deux stratégies de reconnaissance utilisant la modélisation proposée.

La première partie de cette thèse (chapitres 3 à 5) présente une technique de modélisation d'objets par des images les représentant. Pour cela, une stratégie de modélisation des images est étudiée en évaluant plusieurs bases de descripteurs locaux et différentes techniques d'apprentissages des mesures correspondantes dans une base de modèles. L'étude de plusieurs bases de descripteurs locaux a pour but d'évaluer la qualité de la représentation locale d'un signal sur ces bases. Cette qualité est évaluée par la discrimination qui est la faculté d'une base de filtres de différencier les imagettes correspondants à des points physiques différents. L'évaluation est faite en mesurant sur une base de test le nombre de points permettant de reconnaître les objets. Au cours de cette évaluation, il apparaît qu'une partie des points ne sont pas discriminants et peuvent être rejetés. Cette étude aboutit à la sélection d'une base de filtres de dérivées de Gaussiennes pour sa robustesse au bruit et pour son paramétrage direct en orientation et en échelle. L'utilisation d'un calage automatique en échelle et en orientation fondé sur le contenu du voisinage permet d'obtenir des mesures locales invariantes par rapport à ces deux paramètres. Ces aspects sont développés aux chapitres 3 et 4.

Plusieurs stratégies de modélisation des images par descripteurs locaux sont disponibles dans la bibliographie. Ces stratégies peuvent être étudiées indépendamment du choix des descripteurs locaux. Une première classe de techniques propose la modélisation d'une image par l'extraction de points particuliers. Ces points sont projetés sur l'espace de caractéristiques locales. L'appariement de ces points suffit à obtenir un système de reconnaissance. L'inconvénient de cette approche est la sélection de points qui est difficile à obtenir. Souvent, les points corrects ne sont pas sélectionnés. Une deuxième classe de techniques modélise une image par un histogramme des caractéristiques locales visibles dans celle-ci. Cette approche présente l'avantage de modéliser l'ensemble des points de l'image et l'utilisation d'un histogramme réduit fortement la quantité de mémoire nécessaire au stockage du modèle. L'interprétation probabiliste de cet histogramme permet d'obtenir un système de reconnaissance très robuste. Par contre, cette approche ne permet pas de retrouver la pose de l'objet observé. De plus, une grande partie de l'information spatiale est perdue. Nous proposons une stratégie d'apprentissage issue de ces deux approches.

Elle consiste à effectuer un apprentissage de l'ensemble des points observés. Chaque point est stocké comme un couple (vecteur de mesures, identificateur) dans une structure de recherche. Le vecteur est la clé de recherche et l'identificateur donne la position du point dans l'image modèle. Un appariement est possible pour la majorité des points de façon robuste et la conservation de l'information structurelle donne une grande robustesse à la technique (voir chapitre 5).

La seconde partie de l'étude est la stratégie de reconnaissance à partir de la base de modèles obtenue dans la première partie. Deux stratégies de reconnaissance sont proposées : un vote ou transformée de Hough peut être effectué pour obtenir une reconnaissance robuste avec évaluation de la pose simultanée. L'algorithme de vote posant des problèmes importants de paramétrage, un deuxième algorithme a été proposé qui se fonde sur le paradigme prédiction–vérification. Une première recherche génère des hypothèses vraisemblables sur le (ou les) objets présents puis une deuxième phase vérifie ces hypothèses et sélectionne les objets les plus vraisemblables.

1.3 Contributions principales de cette thèse

L'étude théorique et expérimentale de plusieurs bases de filtres et de leurs propriétés a permis de sélectionner une base fondée sur des dérivées de Gaussiennes donnant des mesures invariantes à l'orientation et à l'échelle des images. Une extension de la technique de sélection du paramètre d'échelle proposée par LINDEBERG [Lin98] nous fournit une description invariante à l'échelle. Cette technique novatrice est extensible à d'autres problèmes de vision pour lesquels les variations incontrôlées de l'échelle sont souvent un problème clé.

La modélisation intégrale de la structure des images a été utilisée pour former la base des modèles et sa forte redondance permet de définir une stratégie de reconnaissance très robuste. Elle a été utilisée pour effectuer une étude systématique des bases de descripteurs locaux et de leur discriminabilité. De plus, le choix consistant à enregistrer avec chaque descripteur un identificateur de l'image, de la position de l'imagette et de ses paramètres d'échelle et d'orientation a permis une utilisation directe de contraintes de cohérence spatiale entre appariements distincts.

L'utilisation des stratégies de reconnaissance sur deux applications montre la validité du système complet de modélisation puis de reconnaissance d'objets dans des conditions difficiles : occultation partielle, apprentissage non contrôlé ou bruit important. De plus, en utilisant la stratégie fondée sur le paradigme Prédiction–Vérification, nous avons pu mettre en place une coopération efficace avec un système de reconnaissance fondé sur des caractéristiques spatio–temporels. L'application de la détection de la transformation affine entre deux images d'un objet est abordée et quelques exemples de détection de similitude entre images montrant des variations importantes sont montrés. Nos résultats ouvrent de nouvelles perspectives pour cette application grâce à sa robustesse par rapport

aux variations d'échelle.

1.4 Contenu détaillé du rapport par chapitre

Les chapitres de cette thèse sont résumés ici.

Le chapitre 2 présente un état de l'art partiel de techniques de modélisation d'objets fondées sur leur apparence. L'apparence d'un objet peut être défini comme l'ensemble de tout ce qui observable sur un objet. Cet ensemble peut être échantillonné pour un objet 3D par une collection d'images qui le représente. L'objet de ce chapitre est de présenter cette approche en se fondant sur un ensemble de travaux déjà proposés dans ce domaine. Deux sous–classes sont présentées : la première se fonde sur la modélisation globale de l'apparence d'objets en utilisant la technique statistique de l'Analyse en Composantes Principales. Cette modélisation présente l'inconvénient majeur de nécessiter une segmentation et normalisation des objets pour leur modélisation et leur reconnaissance. Pour limiter ces inconvénients, la deuxième classe est basée sur une modélisation locale. Elle fonde la reconnaissance sur l'appariement de caractéristiques locales. L'utilisation de contraintes de cohérence spatiale permet d'augmenter la robustesse des cette approche. Nous proposons à la fin du chapitre une modélisation d'objets 3D fondée sur leurs apparences par un apprentissage structurel intégral des images.

Le chapitre 3 présente une étude théorique et expérimentale de plusieurs bases de descripteurs locaux d'images. L'objectif est de sélectionner la base de descripteurs caractérisant une image de la façon la plus concise et précise possible. Une base obtenue par une Analyse en Composantes Principales sur des voisinages est évaluée comme extension de l'approche globale. Cette base de descripteurs permet une caractérisation optimale d'images en l'absence de variations des paramètres de points de vue que sont l'échelle et la rotation autour de l'axe optique de la caméra. Par la suite, des bases de filtres fondées sur des dérivées de Gaussiennes sont évaluées. L'utilisation des dérivées de Gaussiennes jusqu'à l'ordre 3 permet une description locale des images robuste aux variations des paramètres d'échelle et d'orientation 2D. La robustesse à l'orientation est obtenue en utilisant la propriété d'orientabilité des dérivées de Gaussiennes et l'orientation locale du Gradient. La robustesse aux variations d'échelle est obtenue par une sélection automatique du paramètre d'échelle des filtres à partir d'un détecteur fondé sur un opérateur Laplacien. Cette base de filtres invariants en échelle et en orientation fournit expérimentalement un taux de reconnaissance par un vecteur isolé très important : plus de 50% des points entraînent une reconnaissance directe sur les bases de test et moins de 10% des points entraînent des faux appariements. Cette stratégie de modélisation est très discriminante et permet de définir dans le chapitre 6 une stratégie de reconnaissance d'objets robuste et rapide.

Le chapitre 4 présente une étude systématique de la sensibilité des descripteurs locaux utilisés par rapport à une large gamme de perturbations. Ces perturbations incluent le bruit lié à la capture d'une image comme la numérisation ou le flou, le changement de point de vue avec les paramètres d'échelle et d'orientation et la variation de l'éclairage avec, dans le cas d'une variation en intensité, la présentation d'une normalisation par l'énergie fondée sur un modèle linéaire de caméra. La distribution des distances entre vecteurs de caractéristiques locales des différentes perturbations est mesurée de façon à évaluer un seuil sur cette distance permettant de décider de la similarité entre vecteurs et ainsi, la génération d'hypothèses pendant la phase de reconnaissance.

Le chapitre 5 introduit le problème de l'apprentissage d'un modèle local pour les images représentant un objet. Cet apprentissage peut être effectué suivant plusieurs stratégies comme la sélection puis l'apprentissage d'un ensemble de points a priori discriminants via l'usage d'un détecteur de points d'intérêts, ou par l'apprentissage complet du modèle soit par l'utilisation d'histogrammes multidimensionnels avec une perte d'information spatiale soit par le stockage de grilles 2D de caractéristiques locales choisi dans cette thèse. Cet apprentissage complet d'un modèle local implique des difficultés sur le stockage et la recherche dans ce modèle. Une structure de données simple est implémentée et permet une détection rapide de vecteurs proches d'un nouveau vecteur de mesures par l'utilisation d'une stratégie par "branch and bound". Le modèle complet présente de nombreux vecteurs similaires entre eux ce qui perturbe la reconnaissance. Plusieurs stratégies sont proposées pour limiter cette redondance comme l'élimination des vecteurs équivalents ou la suppression des vecteurs correspondants à des imagettes non discriminantes car extrêmement fréquentes. Elles permettent de supprimer jusqu'à 40% des points de la structure de données sans perte notable en terme de reconnaissance.

Le chapitre 6 propose une stratégie de reconnaissance d'objets 3D à partir d'un apprentissage structurel d'un modèle des images représentant les objets. Ce chapitre pose, d'abord, le problème de l'évaluation d'une hypothèse fondée sur des appariements multiples. Cette évaluation nécessite de pouvoir regrouper des appariements compatibles puis d'évaluer, à des fins de comparaison entre hypothèses, un score pour chacune des hypothèses. Plusieurs stratégies de sélection des points à reconnaître dans une images sont proposées. En particulier, une stratégie fondée sur la sélection des points correspondants aux vecteurs de caractéristiques les moins fréquents dans la base est utilisée. Puis, une stratégie directe de reconnaissance par vote ou transformée de Hough généralisée est étudiée et permet en particulier une détection de similitude 2D entre deux images d'un même objet. Le chapitre termine par la proposition d'une stratégie fondée sur la paradigme prédiction–vérification pour la reconnaissance d'objets. Cette stratégie se découpe en deux phases. La première phase est une génération d'hypothèses vraisemblables par la recherche de vecteurs de caractéristiques a priori discriminants, puis, la deuxième est

une vérification des hypothèses par la vérification de la présence des voisins des hypothèses. Cette stratégie est très discriminante et génère expérimentalement très peu de faux appariements.

Le chapitre 7 présente deux applications des techniques proposées dans cette thèse : la première est l'estimation de position en robotique mobile. Elle consiste à apparier l'image vue par le robot avec une image modèle prise à la même position. Cet appariement permet l'estimation de la position du robot sur un chemin visuel préalablement effectué. Une deuxième application est un défi : reconnaître des poissons rouges qui sont des objets vivants et donc incontrôlables aussi bien pendant la phase d'apprentissage que pour la reconnaissance. Nous avons mis en place un système faisant coopérer une technique d'analyse du mouvement avec la technique de reconnaissance d'objets proposée dans cette thèse,

Le chapitre 8 analyse les apports de cette thèse sur la modélisation et la reconnaissance par des caractéristiques locales. Ces apports incluent la sélection et l'étude d'une base de descripteurs locaux fondée sur des dérivées de Gaussiennes. Ses propriétés d'équivariance à l'orientation et à l'échelle sont évaluées et un stratégie novatrice de modélisation robuste aux variations d'échelle est proposée. Elle consiste à détecter en chaque point une échelle caractéristique de façon à caler les descripteurs suivant cette échelle et obtenir une invariance à l'échelle. D'autre part, une modélisation d'images par l'ensemble des caractéristiques locales est utilisée et permet la définition d'une stratégie de reconnaissance robuste à l'occultation partielle fondée sur le paradigme prédiction–vérification.

Chapitre 2

Modélisations d'objets fondées sur l'apparence

Ce chapitre présente plusieurs techniques de modélisations fondées sur la notion d'apparence qui ont inspiré ou complètent l'étude de cette thèse. Le principe fondamental de ces techniques est une modélisation automatique des objets par tout ce qui observable sur ceux-ci. Cet ensemble d'observations est appelé apparence de l'objet et peut être décrit par une collection des vues possibles de cet objet lorsque les paramètres de vue comme la position ou l'éclairage varient. Chaque vue est définie par rapport à un capteur qui, dans le cadre de cette thèse, est une caméra. Une vue est une image représentée par un tableau 2D et un objet est représenté par une collection d'images. D'autres capteurs peuvent être utilisés tel un capteur laser. Dans le cas d'objets mobiles, un point de vue peut être défini comme une séquence d'images capturées par une caméra représentée par une tableau 3D. La reconnaissance est obtenue par l'appariement entre une nouvelle vue d'un objet et une vue modèle.

La collection d'images représentant un objet est un échantillonnage d'une fonction théorique donnant toutes les apparences de l'objet suivant les conditions d'observations. Cette fonction donne l'intensité lumineuse pour toutes les positions possibles d'un observateur. ADELSON et BERGEN [AB91] lui donnent le nom de fonction plénoptique[1]. Elle capture toute la variabilité que l'on peut trouver dans les images d'un objet. L'intensité lumineuse sur la rétine d'un observateur varie suivant 7 paramètres : $I(\Theta, \Phi, \lambda, t, V_x, V_y, V_z)$. Le couple (Θ, Φ) est la position sur la rétine, λ est la longueur d'onde pour laquelle est observée l'intensité lumineuse, t est l'instant d'observation et (V_x, V_y, V_z) est la position de l'observateur dans la scène. En pratique, la caméra effectue une intégration sur le paramètre λ. En vision par ordinateur, cette fonction peut être paramètrée de façon équivalente en considérant un plan image fictif à distance unité de la pupille par $I(x, y, \lambda, t, V_x, V_y, V_z)$ avec (x, y) les coordonnées dans ce plan image.

1. Plénoptique est un terme issu des racines latines *plenus* complet et *opticus* voir.

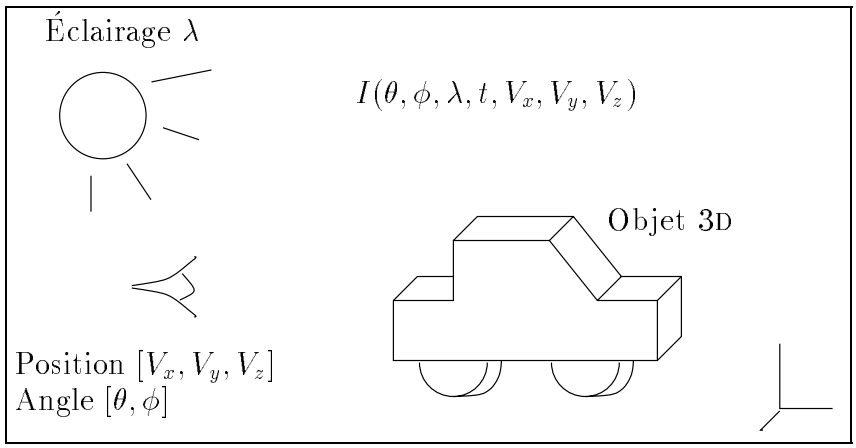


FIG. 2.1 – *Illustration de la Fonction plénoptique I de l'apparence d'une scène. Cette fonction admet 7 paramètres définissant la position de l'observateur, l'instant et la longueur d'onde de l'observation. Cette fonction décrit tout ce qu'il est possible de voir dans la scène.*

Cette modélisation représente un objet par une collection d'images montrant toutes ses apparences. Deux stratégies de modélisation des images sont présentées ici : l'apprentissage global de chacune des images en considérant qu'une image représentant un objet est un modèle complet et indivisible de l'objet ce qui ne permet pas facilement d'appariement si une partie de cette image est modifiée par exemple par la présence d'un deuxième objet ou d'un fond non uniforme. Pour limiter ces inconvénients, d'autres techniques représentent les images par des mesures locales qui suppriment les pré-traitements comme la segmentation et la normalisation nécessaires à une modélisation globale d'images. De plus, l'occultation partielle d'un objet n'est plus un obstacle à la reconnaissance.

2.1 Modélisation d'objets par une collection d'images

Un objet peut être modélisé par une collection d'images formant un échantillonnage de la sphère des vues possibles. La précision de cet échantillonnage doit être fondée sur la vitesse de variation des images de l'objet par rapport aux changements de point de vues. La modélisation est correcte si toute nouvelle prise de vue de l'objet est similaire à l'une des images de l'échantillonnage. Dans le cadre de la reconnaissance, ceci se traduit par le fait qu'il doit être possible d'apparier toute nouvelle image de l'objet à l'une des images de cet échantillonnage. La figure 2.2 montre l'exemple de l'échantillonnage de la sphère des vues possibles d'un objet en considérant une seul paramètre de rotation.

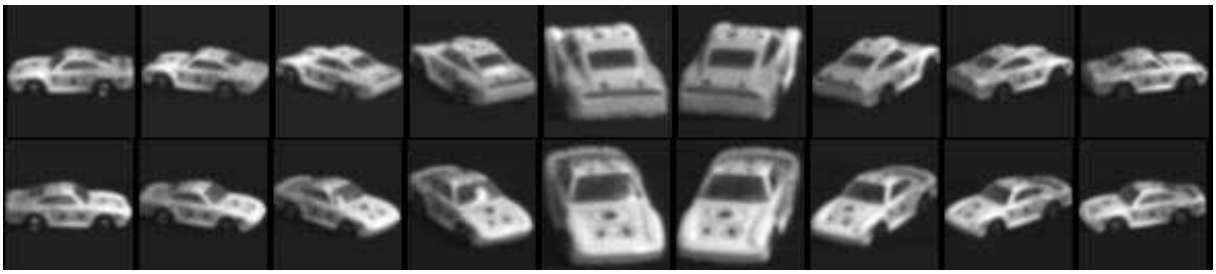


FIG. 2.2 – *Sous-ensemble des 72 images de l'objet 8 de la base Columbia présentée en annexe A.1. La modélisation d'un objet est effectuée par une collection d'images de cet objet formant un échantillonnage de la sphère des vues possibles.*

2.1.1 Appariements visuels

L'appariement direct d'images ou d'imagettes est utilisable sous des restrictions assez sévères des points de vues envisageables et pour un objet simple. Il peut être utilisé efficacement pour l'appariement d'indices visuels comme par LAN [Lan97] entre deux images ou pour le suivi d'objets. Cet appariement direct utilise une mesure de corrélation entre images pour mettre en correspondance une image avec un modèle. Différentes techniques de corrélation peuvent être utilisées suivant les conditions expérimentales (voir MARTIN [MC95]). Néanmoins, cette approche ne peut pour des raisons de complexité, être directement appliquée pour de nombreux objets sous de multiples points de vues. Il faut, en effet, effectuer un parcours exhaustif de toutes les images modèles de façon à déterminer l'image la plus proche de celle observée. Il est possible de remplacer ce parcours exhaustif par un adressage direct dans un espace de caractéristiques de faible dimensionnalité obtenu par une technique statistique comme le montre la section suivante.

2.1.2 Application à la reconnaissance d'objets

La transformation de Karhunen-Lœve (ou Analyse en Composantes Principales) permet de déterminer un sous-espace optimal pour la représentation d'un ensemble de données. En vision par ordinateur, cette technique est adaptée au problème de reconnaissance d'objets. L'espace est optimal dans le sens où il minimise le nombre de dimensions nécessaires pour représenter une proportion fixée de la variance des données. Ainsi, le sous-espace de représentation obtenu décrit les données avec le minimum de dimensions. D'autre part, les distances dans ce sous-espace sont une approximation d'une mesure de corrélation dans l'espace image initial. Cet espace de représentation est très discriminant pour des tâches de reconnaissance d'objets ou de scènes nécessitant un appariement d'images.

La technique statistique de l'ACP permet de définir un nouvel espace de représentation d'images ou, plus généralement, d'un signal numérique vectoriel. Les dimensions

de l'espace de représentation sont définies par des descripteurs appelés *images propres*[2]. SIROVITCH et KIRBY [SK87] ont proposé cette représentation pour caractériser, puis reconnaître des visages. Par la suite, TURK et PENTLAND [TP91, Tur91] ont utilisé cette technique afin de réaliser un système temps-réel de reconnaissance de visages. Une détection et segmentation du visage est effectuée en analysant le mouvement entre images, puis l'image est normalisée (suppression du fond et redimensionnement) et projetée sur l'espace propre. Les points proches dans cet espace correspondent à des visages similaires ce qui permet une reconnaissance fiable. Cette technique peut être généralisée à une large gamme d'objets et peut être utilisée dans les domaines de la reconnaissance d'objets rigides [PPP98], la reconnaissance d'objets non rigides comme des gestes ou des visages [MC97, MHC98, FG98], la compression d'images vidéo [CCBS97, VSC99] ou pour l'estimation de position d'un robot mobile par appariement d'images [PC98, Pou98, WSV99] ou par appariement de données télémétriques issues d'un capteur Laser [CWS98, Wal97].

L'espace de représentation obtenu par ACP permet de généraliser la représentation par une collection d'images en une représentation par une variété où une interpolation est effectuée entre les points représentants les images. Cette généralisation a été proposée dans le cadre d'un système de reconnaissance d'objets 3D. La représentation d'un objets par une variété supprime l'échantillonnage de plusieurs paramètres de la fonction plénoptique comme l'éclairage ou le point de vue.

La section suivante présente une description précise de l'approche de la reconnaissance d'objets par une ACP sur les images représentant les objets.

2.1.3 Reconnaissance d'objets 3D par Images Propres

La technique de reconnaissance d'objets par images propres telle qu'elle est proposée, par exemple, par MURASE et NAYAR [MN95] est fondée sur la technique de l'Analyse en Composantes Principales des vecteurs images représentant les points de vues possibles des objets. Chaque image d'apprentissage est représentée par un point dans un espace de représentation. L'ensemble des points correspondants aux images de la base d'apprentissage forme la base des modèles. L'appariement d'une nouvelle image est obtenu par sa normalisation en luminance et en échelle. Cette normalisation limite le nombre d'apparences possibles pour un objet. Puis l'image est projetée sur l'espace de description et le point obtenu peut être apparié avec ses plus proches voisins dans la base des modèles de façon à obtenir l'identification et la pose de l'objet. Cet appariement est effectué de façon rapide par l'utilisation d'une structure de données adaptée pour stocker la base des modèles.

2. En anglais : EigenImages

Apprentissage des images par Analyse en Composantes Principales : Les statistiques procurent la transformation de Karhunen-Lœve [Fuk90] pour réduire le nombre de dimensions d'un ensemble de données en évaluant les directions de l'espace les plus représentatives (en terme de variance). Ces directions sont appelées les composantes principales des données.

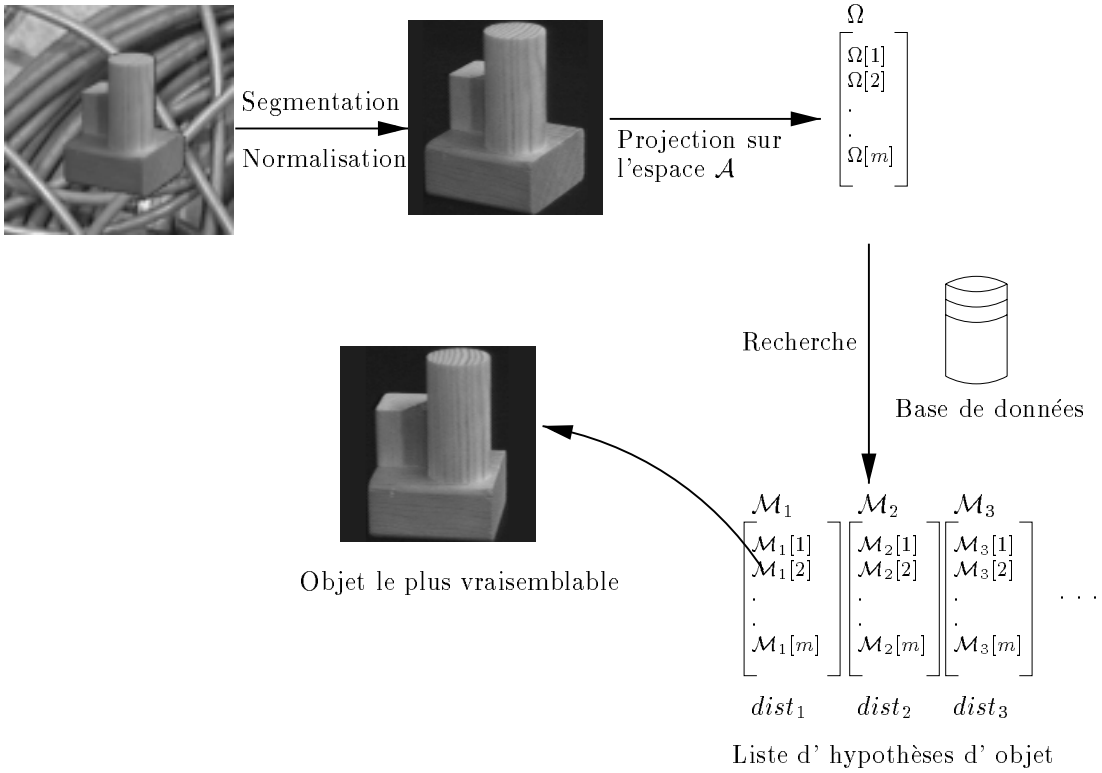


FIG. 2.3 – *Reconnaissance d'une image. Ce schéma illustre la stratégie de reconnaissance d'objets proposée par* MURASE *et* NAYAR . *Une image de la scène est capturée, puis l'objet contenue dans celle-ci est segmenté puis normalisé en taille. L'image obtenue est projetée sur l'espace de description $\mathcal{A}$. Le vecteur Ω obtenu est recherché dans la base de modèles. Les vecteurs $\mathcal{M}_k$ des variétés les plus proches sont appariés à ce vecteur. La distance $dist_k$ entre Ω et $\mathcal{M}_k$ donne un indice de confiance sur les appariements. Sur l'exemple, le vecteur $\mathcal{M}_1$ correspond à une hypothèse d'image modèle correcte.*

– *Données*: un ensemble de N images J_k normalisées de taille $w \times h$. L'objectif est de représenter cet ensemble $\mathcal{J} = \{J_1, J_2, \ldots, J_k, \ldots, J_N\}$ de façon minimale. Plus précisément le nombre d'octets permettant de représenter chaque image doit être le plus petit possible. Les images J_k sont interprétées comme des vecteurs de longueurs

$L = w \times h$. Une matrice $\mathbf{J}$ est définie comme la concaténation des vecteurs colonnes J_k. $\mathbf{J}$ est une matrice de dimensions $L \times N$:

$$\mathbf{J} = [J_1 J_2 \dots J_k \dots J_N]$$

– *Calcul de la matrice de covariance* : soit $\bar{J}$ le vecteur moyen de l'ensemble $\mathcal{J}$ obtenu par :

$$\bar{J} = \frac{1}{N} \sum_{k=1}^{N} J_k$$

Soit $\tilde{J}_k$ l'image J_k dont on soustrait la moyenne $\bar{J}$ et $\tilde{\mathbf{J}}$ la matrice ainsi obtenue.

$$\begin{aligned} \tilde{J}_k &= J_k - \bar{J} \\ \tilde{\mathbf{J}} &= [\tilde{J}_1 \tilde{J}_2 \dots \tilde{J}_k \dots \tilde{J}_N] \end{aligned}$$

Une fois ces éléments définis, il est possible de calculer $\mathbf{Q}$ la matrice de covariance de $\mathcal{J}$:

$$\mathbf{Q} = \frac{1}{N} \tilde{J} \tilde{J}^t$$

– *Calcul des vecteurs propres et des valeurs propres* : les techniques de Householder ou de Jacobi [PFTV86] permettent de diagonaliser la matrice $\mathbf{Q}$ et d'obtenir un ensemble $\mathcal{V}$ de vecteurs propres et de leurs valeurs propres associées : $\mathcal{V} = \{(\phi_1, \lambda_1), (\phi_2, \lambda_2), \dots, (\phi_i, \lambda_i), \dots, (\phi_{\tilde{m}}, \lambda_{\tilde{m}})\}$ définis par l'équation suivante :

$$\begin{aligned} i \in [1 : \tilde{m}], \quad \mathbf{Q}\phi_i &= \lambda_i \phi_i \\ \text{ou, matriciellement,} \quad \mathbf{Q} &= \mathbf{\Phi \Lambda \Phi^t} \end{aligned}$$

avec $\mathbf{\Phi}$ la matrice $L \times \tilde{m}$ composée des vecteurs ϕ_i et $\mathbf{\Lambda} o$ la matrice diagonale composée des valeurs propres λ_i. $\tilde{m}$ est le nombre de dimensions de l'espace propre, il est donné par : $\tilde{m} = \min(L - 1, N - 1)$. Les valeurs propres sont triées par ordre décroissant. Elles représentent les variances sur chacune des dimensions définies par les vecteurs propres.

– *Sélection d'un sous-espace de représentation* : la sélection des vecteurs propres correspondant aux plus grandes valeurs propres permet de représenter la majeure partie de la variance des données sur un sous-espace. Il est possible, de plus, de choisir le nombre de dimensions comme une fonction de la variance conservée par la projection. Le choix d'une conservation de $X\%$ de la variance donne le nombre de dimensions suivant l'équation suivante :

$$\text{soit } m \text{ le plus petit entier tel que } \frac{\sum_1^m \lambda_i}{\sum_1^{\tilde{m}} \lambda_i} > X\%$$

Le choix des m premiers vecteurs ϕ_i définit un espace de représentation $\mathcal{A}$ maintenant $X\%$ de la variance des données de la base d'apprentissage.

– *Apprentissage des images*: chaque image est projetée sur l'espace $\mathcal{A}$, le vecteur de mesures $\mathcal{M}$ obtenu est stocké en association avec un identificateur Id de l'image dans une base de modèles. Cette base doit permettre un accès rapide aux plus proches voisins d'une nouvelle projection pour effectuer la reconnaissance. Le chapitre 5 et l'annexe B présentent une structure de données et les algorithmes associés permettant ce type d'accès. L'apprentissage proposée dans ce paragraphe modélise la variété de l'apparence des objets par une collection de points extraits de cette variété. Il est possible d'effectuer une interpolation entre ces points de façon à représenter la variété elle-même comme MURASE le propose.

Reconnaissance : l'objectif consiste à identifier un objet ainsi que sa pose dans une nouvelle image. Pour cela, il est nécessaire de normaliser cette image puis de la projeter sur l'espace $\mathcal{A}$ en obtenant ainsi un vecteur Ω de représentation. La recherche des vecteurs $\mathcal{M}$ proches de Ω dans la base de modèles permet d'obtenir une reconnaissance associée avec une valeur de confiance : la distance entre Ω et $\mathcal{M}$. Une distance faible correspond à une forte valeur de corrélation entre l'image correspondante de la base et la nouvelle image. En effet, le calcul de la distance euclidienne entre deux vecteurs Ω_1 et Ω_2 de $\mathcal{A}$ est une approximation de la distance entre les images J_1 et J_2 d'origine.

Un schéma de la phase de reconnaissance est présentée sur la figure 2.3. MURASE et NAYAR obtiennent des taux de reconnaissance très élevés sur des bases de l'ordre de 1000 images. Cela prouve la faisabilité d'un système de reconnaissance temps-réel basé sur l'apparence.

2.1.4 Fiabilité d'un système de reconnaissance fondé une modélisation par des images

La technique présentée est fondée sur l'appariement entre images. Néanmoins cet appariement est difficile car il n'est pas possible d'échantillonner la sphère des vues suivant l'ensemble des perturbations envisageables. Il est, en particulier, difficile d'inclure les cas d'occultation partielle d'objets dans l'apprentissage. Généralement, ces techniques sont découpées en deux phases : segmentation et normalisation des images pour réduire l'espace des images envisageables puis modélisation dans l'espace de description. Ces méthodes sont intrinsèquement globales car l'objet à reconnaître doit être nettement segmenté et séparé de son fond pour lui permettre d'être correctement normalisé puis finalement reconnu. Ce processus coûteux est souvent difficile voir impossible à effectuer. Dans un environnement complexe, un objet apparaîtra fréquemment partiellement occulté et ne sera pas reconnaissable. Des techniques statistiques de projection basées sur le principe de Description de Longueur Minimale (ou MDL) peuvent être utilisées si aucune normalisation n'est nécessaire et si l'image est très faiblement occultée [LBE97]. Mais cette technique reste limitée et ne peux permettre une reconnaissance robuste sur de grandes

bases d'objets. La section suivante propose de reconnaître non plus une image complète d'un objet mais des sous-images de cet objet aussi petites que possible. Ainsi, lors de la phase de reconnaissance, il est fortement probable que plusieurs sous-images soient intégralement visibles et permettent ainsi la reconnaissance des objets.

2.2 Modélisation par caractéristiques locales

Une image est modélisée par un ensemble de vecteurs de mesures de caractéristiques locales obtenus à différentes positions. Chacun des vecteurs est évalué à partir d'une portion de l'image ou imagette. Plus les imagettes sont petites, plus la représentation est locale et donc robuste à l'occultation partielle ou aux modifications du fond de l'image. La reconnaissance est obtenue par l'appariement de ces vecteurs de mesures représentant les imagettes. L'utilisation de critères de cohérence spatiale entre les positions d'origine des vecteurs appariés permet de lever les ambiguïtés liées à la présence d'imagettes semblables entre des objets différents.

Plusieurs approches ont été employées pour effectuer une modélisation d'objets par des caractéristiques locales pour la reconnaissance. Trois classes d'approches sont proposées ici :

1. Reconnaissance fondée sur l'appariement de caractéristiques locales.
2. Reconnaissance fondée sur l'appariement de graphes de caractéristiques locales.
3. Reconnaissance fondée sur des statistiques sur les caractéristiques locales.

2.2.1 Reconnaissance par appariement de caractéristiques locales

Cette approche nécessite de définir un espace de représentation des caractéristiques locales noté $\mathcal{A}$. Les images sont représentées par un ensemble de points de l'espace $\mathcal{A}$. L'algorithme de modélisation employé est le suivant :

- Projection d'un ensemble de points sur un espace de descripteurs locaux $\mathcal{A}$. Le choix des points dépend de critères locaux par l'utilisation d'un détecteur de points d'intérêts ou globaux pour sélectionner des points de mesures aussi différentes les unes des autres que possible.
- Apprentissage de ces points associés à des identificateurs dans une base de modèles ou apprentissage simultané des points projetés et des relations spatiales entre eux.

La reconnaissance est fondée sur l'appariement de vecteurs de mesures obtenus sur une nouvelle image avec ceux disponibles dans la base de modèles. Une technique de vote ou

de transformée de Hough peut être employée directement pour sélectionner l'objet le plus vraisemblable.

Une première base de représentation du signal est une extension directe des approches globales par "images propres". Elle est obtenue en décomposant les images d'un objet en imagettes aussi petites que possible de façon à garantir que, dans toute nouvelle image de l'objet, une partie des imagettes sera présente et donc reconnaissable. Cette approche est développée à la section 3.3. IKEUCHI [OI96, OI97] nomme les descripteurs obtenus par l'application de la technique ACP sur ces imagettes des "fenêtres propres" [3]. Une partie des imagettes de chaque image modèle sont sélectionnées puis projetées sur un sous-espace de l'espace propre. IKEUCHI propose un double critère de sélection des fenêtres fondé sur leur qualité locale (détecteur de TOMASI et KANADE [TK91]) et sur leur qualité globale (unicité). Cet aspect est repris à la section 5.1.1. Le système donne des résultats intéressants de reconnaissance sur des bases de quelques objets. Les images de test sont constituées par plusieurs objets simultanés et montrent la robustesse de la technique à l'occultation partielle. La robustesse aux variations en orientation des objets est obtenue par apprentissage des objets sur de multiples points de vues. Seule la translation est possible entre un objet et son modèle. Néanmoins, la perte de la normalisation globale en échelle rend la technique sensible aux variations de l'échelle entre les images modèles et les images de test.

Les vecteurs de mesures obtenus par un apprentissage par ACP sont sensibles aux variations de l'éclairage. Selon KRUMM [Kru97], il est possible de limiter cette sensibilité en utilisant des caractéristiques binaires. Des imagettes carrées binarisées sont extraites des images modèles puis utilisées comme clés dans un dictionnaire[4]. Les expérimentations utilisent un objet unique qui est reconnu parmi d'autres objets relativement similaires. Il s'agit d'un objet 2D dont seuls les paramètres d'orientation et de translation sont variables. La robustesse à l'orientation est obtenue par l'apprentissage simultané de 360 points de vue. Les clés binaires proposées présentent une robustesse à l'orientation inférieure à 1°. Ces résultats pour un seul objet ne sont pas très convainquants et encouragent à ne pas utiliser de données binaires comme caractéristiques locales stables. De plus, la robustesse à l'échelle apparaît, similairement à l'approche par fenêtres propres, difficile à obtenir.

De façon à supprimer la phase d'apprentissage inhérente à l'ACP, il est possible d'utiliser des base de descripteurs Gaussiens. Ces bases présentent des propriétés de paramétrage

3. Eigen Windows

4. Le terme *dictionnaire* est utilisé au sens algorithmique. Il s'agit d'une table dont les éléments sont des couples (clé, identificateur). Cette table permet de retrouver un identificateur correspondant à une clé exacte. Une implémentation classique d'un dictionnaire est une table de *Hash Code*.

en échelle et en orientation très appropriées pour accéder à une représentation invariante à ces deux paramètres et donc moins coûteuse en mémoire. L'objectif consiste à choisir une base de descripteurs qui approche aussi précisément que possible les imagettes observées par un nombre de descripteurs aussi faible que possible. L'approximation est d'autant plus adéquate qu'elle différencie au mieux les imagettes observées. Une approximation classique d'une fonction en un point $A = (x_A, y_A)$ est l'approximation de TAYLOR pour une fonction $J(x, y)$ donnée par la formule suivante :

$$J(x,y) = J(x_A,y_A) + (x - x_A)\frac{\partial J(x_A,y_A)}{\partial x} + (y - y_A)\frac{\partial J(x_A,y_A)}{\partial y} + \ldots + O(x^n, y^n)$$

L'ensemble des dérivées jusqu'à l'ordre n forme un vecteur représentant le signal autour du point A. Ce vecteur est appelé "Jet Local" par KOENDERINK [KvD87] et permet une représentation concise du signal. Les dérivées peuvent être calculées de façon stables en utilisant des descripteurs dérivées de Gaussiennes. Nous étudions cette approche plus précisément à la section 3.4.1.

RAO [RB95] propose de représenter un objet par un ensemble de vecteurs de mesures locales fondées sur des descripteurs *dérivées de Gaussiennes*. Le vecteur proposé est constitué de 45 dimensions: chacunes d'entre elles est définie par une dérivée de Gaussienne. L'utilisation de 5 échelles et 9 dérivées par échelle donne les 45 dimensions. Les 9 dérivées sont réparties sur les ordre 1, 2 et 3. L'utilisation des propriétés d'orientabilité des dérivées de Gaussiennes permet à ces vecteurs d'être invariants à l'orientation 2D. La modélisation est fondée sur l'enregistrement de ces vecteurs dans une base de modèles puis la reconnaissance est effectuée en recherchant des vecteurs similaires à ceux observés dans une nouvelle image. Le grand nombre de dimensions rend cette recherche très discriminante. Les faux appariements sont très peu fréquents.

Les expériences présentées évaluent la reconnaissance en fonction du nombre d'échelles utilisées et non pas en fonction de la valeur de l'échelle utilisée. Les résultats sur la taille des imagette présentés à la section 3.3.3 montrent qu'en absence d'occultation partielle le taux de reconnaissance est d'autant plus élevé que les imagettes sont grandes. Ainsi, il est probable que la discrimination obtenue entre les objets soit principalement due aux imagettes les plus grandes.

SCHMID et MOHR [Sch96] proposent un système de reconnaissance d'objets dans lequel l'apprentissage est effectué suivant trois phases successives : extraction de points d'intérêt par un détecteur de HARRIS , puis caractérisation de ces points par des vecteurs de mesures locales et stockage de ces vecteurs dans une base de modèles. Les vecteurs de mesures utilisés sont des invariants différentiels fondés sur des dérivées de Gaussiennes proposés par KOENDERINK [KvD84]. (voir section 3.4.2). La reconnaissance sur des images inconnues reprend les deux premières phases puis effectue un appariement des vecteurs de mesures similaires. La reconnaissance est obtenue via un algorithme de vote. Cet algorithme s'avère suffisant dans de nombreux cas d'autant plus que pour des bases

plus difficiles (nombreux objets similaires), l'utilisation de critères de cohérence spatiale diminue fortement le nombre de faux appariements.

Néanmoins, la sélection de points d'intérêts est le maillon faible de la méthode, il est difficile d'obtenir un détecteur réellement stable par rapport aux différentes conditions d'enregistrement. Il est, en particulier, difficile de définir un détecteur de points d'intérêts robuste aux variations de l'échelle. Ces approches sont fondées sur l'appariement de vecteurs de mesures locales avec, pour certaines, une augmentation de la discrimination par l'utilisation de critères de cohérence spatiale. Les approches présentées dans la section suivante fondent leur reconnaissance sur des graphes des caractéristiques locales souvent peu discriminantes en elles-mêmes.

2.2.2 Reconnaissance par appariement de graphes de caractéristiques locales

Cette section présente une stratégie de reconnaissance par l'appariement simultané de caractéristiques locales et de graphes de ces caractéristiques locales. La discrimination est principalement obtenue par les graphes de caractéristiques et non par les caractéristiques elles-mêmes. Les appariements de graphes présentent une difficulté combinatoire importante.

CAMPS [CHK97] propose une décomposition des objets en éléments simples. La définition proposée d'un élément simple est fondée sur l'algorithme de segmentation : les éléments simples sont des surfaces polynomiales approximativement fermées, non recouvrantes qui forment une partition optimale de l'image suivant un principe de description de longueur minimale ou MDL[5]. L'apparence de chacun des éléments est apprise en utilisant la technique de MURASE et NAYAR de reconnaissance par images propres. Cette approche remédie à l'inconvénient de l'approche globale de MURASE en optant pour une segmentation en éléments simples des images. Les éléments simples étant appris sous une échelle canonique, leur reconnaissance est indépendante de l'échelle des objets dans l'image. Une image est représentée par un graphe des relations géométriques entre ses éléments simples. L'appariement d'un nouvelle image est obtenu par la reconnaissance simultanée des éléments simples et du graphes des relations entre ces éléments.

La segmentation en éléments simples n'est pas un processus accessible dans le cas général et constitue la faiblesse de cette approche. La reconnaissance de ces éléments nécessite, en effet, leur segmentation préalable correcte pour permettre leur normalisation en échelle.

NELSON et SELINGER [NS98] utilisent des caractéristiques locales fondées sur des contours pour obtenir un système de reconnaissance robuste. Leur technique consiste à appliquer un détecteur de contours sur les images d'apprentissage puis à sélectionner les plus longs contours comme caractéristiques principales. L'adjonction à ces caractéris-

5. Minimum Description Length

tiques principales des contours les intersectant à l'intérieur d'une fenêtre 21×21 permet d'obtenir des indices visuels locaux et robustes aux variations d'éclairage, d'échelle et d'orientation. Le regroupement des indices de bas niveaux (des contours) permet d'obtenir des caractéristiques très discriminantes. Une base de 24 objets 3D assez simples est utilisée pour valider ces descripteurs dans le cadre d'un système de reconnaissance. La reconnaissance apparaît largement fondée sur les contours extérieurs des objets et la majorité de ces contours apparaissent dans les images de test même en présence d'occultation partielle ou d'un fond non uniforme. Le choix de caractéristiques fondées sur des contours paraît assez limitatif pour la gamme d'objets modélisables : des objets fortement texturés risquent, par exemple, de noyer la technique dans un trop grand nombre de contours peu informatifs.

Des caractéristiques locales très simples peuvent être utilisées en mettant l'accent sur la forte discrimination donnée par le graphe des relations géométriques entres ces caractéristiques. Ainsi, FLEURET [JF96] a défini 32 classes d'imagettes 5×5 qui sont déterminées par leur topographie (uniforme, bord, coin, texture). Ceci permet d'associer à chaque pixel d'une image un code dans l'intervalle [1 : 32]. La limitation de ce code à 32 valeurs ne lui permet d'être pas discriminant en lui-même. Par contre, la modélisation des images comme un graphe de relations entre chacun des codes donne la discrimination à cette technique. Une stratégie d'apprentissage automatique des images est utilisée et génère pour chaque image plusieurs graphes de représentation dans une base de modèles. La reconnaissance est obtenue par l'appariement d'un graphe extrait d'une image avec l'un des graphes de la base des modèles. Cet appariement est obtenu grâce à la structuration de la base des modèles en un arbre de décision. Les résultats donnés par l'auteur sont limités à des bases de quelques images et ne semblent pas extensibles à de plus grandes bases pour des raisons de complexité.

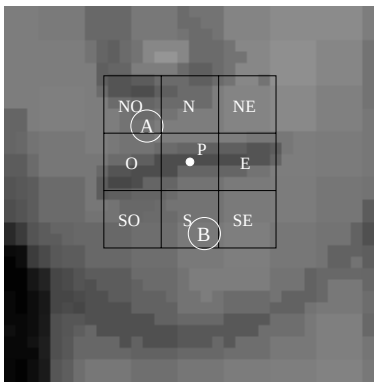


FIG. 2.4 – *Exemple de graphe de détection de la classe "Mouth" Le point P est caractérisé par la présence d'une caractéristique A dans la case MO (pour Nord-ouest) et par l'absence d'une caractéristique B dans la case S (pour Sud).*

GUARDA [GLL98] a étendu la représentation par graphes de caractéristiques à une représentation par programmes sur un ensemble de caractéristiques. Cette extension permet d'accéder à la classification d'objets. Ces programmes sont obtenus par un apprentissage génétique sur des classes d'objets comme les éléments du visages (bouche, yeux, nez). Les caractéristiques elles-mêmes ne sont pas définies à l'avance mais extraites des images par apprentissage. Ces programmes peuvent être symbolisés par des graphes centrés sur les éléments à reconnaître puis étiquetés par des opérateurs logiques qui permettent de modéliser des relations complexes entre caractéristiques locales comme la caractéristique B est présente au sud du point P et la caractéristique A n'est pas présente au nord-ouest de P. Plus précisément, cette relation sera représentée par l'expression :

$$Mouth = (AND(NOT(existe\_A\_dans\_zone\_NO?))(existe\_B\_dans\_zone\_S?))$$

Cette expression est visualisée sur la figure 2.4. Il est important de noter que la notion de présence d'une caractéristique dans une zone est spatialement assez floue et permet d'accepter des variations assez importantes en position.

La caractéristique importante de ces dernières approches est l'inclusion d'informations spatiales dans la modélisation par caractéristiques locales d'objets ou de classes d'objets. Ces approches motivent l'utilisation de critères spatiaux pour une grande discrimination entre objets. L'explosion combinatoire de certaines approches est souvent liée à l'utilisation de caractéristiques locales peu discriminantes comme pour FLEURET qui n'utilise que 32 classes d'indices visuels. Le chapitre 6 montre que l'ajout d'information sur les relations spatiales entre caractéristiques locales *discriminantes* améliore de façon conséquente la robustesse de la technique de reconnaissance proposée sans entraîner une explosion combinatoire.

2.2.3 Reconnaissance par évaluation statistique de caractéristiques locales

Une stratégie orthogonale qui évite l'appariement de caractéristiques peut être utilisée. Elle se fonde sur la comparaison des distributions de caractéristiques locales.

Une technique de reconnaissance fondée sur des statistiques de couleurs a été introduite par SWAIN et BALLARD [SB91] en 1991. Cette technique consiste à représenter un objet par un histogramme tridimensionnel des couleurs présentes dans une image de cet objet. Cette information est apparue, en l'absence de variations de l'éclairage, comme suffisamment discriminante pour effectuer une reconnaissance d'objets. Cette approche a été étendue à la représentation d'une image par des histogrammes multidimensionnels de champs réceptifs (ou caractéristiques locales) par SCHIELE [SC96] qui a obtenu un système de reconnaissance d'objets sans appariements de points très efficace. L'interprétation des histogrammes en terme de probabilités d'apparition des caractéristiques locales liée à l'utilisation de la règle de BAYES permet à cette technique une grande robustesse

à diverses perturbations des images comme l'occultation partielle. Cette approche est développée à la section 5.1.2.

2.3 Motivations pour la définition d'une nouvelle stratégie de reconnaissance

Les stratégies de reconnaissance proposées dans les sections précédentes fournissent des motivations sur ce qu'un système de reconnaissance fondé sur l'apparence doit contenir :

- La modélisation d'un objet par une collection d'images représentant ses différentes apparences est une stratégie très efficace en reconnaissance. Les images de cette collection forment un échantillonnage de la sphère des vues de l'objet.

- La modélisation par caractéristiques locales des images est robuste par rapport à de nombreuses perturbations du signal comme le bruit ou l'occultation partielle. La suppression des étapes de segmentation et normalisation augmente la robustesse.

- Une modélisation de la structure spatiale des images est très discriminante pour la reconnaissance. Elle permet d'utiliser plusieurs caractéristiques simultanées de façon optimale.

- La représentation des caractéristiques locales de l'ensemble d'une image comme un histogramme multidimensionnel démontre l'intérêt d'une modélisation complète de l'image et non d'une modélisation par extraction de points caractéristiques.

La stratégie de la reconnaissance proposée dans cette thèse se fonde sur cet ensemble de leçons. Elle utilise une modélisation de la structure spatiale des images par des caractéristiques locales et de leur relations. Cette modélisation sera appelée *Modélisation structurelle par caractéristiques locales*. Elle peut être résumée suivant trois axes :

- Sélection d'une base de descripteurs locaux qui définissent l'espace de représentation. Ces descripteurs permettent de mesurer des caractéristiques locales présentes sur les images. Chaque image peut être représentée dans l'espace $\mathcal{A}$ défini par cette base de descripteurs locaux. Un point de $\mathcal{A}$ (appelé *vecteur de mesures sur les caractéristiques locales*) est associé à chaque point (ou voisinage) d'une image. Cette étude est effectuée dans le chapitre 3. Le chapitre 4 montre une évaluation de la stabilité des vecteurs de mesures par rapport à différentes perturbations des images.

- Modélisation des images modèles en associant à chaque point d'une image un point de $\mathcal{A}$. Une image est, ainsi, représentée par une grille 2D dans l'espace $\mathcal{A}$. Cette phase de construction d'une base de modèles est appelée *phase d'apprentissage*. Cet aspect est étudié dans le chapitre 5.

- Comme pour les techniques de reconnaissance à base de caractéristiques locales présentées, la reconnaissance est fondée sur l'appariement entre des vecteurs de mesures des images modèles et ceux des images de test. La connaissance des relations géométriques entre les vecteurs de mesures permet d'augmenter la discrimination. Une stratégie directe de reconnaissance par vote ou transformée de Hough est proposée puis une stratégie fondée sur le paradigme prédiction-vérification est étudiée : elle consiste à effectuer une reconnaissance en deux étapes. La première étape consiste à générer des hypothèses vraisemblables d'objets par l'appariement de vecteurs de caractéristiques les plus discriminants puis la deuxième étape est une confirmation des hypothèses précédentes en évaluant si les points voisins des vecteurs précédents confirment ou réfutent les hypothèses générées. Ces aspects sont évalués dans les chapitres 6 et 7.

Chapitre 3

Caractéristiques locales

Reconnaître nécessite de sélectionner des caractéristiques de description. Cette sélection dépend du type d'objets à reconnaître, des conditions de prise de vues et de l'objectif de la reconnaissance. Cette thèse se concentre sur l'utilisation de caractéristiques génériques utilisables pour une large gamme d'applications. Les difficultés liées à l'utilisation de caractéristiques globales en reconnaissance motivent l'utilisation de caractéristiques locales pour obtenir un système de reconnaissance robuste. Ainsi, ce chapitre propose l'évaluation de différentes bases de caractéristiques locales suivant deux critères principaux : la stabilité des mesures de ces caractéristiques locales par rapport à une large gamme de perturbations et la dispersion de ces mesures dans l'espace des caractéristiques. L'évaluation de deux classes de caractéristiques locales est proposée :

1. des bases de filtres obtenues par une Analyse en Composantes Principales sur une décomposition des images d'apprentissage en imagettes,

2. et des bases fondées sur des dérivées de Gaussiennes.

La propriété de stabilité ou répétabilité des mesures peut se définir comme le fait qu'un même point physique d'un objet visible sur deux images différentes soit représenté par des mesures identiques. Ces mesures sont dites *invariantes* par rapport à une classe de transformation si leur valeur théorique ne varie pas à l'intérieur de cette classe, par exemple, la classe des similitudes 2D. L'invariance peut être obtenue sous l'hypothèse de la connaissance du (ou des) paramètres de la transformation. Il s'agit, dans ce cas, d'*équivariance* par rapport à ce (ou ces) paramètres. Les mesures sont dites *robustes* par rapport à une classe de transformations si elles varient faiblement pour une transformation de faible amplitude à l'intérieur de la classe. Cette propriété permet la mise en correspondance de points entre images par l'appariement des vecteurs de mesures.

La propriété de dispersion des mesures est fondée sur une propriété intuitive : deux vecteurs de mesures correspondants à deux imagettes visuellement différentes doivent être différents. La dispersion dans l'espace de représentation doit être aussi importante que

possible pour permettre de distinguer aussi nettement que possible les imagettes différentes. Cette propriété est difficile à mesurer car elle demande d'évaluer la répartition des vecteurs de mesures indépendamment de la répartition des imagettes dans l'espace image. Son évaluation peut être obtenue en mesurant et comparant la qualité de la reconnaissance sur une base de test pour plusieurs bases de descripteurs.

Ce chapitre propose une évaluation systématique de plusieurs bases de descripteurs locaux permettant de décrire l'apparence d'objets 3D du monde réel. La première section introduit les objectifs et les propriétés attendues pour une base de descripteurs locaux pour la reconnaissance, puis la deuxième section propose une étude de filtres fondés sur une Analyse en Composantes Principales d'imagettes. Mais leurs limitations motivent, dans la section suivante, l'évaluation de plusieurs bases de filtres fondées sur l'usage des dérivées de Gaussiennes et de leur propriétés d'équivariance à l'orientation et à l'échelle. La conclusion aborde une comparaison des ces différentes bases de filtres et entraîne le choix de la base de Dérivées de Gaussiennes ajustables en orientation et en échelle pour le cas général où les objets observés ne sont pas contraints.

3.1 Descripteurs Locaux

Les modélisations globales d'objets possèdent des limitations importantes telles que la segmentation, la normalisation ou la sensibilité à l'occultation partielle. Cette thèse propose de favoriser l'utilisation de caractéristiques aussi locales que possibles afin de minimiser autant que possible ces difficultés.

Principes de la technique de reconnaissance proposée Un point comme représentant de son voisinage est projeté sur un espace de caractéristiques noté $\mathcal{A}$. Cette projection associe à tout point d'une image un vecteur de mesures $\mathcal{M}$ de m coordonnées. Puis, la phase d'apprentissage consiste à enregistrer tous les couples (points, vecteurs) que l'on veut pouvoir retrouver et la phase de reconnaissance consiste à projeter un nouveau point ce qui donne un nouveau vecteur puis à retrouver les vecteurs similaires appris et à retourner les points correspondants associés à des valeurs de confiance. Ce chapitre se limite à cette première phase de la reconnaissance, l'utilisation de plusieurs points simultanément est évaluée au chapitre 6.

Les différentes bases de filtres sont évaluées en mesurant le taux de reconnaissance obtenu en utilisant un seul vecteur de mesures par recherche. Plus précisément une requête de recherche d'un vecteur de mesures dans une base de modèles peut aboutir à quatre types de réponses :

- Le système retourne un point issu d'un modèle correct comme réponse la plus probable. La distance entre les deux vecteurs de mesures est la plus faible obtenue sur l'ensemble des vecteurs retourné (cas [a] *succès*).

- Le système retourne un point issu d'un modèle correct parmi ses réponses. La distance entre les vecteurs de mesures est inférieure au seuil de recherche mais d'autres vecteurs sont plus proches et sont donc retournés avant (cas [b] *succès partiel*).
- De nombreux vecteurs similaires au vecteur recherché sont trouvés par la recherche. Le voisinage considéré comme non discriminant est rejeté. Le système ne propose pas un mauvais appariement et il ne s'agit donc pas d'un cas d'échec. Cette catégorie regroupe aussi les cas où la recherche ne donne aucun résultats. (cas [c] *rejet*).
- Des résultats incorrects sont obtenus uniquement : il s'agit du seul cas d'échec de la technique : reconnaissance incorrecte (cas [d] *échec*).

L'évaluation globale de chaque base de filtres est effectuée indépendamment de l'algorithme de reconnaissance complet proposé dans les chapitres suivants.

3.1.1 Définitions

Un *descripteur local* est un opérateur dont le support spatial est faible par rapport à la taille de l'objet (moins de 5% de la taille de l'image). Cet ordre de grandeur sur la taille d'un opérateur par rapport à la taille de l'image est valide pour une échelle fixe. La reconnaissance à échelles multiples implique des dimensions plus importantes après l'application d'un zoom à l'image. Son application en un point d'une image ou imagette permet d'obtenir une mesure sur le voisinage de ce point.

Les dimensions du support d'un descripteur sont définies par le paramètre σ définissant l'échelle des opérateurs fondés sur des Gaussiennes. Pour un filtre Gaussien, un rayon de 3σ autour du point considéré doit être pris en compte pour une bonne approximation du filtre (99% de l'énergie). Cela donne un filtre de taille $6\sigma \times 6\sigma$. L'intérêt de cette localité consiste à n'utiliser qu'une faible partie de l'image pour évaluer les vecteurs de mesures et ainsi obtenir une approche aussi locale que possible. La figure 3.1 est une illustration

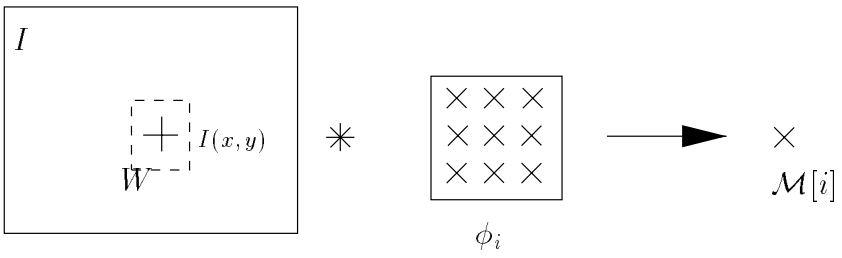


FIG. 3.1 – *Illustration de l'application du filtre ϕ_i sur le point A.*

de l'application d'un opérateur local ϕ_i symbolisé par un masque de convolution sur un voisinage W correspondant au point $J(x, y)$. Le résultat de cette application est un scalaire $\mathcal{M}[i]$. L'indice i numérote le filtre parmi une série de filtres.

Le scalaire obtenu par l'application d'un descripteur local est appelé *mesure* de l'imagette. L'application d'un ensemble d'opérateurs locaux permet d'obtenir un vecteur de mesures. Ce vecteur est la seule représentation de l'imagette dans notre représentation. Le choix des filtres et de leur nombre permettra d'obtenir une représentation plus ou moins précise de l'espace des imagettes. Ce choix définit un sous-espace de l'espace des imagettes dont l'ensemble des filtres est une base. Globalement, chaque filtre décrit une dimension du sous-espace. Ce sous-espace de l'apparence des imagettes est nommé l'espace de description $\mathcal{A}$ dans la suite.

Le choix d'un sous-espace pour représenter les imagettes implique de définir une distance dans ce sous-espace pour évaluer la similitude entre les projections issues d'imagettes différentes. Plusieurs distances sont évaluées : la distance euclidienne et la distance de Mahalanobis [Kan95] qui permet de tenir compte des covariances entre les dimensions. L'évaluation de la similarité entre les imagettes est, dans ce cas, plus précise (voir section 4.1.1). Deux types d'espaces de description sont abordés dans la suite.

3.1.2 Apprendre des filtres ou utiliser des filtres analytiques Gaussiens

Nous avons poursuivi deux approches. Dans un premier temps, par extension des travaux de MURASE , nous avons généré une base de filtres en optimisant un critère statistique sur les imagettes de la base d'apprentissage. L'Analyse en Composantes Principales correspond à cette approche en maximisant la variance sur les dimensions successives. Une extension de l'Analyse en Composantes Principales aux ordres supérieurs à 2 appelé l'Analyse en Composantes Indépendantes (ou ACI) [LAC97] permet d'obtenir une base de descripteurs plus indépendants, donc moins redondants. Nous avons aussi étudiée une base de filtres composée de dérivées de Gaussiennes. Cette base permettent un choix précis des caractéristiques que l'on veut mesurer. Ces descripteurs locaux décomposent le signal suivant des bandes de fréquences et réagissent suivant des orientations particulières du signal 2D. D'autres bases de filtres comme celles proposées par KOENDERINK [KvD84] maintiennent des propriétés particulières comme l'invariance à la rotation. Pour ces différentes approches, le choix d'une description du signal par une base de filtres pose le problème de l'évaluation de la base de filtres pour l'objectif de reconnaissance.

3.2 Évaluation des descripteurs locaux

Ce paragraphe propose de décrire les critères d'évaluation possibles d'une base de filtres locaux pour leur utilisation en reconnaissance. Une base de filtres est optimale si elle permet de discriminer de grandes quantités d'imagettes en utilisant une description aussi concise que possible de chacune des imagettes. Une description est dite *concise* si la quantité d'informations nécessaire à son stockage est faible. Cette quantité est le produit

du nombre de dimensions de la base par la quantité d'informations représentable dans chacune des dimensions ou quantification de la dimension.

Indépendamment du problème de quantification des dimensions, deux critères principaux peuvent permettre d'évaluer une base de filtres :

- la stabilité de la base de filtres par rapport aux variations de l'environnement d'observation.

- la dispersion des vecteurs de description des imagettes sur l'espace décrit par la base de filtres.

Ces deux critères sont abordés indépendamment dans les paragraphes suivants. Puis, la conjonction de ces deux critères permet de prédire la qualité de la reconnaissance induite par l'usage de ces filtres. Cette reconnaissance peut être évaluée en terme de discriminabilité d'une base de descripteurs.

3.2.1 Stabilité des mesures

Un descripteur est une opérateur local défini par un filtre de convolution ou par une fonction analytique. Le résultat de son application en un point d'une image est un scalaire appelé mesure. La stabilité des descripteurs permettant de décrire une imagette est un critère primordial d'évaluation d'une base de filtres. En effet, il apparaît important que deux évaluations d'une même imagette avec de faibles variations des conditions d'observation procurent des vecteurs de mesures proches. Le terme de répétabilité de l'évaluation peut être utilisé.

Deux types de variations peuvent être distingués : D'une part, l'évaluation des mesures est bruitée par la chaîne d'acquisition de l'image (objectif de la caméra imparfait, bruit de numérisation) et par les opérateurs locaux appliqués aux images qui ne sont pas idéaux (anisotropie, effet de bords liés au repliement de spectre). D'autre part, les paramètres d'observation comme l'éclairage ou le point de vue changent et influent sur les images observées. L'objectif est d'utiliser une base de descripteurs qui minimise les variations des mesures par rapport aux différentes perturbations des images. L'évaluation de cette stabilité des mesures est primordiale à leurs utilisation en reconnaissance et son étude expérimentale fait l'objet du chapitre 4 sur la sensibilité des descripteurs locaux.

Parallèlement à l'étude de la stabilité des descripteurs, il est nécessaire d'évaluer dans quelle mesure une base de descripteurs permet de différencier correctement des imagettes distinctes. L'exemple simple d'une mesure identiquement nulle montre que la stabilité n'est pas un critère suffisant d'évaluation d'une base de descripteurs. Le paragraphe suivant évalue cette propriété que nous appelons dispersion des données dans l'espace de description des imagettes.

3.2.2 Dispersion des données

L'espace de description des imagettes doit permettre de différencier les vecteurs de mesures correspondant à des imagettes différentes. Cette différenciation des vecteurs dépend de l'espace de description. Un premier paramètre de cet espace est son nombre de dimensions. Plus ce nombre est important, plus les points sont dispersés : la distance entre deux vecteurs correspondants à des imagettes visuellement différentes augmente avec le nombre de dimensions. Cette amélioration de la dispersion avec le nombre de dimensions nécessite un choix correct des opérateurs définissant chaque dimension. Il est, en particulier, important de choisir un ensemble de descripteurs linéairement indépendants et formant donc une base de l'espace de description. L'augmentation du nombre de dimensions représente un compromis entre la concision et la précision de la représentation.

L'utilisation d'un espace de description composé d'une unique dimension permet de différencier au mieux k imagettes différentes. Ce nombre k peut être évalué en moyenne en évaluant la stabilité du descripteur par rapport aux variations des conditions d'observation. Expérimentalement, l'évaluation statistique de la distance entre mesures correspondants à la même imagette physique permet d'évaluer un seuil de similarité s sur la distance entre mesures. Ce seuil permet a posteriori d'évaluer la similarité entre deux imagettes. Pour une mesure évaluée sur l'intervalle $[0:1]$, le paramètre k vaut : $k = \frac{1}{s}$. Par extension, sur un espace de description à m dimensions en utilisant une distance fondée sur la norme L_∞, le nombre maximum de vecteurs différents est $k_m = k^m = \frac{1}{s^m}$. Soit, pour $s = 0.1$ et $m = 10$, $k_m = 10$ milliards. k_m est le nombre maximum de vecteurs différents représentables, ce nombre est supérieur à la mémoire disponible dans une station de travail et permet théoriquement de distinguer toutes les imagettes possibles mais, malheureusement, cet immense espace de description n'est pas intégralement accessible car les mesures sont réparties de façon non uniforme sur chaque dimension comme le montre la figure 3.2 qui montre une répartition irrégulière des 3 premières dimensions d'une base de descripteurs obtenue par ACP. Une répartition similaire s'observe sur d'autres bases de filtres. Cette distribution permet difficilement d'évaluer la dispersion apportée par une base de filtres car elle montre principalement la répartition irrégulière des imagettes de la base d'apprentissage. A échelle fixée, de nombreuses imagettes sont de niveau de gris constant. Les réponse des filtres sont alors les mêmes pour toutes ces imagettes d'où la présence de pics autour de la valeur 0. Une évaluation indépendante de la base d'apprentissage est difficile car elle implique de pouvoir évaluer si deux imagettes sont similaires. Cette similarité n'est pas évaluable par une simple corrélation car elle correspond à une notion intuitive qui permet de décider si deux imagettes correspondent à une même information. La difficulté de l'évaluation de cette répartition implique d'évaluer une base de descripteurs par la qualité de la reconnaissance induite par son utilisation.

Pour conclure, la répartition irrégulière encourage à l'utilisation d'un nombre de dimensions aussi grand que possible pour disperser le plus les vecteurs. Mais le coût mémoire limite ce nombre et implique un compromis. Le nombre de dimensions est aussi limité par

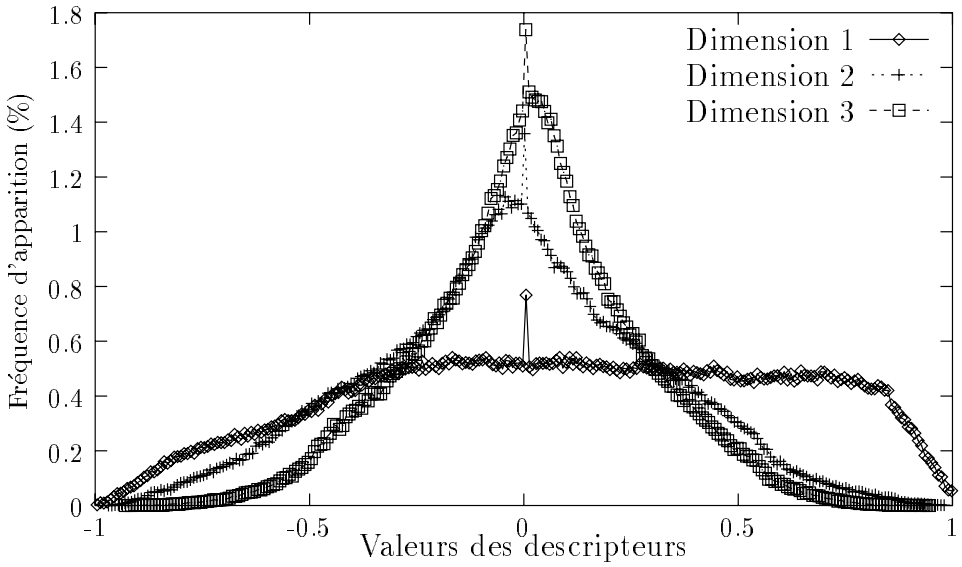


FIG. 3.2 – *Répartition des descripteurs sur les 3 premières dimensions d'un espace de filtres obtenu par* ACP *(voir section 3.3).*

le nombre de filtres stables indépendants qu'il est possible d'évaluer sur les imagettes. Pour des imagettes de petites tailles (9×9 par exemple) en niveau de gris, une dizaine de dimensions est au plus disponible. L'utilisation des trois canaux issus d'une représentation en couleur permet de tripler ce chiffre.

3.2.3 Discriminabilité et Reconnaissance

Le point clé consiste à évaluer la discrimination donnée par un filtre ou une base de filtres sur une base d'images. La difficulté de l'évaluation de ce critère provient de ce qu'il est lié à la base d'objets à indexer. Dans une base d'apprentissage ne contenant que des images très texturées une imagette constante sera très discriminante. De même, lorsque l'information couleur est prise en compte dans l'indexation, une fenêtre de couleur rouge sera d'autant plus discriminante qu'elle sera seule dans la base. Un compromis entre les propriétés de stabilité et de dispersion doit donner une discriminabilité optimale. L'évaluation de cette propriété ne peut être effectuée que par l'évaluation de la reconnaissance elle-même. Il s'agit de compter, pour une base d'apprentissage et une base de test, combien d'imagettes test sont reconnues, avec quelle précision et de plus, combien d'imagettes sont rejetées par l'algorithme de reconnaissance car non discriminantes. Ceci implique que l'évaluation d'une base de filtres ne peut être effectuée dans l'absolu mais doit être liée à une base d'apprentissage ainsi qu'à une base de test.

Ces résultats de reconnaissance sont une première étape dans l'objectif de reconnaissance d'objets, l'étape suivante consiste à faire coopérer des recherches de plusieurs fe-

nêtres pour obtenir une reconnaissance robuste : ceci est le sujet du chapitre 6. Il est important d'observer que l'identification d'une imagette détermine une hypothèse d'objet accompagnée d'une hypothèse de sa position précise par rapport au point de vue dans lequel l'imagette est reconnue. Cette position permet d'évaluer une similitude entre l'image et le modèle. Ceci est l'idée de base des processus de reconnaissance qui sont proposés dans le chapitre 6.

3.3 Descripteurs locaux obtenus par Analyse en Composantes Principales

Cette section présente une extension de la technique de reconnaissance par Analyse en Composantes Principales présentée dans la section 2.1.3 qui permet de définir une base de projection en fondant le choix des vecteurs sur un critère statistique : maximiser la variance conservée par l'extraction d'un sous-espace $\mathcal{A}$ fondé sur la sélection des vecteurs propres les plus informatifs. Cette technique permet de définir un sous-espace de représentation $\mathcal{A}$ des imagettes. Le critère statistique de la variance présente la propriété intéressante de négliger les faibles variations (peu d'énergie ou bruit) par rapport au reste. De plus, la dispersion des données est optimisée par cette technique dans la mesure où les vecteurs obtenus sont orthogonaux (décorellés).

3.3.1 Calcul des filtres ACP

Le principe de la technique de l'ACP est de trouver une nouvelle base de description des données. La sélection des dimensions les plus discriminantes de cette nouvelle base permet d'obtenir la base de représentation. Cette technique s'applique indifféremment sur des vecteurs de données quelconques et, en particulier, aussi bien sur des images de luminance que sur des images couleur. Le paragraphe suivant se propose de décrire le calcul de cette nouvelle base sur des imagettes :

- *Les données* : N imagettes W_k extraites d'un parcours exhaustif des N_J images J_k de la base d'apprentissage. L'objectif est de représenter chacunes des N imagettes de l'ensemble $\mathcal{W} = \{W_1, W_2, \ldots, W_N\}$ de façon minimale (quelques octets). La limitation des effets de bords (repliement de spectre lié au fenêtrage) nécessite à ce stade d'appliquer un masque Gaussien sur les fenêtres extraites. Les imagettes W_k de taille $M = w \times h$ sont interprétées comme des vecteurs. Pour les imagettes couleurs, le vecteur est constitué par la concaténation des vecteurs correspondants aux plans Rouge, Vert et Bleu, ce qui donne : $M = w \times h \times 3$. Une matrice $\mathbf{W}$ est définie comme la concaténation des vecteurs colonnes W_k. $\mathbf{W}$ est une matrice de dimensions $M \times N$:

$$\mathbf{W} = [W_1 W_2 \ldots W_k \ldots W_N] \tag{3.1}$$

– *Calcul des vecteurs propres et des valeurs propres* : la matrice est diagonalisée et un ensemble $\mathcal{V}$ de vecteurs propres et de leurs valeurs propres associées est obtenu : $\mathcal{V} = \{(\phi_1, \lambda_1), \ldots, (\phi_i, \lambda_i), \ldots (\phi_\mathcal{M}, \lambda_\mathcal{M})\}$ définis par l'équation suivante :

$$\begin{aligned} i \in [1 : \mathcal{M}], \quad \mathbf{Q}.\phi_i &= \lambda_i \phi_i \\ \text{ou, en termes matriciels,} \quad \mathbf{Q} &= \mathbf{\Phi \Lambda \Phi^t} \end{aligned}$$

avec $\mathbf{\Phi}$ la matrice $M \times m$ composée des vecteurs ϕ_i et $\mathbf{\Lambda}$ la matrice diagonale composée des valeurs propres λ_i. m le nombre de dimensions de l'espace propre, dans le cas des imagettes $M << N$ et m est maximum : $\mathcal{M} = M - 1$. La figure 3.3 montre les quarante premiers vecteurs propres de l'espace obtenu par apprentissage de toutes les imagettes de taille 9×9 d'une sélection des images (converties en images de luminance) de la base de Columbia A.1. Le choix de la taille 9×9 est issue d'un étude expérimentale (voir section 3.3.3).

L'aspect Gaussien des filtres obtenus s'explique par l'application d'un masque Gaussien préalablement à l'apprentissage pour limiter les effets de bords (repliement de spectre). Le premier vecteur obtenu effectue un filtrage gaussien qui ne permet pas la moindre discrimination entre imagettes. Dans le cas d'une normalisation par l'énergie, sa réponse est constante; il est donc écarté. Les dix dimensions suivantes sont sélectionnées pour former le sous-espace $\mathcal{A}$ de représentation. Le choix de dix dimensions est motivé par deux critères : la variance capturée par ces premières dimensions (98% sur cet exemple) et les hautes fréquences présentes dans les dimensions supérieures ne permettent pas d'envisager une stabilité suffisante des convolutions avec ces filtres. L'approche statistique de l'ACP a permis de définir une base orthogonale pour décrire le signal. Cette base est constituée de dix vecteurs décorrélés dont l'objectif est une indépendance maximale entre ces vecteurs. D'autres approches permettent d'évaluer des vecteurs indépendants à des ordres supérieurs à deux mais présentent des difficultés d'évaluation d'ordre pratique peu envisageables pour l'utilisation en reconnaissance d'objets proposée ici (voir Analyse en Composantes Indépendantes (ou ICA) [LAC97, FA99]).

A partir d'une quantité de données à apprendre très importante, il peut apparaître nécessaire d'utiliser plus de dimensions pour décrire ces données, l'usage de la couleur peut permettre d'augmenter ce nombre de dimensions utiles. La figure 3.4 montre le résultat du calcul ACP en conservant l'information couleur. Les vecteurs colonnes W_k sont ici constitués par la concaténation des plans Rouge, Vert et Bleu des imagettes. La figure fait apparaître 2 types de filtres :

– Des filtres en niveaux de gris similaires à ceux obtenus sans utiliser la couleur.

– Des filtres colorés correspondants à deux couples de couleur (rouge, bleu) et (vert, violet).

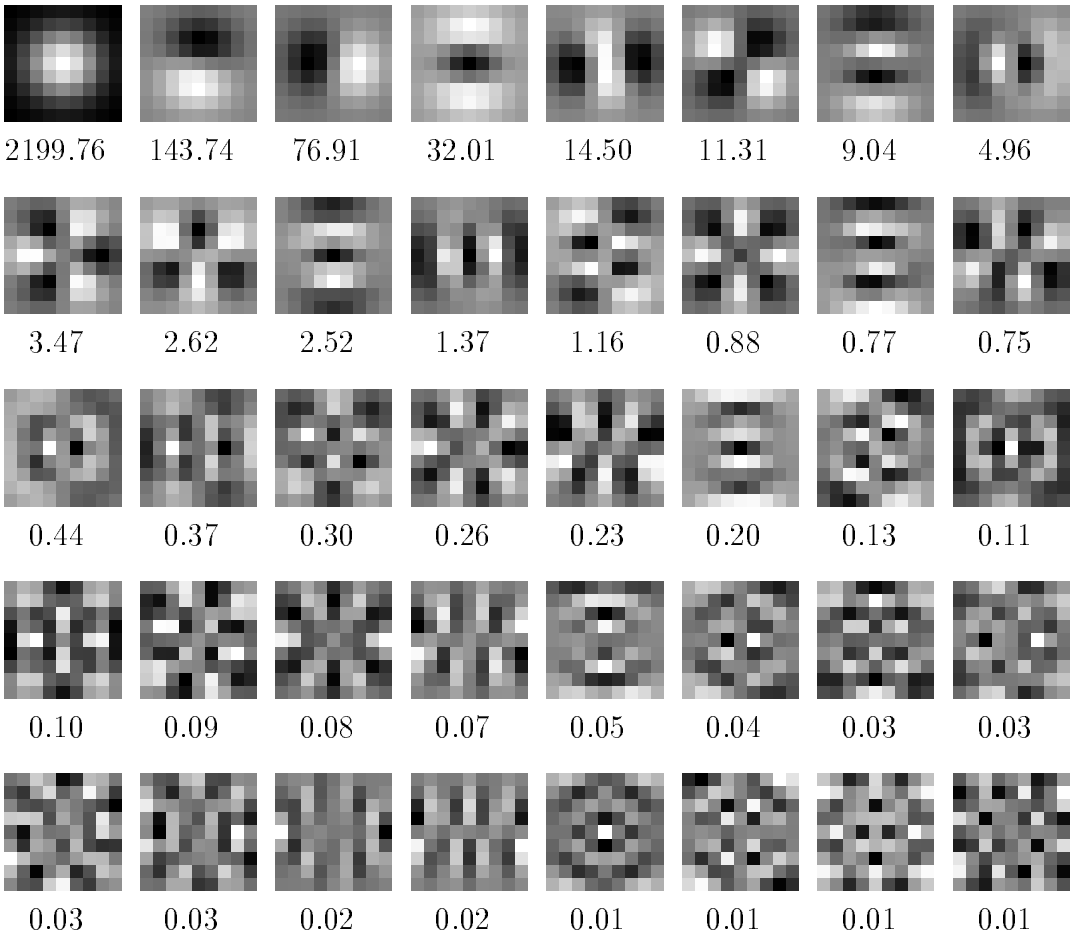


FIG. 3.3 – *Vecteurs Propres et Valeurs Propres associées.*

Les 4 premiers filtres évaluent l'intensité des différentes couleurs puis les suivants correspondent à des dérivées du signal sur un ou plusieurs plan de couleur. En pratique, la reconnaissance obtenue par l'utilisation de filtres couleurs est plus discriminante que la reconnaissance sur les filtres en niveaux de gris.

3.3.2 Quelques résultats

Ce paragraphe propose de montrer quelques résultats expérimentaux validant le choix d'un espace de filtres ACP pour la description locale d'images pour la reconnaissance d'objets.

Les résultats sont obtenus en utilisant une base extraite de la base de Columbia [NNM96b] (voir annexe A.1). La base d'apprentissage comprend, ici, 400 images soit environ 700.000 imagettes. Chaque image est décomposée en un ensemble d'imagettes 9×9 recouvrantes

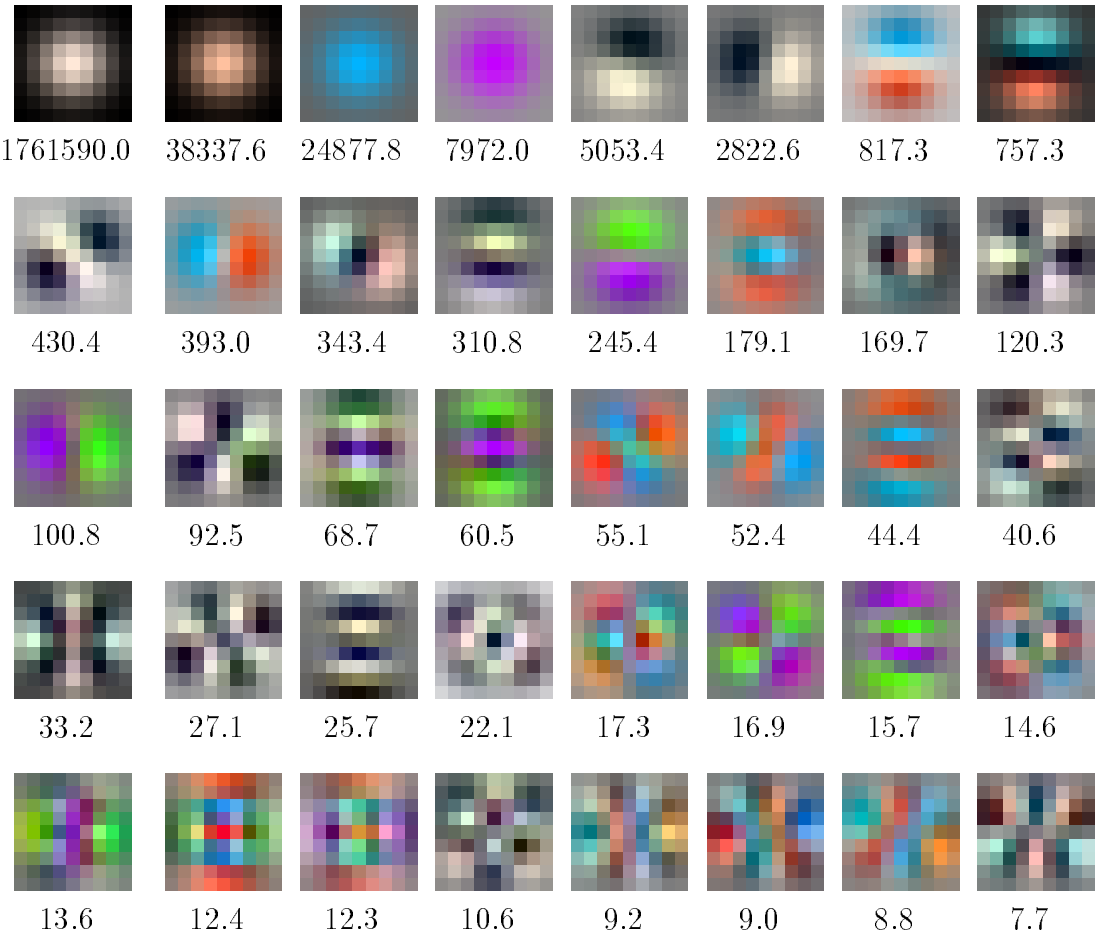


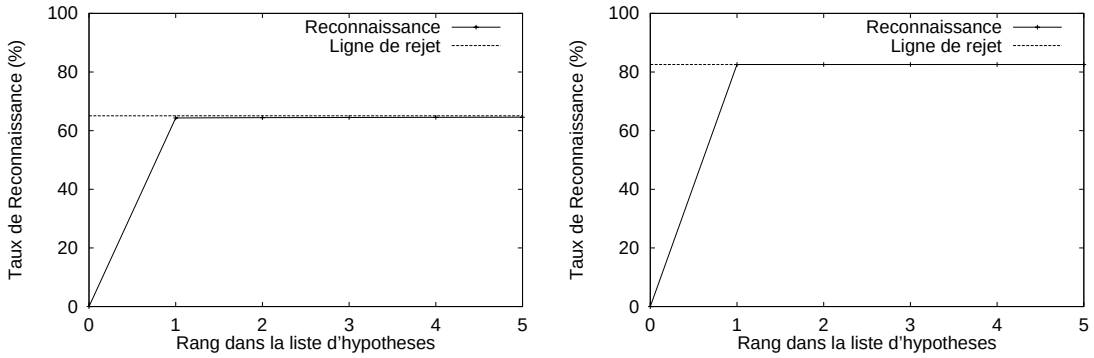
FIG. 3.4 – *Vecteurs Propres et Valeurs Propres associées.*

avec un pas de un pixel entre elles. Les images ont été utilisées avec une résolution moitié soit 64×64. Ces images correspondent à 4 points de vues pour 100 objets différents. Chacune de ces imagettes est projetée sur un sous-espace $\mathcal{A}$ de 10 dimensions. La phase d'évaluation comprends 2 parties : vérification sur les images d'apprentissage puis évaluation de la reconnaissance sur des images extérieures à la base d'apprentissage (points de vues proches) soit 1000 images de test.

L'évaluation de la reconnaissance est effectuée ici de façon directe : une requête de reconnaissance est effectuée sur l'intégralité des imagettes sans présélection (voir paragraphe 3.1). Les figures présentent les résultats correspondants à une base de filtres "niveaux de gris" et à une base de filtres "couleur". La connaissance préalable de la transformation approximative entre images modèles et images de test permet de valider ou rejeter un appariement.

La figure 3.5 présente les résultats de reconnaissance sur les images de la base d'appren-

tissage. Les courbes représente une évaluation statistique des scores de reconnaissance. La liste de résultats (correspond au cas [b]) est triée par ordre de distance croissante et le rang de reconnaissance dans la liste des hypothèse est visualisé sur l'axe des abscisses tandis que le taux de reconnaissance ou de rejet est visualisé sur l'axe des ordonnées. Les

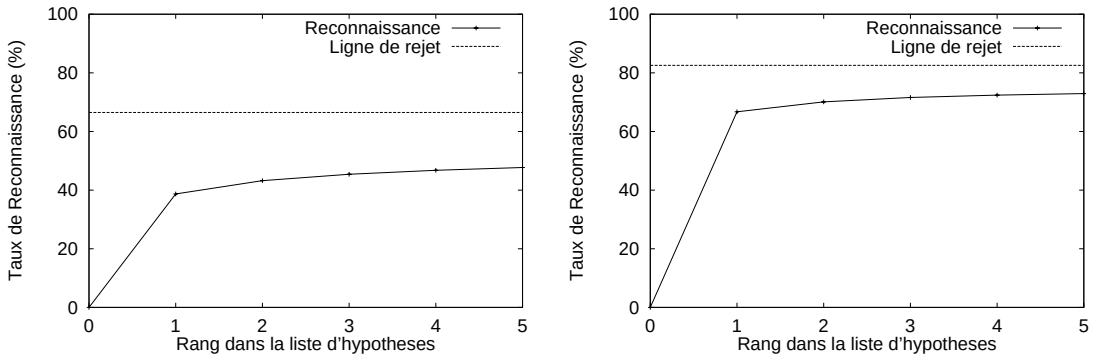


(i) Images de luminance (ii) Images de chrominance

FIG. 3.5 – *Évaluation de la reconnaissance par une imagette issue de la base d'apprentissage. Les graphes présentent le pourcentage de points reconnus aux rangs 1 et 5. 20 à 30 % d'imagettes au dessus de la "ligne de rejet" correspondent au cas [c] (trop d'imagettes similaires).*

deux courbes (i) et (ii) montrent qu'il n'y a dans aucun cas d'échec ou de reconnaissance à un rang supérieur à 1 (la distance est ici toujours nulle). Dans certains rares cas, le rang apparent est supérieur à 1 mais cela correspond à plusieurs imagettes identiques dans la base d'apprentissage. La reconnaissance est donc parfaite sur la base d'apprentissage. Par contre, les courbes montrent que l'utilisation de la couleur diminue beaucoup le taux de rejet car les imagettes sont mieux distribuées dans l'espace de description $\mathcal{A}$.

La figure 3.6 montre les courbes de reconnaissance sur la base de 1000 images de test. Le taux de reconnaissance est de 40% des imagettes reconnues directement et 45% à un rang inférieur à 5 pour les images "niveaux de gris". Pour les images couleurs, le taux de reconnaissance directe est de 65% et 70% sous un rang dans la liste d'hypothèses inférieur à 5. Le taux de rejet apparaît assez important et peut être limité en utlisant deux startégies : élimination de la redondance dans la base (voir section 5.3.2) et sélection pendant la phase de reconnaissance des imagettes les plus discriminantes (voir section 6.2). Le taux d'échec est faible dans les deux cas : 25% des points pour les imagettes en niveaux de gris et 10% dans le cas Couleur. Ainsi, dans le cas de l'utilisation de plusieurs imagettes simultanées (chapitre 6), une reconnaissance d'objets efficace est possible.



(i) Images de luminance (ii) Images de chrominance

FIG. 3.6 – *Évaluation de la reconnaissance par une imagette issue de la base de test. Les graphes présentent le pourcentage de points reconnus en fonction du rang dans la liste des imagettes reconnues. Les graphes correspondent à une base niveau de gris et une base couleur qui donne de meilleurs résultats.*

3.3.3 Évaluation de la taille des filtres

Le paramètre principal de la description par filtres locaux obtenus par ACP est la dimension de ces filtres. Cette dimension doit être évaluée suivant deux critères : la reconnaissance obtenue comme fonction de la taille des filtres et la localité des filtres. Cette localité est primordiale pour obtenir une robustesse à l'occultation partielle ou au changement du fond des images. L'information sur un objet est d'autant plus importante que les imagettes sont grandes et ainsi, la reconnaissance doit, théoriquement, sur une base de test sans occultation, croître avec le paramètre de dimension.

Une évaluation expérimentale sur la figure 3.7 permet de vérifier l'hypothèse d'amélioration de la reconnaissance avec la taille des filtres. Cette évaluation est effectuée sur une base extraite de la base de Columbia. La base utilisée contient moins d'objets que la base utilisée à la section précédente ce qui explique de meilleurs résultats pour la taille 9×9. Sur cette figure, la reconnaissance est évaluée en fonction de la taille des filtres. Le nombre d'imagettes rejetées diminue avec l'augmentation de la taille des filtres. En effet, le nombre d'imagettes identiques est de plus en plus faible. Les taux de reconnaissance en première réponse et dans la liste des réponses augmentent fortement avec la dimension de l'imagette. Le taux d'erreurs pour des imagettes supérieures à 19×19 devient inférieur à 1%. Pour ces dimensions, une grande partie de l'image est présente dans chaque imagette et la reconnaissance devient presque globale et le taux de reconnaissance obtenu est similaire à celui obtenu par MURASE en utilisant les images complètes. Le taux de reconnaissance chute rapidement pour des imagettes de dimensions inférieures à 9×9 et, ainsi, cette taille a été utilisée principalement pour les expériences avec filtres ACP. La reconnaissance reste dans ce cas suffisamment locale et ainsi robuste à l'occultation

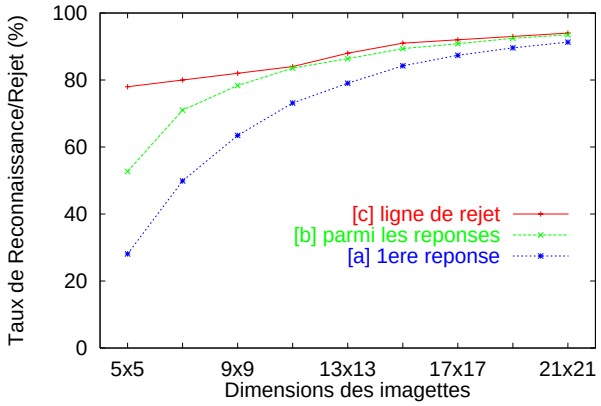


FIG. 3.7 – *Évaluation de l'influence de la taille des fenêtres sur la reconnaissance par une imagette quelconque de la base de test. La reconnaissance augmente avec la taille de l'imagette mais la localité de l'approche diminue avec l'augmentation de la taille. Un compromis peut être choisi autour de* 10×10 .

partielle. Le résultat donnant une augmentation de la reconnaissance avec la tailles des filtres ACP s'étend directement aux autres bases de filtres présentées dans la section 3.4. Dans ce cas, le paramètre de dimension n'est plus la taille de l'imagette mais le paramètre d'échelle σ de l'enveloppe Gaussienne.

3.3.4 Sensibilité à l'orientation 2D

La base de descripteurs obtenu par ACP ne présente pas de robustesse théorique aux paramètres d'échelle et d'orientation. Ce paragraphe évalue, comme cas d'étude, le comportement par rapport à l'orientation. Intuitivement, à l'œil et mise à part les bords, l'information contenue dans deux images d'un même objet différents uniquement par l'orientation de l'objet dans ces images est conservée. Il paraît souhaitable qu'un système automatique de reconnaissance soit capable d'effectuer la reconnaissance indépendamment de ce paramètre. Une étude expérimentale est effectuée sur une base d'images pour lesquelles seul le paramètre d'orientation autour de l'axe optique de la caméra est modifié continuement. La base d'objets utilisée est un ensemble de huit objets 2D photographiés sous un orientation variable. Cette base est présentée à l'annexe A.2. Une image par objet est utilisée pour l'apprentissage puis les autres images sont utilisées pour le test. Le graphe de la figure 3.8 montre une chute de la reconnaissance dés que l'orientation diffère de plus de 10 degrés et implique donc pour une modélisation par descripteurs ACP l'apprentissage de 18 images par point de vue de façon à garantir que toute nouvelle orientation observée de l'objet soit à proximité d'une position apprise. Cette stabilité est plus importante que les 1 degrés obtenus avec des filtres binaires par KRUMM . Ceci peut s'expliquer par

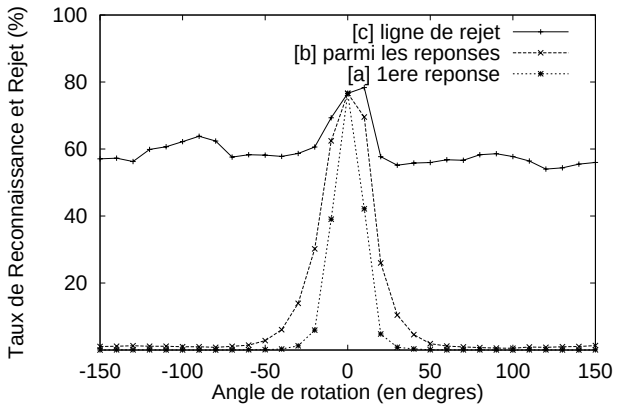


FIG. 3.8 – *Évaluation de la reconnaissance par filtre* ACP *en présence de rotations* 2D *sur des images de la base de test tournées par rapport aux images modèles. Le reconnaissance s'effondre dés que l'angle de rotation dépasse 10 degrés.*

les défauts inhérents aux filtre binaires et par le masque Gaussien appliqué à l'apprentissage. De façon identique, une étude du comportement par rapport aux variations d'échelle donne une robustesse de 10% et implique l'apprentissage d'une image modèle par tranche de 20% en échelle. Une reconnaissance robuste aux variations d'échelle et d'orientation apparaît donc très coûteuse et peu envisageable.

3.3.5 Conclusions

L'apprentissage de filtres par la technique statistique de l'ACP a donné de résultats expérimentaux satisfaisants. Les résultats de reconnaissance montrent un taux d'erreurs pour la reconnaissance par une imagette isolée inférieure à 20% en images de luminance et 10% en images couleurs. Ces résultats permettent d'envisager une reconnaissance d'objets par plusieurs imagettes très robuste.

Par ailleurs, cette technique a fait apparaître une similitude très importante entre les filtres obtenus et les filtres analytiques "Dérivées de Gaussiennes". Ce phénomène a déjà été observé par HANCOCK [HBS91] qui a extrait des composantes principales par un réseau de neurones sur des imagettes extraites d'images naturelles[1]. Il a conclu, en particulier, que, pour ce type d'images, ces filtres caractéristiques sont obtenus indépendamment de l'échelle. La base des dérivées de Gaussiennes peut être considérée comme une base de filtres canonique pour les images naturelles.

La base de descripteurs obtenue par Analyse en Composantes Principales est restreinte à la reconnaissance d'objets vu sous un point de vue similaire à l'un des points de vue

1. Les images naturelles sont des images issues d'objets de la nature comme des arbres ou des paysages. Elles sont appelées ainsi par opposition aux images d'environnements artificiels comme des batiments.

d'apprentissage. Sans un apprentissage à orientation et échelle variables, la reconnaissance ne peut être robuste à ces paramètres que de façon très limitée. La stabilité des projections sur l'espace de description donne une robustesse expérimentale de 10% en échelle et de 10 degrés en orientation. L'utilisation de filtres analytiques Gaussiens et de leur réglages en échelle et en orientation permet de résoudre cette limitation. Ainsi la section suivante présente des base de filtres analytiques Gaussiens avec, en particulier, la base des dérivées de Gaussiennes.

3.4 Descripteurs Gaussiens

La définition analytique d'une base de filtres Gaussiens présente de nombreux avantages, il est possible de contrôler précisément le contenu spatial et fréquentiel des filtres. De plus, l'enveloppe Gaussienne donne une robustesse importante au bruit additif. La synthèse de ces filtres théoriques est une difficulté importante, elle peut être obtenue de deux façons : d'une part, il est possible de générer des masques discrets représentant une approximation de support fini de ces filtres à support infini puis d'obtenir des mesures par des convolutions. D'autre part, ces filtres peuvent être évalués par une approximation récursive sans évaluation de masque, de façon très efficace. Le problème, dans ce cas, est la validation de ces filtres.

Une propriété importante est la possibilité de régler les paramètres d'échelle et d'orientation de ces filtres de façon obtenir une équivariance à ces paramètres puis une invariance grâce à l'utilisation d'un calage en échelle et orientation.

Filtres de Gabor : Les filtres de Gabor fournissent une base générale de description d'un signal image. Ces filtres présentent l'avantage de permettre le paramétrage indépendant de la fréquence et de la largeur de la bande du filtre. Néanmoins, plusieurs auteurs notent des résultats de reconnaissance semblables entre les bases de filtres dérivées de Gaussiennes et filtres de Gabor. La description obtenue apparaît similaire. SCHIELE [Sch97] a préféré l'usage des dérivées de Gaussiennes après avoir observé des résultats similaires. CHOMAT [Cho99] a obtenu des résultats comparables entre ces deux bases de filtres. Dans le cadre de cette étude, le paramétrage indépendant de la fréquence et de la largeur de bande ne présente pas d'avantage et l'étude proposée se limite aux dérivées de Gaussiennes comme cas d'étude. Les résultats peuvent s'étendre aux filtres de Gabor de façon directe.

Cette section présente d'abord la base de filtres *dérivées de Gaussiennes*. Le paragraphe 3.4.2 propose d'utiliser la théorie de FREEMAN sur les filtres orientables[2] pour adapter la base de filtres aux variations de l'orientation 2D. Puis, le paragraphe 3.4.3 étudie la propriété d'équivariance à l'échelle de ces filtres et propose une technique de sélection

2. En anglais : Steerable Filters

automatique de l'échelle locale fondée sur les travaux de LINDEBERG [Lin98]. Par la suite, le paragraphe 3.4.2 propose un traitement de la rotation 2D par l'usage d'invariants différentiels calculés à partir des Dérivées de Gaussiennes.

3.4.1 Base de filtres Dérivées de Gaussiennes

La fonction plénoptique présentée au chapitre 2 proposée par ADELSON et BERGEN permet de décrire tout ce qui est observable sur une scène, c'est-à-dire l'ensemble des apparences possibles de cette scène. Il est possible de représenter l'apparence d'une scène par un sous-échantillonage de cette fonction dont chacun des échantillons peut être analysé par ses dérivées successives puis utilisé comme mesures de l'image ou caractéristiques locales. Ainsi, KOENDERINK [KvD87] a proposé de décomposer le signal image en séries de TAYLOR . Le vecteur des termes successifs de cette décomposition (les dérivées à des ordres croissants) est appelé le *jet local*.

La décomposition de Taylor d'un signal continu à l'ordre n en un point $A = (x_A, y_A)$ est donnée par la formule suivante :

$$J(x,y) = J(x_A,y_A) + (x-x_A)\frac{\partial J(x_A,y_A)}{\partial x} + (y-y_A)\frac{\partial J(x_A,y_A)}{\partial y} + \ldots + O(x^n,y^n)$$

Le "jet local" d'ordre n au point A est le vecteur des dérivées jusqu'à l'ordre n au point A.

Sur un signal discret, il est possible de calculer les dérivées successives de ce signal en utilisant des opérateurs "Dérivées de Gaussiennes" comme alternative à la base ACP proposée précédemment.

Du point de vue traitement d'image, les dérivées de Gaussiennes sont largement utilisées et bien maîtrisées [FA91, Sch97]. Elles sont utilisées pour modéliser le cortex visuel humain (voir YOUNG [You85] par exemple). Elles présentent plusieurs propriétés majeures comme la possibilité de les calculer suivant une orientation et une échelle arbitraires. De plus, les filtres correspondants sont séparables et peuvent être synthétisés de façon très efficace par une implémentation récursive.

Le paragraphe suivant définit les opérateurs dérivées de Gaussiennes puis décrit leur propriété d'équivariance à l'orientation et à l'échelle. Pour décrire l'image indépendamment des paramètres d'orientation et d'échelle, les paragraphes suivants étudient des techniques de détection de l'orientation locale et de détection d'une échelle caractéristique locale.

Dérivées de Gaussiennes : Les descripteurs *dérivées de Gaussiennes* sont obtenus par la dérivation de la fonction gaussienne bidimensionnelle $G(x,y,\sigma)$ selon une direction Θ. La fonction gaussienne bidimensionnelle est définie par :

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\, e^{(-\frac{x^2+y^2}{2\sigma^2})} \tag{3.2}$$

La dérivée d'ordre n de $G(x,y,\sigma)$ suivant la direction Θ avec $\vec{v} = (\cos\Theta \sin\Theta)^T$ est définie par :

$$G^n_\Theta(x,y,\sigma) = \frac{\partial^n}{\partial \vec{v}^n} G(x,y,\sigma) \tag{3.3}$$

Les équations des filtres de dérivation théorique permettent d'évaluer les dérivées $L^n_\theta(x,y,\sigma)$ du signal image en évaluant la convolution de l'image par ce filtre. Cette évaluation est effectuée pour un paramètre d'échelle σ :

$$\begin{aligned} L^n_\theta(x,y,\sigma) &= G^n_\Theta(x,y,\sigma) * J(x,y) &(3.4)\\ &= \int_{x',y'} G^n_\Theta(x-x',y-y',\sigma) J(x',y') dx' dy' &(3.5)\end{aligned}$$

La dérivation suivant une orientation arbitraire ne se calcule pas directement de façon stable sur un signal image. Il est plus simple d'évaluer les dérivées de Gaussiennes correspondants aux axes de l'image. L'indice x correspond à l'angle $\Theta = 0$ et l'indice y correspond à l'angle $\Theta = \frac{\pi}{2}$. Les dérivées de Gaussiennes correspondantes sont données ici :

Les dérivées d'ordre 1 sont données par les équations :

$$\begin{aligned} G_x(x,y,\sigma) &= -\frac{x}{\sigma^2}\, G(x,y,\sigma) &(3.6)\\ G_y(x,y,\sigma) &= -\frac{y}{\sigma^2}\, G(x,y,\sigma)\end{aligned}$$

Les dérivées d'ordre 2 sont données par les équations :

$$\begin{aligned} G_{xx}(x,y,\sigma) &= \frac{x^2-\sigma^2}{\sigma^4}\, G(x,y,\sigma) &(3.7)\\ G_{xy}(x,y,\sigma) &= \frac{x\,y}{\sigma^4}\, G(x,y,\sigma)\\ G_{yy}(x,y,\sigma) &= \frac{y^2-\sigma^2}{\sigma^4}\, G(x,y,\sigma)\end{aligned}$$

Puis, les dérivées d'ordre 3 sont données par les équations :

$$\begin{aligned} G_{xxx}(x,y,\sigma) &= \frac{x\,(x^2 - 3\,\sigma^2)}{\sigma^6}\, G(x,y,\sigma) &(3.8)\\ G_{xxy}(x,y,\sigma) &= \frac{-y\,(x^2 - \sigma^2)}{\sigma^6}\, G(x,y,\sigma)\\ G_{xyy}(x,y,\sigma) &= \frac{-x\,(y^2 - \sigma^2)}{\sigma^6}\, G(x,y,\sigma)\\ G_{yyy}(x,y,\sigma) &= \frac{y\,(y^2 - 3\,\sigma^2)}{\sigma^6}\, G(x,y,\sigma) &(3.9)\\ & &(3.10)\end{aligned}$$

Les dérivées croisées permettent de calculer les dérivées de Gaussiennes suivant des orientations arbitraires (voir paragraphe 3.4.2). Les équations analytiques des filtres permettent d'obtenir les dérivées successives d'un signal image par la convolution de ce signal par les filtres. De plus, les filtres dérivées de Gaussiennes sont séparables et peuvent être programmés de façon récursive pour ou coût très faible: La complexité est, comme pour les convolutions, en $O(n)$ pour n le nombre de pixel dans l'image, mais, par contre, elle est indépendante de la taille des filtres. Plusieurs implémentations sont possibles comme celle de DERICHE [Der92] qui approche les dérivées par des polynômes d'ordre 4 ou celle de YOUNG et VAN VLIET [YvV95] utilisée dans cette thèse qui utilise une approximation d'ordre 3. Cette évaluation est rapide et de complexité indépendante de la valeur du paramètre σ contrairement à l'implémentation classique par un filtre de convolution. L'annexe C donne quelques détails sur l'implémentation récursive et les problèmes induits par cette implémentation.

3.4.2 Équivariance à l'orientation

Ce paragraphe présente la notion d'orientabilité d'un filtre et son application au cas des filtres Dérivées de Gaussiennes. Cette propriété permet d'obtenir à partir d'un nombre fini de convolutions, la valeur d'un filtre suivant une orientation arbitraire Θ en effectuant une combinaison linéaire entre les résultats des convolutions évaluées. Ceci permet d'adapter l'orientation d'un filtre à l'objet observé pour un coût négligeable. De plus, la sélection automatique de l'orientation fondée sur la direction du gradient permet un réglage des filtres et, par conséquent, une invariance à l'orientation.

La propriété d'orientabilité peut être visualisée sur l'exemple simple de la dérivée de Gaussienne d'ordre 1. En effet, $G^{\sigma}_{1,\Theta}$ est définie simplement à partir des filtres dérivées suivant les axes x et y :

$$G^1_\Theta(x,y,\sigma) = \cos\Theta\ G^1_0(x,y,\sigma) + \sin\Theta\ G^1_{\frac{\pi}{2}}(x,y,\sigma) \tag{3.11}$$

L'équation 3.11 permet de calculer le filtre suivant l'orientation voulue. Les fonctions $\cos\Theta$ et $\sin\Theta$ sont des fonctions d'interpolation sur les filtres de base correspondants aux axes x et y. Une propriété capitale de la convolution est sa linéarité: il est équivalent de convoluer une image avec $G^1_\Theta(x,y,\sigma)$ que d'évaluer les convolutions avec $G^1_0(x,y,\sigma)$ et $G^1_{\frac{\pi}{2}}(x,y,\sigma)$ puis d'appliquer les fonctions d'interpolation sur les résultats. Ainsi, pour une image J, les convolutions suivants les axes x et y sont évaluées :

$$\begin{aligned} L^1_0(\sigma) &= G^1_0(x,y,\sigma) * J \\ L^1_{\frac{\pi}{2}}(\sigma) &= G^1_{\frac{\pi}{2}}(x,y,\sigma) * J \end{aligned} \tag{3.12}$$

L'interpolation se fait alors à partir les images résultantes $L^1_0(\sigma)$ et $L^1_{\frac{\pi}{2}}(\sigma)$ pour un angle Θ quelconque :

$$L^1_\Theta(\sigma) = cos\Theta\ L^1_0(\sigma) + sin\Theta\ L^1_{\frac{\pi}{2}}(\sigma) \tag{3.13}$$

Cette propriété classique des dérivées Gaussiennes d'ordre 1 a été étendue, formellement, par FREEMAN et ADELSON [FA91] pour plusieurs gammes de filtres comme les dérivées de Gaussiennes d'ordre quelconque. Les paragraphes suivants retracent leurs résultats généraux puis les appliquent au cas des dérivées de Gaussiennes d'ordre 1 à 3.

Orientabilité d'un filtre Un filtre $f(x,y)$ est dit orientable[3] s'il peut s'écrire comme une combinaison linéaire de lui-même sous des différentes orientations. Le nombre d'orientations nécessaire est fini. Cela donne la contrainte d'orientabilité :

$$f^{\Theta}(x,y) = \sum_{j=1}^{l} k_j(\Theta) f^{\Theta_j}(x,y) \tag{3.14}$$

avec l le nombre de fonctions d'interpolation $k_j(\Theta)$ et $\{f^{\Theta_j}(x,y)/j \in [1:l]\}$ l'ensemble fini de fonctions f orientées suivant les angles Θ_j. Le point clé de cette approche consiste à évaluer le nombre l minimal de fonctions d'interpolation puis à déterminer ces fonctions d'interpolation.

FREEMAN a montré que, pour un filtre orientable, le nombre l minimal de fonctions de base est égal au nombre de coefficients non nuls dans une décomposition de FOURIER sous une représentation polaire. Par exemple, la première dérivée de Gaussienne s'écrit en coordonnées polaires, puis se décompose sur la base de Fourier :

$$\begin{aligned} G_0^1(r,\phi) &= -2re^{-r^2}cos(\phi) \\ &= -re^{-r^2}(e^{\imath\phi} + e^{-\imath\phi}) \end{aligned} \tag{3.15}$$

Cette décomposition admet deux coefficients non nuls et, par conséquent, deux fonctions d'interpolation sont suffisantes pour représenter ce filtre. Les fonctions d'interpolation sont évaluées en résolvant l'équation suivante :

$$(e^{\imath\Theta}) = (e^{\imath\Theta_1}\ e^{\imath\Theta_2}) \begin{pmatrix} k_1(\Theta) \\ k_2(\Theta) \end{pmatrix} \tag{3.16}$$

Des raisons de symétrie et de robustesse au bruit demandent de répartir au mieux les angles de base sur l'intervalle $[0:\pi[$. Ainsi, Θ_1 et Θ_2 sont choisis égaux à 0 et $\frac{\pi}{2}$. Dans ce cas, l'équation 3.16 se résout simplement par $k_1(\Theta) = \cos(\Theta)$ et $k_2(\Theta) = \sin(\Theta)$, ce qui redonne le résultat connu à l'ordre 1 donné sur l'équation 3.11 :

$$G_\Theta^1(x,y,\sigma) = \cos\Theta\ G_0^1(x,y,\sigma) + \sin\Theta\ G_{\frac{\pi}{2}}^1(x,y,\sigma) \tag{3.17}$$

La propriété d'orientabilité est vérifiée pour de nombreuses gammes de filtres et, en particulier, les filtres dérivées de Gaussiennes à tout ordre.

3. en anglais : Steerable

Une deuxième propriété a été étudiée par FREEMAN : l'évaluation du nombre de fonctions d'interpolation XY–séparables permettant d'évaluer un filtre orientable suivant une orientation arbitraire. Cela est possible pour certaines gammes de filtres et pour les dérivées de Gaussiennes en particulier. Ainsi, une dérivée à un ordre et une orientation arbitraire peut être obtenue à partir des résultats des convolutions avec les dérivées XY–séparables à l'ordre correspondant. Les équations 3.6, 3.7 et 3.8 donnent les formules de ces dérivées jusqu'à l'ordre 3. Les équations 3.18 définissent les fonctions d'interpolation permettant de calculer les dérivées aux ordres 1 à 3 sous une orientation Θ.

$$\begin{aligned} L^1_\Theta(x,y,\sigma) &= cos(\Theta)L_x(x,y,\sigma) + sin(\Theta)L_y(x,y,\sigma) \\ L^2_\Theta(x,y,\sigma) &= cos^2(\Theta)L_{xx}(x,y,\sigma) + 2cos(\Theta)sin(\Theta)L_{xy}(x,y,\sigma) + sin^2(\Theta)L_{yy}(x,y,\sigma) \\ L^3_\Theta(x,y,\sigma) &= cos^3(\Theta)L_{xxx}(x,y,\sigma) + 3cos^2(\Theta)sin(\Theta)L_{xxy}(x,y,\sigma) + \\ &\quad 3cos^2(\Theta)sin(\Theta)L_{xyy}(x,y,\sigma) + sin^3(\Theta)L_{yyy}(x,y,\sigma) \end{aligned} \tag{3.18}$$

Finalement, un espace $\mathcal{A}$ de description invariant à l'orientation peut être défini. Pour l'ordre n, $n+1$ descripteurs sont nécessaires et suffisants pour décrire complètement le signal. Ainsi, l'espace est défini par 9 descripteurs $\mathcal{M}[i]$. Le vecteur de mesures $\mathcal{M}$ est défini au point (x,y) pour le paramètre d'échelle σ par :

$$\mathcal{M} = [L^1_0\ L^1_{\frac{\pi}{2}}\ L^2_0\ L^2_{\frac{\pi}{3}}\ L^2_{\frac{2\pi}{3}}\ L^3_0\ L^3_{\frac{\pi}{4}}\ L^3_{\frac{\pi}{2}}\ L^3_{\frac{3\pi}{4}}]^T \tag{3.19}$$

Les orientations des dérivées ont été choisies pour une répartition optimale sur l'intervalle $[0 : \pi[$, soit $\Theta_i = \frac{i\pi}{n+1}$ à l'ordre n. En pratique, les angles Θ_i sont remplacés par $\Theta_i + \alpha$ où α est l'orientation de la scène ou de l'environnement local (voir paragraphe 3.4.2).

Les paragraphes suivants proposent de valider la propriété d'orientabilité expérimentalement puis étudient la détection de l'angle α de base à partir de l'évaluation de la direction du gradient dans l'image.

Validation expérimentale de l'orientabilité : Cette validation est fondée sur la connaissance préalable de l'orientation de la scène observée de façon à utiliser cette information pour vérifier que les vecteurs $\mathcal{M}$ réorientés sont effectivement semblables aux vecteurs correspondants appris sous une autre orientation. Cette propriété est appelée équivariance à l'orientation.

Une série de huit objets (scènes 2D) a été photographiée en faisant varier l'orientation autour de l'axe optique de la caméra. Une image par objet est utilisée comme modèle pour la phase d'apprentissage : chacunes des imagettes de cette image sont projetées sur l'espace $\mathcal{A}$ en utilisant les filtres de convolutions sous une orientation $\alpha = 0$. Puis, des imagettes sont extraites des images restantes pour tester la propriété d'orientabilité des filtres Dérivées de Gaussiennes.

Le taux de reconnaissance obtenu en fonction de l'orientation des images de test est évalué expérimentalement. De façon identique aux expériences utilisant la base de

filtres ACP, la reconnaissance est envisagée sur l'intégralité des imagettes de test possibles sans sélection d'imagettes informatives suivant un critère. La figure 3.9 présente un extrait d'une série d'images utilisée pour cette validation. Ces images proviennent de la base d'images MOVI [Gro98]. L'ensemble des images utilisées est disponible en annexe A.2.

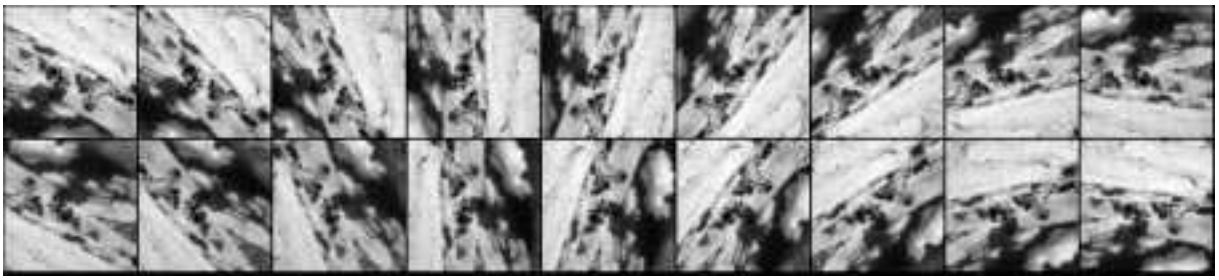


FIG. 3.9 – *Extrait d'une série d'images en rotation (série c2_vp_rz_s3 de la base d'images* MOVI*) utilisées pour évaluer la robustesse à l'orientation. Une image par objet est sélectionnée comme modèle puis les autres images permettent d'évaluer la reconnaissance par rapport à l'angle de rotation*

.

La figure 3.10 montre l'évolution de la reconnaissance en fonction du changement d'orientation avec l'image modèle. Le changement d'orientation est préalablement connu dans cette expérience. La figure présente 3 courbes qui distinguent l'espace des réponses

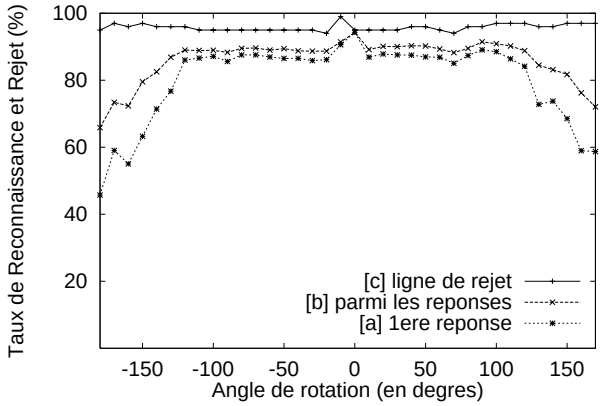


FIG. 3.10 – *Évaluation de la reconnaissance en fonction de l'angle de prise de vue connu. Les vecteurs de mesures sont calés suivant cet angle connu de façon à être similaires aux vecteurs appris sous l'orientation du modèle.*

en quatres classes [a] à [d] (voir paragraphe 3.1) : la première (cas [a]) montre le pourcentage de points aboutissant à la reconnaissance de l'objet et de sa pose comme première hypothèse. La seconde (cas [b]) montre les cas où la pose correcte est retrouvée mais pas

en première réponse. La troisième courbe distingue les classes d'échecs (cas [d]) et les cas de rejets pour les points non discriminants (cas [c]). La figure montre une dégradation de la reconnaissance lorsque l'angle de rotation est maximum (180 degrés d'écart avec les images d'apprentissage). Si l'application exige un taux de reconnaissance plus élevé l'apprentissage de deux images modèles peut être effectué et permettre ainsi une reconnaissance plus robuste. Néanmoins la reconnaissance est en moyenne très suffisante : 80% des points aboutissent à une reconnaissance directe. L'utilisation des filtres orientables apparaît donc très efficace et utilisable pour des taches de reconnaissance.

Ici en l'absence de détection d'une orientation caractéristique, la reconnaissance est fondée sur une connaissance préalable du changement d'orientation. L'appariement nécessite, en effet, que les orientations des caractéristiques correspondent entre les images modèles et les images de test. Cette correspondance entre les orientations peut être obtenue suivant deux stratégies :

- Apprentissage des objets sous des orientations multiples en supposant que toute nouvelle image des objets ressemblera à l'une des images apprises. Cette approche a été utilisé pour la détection de la pose d'un objet 2D unique par KRUMM [Kru97]. Sur cet exemple, les caractéristiques évaluées sont des vecteurs de caractéristiques binaires enregistrées dans un dictionnaire. Puis la reconnaissance se fait directement sur ce dictionnaire. Pour son système, KRUMM a eu besoin d'apprendre l'objet sous 360 orientations (une tous les degrés) ce qui ne paraît pas du tout généralisable pour de nombreux objets pour des raisons de coût mémoire important.

- L'approche duale consiste à effectuer un apprentissage suivant une orientation puis à reporter le problème du choix de l'orientation à la phase de reconnaissance qui se base sur des connaissances externes ou temporelles pour une réduction très importante de la gamme des orientations possibles. Ceci se rapproche du système de suivi de doigt de DEVIN [Dev98] qui utilise une fenêtre coulissante de trois masques de corrélation. Dans ce cas, l'orientation α est connue approximativement à chaque instant et il suffit de valider à chaque nouvelle image si l'objet observé a tourné depuis l'observation précédente en évaluant successivement les trois choix d'orientations possibles $\alpha - \epsilon$, α et $\alpha + \epsilon$ puis en sélectionnant la meilleure. Un tel système ne peut fonctionner que si la vitesse de rotation est faible par rapport à la fréquence d'acquisition des images. Plus précisément, il est indispensable que le changement d'orientation entre deux images soit inférieur au paramètre ϵ. La robustesse des descripteurs est expérimentalement de l'ordre de 10°, il faut donc fixer ϵ à 20° au plus pour que les vecteurs de mesures soient correctement évalués et permettent une mise en correspondance du motif observé avec son modèle. Cette stratégie peut être adaptée pour gérer les changements d'échelle d'un objet pendant un suivi.

Il est possible de résoudre le problème de la connaissance préalable de l'orientation, en détectant une orientation caractéristique en chaque point puis en l'utilisant pour l'ap-

prentissage et la reconnaissance. Le paragraphe suivant propose d'évaluer une technique de détection de l'orientation fondée sur l'évaluation de la direction du Gradient pour permettre une représentation invariante à l'orientation 2D.

Détection de l'orientation et Résultats : L'orientabilité de la base de filtres Dérivées de Gaussiennes est bénéfique si il est possible d'évaluer de façon stable une orientation caractéristique pour une majorité de points des images. Cette orientation peut être évaluée de façon globale sur l'image complète ou alors, localement, par évaluation de la direction du Gradient. Il faut ensuite orienter les filtres suivant cette orientation de base. Les vecteurs obtenus sont *invariants à l'orientation.*

La détection de la direction du gradient d'une fenêtre est faite en évaluant les dérivées premières (filtres G_0^1 et $G_{\frac{\pi}{2}}^1$). Le vecteur gradient $\vec{Grad} = \binom{L_0^1}{L_{\frac{\pi}{2}}^1}$ est obtenu. Puis, la direction α est obtenue en calculant l'arctangente de ces deux dérivées :

$$\alpha = \arctan 2(L_{\frac{\pi}{2}}^1, L_0^1) \tag{3.20}$$

Cette équation permet d'évaluer l'orientation α en tous points d'une image. Un vecteur de mesures $\mathcal{M}$ invariant à l'orientation 2D est obtenu en tournant toutes les coordonnées de $\mathcal{M}$ par cet angle α :

$$\begin{aligned} \mathcal{M} &= \mathcal{M}_\alpha \\ &= [L_\alpha^1 \; L_{\frac{\pi}{2}+\alpha}^1 \; L_\alpha^2 \; L_{\frac{\pi}{3}+\alpha}^2 \; L_{\frac{2\pi}{3}+\alpha}^2 \; L_\alpha^3 \; L_{\frac{\pi}{4}+\alpha}^3 \; L_{\frac{\pi}{2}+\alpha}^3 \; L_{\frac{3\pi}{4}+\alpha}^3]^T \end{aligned} \tag{3.21}$$

La première coordonnée L_α^1 donne l'amplitude du Gradient et la deuxième $L_{\frac{\pi}{2}+\alpha}^1$ se retrouve identiquement nulle et peut donc être supprimée. Le vecteur $\mathcal{M}$ est constitué de 8 coordonnées. Le point correspondant de l'image est modélisé par le couple $(\mathcal{M}, \alpha)$.

L'algorithme de projection d'une imagette aussi bien pendant la phase d'apprentissage que pendant la phase de reconnaissance est le suivant :

- Convolution de l'imagette avec les filtres dérivées de Gaussiennes XY–séparables (équations 3.6, 3.7, 3.8) soient 9 convolutions. Le vecteur $\mathcal{M}_o$ de 9 coordonnées obtenu dépend de l'orientation.
- Évaluation de la direction du Gradient par le calcul de l'arctangente entre les deux dérivées premières.
- Calcul du vecteur $\mathcal{M}$ sur la direction du Gradient en chaque point en utilisant les formules d'interpolation (équations 3.18). Le vecteur de mesures $\mathcal{M}$ est indépendant de l'orientation de l'image.

L'invariance obtenue ici par rapport aux variations de l'orientation n'implique pas la perte de l'orientation qui est conservée pour l'évaluation de la pose. La mise en correspondance

de deux imagettes est effectuée en utilisant les vecteurs invariants puis l'écart entre les orientations détectées donne l'angle de rotation approximatif entre les deux images.

La figure 3.11 montre une image (i) et l'orientation $\alpha(x, y)$ du gradient en chacun de ses points (ii). L'image (iii) présente une évaluation de la stabilité de cette orientation $\alpha(x, y)$ en un point par rapport à ces voisins. Il s'agit de la distance angulaire moyenne entre l'angle détecté en un point et l'angle détecté par ses voisins. Cette stabilité est obtenue par l'équation :

$$stab(x, y) = \alpha(x, y) - \frac{1}{4}(\alpha(x-1, y) + \alpha(x+1, y) + \alpha(x, y-1) + \alpha(x, y+1))$$

Une instabilité de l'angle détecté signifie que sur une nouvelle image du même objet, l'orientation ne serait probablement pas retrouvée.

Sur l'image (ii), le noir correspond à un angle 0 et le blanc correspond à l'angle 2π . La discontinuité du passage entre 0 et 2π n'est qu'apparente et n'est pas prise en compte dans l'évaluation de la stabilité de l'image (iii). L'image (iii) montre des instabilités sur

(i) Image originale

(ii) Images des orientations du gradient

(iii) Stabilité de l'orientation du gradient

FIG. 3.11 – *Une image, l'orientation du gradient détectée sur celle-ci et une évaluation de la stabilité de gradient en chaque point (blanc signifie instable et noir stable). Ces images illustrent le problème de l'instabilité de la direction du gradient en certains points.*

l'orientation détectée pouvant aller jusqu'à un maximum de $\frac{\pi}{2}$. Cette instabilité est très forte et confirme une impossibilité d'évaluation de l'orientation locale pour tous les points d'une image.

L'utilisation de la fonction *Arc-tangente* pour le calcul de l'angle implique une instabilité pour un *Gradient* très faible. Ceci peut être observé sur la figure 3.12 où la reconnaissance est évaluée en fonction de la valeur du Gradient. Il s'agit ici d'une base de 8 objets dont une image a été apprise en détectant l'angle du gradient pour chacune des imagettes. Les imagettes des images de test sont projetées sur l'espace $\mathcal{A}$ puis leur reconnaissance est évaluée. Les classes de reconnaissance [a] à [d] sont définies au paragraphe 3.1. La figure montre que la reconnaissance se dégrade pour un gradient faible et qu'à partir d'un gradient inférieur à 2, la proportion d'échecs et de rejets augmente beaucoup et implique de rejeter les vecteurs de projection ayant un gradient inférieur à 2.

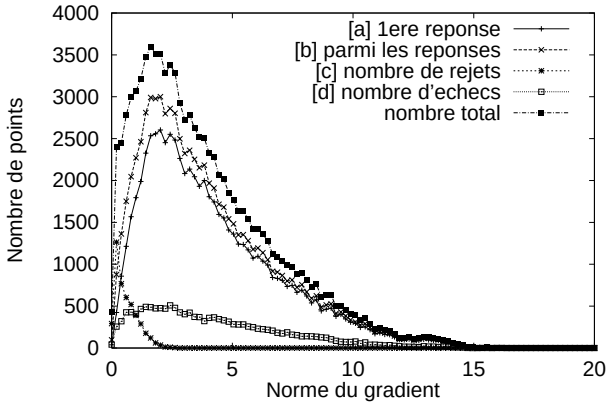


FIG. 3.12 – *Reconnaissance d'une imagette en fonction de la valeur du gradient. Ce graphe confirme la connaissance intuitive qu'un gradient faible implique une détection de sa direction imprécise et par conséquent une reconnaissance souvent incorrecte.*

Évaluation expérimentale de la reconnaissance : La figure 3.13 montre une évaluation expérimentale de la reconnaissance par un vecteur de mesures quelconque en présence de rotations 2D. Le jeu d'images utilisé est identique aux expériences présentées sur les figures 3.8 (p. 51) et 3.10 (p. 58) : huit objets de la base MOVI présentée en annexe A.2 dont une image par objet est apprise tandis que les autres sont utilisées pour valider la reconnaissance. En abscisses, l'orientation correspond à l'angle de rotation entre les images test et les images modèles. En ordonnées, le pourcentage de points donnant une reconnaissance ou un rejet est visualisé. Les classes [a] à [d] de résultats de reconnaissance présentées au paragraphe 3.1 sont utilisées. 25% des imagettes sont rejetées. Ceci peut s'expliquer par la présence de nombreuses imagettes de niveau de gris presque constant. Ces imagettes ne sont pas discriminantes et sont rejetées. La reconnaissance est parfaite pour les images d'apprentissage. Elle est de 40% des points en premier rang pour les autres images. Le taux d'échecs est de l'ordre de 5% ce qui est très faible. Des pics sont observés pour les orientations multiples de $\frac{\pi}{2}$, ceci peut s'expliquer par la représentation des images en tableaux 2D qui privilégie les directions des deux axes par rapport aux autres et les filtrages effectués sur ces images ne sont pas suffisamment isotropiques. L'annexe C donne quelques détails sur l'anisotropie du filtrage récursif.

Cette expérience démontre la validité des dérivées de Gaussiennes orientables pour la reconnaissance. Le taux de rejet est lié principalement à la présence de nombreuses imagettes de niveau de gris quasi-constant dans les images. Ce phénomène peut être corrigé en évitant d'apprendre les points correspondant à cette échelle. Il existe nécessairement une échelle plus importante pour laquelle ces points ne correspondent plus à des imagettes constantes et la sélection automatique présentée dans la section 3.4.3 peut permettre de limiter ce problème. L'utilisation d'invariants différentiels proposée dans le paragraphe

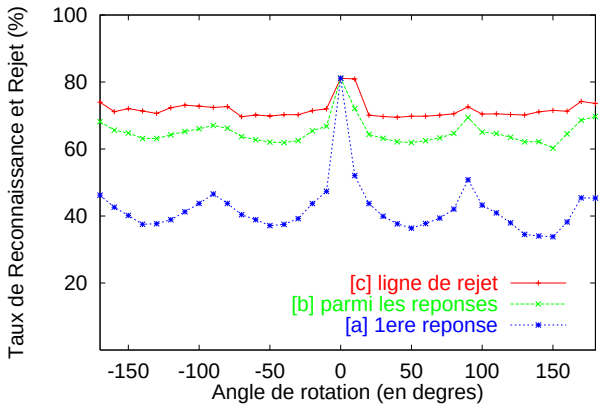


FIG. 3.13 – *Évaluation expérimentale de la reconnaissance en fonction de l'angle de vue avec détection automatique de l'orientation local en chaque point. Ce graphe montre 25% d'imagettes rejetées en moyenne et moins de 5% de faux appariements. Le rejet important s'explique par la présence de nombreuses imagettes de niveau de gris constant.*

suivant peut aussi permettre de résoudre cette difficulté.

Invariants différentiels : Une autre stratégie permet de résoudre le problème de l'orientation 2D. Cette stratégie est fondée sur l'utilisation de descripteurs invariants aux variations de l'orientation proposés par KOENDERINK [KvD84] : des invariants différentiels. Ceci a permis à plusieurs auteurs d'obtenir des taux de reconnaissance élevés comme SCHMID [Sch96] pour un système de reconnaissance d'images ou SCHIELE pour un système de reconnaissance à base d'histogrammes multidimensionnels. Ces invariants sont fondés sur l'évaluation de dérivées de Gaussiennes. Pour des raisons de stabilité numérique, les dérivées sont limitées à l'ordre 3 et permettent de définir huit invariants indépendants. La table suivante présente les quatres premiers invariants correspondants aux ordres 1 et 2:

$$\begin{bmatrix} {G_x}^2 + {G_y}^2 \\ G_{xx}{G_x}^2 + 2G_{xy}G_xG_y + G_{yy}{G_y}^2 \\ G_{xx} + G_{yy} \\ {G_{xx}}^2 + 2{G_{xy}}^2 + {G_{yy}}^2 \end{bmatrix} \tag{3.22}$$

Il est possible de reconnaître le premier comme étant la norme du gradient et le troisième comme le Laplacien. Les résultats expérimentaux attendus par cette base de filtres sont similaires à ceux obtenus par les filtres dérivées de Gaussiennes orientables mais, néanmoins, il est possible d'observer une différence importante entre ces deux bases, l'information angulaire mutuelle entre les vecteurs est perdue dans cette base mais conservée dans la base orientable. De plus, ces invariants théoriques sont obtenus en effectuant plusieurs multiplications sur les dérivées de Gaussiennes ce qui augmentent leur sensibilité

au bruit. Ainsi, cette base donne une discrimination théoriquement similaire à la base des dérivées de Gaussiennes mais son utilisation est rendue délicate par sa sensibilité au bruit. Dans le cadre de l'invariance à l'échelle présentée à la section suivante, le facteur de normalisation appliqué aux dérivées augmentent encore cette sensibilité et ne permet plus leur utilisation. La section suivante propose le réglage local du paramètre d'échelle qui permet de rendre les descripteurs robustes à l'échelle et, de plus, de corriger les problèmes de détection d'angle dans des fenêtres de Gradient faible en sélectionnant des échelles où le Gradient est suffisamment important.

3.4.3 Équivariance à l'échelle

Les dimensions dans l'image des objets observés varient avec la distance de la caméra avec l'objet ou suivant la focale de la caméra. Cette variation est une difficulté importante du problème de la reconnaissance d'objets : comment évaluer une mesure indépendamment de la dimension apparente dans l'image ? Les dérivées de Gaussiennes présentent la propriété d'être calculables à des échelles arbitraires (voir YOUNG [You85]). Sous la contrainte d'une normalisation adaptée comme celle proposée par LINDEBERG [Lin98], elles peuvent être calculées à différentes résolutions en conservant une valeur indépendante de la résolution choisie : cette propriété est appelée équivariance à l'échelle.

L'équivariance à l'échelle est similaire sur un signal 1D et sur un signal 2D. Ces propriétés sont, par souci de clarté, présentées sur un signal monodimensionnel. Soit un signal $J(x)$ et $P(\tilde{x})$ une version de ce même signal à une autre échelle. Analytiquement, ce changement d'échelle se décrit par un changement de variables :

$$sx = \tilde{x}$$

soit :

$$J(x) = P(\tilde{x})$$

Ce changement de variables se répercute sur les dérivations du signal :

$$\begin{aligned} J(x) &= P(sx) \\ \frac{\partial J(x)}{\partial x} &= s\frac{\partial P(sx)}{\partial x} \\ &\vdots \\ \frac{\partial^n J(x)}{\partial x^n} &= s^n\frac{\partial^n P(sx)}{\partial x^n} \end{aligned} \tag{3.23}$$

Ces équations montrent qu'il est possible de calculer les dérivées successives d'un signal $J(x)$ à partir des dérivées successives de $P(sx)$. Le paragraphe 3.4.1 a montré que les dérivées d'un signal discret J peuvent être calculées en utilisant des convolutions par des

dérivées de Gaussiennes. Dans ce cas, les dérivées de $J(x)$ et $P(\tilde{x})$ se calculent suivant les équations suivantes :

$$\begin{aligned} L^n(x,\sigma) = & \frac{\partial^n J(x)}{\partial x^n} = & G^n(x,\sigma) * J(x) \\ \tilde{L}^n(\tilde{x},\tilde{\sigma}) = & \frac{\partial^n P(\tilde{x})}{\partial \tilde{x}^n} = & G^n(\tilde{x},\tilde{\sigma}) * P(\tilde{x}) \end{aligned} \tag{3.24}$$

L'application des équations 3.23 et 3.24 donne pour $\tilde{\sigma} = s\sigma$:

$$\begin{aligned} L^n(x,\sigma) &= s^n \frac{\partial^n P(sx)}{\partial x^n} \\ &= s^n G^n(x,\sigma) * P(sx) \\ &= s^n G^n(x,s\sigma) * P(x) \\ &= s^n \tilde{L}^n(x,\tilde{\sigma}) \end{aligned} \tag{3.25}$$

Ceci montre qu'il est possible de calculer les dérivées successives de J à partir du signal P. Néanmoins, l'équation 3.25 fait apparaître un paramètre s : il s'agit du facteur d'échelle qui est lié à la taille du support du filtre de convolution. De façon générale, ce facteur n'est pas connu et l'invariance à l'échelle n'est pas accessible. Néanmoins, une normalisation adéquate permet de supprimer ce facteur d'échelle.

Une méthode de normalisation a été proposée par LINDEBERG pour obtenir une équivariance à l'échelle. Celle-ci consiste à multiplier chaque dérivée d'ordre n par σ^n pour obtenir une dérivée normalisée équivariante à l'échelle. Les dérivées normalisées ainsi obtenues seront notées $\mathcal{L}^n(x,\sigma)$:

$$\mathcal{L}^n(x,\sigma) = \sigma^n L^n(x,\sigma) \tag{3.26}$$

Cette équation permet d'obtenir l'équivariance à l'échelle :

$$\begin{aligned} \tilde{\mathcal{L}}^n(x,\tilde{\sigma}) &= \tilde{\sigma}^n \tilde{L}^n(x,\tilde{\sigma}) \\ &= (s\sigma)^n \tilde{L}^n(x,\tilde{\sigma}) \\ &= \sigma^n L^n(x,\sigma) \\ &= \mathcal{L}^n(x,\sigma) \end{aligned} \tag{3.27}$$

Nous avons donc :

$$\tilde{\mathcal{L}}^n(x,\tilde{\sigma}) = \mathcal{L}^n(x,\sigma) \tag{3.28}$$

Une dérivée normalisée $\mathcal{L}^n(x,\sigma)$ suivant ce processus est équivariante par rapport au σ utilisé. Il est possible de définir un nouveau vecteur de mesures $\mathcal{M}$ constitué de dérivées de Gaussiennes normalisées :

$$\mathcal{M} = [\mathcal{L}_x \ \mathcal{L}_y \ \mathcal{L}_{xx} \ \mathcal{L}_{xy} \ \mathcal{L}_{yy} \ \mathcal{L}_{xxx} \ \mathcal{L}_{xxy} \ \mathcal{L}_{xyy} \ \mathcal{L}_{yyy}]^T \tag{3.29}$$

Ce vecteur est équivariant à l'échelle: il peut être calculé de façon similaire pour des résolutions différentes.

La multiplication par le facteur σ^n pose, néanmoins, un problème de stabilité par rapport au bruit. Ce facteur multiplicatif augmente le bruit d'autant plus que σ est grand.

Dans le cadre d'une application en reconnaissance où l'échelle est a priori inconnue, il faut que l'échelle des dérivées calculées sur l'image modèle correspondent à l'échelle des dérivées calculées sur les images de test. Ceci, similairement à la rotation peut être obtenu suivant deux stratégies. D'une part, il est possible d'effectuer une modélisation multi–échelles. Dans ce cas, l'apprentissage ou la reconnaissance suivant une large gamme d'échelles permet de garantir pour un intervalle de variation en échelle que certaines mesures sont évaluées à la même échelle dans les images de test et les images modèles. Une alternative consiste à détecter une échelle caractéristique et de caler les filtres suivant cette échelle de façon à obtenir des mesures invariantes à l'échelle. Cette stratégie qui a l'intérêt d'être peu coûteuse est développée par la suite.

Validation expérimentale de l'équivariance à l'échelle La validation de l'équivariance des filtres à l'échelle requiert une évaluation de la reconnaissance obtenue sur une base d'images contenant des variations d'échelles connues. Une base de 28 objets est utilisée (voir base complète en annexe A.3). Chacun des objets est photographié sous 17 échelles différentes. L'échelle intermédiaire est sélectionnée comme modèle et les autres images sont utilisées pour l'évaluation de la reconnaissance. Les transformations affines entre images sont préalablement évaluées. Ceci permet pour tous points de valider ou rejeter une reconnaissance de façon fiable. La figure 3.14 montre les taux de reconnaissance comme fonction du rapport d'échelle entre l'image modèle et les images de test pour la série "chocos". L'échelle 1 correspond à la reconnaissance de l'image modèle soit une reconnaissance sans erreurs mais avec environ 35% de rejets. L'utilisation de points issus d'un fond relativement uniforme peut expliquer ce taux de rejet. Ce taux décroît lorsque la caméra s'approche de l'objet. Ceci peut s'expliquer par la présence d'un fond assez uniforme en arrière plan de l'objet pour les images d'échelles inférieures ou égale à 1. Par contre, les images prises plus proches ne contiennent plus ce fond et l'ensemble des points est alors discriminant. Le taux de fausses reconnaissances est inférieur à 5% ce qui est très faible. La figure 3.15 montre les résultats pour certaines images de test. Une image d'une dérivée seconde normalisée est présentée de façon à illustrer la normalisation. Il est possible d'observer que le niveau de gris représentant une valeur de dérivée pour un point physique est constant entre les images de dérivées. La dernière ligne montre pour chacune des images les points reconnus par l'algorithme proposé. Les autres objets donnent des résultats similaires. Ces courbes et ces images démontrent la validité de l'équivariance à l'échelle de façon expérimentale et permettent d'envisager une reconnaissance robuste aux variations d'échelles par l'utilisation de cette normalisation.

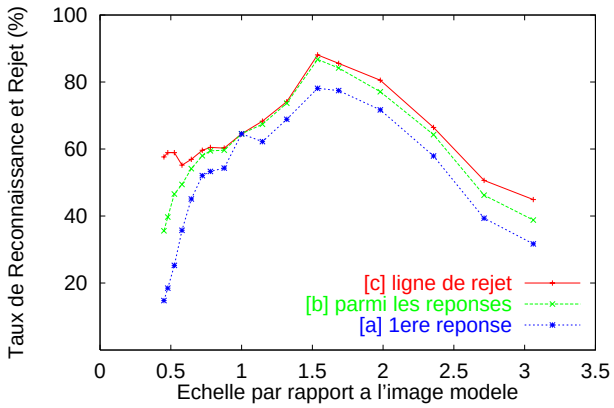


FIG. 3.14 – *Évaluation de la reconnaissance en fonction du changement d'échelle pour la série "Chocos" avec prise en compte du changement d'échelle connu pour l'adaptation du paramètre σ des filtres.*

Approche Multi-échelles Une approche classique de la reconnaissance tenant compte de variations importantes de l'échelle consiste à apprendre chaque objet sous de nombreuses échelles et espérer ainsi que l'échelle d'une nouvelle vue de l'objet soit proche d'une échelle préalablement apprise. Cette approche à été utilisée avec succès par de nombreux auteurs comme RAO [RB95], SCHIELE [Sch97] ou SCHMID [Sch96]. La stabilité des filtres utilisés par ces auteurs par rapport à l'échelle est de l'ordre de 10%. Ils ont choisi de partitionner l'espace des échelles en tranche de 20% : un apprentissage par tranche et ainsi toute nouvelle image présente les objets sous une échelle différent d'au plus 10% par rapport à l'une des échelles d'apprentissage. Cette approche n'est pas très satisfaisante car elle nécessite de dupliquer une grande partie des descripteurs de façon aveugle même s'ils représentent la même information. Il est, de plus, nécessaire d'utiliser plusieurs niveaux simultanément pour accéder à la reconnaissance. Une autre approche est proposée dans cette thèse, elle consiste à sélectionner en chaque point une échelle caractéristique que l'on retrouve quelque soit l'échelle d'observation.

Sélection automatique de l'échelle L'objectif de cette section est de présenter une technique novatrice permettant de détecter automatiquement une échelle caractéristique pour l'ensemble des points d'une image de façon à pouvoir régler le paramètre d'échelle d'une base de descripteurs. Ceci fournit une description invariante à l'échelle et, par conséquent, une stratégie de reconnaissance robuste aux variations d'échelles. Le principe est similaire à la détection de l'orientation locale par la direction du gradient proposée à la section 3.4.2. Cette détection a permis de caler les descripteurs en orientation et d'obtenir ainsi une invariance à l'orientation.

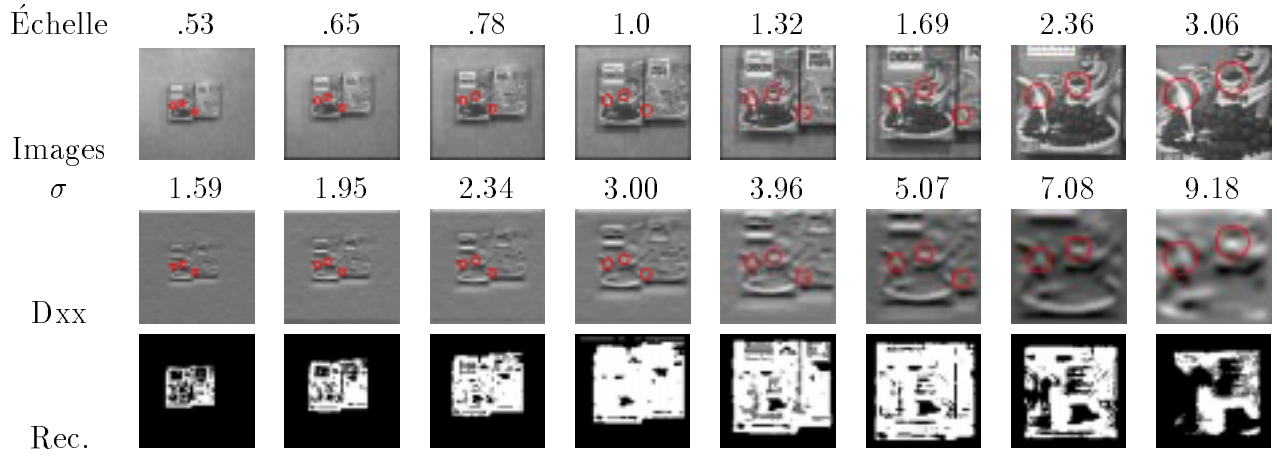


FIG. 3.15 – *Extrait de la série d'images "Chocos" avec changement de focale (voir annexe A.3) La première ligne présente les images brutes, la seconde les images des dérivées secondes normalisées dxx puis la troisième les cartes des points reconnus. Chaque colonne correspond à une échelle indiquée qui a permis de fixer le paramètre σ. Trois points physiques sont suivis sur toutes les images. Le rayon des cercles correspond à 2σ.*

Détection des "blobs" LINDEBERG [Lin98] propose une stratégie pour évaluer une échelle caractéristique locale en se fondant sur l'hypothèse que : les extrema locaux par rapport à l'échelle de dérivées secondes normalisées caractérisent l'échelle des formes observées. L'échelle sélectionnée correspond au paramètre d'échelle pour lequel la convolution entre l'opérateur de dérivation et le signal image donne la réponse maximale. Le détecteur d'échelle proposé est un opérateur Laplacien normalisé de façon à être équivariant à l'échelle :

$$Lap(x, y, \sigma) = (\sigma^2)(\partial_{xx} g(x, y, \sigma) + \partial_{yy} g(x, y, \sigma))$$

Le Laplacien est isotrope et de plus, il est intuitivement maximal sur des caractéristiques de type "blobs". Ces caractéristiques sont visibles indépendamment de l'échelle de l'image et se retrouvent à différentes échelles. En pratique, il s'agit de trouver le (ou les) extrema de la fonction $Lap(x, y, \sigma)$ en l'évaluant pour une large gamme de σ. L'échelle σ_0 sélectionnée correspond à l'extremum $Lap^{max}(x, y, \sigma = \sigma_0)$. L'utilisation du paramètre σ_0 pour régler le vecteur de mesures $\mathcal{M}$ rend ce vecteur *invariant* aux variations d'échelle.

Étude expérimentale du détecteur d'échelle local Ce paragraphe illustre la technique de sélection de l'échelle locale sur un exemple. La figure 3.16 montre un exemple d'appariement de caractéristiques entre deux images d'un objet observé sous deux échelles différentes. La courbe centrale montre pour chacune des deux images et pour les caractéristiques sélectionnées la courbe d'évolution du Laplacien normalisé en fonction de

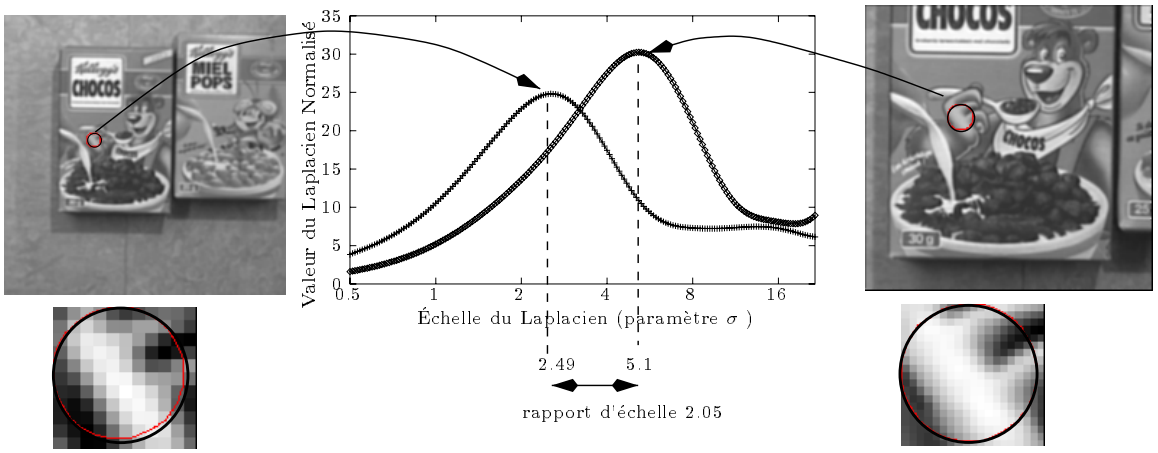


FIG. 3.16 – *Sélection automatique de l'échelle pour évaluer des appariements de points entre deux images. Les courbes présentent l'évolution du Laplacien Normalisé avec σ. Les cercles montrent un rayon $2\sigma_0$ (σ_0 est le paramètre d'échelle sélectionné). Le rapport entre les σ_0 donne un rapport d'échelle approximatif de 2 entre les images.*

l'échelle σ. Les courbes présentent parallèlement des maxima M_0 et M'_0 pour lesquels σ_0 et σ'_0 sont choisis pour calculer les vecteurs de mesures $\mathcal{M}_{\sigma_0}$ et $\mathcal{M}'_{\sigma'_0}$. Le cercle tracé est centré sur la caractéristique et a pour rayon 2σ. Ainsi, la détection d'une échelle caractéristique a permis de caler les filtres des vecteurs de mesure. Cela donne $\mathcal{M}_{\sigma_0} \approx \mathcal{M}'_{\sigma'_0}$. Cette égalité donne l'appariement entre les points des deux images malgré un changement d'échelle important (environ 2).

La figure 3.17 illustre la technique pour trois images du même objet. La première ligne montre les images brutes, la deuxième montre les images des paramètres σ_0 détectés en chaque point puis la troisième montre les images de $\mathcal{M}[0]$ la dérivée première calculée en utilisant les paramètres σ_0 détectés. Les images de σ_0 présentent la même structure pour les trois échelles. De plus, on peut observer sur les images de dérivées que les points se correspondant ont un niveau de gris identique qui correspond à une même valeur de dérivée et permet donc un appariement. Quatre appariements associés avec un cercle de rayon $2\sigma_0$ sont montrés sur les images de dérivées. Les rapports entre σ_0 détectés sont approximativement constant et correspondent aux variations d'échelle effectives entre les images.

Limitations La technique proposée donne, dans la plupart de nos expériences, des résultats satisfaisants : chaque point présente un maximum du Laplacien normalisé qui est utilisé pour calculer le vecteur de mesure $\mathcal{M}$. Néanmoins, cette propriété n'est pas garantie pour l'intervalle des valeurs du paramètre σ calculable. En certains points, aucun maximum n'est disponible. Pour d'autres cas comme celui présenté sur la figure 3.18

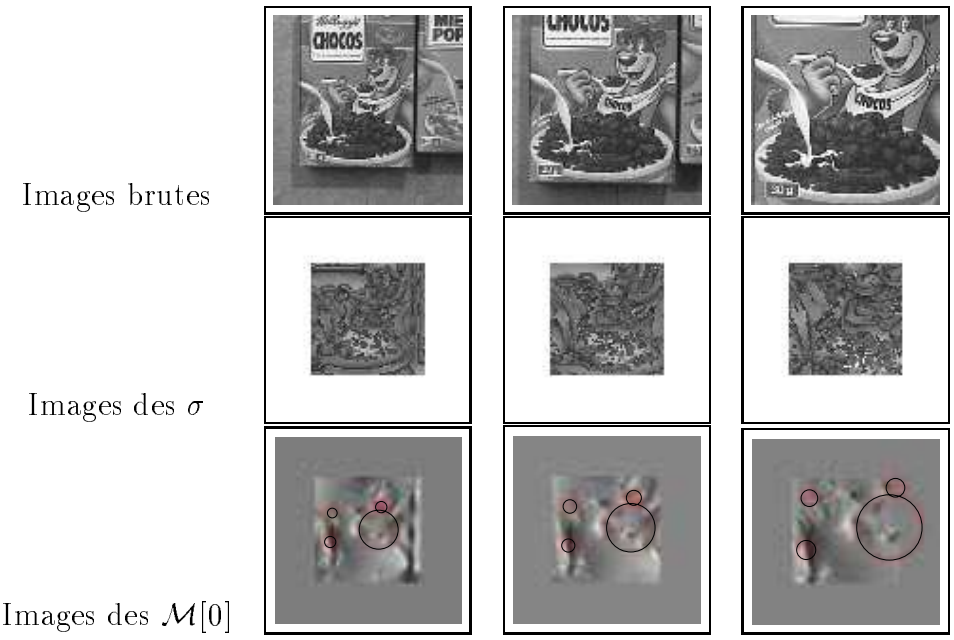


FIG. 3.17 – *(1) Images brutes, (2) Images des σ sélectionnées (ces valeurs correspondent aux maxima du Laplacien normalisé) et (3) Quelques correspondances visualisées sur $\mathcal{M}[0]$ (la dérivées premières après réglage de l'échelle).*

plusieurs maxima sont observés. Dans ce cas, cela signifie que plusieurs échelles caractéristiques sont disponibles et il est raisonnable de représenter ce point sous plusieurs échelles. Les deux cercles représentent les deux échelles visibles sur le graphe.

Ce problème illustre la difficulté du choix des deux paramètres de cette approche : l'intervalle de recherche des maxima et le pas de recherche de ces maxima. L'intervalle doit être le plus grand possible et le pas le plus faible possible. La stabilité importante des dérivées de Gaussiennes par rapport à l'échelle permet un pas relativement important de l'ordre de 10% pendant la phase de reconnaissance. Pour la phase d'apprentissage, un pas inférieur est profitable : 2% dans les expériences. D'autre part, la valeur minimale envisagée est $\sigma = 0.5$ ce qui correspond à une imagette de dimensions très faible 3×3. Un support aussi petit ne permet une évaluation correcte des différentes dérivées à cette échelle. Le choix de la valeur maximale dépend de la taille de l'image : le σ maximal utilisé est $\sigma = 20.0$ mais cela implique des fenêtres de grandes tailles, environ 120×120 , ce qui s'éloigne de la stratégie locale initiale et risque de poser des problèmes en cas d'occultation partielle. De plus, des filtres de grandes tailles ne sont calculables que sur une faible partie de l'image : les bords sont exclus car impliquant un repliement de spectre très important.

L'évaluation des maxima du Laplacien normalisé implique le calcul des images des Laplaciens pour de nombreuses valeurs de σ. De plus, le calcul des vecteurs de mesures

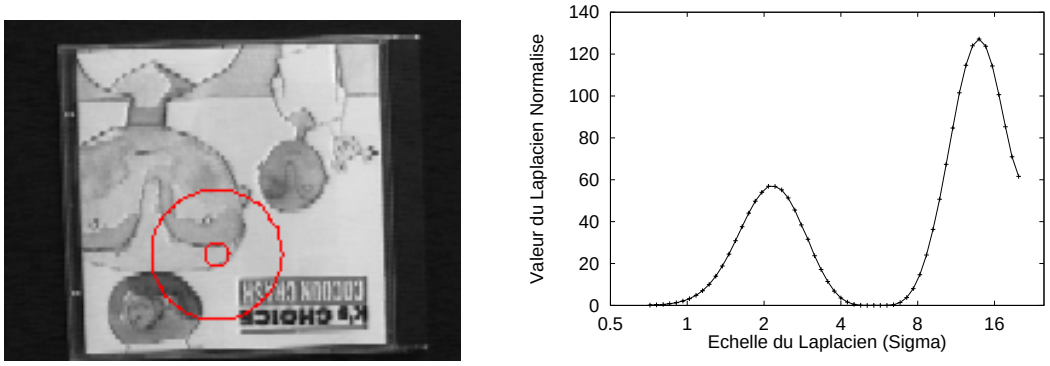


FIG. 3.18 – *Exemple de détection de maxima du Laplacien Normalisé. Cet exemple montre un point où deux maxima sont disponibles et donc deux échelles caractéristiques. Les cercles sont tracés en prenant pour centre le point évalué et pour rayon* 2σ *avec* σ *les paramètres d'échelles sélectionnés.*

pour toutes les échelles sélectionnées demandent de nombreuses convolutions par des filtres dérivées de Gaussiennes. Le calcul par filtres séparables classiques ne permet pas un temps de calcul raisonnable. Les paramètres donnés ici demandent de calculer pour une image 40 convolutions pour 9 filtres différents, soient 360 convolutions avec des filtres dont les tailles sont comprises entre 3×3 et 120×120 . Des techniques de filtrage récursif (voir annexe C) permettent de remédier à ce problème calculatoire sans, toutefois, permettre actuellement une implémentation temps–vidéo de la technique dans le cas général.

De plus, nous avons sélectionné l'échelle de représentation en utilisant un opérateur Laplacien, il s'agit de la maximisation d'une dérivée d'ordre 2. Ce paramètre est utilisé pour caler aussi bien les dérivées d'ordre 2 que les dérivées d'ordre 1 et 3. Ceci n'est pas forcément optimal : en effet, dans le cas 1D, le maximum d'une dérivée d'ordre 2 ne correspond pas intuitivement à des plages correctes pour les autres dérivées. A l'ordre 1, ce maximum donne une grande pente à la dérivée qui est donc plutôt instable. A l'ordre 3, ce maximum correspond à une annulation de la dérivée ce qui n'est pas très discriminant. Ces limitations intuitives permettent néanmoins d'obtenir des résultats expérimentaux corrects.

Validation expérimentale de la reconnaissance La reconnaissance d'objets est évaluée sur une base de 28 objets dont une image modèle est apprise. Les figures 3.20 et 3.21 montrent le taux de reconnaissance obtenu en utilisant, comme pour les expériences précédentes, un seul vecteur de mesure $\mathcal{M}$ pour effectuer la reconnaissance de l'objet. Les graphes correspondent aux objets "Chocos" et "Robot" de la figure 3.19. Les autres objets sont disponibles en annexe A.3.

La première colonne montre l'objet "Chocos" et la deuxième l'objet "Robot". Pour chacun des objets, deux graphes sont présentés : le premier correspond à une reconnais-

FIG. 3.19 – *Images modèles des séries "Chocos" et "Robot".*

sance sans sélection de l'échelle et le second avec sélection automatique de l'échelle. En abscisses, ces graphes montrent le rapport d'échelle avec l'image modèle et permettent donc d'évaluer la reconnaissance en fonction de l'échelle. Chaque graphe montre 3 courbes qui séparent les quatres cas de reconnaissance (voir section 3.1).

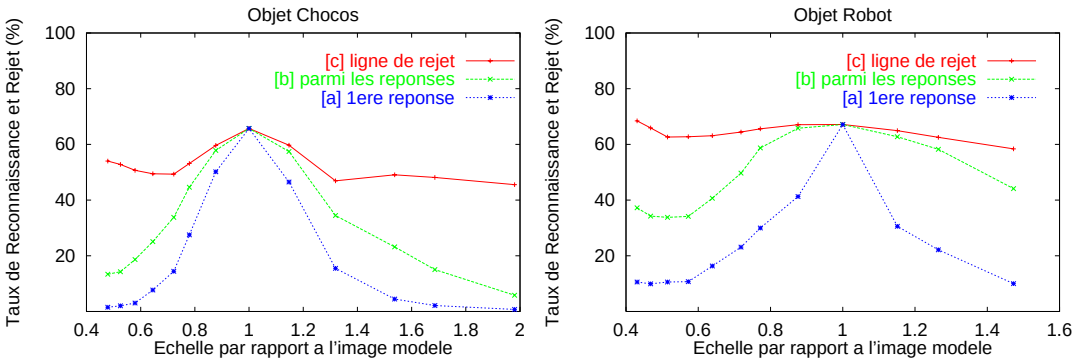


FIG. 3.20 – *Évaluation de la reconnaissance sans utilisation de l'algorithme de sélection de l'échelle. La reconnaissance s'écroule dés que le changement d'échelle dépasse 20%.*

On observe que le nombre de points rejetés est très important pour le cas sans détection d'échelle : en effet, la taille des filtres est fixée et des imagettes de niveau de gris quasiment constant sont fréquemment rencontrées mais ne permettent pas une reconnaissance. La sélection automatique de l'échelle supprime ces imagettes non discriminantes par une échelle plus adaptée. Sur la figure 3.20, la reconnaissance chute très rapidement à partir d'un changement d'échelle de 20% alors que, pour la figure 3.21, la reconnaissance diminue lentement et reste supérieure à 50% pour les images extrêmes (rapport d'échelle de l'ordre de 2). Une proportion de 50% est largement suffisante à un algorithme utilisant plusieurs points simultanément pour effectuer la reconnaissance.

Mesure invariante à l'orientation 2D et à l'échelle

La section 3.4.2 a montrée l'utilisation des filtres dérivées de Gaussiennes orientables pour la reconnaissance indépendante de l'orientation 2D. La section 3.4.3 a montrée l'utilisation d'une sélection de l'échelle locale couplée à une normalisation à l'échelle pour obtenir une reconnaissance robuste à l'échelle. Il est, par extension de ces deux approches,

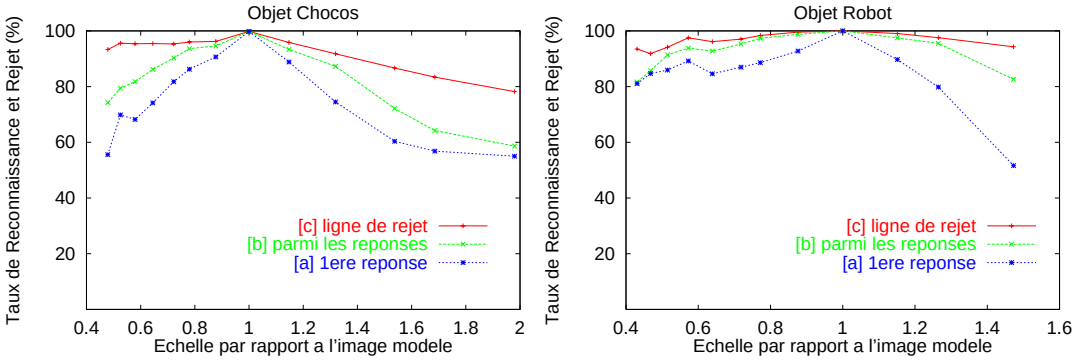


FIG. 3.21 – *Évaluation de la reconnaissance avec sélection automatique de l'échelle (précision 2% en échelle). Le taux de reconnaissance est supérieur à 50% pour un facteur d'échelle maximum de 2.*

possible d'obtenir en tout point d'une image un vecteur de mesures $\mathcal{M}$ invariant à ces deux paramètres.

Ce vecteur invariant $\mathcal{M}$ est calculé en deux étapes : d'abord, l'échelle locale (paramètre σ) est évaluée puis un vecteur de mesures $\mathcal{M}_e$ invariant à l'échelle est obtenu. Le calcul du gradient local sur ce vecteur suivi de l'orientation des mesures suivant sa direction permet d'obtenir le vecteur $\mathcal{M}$ invariant à l'orientation et à l'échelle. Il est calculé en fixant le paramètre d'échelle au σ détecté et en fixant l'orientation de base à α la direction du gradient :

$$\mathcal{M} = [\mathcal{L}^1_0 \ \mathcal{L}^2_0 \ \mathcal{L}^2_{\frac{\pi}{3}} \ \mathcal{L}^2_{\frac{2\pi}{3}} \ \mathcal{L}^3_0 \ \mathcal{L}^3_{\frac{\pi}{4}} \ \mathcal{L}^3_{\frac{\pi}{2}} \ \mathcal{L}^3_{\frac{3\pi}{4}}]^T \tag{3.30}$$

Ce processus peut être illustré par la figure 3.22 pour la projection d'un point $A(x, y)$ d'une image I en vecteur de mesure $\mathcal{M}(x, y)$. Un point est modélisé par un triplet $(\mathcal{M}, \sigma, \alpha)$. Le vecteur $\mathcal{M}$ permet l'appariement des points correspondants puis les deux paramètres (σ, α) permettent d'évaluer la similitude 2D entre les points appariés.

L'application des deux invariances est très profitable à la reconnaissance. Une expérience simple montre une amélioration de la reconnaissance sur une base d'images contenant uniquement des variations de l'orientation. Sans adaptation de l'échelle, la figure 3.13 (page 63) à la section 3.4.2 montre une grande quantité de rejets liés à des imagettes de niveaux de gris constants. La figure 3.23 montre la même expérience mais en utilisant la sélection automatique du paramètre d'échelle avant la détection de l'orientation. Le nombre de points rejetés est fortement diminué et la reconnaissance améliorée. Ceci s'explique par l'absence d'imagettes de niveau de gris constant après le recalage en échelle. La forte amélioration de la reconnaissance s'explique, aussi, par une échelle moyenne plus importante que l'échelle choisie dans l'expérience précédente. Des expériences sur des scènes contenant des variations simultanées en orientation et en échelle sont exposées au cours des chapitres suivants.

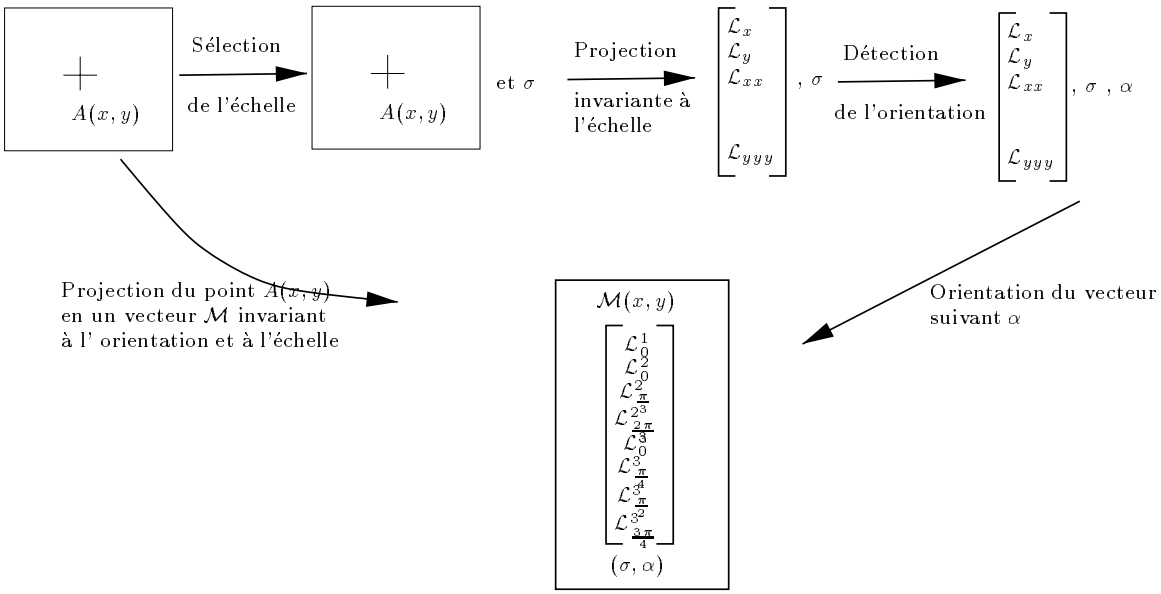


FIG. 3.22 – *Processus de projection de l'imagette englobant un point $A(x,y)$ en un vecteur de mesures invariant à l'échelle et à l'orientation. Ce processus est effectué en quatre étapes : sélection de l'échelle local, projection suivant cette échelle, détection de l'orientation locale puis projection invariante à l'échelle et à l'orientation.*

3.5 Conclusions

Ce chapitre a présenté une large gamme de filtres permettant de capturer une description locale des images. Dans le cadre le plus général où l'échelle et l'orientation sont inconnus, une base invariante à ces paramètres fondée sur les dérivées de Gaussiennes a été utilisée avec succès.

Comparaison expérimentale des bases de filtres : L'objet de ce paragraphe est une étude comparative de la reconnaissance selon la base de filtres utilisée. Les conditions expérimentales ne comportent pas de variations d'orientation de la caméra ou de l'échelle. Dans le cas où ces paramètres varient seules certaines base de filtres sont utilisables. La base utilisée est extraite de la base de Columbia (voir annexe A.1). 40 objets sont extraits avec une image tous les 20°. La reconnaissance est évaluée sur les images intermédiaires pour l'ensemble des points.

Les six bases évaluées sont les suivantes :

1. Base de filtres ACP en niveau de gris (section 3.3). La base est constituée de 10 dimensions et chacun des filtres ont la taille 9×9 (base notée ACP-g).

2. Base de filtres ACP en couleur (section 3.3). La base est constituée de 10 dimensions

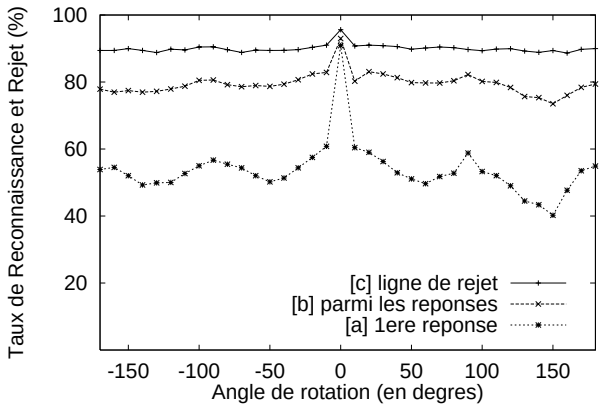


FIG. 3.23 – *Évaluation expérimentale de la reconnaissance en fonction de l'angle de vue avec évaluation automatique de l'échelle et de l'orientation. Le nombre de points rejetés est très inférieur à la même expérience sans sélection de l'échelle. Le taux de reconnaissance est lui aussi augmenté.*

et chacun des filtres ont la taille $9 \times 9 \times 3$ (base notée ACP-c).

3. Base de dérivées de Gaussiennes jusqu'à l'ordre 3 sans normalisation à l'orientation et l'échelle. Cette base est constituée de 9 dimensions. Chaque dérivée est évaluée avec $\sigma = 3$ (base notée DG).

4. Base de dérivées de Gaussiennes orientables jusqu'à l'ordre 3. L'orientation locale est détectée en utilisant la direction du gradient, puis tous les descripteurs sont tournés suivant cette orientation. Cette base est constituée de 8 dimensions. La deuxième dérivée première est identiquement nulle. Chaque filtre est évalué avec $\sigma = 3$ (base notée DG-o).

5. Base de dérivées de Gaussiennes normalisées en échelle jusqu'à l'ordre 3. L échelle locale est sélectionnée en détectant le maximum d'un filtre Laplacien normalisé puis cette échelle est utilisée pour évaluer les 9 dérivées. Le paramètre σ varie entre 0.5 et 20.0 avec un pas d'évaluation de 6%. Le maximum du Laplacien normalisé n'apparaît pas en tous les points des images et les points sans maximum sont supprimés du traitement. Un seul maximum a, de plus, été utilisé pendant l'apprentissage malgré la détection en certains points de plusieurs maxima (base notée DG-e).

6. Base de dérivées de Gaussiennes normalisées en échelle puis en orientation. Les deux normalisations sont appliquées successivement pour obtenir un vecteur de 8 descripteurs (base notée DG-oe).

La figure 3.24 montre les résultats de l'évaluation comparative des différentes bases

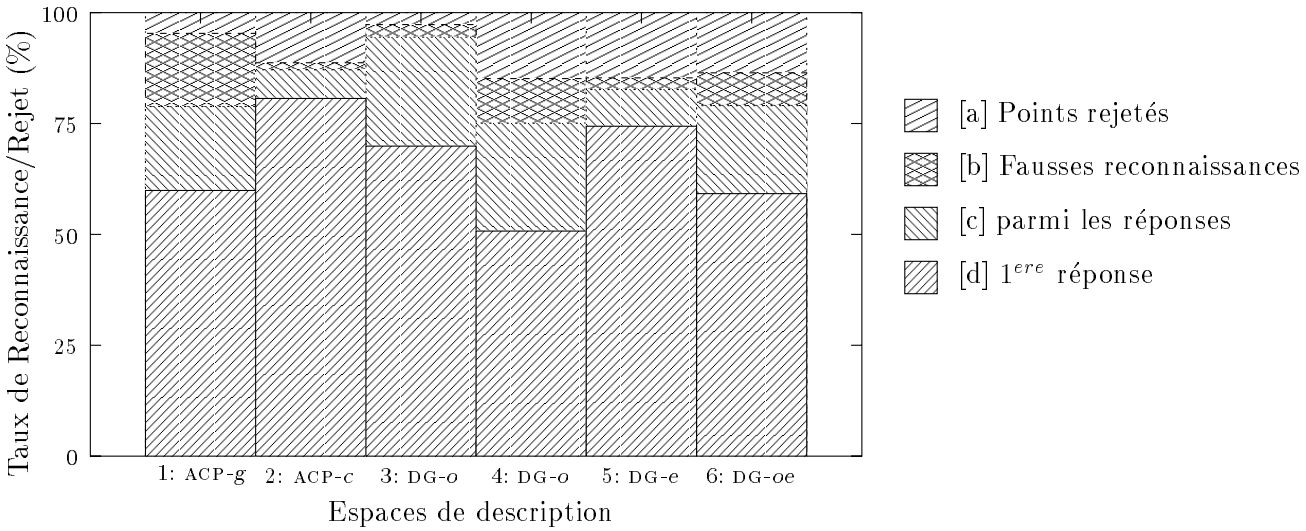


FIG. 3.24 – *Comparaison expérimentale des différentes bases de filtres présentées dans ce chapitre. La reconnaissance est évaluée sur une base de test où l'orientation et l'échelle ne varient pas.*

de filtres sur l'intégralité des imagettes des images de test. La recherche d'une imagette implique l'une des quatre réponses $[a]$ à $[d]$ définies à la section 3.1.

Globalement, la figure montre que, quelque soit la base de filtres utilisée, plus de 50% des imagettes permettent une reconnaissance directe de l'objet et de sa pose. Ce résultat permet d'envisager une reconnaissance d'objets à partir d'appariements multiples très discriminante.

Un taux d'échec important est observé pour la base de filtres ACP (1), ce taux supérieur s'explique principalement par la taille très faible des imagettes utilisée (9×9). L'espace des dérivées de Gaussiennes donne de meilleurs résultats car le paramètre $\sigma = 3.0$ utilisé correspond à des imagettes de plus grandes dimensions.

L'utilisation de la couleur améliore de façon très conséquente la reconnaissance par filtres ACP et motive une utilisation des filtres dérivées de Gaussiennes en couleur. Cette expérience permet de conclure que la base de dérivées de Gaussiennes est la plus adaptée parmi les bases étudiées pour la reconnaissance par mesures de caractéristiques locales. Les invariances éventuelles aux paramètres d'orientation et d'échelle permettent d'obtenir une reconnaissance robuste par rapport à ces paramètres avec une perte très limitée en discrimination.

Conclusions La description locale du signal image est très discriminante pour la reconnaissance. Le taux de reconnaissance obtenu pour une imagette quelconque est élevé. Une technique de reconnaissance d'objets peut se fonder de façon sure sur cette description locale. L'utilisation de plusieurs imagettes et de leur positions spatiales mutuelles permet

de définir une système robuste de reconnaissance d'objets (voir chapitre 6).

Une étude de la sensibilité des filtres par rapport à plusieurs sources de bruit est proposée dans le chapitre suivant. Un objectif de cette évaluation est de déterminer une relation entre la similarité entre deux imagettes et la distance entre les vecteurs de mesures correspondants. La similarité entre deux imagettes est d'autant plus importante que la distance entre les vecteurs de mesures correspondants est faible. Le but de ce chapitre est d'obtenir un seuil sur la distance entre vecteurs qui permet de définir un prédicat sur la similarité de deux imagettes.

Chapitre 4

Sensibilité des descripteurs locaux

Dans le chapitre 3, plusieurs bases de descripteurs locaux ont été étudiées et comparées. Ces descripteurs permettent de représenter une image par une grille de vecteurs de mesures locales. Ces vecteurs de mesures sont évalués sur des images bruitées et leur évaluation est elle–même bruitée : la capture de l'image d'un objet par une caméra puis son échantillonnage entraînent un bruit de numérisation important, puis, les traitements sur l'image (filtrages) sont eux-mêmes bruités et entraînent un biais supplémentaire dans l'évaluation des mesures. L'ensemble de ces sources de bruit ne permet pas un appariement *exact* entre mesures issues d'images différentes d'un même objet en utilisant un dictionnaire. C'est pourquoi, il est nécessaire de définir un critère permettant d'évaluer la similarité entre vecteurs de mesures. Ce critère doit, en particulier, définir si deux vecteurs correspondent à une même caractéristique visuelle et, donc, représentent un même point physique. Cette similarité peut être évaluée en termes de distances entre vecteurs de mesures. L'utilisation d'une distance implique l'étude du comportement de cette distance par rapport aux sources de bruit de façon à pouvoir évaluer un seuil de recherche.

Le problème de l'évaluation de la similarité entre vecteurs de mesures est étudié dans la première section. Il s'agit de choisir une distance permettant de comparer les vecteurs puis de positionner un seuil sur cette distance de façon à obtenir un prédicat sur la similarité de deux vecteurs. Le seuil de similarité par rapport à différentes perturbations comme le bruit numérique, le changement d'éclairage ou des changements de point de vue de la caméra est mesuré dans les sections suivantes. Ces expériences sont effectuées pour une base de dérivées de Gaussiennes mais sont extensibles directement à d'autres bases de filtres.

La connaissance du comportement des filtres par rapport au bruit permet de choisir un seuil de similarité et de définir une prédicat de similarité entre vecteurs de mesures. Il est, par la suite, possible d'évaluer différentes techniques d'apprentissage (par points d'intérêts, statistique ou structurel), puis de définir une stratégie de reconnaissance d'objets.

4.1 Évaluation de la similarité entre vecteurs de mesures

La stratégie de reconnaissance présentée dans cette thèse est fondée sur l'appariement entre vecteurs de mesures. L'appariement de vecteurs est obtenu en mesurant la similarité entre ceux-ci. L'aspect bruité du traitement d'images implique un appariement de vecteurs non identiques mais proches suivant une distance et un seuil à déterminer. Plusieurs choix de distance sont étudiés dans la section 4.1.1, puis le problème de l'évaluation du seuil de recherche est présenté dans la section 4.1.2.

4.1.1 Distance d'évaluation

Plusieurs distances peuvent être choisies pour évaluer la similarité entre deux vecteurs de mesures. Deux distances sont envisagées ici : la distance euclidienne et la distance de Mahalanobis.

La distance L_2 ou euclidienne (somme des carrés des différences entre coordonnées des vecteurs) est adaptée si les différentes coordonnées des vecteurs de mesures ont des ordres de grandeur similaires. Ceci est vérifié pour le cas de l'utilisation de filtres obtenus par apprentissage par Analyse en Composantes Principales. Dans ce cas, la distance euclidienne entre les vecteurs de mesures est une approximation de la distance euclidienne entre les imagettes. Ainsi, la mesure de similarité revient à une corrélation qui est une mesure optimale de similarité sur un signal en présence de bruit Gaussien additif (voir Martin [MC95] et Lan [Lan97]).

La distance de Mahalanobis normalise les réponses entre les dimensions. Cette normalisation est fondée sur une hypothèse simplificatrice de répartition Gaussienne des vecteurs dans l'espace de représentation. Cette répartition est évaluée par une matrice de covariance Q de taille $m \times m$ avec Q_{ii} les variances dans chacunes des dimensions et Q_{ij} avec $i \neq j$ les covariances entre les dimensions. Pour deux vecteurs de mesures $\mathcal{M}_1$ et $\mathcal{M}_2$, la distance $d^2_{Mah}(\mathcal{M}_1, \mathcal{M}_2)$ est définie par :

$$d^2_{Mah}(\mathcal{M}_1, \mathcal{M}_2) = (\mathcal{M}_1 - \mathcal{M}_2)^t Q^{-1} (\mathcal{M}_1 - \mathcal{M}_2) \tag{4.1}$$

Q est évaluée sur la base d'apprentissage. L'hypothèse de répartition Gaussienne n'est pas vérifiée et une évaluation plus précise montrerait que les paramètres de Q ne sont pas constants selon la position dans l'espace.

Les expérimentations décrites ici utilisent principalement la distance de Mahalanobis car elle permet de s'adapter à des données d'ordre de grandeurs différents et de tenir compte des interactions entre les dimensions. L'évaluation de Q est effectuée sur la base d'apprentissage en la considérant indépendante de la position dans l'espace.

4.1.2 Évaluation du seuil

Le bruit sur les mesures implique d'évaluer jusqu'à quelle distance deux mesures peuvent être considérées comme identiques, ou, plus précisément, comme représentant une même donnée physique. Le bruit sur les mesures provient de plusieurs sources simultanées comme l'illustre la figure 4.1.

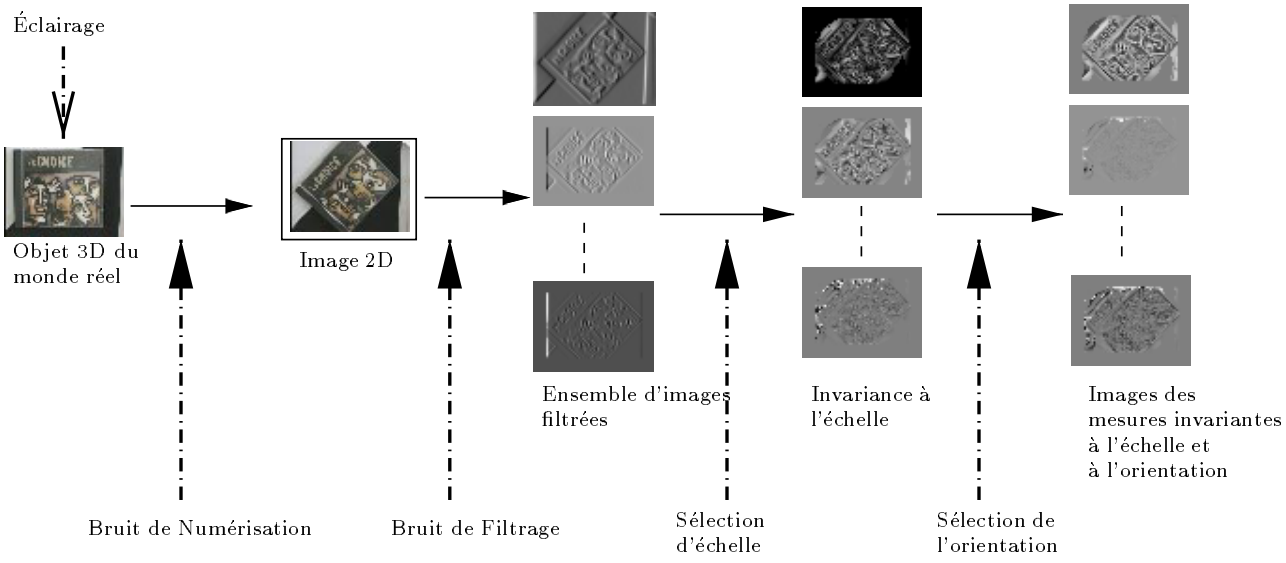


FIG. 4.1 – *Schématisation de l'ajout de bruit cumulé sur le signal par la succession de traitements effectués. A partir d'un objet* 3D *du monde réel, un ensemble de traitements est effectué avant d'aboutir à, pour chacun des points d'une image de cet objet, une mesure* $\mathcal{M}$ *invariante ou, au minimum, stable aux variations de l'éclairage, de l'orientation de la caméra et de l'échelle. Chacun des traitements ajoute un bruit à l'évaluation des vecteurs de mesures et doit être mesuré.*

Il s'agit de mesurer l'évolution des vecteurs de mesures suivant plusieurs perturbations :

- Le bruit de la chaîne d'acquisition et du filtrage.
- Le bruit lié aux propriétés d'invariance des descripteurs (échelle et orientation 2D).
- Le bruit lié aux variations d'éclairage.

Les sections suivantes caractérisent successivement les différentes sources de bruit en évaluant la distribution (histogramme, moyenne et variance) des distances entre points correspondants. Cette évaluation par moyenne et variance est valide sous l'hypothèse d'une distribution Gaussienne. Cette hypothèse n'est pas vérifiée mais permet une visualisation simplifiée de la distribution. L'algorithme d'évaluation consiste à utiliser une base d'images calibrées, c'est-à-dire dont les correspondances points à points entre images sont connues.

Pour les évaluations proposées, la transformation entre une image modèle et une image bruitée est soit l'identité soit une similitude 2D représentée par une matrice. L'évaluation consiste à mesurer pour tous les couples de points appariés la distance entre les vecteurs de mesures correspondants. La connaissance de la transformation affine entre les deux images permet de n'utiliser que les distances correspondants à des appariements valides. Puis, la distribution des distances est évaluée sur plusieurs images.

Deux types de courbes sont proposés : d'une part, des histogrammes de répartition des distances entres points correspondants de deux images et, d'autre part, une courbe de l'évolution de la distance en fonction du paramètre du modèle de bruit étudié. Cette courbe permet d'évaluer la stabilité des vecteurs de mesures par rapport à ce paramètre. Il est alors possible de borner les variations de ce paramètre par rapport à un seuil de reconnaissance ou d'évaluer le seuil de reconnaissance pour un intervalle de variations du paramètre de bruit.

L'appariement pour une perturbation importante demande le choix d'un seuil important. Plus ce seuil est choisi grand, plus le nombre de faux appariements devient important. Il est donc nécessaire d'évaluer le seuil maximal utilisable pour qu'une recherche ne renvoie pas des appariements non fondés. Une borne supérieure peut être mesurée en évaluant la distance moyenne entre deux vecteurs de mesures correspondants à deux points quelconques. Cette évaluation est faite en tirant aléatoirement un grand nombre de points sur des images. La moyenne obtenue sur la distance est 0.15 avec un écart–type de 0.04. Ceci motive à l'utilisation d'un seuil inférieur à 0.1 pour valider un appariement entre deux vecteurs. Plus précisément, sous une approximation de distribution Gaussienne, 95% des couples de vecteurs de points choisis aléatoirement sont distants d'au moins 0.7. Une distance inférieure garantit donc une similarité entre deux vecteurs. Ce seuil permet d'évaluer les évolutions des vecteurs de mesures liés aux différentes perturbations.

4.2 Sensibilité au bruit numérique

Les phases de capture de l'image d'un objet et de sa numérisation entraînent un bruit sur l'image obtenue. Le bruit de numérisation peut être modélisé comme un bruit additif aléatoire de distribution Gaussienne. Le bruit lié à la mise au point de la caméra peut être modélisé par une fonction de flou (filtrage[1] répété par la moyenne). Cette section évalue l'influence de ces deux bruits sur une image réelle.

La figure 4.2 montre une image dégradée par un bruit Gaussien additif de plus en plus important. Son écart-type progresse de 0.0 à 80.0. La figure 4.3 montre la stabilité des vecteurs de mesures obtenues pour l'ensemble des points des images. La distance moyenne et la variance sont représentées en fonction du bruit ajouté. Pour un bruit faible ($\sigma < 10$), les vecteurs de mesures évoluent sous une distance avec le vecteur modèle de 0.06.

1. mean-filtering

FIG. 4.2 – *Série d'images bruitées avec un bruit Gaussien additif. Les images présentées ici ont été obtenues en ajoutant un bruit Gaussien additif à l'image initiale en utilisant les paramètres σ de bruit indiqués sur la deuxième ligne de la figure.*

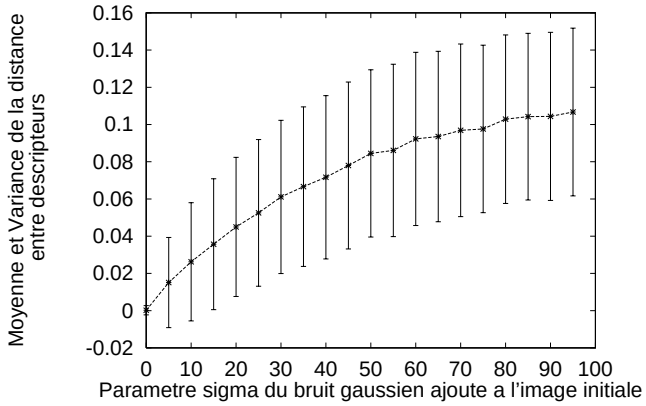


FIG. 4.3 – *Stabilité par rapport à un bruit Gaussien additif. Ce graphe montre l'évolution de la distance entre vecteurs de mesures se correspondant en fonction d'un bruit Gaussien additif croissant. Cette évolution est visualisée par sa moyenne et son écart-type.*

De façon à simuler un flou lié à une mise au point incorrecte de la caméra, il est possible d'appliquer plusieurs fois une filtre de lissage par la moyenne. Ici, un masque 3×3 de moyennage est appliqué de 1 à 10 fois sur une image pour observer le comportement des vecteurs de mesures par rapport à une mise au point imprécise. La figure 4.4 montre les images considérées. La figure 4.5 montre l'évolution de la distance entre vecteurs de mesures comme fonction du flou ajouté. Un flou faible (1 ou 2 applications du masque de moyennage) perturbe de façon faible les mesures (distance inférieure à 0.06). Par contre, un flou plus grand entraîne rapidement une distance souvent supérieure à 0.1 et donc peu utilisable pour la reconnaissance.

4.3 Influence des invariances

L'espace de *dérivées de Gaussiennes* proposé au chapitre 3 permet une invariance aux paramètres d'échelles et d'orientation 2D. Cette invariance n'est pas numériquement

FIG. 4.4 – *Série d'images filtrées plusieurs fois avec un opérateur de moyennage (ajout de flou). L'indice représente le nombre d'application du masque* 3×3 *de moyennage.*

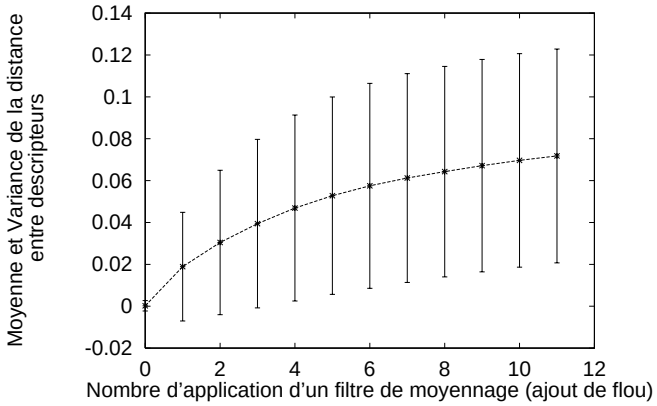


FIG. 4.5 – *Stabilité par rapport à un bruit de flou. Ce graphe montre l'évolution de la distance entre vecteurs de mesures se correspondant par rapport à un flou croissant (simulation d'une mauvaise mise au point de la caméra). Cette évolution est visualisée par sa moyenne et son écart-type.*

exacte et il est nécessaire d'évaluer le bruit inhérent à l'utilisation de cette invariance. Les sections suivantes caractérisent expérimentalement le bruit associé à ces deux perturbations des images.

4.3.1 Échelle

Ce paragraphe analyse la distribution des distances entre vecteurs de mesures correspondants en présence de variations d'échelle. Tout d'abord entre deux images de la série "chocos" avec une variation d'échelle de 65%, la figure 4.6 montre l'histogramme des distances points à points. La stabilité par rapport au changement d'échelle apparaît suffisante avec un pic autour la distance 0.01.

Puis, plus globalement sur une plus grande base d'images, la distance moyenne et sa variance sont évaluées comme fonction du changement d'échelle effectif (figure 4.7). L'échelle 1 correspond à l'évaluation de la reconnaissance sur l'image modèle. Les échelles supérieures et inférieures sont obtenues en faisant varier la focale de la caméra. La distance

Rapport d'échelle
65%

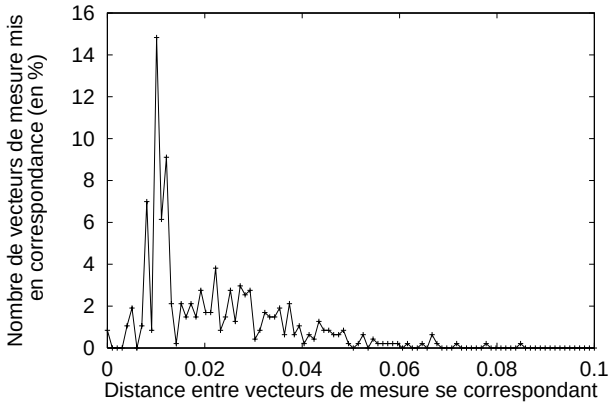
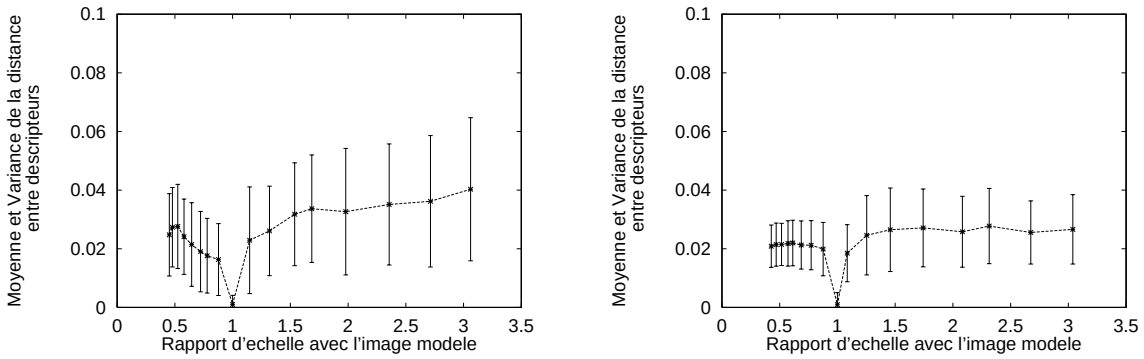


FIG. 4.6 – *Distribution des distances entre vecteurs de mesures invariants à l'échelle entre les deux image de la série "chocos" présentées sur la gauche de la figure. Le rapport d'échelle entre les images est de 65%. La distribution des distances est évaluée par un histogramme des distances entre points correspondants. Seules les distances correspondants à des appariements corrects sont prises en compte.*

apparaît stable par rapport au facteur d'échelle. Les deux séries donnent des résultats équivalents. Suivant ces courbes, le seuil de reconnaissance doit être en présence de variations d'échelle au moins égal à 0.04. Ces graphes ne font pas apparaître le nombre de points mis en correspondance entre les images. Ce nombre diminue avec l'augmentation du facteur d'échelle. Ceci s'explique par la processus de sélection d'échelle qui fait apparaître de nouvelles échelles lorsque la caméra se rapproche de l'objet et fait disparaître certaines échelles lorsqu'elle s'éloigne. Les expériences présentées ici se limite à l'utilisation d'un seul maximum en échelle du Laplacien normalisé (voir section 3.4.3 page 64). Cette limitation simplifie l'évaluation mais implique une perte d'un grand nombre d'appariements pour les facteurs d'échelles les plus importants.

4.3.2 Rotation 2D

La distribution des distances entre vecteurs de mesures correspondants est évaluée en présence de variations de l'orientation 2D. Tout d'abord entre deux images avec une rotation de 90°, la figure 4.8 montre l'histogramme des distances points à points. La stabilité par rapport aux variations d'orientation apparaît assez importante avec un pic autour la distance 0.02. Le pic est plus large que pour les variations d'échelle et le seuil d'appariements doit être choisi plus grand que dans le cas de variations d'échelle. Cette imprécision sur l'évaluation des vecteurs de mesures dans le cas de changement d'orientation peut s'expliquer par l'aspect anisotropique de la chaîne d'acquisition et de filtrage des images. Les images sont numérisées sur des matrices 2D de pixels. Cette représentation n'est pas isotropique. De plus, la technique de calcul récursive des dérivées de Gaussiennes ajoute



Série *Chocos* Série *Cabel*

FIG. 4.7 – *Évolution de la distance (en moyenne et en variance) entre vecteurs de mesures invariants à l'échelle pour une large gamme d'échelles pour deux séries d'images. Les séries d'images Chocos et Cabel sont représentatives d'une large gamme d'objet (voir annexe A.3) Les deux graphes montrent une grande stabilité de la distance entre vecteurs de mesures se correspondant malgré des variations en échelle jusqu'à un facteur 3. Le nombre de points mis en correspondance diminue néanmoins avec le facteur d'échelle qui entraîne des détections d'échelles différentes.*

elle-même un bruit non isotrope (voir annexe C).

Plus globalement sur une plus grande base d'images, la distance moyenne et sa variance sont évaluées comme fonction du changement d'orientation effectif (figure 4.9). La figure montre une distance stable par rapport aux variations de l'orientation. La variance est assez importante et nécessite le choix d'un seuil minimal de l'ordre de 0.06 pour effectuer une mise en correspondance correcte des vecteurs de mesures entre images.

4.4 Invariance à l'éclairage par la normalisation

La vision et, plus particulièrement, la reconnaissance d'objets se heurte au problème majeur de la variation des images des objets avec l'éclairage. L'objet de cette section est une étude succincte de quelques techniques permettant de pallier dans une certaine mesure aux problèmes de variations de l'éclairage sous un modèle linéaire de caméra.

Indépendamment du choix de la base de filtres, les projections sur l'espace de représentation $\mathcal{A}$ sont effectuées par une convolution entre l'imagette et les filtres. Une normalisation de cette convolution peut permettre d'augmenter la robustesse aux variations de l'éclairage.

Le premier paragraphe étudie la normalisation des convolutions puis le second paragraphe montre des résultats expérimentaux sur cette normalisation pour atténuer les variations des images liées à l'intensité de l'éclairage. Le dernier paragraphe propose l'uti-

Rotation 90°

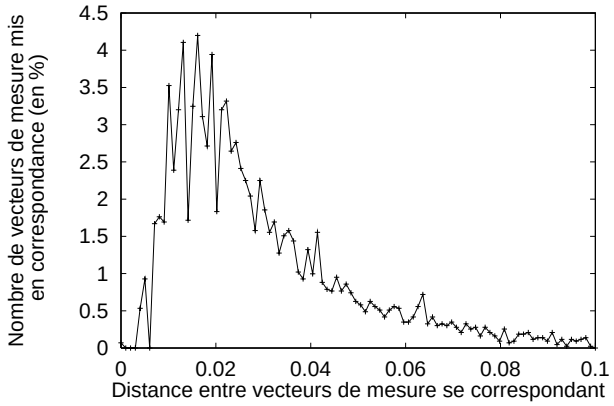


FIG. 4.8 – *Distribution des distances entre vecteurs de mesures invariants à l'orientation entre deux image présentées sur la gauche de la figure. La distribution des distances est évaluée par un histogramme des distances entre points correspondants. Seules les distances correspondants à des appariements corrects sont prises en compte.*

lisation de techniques de normalisations plus évoluées fondées sur l'utilisation de la couleur pour accéder à une plus grande invariance aux variations d'éclairage.

4.4.1 Normalisation des convolutions

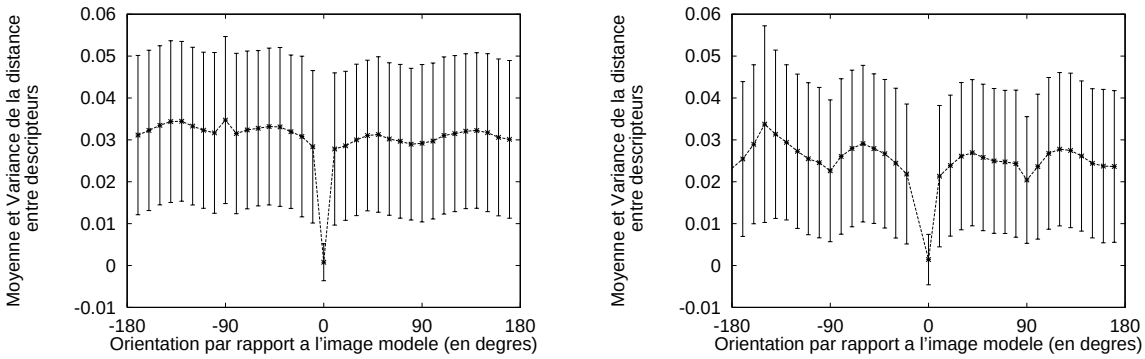
Pour limiter la sensibilité aux variations d'éclairage des mesures, il est possible de remplacer la convolution par un produit scalaire normalisé. Différents produits scalaires sont proposés par MARTIN [MC95] et SCHIELE [Sch96]. Tout d'abord, la convolution est la corrélation CC pour "Cross–Correlation" :

$$CC(\phi_i, J(x,y)) = \sum_{x'=x-M}^{x+M} \sum_{y'=y-N}^{y+N} \phi_i(x-x', y-y') W(x', y') \tag{4.2}$$

J est l'image et ϕ_i est le filtre de convolution. La convolution présente l'inconvénient d'une forte sensibilité aux variations d'éclairage. Si l'intensité moyenne double sur le voisinage de $J(x,y)$ alors le résultat double aussi. Pour pallier à cet inconvénient, sous un modèle de caméra linéaire, une convolution normalisée par l'énergie (NCC pour "Normalized Cross–Correlation") peut être utilisée avec $E()$ la fonction Énergie du signal :

$$E(J(x,y)) = \frac{1}{(2M+1)(2N+1)} \sqrt{\sum_{x'=x-M}^{x+M} \sum_{y'=y-N}^{y+N} J(x', y')^2}$$

pour une fenêtre centrée sur $J(x,y)$ de taille $[2M+1] \times [2N+1]$.



Série *c2_vp_rz_s2* Série *c2_vp_rz_s8*

FIG. 4.9 – *Évolution de la distance (en moyenne et en variance) entre vecteurs de mesures invariants à l'orientation. Deux séries d'images différentes sont représentées. Les deux graphes montrent une stabilité importante de la distance entre vecteurs de mesures quelque soit le changement de l'orientation observée. Les variations de la distance moyenne semble liées à l'orientation : les orientations autour de* $k\frac{\pi}{2}$ *donnent une distance moyenne inférieure. Ceci peut s'expliquer par l'aspect non isotropique des filtres utilisées.*

Dans le cadre des filtres dérivées de Gaussiennes, il y a un support a priori infini d'où la nécessité de masquer avec un filtre Gaussien $G(x,y)$ le support pour obtenir l'énergie correspondant au voisinage centré sur le point (x,y) pour un paramètre d'échelle σ :

$$E(J(x,y)) = \frac{1}{\sigma^2}\sqrt{\sum_{x'=x-M}^{x+M}\sum_{y'=y-N}^{y+N}[G(x-x',y-y')J(x',y')]^2dx'dy'}$$

Ce qui donne :

$$NCC(\phi_i, J(x,y)) = \frac{CC(\phi_i, J(x,y))}{E(J(x,y))} \tag{4.3}$$

Le modèle de caméra linéaire consiste à supposer que la réponse de la caméra est proportionnelle à l'intensité de l'éclairage soit, pour un point (x,y), une valeur $J(x,y)$ dépendant de l'intensité $I(x,y)$ de l'éclairage et de la réflectance $R(x,y)$ du point physique.

$$J(x,y) = R(x,y)I(x,y) \tag{4.4}$$

L'ajout de l'hypothèse d'un éclairage localement constant permet de supposer $I(x,y)$ constant à l'intérieur de la fenêtre W. Ainsi, NCC devient :

$$NCC(\phi_i, J(x,y)) = \frac{CC(\phi_i, R(x,y))}{E(\phi_i)E(R(x,y))} \tag{4.5}$$

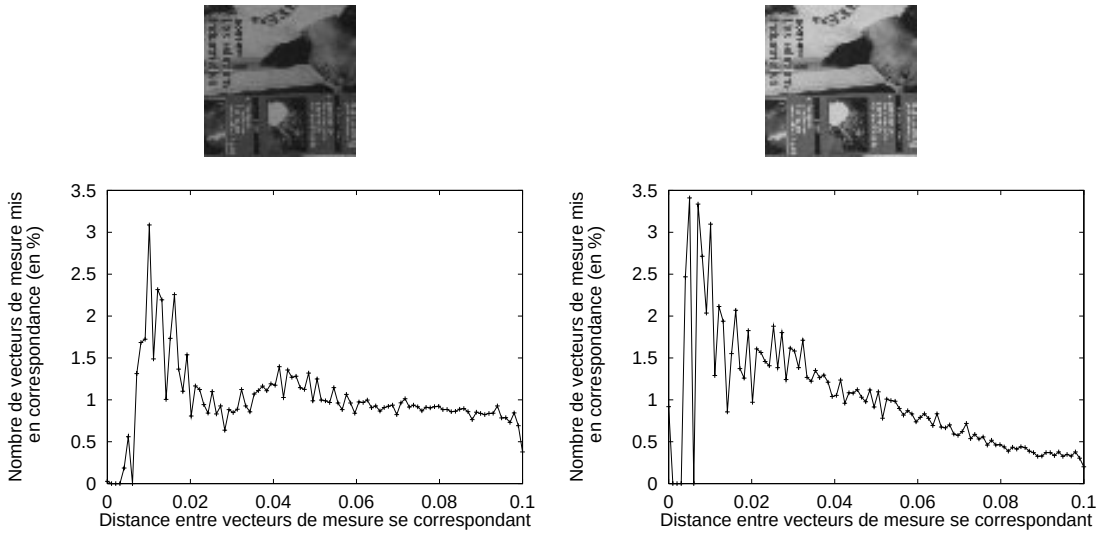
Cette équation ne fait plus apparaître le paramètre d'éclairage I et la convolution NCC admet donc une certaine invariance aux variations d'éclairage. Il faut néanmoins constater que la division par $E(R(x,y))$ diminue d'une dimension l'espace de l'apparence des imagettes.

D'autres normalisations peuvent être utilisées comme une normalisation par la variance ou une normalisation max-min (BOBET [Bob95]). Des expérimentations ont montrés que ces normalisations étaient moins intéressantes que la normalisation par l'énergie (voir SCHIELE [Sch96])

Le comportement expérimental de la normalisation par l'énergie par rapport à l'éclairage est étudié à la section suivante. L'usage de ces normalisations est profitable mais doit être nuancé par la perte d'information toujours liée à une normalisation.

4.4.2 Sensibilité aux variations de l'éclairage

La normalisation par l'énergie (NCC) permet une certaine robustesse aux variations de l'intensité lumineuse. Cette propriété est évaluée dans cette section. La figure 4.11 montre une série d'image pour laquelle l'intensité de la source de lumière a été continuement modifiée. Cette série d'images est issue de la base MOVI [Gro98].



Sans normalisation (CC) Normalisation par énergie (NCC)

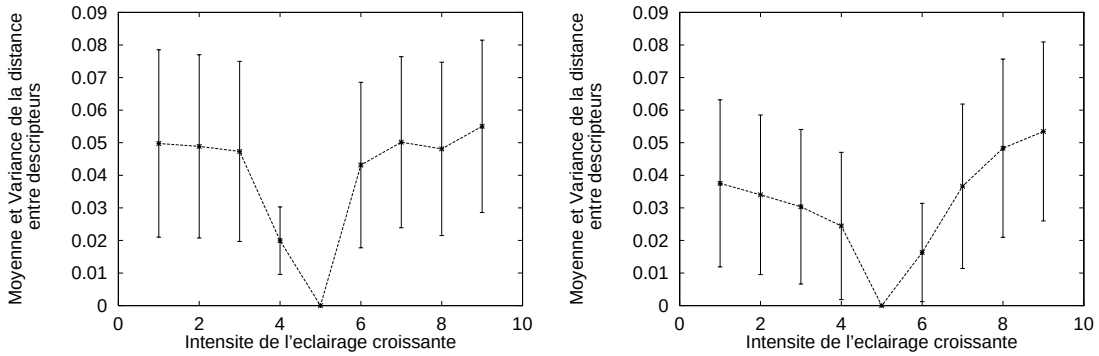
FIG. 4.10 – *Comparaison des distributions de distances entre vecteurs de mesures se correspondant avec et sans normalisation par l'énergie. Les graphes présentent les histogrammes des distances. La comparaison des deux histogrammes montrent une stabilité plus importante des vecteurs de mesures grâce à l'utilisation de la normalisation par l'énergie.*

Sur cette base d'images, la moyenne et la variance des distances entre vecteurs de

mesures se correspondants sont évaluées pour deux convolutions différentes : CC (sans normalisation) et NCC (normalisation par l'énergie). La figure 4.10 montre la répartition des distances points à points entre deux images présentant une variation de l'intensité lumineuse importante. Cette figure montre que la normalisation par l'énergie implique une plus grande partie des points proches de la distance nulle.

FIG. 4.11 – *Série d'images sous un éclairage d'intensité croissante (référence c2_1l_li_s1). L'éclairage est constitué d'un source de lumière dont l'intensité est variable.*

De façon plus globale, il est possible en utilisant la base complète (figure 4.11) de visualiser le comportement des distances par rapport à l'intensité. La figure 4.12 montre



Sans normalisation (CC) Normalisation par énergie (NCC)

FIG. 4.12 – *Évolution de la distance entre vecteurs de mesures comme fonction de l'intensité croissante de l'éclairage. Cette évolution est visualisée par sa moyenne et son écart-type.*

l'évolution de la distance moyenne et de sa variance (approximation Gaussienne) comme fonction de l'intensité de l'éclairage. Les courbes montrent une stabilité légèrement plus importante aux variations de l'éclairage par l'utilisation de la normalisation à l'énergie. Néanmoins, cette normalisation reste limitée et pour aller plus loin, il est nécessaire d'utiliser des invariants de la couleur.

4.4.3 Invariants Couleur

Cette section s'est jusqu'ici limitée à l'étude de descripteurs d'images de luminance (souvent appelées images de niveaux de gris). Une extension naturelle de ces études consiste à utiliser des images de chrominance (RVB) pour inclure l'information couleur dans la description des imagettes. Cette extension est faite pour les descripteurs calculés par Analyse en Composantes Principales . La discrimination des filtres obtenus est beaucoup plus importante que celle des filtres de luminance. De plus, l'utilisation de la couleur a permis à de nombreux auteurs de définir ou d'améliorer leurs systèmes de reconnaissance. En particulier, le système de SWAIN et BALLARD [SB91] fondé sur une modélisation par histogrammes de couleur démontre la pertinence de la couleur pour la reconnaissance. Par ailleurs, l'usage de la couleur permet d'accéder à une invariance plus importante par rapport aux variations de l'éclairage. Ceci a été démontré par FINLAYSON [FF95, FCF96, FSC98] puis par l'auteur de cette thèse et SCHIELE [Col96, Sch97]. Une revue des différents invariants colorimétriques est donnée dans le rapport technique de GROS et al. [GMD+97]. L'utilisation de la couleur est une extension directe de cette thèse qui permettra une amélioration de la discrimination entre vecteurs de mesures locaux.

4.5 Conclusions

Dans ce chapitre, nous avons présenté une stratégie systématique d'évaluation de la sensibilité des descripteurs locaux. Cette stratégie a été utilisée pour évaluer l'évolution des vecteurs de mesures par rapport aux différentes perturbations du signal : le bruit lié à la chaîne d'acquisition, le bruit lié au filtrage des images et le bruit lié aux propriétés d'invariance des descripteurs par rapport à l'orientation et l'échelle. Globalement, cette étude de la sensibilité des descripteurs locaux permet de choisir le paramètre principal d'une recherche : le seuil S sur la distance qui valide ou rejette un appariement. Ainsi, nous obtenons un prédicat P de similitude entre vecteurs : $P(V_1, V_2) = (dist(V_1, V_2) < S)$. Ce prédicat est à la base de toute requête de recherche. Deux algorithmes de recherche peuvent être utilisés : il s'agit de la recherche de l'intégralité des points répondants au prédicat. Cet algorithme présente l'inconvénient de retourner dans certains cas peu discriminants un nombre de points très importants. L'autre algorithme est la recherche des k meilleurs voisins. Cette recherche est plus efficace mais présente l'inconvénient de ne plus garantir que tous les points répondant au prédicat sont retournés. Ceci dépend de l'environnement des points et donc des autres points de la base. Ces stratégies sont présentées plus précisément au chapitre suivant.

Chapitre 5

Apprentissage d'un modèle d'objet

Ce chapitre propose une étude du problème de la modélisation d'un objet par son apparence en utilisant des caractéristiques locales. Un objet 3D est représenté par une collection d'images formant un échantillonnage de la sphère des vues. Pour chaque image, nous proposons une modélisation fondée sur des caractéristiques locales étudiées au cours des chapitres précédents. Ces caractéristiques peuvent être mesurées en chacun des points des images et plusieurs stratégies de modélisation les utilisant sont étudiées dans ce chapitre.

Deux stratégies classiques de modélisation fondées sur des caractéristiques locales sont présentées dans la première section : d'une part, un détecteur de points d'intérêt peut être utilisé pour sélectionner un sous-ensemble de points de l'image avant l'apprentissage : ces points sont a priori discriminants pour la reconnaissance. D'autre part, une évaluation statistique des mesures locales par un histogramme multidimensionnel peut être utilisée. Dans ce cas, chaque image est représentée par un histogramme. Ce modèle est discriminant pour la reconnaissance malgré l'abandon de l'information structurelle. Cet histogramme permet aussi de sélectionner des points dont les caractéristiques sont peu fréquentes dans la base d'apprentissage et donc, a posteriori, discriminants pour la reconnaissance : il s'agit dans ce cas d'un détecteur de points discriminants.

La deuxième section propose une nouvelle modélisation qui consiste en un apprentissage structurel intégral des vecteurs de mesures des caractéristiques locales. L'apprentissage intégral permet une reconnaissance d'objets très robuste à l'occultation partielle et au bruit. En effet, cette modélisation est fortement redondante et permet un appariement de tous les points. Une illustration de cet apprentissage est donnée sur la figure 5.1. Une image est décomposée en une grille 2D d'imagettes recouvrantes entre elles. La grille correspondante est visualisée sur un sous-espace 3D de l'espace de représentation $\mathcal{A}$. Il s'agit d'une visualisation de la fonction plénoptique avec uniquement une variation des paramètres de position x et y, et la grille présentée quantifie la fonction sous-jacente. Dans le cadre de la reconnaissance, l'objectif est de mettre en correspondance une surface nouvelle (ou plutôt sa version quantifiée sur une nouvelle image) avec une surface

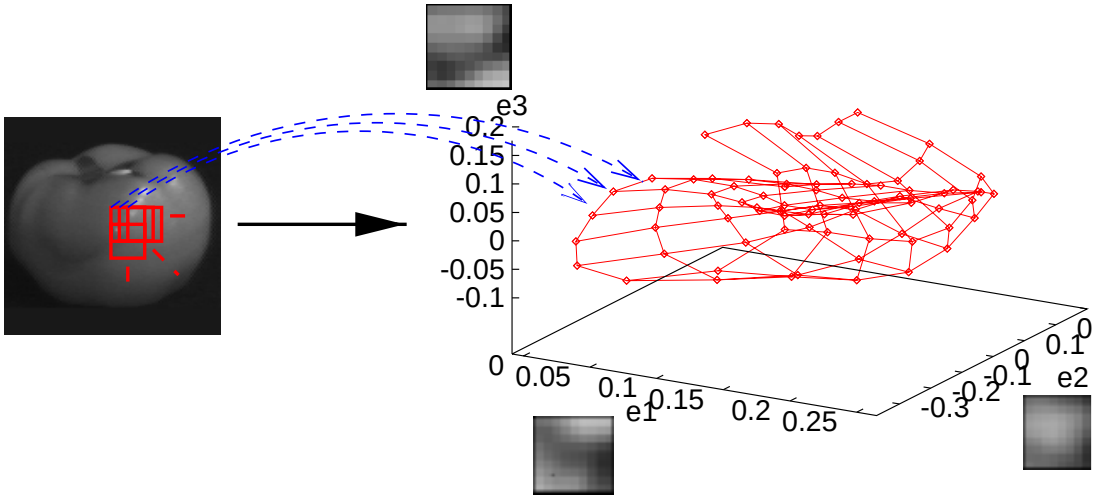


FIG. 5.1 – *Représentation d'une image comme une grille* 2D *dans un sous-espace* 3D *de* $\mathcal{A}$.

apprise. Cette mise en correspondance permet, simultanément, l'identification de l'objet et la détermination de sa pose dans l'image.

La dernière section aborde les problèmes liés à l'apprentissage intégral des caractéristiques locales : l'indexation dans un espace de grande dimension et la redondance de certains vecteurs de caractéristiques qui bloque leur reconnaissance.

5.1 Modélisation par caractéristiques locales

Plusieurs techniques sont utilisées pour capturer l'apparence d'objets observés par une caméra en utilisant des mesures de caractéristiques locales. Deux approches classiques sont présentées dans cette section. La première consiste à sélectionner dans l'image observée un ensemble de points caractéristiques (ou point d'intérêts) puis à projeter les imagettes centrées sur ces points sur l'espace de représentation $\mathcal{A}$. La deuxième approche consiste à représenter l'ensemble de l'image par une mesure statistique sur tous les points de l'image projetés sur $\mathcal{A}$. Cette mesure peut être représentée par un histogramme ou par une collection de moments à divers ordres.

5.1.1 Extraction de points discriminants

Une image peut être représentée par un sous-ensemble de ses points : ces points sont sélectionnés suivant des mesures sur leur environnement local. La modélisation d'une

image est effectuée suivant l'approche suivante :

1. Détection des points caractéristiques (ou point d'intérêts) de l'image.
2. Projection de ces points dans l'espace de description $\mathcal{A}$ (voir le chapitre 3 pour une description des différentes bases utilisées).
3. Apprentissage des vecteurs obtenus associés à un identificateur de leur image et position d'origine (voir section 5.2 pour une technique de stockage d'un grand nombre de ces projections).

Pour la phase de reconnaissance, les étapes (1) et (2) sont effectuées sur une nouvelle image, puis l'étape (3) est remplacée par la recherche des projections similaires préalablement apprises pour obtenir l'identification des objets observés sur l'image.

L'objectif des détecteurs est la sélection de points discriminants pour la reconnaissance. Pour cela, un détecteur doit présenter deux propriétés principales: la *Répétabilité* et l'*Unicité*. La répétabilité signifie que pour deux images d'un même objet, les mêmes points doivent être sélectionnés et l'unicité signifie que les points sélectionnés doivent se différencier le plus possible les uns des autres pour éviter des ambiguïtés de reconnaissance. Ces deux propriétés sont similaires aux propriétés de stabilité et de dispersion proposées à la section 3.2 pour l'évaluation des différentes bases de descripteurs locaux. Elles ont été proposées par IKEUCHI [OI97][1] pour sélectionner des imagettes informatives sur des images. La propriété de répétabilité est une qualité locale du signal image tandis que la propriété d'unicité est une qualité globale qui ne peut être évaluée que par rapport aux autres points de l'image ou de la base d'images.

Répétabilité : le détecteur doit sélectionner des points de façon la plus répétable possible : c'est à dire que deux applications successives du détecteur sur un même objet mais sous des conditions d'observation différentes doit sélectionner les mêmes points physiques de façon à permettre leur appariement pendant la phase de reconnaissance. Cette propriété capitale est difficile à obtenir de façon précise par rapport à des variations importantes du point de vue d'observation comme un changement d'échelle ou de l'éclairage.

Les points sélectionnés par ces détecteurs correspondent généralement à des zones très texturées des images. Ainsi, une grande partie des détecteurs sont fondés sur la matrice G des dérivées locales du signal. Le vecteur L^1 des dérivées premières est calculé de façon stable par l'utilisation de dérivées de Gaussiennes (voir section 3.4.1). $I(x,y)$ est la fonction image au point $(x,y)^T$.

$$L^1(x,y) = \left(\frac{\partial I(x,y)}{\partial x}, \frac{\partial I(x,y)}{\partial y}\right)^T = (L_x(x,y), L_y(x,y))^T \tag{5.1}$$

$$G(x,y) = L^1(x,y)L^1(x,y)^T = \begin{bmatrix} L_x^2(x,y) & L_x(x,y)L_y(x,y) \\ L_x(x,y)L_y(x,y) & L_y^2(x,y) \end{bmatrix} \tag{5.2}$$

1. En anglais, les termes utilisés sont "Local Goodness" et "Global Goodness"

SCHMID [Sch96] propose dans sa thèse une étude de plusieurs détecteurs et a sélectionné pour ses qualités de répétabilité celui défini par HARRIS [HS88]. Ce détecteur est une amélioration de celui proposé par MORAVEC [Mor81] consistant à évaluer la fonction d'auto-corrélation du signal. En particulier, l'aspect anisotropique de son approche a été résolu par l'utilisation des dérivées premières du signal (matrice G). Par ailleurs, TOMASI [ST94] a défini un détecteur basé sur la stabilité des caractéristiques détectées pendant un déplacement de la caméra. Celui-ci est similaire au détecteur de HARRIS car fondé sur la matrice G et sur ses valeurs propres. Une imagette est stable aux déplacements faibles de la caméra si la matrice G correspondante présente 2 propriétés :

- Ses composantes doivent être largement supérieures au niveau de bruit de l'image; cela se traduit par de grandes valeurs propres λ_1 et λ_2.

- Elle doit être bien conditionnée, ce qui signifie que les deux valeurs propres sont du même ordre de grandeur. Une valeur propre largement supérieure à l'autre signifierait une imagette avec une seul orientation très marquée qui ne permet un suivi précis que le long de cette orientation. Ceci se produit sur un front où la position perpendiculaire au front est précise mais la position sur la direction parallèle à ce front est floue.

Comme les valeurs des pixels sont bornés par le numériseur (255 en général), il suffit de vérifier que les 2 valeurs propres λ_1 et λ_2 sont supérieures à un seuil λ_s :

$$\min(\lambda_1, \lambda_2) > \lambda_s$$

Le détecteur de TOMASI a été appliqué sur trois images de la base de Columbia sur la figure 5.2. La figure montre les trois images, les images des valeurs minimales ($\min(\lambda_1, \lambda_2)$) puis la sélection des 10% des points les plus discriminants suivant le critère de TOMASI . Cette dernière série d'images montre principalement les contours des objets comme points discriminants, ceci n'est pas très positif car cela implique une segmentation correcte des objets par rapport au fond pour obtenir une reconnaissance de ces objets. D'autre part, un bord ou un coin n'est pas nécessairement discriminant par rapport aux autres objets de la base car il s'agit d'une caractéristique assez commune surtout dans le cas où l'orientation 2D n'est pas connue.

Un deuxième aspect de la répétabilité est la stabilité du calcul des descripteurs sur les points sélectionnés. Il est, en effet, raisonnable de sélectionner des points dont les descripteurs associés sont calculables de façon la plus stable possible. IKEUCHI [OI97], par exemple, propose d'évaluer la dérivée des caractéristiques des points sélectionnés par rapport à un bruit ajouté (voir problème de sensibilité des descripteurs locaux par rapport au bruit au chapitre 4).

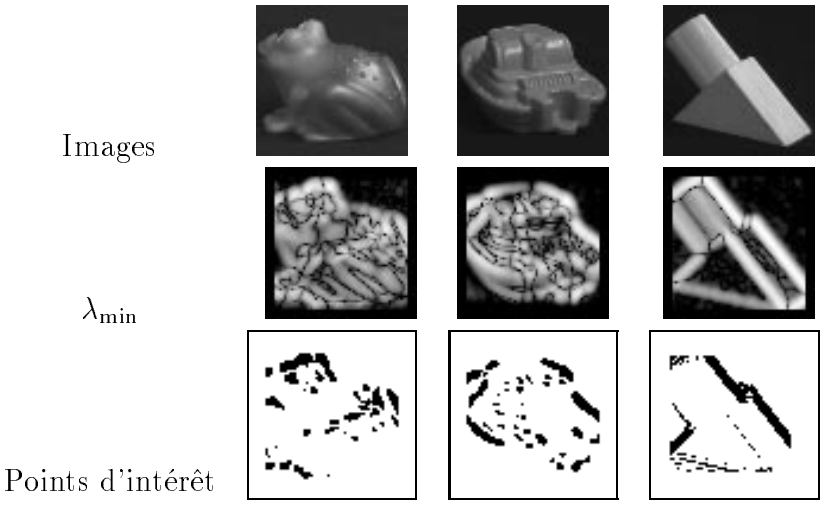


FIG. 5.2 – *Trois images de la base de Columbia(voir annexe A.1) associées aux images du détecteur de points d'intérêts de Tomasi (noir signifie "peu significatif" et blanc "très significatif") La valeur $\lambda_{\min}$ est la plus petite des deux valeurs propres de la matrice G des dérivées premières de l'image. La troisième ligne montre les 10% des points dont la valeur $\lambda_{\min}$ est la plus importante dans l'image, il s'agit des points sélectionnés par le détecteur de points d'intérêts.*

Unicité (ou discriminabilité) : le détecteur doit sélectionner des points de caractéristiques uniques ou, tout au moins, de caractéristiques assez rares de façon à limiter autant que possible les ambiguïtés. Cette notion est globale et ne peut être garantie par un détecteur local. IKEUCHI propose de supprimer les points dont les descripteurs sont similaires dans la base d'apprentissage. KRUMM [Kru97] propose de conserver les points localement uniques : aucun voisin proche ne ressemble au point évalué. Il est possible de gérer ces points similaires de façon plus fine en évaluant le nombre d'occurences et l'information effective contenue dans ces points par l'étude statistiques de l'occurence de tels points (voir section 5.1.2). De plus, il faut noter que ce critère global peut remettre en cause le choix des points à fort contraste comme point d'intérêt. En effet, certaines caractéristiques comme les coins apparaissent sur de nombreux objets et ne sont pas, pour cette raison, discriminants. Par contre, certaines imagettes de faible contraste peuvent être très discriminantes.

L'approche par points d'intérêt présente l'avantage important de permettre une forte compression de la base d'apprentissage en représentant chaque image modèle par un sous-ensemble de points. Cela permet, en particulier, l'utilisation de cette technique sur de très grandes bases d'objets. Par contre, l'inconvénient associé est la difficulté pour détecter des points d'intérêt de façon répétable. Souvent, deux images de points de vue proches d'un même objet ne font pas apparaître les mêmes points d'intérêts ou alors seul

un sous-ensemble des points d'intérêts est commun aux deux images. D'autre part, les performances de cette approche se dégradent beaucoup pour un objet trop ou pas assez texturé car le détecteur de points est submergé de points ou alors n'en a pas assez. Des objets constitués principalement par des cylindres ou des sphères sont assez difficiles à reconnaître. De plus, les zones de fort contraste risquent dans ce cas de ne pas correspondre à des caractéristiques de l'objet comme par exemple une zone de forte spécularité. Dans ces cas, la stratégie de reconnaissance doit être remise en cause en évitant de sélectionner des informations a priori discriminantes et en augmentant ainsi la redondance dans les modèles.

5.1.2 Modélisation statistique

La sélection de points particuliers implique une perte d'une partie de l'information contenue dans les images modèles. Celle-ci peut rendre difficile la reconnaissance. Par exemple, la présence de nombreuses fenêtres de couleur constante peut être en elle-même une information très importante pour la reconnaissance d'un objet. Cette sélection peut être remplacée par une évaluation statistique des projections dans l'espace de description $\mathcal{A}$. Cette technique a été initialement proposée par SWAIN et BALLARD [SB91] sur des images couleurs. L'espace de description est constitué de trois dimensions : les trois canaux de couleurs Rouge, Vert et Bleu. Une image couleur est représentée par un histogramme des couleurs présentes dans celle-ci. Cet histogramme est obtenu en parcourant les points de l'image : chaque point incrémente la case de l'histogramme correspondant à sa couleur. Ainsi, une image est modélisée par le nombre de pixels dont elle dispose dans chacune des couleurs. Cette seule information statistique des couleurs présentes et de leur quantité suffit à obtenir un système de reconnaissance d'images très efficace. Sur la figure 5.3, l'histogramme modèle de l'objet *Miel3* est présenté. La reconnaissance est

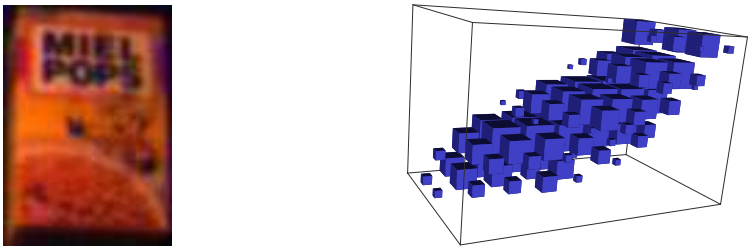


FIG. 5.3 – *Image Miel3 et histogramme associé.*

fondée sur la ressemblance entre l'histogramme d'une image modèle et l'histogramme de l'image de test. La caractéristique locale utilisée est le vecteur des trois coordonnées RVB de chaque pixel de l'image. Cette représentation est très robuste aux changements de point de vue. Elle très efficace en l'absence de variations d'éclairage mais peut être étendue au cas où l'éclairage varie par l'utilisation d'invariants de la couleur. Ainsi, une extension

de cette technique pour obtenir une robustesse par rapport aux variations d'éclairage est présentée dans [Col96, FF95]. Parallèlement à l'utilisation d'histogrammes pour évaluer les statistiques des couleurs ou des invariants de la couleur, des moments d'ordre 3 et 4 sur ces histogrammes ont été utilisés par HEALEY [HS94] et FINLAYSON [FCF96] pour représenter des images par un court vecteur de caractéristiques.

SCHIELE [SC96, Sch97] a généralisé la modélisation par histogrammes de couleurs à des histogrammes multidimensionnels des champs réceptifs. Il s'agit de modéliser les images par des histogrammes des mesures de caractéristiques locales comme les dérivées de Gaussiennes (voir paragraphe 3.4.1) ou des filtres de Gabor. Cette modélisation par histogrammes est très discriminante et a permis à SCHIELE d'obtenir un système de reconnaissance d'images très robuste. L'interprétation en termes de probabilités d'apparition

FIG. 5.4 – *Exemple d'une image Chocos10 et d'un histogramme Dx-Dy-Lap associé.*

des caractéristiques locales donne une reconnaissance robuste à l'occultation partielle des objets.

Dans le cadre de cette thèse, les statistiques sur la présence de vecteurs de caractéristiques sont disponibles facilement en utilisant la structure de données présentée à la section 5.2. Il est possible d'obtenir pour une quantification donnée, le nombre de vecteurs contenus dans une case de l'histogramme multidimensionnel correspondant. Ce nombre donne la fréquence d'apparition d'un vecteur de mesure dans la base d'apprentissage et sous l'hypothèse d'une équiprobabilité des objets et de leurs points, cet histogramme peut être interprété en termes de probabilité d'apparition de vecteurs de mesures : soit $h_q(\mathcal{M})$ le nombre de points dans la cellule de l'histogramme contenant le vecteur $\mathcal{M}$ pour une quantification de q bits par dimensions. La probabilité d'apparition de $\mathcal{M}$ pendant la phase d'apprentissage est :

$$P(\mathcal{M}) = \frac{1}{H} h_q(\mathcal{M})$$

avec H le nombres de points dans la structure de données. Cette probabilité a priori peut être utilisée pendant la phase de reconnaissance pour sélectionner les points les plus susceptibles d'être reconnus.

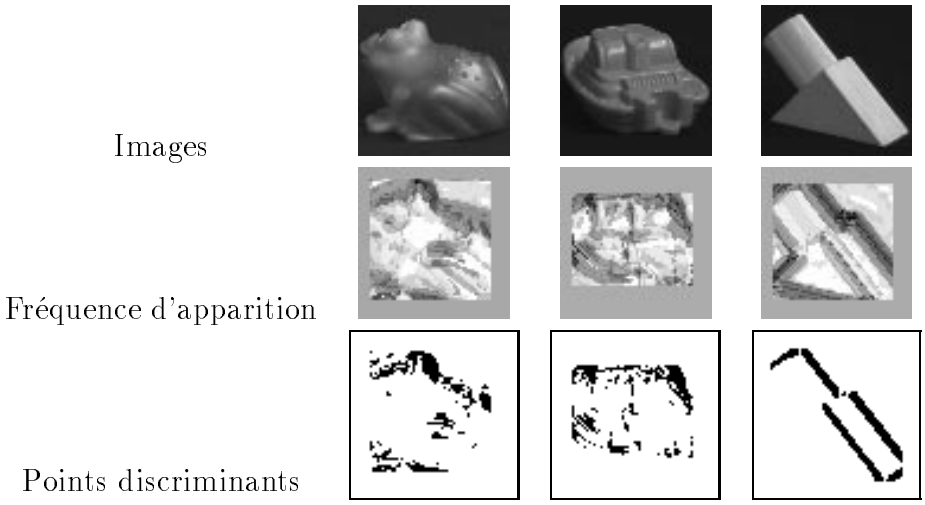


FIG. 5.5 – *Trois images de la base de Columbia(voir annexe A.1) associées aux images de fréquence (probabilité d'apparition) dans la base d'apprentissage (noir signifie peu fréquent et blanc très fréquent). La troisième ligne montre la sélection des 10% de points les moins fréquents des images par rapport à la base d'apprentissage. Ces points peu fréquents peuvent être utilisés, en priorité, pendant la phase de reconnaissance.*

La figure 5.5 montre les images de probabilité d'apparition pour trois images de la base de Columbia ainsi que les points sélectionnés qui correspondent au 10% des points les moins fréquents (de probabilité la plus faible) dans la base d'apprentissage.

Nous proposons dans cette thèse une extension des deux techniques présentées dans cette section : il s'agit de l'apprentissage des vecteurs de mesures issus de l'ensemble des points en conservant la structure spatiale de l'image. Pendant la phase de reconnaissance, les vecteurs de mesures locales fournissent des hypothèses d'appariements et l'utilisation de l'information structurelle permet de rejeter les faux appariements de façon très discriminante. Cette représentation est très redondante de façon à permettre une robustesse aussi importante que possible.

5.2 La variété de l'apparence

Dans cette thèse, la modélisation des images est fondée sur l'existence d'une fonction abstraite continue du niveau de luminance des points en fonction des différents paramètres de points de vue ou d'éclairage. Cette fonction est la fonction plénoptique présentée au chapitre 2. Elle est évaluée par sa projection sur différentes bases de descripteurs locaux (voir chapitre 3). La variation lente des paramètres de cette fonction décrit une forme dans l'espace des descripteurs $\mathcal{A}$. Il est possible de modéliser cette forme comme étant

une surface appelée variété de l'apparence[2]. Le paragraphe suivant montre deux exemples de synthèse de cette variété. Le premier est une synthèse complète pour une forme simple à partir d'une modélisation complète du processus d'acquisition d'image et le deuxième montre une approximation d'une variété par un sous ensemble de ces points.

Variétés de formes simples : BAKER et NAYAR [BNM98] ont synthétisé les variétés associées à des formes simples telles qu'une marche ou un front en se basant sur les fonctions analytiques de ces formes et une modélisation de la caméra. La figure 5.6 montre l'exemple de la variété correspondant à une marche projetée sur un sous-espace à trois dimensions. La variété présente deux paramètres ρ et θ définissant les différents points de vue possible d'une marche. Cette représentation permet de détecter la présence et la

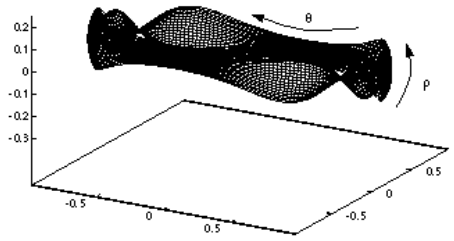


FIG. 5.6 – *La variété d'apparence d'une marche sur sous-espace* 3D *obtenu par* ACP *par* BAKER *[BNM98].*

pose d'une forme particulière par un accès direct aux paramètres ρ et θ sur la variété. Les auteurs étendent cette approche à des caractéristiques non synthétiques pour obtenir un système de reconnaissance robuste.

Un deuxième exemple est l'évaluation de la variété d'apparence et son approximation linéaire ou par splines. Cette approximation a été proposée, dans ses travaux effectués dans l'équipe PRIMA, par POURRAZ [Pou98] qui décrit une scène intérieure par l'image produite par une caméra : cette image se déforme continuement lorsque la caméra se déplace ou lorsque l'éclairage varie. A partir d'une discrétisation de l'espace des positions possibles de la caméra, une variété de l'apparence d'une scène est approchée. La figure 5.7 montre la variété unidimensionnelle induite par le déplacement d'un robot suivant une direction de l'espace et la projection des images obtenues sur l'espace d'apparence. L'objectif de ce travail était, ici, à partir de l'évaluation de la variété de pouvoir trouver la position effective du robot sur de nouvelles images (voir chapitre 7). La représentation est une forme d'interpolation entre les points de vues observées. La forme de la trajectoire dans l'espace d'apparence ne permet pas une extrapolation à des points extérieurs.

2. Le terme variété de l'apparence est la traduction de l'anglais *Appearence Manifold* proposé par NAYAR

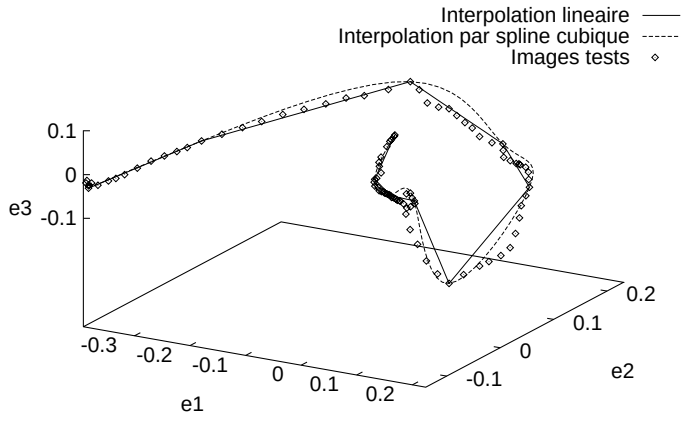


FIG. 5.7 – *Visualisation, dans un sous-espace des Composantes Principales de dimensions 3, de la représentation paramétrique de l'apparence de la scène et des projections des images tests. Images d'apprentissage : tous les 20 cm. Images de test : tous les 2 cm.*

Représentation ponctuelle : Cette section a montrée deux exemples de représentation des variétés d'apparence. Ces représentations sont difficiles et coûteuses à obtenir mais il est, par contre, facile d'obtenir un échantillonnage de la variété en faisant l'hypothèse que celui–ci est suffisamment dense pour garantir que tout nouveau point sera situé à distance faible de l'un des points de cet échantillonnage. On peut remarquer qu'il s'agit d'une modélisation par interpolation. Seuls des points intermédiaires et proches de ceux appris peuvent être retrouvés. La section suivante décrit une structure de données permettant cette représentation ponctuelle. Les algorithmes et les problèmes liés à cette représentation sont étudiés.

5.3 Structure de données pour une représentation ponctuelle de la variété de l'apparence

Dans le contexte de la reconnaissance d'objets 3D par une technique fondée sur la mesure de caractéristiques locales : (1) un objet est représenté par une collection d'images représentant une discrétisation de la sphère des vues de l'objet. (2) Une image représentant un point de vue est décomposée en imagettes recouvrantes. (3) Chacune des ses imagettes est représentée par un vecteur de mesures $\mathcal{M}$ de m coordonnées dans le sous-espace de représentation $\mathcal{A}$. Une image est représentée par une grille 2D de vecteurs $\mathcal{M}$ et un objet par une famille de grilles 2D.

La reconnaissance est obtenue en associant une imagette nouvelle aux variétés les

plus proches dans $\mathcal{A}$ ce qui correspond pour une représentation ponctuelle à trouver les descripteurs $\mathcal{M}$ les plus proches d'une mesure représentant une nouvelle imagette. La reconnaissance est obtenue sous l'hypothèse de répétabilité de l'évaluation des vecteurs de mesures $\mathcal{M}$.

Dans ce contexte, l'objectif de cette section est de proposer une structure de données et des algorithmes permettant d'effectuer les opérations suivantes :

- Stocker l'ensemble de couples ($\mathcal{M}$; Id) des images modèles des objets dans la base d'apprentissage avec Id un identificateur (objet, point de vue, position) de l'imagette se projetant sur le vecteur de mesures $\mathcal{M}$.

- Trouver les couples ($\mathcal{M}$; Id) de la base d'apprentissage qui sont "proches" d'un nouveau vecteur de mesures.

- Vérifier la présence d'un couple "compatible" avec un nouveau couple ($\mathcal{M}$; Id) dans la base d'apprentissage. La notion de compatibilité intègre la similitude des descripteurs ainsi que la similitude des identificateurs. Cette notion est développée plus précisément dans le chapitre 6.

L'objectif présenté nécessite de définir le prédicat "proche" indiquant si deux descripteurs sont à mettre en correspondance. Ce prédicat est un seuillage sur une distance de Mahalanobis (voir chapitre 4). Le résultat d'une recherche est l'intersection entre une m –boule centrée sur le vecteur de mesures $\mathcal{M}$ cherché avec l'ensemble des vecteurs de la base d'apprentissage. Plusieurs algorithmes sont possibles : un algorithme de recherche exhaustive dans la m–boule ou une recherche des k plus proches voisins du vecteur (voir paragraphe 5.3.1).

Pendant la phase de recherche, la valeur de la distance entre le vecteur de test et le vecteur modèle n'est utilisée que pour la comparer au seuil $\mathcal{S}$ mais, dans la suite, la valeur de la distance donne aussi une information sur la fiabilité de l'appariement des vecteurs correspondants (voir section 6.1.2).

Le problème étudié ici est l'indexation de points dans un espace de grande dimensionnalité. Ce problème a été largement étudié par des spécialistes du domaine des bases de données : les structures de *B-tree*, de *R-tree* ou de *sparse distributed memory* peuvent être utilisées. Dans le domaine de la vision, LAMIROY [Lam98] a étudié, dans le 5^{e} chapitre de sa thèse, des algorithmes d'indexation ainsi que leur complexité. NAYAR [NN97] propose un algorithme simple basé sur des projections sur les dimensions pour effectuer cette recherche des plus proches voisins sous un seuil. Ce sujet n'est pas le domaine d'étude de cette thèse et n'est pas développé ici. Néanmoins, pour permettre l'étude de la reconnaissance d'objets sous une telle approche, une structure d'indexation hiérarchique simple a été implémentée : elle est présentée, brièvement, dans le paragraphe suivant avec ses caractéristiques.

5.3.1 Indexation dans un espace de grande dimension

Ce paragraphe présente la technique d'indexation utilisée pour cette thèse. Les algorithmes et la structure de données utilisés sont développés plus précisément en annexe B.

Structure de donnée : Par simplicité, une structure de données arborescente a été choisie. Son principe consiste à diviser successivement et statiquement les axes des différentes dimensions de l'espace en quatre sections de longueurs égales. Il s'agit d'un arbre dont chaque noeud a quatre fils. Une feuille de cet arbre est une liste des points présents dans cette partie de l'espace. La structure est développée au fur et à mesure que l'ajout de nouveaux points forme des listes de points trop longues dans les feuilles. La figure 5.8

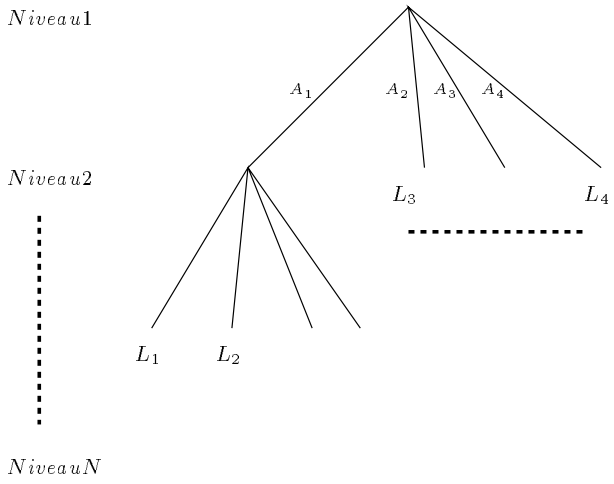


FIG. 5.8 – *Exemple d'un arbre d'indexation.*

illustre un arbre de représentation. Les feuilles L_i sont des listes assez courtes de points. De façon à séparer suffisamment les données, le sous-arbre A_1 est divisé jusqu'à la deuxième dimension tandis que les sous-arbres A_2, A_3 et A_4 sont des feuilles divisées uniquement suivant la première dimension de l'arbre.

Algorithme d'ajout d'un nouveau point à l'arbre : Pour ajouter un nouveau point, il suffit de suivre les branches de l'arbre jusqu'à obtenir une feuille, puis d'ajouter ce point à cette feuille. Si le nombre de points de cette feuille dépasse un seuil préalablement fixé, cette feuille est divisée suivant la dimension suivante. Aucun choix de la dimension la plus adéquate à diviser n'est effectué et cette division ne sépare pas optimalement les points de cette feuille dans le cas général. Néanmoins, pour la base de descripteurs obtenue par ACP, cette division par dimensions successives de l'espace est optimale car les dimensions sont classées par ordre de variance décroissante.

Algorithme de recherche des points proches d'un nouveau point en utilisant une stratégie de "Branch and Bound" : La présence de bruit sur les mesures implique que les points de la base d'apprentissage correspondants au nouveau point $\mathcal{M}$ ne sont pas uniquement dans la même feuille de l'arbre. Il faut effectuer un parcours de toutes les feuilles susceptibles de contenir un élément appartenant à la m–boule de recherche centrée sur $\mathcal{M}$. Il peut être utile d'explorer jusqu'à trois des quatre branches d'un noeud. Une recherche naïve entraîne l'exploration de 3^m feuilles dans le cas d'un arbre complet. Néanmoins, l'utilisation d'une technique type "Branch and Bound" permet de n'accéder qu'aux feuilles qui intersectent effectivement la m–boule. Pour cela, lors de la recherche, en chaque noeud visité, une distance optimiste entre $\mathcal{M}$ et le noeud est évaluée. Si cette distance dépasse le seuil de recherche, le noeud est rejeté.

Cet algorithme est lent si de très nombreux points sont susceptibles d'être sélectionnés. Pour limiter cet inconvénient, il est possible de définir un paramètre K fixant le nombre maximum de points pour une requête. Ce paramètre K est utilisé pour une recherche des K plus proches voisins du point $\mathcal{M}$ à apparier. Dés que K points sont trouvés, le seuil de recherche est abaissé à la distance avec le kième point trouvé puis, est mis à jour en fonction des nouveaux points trouvés. Cette mise à jour évite de parcourir des feuilles dont la distance est plus importante que la distance du kième élément.

Quelques remarques sur le coût de cette structure et des algorithmes

- Coût Mémoire : l'ensemble des points est stocké en mémoire vive. La mémoire disponible est la facteur limitant de cette approche. Le coût ajouté par la structure elle-même par rapport à un stockage direct des données est faible. En effet, le nombre de noeuds dans l'arbre est négligeable devant le nombre de feuilles qui est borné par le nombre de points stockés mais reste bien inférieur à celui-ci. Globalement, le surcoût dû à la structuration des données apprises est faible et augmente très faiblement avec l'ajout de nouveaux points.

- Temps de recherche : le temps nécessaire pour effectuer une recherche dépend linéairement du nombre de feuilles évaluées. Ce nombre de feuilles nécessitant d'être explorées pendant la recherche peut, dans le cas d'un arbre complet, être évalué par une méthode de Monte-Carlo. La figure 5.9 présente des courbes représentant le nombre de feuilles à explorer en fonction du seuil de recherche utilisé. Pour évaluer ce nombre moyen, il suffit de générer aléatoirement un point dans un cube de m dimensions puis de compter le nombre de cubes voisins que la m–boule centrée sur ce point intersecte. L'algorithme de recherche optimisé par la technique du branch and bound n'explore que les noeuds et feuilles nécessaires, soit un nombre de feuilles égal au nombre de cubes intersectée. La figure présente les courbes correspondants à des espaces de 5 à 20 dimensions. L'axe des abscisses donne le rayon de m–boule de recherche pour une distance euclidienne. La valeur 1.0 correspond à la longueur

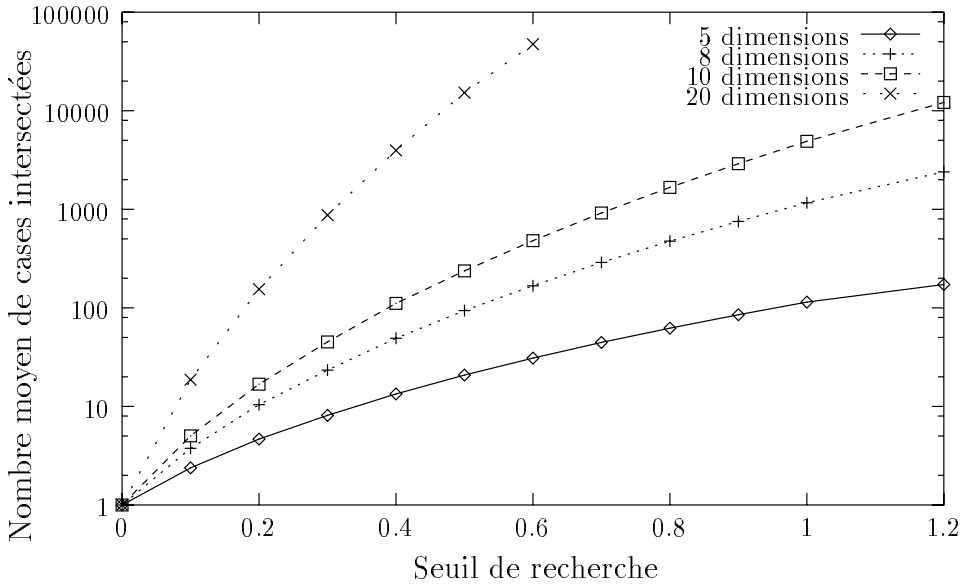


FIG. 5.9 – *Nombre de cases intersectées par une m–boule en fonction du seuil de recherche et du nombre de dimensions.*

d'un coté de l'hypercube. En pratique, les seuils de recherche utilisés sont de l'ordre de la moitié de la largeur d'une cellule, et la figure montre que le nombre moyen de feuilles intersectées pour un arbre de dix dimensions est d'environ 200. Les arbres utilisés n'étant pas complets, le nombre de feuilles recherchées est expérimentalement de l'ordre de 30 pour une base de 300.000 points. Cette structure de données a permis d'évaluer notre technique de reconnaissance pour des bases de modèles constituées de plus de 3 millions de points.

L'augmentation très importante du nombre de feuilles avec le nombre de dimensions est à modérer par l'aspect d'autant plus creux de l'espace lié au nombre de dimensions. Le nombre de cases augmente aussi de façon exponentielle. Expérimentalement, la structure est apparue suffisamment efficace pour l'objectif de reconnaissance.

En conclusion, l'étude de la structure obtenue fait apparaître un problème important : certaines projections sont fortement redondantes et se produisent plusieurs milliers de fois. Leur discriminabilité en terme de position est nulle, par contre leur interprétation en terme d'histogramme peut être informative. La figure 5.10 montre le nombre de projections par feuilles de l'arbre de représentation. Il est possible d'observer sur cette figure que très peu de feuilles sont très pleines : elles correspondent à des imagettes de niveau de luminance constant. L'utilisation de la couleur apparaît très profitable car les points sont, dans ce cas, mieux répartis sur l'espace de projection $\mathcal{A}$. Ainsi, le coût de la recherche est faible dans la plupart des cas sauf si l'algorithme doit accéder aux feuilles non discriminantes.

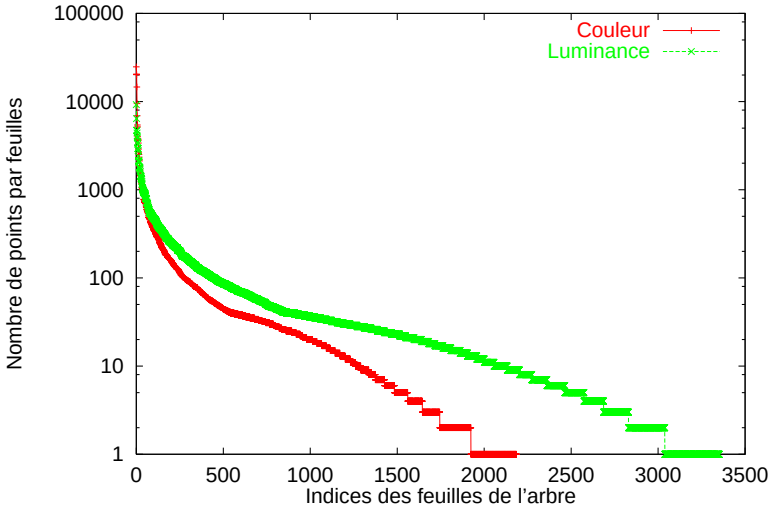


FIG. 5.10 – *Nombre de points par feuille de l'arbre classés de façon décroissante pour une base 160 images (environ 270.000 points) sur un espace $\mathcal{A}$ obtenu par une ACP en images de luminance et en images couleur.*

Le paragraphe suivant pose ce problème de redondance et envisage certaines solutions.

5.3.2 Redondance des projections

Le choix de l'apprentissage complet de la variété de l'apparence pose un problème de redondance de l'information stockée. Il est possible de distinguer deux types de redondance : (a) la redondance liée à la similarité importante entre points voisins d'une image et (b) la redondance des projections très courantes sur les images observées (imagettes de luminance constante par exemple).

Ce premier type de redondance des projections dans la structure de recherche implique pendant la phase de recherche que plusieurs imagettes équivalentes sont reconnues simultanément. En effet, les imagettes voisines dans une image de la base d'apprentissage se projettent sur des points voisins dans l'espace $\mathcal{A}$. La figure 5.11 illustre le problème. Cette figure montre sur un sous-espace à deux dimensions de $\mathcal{A}$ que de nombreux points d'une même variété peuvent être mis en correspondance avec un point X : tous ces points représentent la même caractéristique de l'objet observé. Le rayon $\mathcal{S}$ de la m–boule de recherche est le seuil de recherche. Sur l'exemple, une solution consiste à détecter les points intérieurs à la m–boule et à choisir le plus proche de X.

Le deuxième type de redondance (b) correspond à la redondance physique de certaines caractéristiques : les imagettes de luminance constante par exemple. Une recherche sur de telles imagettes donne de nombreuses correspondances qui ne permettent pas une reconnaissance directe. Le recherche de ces imagettes présente, de plus, l'inconvénient

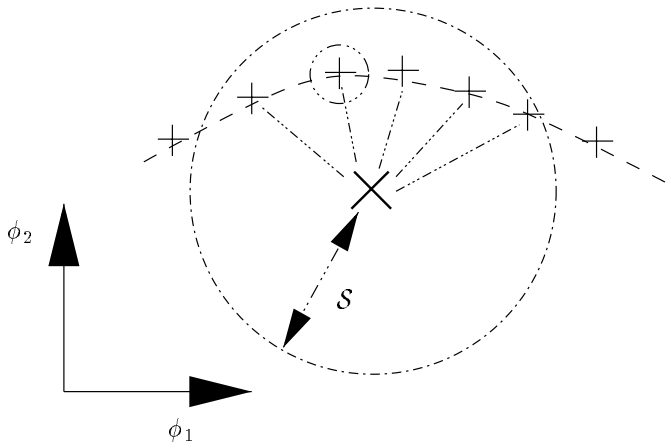


FIG. 5.11 – *Redondances des solutions sur une sous-espace* 2D *de* $\mathcal{A}$.

d'un temps de réponse important lié au nombre d'hypothèses à générer puis à trier pour supprimer les hypothèses redondantes. Le nombre de réponses est trop important pour atteindre une hypothèse d'objet fiable mais, par contre, un algorithme de vérification d'hypothèses peut utiliser cette information pour confirmer ou rejeter une hypothèse (voir chapitre suivant). Il n'est donc pas intéressant de supprimer complètement l'information correspondant à ces imagettes. Ce type de redondance correspond à la propriété d'unicité des points proposée au paragraphe 5.1.1.

Sélection de la meilleure hypothèse en ligne La gestion du problème (a) peut se faire aisément en ligne pendant la recherche. Pour ceci, un algorithme souple de sélection d'une bonne hypothèse est proposée dans ce paragraphe. La sélection de l'hypothèse correcte est apparue assez évidente sur le graphe présenté sur la figure 5.11 mais la forme de la grille 2D modélisant l'objet peut être très variée et entraîner des problèmes de sélection plus difficile.

Un prédicat $MemePoint(H_1, H_2)$ définissant si une hypothèse H_1 est équivalente à une hypothèse H_2 doit préalablement être défini. Ce prédicat est un cas trivial du prédicat de Compatibilité entre recherches successives proposé au chapitre 6. Une hypothèse H est un triplet (Id, p, d) (voir aussi section 6.1.1) où Id est un identificateur de l'image modèle (un objet et un point de vue), $p = (x, y, \theta, \sigma)$ est la position dans l'image modèle et d la distance entre la projection correspondant à cette hypothèse et la projection recherchée. Dans le cas particulier où Id_1 et Id_2 sont les mêmes et où les paramètres θ et σ sont inchangés, le prédicat s'écrit simplement :

$$MemePoint(H_1, H_2) = (Dist(p_1, p_2) < Seuil)$$

De façon plus générale, la connaissance des transformations affines entre les différentes

images modèles peut permettre de définir ce prédicat pour des hypothèses provenant d'images modèles différentes mais ayant la même apparence et entraînant donc une double reconnaissance.

L'algorithme proposé sur la table 5.1 consiste à détecter les composantes connexes parmi l'ensemble des hypothèses détectées, puis à renvoyer les meilleurs représentants de ces composantes au sens de la distance avec la projection recherchée. L'algorithme présenté est appelé pour chacun des objets présents dans la liste des hypothèses. Il utilise un tri préalable des listes d'hypothèses par distance croissante pour garantir que le représentant d'une composante connexe est le meilleur possible. Il faut remarquer que cet algorithme trouve des composantes connexes sous la condition du prédicat $MemePoint(H_1, H_2)$ et ne garantit pas que chaque composante connexe effective soit représentée par un point unique ni qu'une composante ne corresponde à une unique caractéristique physique.

```
Soit L_hypo = (H_1, H_2, ...)
// Liste des hypothèses trouvées triées par ordre de distance croissante
Fonction Sélectionner(L_hypo)
Début
    Soit L_cc = () // Liste des centre des composantes trouvées
    Pour chaque hypothèse H_i de L_hypo, faire :
        drapeau_iter = faux
            Pour chaque composante C_j de L_cc, faire :
                Si MemePoint(H_i, C_j) alors
                    drapeau_iter = vrai
                    ajouter H_i à la composante C_j
        Si drapeau_iter = faux alors
            ajouter une nouvelle composante {H_i} à la fin de L_cc
    Retourner L_cc // Liste triée des hypothèses sélectionnées
Fin.
```

TAB. 5.1 – *Algorithme de sélection des hypothèses les plus représentatives d'un objet (pseudo pascal).*

Cet algorithme a une complexité importante : $o(nm)$ avec n le nombre d'hypothèses dans L_{hypo} et m le nombre d'hypothèses sélectionnées. Ceci implique un temps de calcul important pour des listes d'hypothèses de grandes tailles. Ce cas se produit pour les imagettes peu discriminantes (cas (b)). En pratique, pour éviter un ralentissement de la reconnaissance par ces imagettes redondantes non discriminantes, il est possible de rejeter ces imagettes pendant la phase de recherche ou hors-ligne.

Sans l'étape de sélection, la reconnaissance est souvent noyée sous de très nombreuses hypothèses équivalentes qui ralentissent le processus de reconnaissance complet : les tech-

niques de votes ou de prédiction–vérification proposées au chapitre suivant sont très ralenties par la redondance des hypothèses trouvées. Pour des paramètres identiques, les taux de reconnaissance restent similaires mais avec une grande réduction du coût. Expérimentalement, le gain obtenu est de l'ordre de 50%. Néanmoins l'algorithme ralentit beaucoup la reconnaissance dans le cas de feuilles très pleines, il est donc nécessaire de rejeter les listes trop longues et de plus, une stratégie hors-ligne permet de limiter ces inconvénients.

Suppression de la redondance hors-ligne : Après l'apprentissage, il est possible de détecter les feuilles de l'arbre qui ont trop d'éléments (voir figure 5.10 en début de section). Les feuilles ainsi détectées peuvent être supprimées de la base car globalement non discriminantes : du point de vue des différents algorithmes de reconnaissance envisagés, une absence de réponse ou trop de réponses donnent le même résultat : rejet du point considéré. Ainsi, un premier algorithme de réduction de la redondance consiste simplement à effacer les feuilles de l'arbre qui contiennent plus que M projections avec M de l'ordre de 10 000. Cet algorithme implique un risque de perdre certains appariements de façon non contrôlée. Un deuxième algorithme permet de contrôler les projections qui sont supprimées de la structure.

Cet algorithme consiste à supprimer de la base les points redondants tout en vérifiant qu'un représentant des points supprimés est toujours présent dans la base. Pour chaque projection de la base, une recherche est effectuée, puis parmi les résultats correspondants au même objet seul un représentant est conservé. L'information du nombre de points regroupés est conservée de façon à pouvoir les utiliser dans un cadre de vérification simple d'hypothèses.

La figure 5.12 présente une comparaison du score de reconnaissance par une imagette pour une base extraite de la base de Columbia (100 objets pour 4 points de vue d'apprentissage) soumise à plusieurs post-traitements. Les quatre colonnes correspondent à quatre post-traitements de la base des modèles :

- La colonne (1) correspond à la base des modèles sans post-traitements.
- La colonne (2) donne les résultats pour la base après application de l'algorithme conservatif (chaque projection conserve un représentant dans la nouvelle base).
- Pour la colonne (3), l'ensemble des feuilles de plus de 15.000 points ont été supprimés de la base initiale.
- Pour la colonne (4), la suppression des feuilles de plus de 15.000 points a été effectuée sur la base (2).

La vitesse et la consommation mémoire sont directement liées au nombre de points

dans chacunes des bases. Ceux-ci sont donnés dans la table ci dessous :

base	(1)	(2)	(3)	(4)
Nombre de projections	712444	537531	443870	424627
Pourcentages	100.0%	75.4%	62.3%	59.6%

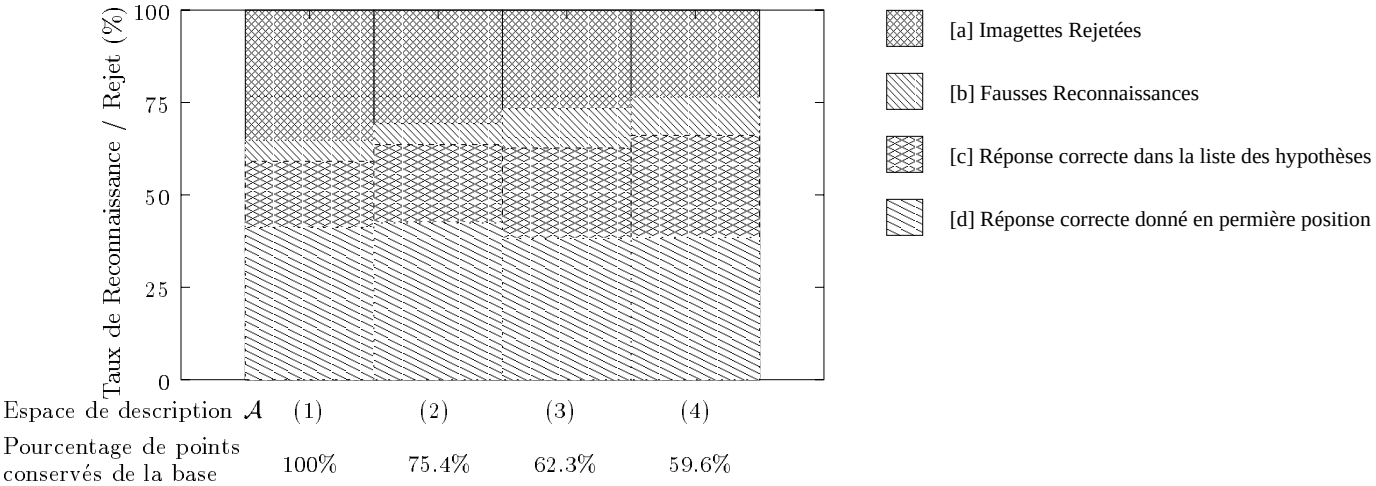


FIG. 5.12 – *Comparaison entre plusieurs post-traitements de la base de modèles.*

Les résultats de la reconnaissance [a], [b], [c] et [d] ont été expliqués à la section 3.1 (page 38). Il transparaît que les résultats sont relativement équivalents entre les différentes bases. Le meilleur taux de reconnaissance est obtenu pour la base (2) car des projections correctes étaient perdues dans le cas (1) par la redondance qui entraînait des rejets. Les bases (3) et (4) donnent des résultats un peu inférieurs mais pour un coût largement plus faible. Dans l'ensemble, ces post-traitements permettent une accélération de la recherche et un gain en mémoire sans perte en reconnaissance. L'usage des ces post-traitements apparaît donc nécessaire pour une utilisation optimale de la technique proposée dans cette thèse. Il est envisageable, pour l'accélération de l'apprentissage, d'effectuer une élimination de la redondance pendant la phase d'apprentissage.

Cette stratégie de suppression des points redondants ou peu discriminants permet de contrôler la taille de la base d'apprentissage. Il est possible de diminuer celle-ci en s'autorisant une perte d'informations. Dans ce cas, notre technique se rapproche des stratégies de modélisation par extraction de points d'intérêts. La différence étant le choix de ces points par leur discrimination dans la base.

5.4 Conclusions

La modélisation d'objets par l'apprentissage de la totalité des vecteurs de mesures disponibles sur l'image en associant à l'information spatiale entre eux est apparue possible

malgré le coût mémoire important. Celui-ci est acceptable car les progrès techniques nous fournissent actuellement des stations de travail dotées d'une mémoire suffisante. Cette modélisation a été appliquée avec succès sur une base de 1800 images représentée par 3 millions de vecteurs. La redondance des modèles obtenus permet d'obtenir une reconnaissance très robuste à une large gamme de perturbations des images comme le bruit de la chaîne d'acquisition ou l'occultation partielle.

L'information structurelle (position relative des points entre eux) a été conservée (par opposition aux histogrammes) et l'ensemble de l'apparence observée a été capturée (par opposition aux techniques à base de points d'intérêts). La structure de données proposée permet un apprentissage des objets suivant la modélisation envisagée. L'ensemble des résultats expérimentaux de cette thèse est fondé sur l'utilisation de cette structure de données. L'étude de celle-ci a permis de détecter des problèmes de densité très importants en certains points de l'espace des caractéristiques. Cette densité a été réduite en supprimant les vecteurs équivalents et les vecteurs non discriminants vis-à-vis de la base d'apprentissage.

Chapitre 6

Reconnaissance d'objets

Cette thèse traite du problème de la reconnaissance d'objets 3D indépendamment du point de vue de la caméra. L'approche choisie consiste à représenter un objet par une collection d'images formant un échantillonage de la sphère des vues. Le problème de la reconnaissance d'objets est traité comme un problème d'appariement d'une image avec une ou plusieurs images modèles.

L'utilisation de caractéristiques locales et l'absence de modélisation globale des objets permet d'accepter pour cette technique une large gamme d'objets : des objets polyhédriques et non–polyhédriques, des objets déformables et des objets non compacts.

Une technique d'appariement de vecteurs de mesures de caractéristiques locales a été présentée dans les chapitres précédents. Cette technique permet d'associer à un point (et son environnement local) d'une image, un ensemble de points issus des images modèles. Les appariements trouvés correspondent à des fenêtres englobants les points fortement similaires : très peu de faux appariements sont observés. Les vecteurs de mesures locales utilisés forment une approximation précise et concise de l'espace des imagettes possibles et en l'absence d'occultation partielle, sur les bases d'images présentées, un minimum de 50% des points donnent un appariement correct. Un appariement entre deux points est donné avec une évaluation de sa qualité (distance entre les vecteurs de mesure) et avec les paramètres de position, orientation et échelle des deux points qui permettent d'évaluer la pose du nouveau point par rapport au point de l'image modèle dans le cadre d'une approximation par une similitude.

La reconnaissance d'objets doit, pour être robuste, se fonder sur l'utilisation de plusieurs caractéristiques locales simultanées. Pour cela, plusieurs questions se posent :

– Comment évaluer une hypothèse basée sur les résultats de plusieurs recherches ?

  1. Compatibilité entre recherches.
  2. Évaluation d'une similitude 2D.
  3. Définition d'un score permettant de classer les hypothèses d'objet.

- Comment sélectionner les points où évaluer puis rechercher les caractéristiques locales ?
- Quel algorithme de reconnaissance pouvons-nous utiliser ?
  1. Approche directe : vote ou transformée de Hough.
  2. Approche fondée sur le principe de prédiction–vérification.

Les questions proposées dans cette introduction sont évaluées successivement dans ce chapitre de façon à obtenir un système de reconnaissance complet.

6.1 Évaluation d'une solution à base de recherches multiples

Pour obtenir une reconnaissance robuste, il est indispensable d'apparier simultanément plusieurs points de l'image. L'évaluation et la comparaison entre hypothèses fondées sur plusieurs appariements est difficile : il faut classer les hypothèses par ordre de vraisemblance ce qui peut être fait par l'évaluation d'un score pour chaque hypothèse. Dans un premier temps, il faut regrouper des résultats cohérents entre eux issus de recherches différentes. Nous proposons deux stratégies : mesurer la compatibilité entre couples d'appariements ou évaluer pour chaque appariement une similitude puis regrouper les similitudes proches. Dans les deux cas, l'évaluation de la compatibilité entre hypothèses est fondée sur une approximation : l'existence d'une similitude 2D entre les deux imagettes. Ce choix n'est pas juste de manière générale et l'évaluation de la similitude apparaît dangereuse dans le cas de transformations perspectives. Par contre, la technique évaluant les hypothèses deux à deux limite l'approximation à ce couple de points et non pas à l'image entière et paraît plus fiable. Après le problème du regroupement des hypothèses compatibles apparaît le problème de l'évaluation du score associé à une hypothèse et de leurs comparaisons.

6.1.1 Compatibilité entre recherches

Cette section aborde le problème de la définition d'un prédicat décidant de la compatibilité entre deux appariements. Le premier paragraphe illustre le problème sur un exemple puis les paragraphes suivants montrent une technique générale pour évaluer cette compatibilité sous l'hypothèse d'une similitude.

Exemple de cohérence spatiale Les chapitres précédents ont proposé une technique de reconnaissance qui permet d'associer à une imagette extraite d'une image une liste des imagettes modèles semblables à celle-ci. Chaque imagette est identifiée par son origine

(objet, image, position). La position est un quadruplet (abscisse, ordonnée, orientation, échelle).

Une technique directe de reconnaissance d'objets consiste à extraire de chaque hypothèse un identificateur de l'objet et de sa prise de vue, puis à en incrémenter le score. L'appel successif de la technique de recherche pour plusieurs imagettes permet de déterminer l'objet de la base le plus similaire à l'image testée. Cette technique basée sur un vote pose, ici, le problème que deux hypothèses géométriquement incohérentes entre elles peuvent renforcer le score de reconnaissance de celui-ci et entraîner de fausses reconnaissances. Il apparaît donc intéressant d'ajouter un critère de cohérence spatiale entre les hypothèses pour supprimer ce problème et rendre ainsi la reconnaissance plus discriminante. Ce critère doit tenir compte des transformations éventuelles subies par l'objet entre l'image modèle et l'image observée : rotation 2D, translation, zoom, perspective ou déformation de l'objet.

La figure 6.1 montre une image de l'objet 89 de la base de Columbia (voir Annexe A.1). Cette image n'a pas été utilisée pendant la phase d'apprentissage. Une recherche des

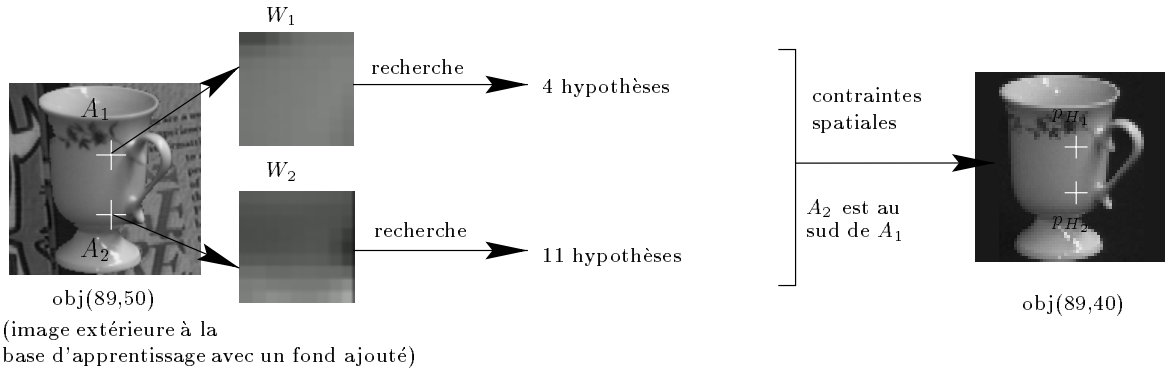


FIG. 6.1 – *Exemple d'une recherche sur deux imagettes d'une image extérieure à la base d'apprentissage. Chacune des deux imagettes donne une liste d'hypothèses. L'utilisation de la cohérence spatiale entre les deux imagettes permet de rejeter l'intégralité des hypothèses incorrectes et permet de sélectionner l'image modèle la plus proche de l'image de test. La mise en correspondance est faite ici en utilisant un e base de filtre* ACP *de 10 dimensions et de taille* 9×9 .

imagettes W_1 et W_2 extraites aux points A_1 et A_2 est effectuée. La recherche de l'imagette W_1 dans la base d'apprentissage transmet 4 hypothèses vraisemblables de solutions et celle de W_2 en donne 11. Ces deux seules imagettes autorisent plusieurs objets possibles, mais l'information supplémentaire que le point A_2 est au sud du point A_1 avec une distance $d = \| \vec{A_1 A_2} \|$ mesurable directement sur l'image restreint les objets possibles à un point de vue appris de l'objet 89 de la base et permet donc sa reconnaissance. Formellement, la position relative des points recherchés peut être vérifiée en comparant le vecteur déplacement $\vec{A_1 A_2}$ reliant les deux points cherchés et le vecteur déplacement $\vec{p_{H_1} p_{H_2}}$ reliant les hypothèses.

Les vecteurs déplacements $\vec{A_1A_2}$ et $\vec{p_{H_1}p_{H_2}}$ doivent être égaux. La position relative des imagettes cherchées a permis une reconnaissance très discriminante. En conclusion, il apparaît nécessaire de définir une mesure de compatibilité entre deux recherches.

Mesures de compatibilité entre recherches

Pour une reconnaissance très discriminante, il est possible d'inclure une information de cohérence spatiale entre hypothèses pour la reconnaissance. Cette information peut être donnée par une mesure de compatibilité entre deux hypothèses. Cette mesure est un prédicat logique qui répond `vrai` si deux hypothèses sont compatibles et `faux` sinon. Cette section propose, après la définition de quelques notations, deux mesures de compatibilité qui peuvent être utilisées pendant la phase de reconnaissance.

Quelques Notations :

- La recherche d'une imagette extraite d'une image ou *requête de recherche* est notée $R = (x_R, y_R, \alpha_R, \sigma_R)$,
- $A = (x_R, y_R)$ est la position du centre de l'imagette dans l'image de test et
- (α_R, σ_R) est le couple des paramètres d'orientation et d'échelle utilisé pour évaluer le vecteur de mesure $\mathcal{M}_R$ au point A.
- Une hypothèse (ou solution) de la requête R est un triplet $H = (Id_H, p_H, dist_H)$,
- Id_H est un identificateur de l'image modèle (un objet et un point de vue),
- $p_H = (x_H, y_H, \alpha_H, \sigma_H)$ est la position de l'imagette dans l'image modèle et
- $dist_H$ est la distance entre le vecteur modèle $\mathcal{M}_H$ et le vecteur $\mathcal{M}_R$. Cette distance donne un indice de confiance sur la qualité de l'appariement.
- Un appariement est un couple $C_k = (R_k; H_k)$.

Une mesure de compatibilité simple : la même image modèle La reconnaissance peut être basée sur le prédicat $MemeImage?$ qui détermine si deux hypothèses H_1 et H_2 sont compatibles. Le prédicat vérifie uniquement que les deux hypothèses correspondent à la même image modèle :

$$MemeImage?((R_1; H_1), (R_2; H_2)) = (Id_{H_1} =_{Id} Id_{H_2})$$

Le choix d'un prédicat aussi simple permet d'associer des hypothèses de façon directe. L'égalité $=_{Id}$ entre deux images modèles peut être assouplie pour permettre de rendre

compatible des images correspondants à des points de vues proches en particulier dans le cas où le point de vue observé est intermédiaire entre deux points de vue modèles.

Cette compatibilité simple présente un inconvénient majeur, elle autorise la coopération entre des hypothèses potentiellement non compatibles. En effet, les imagettes ne sont pas uniques et peuvent apparaître sur plusieurs objets ou plusieurs fois sur le même objet sans toutefois correspondre à la même caractéristique physique. Ainsi, pour une plus grande discrimination, il est intéressant d'ajouter à cette compatibilité simple une vérification des positions relatives entre les points de l'image de test d'une part et les hypothèses correspondantes d'autre part.

Une compatibilité sur la position relative des hypothèses La compatibilité entre deux couples (requête, hypothèse) C_1 et C_2 peut être vérifiée en définissant une mesure de compatibilité entre hypothèses qui tienne compte des positions relatives des hypothèses d'un même objet. Plus précisément, l'objectif consiste à définir une mesure $Compat?$ qui, à partir de deux couples $C_1 = (R_1, H_1)$ et $C_2 = (R_2, H_2)$, détermine la compatibilité des hypothèses H_1 et H_2 avec les requêtes R_1 et R_2 avec $R_k = (x_{R_k}, y_{R_k}, \alpha_{R_k}, \sigma_{R_k})$ et $H_k = (Id_{H_k}, p_{H_k}, dist_{H_k})$ pour $k \in \{1;2\}$.

Sous l'hypothèse d'une orientation et d'une échelle constantes, la mesure de compatibilité consiste à vérifier que l'image modèle est la même pour deux hypothèses et que le vecteur déplacement $p_{H_1}\vec{p}_{H_2}$ entre les hypothèses de l'image modèle est similaire au vecteur déplacement $R_1\vec{R}_2$ de l'image test. La mesure de compatibilité s'écrit :

$$Compat?((R_1, H_1), (R_2, H_2)) = MemeImage?(H_1, H_2) \wedge MemeVecteur?(p_{H_1}\vec{p}_{H_2}, R_1\vec{R}_2)$$

$MemeVecteur?$ est une mesure de similarité entre deux vecteurs déplacements. Cette mesure peut être évaluée avec plusieurs niveaux de précision. Le premier niveau consiste à vérifier uniquement que les deux vecteurs ont la même direction, puis un deuxième niveau consiste à évaluer, en plus, l'égalité des normes de ces deux vecteurs déplacements. Ce deuxième cas consiste à évaluer si les deux vecteurs de déplacements sont très proches :

$$MemeVecteur?(\vec{V_1} \left| \begin{matrix} x_1 \\ y_1 \end{matrix} \right. , \vec{V_2} \left| \begin{matrix} x_2 \\ y_2 \end{matrix} \right.) \iff \frac{(x_1 - x_2)^2 + (y_1 - y_2)^2}{(x_1 + x_2)^2 + (y_1 + y_2)^2} < S^2$$

Le seuil S d'égalité doit être choisi de façon à tenir compte des déformations éventuelles de l'objet observé dues aussi bien au déplacement de la caméra qu'à des déformations effectives de l'objet ou d'autres bruits. Ainsi, ce seuil dépend de la finesse de l'apprentissage de la sphère des vues possibles de l'objet. Cette équation effectue un seuillage sur l'erreur relative des vecteurs déplacements. Il est, par exemple, possible de définir $S = 20\%$ pour autoriser une variation de 20% du vecteur entre le modèle et l'image test. La discrimination ajoutée par ce prédicat est expérimentalement très importante même pour le choix d'un seuil S plus important.

Dans le cas où l'orientation et l'échelle ne sont pas fixées, la mesure de similarité doit tenir compte des variations en échelle et en orientation. L'ajout des paramètres d'orientation et d'échelle transforme les équations précédentes ainsi :

$$\begin{aligned} Compat?((R_1,H_1),(R_2,H_2)) &= MemeImage?(H_1,H_2) \\ &\wedge\ MemeVecteur?(p_{H_1}\vec{p}_{H_2}, R_1\vec{R}_2) \\ &\wedge\ MemeTransf?((R_1 \to H_1),(R_2 \to H_2)) \end{aligned}$$

La fonction $MemeTransf?((R_1 \to H_1),(R_2 \to H_2))$ évalue si la similitude évaluée entre la requête R_1 et l'hypothèse associée H_1 est compatible avec la similitude évaluée entre la requête R_2 et l'hypothèse associée H_2. En pratique, il faut vérifier que :

- La rotation entre R_1 et H_1 est égale à la rotation entre R_2 et H_2 :

  $$\alpha_{H_1} - \alpha_{R_1} =_{ang} \alpha_{H_2} - \alpha_{R_2}$$

  L'égalité $=_{ang}$ est définie "$modulo\,2\pi$" et doit tenir compte de l'imprécision de la détection angulaire.

- Les changements d'échelle observés pour R_1 et R_2 sont identiques :

  $$\sigma_{H_1}/\sigma_{R_1} =_{ech} \sigma_{H_2}/\sigma_{R_2}$$

Si l'angle de rotation ou le rapport d'échelle sont changés entre les recherches, les hypothèses sont incompatibles.

La mesure $MemeVecteur?$ est aussi étendue pour des changements d'échelles et d'orientations : un facteur d'échelle $\sigma_{H_1}/\sigma_{R_1}$ ainsi qu'une rotation $\alpha_{H_1} - \alpha_{R_1}$ sont appliqués à $R_1\vec{R}_2$ puis celui-ci est comparé au vecteur $p_{H_1}\vec{p}_{H_2}$ de l'image modèle.

En conclusion, le choix de la mesure de compatibilité permet d'utiliser un algorithme de reconnaissance d'objets incluant plusieurs recherches d'imagettes. Deux algorithmes sont proposés dans les sections 6.3 et 6.4. L'intérêt majeur de l'utilisation d'une mesure de compatibilité est la souplesse du paramétrage de cette mesure. Dans un cadre particulier, il est possible, par exemple, d'abandonner un paramètre non discriminant. De plus, les transformations envisageables étant perspectives, l'utilisation de l'évaluation de la similitude pour regrouper les appariements est risquée. Le mesure présentée ici se limite à effectuer une approximation sur une similitude entre couples de recherches. Pour le cas d'une similitude, la section suivante décrit comment évaluer cette transformation à partir d'un appariement unique.

Évaluation de la similitude entre deux imagettes

Une imagette est identifiée dans une image par un quadruplet $R = (x_R, y_R, \alpha_R, \sigma_R)$. L'appariement de deux quadruplets R_t et R_m permet de définir une similitude 2D unique.

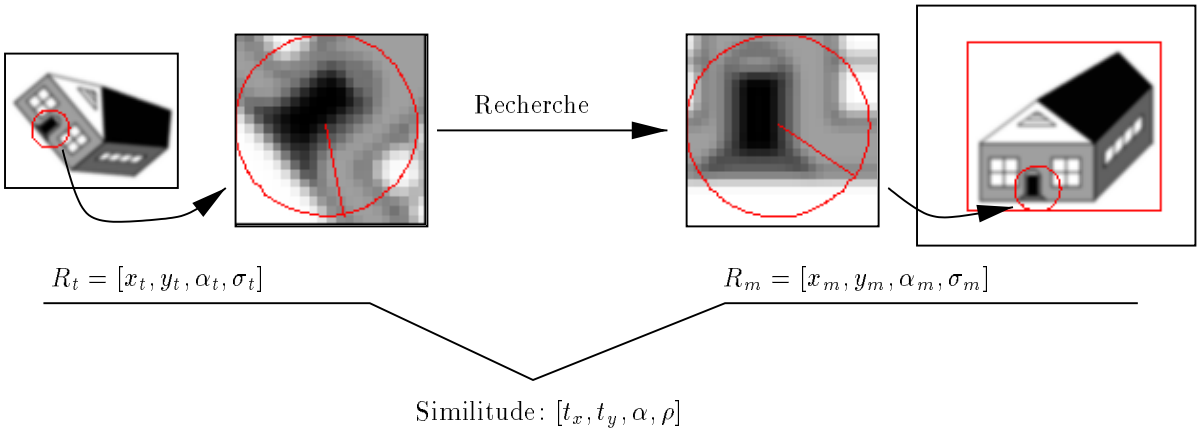


FIG. 6.2 – *Détection de similitude à partir d'un appariement entre deux imagettes. Ce schéma illustre l'évaluation de la similitude* 2D *entre deux images à partir de l'appariement de deux imagettes. Chaque imagette est représentée par un vecteur de ses position, orientation et échelle. Ces deux vecteurs permettent d'évaluer une similitude entre ces deux imagettes. Pour un objet rigide* 2D*, cette similitude est valide pour l'image complète. Cette évaluation peut être obtenue de façon robuste à partir d'appariements multiples par l'utilisation d'un vote sur l'espace* 4D *des paramètres de la similitude.*

Ceci est fait, classiquement, pour des segments de droites orientés (AYACHE [Aya83]). La figure 6.2 illustre la technique d'estimation de similitude à partir d'un appariement de deux imagettes. L'estimation de la similitude $T = [t_x, t_y, \alpha, \rho]^t$ entre une imagette R_t d'une image test et une imagette R_m d'une imagette modèle se fait ainsi :

$$\begin{aligned} \rho &= \sigma_t/\sigma_m \\ \alpha &= \alpha_t - \alpha_m \\ t_x &= x_t - \rho[x_m \cos\alpha - y_m \sin\alpha] \\ t_y &= y_t - \rho[x_m \sin\alpha + y_m \cos\alpha] \end{aligned} \tag{6.1}$$

La précision de la similitude ainsi détectée dépend fortement de la qualité de détection des paramètres d'échelle et d'orientation des imagettes. Cette transformation ne peut être calculée de façon stable à partir de deux imagettes de faible taille. Néanmoins, cette évaluation peut permettre par un vote (ou transformée de Hough [Hou62]) d'évaluer un point d'accumulation dans l'espace 4D des similitudes à partir d'un ensemble de recherches et obtenir ainsi la similitude entre deux images (voir section 6.3).

6.1.2 Évaluation d'une hypothèse fondée sur des appariements multiples

Cette section a pour but de donner un score à une liste d'appariements compatibles entre eux de façon à donner un score de reconnaissance à un algorithme fondé sur une utilisation simultanée de plusieurs points. Cette étude est inspirée de FAUGERAS [Fau93].

- Entrée : $L_n^{(i)} = ((R_1, H_{i_1}), \ldots, (R_n, H_{i_n}))$ avec R_k les points de l'image de test et H_{i_k} les points modèles appariés avec les R_k. L'ensemble de ces couples est compatible au sens proposé dans la section 6.1.1. Il est possible d'associer à cette liste une similitude entre l'image test et l'image modèle. Chacun des appariements peut être évalué par sa distance au point modèle associé $d_j^{(i)} = dist(R_j, H_{i_j})$ (voir section 4.1.1 pour plus de détails).

- Sortie : un score permettant de comparer plusieurs liste $L_n^{(k)}$ de résultats.

Deux scores peuvent être évalués : d'une part, la distance moyenne entre points de l'image de test et points modèle, soit $\epsilon_n^{(i)} = \frac{1}{n}\sum_{j=1}^{n} d_j^{(i)}$. La minimisation de ce score $\epsilon_n^{(i)}$ permet de sélectionner l'ensemble d'appariements les plus similaires entre eux. Néanmoins, on observe la plupart du temps que un ou plusieurs appariements ne sont pas trouvés. On associe aux termes H_{i_j} non appariés la valeur NIL et ils ne peuvent pas être utilisés pour le calcul de $\epsilon_n^{(i)}$. Ceci implique de définir un deuxième score $p^{(i)}$ qui est le nombre d'hypothèses NIL dans la liste.

L'évaluation de la liste est fondée sur une combinaison des scores $p^{(i)}$ et $\epsilon_n^{(i)}$. Cette combinaison est un problème difficile qui dépend des conditions expérimentales et de l'objectif de la reconnaissance. En général, un ordre simple est utilisé pour trouver la meilleure hypothèse pour un ensemble de recherche : un tri est effectué sur le nombres d'appariements manqués $p^{(i)}$ puis la distance $\epsilon_n^{(i)}$ départage les hypothèses ayant le même nombre d'appariements. En conclusion de cette thèse, nous donnons un point de départ pour une évaluation de ce score de manière probabiliste qui peut permettre de limiter les inconvénients de l'évaluation proposée ici : score basé sur un couple et utilisation de prédicats logiques pour valider un appariement ou la cohérence spatiale entre appariements.

Deux algorithmes ont été implémentés pour effectuer la reconnaissance à partir d'appariements multiples. Ils évaluent ce score de façon à sélectionner l'objet le plus vraisemblablement présent dans l'image.

- Algorithme par vote : chaque appariement incrémente un accumulateur. L'accumulateur obtenant le maximum de votes correspond à l'objet le plus vraisemblable. Le calcul simultanée de la distance moyenne $\epsilon_n^{(i)}$ permet de départager d'éventuels ex-aequo.

– Algorithme fondé sur le paradigme prédiction–vérification : le principe de cet algorithme consiste à générer des hypothèses d'objets à partir d'une projection puis à confirmer ces hypothèses en vérifiant la présence de leurs voisins dans la base d'apprentissage. La même évaluation peut être utilisée pour comparer les hypothèses d'objets mais cet algorithme permet une optimisation combinatoire en supprimant les hypothèses peu vraisemblables très rapidement.

6.2 Sélection des points à rechercher

L'apprentissage de l'intégralité des images sans sélection de points d'intérêts permet, a priori, de mettre en correspondance n'importe quel point d'une image de test avec les images modèles. Néanmoins, certains points ou certains parcours de l'image de test peuvent améliorer, et surtout accélérer, la phase de reconnaissance. Plusieurs stratégies de sélection de points sont proposées et exposés sur la figure 6.3.

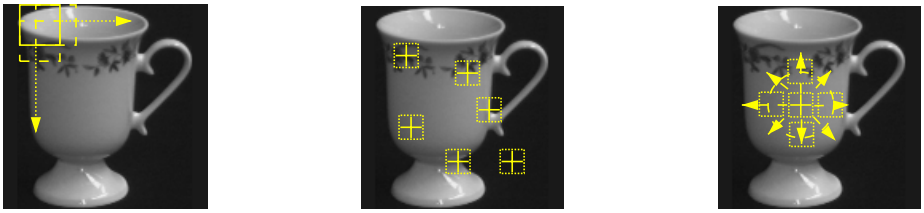


FIG. 6.3 – *Plusieurs parcours sont envisageables pour reconnaître le contenu d'une image inconnue : Parcours exhaustif, Parcours aléatoire ou Parcours centré sur un point de focalisation.*

En l'absence d'informations sur le contenu de l'image ou de pré-traitements sur l'image, deux stratégies directes sont possibles. Un parcours exhaustif de toutes les imagettes disponibles permet une reconnaissance la plus correcte possible : il s'agit d'extraire successivement l'ensemble des imagettes de l'image ou d'une région d'intérêt de l'image. Cette technique est très discriminante mais aussi très coûteuse. Le fort recouvrement de deux imagettes successives motive un pas de parcours supérieur à 1 pixel. Les expérimentations utilisent un pas p de parcours de valeur $n/2$ où n est la taille des imagettes dans le cas d'une base obtenue par ACP et un pas σ pour les bases dérivées de Gaussienne. Ainsi, la redondance de l'information entre les imagettes successives est assez faible. Ce parcours devra être utilisé si plusieurs objets sont présents simultanément dans l'image ou si une forte occultation des objets est possible. Pour un coût plus faible et de façon assez équivalente, il est possible de sélectionner des imagettes au hasard dans l'image et pour un nombre d'imagettes sélectionnées suffisant obtenir une reconnaissance robuste. Ce parcours a été utilisé pour l'appariement de scène pour l'estimation de position en robotique mobile (voir section 7.1).

Une deuxième stratégie consiste à sélectionner des points a priori discriminants pour effectuer une recherche initiale sur ces points. Les points recherchés seront, d'une part, les points d'intérêts sélectionnés et d'autre part, les voisins de ses points. Ainsi, à partir d'un point discriminant, un parcours possible consiste à évaluer ses voisins en s'éloignant progressivement de celui-ci jusqu'à obtenir la confirmation des hypothèses émises par la recherche du point discriminant. Ce parcours présente l'avantage de conserver une certaine localité et donc de permettre une grande robustesse par rapport à l'occultation partielle.

Les points a priori discriminants peuvent être sélectionnés suivant les deux techniques présentées au chapitre précédent (sections 5.1.1 et 5.1.2) : soit par l'utilisation d'un détecteur de points d'intérêts comme le détecteur de TOMASI , soit par l'utilisation d'un détecteur probabiliste fondé sur la fréquence d'apparition des points dans la base d'apprentissage. Il est, aussi, possible d'obtenir des points discriminants depuis des processus externes de suivi ou de reconnaissance (voir section 7.2 pour un exemple dans le cadre de la reconnaissance de poissons rouges).

Les deux sections suivantes décrivent deux stratégies de reconnaissance.

6.3 Algorithme de vote

Un algorithme de vote peut être utilisé pour regrouper des hypothèses obtenues par des recherches distinctes. Il s'agit d'un algorithme similaire à la transformée de Hough : l'objectif est de trouver un point d'attraction dans un espace de paramètres. Chaque recherche génère une liste d'hypothèses. Chacunes d'entre elles votent pour un objet (incrément d'un accumulateur) puis après quelques recherches le (ou les) objets obtenant le maximum de votes sont reconnus. La difficulté principale de cet algorithme est le choix de la discrétisation de l'espace des solutions. Ce choix du pas de discrétisation dans les différentes dimensions de cet espace est difficile car une discrétisation trop fine génére de multiples hypothèses peu probables et une discrétisation trop large rend similaire des hypothèses différentes ce qui risque d'entraîner des fausses reconnaissances. Un problème important est l'aspect arbitraire des frontières posées entre les cases de l'espace discrétisé : deux éléments très proches peuvent voter pour des cases différentes. L'idée de l'algorithme est de choisir une taille de case telle que ce problème se produise aussi rarement que possible.

Dans le cadre d'un espace de solutions ayant plus de deux dimensions comme pour le cas du vote basé sur la mesure *Compat*? (voir paragraphe 6.3). Une discrétisation trop fine posera, en plus, des problèmes de mémorisation et d'accès à un espace de grande dimension.

Deux choix principaux d'espaces de solutions peuvent être faits ici :

- Un espace 1D fondé sur la mesure simple *MemeImage*? peut être utilisé. Dans ce cas, un accumulateur est associé à chaque image modèle. Les images obtenant les

accumulateurs maximums sont reconnues. Cet espace a l'inconvénient de regrouper des hypothèses incompatibles.

- Un espace 5D fondé sur la mesure *Compat?*. Dans ce cas, chaque hypothèse d'appariement implique une hypothèse d'image et une similitude 2D entre le point de l'image test et le point de l'image modèle soit le vecteur à 5 paramètres $v = (img, t_x, t_y, \alpha, sc)$. Une hypothèse de similitude 2D globale entre l'image de test et une image modèle est faite. Dans le cas d'une transformation perspective importante, cet algorithme sera en échec.

L'algorithme de vote est présenté précisément sur la table 6.3. Un des points clés de cet algorithme est l'évitemment du vote multiple : la recherche d'un vecteur de mesure peut impliquer plusieurs votes simultanés pour le même accumulateur et, dans ce cas, il est nécessaire d'éviter d'incrémenter plusieurs fois cet accumulateur sous peine de fausser le vote.

Soit $L_A = (A_1, A_2, \ldots)$ *// liste des points à rechercher*
Fonction ReconnaîtreVote(L_A)
Début
  Soit A_{obj} un tableau multidimensionnel dont chaque case correspond à un vote possible.
    // Une case est composée d'un compteur initialement nul et d'un drapeau booléen
    // initialement faux décrivant si la case a été modifiée pendant l'itération courante.
  Pour chaque point A_j de L_A, **faire :**
    L_{hypos} = Recherche(A_j)
    Pour chaque hypothèse H_k de L_{hypos}, **faire :**
      Affecter à $case$, le vecteur décrivant la case de H_k
      Si $drapeau(A_{obj}(case)) = faux$ **alors**
        $drapeau(A_{obj}(case)) = vrai$
        Incrémenter le compteur de $A_{obj}(case)$
    Positionner tous les drapeaux de A_{obj} à $faux$.
  Retourner les $case$ correspondants aux compteurs maximums de A_{obj}.
Fin.

TAB. 6.1 – *Algorithme de reconnaissance par vote.*

Dans le cas d'une similitude quelconque, l'espace des solutions est constitué de 5 dimensions. La première correspond à l'image modèle et les suivantes donnent les paramètres de la similitude entre l'image modèle et l'image de test. Ces quatres paramètres appartiennent à un espace continu qu'il faut discrétiser pour effectuer un vote : il faut définir la largeur des cases dans chacunes des dimensions. Cette largeur correspond à la précision maximale de la transformation détectée.

Une expérience simple valide l'algorithme de transformation de Hough pour l'évaluation de la similitude 2D entre deux images d'un même objet comme cas d'étude. Les images présentées sur les figures 6.4 et 6.5 présentent des variations importantes du zoom et de l'orientation. L'évaluation de la similitude se fait en trois phases :

1. Calcul des vecteurs de mesures invariants à l'échelle et à l'orientation en tout points des deux images (voir chapitre 3).

2. Apprentissage de tous les points de la première image dans la structure de recherche (voir section 5.3.1). Ainsi, il est possible d'effectuer des appariements rapides de vecteurs de mesure.

3. Appariement de chaque vecteur de mesure de la deuxième image avec le vecteur le plus proche de la première image. Chaque appariement propose une hypothèse de similitude (t_x, t_y, θ, ρ). De façon similaire à la transformée de Hough, il suffit alors de trouver le point d'attraction de l'espace de ces quatres paramètres pour obtenir la similitude. Ceci est fait par un vote.

Les figures 6.4 et 6.5 donnent deux exemples d'évaluation de similitude sur des couples d'images. Pour la première, la similitude $(t_x = 15, t_y = 80, \theta = 44°, \rho = 43\%)$ est trouvée entre les deux images et pour la seconde, nous obtenons $(t_x = -510, t_y = 420, \theta = 70°, \rho = 369\%)$. La validité des similitudes trouvées est visualisée par quelques appariements de points calculés à partir de cette transformation.

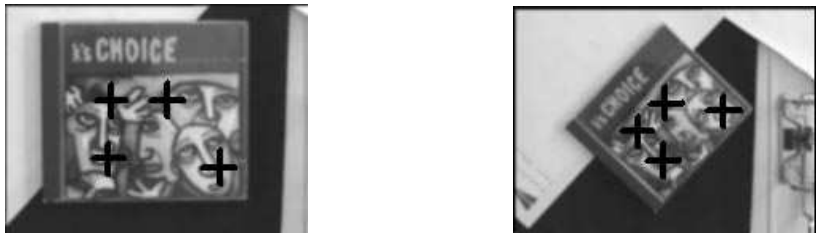


FIG. 6.4 – *Évaluation de la similitude par vote sur l'espace* 4D *des similitude* 2D. *Quatre appariements de points sont visualisés pour valider la transformation trouvée.*

Cet algorithme peut être étendu à la reconnaissance d'objets en ajoutant une dimension à l'espace des solutions : le numéro de l'image modèle mais les problèmes de discrétisation et de réponses multiples en chaque point rendent ce processus instable. Néanmoins, cet algorithme a été utilisé pour la reconnaissance automatique de scènes pour l'estimation de position en robotique mobile. Dans ce cadre, l'espace des paramètres est limité à trois dimensions. L'identité de l'image est la première dimension. Cette dimension est discrète. Les deux dimensions suivantes sont les paramètres de translation entre l'image courante et l'image modèle. L'évaluation de l'identité de l'image donne la position approximative du robot. L'utilisation des paramètres de translation renforcent

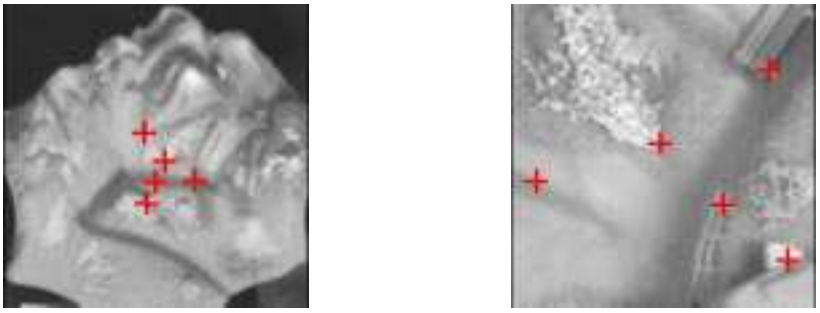


FIG. 6.5 – *Évaluation de la similitude par vote sur l'espace* 4D *des similitude* 2D. *Cinq appariements de points sont visualisés pour valider la transformation trouvée.*

la discrimination entre images en supprimant les votes non cohérents entre eux. Le paramètre de translation suivant l'axe des abscisses permet de corriger une trajectoire. Cette application est développée à la section 7.1.

De façon générale, la transformée de Hough nécessite de trouver un compromis entre la précision de l'espace des paramètres et l'incertitude sur les données. Plus la précision voulue est grande, plus le point d'attraction risque d'être manqué. L'incertitude sur les données implique une précision peu importante. Ces problèmes impliquent pour un espace 5D une reconnaissance très difficile. Néanmoins, le point d'attraction dans cet espace des paramètres existe et peut être obtenu par l'utilisation d'une transformée de Hough flou comme proposée par STRAUSS [Str99]. Cette transformée tient compte de l'incertitude sur les données pour effectuer un vote pondéré par la distance entre les points de l'espace. la précision est conservée sans risquer la perte du pic par la phase de vote. Une autre stratégie fondée sur le paradigme Prédiction–Vérification est proposée à la section suivante.

6.4 Une stratégie Prédiction–Vérification

La technique de reconnaissance d'objets de la section précédente est fondée uniquement sur une approche ascendante. Il s'agit d'obtenir une information sur l'identité et la pose de l'objet à partir des informations extraites directement de l'image. En opposition, l'approche descendante consiste à utiliser des connaissances abstraites sur la scène observée pour spécialiser la recherche d'éléments dans l'image confirmant ces connaissances. Pour l'approche ascendante, les vecteurs de mesures extraits des images sont recherchés puis regroupés par un vote pour accéder à la reconnaissance.

Plusieurs systèmes de vision proposent d'utiliser conjointement une approche ascendante et descendante : dés qu'une connaissance est obtenue sur la scène, elle est utilisée pour diriger l'extraction et la recherche des autres caractéristiques locales dans la scène. Il est alors plus simple de confirmer ou de réfuter la connaissance initiale. Les avantages

de tels systèmes sont la possibilité de prendre en compte des données très hétérogènes et la possibilité d'utiliser directement des connaissances préalables sur la scène observée. De plus, dans le cas d'indices visuels de type homogène l'utilisation de toute la connaissance courante permet souvent d'accélérer la reconnaissance en la spécialisant sur l'état courant. Plusieurs systèmes de vision peuvent être cités comme le système PVV proposée par SOUVIGNIER et LUX [Sou83, Lux86] ou le système proposé par AYACHE [Aya83]. Ces systèmes proposent d'utiliser le paradigme général prédiction–vérification pour obtenir une reconnaissance robuste et rapide. La reconnaissance est basée sur deux modules principaux : un premier module génère des hypothèses vraisemblables sur la scène observée, puis un deuxième module confirme ou réfute ces hypothèses en extrayant d'autres caractéristiques de bas-niveau. Ces deux modules doivent être contrôlés par un module de décision de plus haut niveau qui activera ces modules successivement jusqu'à obtenir un score de reconnaissance acceptable.

De nombreuses optimisations pourront être ajoutées au module de décision en supprimant des branches de l'arbre de confirmation dés que la confirmation devient catastrophique et inversement l'arrêt de la phase de confirmation dés qu'une hypothèse exceptionnellement valide est obtenue Ces systèmes utilisent des caractéristiques visuelles assez instables et peu discriminantes comme des segments de droites. La technique proposée ici est fondée sur l'approche de AYACHE en remplaçant les indices segments par les caractéristiques locales présentées au cours des chapitres précédents. L'utilisation de caractéristiques très discriminantes permet d'étendre sa technique à la reconnaissance d'objets parmi une grande base d'objets.

6.4.1 L'algorithme prédiction–vérification

Un algorithme fondé sur le paradigme prédiction–vérification opère en deux phases :

1. La *Prédiction* génère des hypothèses de reconnaissance. Ceci peut être fait en utilisant les descripteurs locaux présentés au chapitre 3 et la technique de recherche proposée. La grande discrimination obtenue par l'utilisation de ces descripteurs permet de limiter fortement le nombre de fausses hypothèses. Il est, de plus, possible de sélectionner des points très discriminants pour générer ces hypothèses initiales. Cette sélection peut être basée sur un détecteur de points d'intérêts (voir section 5.1.1), sur un critère statistique (voir section 5.1.2) ou sur le résultat d'un autre algorithme extérieur (voir section 7.2).

2. La *Vérification* confirme ou réfute une hypothèse en vérifiant d'autres indices visuels sur l'image. Confirmer la présence d'un vecteur de mesure pour une image modèle à une certaine position est un processus beaucoup plus simple et rapide qu'une recherche complète : dés que le vecteur est trouvée, le parcours s'arrête !

Ainsi, cet algorithme de mise en correspondance de deux images est étendu ici à la reconnaissance d'objets 3D grâce à l'utilisation de caractéristiques locales très discriminantes.

L'algorithme proposé nécessite de définir un prédicat de confirmation qui réponde à la question : Est-ce que le point R de l'image est compatible avec l'hypothèse H. Cette évaluation nécessite un parcours de la base de modèles qui, souvent, s'arrête beaucoup plus vite qu'une recherche générale. Dés que le point est trouvé, le parcours est interrompu. La fonction de confirmation est définie plus précisément en annexe sur la table B.3 (page 152).

L'algorithme précis est présenté sur la table 6.2.

// Phase de Prédiction
Fonction PredictionHypothèses(A un point d'intérêt initial)
// ce point d'intérêt doit être choisi discriminant
Début
    Retourner (Recherche(A))
Fin.

// Phase de Vérification
Fonction VerificationHypothèse(L_A, L_pred)
// L_A est une liste de points de l'image. // L_pred est une liste d'hypothèses à vérifier. **Début**
    Initialiser une table de compteurs des éléments de L_pred à zéro.
    Pour chaque point A_j de L_A, **faire :**
        Pour chaque prédiction H_k de L_{pred}, **faire :**
            Si Confirmation(A_j, H_k) **alors**
                Incrémenter compteur de H_k
    Retourner les hypothèses ayant le compteur maximum
Fin.

TAB. 6.2 – *Algorithme de reconnaissance par pr1édiction/vérification.*

Après sélection d'une fenêtre a priori discriminante, une recherche sur cet élément permet de générer une série d'hypothèses vraisemblables. Pour chacune des hypothèses, l'algorithme tente de confirmer l'hypothèse en vérifiant que ses voisins sont reconnus. Un parcours des 8 voisins de la fenêtre est utilisé.

6.4.2 Résultats expérimentaux

La validation de cette stratégie de reconnaissance par prédiction–vérification est effectuée sur un problème simple : la base d'images Columbia ou COIL [NNM96b] (voir détails en annexe A.1). La base d'apprentissage utilisée pour la construction de la base des modèles est constituée de 18 images par objet pour 100 objets. Deux images d'un même objet sont séparées de 20°. La base de descripteurs utilisés est une base de 10 filtres ACP couleur de petite taille 9×9. La base des modèles est constituée d'environ 3 millions de

vecteurs de mesures associés aux identificateurs les référençant. Les images intermédiaires sont utilisées pour former la base de test, soit 5400 images. Le choix de cette base se justifie par l'absence de changement d'échelle ou d'orientation 2D entre les images de test et les images modèles.

L'algorithme de reconnaissance se fonde sur deux hypothèses simplificatrice : un objet unique est présent sur une image et cet objet est centré dans l'image. Ceci permet de sélectionner le point central de l'image pour la phase de prédiction de l'algorithme. L'imagette centrée sur ce point est recherchée dans la base des modèles puis les 20 modèles les plus proches de l'imagette sont sélectionnés. La phase de vérification sélectionne le modèle le plus proche par validation sur les 8 imagettes voisines. Le modèle reconnu est sélectionné sur le nombre d'imagettes voisines retrouvées dans l'image de test. En cas d'égalité sur ce nombre, la distance moyenne entre imagettes permet de sélectionner le modèle le plus proche.

Suivant la technique proposée, nous avons obtenu pour 5400 images de test, 5312 reconnaissances correctes et 88 échecs soit un taux de reconnaissance de 98.4%. L'analyse des cas d'échecs permet de distinguer deux causes liées principalement à la simplicité de l'algorithme utilisé. Une première cause d'échec est la présence d'objets localement similaires et donc non distinguable localement comme le montre l'exemple de la figure 6.6. Dans ce cas, l'objet sélectionné parmi les modèles proches n'est pas sûr et aboutit dans certains cas à un mauvais choix. Le modèle correct apparaît néanmoins en deuxième ou troisième reconnaissance. La deuxième cause d'échecs est l'utilisation comme première imagette pour la prédiction d'une imagette peu discriminante. Dans ce cas, le modèle correct n'est pas nécessairement sélectionné parmi les 20 modèles les plus proches et ceci aboutit à un échec de la reconnaissance.

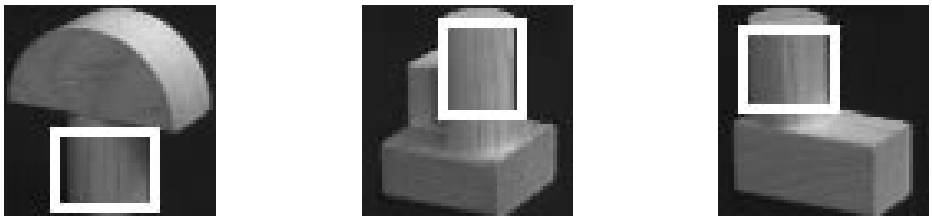


FIG. 6.6 – *3 objets localement similaires.*

Cette expérience prouve la validité de l'algorithme sur une grande base de modèles (3 millions de points). Néanmoins, la base de Columbia est considérée comme simple à reconnaître : le choix de l'imagette centrale pour la phase de prédiction s'est avéré suffisant. Pour d'autres bases d'objets de moindre qualité et contenant des variations en échelle et en orientation, ce choix n'est plus suffisant et il est nécessaire de mettre en place un superviseur pour effectuer cette reconnaissance de façon systèmatique. Les expériences sur la base de 28 objets avec variations de l'échelle importantes ont montré que, pour un choix adapté du point initial, la phase de vérification sélectionne l'objet correct mais que

si le point est peu discriminant, aucun objet n'est reconnu avec une certitude suffisante. L'utilisation de ces informations doit permettre de mettre en place un superviseur capable d'exécuter successivement les phases de prédiction et de vérification jusqu'à obtention d'un score de reconnaissance suffisamment important. Ceci a été fait, dans le cas particulier de la reconnaissance de poissons rouges proposé dans le chapitre 7, les points initiaux sont donnés par un autre algorithme et, dans ce cas, l'algorithme par prédiction–vérification est appelé directement sur ces points.

6.5 Conclusions et Perspectives

Ce chapitre a abordé le problème complexe de la reconnaissance d'objets en utilisant comme indices de reconnaissances des vecteurs de mesures sur la présence de caractéristiques locales. Les stratégies proposées ont pour objet d'aboutir à l'appariement d'une image de test avec une image modèle en utilisant des appariements sur ces vecteurs de mesures. Deux stratégies sont proposées :

- Un vote ou transformée de Hough associe à chaque appariement entre vecteurs une similitude 2D. La détection d'un point d'accumulation dans l'espace des similitudes fournit une hypothèse de reconnaissance.

- La reconnaissance peut être obtenue suivant une stratégie de prédiction-vérification en deux phases. La première génère des hypothèses d'objets vraisemblables à partir de la recherche d'une imagette puis la seconde confirme ou réfute chaque hypothèse pour aboutir à la reconnaissance.

Ces stratégies ont leur limitations et nécessitent une adaptation au problème précis envisagé. En particulier, il est très profitable de tenir compte des informations sur l'occultation partielle ou la présence d'objets multiples pour permettre une reconnaissance fiable. De plus, le problème de la localité implique des faux appariements qu'il n'est possible de corriger que par une approche globale ou, au moins, un apprentissage des similarités entre objets.

Pour obtenir un système de reconnaissance robuste, l'accent doit être mis par la suite sur la mise en place d'un superviseur qui doit contrôler l'execution des phases de prédiction et de vérification : l'algorithme actuel se limite à une phase de prédiction suivi d'un phase de vérification et ceci implique une forte dépendance sur la prédiction initiale. L'utilisation des scores de reconnaissance permet d'évaluer la qualité des reconnaissances et, ainsi, de choisir de relancer une nouvelle prédiction ou de s'arrêter. De plus, l'utilisation d'un superviseur basé sur une stratégie par prédiction–vérification permet facilement d'utiliser d'autres caractéristiques pour effectuer la reconnaissance. Ainsi, à la section 7.2, l'utilisation de descripteurs fondés sur le mouvement en association avec la modélisation par dérivées de Gaussiennes proposée dans cette thèse a permis de mettre en place un système de reconnaissance robuste et rapide.

Chapitre 7

Applications à des problèmes de vision

7.1 Reconnaissance de scènes pour l'estimation de position

L'estimation de position est un problème clé pour la navigation autonome d'un robot mobile. En effet, un processus de navigation peut être décomposé en trois tâches : l'estimation de position, la planification de chemin et le contrôle d'exécution. L'objet de cette section est de décrire une technique d'estimation de position à partir de données capteur brutes.

L'estimation de position d'un robot mobile dans un environnement connu et contrôlé est fréquemment résolue par la mise en place de repères visuels caractéristiques que le robot peut facilement détecter et suivre pour se positionner. Afin de supprimer la nécessité de modifier l'environnement du robot, MATSUMO [MII96] a proposé de définir une nouvelle représentation visuelle d'un chemin du robot : la VSRR (pour "View–Sequenced Route Representation"). Cette représentation permet au robot de se localiser sur un trajet préalablement enregistré par l'utilisation de corrélations entre l'image observée par le robot et les images apprises sur l'itinéraire.

De façon similaire JONES et ANDERSEN [AJC97, Jon97] ont mis en place un système de navigation visuelle par pré-apprentissage de la route à suivre. Ils ont mis en place des processus visuels qui convertissent les images issues de la caméra en commandes robot. Une corrélation est utilisée pour maintenir et corriger la direction du véhicule.

Nous avons menés des expériences avec la technique de reconnaissance d'objets proposée dans cette thèse. Pour l'estimation de position, les objets à reconnaître sont les scènes observées par la caméra posée sur un robot mobile. Une approche similaire a été proposée par NAYAR [NNM96a]. Les chemins visuels proposés par les auteurs précédents souffrent du risque d'occultation partielle des images d'apprentissage qui ne permettent

FIG. 7.1 – *Tour visuel de la halle robotique du* LIFIA.

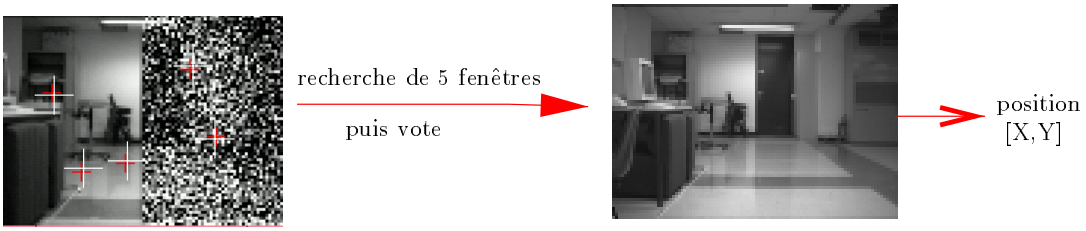


FIG. 7.2 – *Exemple d'une recherche réussie en utilisant 5 fenêtres pour une image à moitié cachée.*

pas le suivi sûr d'une trajectoire, en particulier, dans le cas d'évitement d'obstacles imprévus. L'utilisation de caractéristiques locales permet d'estimer la position de façon robuste à l'occultation partielle et aux incertitudes en translation. Globalement, l'estimation de position est fondée sur l'appariement d'une image apprise avec l'image observée par le robot en utilisant le système de reconnaissance d'objets proposé dans cette thèse.

Un chemin visuel (figure 7.1) dans le laboratoire LIFIA est défini par une séquence d'images : il s'agit de séquences vidéo enregistrées par ANDERSEN et JONES pour l'évaluation de leur système de navigation. L'objectif de notre système est de détecter dans une nouvelle image l'image modèle qui lui est la plus similaire et obtenir ainsi une position approximative du robot. Une position plus précise peut être obtenue par l'usage de techniques d'interpolation comme POURRAZ [PC98, Pou98]. La figure 7.2 montre un exemple d'une image fortement dégradée. L'image a été reconnue par la recherche de cinq imagettes choisies aléatoirement qui a permis, par un vote, de sélectionner l'image d'apprentissage correcte. Sur cette application, les images modèles ont été projetées sur une base de descripteurs locaux ACP niveaux de gris de taille 9×9 . Sur une base d'images de test, la figure 7.3 évalue la qualité de la reconnaissance en fonction du nombre d'imagettes utilisées. Les images de test utilisées sont prises à des positions intermédiaires des images modèles.

Cette figure montre une reconnaissance presque parfaite par l'utilisation de plusieurs fenêtres sur des images non occultées. Quatre fenêtres fournissent 97% de reconnaissance

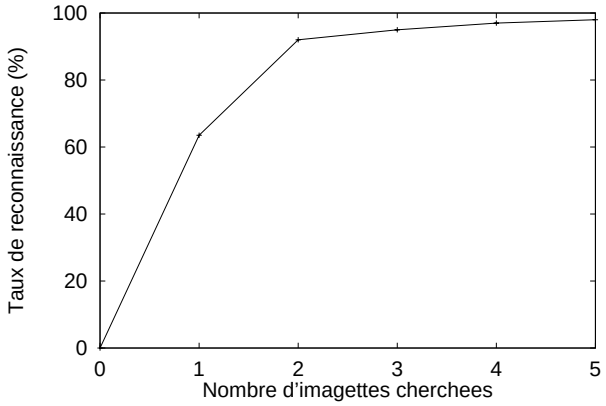


FIG. 7.3 – *Reconnaissance en fonction du nombre de fenêtres utilisées.*

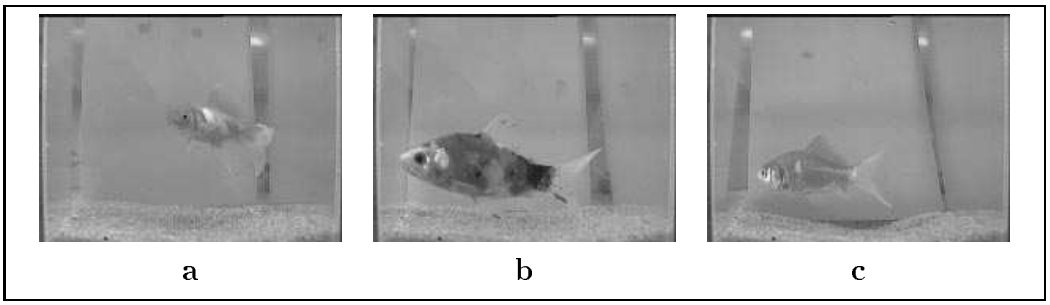


FIG. 7.4 – Les trois poissons rouges utilisés dans les expériences.

et pour les 3% restants, la réponse correcte est trouvée en 2ème ou 3ème position. L'algorithme de vote proposé au chapitre précédent donne sur un cas simple où les paramètres d'échelles et d'orientation ne sont pas variables de très bons résultats de reconnaissance. L'ajout de la sélection automatique du paramètre d'échelle peut permettre d'utiliser moins d'images pour représenter un parcours du robot mobile.

7.2 Reconnaissance de poissons rouges

La technique de reconnaissance proposée dans cette thèse est très générale, elle peut être appliquée sur une large gamme d'objets. La reconnaissance de différentes espèces de poissons rouges est un défi pour les techniques de reconnaissance d'objets par ordinateur. La figure 7.4 montre les images de trois poissons rouges différents. L'objectif est de les distinguer automatiquement sans utilisation d'une technique ad-hoc.

Cette application est difficile pour plusieurs raisons :

– Les poissons sont des objets vivants et non contrôlables. Pendant la phase d'ap-

prentissage, il est très difficile d'obtenir un échantillonnage suffisamment dense de la sphère des points de vues possibles de chaque poisson. Une sélection manuelle des images les plus représentatives de l'apparence de chacun des poissons n'est pas envisageable.

- Les poissons sont des objets déformables et non polyhédriques : ceci exclut l'utilisation de techniques à base de modèles géométriques et encourage l'utilisation de techniques fondées sur l'apparence.

- Le milieu aqueux entraîne des problèmes de reflets et des images très bruitées. Les bords de l'aquarium forment, souvent, des reflets multiples des poissons.

Le problème de l'apprentissage non contrôlé est abordé dans la section 7.2.1. Puis, la section 7.2.2 aborde le problème de la reconnaissance qui est obtenue par l'utilisation d'une stratégie coopérative fondée sur une évaluation de l'activité des poissons et sur l'évaluation de leur apparence statique par la technique présentée dans cette thèse.

7.2.1 Apprentissage non contrôlé

L'apprentissage de chacun des poissons nécessite de représenter chacun d'entre eux par une collection d'images échantillonnant leurs points de vue possibles.

Apprentissage d'un poisson L'apprentissage est effectué en filmant un poisson isolé des autres pendant quelques minutes. Chacune des images est segmentée en se fondant sur la détection du mouvement entre images successives. Un échantillonnage des images est extrait. Pour chacune des images extraites, le poisson est projeté sur l'espace $\mathcal{A}$ de représentation. L'espace $\mathcal{A}$ utilisé dans cette expérience est l'espace de dérivées de Gaussiennes avec utilisation de la détection automatique de l'orientation et de l'échelle. Puis, l'ensemble des vecteurs de mesures est enregistré dans une base de données. L'inconvénient de cette approche est la présence probable de plusieurs images similaires d'un même poisson ainsi que le risque de manquer certains points de vues possibles. Le premier inconvénient peut être levé en effectuant avant chaque ajout d'une nouvelle image la recherche de l'image dans la base. Si suffisamment de points donnent un appariement correcte, l'image est rejetée car similaire à un modèle. Le deuxième inconvénient ne peut être corrigé que par l'utilisation d'autant d'images que possible ou par la génération automatique de nouvelles images.

Génération automatique de points de vues non observés De manière générale et pour cette application en particulier, il est possible à partir de quelques images d'un objet 3D de synthétiser de nouvelles images de cet objet. Ce problème a été, par exemple, étudié par Blanc [Bla98]. Pour cela, une méthode précise consiste à évaluer les transformations projectives liant les images entre elles et une fois la géométrie connue, il est

possible de reprojeter les objets suivant de nouveaux points de vues. Cette technique permet de générer un échantillonage plus complet de la sphère des vues. La collection d'images représentant l'objet devient suffisante pour le reconnaître même sous des points de vue non observés.

7.2.2 Reconnaissance

Les poissons rouges sont caractérisés par deux aspects principaux: leur apparence statique et leur comportement dynamique. Cette section propose une technique de reconnaissance fondée sur la modélisation de ces deux aspects. Le comportement dynamique ou activité est capturé par une technique statistique proposée par CHOMAT et l'apparence statique est capturée suivant la stratégie proposée dans cette thèse. Un algorithme coopératif simple permet de conjuguer les résultats des deux techniques suivant le paradigme prédiction–vérification. L'algorithme fondé sur l'activité génère des hypothèses de reconnaissance puis, l'algorithme statique vérifie ces hypothèses en appariant une image modèle à l'image observée.

Reconnaissance par modélisation statistique de l'activité Les poissons peuvent être caractérisés par une modélisation de leur mouvement : CHOMAT [CC99a] propose une technique de reconnaissance probabiliste d'éléments d'activités par leur caractéristiques spatio-temporelles locales. Un espace de description est défini par une base de champs réceptifs sensibles à l'énergie du mouvement. L'analyse statistique de l'espace de caractéristiques locales est effectuée par la construction d'histogrammes multi-dimensionnels qui donnent une estimation de la densité de probabilité nécessaire à un processus de reconnaissance probabiliste basé sur une règle de Bayes. La technique résultante permet une reconnaissance d'éléments d'activités qui est relativement indépendante de la texture de l'objet en activité et qui est robuste aux occlusions et aux changements d'illumination. En illustration de cette technique, la figure 7.5 montre des cartes de probabilités obtenues sur des séquences extérieures à celles d'apprentissage. Les zones blanches correspondent à une forte probabilité.

Algorithme de coopération La reconnaissance des poissons rouges étant un problème difficile, nous avons choisi d'utiliser deux techniques de reconnaissance simultanées qui sont fondées sur deux modélisations différentes: la première est fondée sur une modélisation statistique des mouvements des poissons et la seconde sur leur apparence statique. La coopération entre ces deux techniques est fondée sur le principe prédiction–vérification.

La phase de prédiction d'hypothèses est effectuée par l'algorithme statistique qui, grâce au mouvement, détecte la position du poisson présent dans l'image et par moyennage des probabilités des différents poissons génère une ou plusieurs hypothèses sur le poisson présent dans l'image. Ces hypothèses sont données avec une localisation précise.

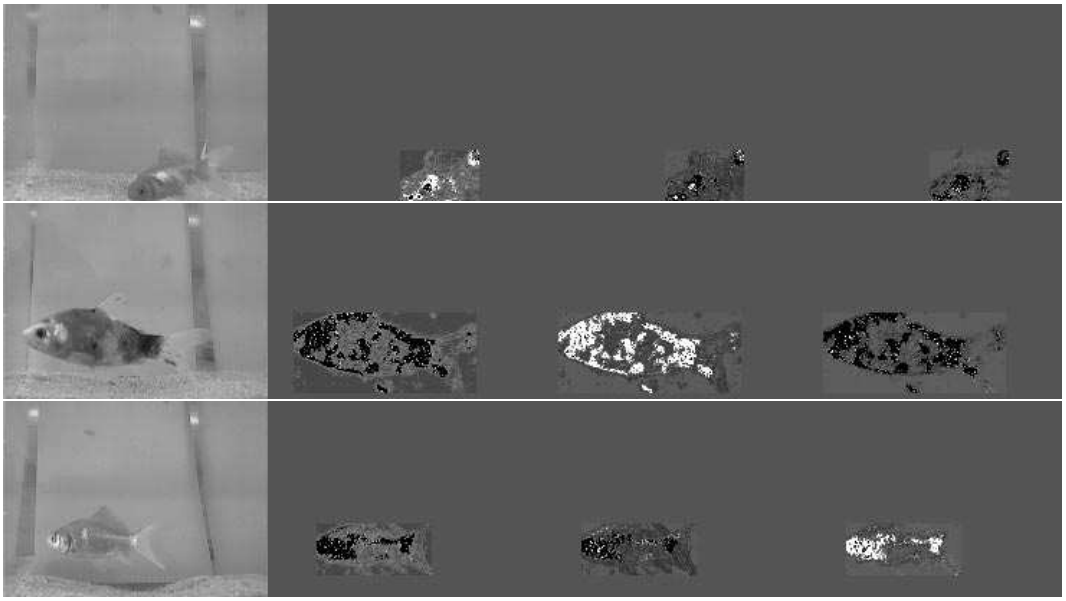


FIG. 7.5 – *Exemples des cartes des probabilités calculées sur des extraits des séquences de test. Les images originales apparaissent dans la première colonne. Les colonnes suivantes correspondent respectivement aux cartes de probabilité que l'action observée corresponde au premier, deuxième ou troisième poisson. Les zones sombres indiquent une probabilité faible, et les zones blanches indiquent une probabilité proche de un.*

La phase de vérification évalue les poissons présents dans l'image en se fondant sur la localisation donnée par la phase de prédiction. L'algorithme de reconnaissance fondé sur le même paradigme de prédiction–vérification et présenté à la section 6.4.2 est appliqué aux points sélectionnés pendant la première phase de l'algorithme.

Le résultat obtenu est une intersection des hypothèses prédites pendant la première phase et des hypothèses générées pendant la deuxième phase. Le paragraphe suivant décrit plus précisément un expérience de reconnaissance.

Expérimentations La figure 7.6 montre une exemple de reconnaissance correcte d'un poisson. Une caractéristique locale du poisson est retrouvée sur l'image de test et une pose approximative du poisson est obtenue.

Plus généralement, la reconnaissance peut être évaluée sur l'ensemble des images d'une base de test qui est différente des bases d'apprentissage. L'éclairage n'est pas spécifiquement contrôlé. Les images de test ne font apparaître qu'un seul poisson à la fois. Les poissons sont nommés **a**, **b** et **c** (voir figure 7.4). La première phase de la reconnaissance, la prédiction, retourne, pour chaque image, l'une des hypothèses suivantes : (**a**), (**b**), (**c**), (**a** or **b**), (**a** or **c**), (**b** or **c**), and (**a** or **b** or **c**). Le processus de reconnaissance est évalué suivant quatre classes pour chaque image : la reconnaissance est correcte (*succès*), la

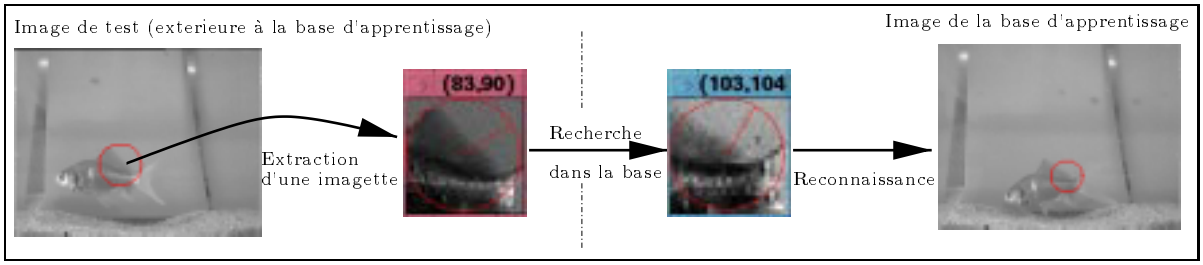


FIG. 7.6 – *Reconnaissance d'une image de test par appariement d'une sous-fenêtre d'une image modèle.*

reconnaissance est partielle (*succès partiel*), la reconnaissance est fausse (*échec*) ou un abandon (*rejet*). Par exemple, sur une image du poisson **a**, l'hypothèse **a** est un succès, les hypothèses (**a** ou **b**) et (**a** ou **c**) sont des succès partiels, l'hypothèse (**a** or **b** or **c**) est un rejet et toute autre hypothèse est un échec. La reconnaissance est évaluée suivant ces quatres classes de résultats.

Un premier tableau montre les taux de reconnaissance aux deux stades de l'algorithme : après la phase de génération d'hypothèses et après la phase vérification de ces hypothèses.

Technique	*succès*	*succès partiel*	*échec*	*rejet*
Prédiction	25.7	41.7	29.3	3.3
Vérification	46.7	19.7	10.3	22.1

La reconnaissance est fortement améliorée par la phase de vérification des hypothèses fournies par la technique statistique. Cette amélioration est obtenue par la mise en correspondance de l'image avec l'une des images de la base d'apprentissage. Cet appariement fournit une hypothèse sur la pose du poisson observé. En cas d'incohérence entre les deux phases de la reconnaissance, l'image est rejetée. L'utilisation des deux techniques supprime la moitié des ambiguïtés liées aux *succès partiels*.

Un deuxième tableau montre les résultats différenciés des trois poissons :

Technique	*succès*	*succès partiel*	*échec*	*rejet*
a	46.2	22.0	9.2	20.8
b	59.4	16.8	3.2	19.2
c	34.4	20.2	18.6	26.2

Ce tableau montre que le dernier poisson donne des résultats très inférieurs aux autres. Ceci peut s'expliquer par sa texture qui est très spéculaire et donc instable. Il faut noter que l'apprentissage non contrôlé ne garantit pas que chacun des poissons aient été observés sous un nombre de points de vue suffisant. Ainsi, une partie des échecs est lié à la présence de points de vue inconnus dans les bases de test. De plus, la reconnaissance est évaluée en continu sur des séquences d'images mais, ici, aucun filtrage temporel n'est effectué pour

supprimer les hypothèses aberrantes. L'utilisation d'un filtrage approprié sur les données comme un filtrage de KALMAN [BS89] améliorerait fortement les résultats.

Le succès de cette expérience difficile démontre les capacités importantes des stratégies de reconnaissance fondées sur les réponses de champs réceptifs locaux spatiaux et spatiaux–temporels. La stratégie de coopération fondée sur le paradigme prédiction–vérification a fortement amélioré le score de reconnaissance.

Chapitre 8

Conclusions et perspectives

8.1 Contributions principales

Dans cette thèse, nous avons étudié et mis en place un système complet de modélisation et de reconnaissance d'objets 3D quelconques. Les performances de ce système sont évaluées expérimentalement sur des bases contenant jusqu'à une centaine d'objets représentés par plus de 1800 images dans une base de modèles.[1] Pour un tel nombre d'objets, le coût mémoire lié au stockage de la base des modèles n'est pas apparu gênant pour les expériences. Deux applications ont validées l'approche proposée: un système de localisation automatique de robot mobile sur un chemin visuel préalablement enregistré sans ajout de repères artificiels et un système coopératif de reconnaissance de poissons rouges. Quelques résultats utilisant des filtres couleurs motivent une extension directe consistant à utiliser la technique sur une base de dérivées de Gaussiennes en couleur pour augmenter le nombre d'objets modélisables simultanément et pour atteindre une robustesse plus importante aux variations d'éclairage.

L'étude prépondérante de cette thèse concerne l'évaluation et le choix d'une base de descripteurs locaux pour la modélisation d'images par caractéristiques locales. Cette étude a permis d'écarter la base de filtres obtenus par Analyse en Composantes Principales car limités par l'absence d'invariance par rapport à la position de la caméra malgré la robustesse obtenue par l'utilisation d'un masquage par une Gaussienne. Une base très discriminante est sélectionnée pour ses propriétés d'invariance à l'orientation et à l'échelle, elle est fondée sur les dérivées de Gaussiennes jusqu'à l'ordre 3. L'utilisation de la théorie de Freeman [FA91] sur les filtres orientables permet d'atteindre l'invariance à l'orientation et une extension novatrice de la théorie de Lindeberg [Lin98] permet d'obtenir une invariance par rapport aux variations d'échelle jusqu'à un facteur 3. La détection d'une ou plusieurs échelles caractéristiques pour chaque point d'une image permet de position-

1. Les expériences sont effectuées sur une station de travail standard : un PC-Pentium II à 333Mhz doté de 256 Mo de mémoire vive.

ner le paramètre d'échelle des dérivées de Gaussiennes les rendant ainsi invariantes. La technique de sélection d'échelle proposée peut être étendue à une large gamme d'applications en vision par ordinateur pour lesquels l'échelle doit souvent être connue a priori. Des travaux actuels étudient l'extension de cette technique pour la sélection d'échelle sur l'axe temporel de façon à obtenir une invariance à la vitesse des activités.

De plus, la modélisation structurelle intégrale par caractéristiques locales des images a permis de définir deux stratégies de reconnaissance : l'une basée sur un vote et l'autre basée sur le paradigme prédiction–vérification. L'algorithme utilisé est une adaptation de travaux plus anciens [Aya83, Sou83] sur la reconnaissance par appariement de segments qui n'avaient pas aboutit par cause d'instabilité des caractéristiques de type segment. Ces techniques ont permis un taux de reconnaissance important sur les images de test mais il apparaît utile d'approfondir cet aspect en mettant l'accent sur le superviseur du système de reconnaissance qui doit diriger les recherches et mieux gérer le problème de l'occultation.

8.2 Perspectives

Un problème important est la définition d'un superviseur permettant de diriger l'analyse et la reconnaissance d'une image à partir des outils fournis par cette thèse. Cette définition est peu évidente car les scores de reconnaissance associés aux différentes hypothèses sont difficiles à quantifier et comparer. Une évaluation minimale de ce score est proposée, elle est fondée sur des prédicats logiques qui prennent une décision binaire quand à l'appartenance d'une caractéristique à un modèle et sur la cohérence spatiale entre appariements. La sélection des hypothèses par ce système de prédicats permet de trier les hypothèses en se fondant sur un ordre lexicographique sur un couple (nombre d'appariement, distance moyenne). Ceci aboutit à une expertise assez difficile de ce qu'est un bon score : comment comparer une hypothèse comprenant 5 appariements de faible qualité avec une hypothèse comprenant 3 appariements de très bonne qualité ? Une solution pour évaluer ce score consiste à modéliser intégralement le processus de reconnaissance de façon probabiliste comme SCHMID [Sch99] l'a proposé pour son système de reconnaissance. Dans ce cadre, les prédicats logiques sont remplacés par des probabilités d'appartenance et le score est donné par une probabilité cumulée. Ceci pose néanmoins des difficultés pratiques : quelle valeur donner à une absence d'appariement, de façon théorique, une valeur nulle pourrait être donnée mais de façon pratique, il est nécessaire d'évaluer une probabilité d'absence d'appariement. D'autre part, il est difficile de paramètrer les fonctions de probabilités correspondants aux différentes phases de la reconnaissance : certaines probabilités sont prépondérantes sur d'autres dans un tel produit et donc une normalisation est nécessaire mais difficile à évaluer et dépendant du contexte. Dans le cas de grandes occultations partielles éventuelles, une absence d'appariement doit modérément influencer une reconnaissance : le nombre d'appariements n'est plus un critère fiable.

La phase d'apprentissage des modèles des objets est une faiblesse importante des approches fondées sur l'apparence. En effet, ces approches nécessitent d'enregistrer les images correspondants à un échantillonnage fin de la sphère des vues d'un objet pour autoriser sa reconnaissance pour n'importe quel point de cette sphère. En pratique, cet enregistrement de toutes les apparences d'un objet n'est possible que sous des restrictions importantes de cette sphère ou alors en laboratoire. L'application à la modélisation de poissons rouges a montré la faisabilité de l'approche sans contrôle sur les objets mais ceci implique une représentation fortement redondante par l'absence de sélection des points de vue pertinents et différents entre eux. Une deuxième approche peut être employée qui s'appuie sur des résultats récents consistant à construire et générer un modèle de type géométrique de l'objet à reconnaître puis à générer automatiquement toutes les images d'un échantillonnage dense de la sphère des vues. Ainsi, la phase d'enregistrement des images de l'objet peut être largement diminuée : les images nécessaires à la modélisation de l'objet peuvent, alors, être générées automatiquement. Cette approche a déjà été proposée sur le cas, plus simple, d'un capteur laser 1D par WALLNER [Wal97]. À partir de quelques scans d'un environnement, il a crée une carte de celui–ci puis il a pu générer des scans synthétiques correspondants à un échantillonnage dense de la scène à représenter.

Une extension intéressante de ce travail est son extension à la reconnaissance de classes d'objets plus larges. Ces classes doivent être des classes visuelles et non des classes fonctionnelles comme, par exemple, une chaise. Il est possible d'envisager des classes comme des visages pour lesquelles la variabilité visuelle interne à la classe reste faible. Une stratégie possible consiste à utiliser un apprentissage supervisé des éléments caractéristiques du visage sur une grande base de personnes différentes en utilisant les descripteurs locaux proposés dans cette thèse. La reconnaissance et, surtout, le positionnement d'un nouveau visage sera, dans ce cas, obtenu par appariement de descripteurs et utilisation de contraintes spatiales entre les éléments du visage.

Une autre perspective de ce travail est son extension à des problèmes d'Interface Homme–Machine (ou IHM). Ceci demande une accélération du système qui peut être obtenue sur deux points : le filtrage systématique sur une large gamme d'échelles peut être implémentée à fréquence vidéo sur une architecture matérielle spécifique et la phase de recherche peut facilement être accélérée par une division de la structure de données contenant la base des modèles sur plusieurs processeurs. Actuellement, une application est développée dans le projet qui consiste à reconnaître des objets manipulés par un utilisateur comme commandes à un système interactif.

Annexe A

Bases d'Images

Ce chapitre présente les bases d'images utilisées pour l'évaluation des techniques proposées dans cette thèse.

A.1 La base de Columbia [NNM96b]

La base de Columbia (COIL) est une base de 100 objets 3D par chacun desquels 72 images ont été enregistrées. Cette base a la particularité d'être prise de façon très contrôlée au niveau éclairage (pas d'ombre). De plus, il n'y a pas de rotation autour de l'axe optique de la caméra (appelée ici rotation 2D). Un axe de rotation 3D est géré par l'utilisation de multiples images pour chaque objet. Cette base est connue pour ses qualités et sa facilité : son utilisation, insuffisante par elle-même, démontre l'application possible des algorithmes de reconnaissances avec l'extension à d'autres bases plus difficiles. Les images originales sont en couleur mais la majorité des tests effectués dans cette thèse utilisent une version de cette base convertie en niveaux de luminance. La figure A.1 montre l'ensemble des 100 objets de la base puis la figure A.2 montre pour un objet une série d'images extraites.

Cette base permet une validation simples des algorithmes de reconnaissance. La validation est effectuée par l'apprentissage d'une partie des images puis par l'évaluation de la reconnaissance sur le reste des images. 72 points de vues sont disponibles pour chacun des objets, soit une image tous les 5°. L'apprentissage est effectué en utilisant un écart de 20° entre images soit 18 images par objet. Les 54 images restantes sont disponibles pour l'évaluation de la reconnaissance.

A.2 Base MOVI [Gro98]

La base d'images MOVI contient des séries d'images présentant un paramètre variable. Deux séries d'images ont été utilisées pour les expérimentations de cette thèse. D'une part, des images pour lesquelles une rotation autour de l'axe optique de la caméra est

FIG. A.1 – *Les 100 objets de la base de Columbia.*

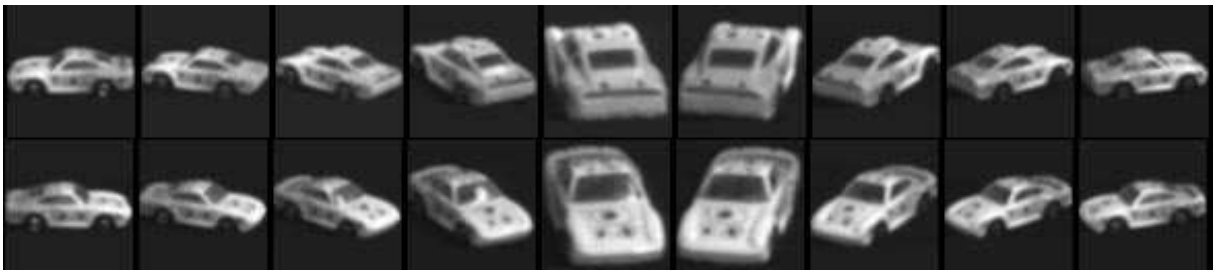


FIG. A.2 – *Sous-ensemble des 72 images de l'objet 8 de la base Columbia.*

effectuée (validation de l'invariance à l'orientation 2D) et d'autre part, une série d'images où l'intensité de l'éclairage est continuement modifiée.

La figure A.3 présente les 8 objets des séquences *c2_vp_rz_s** et la figure A.4 présente un extrait de la séquence *c2_vp_rz_s2*. La figure A.4 présente un extrait d'une séquence complète. Cette base d'images avec rotation autour de l'axe optique permet de valider la robustesse des algorithmes proposés aux variations de l'orientation 2D. Une image par objet est apprise puis la reconnaissance est évaluée en fonction du changement d'orientation.

Une série d'images où l'intensité de l'éclairage est variable a été utilisée pour valider la normalisation par l'énergie comme technique permettant de rendre la reconnaissance plus robuste aux variation d'éclairage. La figure A.5 montre l'ensemble des images de cette série.

FIG. A.3 – *sept objets des séquences c2_vp_rz_s*.*

FIG. A.4 – *Extrait de la séquence c2_vp_rz_s2.*

A.3 Base d'images avec variation d'échelle

Une contribution majeure de cette thèse est la sélection automatique des paramètres d'échelle de descripteurs locaux. Cette sélection automatique est particulièrement intéressante dans le cas de la reconnaissance d'objets avec des variations d'échelle importantes. L'évaluation de la robustesse de la reconnaissance par rapport aux variations d'échelle est évaluée sur la base montrée sur la figure A.6. Cette base est issue des expérimentations de SCHIELE [Sch97] pour sa thèse. Cette base présente 28 objets de caractéristiques très différentes : rigides et non rigides, planaires et tridimensionnels avec pour certaines images de grandes différences de profondeurs impliquant des déformations projectives importantes. La figure A.7 montre une série d'images pour l'un des objets de la base : l'objet "chocos". L'évaluation de la robustesse par rapport aux variations d'échelles est effectuée par un apprentissage des images d'échelle intermédiaire puis par l'évaluation de la reconnaissance sur les images restantes comme fonction du changement d'échelle observé.

FIG. A.5 – *Séquence c2_1l_li_s1.*

FIG. A.6 – *28 objets.*

FIG. A.7 – *images de l'objet Chocos.*

Annexe B

Détails d'implémentation

B.1 Structure Arborescente de stockage de points nD

La structure de données est implémentée comme une classe *C++* récursive. Une instance de la classe est soit une feuille (une liste de points de l'espace de description $\mathcal{A}$), soit un noeud (un tableau de 4 arbres). La section B.1.1 présente le format de la classe utilisée pour stocker l'arbre. La section B.1.2 présente succintement l'algorithme d'ajout d'un point à l'arbre. Puis les algorithmes de recherche et de confirmation d'hypothèse dans l'arbre sont présentés.

B.1.1 Attributs de la classe Arbre :

- *entier* niveau : niveau du sous-arbre dans l'arbre complet. Le niveau de la racine de l'arbre est un.
- *entier* nb : nombre de points dans le sous-arbre.
- *type-arbre* type : type énuméré stockant le type du sous-arbre (un tableau de sous-arbre T , une liste de points L ou rien).
- *Union (tableau de size arbres | liste de nb projections)* data : liste de points stockés soit somme une liste, soit, dès que la liste dépasse une certaine taille, par un tableau de 4 sous-arbres.

B.1.2 Algorithme d'ajout d'un point $\mathcal{M}$:

Un point $\mathcal{M}$ est composée de *nbdim* coordonnées $\mathcal{M}[i]$. Chaque $\mathcal{M}[i]$ est stockée sur un octet de 0 à 255. L'intervalle $[0:255]$ est décomposé en 4 sous-intervalles $[0:63], [64:$

127], [128 : 191], [192 : 255]. Les coordonnées sont sélectionnées par ordre croissant. L'intervalle [0:255] est une représentation à virgule fixe de la projection sur une dimension. L'ajout d'un point est fait simplement en l'ajoutant à la feuille correspondante de l'arbre. Un seuil sur le nombre de points dans une feuille permet de subdiviser chaque feuille dès qu'elle dépasse une taille donnée. L'algorithme précis est présenté sur la table B.1.

```
Procédure add(arbre A, point M , niveau i)
Début
   Selon
      ( (A.type = L et nb < seuil) ou i = nbdim)
         ajoute M à la liste A.data
      (A.type = T)
         déterminer le sous-intervalle de M[i] → s = 0,1,2 ou 3.
         appeler add(A.sous − arbre[s], M, i + 1)
      sinon Erreur()
   fin Selon
Fin
```

TAB. B.1 – *Algorithme d'ajout d'un point à la structure de donnée (pseudo pascal).*

B.1.3 Algorithme de recherche d'un point $\mathcal{M}$:

L'algorithme nécessite en entrée une structure arbre A, un point $\mathcal{M}$ et un seuil $distanceMax$. Pour limiter la recherche dans des zones de l'espace de représentation très denses, une seuil K limite la recherche aux K points les plus proches du point cherché.

L'algorithme précis est présenté sur la table B.2.

B.1.4 Algorithme de confirmation d'une hypothèse par un point $\mathcal{M}$:

La stratégie de prédiction–vérification proposée dans ce chapitre implique de pouvoir confirmer la présence d'un point à une position donnée d'un objet. L'algorithme de confirmation parcourt l'arbre de façon similaire à une recherche mais s'arrête dès qu'une hypothèse correcte est trouvée. Il est présenté sur la table B.3.

B.2 Algorithmes de détection des maxima pour la sélection automatique de l'échelle

La courbe représentant l'évolution du Laplacien normalisé présente dans certains cas plusieurs maxima locaux distincts : l'exemple proposé sur la figure B.1 montre une image dont un point a deux maxima visibles. Les cercles représentent l'envergure des laplaciens maxima détectées. Le rayon des cercles est 2σ. La détection des maxima locaux correspondants aux échelles caractéristiques peut être effectuée grâce l'utilisation d'un automate d'états finis.

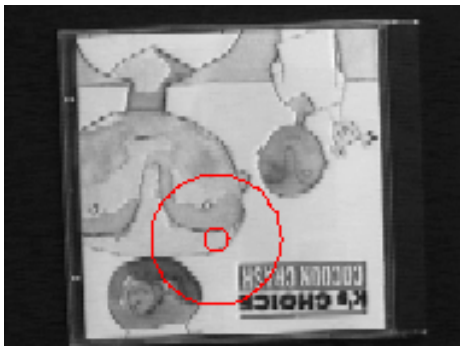


FIG. B.1 – *Deux échelles caractéristiques sont disponibles pour le point présenté.*

B.2.1 Algorithme par Automate d'États Finis

La figure B.2 présente l'automate d'états finis utilisé pour la détection des maxima. Cette automate dispose de 3 états : un état d'**A**ttente, un état **M**ontée et un état **D**escente. La séquence des valeurs du Laplacien normalisé est parcourue de façon croissante. A chaque étape, une valeur courante C et une valeur précédente P sont définies ainsi que l'état de l'automate et un triplet (Gauche, Centre, Droite) en construction. L'état est modifiée en comparant les valeurs C et P et en vérifiant la validé globale de la descente ou de la montée précédente. Cette validité est définie comme la hauteur minimal d'une montée ou d'une descente. La figure B.2 montre une illustration de cet automate et son application à la détection des deux maxima locaux de la figure B.1. Nous utilisons cet automate de façon systématique pour détecter les échelles caractéristiques des images observées et obtenir ainsi une invariance à l'échelle. Cet algorithme présente l'inconvénient de nécessiter une évaluation dense des valeurs du Laplacien. L'aspect Gaussien de la courbe d'évolution du Laplacien motive un algorithme plus précis : une approximation de cette courbe par une mixture de Gaussienne qui permet l'évaluation des maxima à partir d'un échantillonnage plus faible de la courbe en conservant une grande précision (voir [ZHPZ96] par exemple).

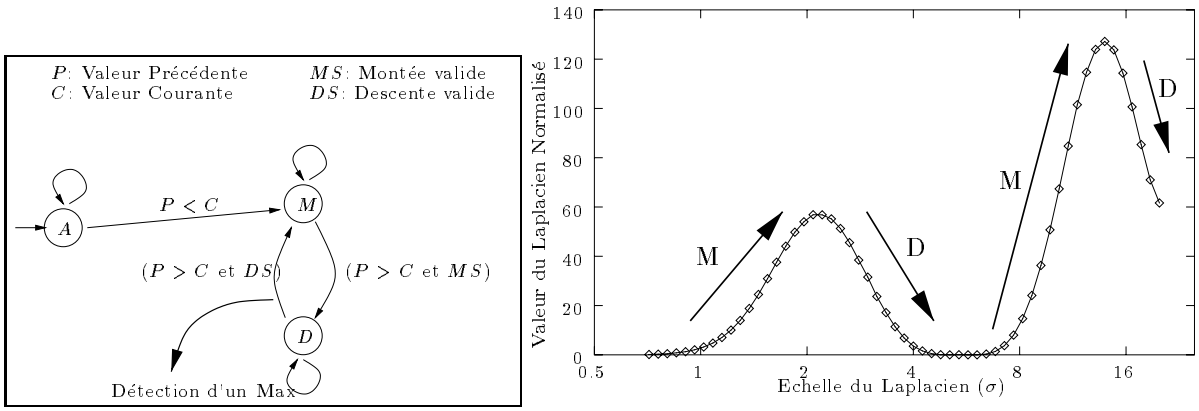


FIG. B.2 – *Exemple de détection de maxima du Laplacien Normalisé par un automate d'États Finis. Les états successifs de l'automate sont indiqués sur la courbe à droite.*

```
listePoints Fonction Search(arbre A, point M, double SeuilDist, entier K)
Début
    initialiser la liste resultat comme liste vide.
    Selon
        (A.type = L)
            Pour chaque couple (point Ω_k, hypothèse H_k) de la list A.data
                Calculer d = Distance(Ω_k , M)
                Si d < SeuilDist alors
                    Ajouter le couple (H_k, d) à resultats
                    Si ( Longueur(resultats) > K alors
                        supprimer le K_ieme élément
                        mettre à jour le seuil SeuilDist.
        (A.type = T)
            Pour chacun des 4 sous-arbres (A_1, A_2, A_3, A_4) de A
                Calcul d = distance partielle optimiste de M avec A_i
                // c'est-à-dire distance entre M et le point de
                // A_i le plus proche possible de M .
                Si (d < SeuilDist) alors
                    Appeler Search(A_i, M , SeuilDist, K)
                    Ajouter le résultat du Search() à resultat.
        sinon Erreur()
    fin Selon
    Retourner resultat.
Fin.

Appel de: Search(A, M , SeuilDist, K)
```

TAB. B.2 – *Algorithme de recherche d'un point $\mathcal{M}$ (pseudo pascal).*

```
Booléen Fonction Confirm(arbre A, point M , double SeuilDist,
        hypothèse H_pred, Fonction Booléenne Compatible())
Début
   Selon
      (A.type = L)
         Pour chaque couple (point Ω_k, hypothèse H_k) de la list A.data
            Calculer d = Distance(Ω_k , M)
            Si d < SeuilDist et Compatible(H_k , H_pred) alors
               Retourner Vrai
      (A.type = T)
         Pour chacun des 4 sous-arbres (A_1, A_2, A_3, A_4) de A
            Calcul d = distance partielle optimiste de M avec A_i
            // c'est-à-dire distance entre M et le point de
            // A_i le plus proche possible de M .
            Si ( (d < SeuilDist) et
               (Confirm(A_i, M , SeuilDist, H_pred, Compatible())) ) alors
               retourner Vrai
      sinon Faux
   fin Selon
Fin.

Appel de: Confirm(A, M , SeuilDist, H_pred, Compatible())
```

TAB. B.3 – *Algorithme de confirmation d'une hypothèse H_{pred} par une fenêtre $\mathcal{M}$ (pseudo pascal).*

Annexe C

Évaluation des dérivées de Gaussiennes par filtrage récursif

La représentation des objets par caractéristiques locales nécessite de pouvoir évaluer efficacement les dérivées de Gaussiennes en tout point des images. Ces filtres sont définis par leur formule analytique et leur synthèse pour un signal discret est une approximation. Elles présentent des propriétés intéressantes pour leur calcul : elles sont séparables suivant les axes x et y ce qui permet d'évaluer ces dérivées par deux filtrages unidimensionnels plus efficaces qu'un filtrage bidimensionnel. Une technique directe pour évaluer ces dérivées consiste à évaluer des masques de convolutions 1D correspondants à ces dérivées : il s'agit de leur réponse impulsionnelle. La convolution par les masques obtenus donnent les dérivées d'une image. Cette technique présente l'inconvénient d'une complexité très importante proportionnelle à la taille du masque ce qui rend très longue l'évaluation de dérivées pour un paramètre d'échelle σ important. Une deuxième propriété des dérivées de Gaussiennes est la possibilité de les synthétiser de façon récursive : complexité très faible et indépendante du paramètre σ.

L'évaluation récursive des dérivées de Gaussiennes est fondée sur une approximation polynomiale de leurs fonctions de transfert. DERICHE [Der92] a proposé un algorithme fondé sur un polynôme d'ordre 4. YOUNG et VAN VLIET [YvV95] proposent un algorithme similaire fondé sur un polynôme d'ordre 3. Cette approximation plus rapide a été utilisée pour évaluer les dérivées de Gaussiennes. Notre implémentation permet sur un PC-pentium à 333Mhz d'effectuer un filtrage par une dérivée de Gaussienne d'une image 512×512 en 0.2 secondes. Néanmoins cette technique présente deux difficultés : discontinuité en $\sigma = 2.5$ et anisotropie.

Correction de la discontinuité en $\sigma = 2.5$ L'algorithme de YOUNG nécessite d'évaluer un paramètre q en fonction du paramètre d'échelle σ. YOUNG propose une approxi-

mation de la fonction $q(\sigma)$ qui n'est pas continue en $\sigma = 2.5$.

$$q(\sigma) = \begin{cases} 0.98711\sigma_0, & \sigma_0 \geq 2.5, \\ 3.97156 - 4.14554\sqrt{1 - 0.26891\sigma_0}, & 0.5 \leq \sigma_0 \leq 2.5 \end{cases} \tag{C.1}$$

Cette équation donne :

$$\lim_{\sigma \to 2.5, \sigma < 2.5} q = 1.598 \tag{C.2}$$

$$\lim_{\sigma \to 2.5, \sigma > 2.5} q = 1.504 \tag{C.3}$$

La figure C.1 montre la courbe $q(\sigma)$ réévaluée et l'approximation de YOUNG et VAN VLIET . L'évaluation de $q(\sigma)$ est effectuée en recherchant pour chaque valeur de σ la valeur optimale du paramètre q. Cette valeur est obtenue en comparant la réponse impulsionnelle du filtre avec le filtre théorique. Une recherche dichotomique en optimisant l'approximation par les moindres carrés donne la valeur de q. L'évaluation de cette fonction nous permet

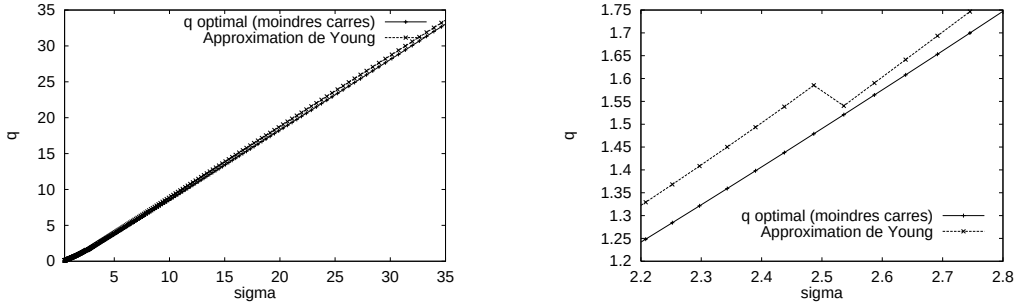


FIG. C.1 – *fonction $q(\sigma)$ avec zoom autour de $\sigma = 2.5$: approximation de* YOUNG *et réévaluation suivant un critère par moindres carrés.*

d'obtenir le paramètre q du filtrage récursif par l'utilisation d'une table de correspondance.

Anisotropie du filtrage récursif La détection de l'échelle est fondée sur une évaluation d'un opérateur Laplacien normalisé sur une large gamme de son paramètre σ. Son évaluation est obtenue en utilisant un filtrage récursif dont la réponse impulsionnelle n'est pas anisotrope comme le montre la figure C.2. Cette anisotropie dégrade la reconnaissance en présence de variations de l'orientation de la caméra. Une amélioration de l'isotropie du filtrage est proposée par les auteurs via l'utilisation d'une approximation de la fonction de transfert des filtres d'ordre 4 et 5. Cette nouvelle approximation permet, selon les auteurs [vVYV98] et avec une augmentation de la complexité algorithmique, de réduire cette anisotropie.

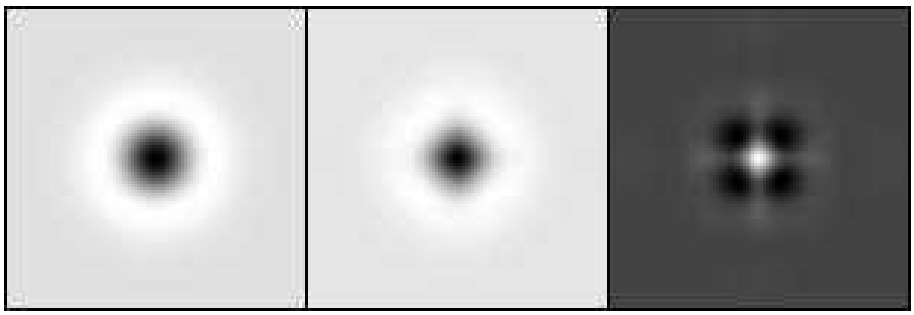


FIG. C.2 – *Réponse impulsionnel d'un Laplacien théorique et récursif, puis images de l'erreur.*

Annexe D

Quelques notations utilisées dans cette thèse

D.1 Notations

- N : nombre d'imagettes utilisées pour l'apprentissage.
- $\mathcal{A}$: Espace de projection ou de description des caractéristiques locales. Cet espace est constitué de 8 à 10 dimensions..
- $\mathcal{M}$: Vecteur de mesures locales. (coordonnées $\mathcal{M}[0-m]$). Il s'agit d'un point de l'espace $\mathcal{A}$ qui représente une imagette.
- m: nombre de dimensions de l'espace de projection $\mathcal{A}$ et, donc, nombre de coordonnées de $\mathcal{M}$.
- $G_\theta^k(x,y,\sigma)$: Filtre de dérivée de Gaussienne. k est l'ordre dérivation. θ est la direction de dérivation. k et θ peuvent être remplacés par $(x, y, xx, xy, yy, xxx, xxy, xyy, yyy)$ pour les dérivations d'ordre 1 à 3 suivant les axes x et y.
- $\Phi = \{\phi_1 \dots \phi_m\}$: filtres obtenus par une Analyse en Composantes Principales sur des imagettes.
- $\Lambda = (\lambda_1 \dots \lambda_m)$: valeurs propres correspondants aux filtres ACP.
- W : fenêtre extraite d'une image ou imagette.
- J : Image. $J(x,y)$ est le point (x,y) de l'image J.
- $\mathcal{J}$: Base d'images J_k utilisée pour la phase d'apprentissage $\mathcal{J} = \{J_1 \dots J_N\}$.

- $L^k_\theta(x,y,\sigma) = G^k_\Theta(x,y,\sigma) * J(x,y)$ est la valeur de la convolution au point (x,y) par le filtre gaussien sur l'image J.
- $\mathcal{L}^k_\theta(x,y,\sigma)$: valeur de la convolution au point (x,y) normalisée par rapport à l'échelle.
- Id : identificateur d'une image et de l'objet correspondant. En pratique Id est un couple (numéro d'objet, numéro de point de vue).

D.2 Vocabulaire

- Fenêtre ou Imagette : l'approche locale utilisée dans cette thèse signifie que les caractéristiques des images sont évaluée sur des sous-images des images complètes. Ces sous-images sont appelées imagettes ou fenêtres selon les cas. Le terme de point et de son voisinage est aussi utilisé.
- Invariance à une classe de transformations T signifie qu'une description ne dépend pas des variations à l'intérieur des classes de T.
- Équivariance à un paramètre P signifie que, sous l'hypothèse d'une connaissance de ce paramètre, il est possible de caller la description de façon à obtenir une description indépendante de ce paramètre. Il s'agit généralement d'un prérequis à l'invariance.
- Champs Réceptif : capteur local mesurant la présence d'une caractéristique particulière.
- Descripteur: filtre ou opérateur définissant une dimension de l'espace de description $\mathcal{A}$.
- Mesure ou Description: scalaire décrivant un voisinage.
- Vecteur de mesures (ou de description) souvent noté $\mathcal{M}$: vecteur de m coordonnées représentant une imagette ou le voisinage d'un point.

Références bibliographiques

[AB91] E. H. Adelson et J. R. Bergen. The plenoptic function and the elements of early vision. Dans M.Landy et J.A.Movshons, éditeurs, *Computational Models of Visual Processing*. MIT Press, Cambridge, 1991.

[AJC97] C. S. Andersen, S. D. Jones, et J. L. Crowley. « Appearance Based Processes for Visual Navigation ». Dans *5th International Symposium on Intelligent Robotic Systems, SIRS'97*, pages 227–236, Royal Institute of Technology, Stockholm, Sweden, juillet 1997.

[Aya83] N. Ayache. « *Un système de vision bidimensionnelle en robotique industrielle* ». Thèse de doctorat , Université de Paris-Sud, Centre d'Orsay, juin 1983. in french.

[BCC99] E. Backmann, V. Colin de Verdière, et J. L. Crowley. « Dense Stereo Matching using Local Appearance ». Dans *Seventh International Symposium for Intelligent Robotics System (SIRS'99)*, University of Coimbra, Portugal, juillet 1999.

[Bla98] J. Blanc. « *Synthèse de nouvelles vues d'une scène 3D à partir d'images existantes* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, GRAVIR – IMAG, janvier 1998.

[BNM98] S. Baker, S. K. Nayar, et H. Murase. « Parametric Feature Detection ». *International Journal of Computer Vision*, 27(1):27–50, 1998.

[Bob95] P. Bobet. « *Tête stéréoscopique, Réflexes oculaires et Vision* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, LIFIA – IMAG, 1995.

[BS89] K. Brammer et G. Siffling. *Kalman–Bucy Filters*. Artech House, 1989.

[CC98a] V. Colin de Verdière et J. L. Crowley. « Reconnaissance d'Objets par Apparence Locale ». Dans *11ème Congrès de Reconnaissance des Formes et d'Intelligence Artificielle, RFIA'98*, volume 2, pages 129–136, janvier 1998. In french.

[CC98b] V. Colin de Verdière et J. L. Crowley. « Visual Recognition using Local Appearance ». Dans *Fifth European Conference on Computer Vision, ECCV'98*, volume 1 de *Lecture Notes in Computer Science*, pages 640–654, Freiburg, Germany, juin 1998. Springer Verlag.

[CC99a] O. Chomat et J. L. Crowley. « Probabilistic Recognition of Activity using Local Appearance ». Dans *Computer Vision and Pattern Recognition (CVPR'99)*, volume 2, 1999.

[CC99b] V. Colin de Verdière et J. L. Crowley. « Local Appearance Space for Recognition of Navigation Landmarks ». *Journal of Robotics and Autonomous System - special issue*, 1999. also published at the Sixth International Symposium for Intelligent Robotic Systems, SIRS'98.

[CCBS97] J. Coutaz, J. L. Crowley, F. Berard, et D. Salber. « EigenSpace Coding as a Means to Support Privacy in Computer Mediated Communication ». Dans *Interact 97*, juillet 1997.

[CCC99] O. Chomat, V. Colin de Verdière, et J. L. Crowley. « Recognizing GoldFish? or Local Scale Selection for Recognition Techniques ». Dans *Seventh International Symposium for Intelligent Robotics System (SIRS'99)*, pages 197–206, University of Coimbra, Portugal, juillet 1999.

[CCC00] V. Colin de Verdière, O. Chomat, et J. L. Crowley. « Sélection Automatique de l'Échelle pour la Reconnaissance par Caractéristiques Locales ». Dans *12ème Congrès de Reconnaissance des Formes et d'Intelligence Artificielle, RFIA'2000*, février 2000. Submission,In french.

[CHK97] O. I. Camps, C-Y. Huang, et T. Kanungo. « Hierarchical Organization of Appearance-Based Parts and Relations for Object Recognition ». Dans *International Conference on Computer Vision and Pattern Recognition*, pages 877–883, 1997.

[Cho99] O. Chomat. « Comparaisons entre filtres de Gabor et Dérivées de Gaussiennes ». Discussions, 1999. Article à paraître dans PAMI.

[Col96] V. Colin de Verdière. « Reconnaissance d'Objets par leurs Statistiques de Couleurs ». Rapport de DEA, Imagerie, Vision et Robotique. ENSIMAG, Projet PRIMA, Laboratoire GRAVIR - IMAG, I.N.P. Grenoble, France, juin 1996. In french.

[CWS98] J. L. Crowley, F. Wallner, et B. Schiele. « Position Estimation using Principal Components of Range Data ». *Robotics and Autonomous Systems*, 23(4):267–276, juillet 1998.

[Der92] R. DERICHE. « Recursively Implementing the Gaussian ant its Derivatives ». Dans *2nd Singapore International Conference on Image Processing*, pages 263–267, septembre 1992.

[Dev98] V. DEVIN. « Techniques Visuelles d'Observation de Gestes Ergotiques ». Rapport de DEA, Imagerie, Vision et Robotique, ENSIMAG, Projet PRIMA, Laboratoire GRAVIR - IMAG, INRIA Montbonnot, France, juin 1998. in french.

[FA91] W. T. FREEMAN et E. H. ADELSON. « The Design and Use of Steerable Filters ». *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 13(9):891–906, septembre 1991.

[FA99] H. FARID et E.H. ADELSON. « Separating Reflections and Lighting Using Independant Components Analysis ». Dans *Computer Vision and Pattern Recognition (CVPR'99)*, volume 1, 1999.

[Fau93] O. FAUGERAS. « *Three-Dimensional Computer Vision : A Geometric Viewpoint* », Chapitre 11. Recognizing and Locating Objects and Places, pages 483–558. The MIT Press, 1993.

[FCF96] G. D. FINLAYSON, S. C. CHATTERJEE, et B. V. FUNT. « Color Angular Indexing ». Dans *ECCV'96, Fourth European Conference on Computer Vision, Volume II*, 1996.

[FF95] B. V. FUNT et G. D. FINLAYSON. « Color Constant Color Indexing ». *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 17(5):522–529, mai 1995.

[FG98] *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition (FG'98)*, Nara, Japan, avril 1998. IEEE Computer Society.

[FSC98] G. D FINLAYSON, B. SCHIELE, et J. L. CROWLEY. « Comprehensive Colour Image Normalization ». Dans *Fifth European Conference on Computer Vision, ECCV'98*, volume 1 de *Lecture Notes in Computer Science*, pages 475–490, Freiburg, Germany, juin 1998. Springer Verlag.

[Fuk90] K. FUKUNAGA. « *Statistical Pattern Recognition* », Chapitre Feature Extraction and Linear Mapping for Signal Representation. Academic Press, School of Electrical Engineering, West Lafayette, Indiana, 1990.

[GLL98] A. GUARDA, C. LE GAL, et A. LUX. « Evolving Visual Features and Detectors ». Dans *International Symposium on Computer Graphics, Image Processing and Vision*, Rio de Janeiro, Brazil, octobre 1998.

[GMD+97] P. Gros, G. Mclean, R. Delon, R. Mohr, C. Schmidt, et G. Mistler. « Utilisation de la couleur pour l'appariement et l'indexation d'images ». Rapport Technique 3269, INRIA, septembre 1997.

[Gro98] P. Gros. « The movi Image Base ». see `http://www.inrialpes.fr/movi/Images/`, 1998.

[HBS91] P. Hancock, R. Baddeley, et L. Smith. « The principal components of natural images ». *Neural: Computation in Neural Systems*, 3:61–70, septembre 1991.

[Hou62] P. V. C Hough. « Method and Means for recognizing complex patterns ». United States Patent 3069654, 1962.

[HS88] C. Harris et M. Stephens. « A combined corner and edge detector ». Dans *Proc. 4th Alvey Vision Conference*, pages 147–151, 1988.

[HS94] G. Healey et D. Slater. « Using Illumination Invariant Color Histogram Descriptors for Recognition ». Dans *International Conference on Computer Vision and Pattern Recognition*, pages 355–360, 1994.

[JF96] B. Jedynak et F. Fleuret. « Reconnaissance d'objets 3D à l'aide d'arbres de classification ». Dans *Image'com 96*, Bordeaux, France, mai 1996. in french.

[Jon97] S. D. Jones. « *Robust Task Achievement* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, GRAVIR – IMAG, mai 1997.

[Kan95] K. Kanatani. *Statistical Optimization for Geometric Computation : Theory and Practice.* Artificial Intelligence Laboratory, Department of Computer Science, Gunma University, Japan, juin 1995.

[Kru97] J. Krumm. « Object Detection with Vector Quantized Binary Features ». Dans *International Conference on Computer Vision and Pattern Recognition*, pages 179–185, 1997.

[KvD84] J. J. Koenderink et A. J. van Doorn. « The structure of images ». *Biological Cybernetics*, 50:363–370, 1984.

[KvD87] J. J. Koenderink et A. J. van Doorn. « Representation of Local Geometry in the Visual System ». *Biological Cybernetics*, 55:367–375, 1987.

[LAC97] J.-L. Lacoume, P.-O Amblard, et P. Comon. *Statistiques d'ordre supérieur pour le traitement du signal.* Traitement du Signal. Masson, 1997.

[Lam98] B. LAMIROY. « *Reconnaissance et modélisation d'objets 3D à l'aide d'invariants projectifs et affines* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, GRAVIR – IMAG, 1998.

[Lan97] Z-D. LAN. « *Méthodes Robustes en Vision: Application aux Appariements Visuels* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, GRAVIR – IMAG, mai 1997.

[LBE97] A. LEONARDIS, H. BISCHOF, et R. EBENSBERGER. « Robust Recognition using Eigimages ». Rapport Technique PRIP-TR-47, Pattern Recogntion and Image Processing group - Vienna University of Technology, juin 1997.

[Lin98] T. LINDEBERG. « Feature Detection with Automatic Scale Detection ». *International Journal of Computer Vision*, 30(2):79–116, 1998.

[Lux86] A. LUX. « La Prédiction - Vérification : une technique d'Intelligence Artificielle pour la Vision par Ordinateur ». Dans *Congrés CIIAM*, pages 49–63, Marseille, France, décembre 1986. in french.

[LZ97] A. LUX et B. ZOPPIS. « An Experimental Multi-language Environment for the Development of Intelligent Robot Systems ». Dans *5th International Symposium on Intelligent Robotic Systems, SIRS'97*, pages 169–174, 1997. more informations at `http://www-prima.imag.fr/Ravi/`.

[MC95] J. MARTIN et J. L. CROWLEY. « Comparison of Correlation Techniques ». Dans *Intelligent Autonomous Systems, IAS'95*, pages 86–93, Karlsruhe, Germany, mars 1995.

[MC97] J. MARTIN et J. L. CROWLEY. « An Appearance–Based Approach to Gesture–Recognition ». Dans *International Conference on Image Analysis and Processing*, Lecture Notes in Computer Science, Florence, Italia, septembre 1997. Springer Verlag.

[MHC98] J. MARTIN, D. HALL, et J. L. CROWLEY. « Statistical Recognition of Parameter Trajectories for Hand Gestures and Face Expressions ». Dans *Workshop on Perception of Human Actions*, Freiburg, Germany, juin 1998.

[MII96] Y. MATSUMOTO, M. INABA, et H. INOUE. « Visual Navigation using View-Sequenced Route Representation ». Dans *International Conference on Robotics and Automation*, volume 1, pages 83–88. IEEE, 1996.

[MN95] H. MURASE et S. K. NAYAR. « Visual Learning and Recognition of 3D Objects from Appearance ». *International Journal of Computer Vision*, 14:5–24, 1995.

[Mor81] H. MORAVEC. « Rover Visual Obstacle Avoidance ». Dans *IJCAI*, pages 785–790, 1981.

[NN97] S. A. NENE et S. K. NAYAR. « A Simple Algorithm for Nearest Neighbor Search in High Dimensions ». *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 19(9):989–1003, septembre 1997.

[NNM96a] S. K. NAYAR, S. A. NENE, et H. MURASE. « Subspace Methods for Robot Vision ». *IEEE Transactions on Robotics and Automation*, 12(5):750–758, octobre 1996.

[NNM96b] S. A. NENE, S. K. NAYAR, et H. MURASE. « Columbia Object Image Library (COIL-100) ». Rapport Technique, Columbia University, New York, février 1996.

[NS98] R. C. NELSON et A. SELINGER. « A Cubist approach to Object Recognition ». Dans *International Conference on Computer Vision (ICCV'98)*, janvier 1998.

[OI96] K. OHBA et K. IKEUCHI. « Recognition of the Multi Specularity Objects for Bin-picking Task ». Dans *IEEE Internationnal Conference on Intelligent Robotic Systems, IROS'96*, volume 3, pages 1440–1448, 1996.

[OI97] K. OHBA et K. IKEUCHI. « Detectability, Uniqueness, and Reliability of Eigen Windows for Stable Verification of Partially Occluded Objects ». *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 19(9):1043–1048, septembre 1997.

[PC98] F. POURRAZ et J. L. CROWLEY. « Continuity Properties of the Appearance Manifold for Mobile Robots Position Estimation ». Dans *Symposium for Intelligent Robotics Systems, SIRS'98*, pages 251–260, Edinburgh, United Kingdom, juillet 1998.

[PFTV86] W. H. PRESS, B. P. FLANNERY, S. A. TEUKOLSKY, et W. T. VETTERLING. « *NUMERICAL RECIPES The Art of Scientific Computing* », Chapitre 11. EigenSystems. Cambridge University Press, 1986.

[Pou98] F. POURRAZ. « Estimation de position d'un robot mobile par projection dans un espace de composantes principales ». Rapport de DEA, Imagerie, Vision et Robotique. ENSIMAG, juin 1998. In french.

[PPP98] L. PALETZTA, M. PRANTL, et A. PINZ. « Reinforcement Learning for Autonomous Three-Dimensional Object Recognition ». Dans *Symposium on Intelligent Robotics Systems, SIRS'98*, pages 63–81, Edinburgh, United Kingdom, juillet 1998.

[RB95] R. P. N Rao et D. H Ballard. « An Active Vision Architecture based on Iconic Representations ». *Artificial Intelligence Journal*, 78:461–505, 1995.

[Rob65] L. Roberts. « *Optical and electro-optical information processing* », Chapitre Machine perception of three-dimensional solids, pages 159–197. MIT press, 1965.

[SB91] M. J. Swain et D. H. Ballard. « Color Indexing ». *International Journal of Computer Vision*, 7(1):11–32, 1991.

[SC96] B. Schiele et J. L. Crowley. « Object Recognition using Multidimensional Receptive Field Histograms ». Dans *ECCV'96, Fourth European Conference on Computer Vision, Volume I*, pages 610–619, avril 1996.

[Sch96] C. Schmid. « *Appariement d'images par invariants locaux de niveaux de gris* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, GRAVIR – IMAG, 1996.

[Sch97] B. Schiele. « *Reconnaissance d'Objets utilisant des Histogrammes Multidimensionnels de Champs Réceptifs* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, GRAVIR – IMAG, juillet 1997.

[Sch99] C. Schmid. « A Structural Probabilistic Model for Rrecognition ». Dans *Computer Vision and Pattern Recognition (CVPR'99)*, volume 2, pages 485–490, juin 1999.

[SK87] I. Sirovich et M. Kirby. « Low-Dimensional Procedure for the Caracterization of Human Faces ». *J. Opt. Soc Am. A*, 4(3):519–524, mars 1987.

[Sou83] V. Souvignier. « *PVV. un système d'interprétation d'images par prédiction et vérification* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, LIFIA – IMAG, juin 1983.

[ST94] J. Shi et C. Tomasi. « Good Features to Track ». Dans *International Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[Str99] O. Strauss. « Use of the Fuzzy Hough transform towards reduction of the precision/uncertainty duality ». *Pattern Recognition*, 32(11):1911–1922, 1999.

[TK91] C. Tomasi et T. Kanade. « Shape and Motion from Image Streams: a Factorization Method, 2. Point Features in 3D Motion ». Rapport Technique CMU-CS-91-105, School of Computer Science, Carnegie Mellon University, janvier 1991.

[TP91] M. Turk et A. Pentland. « Eigenfaces for Recognition ». *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[Tur91] M. A. Turk. « *Interactive-Time Vision: Face Recognition as a Visual Behavior* ». Thèse de doctorat , Carnegie Mellon University, septembre 1991.

[VSC99] W. E. Vieux, K. Schwerdt, et J. L. Crowley. « Face-tracking and Coding for Video-Compression ». Dans *First International Conference on Computer Vision Systems, ICVS'99*, Las Palmas, Spain, janvier 1999.

[vVYV98] L. J. van Vliet, I. T. Young, et P. W. Verbeek. « Recursive Gaussian Derivatives Filters ». Dans *14th International Conference on Pattern Recognition (ICPR'98)*, volume 1, pages 509–514, Brisbane, Australia, août 1998. IEEE Computer Society Press.

[Wal97] F. Wallner. « *Position estimation for a mobile robot from principal components of laser range data* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, GRAVIR – IMAG, octobre 1997.

[WSV99] N. Winters et J. Santos-Victor. « Omni-directional Visual Navigation ». Dans *Symposium for Intelligent Robotics Systems, SIRS'99*, pages 109–118, University of Coimbra, Portugal, juillet 1999.

[You85] R. A. Young. « The Gaussian Derivative Theory of Spatial Vision: Analysis of Cortical Cell Receptive Field Line–Weighting Profiles ». Rapport Technique GMR-4920, General Motors Research Laboratories, mai 1985.

[YvV95] I. Young et L. van Vliet. « Recursive implementation of the Gaussian filter ». *Signal Processing*, 44:139–151, 1995.

[ZHPZ96] X. Zhuang, Y. Huang, H. Palaniappan, et Y. Zhao. « Gaussian Mixture Density Modeling, Decomposition, and Applications ». *IEEE Transactions on Image Processing*, 5(9):1293–1302, septembre 1996.

[Zop97] B. Zoppis. « *Outils pour l'Intégration et le Contrôle en Vision et Robotique Mobile* ». Thèse de doctorat , Institut National Polytechnique de Grenoble, juin 1997.

L'implémentation de la technique présentée ainsi que les expériences effectuées pour cette thèse sont basées sur l'utilisation de l'environnement multi-languages ravi développé au sein du projet prima par Augustin Lux , Bruno Zoppis , Claude Poizat et Christophe Le gal [LZ97, Zop97].

Index des auteurs cités