Fiabilité et précision en stéréoscopie : application à l'imagerie aérienne et satellitaire à haute résolution

Neus Sabater

École Normale Supérieure de Cachan

THÈSE

présentée par

Neus SABATER

pour obtenir le grade de

DOCTEUR DE L'ÉCOLE NORMALE SUPÉRIEURE DE CACHAN

Spécialité : Mathématiques Appliquées

Fiabilité et précision en stéréoscopie Application à l'imagerie aérienne et satellitaire à haute résolution

Reliability and accuracy in stereovision Application to aerial and satellite high resolution images

---

Soutenue le 7 décembre 2009 devant le jury composé de :

Andrés ALMANSA	Telecom ParisTech	Directeur
Gwendoline BLANCHET	CNES	Examinateur
Antonin CHAMBOLLE	École Polytechnique	Rapporteur
Tomeu COLL	Universitat de les Illes Balears, Espagne	Invité
Jean-Marc DELVIT	CNES	Examinateur
Yann GOUSSEAU	Telecom ParisTech	Invité
Renaud KERIVEN	École des Ponts ParisTech	Rapporteur
Yves MEYER	École Normale Supérieure de Cachan	Président
Lionel MOISAN	Université Paris Descartes	Invité
Jean-Michel MOREL	École Normale Supérieure de Cachan	Directeur
Bernard ROUGÉ	CESBIO	Examinateur
Lenny RUDIN	Cognitech, Inc., Étas-Unis	Rapporteur

Résumé

Cette thèse se situe dans le cadre du projet MISS (Mathématiques de l'Imagerie Stéréoscopique Spatiale) monté par le CNES en collaboration avec plusieurs laboratoires universitaires en 2007. Ce projet se donne l'objectif ambitieux de modéliser un satellite stéréoscopique, prenant deux vues non simultanées mais très rapprochées de la Terre en milieu urbain. Son but principal est d'obtenir une chaîne automatique de reconstruction urbaine à haute résolution à partir de ces deux vues.

Ce projet se heurte toutefois à des problèmes de fond que la présente thèse s'attache à résoudre. Le premier problème est le rejet des matches qui pourraient se produire par hasard, notamment dans les zones d'ombres ou d'occlusion, et le rejet également des mouvements au sol (véhicules, piétons, etc.) La thèse propose une méthode de rejet de faux matches basée sur la méthodologie dite *a contrario*. On montre la consistance mathématique de cette méthode de rejet, et elle est validée sur des paires simulées exactes, sur des vérités terrain fournies par le CNES, et sur des paires classiques de benchmark (Middlebury). Les matches fiables restants représentent entre 40% et 90% des pixels selon les paires testées.

Le second problème de fond abordé est la précision. En effet le type de stéréoscopie envisagé exige un très faible angle entre les deux vues, qui sont visuellement presque identiques. Pour obtenir un relief correct, il faut effectuer un recalage extrêmement précis, et calibrer le niveau de bruit qui permet un tel recalage. La thèse met en place une méthode de recalage subpixélien, qui sera démontrée être optimale par des arguments mathématiques et expérimentaux. Ces résultats étendent et améliorent les résultats obtenus au CNES par la méthode MARC. En particulier, il sera montré sur les images de benchmark Middlebury que la précision théorique permise par le bruit correspond bien à celle obtenue sur les matches fiables.

Bien que ces résultats soient obtenus dans le cadre d'un dispositif d'acquisition précis (stéréoscopie aérienne ou satellitaire à faible angle), tous les résultats sont utilisables en stéréoscopie quelconque, comme montré dans beaucoup d'expériences.

Abstract

This thesis is a contribution to stereovision written in the framework of the MISS (Mathematics for Stereoscopic Space Imaging) project launched by CNES in cooperation with several university laboratories in 2007. This project has the ambitious goal to model a stereo satellite, using two almost simultaneous views of the Earth with small baseline in urban areas. Its main goal is to get an automatic chain of urban reconstruction at high resolution from such pairs of views.

The project faces fundamental problems that this thesis aims at solving. The first problem is the rejection of matches that could occur just by chance, particularly in shadows or occlusions, and the rejection of moving objects (vehicles, pedestrians, etc.). This thesis proposes a method for rejecting false matches based on the *a contrario* methodology. The mathematical consistency of this rejection method will be shown and it will be validated on exact simulated pairs, on ground truths provided by CNES, and pairs of classical benchmark (Middlebury). The reliable accepted matches reach a 40% to 90% density in the tested pairs.

The second issue is the accuracy. Indeed, the type of considered stereoscopy requires a very low baseline between the two views, which are visually almost identical. To get a proper relief, an extremely accurate shift must be estimated, and the noise level that allows this accuracy must be calibrated. In this thesis a subpixel disparity estimation method is proposed, which will be proved optimal by experimental and mathematical arguments. These results extend and improve the results obtained by the CNES method MARC. In particular, it will be shown on the Middlebury benchmark that the theoretical accuracy allowed by the noise exactly corresponds to the accuracy obtained on the reliable matches.

Although these results are obtained within the framework of a specific acquisition system (low baseline stereoscopy on aerial or satellite images), all results are used in a general stereo framework, as shown in many experiments.

Als meus pares,
A Sébastien.

Remerciements

En premier lieu, je tiens à exprimer toute ma gratitude à Andrés Almansa et Jean-Michel Morel pour m'avoir dirigée pendant ces trois années. Ils ont été le binôme parfait. Je n'aurais pas pu mener ce travail à bien sans leur exigence, leur disponibilité et leur patience. Leurs grandes qualités scientifiques et humaines m'ont beaucoup apporté et, chacun à sa façon, a su me motiver et m'orienter tout en me donnant une grande liberté. Pour tout cela, je leur en suis très reconnaissante.

J'aimerais remercier chaleureusement mon jury. D'abord Antonin Chambolle, Renaud Keriven et Lenny Rudin qui ont accepté de rapporter ma thèse. Je voudrais aussi remercier sincèrement Gwendoline Blanchet, Tomeu Coll, Jean-Marc Delvit, Yann Gousseau, Yves Meyer, Lionel Moisan et Bernard Rougé pour avoir accepté de participer au jury.

Tout au long de ma thèse, j'ai eu la chance de croiser un bon nombre de chercheurs et cette thèse doit beaucoup à toutes ces rencontres. En particulier, je tiens à remercier tous les membres du Projet MISS. Depuis le début, ils ont suivi mes travaux grâce à d'innombrables présentations. Les discussions et retours avec tous les membres ont été très enrichissants. J'espère qu'ils seront encore courageux pour venir à la soutenance m'écouter une fois de plus, cette fois-ci au moins il y aura un pot qui suivra (ils l'auront bien mérité !)

Je remercie Bernard Rougé pour avoir lu aussi attentivement le chapitre sur la précision subpixelienne et pour toutes les conversations sur la corrélation qui m'ont permis d'avoir un autre regard sur mon sujet de thèse. Gwendoline Blanchet a été d'une grande aide lors des test avec MARC. De plus, elle a été une excellente organisatrice d'activités chaque fois que nous sommes allés à Toulouse.

Je voudrais remercier très chalereusement Tomeu Coll. C'est grâce à lui que je me suis orientée vers le traitement d'images. Il a toujours été de très bon conseil et il m'a toujours encouragé tout au long de la thèse. Je profite aussi de ses lignes pour le remercier pour les nombreuses fois où il a fait le "transporteur" de paquets depuis Majorque. Je remercie également un autre compatriote de "sa roqueta", Toni Buades. Il s'est toujours beaucoup intéressé à ce que j'ai fait et ses critiques constructives ont été précieuses. Il m'a donné un sacré coup de main avec le code et mes plus ou moins douloureux débuts avec Megawave. Tomeu et lui m'ont très bien reçu chaque fois que je suis allée squatter à la UIB à Majorque.

Mes remerciements vont aussi à Vicent Caselles pour son soutien et ses encouragements. Il m'a très bien accueillie lors de mon séjour à la Pompeu Fabra à Barcelone où j'ai eu l'occasion de lui montrer mon travail. Je regrette son absence le jour de ma soutenance. Merci à Pascal Monasse pour son intérêt et ses commentaires sur mon travail, ainsi qu'à Lionel Moisan, qui de plus m'a permis de rencontrer d'autres chercheurs dans un cadre très sympathique lors de ses invitations au CIRM.

I would like to thanks Freddy Bruckstein, Guillermo Sapiro and Lenny Rudin for the fruitful conversations and the time they spend with me when they visited the CMLA.

J'aimerais particulièrement remercier Rafael Grompone pour ses lectures attentives de cette thèse. Il a toujours été là quand j'en ai eu besoin. Jérémy Jakubowicz a joué le rôle de grand frère de thèse quand je venais de commencer. Également Gabriele Facciolo, pour avoir testé mon code et tous ses commentaires sur mon travail. Gabriele, eres un pro-fesional! Je n'oublie pas (Saint) Nicolas Limare qui a fait des miracles même depuis le Japon. Je le remercie pour sa patience et sa pédagogie pour m'apprendre un peu d'informatique. Zhongwei Tang qui a fait toutes les rectifications d'images même quand "c'était pour hier" mais aussi Eric Bughin et Julie Digne pour les discussions qui m'ont fait penser en 3D et non en 2D et demie !

Je voudrais sincèrement remercier Jean-François Aujol qui, avec son éternelle bonne humeur, a toujours été de très bon conseil. Merci à Jerôme Darbon pour tous ses bons conseils lors de la préparation des candidatures, à Julie Delon pour m'avoir passé nombre de ses codes et à Jean-Denis Durou pour avoir relancé la collaboration avec l'IPGP et Antoine Lucas.

Je voudrais remercier aussi A. Desolneux et F. Richard qui ont été d'excellents encadrants de stage de M2.

Merci à Nick the Bear car, malgré ses méthodes un peu "esclavagistes", je crois avoir appris un peu d'anglais. Au moins il aura bien réussi à me changer les idées une fois par semaine !

Je ne serais sans aucun doute jamais arrivée jusqu'ici sans la bienveillance de quelques uns de mes enseignants. M'agradaria donar les gràcies a n'Antònia Rigo per les seves classes de matemàtiques tan excepcionals a Joan Alcover que varen fer que volgués seguir aquest camí. Igualment m'agradaria donar les gràcies a n'en Juan Carlos Naranjo del Val i en José Ignacio Burgos Gil per haver-me iniciat amb la geometria projectiva aplicada a la visió per ordinador a més a més d'haver-me ajudat a venir a Paris.

Les conditions de travail au CMLA ont été excellentes et j'en remercie tous les membres. Un grand merci au secrétariat avec Carine Saint-Prix, Micheline Brunetti, Sandra Doucet, Virgine Pauchont et Veronique Almadovar pour leur efficacité qui a fait que les tâches bureaucratiques soient un jeu d'enfant mais surtout pour leur gentillesse et leur bonne humeur. Je promets de faire un TP cuisine avec elles maintenant que la thèse est finie. Je leur dois bien ça ! Un autre grand merci à Christophe Labourdette pour son aide et ses bons conseils mais surtout pour être toujours là quand "au secours, mon ordi ne marche plus". Merci aussi à Pascal Bringas pour sa bonne volonté et pour être le meilleur testeur de wifi du labo au moment du gâteau des thésards.

Une mention très spéciale aux doctorants qui m'ont accompagné tout au long de ces années et qui ont fait que mon quotidien au CMLA ait été aussi agréable. Je pense à nos cafés au pavillon des Jardins où on a arrangé le monde, à nos goûters gourmands, à nos repas de doctorants, à nos soirées sportives à la piscine (les nageurs se reconnaîtront), au dancing-floor de la Colle sur Loup, aux parties de loup-garous, times-up et poker. Mais merci surtout pour votre aide précieuse lors de la préparation de ma soutenance et pour tous les encouragements que j'ai eu pendant cette très pénible dernière ligne droite. Votre soutien a fait que ça s'est passé avec un peu plus de douceur. Merci à vous ! Je pense aussi aux Devillettes' thesards qui m'ont accompagné lors de mes repas-picard les samedis au labo. J'espère n'oublier personne : Adina, Aude, Ayman, Benjamin, Bruno, Eric, F-X, Frédéric, Frédérique, Gaël, Jérémie, Jean-Pascal, Julie, Nicolas (et bis), Rafa, Stanley, Yen, Yohann, Zhongwei et aussi à ceux qui ont été de passage par le CMLA : Alex, Barbara, Gabriele, Hugo, Mauricio, Magalie, Mariano, Mariella, Rodrigo, Yaxin et Yifei.

Je n'oublie pas les gens de Paris 5, où j'ai fait mon stage de DEA et mon monitorat, qui m'ont très bien accueilli au 4ème étage : Arno, Baptiste, Benjamin, Cécile, Emeline, Nathalie, Mahendra, Mohamed, Sylvain et Sandra.

Lors de mon arrivée en France, j'ai eu la chance de partager une expérience inoubliable avec tous les habitants de Victor Lyon. Parmi eux, je tiens particulièrement à remercier Hernando et Jean-Philippe qui m'ont énormément aidé quand la belle vie fût finie. Mais je n'oublie pas toute la "troupe" avec qui j'ai vécu tous ses moments : Carmencita, Christian (merci pour nous avoir mis dans tes bagages quand tu es rentré en Allemagne), Dino (je n'arriverai jamais à t'appeler Leo), Khemila, Luca, Marie, Marco et Marco Paperino (ça me fait trop plaisir de te revoir en France), Michalis (je n'aurais pas pu avoir un meilleur guide

en Grèce), Raul, Sauro, Stef et Tom (merci pour m'avoir accueillie si bien en Belgique).

Je n'oublie pas mes amis qui, malgré la distance, m'ont toujours soutenue : Ferran, Francesc, Joaquin et Vanesa. Gràcies per rebre'm sempre a ca vostra amb la porta oberta. Gràcies pels moments que hem passat junts. Mai olvidaré ni Praga ni Guatemala ni tampoc aquella primavera mítica. A veure quan aconseguiré que vengueu a Mallorca. Ho tenim pendent! També vull aprofitar per donar les gràcies a na M. Àngels per haver-me ensenyat Leeds però sobretot per haver compartit amb jo tota l'aventura a Barcelona des de l'institut. No m'oblid de na Mònica amb qui he viscut un munt de peripècies des de que anàvem als escoltes passant pel pis d'Hospital Sant Pau.

Merci aux copains : Gigi, Titi, Cécé, Fifi, Toto et Magalie. Ils m'ont souvent insufflé des bouffées d'énergie notamment lors de nos excursions normandes. Toujours prêts à faire les cobayes pour mes tests culinaires ou pour m'aider à faire des lettres de motivation en français.

Florence et Aurélien ont eu le don de me ressourcer à chaque passage par Grenoble. C'est sûr qu'un peu de ski et un tour de moto dans les Alpes, ça vous remonte une thésarde. J'attends avec impatience l'heureux événement.

Merci enfin à ma famille, à Marieneiges pour son aide lors de mon départ et pour les rencontres à Paris et à New York, a les meves padrines per haver-me ensenyat tantes coses que no s'aprenen als llibres, a n'en Gaspar i na Lara que tant enyor tots els dies i que tant contenta me fan cada vegada que venen a veure'm a França, però sobretot gràcies als meus pares per haver-me recolzat incondicionalment des del principi encara que la meva absència no ha estat sempre fàcil. Sense ells no hagués arribat mai on he arribat. Simplement gràcies per tot!

Enfin, merci à toi Sébastien, merci pour avoir fait ce bout de chemin à mon côté. Merci pour les corrections de français à pas d'heure. Merci de m'avoir soutenue et épaulée toutes ces années même si je n'ai pas été d'une humeur facile tout le temps. Du fond du coeur, merci.

Contents

Notations

- Reference image

$$\begin{array}{rccc} u_1: & I & \longrightarrow & \mathbb{R} \\ & \mathbf{q}=(q_1,q_2) & \longmapsto & u_1(\mathbf{q}) \end{array}$$

- Secondary image

$$\begin{array}{rccc} u_2: & I' & \longrightarrow & \mathbb{R} \\ & \mathbf{q}=(q_1,q_2) & \longmapsto & u_2(\mathbf{q}) \end{array}$$

- Estimated disparity map

$$\begin{array}{rccc} \mu: & I & \longrightarrow & \mathbb{R} \\ & \mathbf{q} & \longmapsto & \mu(\mathbf{q}) \end{array}$$

- Real disparity map (ground truth)

$$\begin{array}{rccc} \varepsilon: & I & \longrightarrow & \mathbb{R} \\ & \mathbf{q} & \longmapsto & \varepsilon(\mathbf{q}) \end{array}$$

- $\mu(\mathbf{q}) = \emptyset$ means that no match has been accepted for $\mathbf{q}$. These points are colored in red in our resulting disparity maps.
- Depending on the context, the pair of images is called I and I' or u_1 and u_2.
- It is assumed that u_1 and u_2 are gray level images if the contrary is not specified.

$B_{\mathbf{q}}$	squared patch centered at $\mathbf{q}$.
u_x	derivative of u in the x axis (the epipolar direction).
$\mathbb{P}$	Probability.
$\mathbb{E}$	Expectation.
Var	Variance.
Cov	Covariance.
$\#A$	cardinal of A.

Introduction

Contexte de la thèse : le projet MISS

Le projet MISS (Mathématiques de l'imagerie stéréoscopique spatiale) est un projet de collaboration entre plusieurs universités et institutions lancé en 2007. Il rassemble le centre national d'études spatiales (CNES), le centre de mathématiques et de leurs applications (CMLA - ENS de Cachan), l'université Paris Descartes (Paris V), l'universitat de les illes Balears (UIB), l'universitat Pompeu Fabra (UPF), l'école d'ingénieurs Telecom Paris Tech.

Le principal objectif de MISS est la restitution automatique de modèles numériques d'élevation (MNE) à partir de deux images d'une scène sous faible différence angulaire, particulièrement en milieu urbain. Ce projet demande la conception d'une chaîne de traitement complètement maîtrisée et fiable qui commence avec l'acquisition des deux images et termine avec la restitution 3D de la scène étudiée. Le projet MISS s'intéresse également à l'étude de problèmes fortement liés au calcul du MNE tels que l'échantillonnage irrégulier, la restauration (bruit et flou), et la compression d'images.

La mise en correspondance point à point d'une paire images stéréoscopiques est un chaînon essentiel de cette chaîne. Il s'agit d'un problème ardu et pas toujours bien posé, en particulier dans les zones urbaines. La présence de surfaces cachées (occlusions), les objets en mouvement ou les surfaces qui renvoient la lumière différemment selon l'angle de vue rendent la tâche de mise en correspondance difficile. Depuis plus d'une dizaine d'années, le groupe du CNES autour de B. Rougé a étudié la viabilité de la faible différence angulaire stéréoscopique pour l'intégrer dans les futurs satellites. C'est ce qu'on appelle des images à B/H faible (où B est la distance entre les deux prises de vue et H est la hauteur du satellite). Ce modèle d'acquisition réduit considérablement une grande partie des difficultés rencontrées quand le ratio B sur H est trop important [Delon and Rougé, 2007].

En 2010, Pléiades, satellite d'observation de la Terre à très haute résolution (THR) sera lancé et fournira des paires d'images stéréoscopiques à (relativement) faible B/H et presque simultanées.

Contributions de la thèse

La stéréovision binoculaire, qui consiste à retrouver la profondeur d'une scène à partir de deux vues, est depuis son origine un des problèmes centraux de la vision par ordinateur. Dans le domaine de l'imagerie satellitaire et aérienne, ce problème fait depuis trente ans l'objet de recherches très actives.

Il y a deux types d'approches en vision stéréoscopique. Les méthodes locales réalisent

la mise en correspondance en comparant les images point à point. Les plus connues sont les méthodes de *block-matching* qui comparent des blocs ou fenêtres autour de chaque point d'une image aux blocs de l'autre image. Les méthodes globales utilisent simultanément tous les pixels de l'image pour minimiser une énergie composée d'un terme d'attache aux données, d'un terme de régularité et d'un terme régissant la possibilité d'occlusion. Ces dernières années, les méthodes globales sont devenues très populaires et elles réalisent les meilleures performances dans les *benchmarks*.

Les méthodes *locales* sont plus simples mais elles ont trois inconvénients : elles peuvent commettre des erreurs liées à la présence de motifs répétés dans les images, elles sont sensibles au bruit, et elles souffrent de ce que l'on appelle effet d'adhérence, qui cause des erreurs près des bords de relief abrupts.

Les méthodes *globales* n'ont pas d'effet d'adhérence. Elles prennent souvent la bonne décision en présence de motifs répétitifs et sont moins sensibles au bruit. De plus, elles proposent un appariement de tous les pixels, alors que les méthodes locales font un appariement non dense, qu'il faut donc compléter: et par quoi, sinon par une méthode globale ?

On devrait déduire de la précédente comparaison que les méthodes globales sont en définitive les seules à considérer.

Le but de ce mémoire est de montrer qu'il n'en est pas ainsi. Les méthodes locales ont deux avantages qui leur sont propres, et que nous allons tenter de pousser à leurs ultimes conséquences. Le premier est que l'appariement local peut mener à des règles statistiques de décision nous disant si un appariement est fiable ou non. Le second est que la précision d'un appariement local peut être complètement caractérisée en fonction du bruit.

Nous allons donc traiter trois questions fondamentales posés par la stéréo locale :

1. le contrôle des fausses alarmes,
2. le problème d'adhérence,
3. la précision subpixélienne.

Le contrôle des fausses alarmes

Cette thèse introduit un modèle stochastique de mise en correspondance *a contrario* de blocs pour le calcul de disparités (décalages). Ce modèle *a contrario* (AC) repose sur la théorie de la Gestalt et compare des fenêtres des deux images pour accepter ou rejeter un possible appariement. Définissant la notion de *correspondance significative* entre points des deux images de la paire stéréo, ce modèle garantit qu'en moyenne pas plus d'un mauvais appariement dû au bruit de fond ne peut se produire sur toute l'image. Toutefois le modèle *a contrario* n'élimine pas les faux matches dus à des formes répétées. Pour les éliminer, un seuil d'autosimilarité (SS) sera aussi implémenté, qui ne retient une correspondance significative que si elle est meilleure que tout autre appariement entre le point de référence et un point dans son voisinage. La combinaison des deux seuils (AC+SS) permet de contrôler toutes les fausses alarmes possibles en stéréoscopie locale. En bref:

- Le modèle *a contrario* (AC) permet de détecter les occlusions et de contrôles les faux matches dans le bruit. De plus, la méthode proposée est capable d'éliminer de façon fiable tout mouvement incohérent d'objets de la scène, ce qui s'avère indispensable pour les couples d'images aériennes ou satellitaires non simultanées. En effet, elles contiennent souvent des véhicules et des piétons qui ont changé de position d'un cliché à l'autre.

- Les zones très peu texturées, comme les zones d'ombre ou les régions saturées, sont complètement maîtrisées par une analyse en composantes principales (ACP) locale intégrée dans le modèle (AC). En effet, cette approche permet de mettre en correspondance avec fiabilité des points dans des régions de l'image inattendues, généralement des ombres, tout en rejetant tous les pixels qui n'ont aucune information utile pour la mise en correspondance.

- Les effets stroboscopiques dûs aux structures répétitives dans les images sont par contre évités par un seuil très simple d'auto-similarité (SS).

Implicitement, l'algorithme qui fait la mise en correspondance significative est adapté à chaque point et réagit différemment selon que le point se trouve dans une zone texturée ou dans une zone dite "plate" comme par exemple une ombre. Cette nouvelle approche permet de mettre en correspondance un nombre important de points de l'image avec un minimum de nombre de fausses alarmes et **sans problème de réglage de paramètres.** L'approche est applicable à n'importe quel couple stéréo, que le B/H soit faible ou fort.

Pour donner un exemple de l'amélioration observée par rapport à un algorithme de stéréo classique, la figure 1 compare les différentes cartes de disparités obtenues avec notre algorithme et celui du CNES (MARC) *Multiresolution Algorithm for Refined Correlation* pour un couple d'images aériennes de Marseille non simultanées.

La solution du problème d'adhérence

La méthode de mise en correspondance stéréo (AC + SS) souffre quand même du phénomène d'adhérence, comme toutes les autres méthodes de *block-matching*. Les erreurs d'adhérence les plus choquantes se produisent près des contours de l'image coïncidant avec une discontinuité du relief. Ces erreurs se manifestent par une dilatation des objets de la scène sur les cartes de disparité.

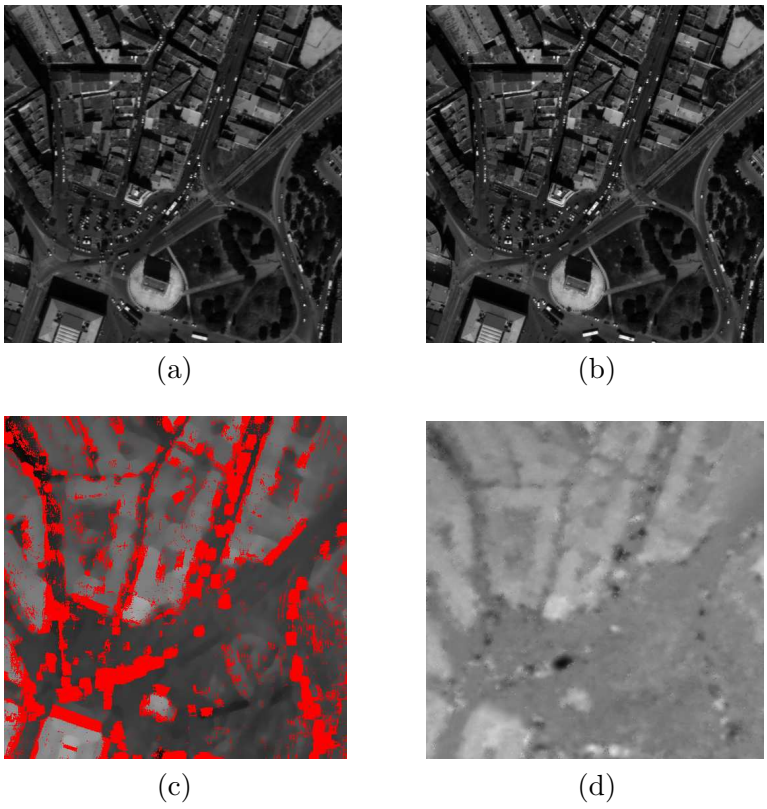
Il ne s'agit pourtant pas de faux matches: les erreurs sont causées par le fait que la mise en correspondance (correcte et significative) d'un bloc est attribuée à son centre, alors quand en fait cette mise en correspondance est causée par d'autres points du bloc que le centre, parfois situés sur la périphérie du bloc. Le phénomène d'adhérence est donc une sorte de myopie causée par la comparaison de blocs. Sa portée est égale à la taille de la fenêtre. Aussi, de nombreux auteurs ont-ils essayé de réduire le phénomène d'adhérence en modulant la forme des fenêtres. Mais cela n'élimine pas réllement le phénomène.

Dans cette étude, une solution nouvelle est proposée à ce problème classique. La détection des pixels risquant l'erreur d'adhérence sera basée sur l'appariement fin du gradient de l'image à l'intérieur de chaque bloc, de manière à attribuer la disparité aux points qui la causent.

La méthode de mise en correspondances fiables (AC + SS) complémentée par la correction d'adhérence sera testée sur des exemples de *benchmarks* classiques et sur des scènes urbaines avec faible B/H. Tous les tests confirment que l'algorithme en trois étapes proposé fournit des **nappes de disparité assez denses (40% - 90%) et contenant moins de 0,4% de faux appariements**.

Le raffinement subpixélien

La stéréo vision a eu tendance pendant longtemps à considérer des couples d'images avec un fort B/H afin d'obtenir une bonne précision sur la reconstruction 3D. Cependant, comme on



(a) (b)

(c) (d)

Figure 1: Marseille : (a) Image de référence. (b) Image secondaire. (c) Carte de disparités non dense obtenue avec notre algorithme. Les points rouges sont des points rejetés par (AC+SS). (d) Carte de disparités dense obtenue avec MARC (Multiresolution Algorithm for Refined Correlation) sans régularisation. Des nombreuses erreurs se produisent à cause des véhicules en mouvement. Nos résultats sont moins denses car toute fausse correspondance possible a été rejetée.

l'a mentionné précédemment, avec un fort B/H les zones d'occlusion augmentent, la forme et la couleur apparente des objets change beaucoup plus, et la mise en correspondance devient donc plus hasardeuse. Dans le cas des images satellitaires, quand les images sont prises par deux passages du même satellite beaucoup d'objets, et même les ombres, ont bougé.

Ceci explique pourquoi [Delon and Rougé, 2007] ont considéré des images avec un B/H faible de l'ordre de 0.1 au lieu des rapports classiquement utilisés de l'ordre de 1. Ce modèle réduit les difficultés rencontrées avec un fort B/H et le modèle de déformation

$$u_1(\mathbf{x}) = u_2(x + \varepsilon(\mathbf{x}), y)\,,$$

entre le couple stéréoscopique u_1 et u_2 (ε est la fonction de disparité) est bien plus juste. Par contre, pour obtenir la même précision en hauteur, une plus grande précision sur le calcul des disparités s'impose, d'où le besoin de calculer des disparités subpixéliennes très précises. Ces problèmes de précision ont été défrichés par Camlong, Delon, Giros et Rougé notamment, aboutissant au logiciel MARC de reconstruction du relief à partir d'une paire stéréo à faible B/H.

Comment réaliser une précision subpixélienne ? [Szeliski and Scharstein, 2004], et plus

récemment [Delon and Rougé, 2007], ont démontré qu'il est nécessaire de pratiquer un zoom préalable d'ordre 2 pour que la corrélation soit bien échantillonnée. D'autre part, il a été remarqué que l'interpolation de la corrélation doit être faite par *zero-padding* (interpolation exacte) si des erreurs inférieures à quelques centièmes de pixel sont envisagées. Notre approche consiste à raffiner l'appariement accepté par (AC+SS) avec la corrélation tout en respectant l'interpolation et l'échantillonnage de Shannon. Un tel raffinement de la disparité n'est pas nouveau, puisqu'il était explicité dans MARC.

Mais, d'abord, le présent travail fournit une formule mathématique exacte de l'erreur de disparité due au bruit et de son écart type. Ensuite cette estimation exacte est confirmée par la nouvelle implémentation sans fausses alarmes, dans laquelle les erreurs résiduelles ne sont dues qu'au bruit.

En nous appuyant sur plusieurs exemples dont la base de donnée de *benchmark* publique Middlebury, nous avons pu démontrer que, dans un cadre complètement réaliste, **de 40% à 90% des pixels d'une image pouvaient être mis en correspondance avec une précision de l'ordre de $\frac{5}{100}$ de pixel**.

De plus, la prédiction théorique des erreurs dues au bruit est à peine inférieure à l'erreur atteinte en pratique par notre algorithme sur des couples d'images simulées et réelles. Ce fait démontre que la méthode est optimale en précision et démontre aussi *a posteriori* que toutes les causes d'erreur ont été éliminées.

La vérité terrain obtenue par validation croisée

Une des principales difficultés rencontrées dans ce travail a été la validation des résultats, et en particulier, la validation de la précision subpixélienne due à l'inexistence d'une vérité terrain bien précise. La validation réalisée par notre algorithme est en fait également une méthode de constitution de vérité terrain par validation croisée. En effet, la base de données stéréo publique Middlebury, qui a été établie par des chercheurs en stéréovision, contient 9 images de la même scène prises par une caméra sur un rail sur une ligne droite et à des intervalles constants. En utilisant l'image centrale comme image de référence, une comparaison des différentes cartes de disparités a été rendue possible.

Il est clair que le projet MISS aurait besoin de définir sa vérité terrain dans le cadre rigoureux permis par un B/H multiple et (presque) simultanné. La vérité terrain serait alors donnée par l'algorithme de disparité lui-même, puisque la validation croisée permettrait d'avoir des points justes à 100% et que de plus la vérité terrain serait alors associée à une précision. C'est ce que nous avons pu faire avec des données simulées, mais aussi avec la base Middlebury.

Une des preuves de l'efficacité de la méthode a été la mise évidence d'une erreur sur la vérité terrain officielle de Middlebury, que nous avons pu rectifier.

Densité de la carte de disparités

Dans toutes ces situations, l'algorithme s'abstient de mettre en correspondance deux points s'il y a une possibilité probable de faux match. La carte de disparités finale obtenue, ne pouvant plus être complètement dense, nécessite une interpolation postérieure pour arriver à un modèle numérique d'élévation complet.

Aucune interpolation fiable ne sera jamais possible dans des zones où presque aucun match n'a été trouvé : dans ces zones, on fera seulement des reconstructions plausibles et

non prouvées. Pour garantir la quantité de correspondances fiables dans les zones d'ombre, il faudrait s'arranger pour avoir des images d'encore meilleure qualité. C'est le nombre de bits/pixels que l'on pourra obtenir qui décidera de l'information accessible dans les zones d'ombres. Une des questions qui se pose donc pour un instrument stéréoscopique est de savoir si une dynamique supérieure peut être atteinte, notamment dans les ombres, et éventuellement en reconsidérant la méthode d'acquisition : élargissement du TDI, augmentation du gain. Cette question semble cruciale pour l'avenir de la stéréoscopie spatiale en milieu urbain. Il est possible que l'instrument d'observation de la Terre puisse bénéficier d'une dernière remise en question, pour être pleinement adapté à une fonction stéréoscopique.

Une autre solution naturelle à ce problème serait l'utilisation de plusieurs images, avec éventuellement des rapports B/H différents. Ceci permettrait de trouver des MNE denses et à très haute précision à condition que chaque point soit visible sur plusieurs images. Ceci est possible car, pour des satellites en orbite héliosynchrone et phasée (comme Spot ou Pléiades), une zone de la Terre est visible plusieurs fois par cycle. Par exemple une région de la Terre à latitude 45° est visible 157 fois par an. De la même façon, des images obtenues avec différents satellites pourraient être utilisées. Dans tous les cas, il est clair que le modèle multi-images permettrait d'obtenir des résultats plus denses à haute résolution.

Organisation du rapport

Le chapitre 1 présente la problématique de la stéréoscopie binoculaire en s'appuyant sur une étude bibliographique. On s'intéresse particulièrement à la stéréoscopie en imagerie aérienne ou satellitaire en zones urbaines. Une méthode d'appariement de points fiables dans les images stéréoscopiques est présentée dans le chapitre 2. En particulier, le modèle *a contrario* pour les blocs d'images est étudié. Dans le chapitre 3, on propose une correction de l'effet d'adhérence consistant en la suppression de la nappe de disparités des pixels à risque. Ce chapitre compare également les résultats avec d'autres méthodes de stéréoscopie non dense. Le chapitre 4 aborde la précision subpixélienne extrême de la corrélation et propose la validation croisée comme méthode de validation. Le chapitre 5 présente les détails de notre algorithme. Le chapitre 6 montre plusieurs résultats obtenus avec la méthode présentée dans cette thèse avec des images de nature très différente. Enfin, le chapitre 7 conclut cette thèse et donne quelques perspectives pour la suite.

Introduction (in English)

Context of the thesis: the MISS Project

The MISS Project (Mathematics for space satellite imaging) is a collaborative project launched in 2007 between several universities and institutions, namely the French space agency (CNES), the Center for Mathematical Studies and its Applications (CMLA - ENS de Cachan), the Paris Descartes University (Paris V), the Illes Balears University (UIB), the Pompeu Fabra University (UPF) and the School of Engineering Telecom Paris Tech.

The main goal of MISS is the automatic reconstruction of Digital Elevation Models (DEM) from two images of a scene under a low angular difference, especially in urban areas. This project requires the design of a completely controlled and reliable processing chain, starting with the acquisition of two images and finishing with the reconstruction of the corresponding 3D scene. The MISS project is also interested in issues strongly related to the DEM computation as irregular sampling, restoration (noise and blurring), and image compression.

Stereoscopic image matching is an essential link of this chain. This is a difficult ill-posed problem, particularly in urban areas. The presence of hidden surfaces (occlusions), moving objects or surfaces that reflect the light differently depending on the viewing angle makes the matching task difficult. For over a decade, the group at CNES around B. Rougé studied the viability of the small angular difference concept in stereoscopy, with the aim of integrating it in future remote sensing systems. This concept is also known as "low B/H stereo (where B is the distance between the two shots and H is the height of the satellite). This acquisition model reduces significantly much of the difficulties encountered when the B/H ratio is too large.

In 2010, Pléiades, an Earth observation satellite at very high resolution (VHR) will be launched and will provide stereo pairs of images with (relatively) low B/H ratio and almost simultaneous views.

Contributions of the thesis

Binocular stereovision, which tries to find the depth of a scene from two views, is since its origin one of the central problems of computer vision. In the field of satellite and aerial imagery, this has been an active research subject throughout the last thirty years.

There are two types of approaches in stereovision. Local methods look for image matches by comparing the images pointwise. The best-known local methods are the block-matching methods comparing blocks or windows around each point of an image to blocks of another image. Global methods use simultaneously all pixels of the image to minimize an energy term consisting of a data term, a regularizing term and a term governing the possibility of occlusion. In recent years, global methods have become very popular and usually obtain the best scores in common benchmarks.

Local methods are simpler but they have three disadvantages: they may make errors in the presence of repeated patterns in the images, they are sensitive to noise, and they suffer from the so-called fattening phenomenon, which causes errors near the edges of discontinuous terrains.

Global methods are not affected by fattening errors. They often take the right decision in the presence of repetitive patterns and are less sensitive to noise. In addition, they offer a matching of all pixels whereas local methods only provide semi-dense mappings which must be completed. How should this be accomplished if not by a global method?

We should infer from the previous discussion that global methods are the only ones to be considered.

The purpose of this work is to show that this is not the case. Local methods have two advantages of their own, and we will try to push them forward to their ultimate limits. First, local matching can lead to statistical decision rules telling us if a match is reliable or not. Second, the precision of each local match can be accurately predicted by a formula which depends on the noise level and on local image characteristics.

Therefore we will address to three fundamental questions posed by local stereo:

1. control of false alarms,

2. the fattening problem,

3. subpixel accuracy.

This thesis introduces an *a contrario* stochastic matching model for image blocks which is useful for the computation of disparities (shifts). This *a contrario* (AC) model is based on the Gestalt theory and compares patches of both images in order to accept or reject a possible match. Defining the concept of *meaningful match* between points of the two images of the stereo pair, this model ensures in theory that not more than a single mismatch due to background noise can occur on average. However, the *a contrario* model does not eliminate false matches due to repeated structures. In order to eliminate them, a self-similarity (SS) threshold will also be implemented, which retains a meaningful match if it is a better match than any other match between the reference point and a point in its neighborhood. The combination of both thresholds (AC + SS) can control all the possible false alarms in local stereoscopy. In brief:

- The *a contrario* (AC) model is used to detect occlusions and control false matches in the presence of noise and other distortions. Moreover, the proposed method is able to reliably eliminate all inconsistent movements of objects in the scene, which is essential for non-simultaneous couples of aerial or satellite images. Indeed, they often contain vehicles and pedestrians who have changed their position from a snapshot to the other.

- The poor textured areas, like shadows or saturated areas, are completely controlled by a local principal component analysis (PCA) which is integrated in the (AC) model. Indeed, this approach makes it possible to reliably match points in unexpected regions of the image, usually shadows, while rejecting all pixels that do not have any useful matching information.

- Stroboscopic effects due to repetitive structures in the images are avoided by a simple self-similarity (SS) threshold.

Implicitly, the matching algorithm is automatically adapted to each point and reacts differently depending on whether it is located, in a textured area or in a “flat” area such as a shadow. This new approach allows one to match a large number of points in the image with

a minimum number of false alarms and **without any parameter tuning**. The approach is actually applicable to any stereo pair, with a low or large B/H ratios.

As an illustration of the improvement that the technique we propose may reach when compared to classical stereo algorithms, figure 1 compares disparity maps that were obtained with our algorithm and the CNES algorithm (MARC) Multiresolution Algorithm for Refined Correlation for a couple of not simultaneous aerial images of Marseille.

The solution to the fattening problem

The stereo matching method (AC + SS) suffers from the fattening phenomenon like other *block-matching* methods. The most shocking fattening errors occur near the edges of the image coinciding with a relief discontinuity. These errors appear in the disparity maps as a dilation of the foreground objects in the scene.

Yet, fattening errors are not false matches. They are just errors caused by the fact that the disparity of a block match (meaningful and correct) is assigned to its center, when in fact this match is caused by other parts of the block than the center, these parts being located on the periphery of the block. Thus, the fattening phenomenon is a kind of myopia caused by the block comparison. Its range is equal to the window size. Hence, many authors have tried to reduce it by modulating the shape of the windows. But this does not really eliminate the phenomenon.

In this study, a new solution is proposed to this classic problem. The detection of pixels risking fattening will be based on the accurate gradient matching of the image within each block, in such a manner that the disparity is assigned to the points that cause fattening.

The reliable matching method (AC + SS) complemented by the fattening correction will be tested on examples of classic benchmarks and urban scenes with low B/H. All tests confirm that the proposed three step algorithm provides **fairly dense disparity maps (40% - 90%) containing less than 0.4 % of false matches**.

Subpixel refinement

Stereovision has tended to consider pairs of images with a large B/H to get an accurate 3D reconstruction. However, as mentioned above, with large B/H the occlusion areas increase, the shape and the apparent color of objects change more, and the matching becomes more hazardous. In the case of satellite images, when images are taken by two sweeps of the same satellite (separated by one or more orbits), lots of objects, and even the shadows have moved.

To avoid this, [Delon and Rougé, 2007] have considered pictures with a low B/H of around 0.1 instead of the more conventional ratios of the order of 1. This model reduces the difficulties encountered with a large B/H and the distortion model

$$u_1(\mathbf{x}) = u_2(x + \varepsilon(\mathbf{x}), y)\,,$$

between the stereo pair u_1 and u_2 (ε is the disparity function) is quite right. The only drawback is that for obtaining the same height accuracy a higher precision in the calculation of disparities is needed. Whence the need to calculate very accurate subpixel disparities. These accuracy problems have been addressed for the first time systematically by Camlong, Delon, Giros and Rougé in the MARC software, a depth map reconstruction algorithm from a low B/H stereo pair.

How to achieve this subpixel accuracy? [Szeliski and Scharstein, 2004] and more recently [Delon and Rougé, 2007] had demonstrated the need of an initial zoom ($\times 2$) for the correct correlation sampling. On the other hand, it was noted that the interpolation of the correlation must be performed by zero-padding (exact interpolation). In that way errors below a few hundredths of pixel become reachable. Our approach has been to refine the accepted matches by (AC + SS) with the correlation, while respecting the Shannon interpolation and sampling. Such refinement of the disparity is not new, as was explained in MARC.

This work provides an exact mathematical formula to estimate the disparity error caused by noise. Then, this exact estimate is confirmed by a new implementation of block matching eliminating most false alarms, where the residual errors are therefore mainly due to the noise.

Based on several examples, including the public database of Middlebury, we have shown that in a completely realistic setting **40% to 90% of pixels of an image could be matched with an accuracy of about $\frac{5}{100}$ pixels**.

Moreover, the predicted theoretical error due to noise is nearly equal to the error achieved by our algorithm on simulated and real images pairs. This shows that the method is optimally accurate, and also shows *a posteriori* that all sources of error have been eliminated.

The ground truth obtained by cross-validation

One of the major difficulties encountered in this work was the validation of results due to the lack of a precise ground truth, and in particular, the validation of subpixel accuracy. Our proposed cross-validation method will be proved to also be a method for establishing a ground truth. The public stereo Middlebury data set contains 9 images of the same scene taken by a camera on a rail on a straight line at constant intervals. By using the central image as reference image, a cross-validation of the obtained disparity maps for different pairs has been done and has confirmed our theoretically predicted error. Thus, one of the proofs of the effectiveness of the proposed method is the refinement of the Middlebury ground truth.

The disparity map density

Our proposed algorithm fails to match two points if a false match is likely. The obtained final disparity map cannot be completely dense, requiring a subsequent error prone interpolation to get a complete digital elevation model. The question that arises for an satellite stereoscopic instrument is whether a higher dynamic range could be attained, especially in shadows. This question seems crucial for the future of space stereoscopy in urban areas.

Another natural solution to this problem is the use of multiple pairs, possibly with several B/H ratios. This would lead to denser DEM's. This is possible because, for satellites in a sun-synchronous and phased orbit (such as SPOT or Pléiades), an area of the Earth is visible several times per cycle. For example, an Earth region at a 45° latitude is visible 157 times per year. In all cases, it is clear that the multi-image model would produce denser results at high resolution. But this requires each stereo pair to keep only reliable points, which is exactly what this thesis has been about.

Plan

Chapter 1 presents the problem of binocular stereoscopy based on a literature review. In particular, we are interested in aerial or satellite stereoscopic imagery in urban areas. A method for matching points reliably in stereovision is presented in Chapter 2. In particular, the *a contrario* model for image blocks is studied. In Chapter 3, we propose a correction of the fattening effect. This Chapter also details our algorithm and compares the results with other methods of semi-dense stereo. Chapter 4 addresses to the optimal subpixel precision of the correlation and proposes cross-validation as a validation method. Chapter 6 shows several results obtained with the presented method in this thesis with images of very different nature. Finally Chapter 7 concludes this work and gives some perspectives.

Chapter 1

Généralités et état de l'art

Contents

Résumé : Dans ce chapitre, on présente la stéréoscopie sous forme générale et on soulève les difficultés et objectifs de notre étude. En particulier, une étude bibliographique des différentes techniques de mise en correspondance de couples d'images stéréoscopiques est faite pour présenter l'état de l'art.

Abstract: In this Chapter we refer to sterevision and we raise the main difficulties and goals of our study. In particular, we review the main stereo matching methods appearing in the literature and we study the state-of-the-art.

1.1 Introduction

La reconstruction 3D, qui vise à représenter une scène en trois dimensions, reçoit un intérêt particulier dû aux nombreuses applications. Grâce aux avancées technologiques, les recherches actuelles portent sur l'acquisition de modèles tridimensionnels de très haute qualité de plus en plus précis.

Essentiellement, il y a deux méthodes pour l'acquisition de données 3D. D'un côté, les méthodes actives acquièrent la profondeur d'une scène à partir d'une source de lumière contrôlée comme les faisceaux lasers. De l'autre, les méthodes passives calculent le relief à partir d'un jeu d'images de la scène. Cette deuxième méthode est étudiée par la communauté de vision par ordinateur. Certaines approches n'utilisent qu'une image comme le *shape-from-shading* [Prados, 2004] [Durou, 2007]; d'autres, comme la stéréophotométrie, exploitent plusieurs images prises sous le même angle et différentes illuminations [Durou and Courteille, 2007]. Il y a une autre méthode qui se trouve entre les méthodes actives et passives, le *structured-light 3D scanner* qui utilise à la fois une lumière (active) structurée et des images de la scène : des motifs lumineux (comme des rayures) sont projettés sur la scène au moment de l'acquisition d'images créant une texture supplémentaire sur la surface.

Dans cette thèse, on s'est intéressé à la stéréoscopie binoculaire qui utilise deux images d'une scène vue sous des angles légèrement différents, comme la vision humaine. Chaque oeil fournit en effet au cerveau deux vues de la scène avec un point de vue différent, ce qui contribue à une sensation de relief instantanée. Toutefois, la perception humaine est basée sur au moins cinq processus distincts en plus de la disparité binoculaire: perception à partir des ombres et dégradés, perspective atmosphérique, perspective géométrique, perspective par déformation de la texture, et perception en couches par l'analyse des occlusions et jonctions en T. La perception du relief par la disparité binoculaire seule fait l'objet de nombreux travaux de recherche depuis l'apparition de la vision par ordinateur dans les années 60 [Julesz, 1960] [Marr and Poggio, 1976] [Marr and Poggio, 1979].

La dernière décennie a vu une explosion de travaux de stéréo vision motivés par de très nombreuses applications. Par exemple, pour n'en citer que quelques uns : le IBMR (*Image-based modeling and rendering*) cherche une représentation 3D d'une scène afin de générer le rendu d'un nouveau point de vue, et la téléprésence en robotique s'intéresse à la reconstruction du relief avec des robots munis de deux caméras pour faciliter les travaux en environnements hostiles ou d'accès difficiles. Une autre application bien connue est la topographie, qui cherche à représenter le terrain à partir d'images aériennes ou satellitaires.

La construction de modèles numériques d'élévation (MNE) en zones urbaines est un point clé pour des applications comme le placement d'antennes de télécommunications, la télésurveillance ou le cadastre. En effet, le MNE décrit tous les objets présents dans une scène : les bâtiments, la végétation, mais aussi le mobilier urbain. Le cas de Google Earth

est un exemple représentatif de l'intérêt du grand public pour la construction de MNE. Pour l'acquisition de données 3D d'une zone urbaine, les méthodes actives et passives sont possibles. Les appareils LIDAR (*light detection and ranging*) sont très précis mais ils fournissent des modèles épars et ils sont très onéreux. Les méthodes qui utilisent des images d'interférométrie radar sont limitées au niveau de la précision, ce qui est notamment gênant en zones urbaines. Aussi la reconstruction à partir des images optiques est-elle souvent la meilleure option. Dans cette thèse, on s'est particulièrement intéressé au calcul automatique de MNE en zones urbaines, qui fait l'objet d'une collaboration suivie avec le CNES (Projet MISS).

1.2 De l'acquisition des images au 3D

Pour une reconstruction 3D complète d'une scène, plusieurs étapes sont nécessaires : l'acquisition des images, la calibration, la rectification, la mise en correspondance et la reconstruction. Les travaux présentés dans cette thèse portent sur l'étape de la mise en correspondance, mais il est important d'avoir une vision générale de la chaîne de traitement. Donc, sans entrer dans les détails, nous allons expliquer ces étapes.

Acquisition

Le modèle sténopé ou pin-hole (fig. 1.1) est habituellement le modèle plus simple considéré pour décrire la formation des images. Dans ce modèle, l'image se forme par projection sur le plan image

$$\begin{array}{rccc} P: & \mathbb{R}^3 & \longrightarrow & \mathbb{R}^2 \\ & (x,y,z) & \longmapsto & (f\dfrac{x}{z}, f\dfrac{y}{z}) \end{array}$$

où C, le centre de projection, est aussi l'origine du repère, f est la longueur focale et (x, y, z) le point de la scène.

Les angles et les distances dans l'image dépendent du relief du terrain et de l'angle d'inclinaison de l'axe de prise de vue. En imagerie aérienne, les images sont usuellement proches du nadir, c'est-à-dire que l'axe principal a la même direction que la normale du terrain.

Calibration et rectification

La calibration est la première étape avant et après l'acquisition des images dans la chaîne, et il est d'une grande importance. Elle consiste à trouver la géométrie interne et externe du système d'acquisition. Pour la calibration interne, il est nécessaire de trouver la focale, le centre optique de la caméra, les dimensions du pixel et l'angle d'obliquité du pixel (5 paramètres). Pour la calibration externe, il s'agit de retrouver les rotations et translations dans l'espace qui ramènent la position de la caméra dans un repère externe usuellement appelé le "repère monde" (6 paramètres). Le problème de calibration est souvent considéré comme résolu, mais des recherches actuelles portent encore sur le sujet car la calibration est une étape cruciale pour la reconstruction 3D. Dans ce domaine, citons les travaux fondateurs de [Hartley and Zisserman, 2000], [Faugeras and Luong, 2001] et de [Lavest et al., 1998]. Le site web hébergé à Caltech[1] contient une importante boite à outils de calibration.

[1] http://www.vision.caltech.edu/bouguetj/calib_doc/

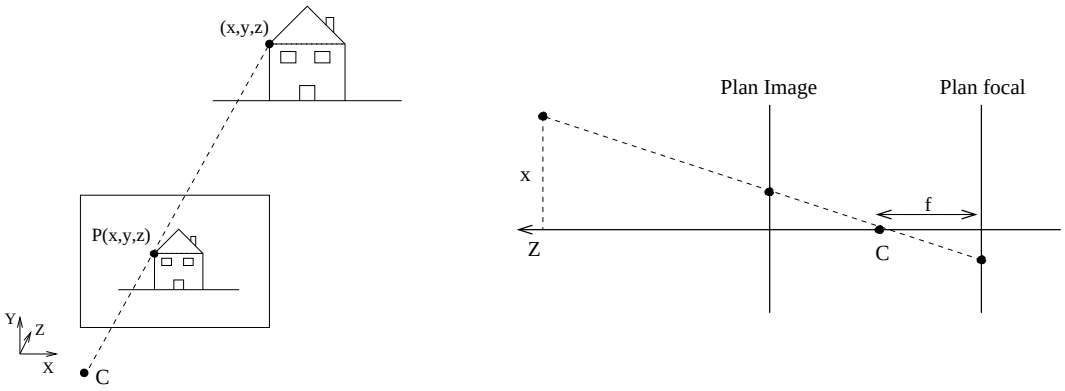


Figure 1.1: Le modèle *pin-hole*. Une caméra *pin-hole* est caractérisée par le choix d'un point C et d'un plan de projection dans $\mathbb{R}^3$. Un rayon de lumière passant par un point (x, y, z) de la scène est obligé de passer par C, appelé centre otique, qui est considéré comme infiniment petit. La formation de l'image est une projection perspective de tout point (x, y, z) sur le plan image. La projection de C sur le plan image est le point principal et la droite passant par ces deux points est la droite principale. C est situé à une distance f du plan focal.

Dans le modèle sténopé, on dit que deux points issus de chacune des images sont homologues s'ils sont la projection du même point de la scène sur les deux plans image. Le plan passant par les deux centres optiques et le point de la scène s'appelle plan épipolaire (fig. 1.2 - gauche). L'intersection du plan épipolaire avec les deux plans images donne les droites épipolaires conjuguées. Ainsi, par définition, les points homologues se trouvent sur ces deux droites. Les projections de chaque centre optique sur l'autre plan image forment deux points appelés épipoles qui se trouvent sur les droites épipolaires. Pour chaque nouveau point de la scène, il existe un nouveau plan épipolaire. L'ensemble des plans épipolaires crée deux faisceaux de droites épipolaires contenues dans chaque plan image et passant par les épipoles.

La rectification consiste à modifier les images originales pour les mettre en géométrie épipolaire (fig. 1.2 - droite). Les images sont ré-échantillonnées de manière à ce que les épipoles soient à l'infini et les droites épipolaires conjuguées soient parallèles et correspondent aux lignes des images. Citons [Loop and Zhang, 1999] qui rectifie les images en minimisant la distortion des images et [Zhang, 1998] qui fait une évaluation de toutes les techniques de rectification.

Il y a essentiellement deux façons de faire la rectification épipolaire : soit on connaît les paramètres de calibration, soit on ne les connaît pas. Dans le cas où les paramètres sont inconnus, il est toujours possible de faire une rectification en aveugle de l'image à partir de quelques correspondances de points homologues entre les images. Il s'agit de trouver la matrice 3×3, F de rang 2 qui satisfait

$$x^T F x' = 0\,,$$

où x et x' sont des vecteurs colonne contenant les coordonnées homogènes des points homologues. F est appelée matrice fondamentale. Quand on utilise cette matrice, la reconstruction 3D finale sera correcte modulo une transformation projective. Si la matrice K contenant les paramètres de calibration internes est connue, alors on connaît la matrice essentielle E qui

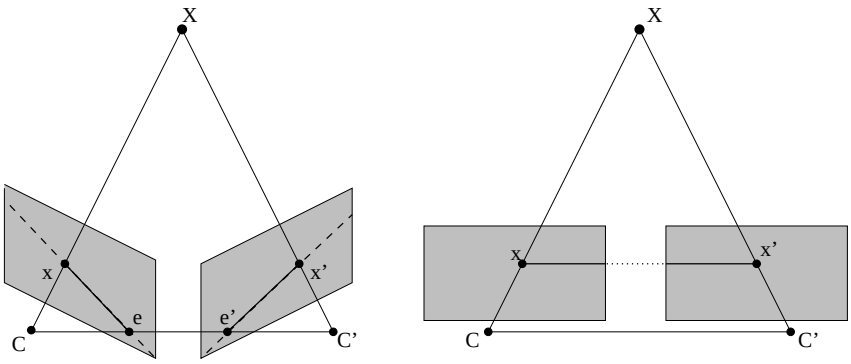


Figure 1.2: A gauche : deux caméras avec centres optiques C et C'. Le segment entre C et C' est la *baseline*. Les points $\mathbf{x}$ et $\mathbf{x}'$ sont les projections du point X sur les deux plans images : ce sont les points homologues. Les points e et e' sont les points d'intersection des plans images avec la *baseline* : ce sont les épipoles. Pour chaque point X de la scène, les droites $(e\,\mathbf{x})$ et $(e'\,\mathbf{x}')$ se correspondent : ce sont les droites épipolaires. En changeant la position de X, on obtient deux faisceaux de droites en e et e'. A droite : position des plans images après rectification. Les épipoles se trouvent à l'infini et toutes les droites épipolaires sont parallèles. Les points homologues se trouvent sur une même droite.

satisfait

$$E = K^T F K.$$

Tout au long de cette thèse, on supposera que les images ont préalablement été rectifiées et satisfont la contrainte d'épipolarité. La mise en correspondance de points homologues est alors limitée à la recherche dans la direction épipolaire. Le problème a été ramené de deux à une dimension.

En imagerie satellitaire, le modèle d'acquisition est un modèle pousse-balai (*push-broom*) où une barrette de capteurs balaie le sol pour former l'image. Ce dispositif, adapté au mouvement du satellite, remplace la matrice de capteurs classique (CCD). Ainsi, les satellites d'observation terrestre capturent les images au fur et à mesure qu'ils avancent sur leurs orbites. Le système d'acquisition est connu précisément, et donc la calibration n'est pas nécessaire. Malheureusement, des microvibrations du satellite peuvent se produire pendant le temps de balayage-acquisition. La correction de ces microvibrations est encore une question ouverte mais les récents travaux sur le sujet [Grompone, 2009] sont très prometteurs.

Principe fondamental de la vision stéréoscopique

Après rectification, on peut considérer que la configuration des caméras est celle de la figure 1.3. Par similarité des triangles, on peut en déduire la relation entre la profondeur d'un point de la scène et le décalage entre les points homologues :

$$D = f\frac{B}{H},$$

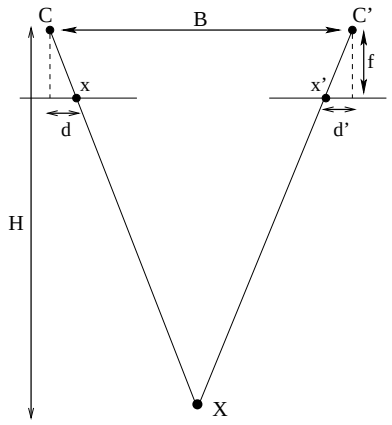


Figure 1.3: Images rectifiées. f est la focale, B la baseline et H la distance du point X à la baseline. On a $B/H = (d + d')/f$.

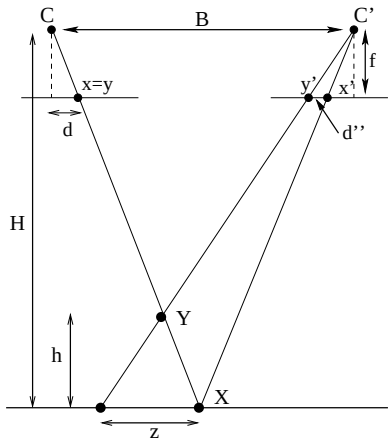


Figure 1.4: Précision.

où f est la focale de la caméra, B est la baseline, H est la distance du point X à la baseline et D est le décalage de position des projections du point X dans les deux images[2].

Cette relation donne le principe fondamental de la vision stéréoscopique. En effet, les changements de profondeur dans une scène en trois dimensions créent des disparités (décalages) géométriques entre les différentes images si celles-ci sont prises de points de vue différents. Ainsi, étant donné deux images, le fait de déterminer les points de chaque image qui correspondent au même point tridimensionnel de la scène permet de trouver la profondeur relative de ce point. Cette procédure est souvent appelée triangulation. Ainsi, le relief de la scène peut être reconstruit à l'aide des paramètres de calibration. Si l'on veut calculer la distance entre deux points de la scène, comme par exemple la hauteur d'un bâtiment par rapport au sol, on a (fig. 1.4)

[2]Remarque: Le décalage $D = 0$ correspond donc à une hauteur $H = \infty$ pour une paire calibrée et rectifiée en géométrie épipolaire. On verra plus bas comment calculer des décalages relatifs à une hauteur moyenne H finie.

$$\frac{B}{z} = \frac{H-h}{h}\,,$$
$$\frac{H}{f} = \frac{z}{d''}\,.$$

Ainsi, comme la résolution de l'image vérifie $R = {}^{H}/_{f}$, on a

$$d'' = \frac{B}{H-h}\frac{h}{R}\,.$$

En imagerie aérienne ou satellitaire, h est négligeable par rapport à H donc on peut simplifier la relation précédente en

$$d'' \backsimeq \frac{B}{H}\frac{h}{R}\,. \tag{1.1}$$

On peut en déduire que la précision sur la profondeur h de Y est proportionnelle à la précision du décalage d''. Si une erreur se produit dans le calcul de d'', ceci induit une erreur dans l'estimation de la profondeur de Y. Remarquons que, pour une erreur de décalage fixe, plus le rapport B/H est grand, plus l'erreur commise sur l'estimation de h est petite. C'est pour cette raison, qu'en général, le couple d'images stéréoscopiques est acquis avec un fort B/H. On verra que ce choix n'est pas toujours le plus judicieux.

1.3 Mise en correspondance des images

La mise en correspondance d'images est une étape importante du processus de reconstruction 3D et il conduit au problème suivant : étant données deux images u_1 et u_2 d'une paire stéréoscopique et $\mathbf{x} = (x, y) \in I$, on cherche la fonction décalage ε telle que

$$\begin{aligned} u_1(\mathbf{x}) &= u(x + \varepsilon(\mathbf{x}), y) + n_1(\mathbf{x})\ , \\ u_2(\mathbf{x}) &= u(\mathbf{x}) + n_2(\mathbf{x}). \end{aligned} \tag{1.2}$$

où n_i est le bruit d'acquisition, $u(\mathbf{x})$ est l'image secondaire idéale sans bruit et ε est la fonction de décalage appelée disparité. Cette fonction est souvent représentée comme une image de la même taille que u_1 et u_2 où, pour chaque point $\mathbf{x}$, on représente la valeur $\varepsilon(\mathbf{x})$ avec un niveau de gris. Cette image est usuellement appelée carte (ou nappe) de disparités.

La mise en correspondance d'images est un problème ardu et pas toujours bien posé. Les difficultés principales sont les suivantes :

- Le phénomène d'occlusion (fig. 1.5) : comme les images sont prises de points de vue différents, il y a des parties de la scène qui ne sont pas visibles simultanément dans les deux images. Aussi, l'existence d'un point homologue pour chaque point de l'image de référence n'est pas garantie. Les surfaces cachées sont plus ou moins grandes selon la complexité de la scène, mais aussi selon la distance B entre les centres optiques. Plus cette distance est grande relativement à la distance aux objets, et plus il y aura d'occlusions dans la scène.

- Changements radiométriques : il se peut que les images possèdent des changements de contraste locaux ou globaux ou que l'illumination de la scène change considérablement entre les deux images, ce qui rend la tâche de mise en correspondance difficile. D'abord, les surfaces ne sont pas toujours lambertiennes (i.e. la radiométrie n'est pas indépendante de la direction de l'observation). Par exemple, les surfaces réfléchissantes, très présentes en zones urbaines à cause des matériaux de construction utilisés, renvoient une partie des rayons lumineux spéculairement et ne se diffusent pas. Ensuite, la source d'illumination peut avoir changé entre les deux prises de vue. En effet, si le couple d'images est acquis non simultanément, la position du soleil, et donc celle des ombres, change entre les deux prises. Plus le temps et l'angle entre les deux prises de vue varie, et plus le changement de contraste risque d'être important.

- Manque d'information : la présence de zones non texturées dans les images est un inconvénient majeur. Elle provoque des ambiguïtés sources d'erreurs. Les ombres, souvent complètement dépourvues de texture, sont un des principaux problèmes de la mise en correspondance. Or, en milieu urbain, les ombres peuvent occuper presque la moitié de l'image. De plus, certaines zones de l'image peuvent être saturées, provoquant ainsi une perte de texture totale de la zone.

- Modifications de l'environnement : si la scène dont on souhaite trouver la profondeur change entre les prises de vue, il se produit un phénomène semblable au phénomène d'occlusion : certains points de l'image n'ont pas de point homologue. C'est le cas des objets disparaissants ou en mouvement. L'exemple le plus frappant est celui des véhicules et des piétons dans les images aériennes ou satellitaires.

- Modification de la géométrie locale : les objets de la scène peuvent apparaître déformés dans les images à cause des projections perspectives. C'est pourquoi les images sont souvent prises le plus perpendiculairement possible à la scène (le nadir en imagerie satellitaire ou aérienne). Néanmoins, il y a toujours des surfaces non parallèles au plan image, comme les façades de bâtiments en zones urbaines. Ces surfaces, si elles sont visibles sur les deux images, changent très fortement d'apparence.

- Le phénomène stroboscopique : une source d'erreurs classique est la présence de structures répétitives dans la direction épipolaire. Une texture uniforme ou plusieurs objets identiques apparaissant dans cette direction (tuiles sur les maisons, fenêtres, etc.) peuvent prêter à confusion car il devient alors très difficile d'identifier correctement un point et son homologue.

Le B/H faible

Si l'on se place dans le cadre d'un très petit angle entre les deux prises de vue, les occlusions, les changements radiométriques et les changements de géométrie locale sont considérablement réduits. Ainsi le modèle de déformation décrit en (1.2) est valide. [Delon and Rougé, 2007] ont proposé un modèle d'acquisition d'images avec un B/H faible de l'ordre de 0.1 ou même inférieur, au lieu des rapports classiquement utilisés de l'ordre de 1. Ce modèle réduit considérablement toutes ces difficultés rencontrées avec un B/H trop important. Par contre, dans ce cas, pour obtenir la même précision en hauteur, une plus grande précision sur le calcul des disparités s'impose comme on l'a observé à partir de la relation (1.1).

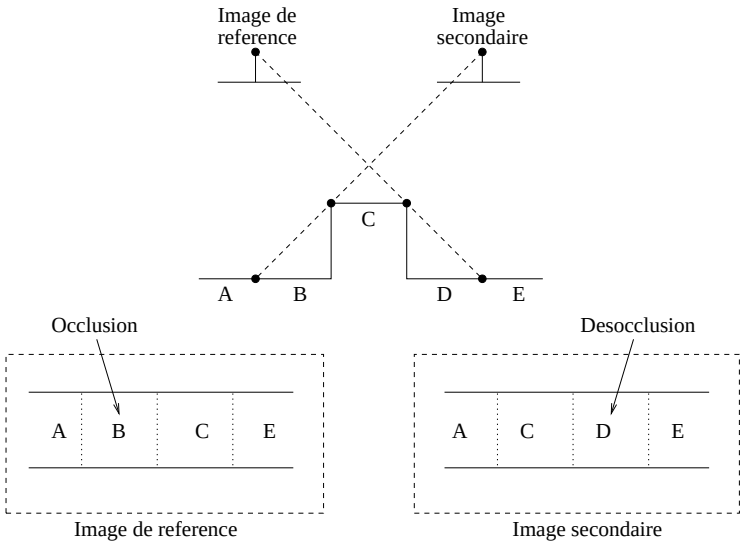


Figure 1.5: Phénomène d'occlusion. La partie B de la scène est visible sur l'image de référence mais pas sur l'image secondaire. B est une partie d'occlusion. Vice versa, la partie D apparaît seulement sur l'image secondaire. D est une zone de désocclusion.

Dans la littérature sur la vision stéréoscopique, le calcul de disparités subpixeliennes est un sujet rare. Le manque d'intérêt pour la précision subpixelienne s'explique par le manque de données bien échantillonnées. Sur ce point, les avancées réalisées par le CNES [Carfantan and Rougé, 2001] permettent d'approcher le sujet avec un nouveau regard. Par le passé, des auteurs comme [Birchfield and Tomasi, 1998b] ont proposé de faire un zoom de l'image avant d'utiliser un algorithme avec une précision au pixel près. Ceci s'avère être extrêmement coûteux quand la taille de l'image est importante.

Mais il y a eu aussi des propositions plus fines comme [Tian and Huhns, 1986] qui ont comparé plusieurs algorithmes de recalage d'images subpixéliens où l'on calcule les maxima d'une surface calculée sur la grille d'échantillonnage. Ces algorithmes peuvent être utilisés pour le raffinement de disparités en utilisant une méthode de descente de gradient itérative [Lucas and Kanade, 1981]. De façon similaire, on peut calculer la parabole interpolant le coefficient de corrélation mais, quand elle est calculée sur les positions entières, les décalages sont systématiquement biaisés. Cet effet est appellé *pixel-locking* [Shimizu and Okutomi, 2001]. Les travaux les plus avancés sur le sujet sont ceux de [Scharstein and Szeliski, 2002; Szeliski and Scharstein, 2004] et de [Delon and Rougé, 2007] qui nous ont servi de point de départ.

L'état de l'art

Le sujet de la mise en correspondance fait depuis plusieurs années l'objet de recherches variées, comme le témoigne la revue sur la stéréo faite dans [Dhond and Aggarwal, 1989]. Les articles [Szeliski and Zabih, 1999], [Brown et al., 2003] et [Scharstein and Szeliski, 2002] résument l'état de l'art. Le but de ce dernier est de faire une étude comparative des algorithmes

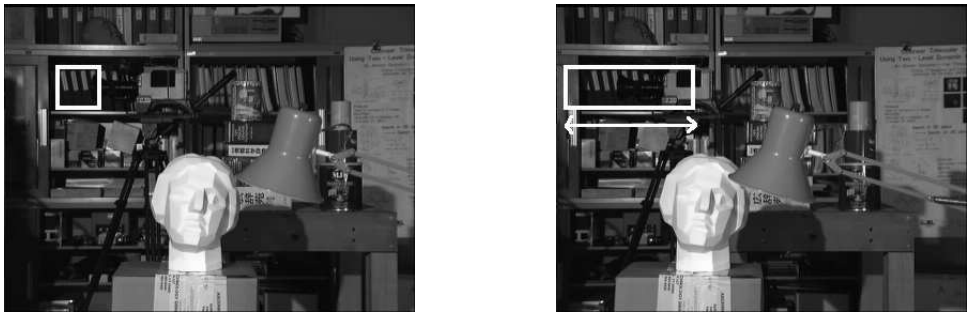


Figure 1.6: A gauche, l'image de référence, à droite l'image secondaire. Pour le calcul d'une carte de disparités par *block-matching*, pour chaque voisinage carré dans l'image de référence, on cherche le voisinage dans la seconde image qui ressemble le plus. Sous la contrainte d'épipolarité, la recherche peut être faite dans une seule direction.

publiés jusqu'en 2002 mais aussi de créer une base de données de référence sur le web[3] pour l'évaluation qualitative d'algorithmes en stéréo binoculaire fournissant des résultats denses à partir d'images rectifiées. De toute évidence, les images stéréo de Middlebury sont devenues le *benchmark* par excellence dans la communauté (plus de 60 articles publiés entre 1999 et 2009 se retrouvent dans leur tableau d'évaluation).

On trouve essentiellement deux stratégies pour la mise en correspondance. D'un côté, les approches locales qui estiment les disparités d'un point en utilisant son voisinage immédiat et, d'un autre côté, les approches globales (actuellement dominantes), qui utilisent une ligne de l'image ou toute l'image.

1.4 Les approches locales

Les approches locales souffrent du manque de texture locale, des occlusions et des ambiguïtés des effets stroboscopiques. Sur ces régions, les cartes de disparité ont souvent des fausses correspondances et sont peu fiables. Cependant, ces approches ne sont pas complexes et elles sont rapides, ce qui fait qu'elles conservent une place en stéréo.

1.4.1 *Block-matching*

Parmi les méthodes locales les plus connues se trouvent les méthodes dites de *block-matching*. Le principe de mise en correspondance par blocs est illustré sur la figure 1.6. Essentiellement, ces approches fonctionnent en deux étapes. D'abord, le calcul d'une fonction de coût C qui est la mesure de similarité entre les fenêtres ou blocs de l'image et ensuite l'estimation μ de la disparité ε en chaque point $\mathbf{x}_0$ comme étant le décalage qui minimise la fonction de coût

$$\mu(\mathbf{x}_0) = \arg\min_d C(\mathbf{x}_0, d).$$

Cette optimisation est appelée WTA (*winner-take-all*). Comme ni l'unicité ni l'existence de correspondances n'est garantie, cette optimisation peut donner des mauvais résultats car il

[3] http://vision.middlebury.edu/stereo/

peut y avoir une convergence vers le mauvais minimum. Les différences de performance entre les méthodes de *block-matching* résident dans le choix de la fonction de coût. Le choix de cette fonction est très large mais les plus classiques sont :

- SAD (Sum of Absolute Diferences) :

$$C(\mathbf{x}_0, d) = \sum_{\mathbf{x} \in B_{\mathbf{x}_0}} |u_1(\mathbf{x}) - u_2(\mathbf{x} + (d, 0))|,$$

  où $B_{\mathbf{x}_0}$ est le bloc associé au point $\mathbf{x}_0$. Il s'agit de la plus simple des mesures en terme de complexité de calcul. Des exemples de travaux où elles ont été utilisées sont [Kanade et al., 1995] et [Mühlmann et al., 2001].

- SSD (Sum of Squared Diferences)

$$C(\mathbf{x}_0, d) = \sum_{\mathbf{x} \in B_{\mathbf{x}_0}} \big(u_1(\mathbf{x}) - u_2(\mathbf{x} + (d, 0))\big)^2,$$

  comme une variante de la précédente. Elle est utilisée dans [Okutomi and Kanade, 1993].

- NCC (Normalized Cross Correlation)

$$C(\mathbf{x}_0, d) = \frac{\sum_{\mathbf{x} \in B_{\mathbf{x}_0}} \big(u_1(\mathbf{x}) - \bar{u}_1\big)\big(u_2(\mathbf{x} + (d, 0)) - \bar{u}_2\big)}{\sqrt{\sum_{\mathbf{x} \in B_{\mathbf{x}_0}} \big(u_1(\mathbf{x}) - \bar{u}_1\big)^2 \big(u_2(\mathbf{x} + (d, 0)) - \bar{u}_2\big)^2}},$$

  où $\bar{u}_1$ et $\bar{u}_2$ sont les moyennes de l'image dans le voisinage B. Il s'agit sûrement de la mesure de similarité la plus couramment utilisée en stéréo. Grâce à la normalisation et à la soustraction des moyennes locales, cette mesure est invariante aux changements de contraste affines entre les images. On utilise aussi la version normalisée sans soustraction des moyennes mais l'invariance est alors seulement aux changements de contraste linéaires. Remarquons que la mesure NCC requiert une maximisation de la fonction de coût, alors que pour SSD et SAD c'est une minimisation. [Faugeras et al., 1993] [Hirschmuller et al., 2002].

Les fonctions de coût précédentes sont basées sur les niveaux de gris de l'image, et elles sont performantes sous réserve qu'il n'y ait pas de changement de contraste entre les images (où de changement de contraste affine pour NCC). Quand le couple d'images ne satisfait pas cette hypothèse, l'utilisation d'autres mesures de similarité s'impose. [Caselles et al., 2005] ont défini une nouvelle mesure qui compare les directions du gradient le long des courbes de niveau. Comme les courbes de niveau d'une image sont invariantes aux changements de contraste, la mesure de similarité proposée l'est aussi. [Hirschmuller and Scharstein, 2007] ont évalué l'insensibilité aux changements radiométriques de différentes fonctions de coût.

Dans la littérature, on trouve encore d'autres mesures de similarité qui essaient de remédier au problème des fausses correspondances dont souffrent souvent les mesures de similarité classiques. [Bhat and Nayar, 1998] ont proposé une mesure de similarité qui est plus robuste

aux distortions dues aux projections perspectives. [Birchfield and Tomasi, 1998a] ont suggéré une fonction de coût qui pénalise les occlusions et qui est moins sensible à l'échantillonnage des images en comparant les niveaux de gris de la première image avec l'interpolation linéaire de la seconde. Enfin, [Aschwanden and Guggenbuhl, 1993] et [Chambon and Crouzil, 2003] sont des études comparatives approfondies de toutes ces mesures.

Le phénomène d'adhérence

Le phénomène d'adhérence est un problème inhérent aux méthodes de *block-matching*. (cf. chapitre 3). Ce phénomène se produit lorsque les décalages entre u_1 et u_2 ne sont pas constants sur le support de la fenêtre de comparaison B et qu'il y a une variation importante des niveaux de gris dans ce même voisinage. Le pire des cas se produit quand les contours des objets présentent un fort contraste et qu'il y a un saut de disparité (fig. 1.7). Ce phénomène se caractérise par une mauvaise estimation du relief au voisinage des bords contrastés, ce qui peut conduire à une dilatation (*fattening*) des objets de la scène. Les images de zones urbaines sont fortement atteintes par ce phénomène. La reconstruction de la scène fait grossir les bâtiments. Ces erreurs dépendent de la fonction de coût [Delon and B. Rougé, 2001] et du voisinage B choisi.

Pour réduire l'adhérence, on aurait tendance à réduire la taille du voisinage, mais plus la fenêtre est petite plus elle est sensible au bruit et aux changements de contraste. C'est ce dilemne qui conduit à l'utilisation de fenêtres adaptatives en taille et en forme. Cette idée est très développée dans la littérature : [Kanade and Okutomi, 1994] proposent une méthode itérative pour trouver la taille optimale de la fenêtre en chaque point en partant d'une fenêtre de taille très petite, et [Lotti and Giraudon, 1994] établissent la taille de la fenêtre sous la contrainte de ne pas superposer les contours de l'image. Il y a aussi les méthodes dites à plusieurs fenêtres, c'est-à-dire qu'un ensemble initial de fenêtres fixes est considéré en chaque point. [Fusiello et al., 1997] choisissent la fenêtre qui minimise la fonction de coût parmi un ensemble de 9 fenêtres fixes et [Kang et al., 2001] prennent en compte toutes les fenêtres contenant le pixel (*shiftable windows*). Au lieu de considérer un nombre limité de fenêtres, [Veksler, 2001] évalue en chaque point un large éventail de fenêtres qui différent en taille et forme grâce à une optimisation efficiente via un algorithme de type MRC (*minimum ratio cycle*).

Une autre variante des fenêtres adaptatives sont les fenêtres pondérées. [Gong and Yang, 2005] utilisent des fenêtres de taille 5×5 et donnent des valeurs dans $\{0, 1, 2, 4\}$ à chaque pixel de la fenêtre. Mais il existe des variantes où la taille de la fenêtre n'est pas fixe. L'idée est que chaque pixel dans le voisinage d'un pixel d'intérêt reçoit un poids différent, et la fonction de coût est calculée avec ces pondérations. Ainsi, le concept de forme et taille de la fenêtre perd son sens, car on ne distingue pas les points appartenant à la fenêtre ou pas. [Yoon, 2006] calcule les poids selon les lois de la Gestalt pour grouper des points. Avec cette idée, les poids sont calculés selon la similarité des couleurs et la distance entre les points. [Scharstein and Szeliski, 1998] utilisent une fenêtre où la taille est calculée par diffusion non linéaire et [Xu et al., 2002] présentent un algorithme qui utilise une carte de disparités initiale et la similarité des couleurs pour déterminer le support pondéré de la fenêtre.

L'utilisation de mesures plus robustes est une autre option pour la correction de l'effet d'adhérence : [Zabih and Woodfill, 1994] ont proposé de nouvelles fonctions de coût combinées avec la corrélation pour obtenir de meilleurs résultats au niveau des contours des objets. Il s'agit de transformations locales non-paramétriques (*rank transform* et *census tranform*).

[Gong et al., 2007] est une étude comparative récente de méthodes de *block-matching* en

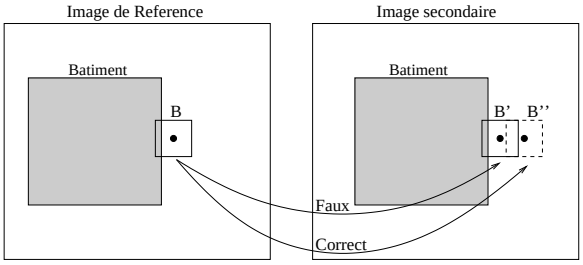


Figure 1.7: Couple d'images stéréoscopiques avec erreur d'adhérence. Les pixels sur le bâtiment ont un décalage différent des pixels au sol. Le patch B avec son pixel central sur le sol contient des pixels sur le bâtiment. Le patch dans l'autre image qui ressemble le plus à B est B' car il contient le contour du bâtiment, qui est très significatif, au même endroit. Le pixel central de B sera faussement attribué de la même disparité que les pixels du bâtiment. La correspondance correcte de B est B''.

temps réel où la performance est évaluée en termes de qualité de la carte de disparités mais aussi en temps de calcul. En particulier, ils comparent la performance des fenêtres fixes avec des fenêtres pondérées ou adaptatives sur les zones de non occlusion et sur les discontinuités de la carte de disparités. Une autre étude récente sur la comparaison de fenêtres adaptatives est [Tombari et al., 2008].

Le phénomène d'occlusion

Dans la plupart des cas, le phénomène d'adhérence est traité en même temps que le phénomène d'occlusion car ils sont extrêmement liés [Sara and Bajcsy, 1997]. Mais certains auteurs ne font même pas de distinction entre l'un et l'autre, et appellent naïvement phénomène d'occlusion l'adhérence.

Pour le traitement des occlusions, plusieurs approches sont possibles. L'option la plus simple est de détecter les pixels occlus avant ou après le calcul de disparités et de remplacer leur disparité par une interpolation des pixels non occlus. Une manière de détecter ces pixels est de vérifier la cohérence symétrique (droite-gauche) de la mise en correspondance. C'est-à-dire que si on calcule deux cartes de disparités en prenant chacune des images comme image de référence, on peut marquer comme pixels occlus les pixels avec une disparité inconsistante dans les deux cartes de disparités [Chang et al., 1991] [Fua, 1993]. Cependant, dans les méthodes de *block-matching* classiques où les cartes de disparités sont assez bruitées, de nombreux pixels peuvent être incohérents pour une autre raison que l'occlusion. L'utilisation de plusieurs caméras permet de modéliser géométriquement l'occlusion. C'est ce que font [Okutomi and Kanade, 1993]. Ils utilisent une caméra bougeant dans une direction et capable d'obtenir plusieurs images d'une même scène sous plusieurs rapports B/H.

1.4.2 *Feature-matching*

Parmi les méthodes locales, il existe aussi des méthodes de *feature-matching* qui évitent tous les problèmes liés à la taille de la fenêtre de comparaison, mais qui produisent des cartes de disparités plus éparses que les méthodes de *block-matching*. Les objets qui peuvent être mis en correspondance sont : des contours [Bignone et al., 1996], des courbes [Marr and Poggio,

1976], des morceaux de ligne de niveau [Musé et al., 2006a], des lignes [Schmid and Zisserman, 2000], des coins [Harris and Stephens, 1988] [Cao, 2004], des descripteurs locaux [Mikolajczyk and Schmid, 2003], des descripteurs SIFT [Lowe, 2004] [Rabin et al., 2008] ou ASIFT [Morel and Yu, 2009], etc. Parfois il s'agit de la combinaison de plusieurs objets, comme [Venkateswar and Chellappa, 1995] qui ont mis en place un algorithme hiérarchique (*coarse-to-fine*) de mise en correspondance de lignes, coins, contours et surfaces.

Les méthodes les plus denses de *feature-matching* sont sans doute les méthodes dites *segmentation-based* ou *region-based* où l'objet mis en correspondance est une région de l'image. Les différences se trouvent dans la façon dont les régions d'intérêt sont définies, et la stratégie utilisée pour le matching des régions. Par exemple, [Randriamasy, 1991] décrit un algorithme récursif qui détermine les régions en utilisant un seuil sur le contraste suivi d'un traitement morphologique sur la région. [Veksler, 2002a] trouve des régions connexes qui se correspondent dans les deux images. Les régions sont définies de façon que les niveaux de gris sur les bords de la région soient plus importants que la différence de niveaux de gris entre les deux images sur ces mêmes pixels. D'autres méthodes font une segmentation de l'image pour ensuite mettre en correspondance chaque région. Parfois, il peut y avoir une étape de fusion (*merging*) de régions initiales [Garrido and Salembier, 1998].

Un certain nombre d'algorithmes imposent une transformation affine pour chaque région de l'image. Chaque région se voit associée à une transformation T affine à 3 paramètres (la contrainte d'épipolarité permet la réduction de paramètres)

$$T : \begin{pmatrix} x \\ y \end{pmatrix} \longmapsto \begin{pmatrix} a & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} c \\ 0 \end{pmatrix} .$$

C'est le cas de [Birchfield and Tomasi, 1999] qui segmentent l'image en petites surfaces plates et fusionnent les régions qui ont une transformation similaire. Semblablement [Igual et al., 2007] calculent des cartes de disparité denses pour des images de zones urbaines mais avec un critère de fusion plus sophistiqué basé sur les méthodes *a contrario* et la théorie de la Gestalt permettant de rejeter le modèle affine sur les zones de végétation.

Les méthodes de segmentation ne souffrent pas d'adhérence et estiment avec bonne précision la profondeur des plans penchés de la scène. Mais elles sont très sensibles à la segmentation initiale de l'image.

1.4.3 Méthodes de gradient

Les méthodes d'*optical flow* sont bien connues en estimation de mouvement dans les séquences vidéo. Elles supposent que l'intensité d'un objet ne change pas entre une image et la suivante. Ces méthodes sont adaptées à la stéréo pour calculer les décalages entre points homologues en résolvant l'équation

$$(\nabla_x u)v + u_t = 0 \,, \tag{1.3}$$

où $\nabla_x u$ est la composante horizontale du gradient de l'image, v est le décalage entre les deux images et u_t est la dérivée du temps (différence d'intensités du couple d'images en stéréo). Trouver v avec cette équation est un problème mal posé quand le gradient est proche de 0. Dans ce cas, des contraintes supplémentaires sont nécessaires.

Par exemple, [Lucas and Kanade, 1981] ont ajouté une contrainte de régularité locale dans le champ de vecteurs. Cette méthode ne donne pas une nappe dense de disparités car elle

dépend de la taille du voisinage choisi. Le choix d'un plus grand voisinage donne des nappes de disparités trop régulières.

La contrainte d'épipolarité simplifie les calculs car le gradient doit seulement être calculé dans une direction mais le fait de n'avoir que deux images peut être contraignant pour les méthodes qui supposent une continuité dans le temps (équation 1.3). En principe, la précision des méthodes de gradient est d'un demi pixel car le gradient d'une image n'est précis que sur la demi-grille $\mathbb{Z}^{1/2}$. En pratique, si l'image est bien échantillonnée, la dérivée peut être calculée plus précisément [Farid and Simoncelli, 2004].

1.4.4 Méthodes de phase

Les méthodes de phase se basent sur le fait que la disparité entre les images crée un décalage de la phase dans le domaine de Fourier. Le principal avantage du calcul de disparités dans le domaine de Fourier est la facilité avec laquelle on peut calculer les décalages subpixeliens très précis. Ces méthodes sont très performantes sur les zones texturées. [Fleet et al., 1991] [Weng, 1993] [Frohlinghaus and Buhmann, 1996] [El-Etriby et al., 2006]. En revanche, ce type de méthodes ne permettent de calculer que un décalage constant à l'intérieur de la fenêtre à laquelle on applique la Transformée de Fourier (TF). Sauf dans le cas où le décalage est globalement constant ceci oblige à l'application de TF locales, avec la perte de performance associée aux artifices necessaires pour éviter les problèmes liées à la périodisation implicite dans la TF.

1.5 Les approches globales

Les approches globales imposent des contraintes de régularité à la carte de disparités ce qui fait qu'elles soient moins sensibles à l'adhérence et aux ambiguïtés locales de l'image (zones sans texture, structures répétitives, etc.) Néanmoins, à cause de leur complexité calculatoire, elles n'ont pas complètement supplanté les méthodes locales. De plus, elles ont l'inconvénient de propager les erreurs. Alternativement, les méthodes globales peuvent être utilisées comme une seconde étape pour régulariser ou interpoler une carte de disparités obtenue avec une méthode locale.

Ces approches se présentent sous la forme de méthodes d'optimisation globale qui calculent la fonction de disparité sur tous les pixels de l'image ou toute une ligne de l'image à la fois. En effet, la carte de disparités est calculée en minimisant une énergie de la forme

$$E_{\text{total}}(d) = E_{\text{données}}(d) + \alpha E_{\text{régul}}(d)\,. \tag{1.4}$$

où $E_{\text{données}} = \sum_{\mathbf{x}_0} C(\mathbf{x}_0, d)$ est le terme d'attache aux données (C est une fonction de coût, comme celles qu'on a présenté précédemment), $E_{\text{régul}}$ est le terme de régularisation et le paramètre α contrôle le niveau de régularisation dans la nappe de disparités. Quand le terme de régularisation est quadratique, la surface reconstruite peut être parfois trop lisse et donc mal adaptée aux contours des objets. Certains auteurs comme [Black and Rangarajan, 1996] ont essayé de remédier à ce problème.

Une fois que l'énergie a été définie, il y a plusieurs méthodes d'optimisation pour en trouver le minimum comme la programmation dynamique ou les *Graph-Cuts*.

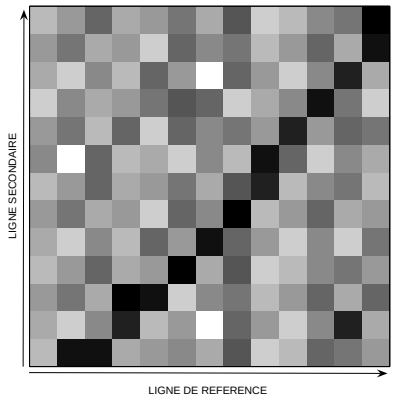


Figure 1.8: Schéma de la mise en correspondance stéréo par programmation dynamique. Pour chaque couple de lignes épipolaires associées, une matrice M de coût est définie. Chaque coefficient de cette matrice M(i,j) se calcule comme le coût partiel entre les deux positions, i de la ligne de référence et j de la ligne secondaire. Le chemin prenant les valeurs minimales (pixels plus foncés) dans la matrice, partant du coin inférieur gauche jusqu'au coin supérieur droit, est le chemin qui correspond à la disparité de la ligne de référence.

1.5.1 Programmation dynamique

Les algorithmes basés sur la programmation dynamique cherchent à résoudre (1.4) ligne par ligne, ce qui fait qu'ils sont plus rapides que le reste des méthodes globales. En effet, ils profitent de la rectification épipolaire des images pour déterminer la disparité d'une ligne de l'image en cherchant le chemin minimal dans une matrice de coût M entre cette ligne et la ligne épipolaire associée (illustration sur la fig. 1.8).

En général, aucune condition n'est imposée dans l'axe vertical (perpendiculaire à l'axe épipolaire) pour le calcul des disparités, ce qui peut créer des discontinuités le long des lignes épipolaires. C'est pourquoi certaines approches prennent en compte la cohérence inter-ligne dans la fonction de coût [Wu and Maître, 1989]. En revanche, la contrainte d'ordre est souvent imposée. C'est-à-dire que le chemin dans M doit être croissant (un chemin décroissant impliquant une occlusion). Cette contrainte conduit à des erreurs quand il y a des objets très fins dans la scène.

La matrice de coût est calculée avec une des mesures de similarité mentionnées précédemment. Quand cette matrice est calculée pour toutes les lignes épipolaires d'une image, le cube qui se forme est équivalent au DSI (*Disparity Space Image*) défini dans [Szeliski and Scharstein, 2004]. Parmi les algorithmes adoptant cette optimisation, on trouve le papier fondamental [Ohta and Kanade, 1985] mais aussi des travaux plus récents [Forstmann et al., 2004], [Wang et al., 2006], [Deng and Lin, 2006] et [Lei et al., 2006]. Ces algorithmes sont, pour la plupart en temps réel, et ils peuvent inclure des conditions spécifiques pour les occlusions et pour les contours. Cependant, ils ne sont pas les algorithmes les plus performants selon le critère d'évaluation Middlebury[4].

[4] http://vision.middlebury.edu/stereo

1.5.2 Graph-Cuts

Le principe des algorithmes de *Graph-Cuts* est de chercher un chemin de flot maximal (ou une coupe de poids minimal) dans un graphe pondéré et orienté. Ces méthodes connues depuis les années 50 ont été adaptées au domaine du traitement d'images récemment. La première adaptation à la stéréo est due à [Roy and Cox, 1998]. Le sujet a progressé avec [Boykov et al., 2001], [Boykov and Kolmogorov, 2004], [Kolmogorov and Zabih, 2001] et [Kolmogorov and Zabih, 2005] qui montrent comment les *Graph-Cuts* peuvent être utilisés de manière efficace. Cependant, leur utilisation en stéréo aérienne et satellitaire est encore limitée par deux facteurs. D'une part, étant donné la très grande taille des images à traiter et les précisions subpixeliennes désirées dans ces domaines, les graphes qui en résultent deviennent très difficiles à exploiter. D'autre part, le terme de régularité utilisé pénalise généralement le fait que deux pixels voisins n'aient pas la même disparité. Les cartes de disparités reconstruites sont donc constantes par morceaux, ce qui limite leur intérêt pour la reconstruction de zones urbaines, surtout pour la très haute résolution (THR). La figure 1.9 montre la carte de disparités obtenue avec *Graph-Cuts* sur une paire simulée d'images.

1.6 Evaluation de résultats

Soit μ la carte de disparités estimée et soit ε la carte de disparités réelle (la vérité terrain). Pour l'évaluation qualitative de l'estimation μ de ε, il y a le choix de la mesure de comparaison. Les mesures les plus souvent utilisées sont:

- RMSE (*Root Mean Squared Error*) :

$$\frac{\sum_{\mathbf{x}\in M}\left(|\mu(\mathbf{x})-\varepsilon(\mathbf{x})|^2\right)^{1/2}}{\#M},$$

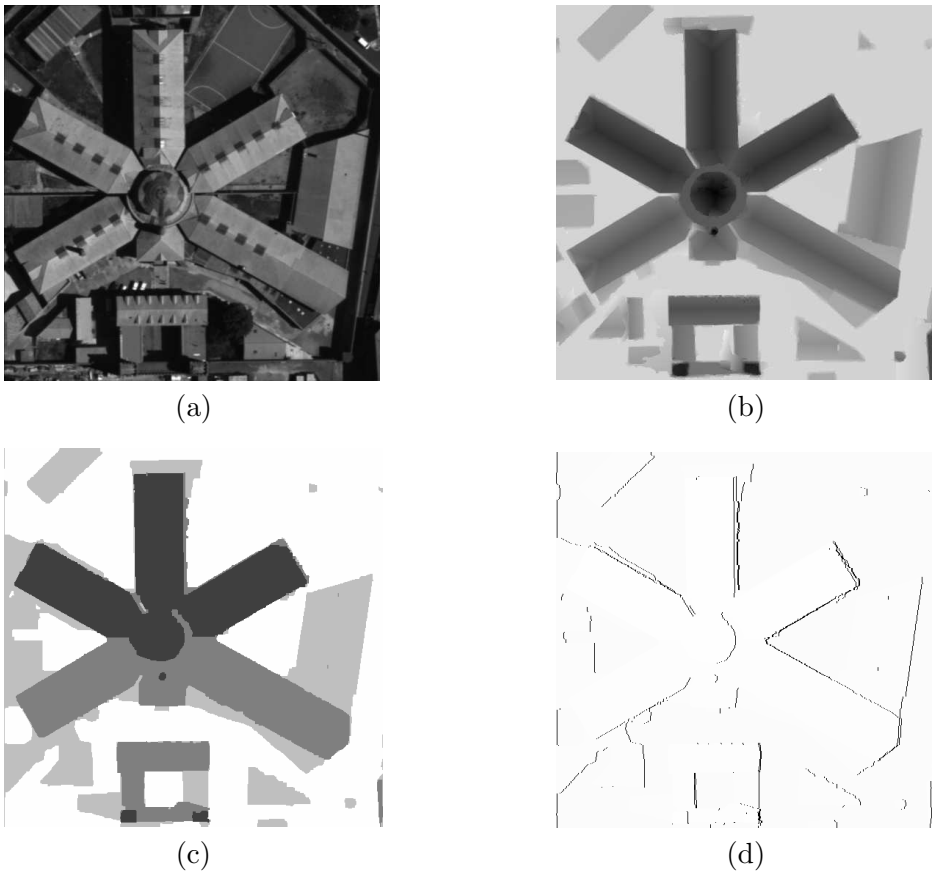
  où M est l'ensemble de pixels considéré.

- Pourcentage de fausses correspondances :

$$\frac{100}{\#M}\sum_{\mathbf{x}\in M}\left(|\mu(\mathbf{x})-\varepsilon(\mathbf{x})|>\lambda\right),$$

  où λ est la tolérance de l'erreur. Dans nos travaux, on a considéré $\lambda = 1$.

L'évaluation de résultats en stéréo est assez délicate car la vérité terrain n'est pas toujours connue et dans les cas où elle l'est, elle n'est pas très précise. Pour remédier à ce problème, on fera une validation croisée de résultats pourvu qu'on ait plusieurs images orthorectifiées de la scène.



(a) (b) (c) (d)

Figure 1.9: (a) Image de référence de la paire simulée. (b) Vérité terrain utilisée pour créer l'image secondaire. (c) Carte de disparités obtenue avec *Graph-Cuts*. Chaque pixel de l'image est labelisé avec un niveau de gris représentant la hauteur en chaque point parmi un ensemble de valeurs discrètes préétablies au départ. La discrétisation imposée par la méthode fait que beaucoup de détails sont perdus, c'est le cas des pentes des toits. (d) Les pixels noirs sont les pixels qui ont été marqués comme occlusions.

Chapter 2

Meaningful Matches in Stereo

Contents

Résumé : Il y a deux classes d'algorithmes de vision stéréoscopique binoculaire. D'un côté, les méthodes locales qui effectuent la mise en correspondance par blocs et d'un autre côte, les méthodes globales qui minimisent une énergie avec un terme d'attache aux données, un terme de régularité et parfois un terme qui contrôle la quantité d'occlusion. Ces dernières années, il y a eu un déferlement de méthodes globales qui obtiennent les meilleures performances dans les *benchmarks* actuels. Par contre, les méthodes locales souffrent de quatre inconvénients: elles peuvent se laisser berner par les motifs répétitifs des images, elles sont sensibles à des erreurs dûes au bruit, elles souffrent de l'effet d'adhérence qui peut propager des erreurs à une distance d'un demi-bloc, enfin, elles ne sont pas denses, car plusieurs blocs peuvent manquer d'information pour être appariés de façon fiable. En revanche, les méthodes globales sont denses, sont capables d'éviter l'effet d'adhérence, prennent souvent la bonne décision avec les motifs répétitifs, et sont moins sensibles au bruit.

Néanmoins, notre objectif est de montrer que l'appariement par blocs a au moins une utilité: elle peut mener à des règles de décision théoriquement fiables. Ce chapitre propose une nouvelle méthode qui s'assure que les appariements ne sont pas des blocs sélectionnés par hasard. La méthode est basée sur la création d'un modèle de fond statistique simple mais fidèle pour les blocs d'une image. La règle de rejet/acceptation de correspondances utilise une méthode (*a contrario*) garantissant que, sous le modèle de fond, pas plus d'un mauvais appariement se produit en moyenne sur toute l'image. La méthode *a contrario* (AC) de rejet est beaucoup plus précise qu'un simple seuil d'auto-similarité (SS). (Bien que, le seuil SS montre une certaine utilité complémentaire pour éviter les erreurs stroboscopiques dûes à des formes répétitives.)

Plusieurs applications sont envisagées. La première est la détection de correspondances fiables dans des régions de l'image inattendues, généralement des ombres. La seconde consiste à détecter tous les pixels qui n'ont aucune information utile pour la mise en correspondance. Une telle information est certainement importante, non seulement pour les méthodes d'appariement par blocs, mais aussi pour toutes les méthodes globales. Il existe une dernière application pour les couples d'images stéréoscopiques non simultanées, elle sera illustré par des images aériennes: l'élimination fiable de tout mouvement incohérent dûs aux véhicules et aux personnes.

Abstract: There are roughly two classes of algorithms in binocular stereo vision. Local methods perform block-matching, and global methods minimize a cost functional with a comparison term, a regularity term and sometimes a term controlling the amount of occlusion. Recent years have actually seen a blooming of global methods, which reach the best performance in the recent benchmarks. Local methods suffer from four drawbacks: they can be fooled by repetitive patterns in the pair; they are prone to errors due to noise; they suffer the fattening effect which can propagate a depth estimate at a half block distance; finally they are not dense, since many blocks can be too flat to be matched reliably. In contrast, global methods are dense, avoid the fattening effect, often take the right decision with repetitive patterns, and are less sensitive to noise.

Yet, our goal is to show that block-matching has at least one function left: it can lead to information-theoretically reliable decision rules. This chapter proposes a new method ensuring that selected block matches are not likely to have occurred "just by chance". The method is based on the generation of a simple but faithful statistical *background model* for image blocks. The ensuing rejection/acceptation process uses an *a contrario*

method guaranteeing that, under the background model, no more than one wrong block match occurs on average for the whole image. The *a contrario* (AC) rejection method is much more accurate than a simple *self-similarity threshold (SS)*. (Still, the SS threshold shows some complementary usefulness to avoid stroboscopic errors due to repetitive shapes.)

Several applications are considered. The first one is the detection of reliable matching points in unexpected image regions, typically shadows. The second is to mark all pixels which retain no useful stereo-matching information. Such an information is definitely relevant, not only to block-matching methods, but actually to all global methods. A final application to *non simultaneous stereo* will be illustrated with aerial imagery: the reliable elimination of incoherent motions due to vehicles and people.

2.1 Introduction

2.1.1 Stereo in Urban Areas

The matching of digital stereo images has been studied in depth for four decades. We refer to [Brown et al., 2003] and [Scharstein and Szeliski, 2002] for a fairly complete comparison of the main methods. Stereo algorithms aim at reconstructing a 3D model from two or more images of the same scene acquired from different angles. Assuming for a sake of simplicity that the cameras are calibrated, and that the image pair has been stereo-rectified, our work will focus on the matching process (Fig.2.1). Our main goal is to build an information theoretic method guaranteeing a very small false matches number. The proposed method will be tested with the simplest of all stereo algorithms, block-matching.

Stereo depth reconstruction algorithms are of very different nature. Global methods aim at a global and coherent solution obtained by minimizing an energy functional containing matching fidelity terms and regularity constraints. The most efficient ones seem to be Belief Propagation [Klaus and Sormann, 2006] [Yang et al., 2006], Graph Cuts [Kolmogorov and Zabih, 2005] and Dynamic Programming [Ohta and Kanade, 1985],[Forstmann et al., 2004]. These methods are much less sensitive to the fattening problem than block-matching. They often resolve ambiguous matches by maintaining a coherence along the epipolar line. They rely on a regularization term to eliminate outliers and reduce the noise. They are, however, at risk to propagate errors, or introduce new ones if the regularization term is not in accordance with the underlying surface. Another drawback of energy minimization methods is that usually there are parameters which are difficult to set. Local methods are simpler but more sensitive to local ambiguities. Such algorithms start by comparing features of the right and left images. These features can be blocks in block-matching methods, or even non-dense local descriptors [Mikolajczyk and Schmid, 2003] in the SIFT based methods [Lowe, 2004] [Rabin et al., 2008], or corners [Harris and Stephens, 1988] [Cao, 2004], etc.

The most common local method is block-matching, which compares blocks by Normalized Cross Correlation (NCC), or Sum of Squared Differences (SSD). Block-matching methods suffer from three mismatching causes that must be tackled one by one:

1. The main mismatch cause is the absence of a theoretically well founded threshold to decide whether two blocks really match or not, or if the match is merely casual and due to noise. Our main goal here will be to define such a threshold by an *a contrario* (AC) rejection rule.

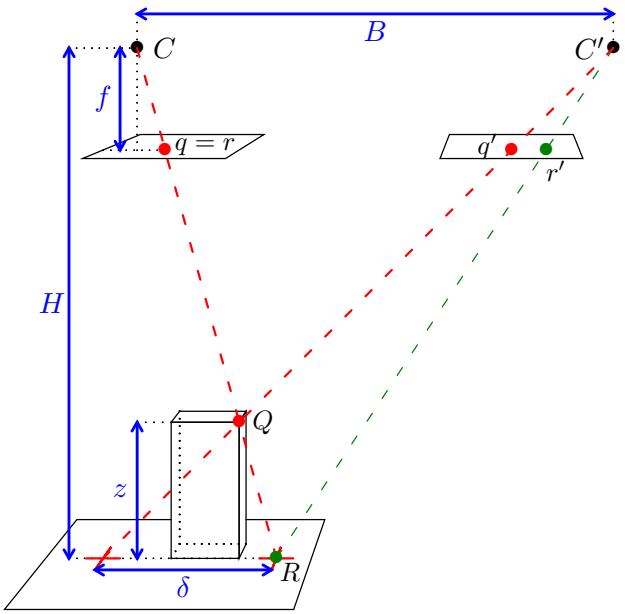


Figure 2.1: C and C' are the optical centers of two cameras. The distance between them is the baseline B and f is the focal length. $\mathbf{q}$ and $\mathbf{q}'$ are the projections of the scene point Q, and $\mathbf{r}$ and $\mathbf{r}'$ are the projections of R. By similar triangles we have $\mathbf{q}'\mathbf{r}' = \frac{f}{H}\delta$ and $\delta \simeq \frac{B}{H}z$.

2. The second mismatch cause is the presence on the epipolar line of repetitive shapes or textures, a problem sometimes called "stroboscopic phenomenon", or "self-similarity". Once AC has been applied, only a few repetitive patterns are at risk, and an elementary self-similarity rule (SS) will eliminate them. We shall also verify that this rule by itself is far from reaching the AC performance. It must be applied only as a complementary safeguard.

3. Block matching computes wrong disparities near intensity discontinuity edges in the image coinciding to depth discontinuities. This phenomenon is called "adhesion", "fattening effect" or "boundary problem" and is acute in urban scenes where it can lead to the apparent dilation of buildings. Thus no study on block-matching is complete without a fattening elimination step, and we will provide an efficient one in Chapter 3.

The elimination of these three sorts of mismatches is a key issue in block-matching methods. This problem has of course been addressed many times. We shall discuss a choice of the significant contributions. In [Sara, 2002], the two first causes of mismatch are considered, namely the mismatches on weakly textured objects and periodic structures. A *confidently stable matching* is defined in order to establish the largest possible unambiguous matching at a given confidence level. The method has two parameters that control the compromise between the percentage of bad matches and the density of the map. Yet, the density falls dramatically when the percentage of mismatches decreases. We will see that the method presented here is able to get denser disparity maps with less mismatches. Similarly, [Manduchi and Tomasi, 1999] try to eliminate errors on repeated patterns. Yet their density of matches seems to

concentrate mainly on image edges. This is problematic, since edges are precisely where the bigger fattening errors can occur. Actually neither [Sara, 2002] nor [Manduchi and Tomasi, 1999] propose a solution to the fattening problem. We will deal with the fattening phenomenon on Chapter 3.

Small Baseline

It is usually assumed that the angles between the views have to be large enough to have a good reconstruction. However, occlusions, large geometric deformations and brightness changes make the matching process difficult and hinder anyway the obtaining of a dense disparity map. These difficulties can be overcome with small baseline stereo pairs, provided highly accurate sub-pixel matching compensates the small B/H ratio, as proposed in [Giros et al., 2004]. The sub-pixel accuracy is not the aim of this chapter (see Chapter 4) but its feasibility has been studied in [Szeliski and Scharstein, 2002]. Our method will be mainly tested on small base-line stereo pairs, because in such methods the validity of the block-matching approach is maximal, and their applicability to satellite imaging promising.

2.1.2 An *A Contrario* Methodology

We shall focus on the simplest and probably the most popular stereo-matching method, namely block-matching, which iteratively compares patches, or blocks, of the left image with blocks of the right image in an epipolar neighborhood. However, our goal throughout this chapter is to eliminate unreliable matches, no matter how they have been obtained. Thus, nothing hinders the *a posteriori* check of matches obtained by any other method than block-matching. Only, this *a posteriori* check will be done by comparing a block around the given pixel to its best matching block, and computing whether the match is meaningful, or not. Because of occlusions and flat areas, we cannot presuppose the existence of uniquely determined correspondences for all pixels in the image. Thus, a decision must be taken on whether a block in the left image actually meaningfully matches or not its best match in the right image.

This problem will be addressed by the *a contrario* approach proposed in [Desolneux et al., 2007]. This method is an adaptation to image analysis of classic hypothesis testing. The basic assumption of this method is the so called Helmholtz principle, according to which all perceived structures can be characterized as having a low probability of occurring in noise. Detecting events in images against a background noise model was first proposed in Computer Vision by [Lowe, 1985], [Grimson and Huttenlocher, 1991], and [Stewart, 1995]. The *a contrario* approach for image comparison has anterior works, which will be discussed now.

2.1.3 Precursors, Previous Statistical Decision Methods

Robin's a contrario change detection. [Robin et al., 2009] describe a method for change detection in a set (a time series) of earth observation images. Like our method, Robin's method is based on an *a contrario approach* which allows to limit the expected number of false detections (NFA) that can occur by chance. In addition, Robin *et al.* detects changes as the complement of a maximal region where the time series does not change significantly. Thus, what is controlled by the *a contrario* method is the NFA of this maximal region of no-change. Hence, Robin's method could also be regarded as a method for matching (with

controlled false positive rates) image regions between two or more images. It is fundamentally different from the method we shall present in several ways:

- *Reference classification.* First of all, Robin's method assumes (in addition to the statistical background model, or H_0 hypothesis) a statistical image model that the time series follows in the regions where no change occurs (H_1 hypothesis), which is not feasible in stereo matching. This image model comes in the form of a reference classification, with respect to which the time series is assumed to follow a "piecewise constant + noise" model, where the noise level σ_b is assumed to be relatively weak with respect to the gray-level range σ_I of the images. In addition the reference is assumed to be known at a higher resolution than the time series, but this does not seem to be a crucial requirement.

- *Region shape and size.* Unlike our method which finds matches between windows of a fixed shape and size, Robin's method does not require the sizes or shapes of the regions to be specified in advance, which makes it very appealing at first sight. Yet, it has been conceived for detecting relatively large regions (comprising the majority of the pixels in an image) where no change occurs. In stereo matching we are in the opposite situation because the set of zero-disparity pixels is in principle a very small fraction of the image domain, especially if highly sub-pixel accuracy is required. Figure 1 in [Robin et al., 2009] seems to suggest that no meaningful detection is possible for relatively small regions, unless the contrast level $c = \sigma_I/\sigma_b$ is very high.

- *Background model.* Robin's method uses a very simple background model, which consists of assuming the gray-levels in all pixels independent and identically distributed (zero-mean Gaussians with variance σ_I^2). In the context of change detection, such a simple background model seems to be useful thanks to the use of a statistical image model in the form of a reference classification. However, when no image model can be assumed, many false matches would appear if we based the detection thresholds for similarity between square windows on such a naive model. A more elaborate background model that reflects more closely natural image statistics is required for block-matching.

Née's statistical tests for region similarity. In [Née et al., 2008] an *a contrario* method for detecting similar regions between two images was presented. However, their method is a classic statistical test, rather than an *a contrario* detection method in the sense of [Desolneux et al., 2007]. Indeed:

- the role of the background model (H_0 hypothesis) and the structure to be tested (H_1 hypothesis) are inverted with respect to computational Gestalt theory; and

- the significance level of the statistical test is set to $\alpha \approx 0.1$ in accordance with classical statistical testing, whereas in computational Gestalt theory this is usually set to a much smaller value (in the order of 10^{-6}).

In fact the method proposed by Née defines the null hypothesis as H_0 = *"the two regions are similar"*, and rejects H_0 (with probability $\alpha \leq 10\%$) when the L^2 norm between the two regions is too large. Thus, this method only controls the false negative rate ($\alpha \leq 10\%$), not the false positive rate (as in typical *a contrario* methods).
This second problem is much more difficult, because it requires defining H_1 = *"the two regions are similar"*, and searching for a suitable H_0 hypothesis. Our experiments indicate that

naively assuming one of the regions, or (as in Née's method) the region difference to be white noise leads to an extremely large false positive rate. A more elaborate null hypothesis closely approximating natural image statistics is required.

Caselles' motion estimation and validation. A more robust null hypothesis was used in [Igual et al., 2007], where gradient orientations (not gray-levels) are assumed to be independent and uniformly distributed. A more elaborate version of this algorithm learns the probability distribution of gradient orientation differences under the hypothesis that disparity (or motion) is zero, and uses this distribution as background model, but still pixels are all considered as independent under the background model. Once this background model was learnt, a given disparity (or motion model) is considered as meaningful if the number of aligned gradient orientations is sufficiently large within the tested region.
This method seems quite useful, but still has two drawbacks:

- the need for an initial over-segmentation of the gray-level image which is later refined by an *a contrario* region merging procedure; and
- a still moderate number of false positives in region matching.

To further reduce the number of false positives a more elaborate *a contrario* model is required, which more closely models natural image statistics.

Musé's shape matching. Learning a probability distribution in a high-dimensional space such as image patches is a difficult problem. As was shown in [Musé et al., 2006a] and [Cao et al., 2008] in the context of shape matching (where shapes are represented as pieces of level lines of a fixed size), high-dimensional distributions can be approximated by the tensor product of correctly chosen marginal distributions. Such marginal laws being one-dimensional are much easier to learn. In [Musé et al., 2003] the orientations along which marginal distributions are learnt are chosen to be the principal components of the whole learning set.
In the present work we shall adapt [Musé et al., 2003] (which was formulated for curve matching) to the context of block-matching.

Burrus' ***a contrario*** **simulations.** [Burrus et al., 2009] proposed an alternative way of choosing detection thresholds in such a way that the number of false detections under a given background model is warranted to stay below a given threshold. The procedure does not require analytical computations or decomposing the probability as a tensor product of marginal distributions. Instead, detection thresholds are learnt by Monte-Carlo simulations in a way that ensures the target NFA rate. Their method, that was developed in the context of image segmentation, involves the definition of a set of thresholds to determine whether two neighboring regions are similar or not. However, as in [Née et al., 2008], the detected event whose false positive rate is controlled is *"the two regions are different"*, and not the one we are interested in in the case of region matching, namely *"the two regions are similar"*.
No obvious way is presented in Burrus' paper that suggests how to extend his technique to the case of region matching.

Among influential related works, we must mention [Lowe, 2004] who presented a method for extracting distinctive invariant features from images that can be matched to different views of an object or scene (the SIFT-features). The SIFT method includes a rejection threshold that is empirical but universal. A match between two descriptors S_1 and S_1' is rejected if the second

closest match S_2' to S_1 is actually almost as close to S_1 as S_2' is. The typical distance ratio rejection threshold is 0.8, which means that S_2 is accepted if $dist(S_1', S_1) \leq 0.8 \times dist(S_2', S_1)$ and rejected otherwise. Interestingly, Lowe justifies this threshold by a probabilistic argument: if the second best match is almost as good as the first, this only means that both matches are likely to occur casually. Thus, they are rejected. Recently, [Rabin et al., 2008] improved the SIFT detector by rejecting SIFT matches with an *a contrario* methodology involving the Earth mover distance. The *a contrario* methodology has also already been used in stereo matching. [Moisan and Stival, 2004] proposed a probabilistic criterion to detect a rigid motion between two point sets taken from a stereo pair, and to estimate the fundamental matrix. This method, ORSA, shows much improved robustness with respect to RANSAC. In the context of foreground detection in video.

[Mittal and Paragios, 2004] proposed an *a contrario* method for discriminating foreground from background pixels, that was later refined by [Patwardhan et al., 2008]. Even though this problem has some points in common with stereo matching, it is in a way less strict, since it only needs to learn to discriminate two classes of pixels. Hence they do not need to resort to image blocks, but rely only on a 5 dimensional feature vector composed of the color and motion vector of each pixel. In conclusion, the *a contrario* methodology is expanding to many matching decision rules, but does not seem to have been previously applied to block matching algorithms.

We shall now proceed to describe the *a contrario* or background model for block-matching. The model is the simplest that work, but the reader may wonder if a still simpler model could actually work. Appendix A analyses a list of simpler proposals, and explains why they must be discarded.

Plan

This chapter is organized as follows: Section 2.2 introduces a global block model. Section 2.3 presents the *a contrario* method applied to disparity estimation in stereo pairs and treats the main problem of deciding whether two pixels match or not. Section 2.4 tackles the stroboscopic problem by adding a self-similarity threshold. Section 2.5 compares the *a contrario* and self-similarity thresholds and shows the usefulness of each other. Section 2.6 concludes the Chapter.

2.2 Block-Matching

We shall denote by $\mathbf{q}=(q_1, q_2)$ a pixel in the reference image and $B_{\mathbf{q}}$ a block centered at $\mathbf{q}$. To fix ideas, the block will be a square throughout this paper, but this is by no means a restriction. A different shape (rectangle, disk) is possible and even a variable shape. This last point is important, as several stereo algorithms try to overcome the fattening effect by adjusting the shape of a rectangular neighborhood. Given a point $\mathbf{q}$ and its block $B_{\mathbf{q}}$ in the reference image, block-matching algorithms look for a point $\mathbf{q}'$ in the second image whose block $B_{\mathbf{q}'}$ is similar to $B_{\mathbf{q}}$.

2.2.1 Principal Component Analysis

Patch comparison methods involve the quadratic distance or variants like the correlation. Since the quadratic distance can be reliably computed in a much lower dimension than the block dimension, we shall systematically reduce the block dimension by Principal Component

Analysis (PCA). The few first components in PCA represent more than 95% of the variance. We shall use these components as the more meaningful ones for a statistical match decision rule.

Let $B_{\mathbf{q}}$ be the block of a pixel $\mathbf{q}$ in the reference image and $(x_1^{\mathbf{q}}, \dots, x_s^{\mathbf{q}})$ the intensity gray levels in $B_{\mathbf{q}}$, where s is the number of pixels in $B_{\mathbf{q}}$. Let n be the number of pixels in the image. We consider the matrix $X = (x_i^j)\ 1 \leq i \leq s,\ 1 \leq j \leq n$ consisting of the set of all data vectors, one column per pixel in the image. Then, the covariance matrix $C = \mathrm{Cov}(X) = \mathbb{E}(X - \bar{x}\mathbf{1})(X - \bar{x}\mathbf{1})^T$ is computed, where $\bar{x}$ is the column vector of size $s \times 1$ storing the mean values of matrix X and $\mathbf{1} = (1, \cdots, 1)$ a row vector of size $1 \times n$. Notice that $\bar{x}$ is a block whose k-th pixel is the average of all k-th pixels of all blocks of the whole image. Thus, $\bar{x}$ is very close to a constant block, with the constant equal to the image average. The loss of the image average is of no consequence to compare blocks, since we compare two blocks by taking their difference.

The eigenvectors of the covariance matrix are called principal components and are orthogonal. Each block is projected onto the principal components in order to transform the original data to the new coordinate system.

Usually, the eigenvectors are sorted in order of decreasing eigenvalue. In that way the first principal components are the ones that contribute most to the variance of the data set. By keeping the first $N < s$ components with larger eigenvalues, the dimension is reduced but the most significant information retained. While this global ordering is used to select the main components, a local ordering will be used for the statistical matching rule. The PCA coordinates of each block will be ordered in decreasing order. In that way, comparisons of these components will be made from the most meaningful to the least meaningful one for this particular block.

Some details about the generalization to color images are given in Appendix C.

2.2.2 A Similarity Measure

Let $\mathbf{q}$ be a point in the reference image I. We look for a pixel $\mathbf{q}'$ in the secondary image I' such that $B_{\mathbf{q}}$ and $B_{\mathbf{q}'}$ are similar. Each block is a square centered at the pixel of interest and is represented by N ordered coefficients $(c_{\sigma_{\mathbf{q}}(1)}(\mathbf{q}), \dots, c_{\sigma_{\mathbf{q}}(N)}(\mathbf{q}))$. Let $c_i(\mathbf{q})$ be the resulting coefficient after projecting $B_{\mathbf{q}}$ onto the principal component $i \in \{1, \dots, s\}$ and $\sigma_{\mathbf{q}}$ the permutation representing the final order when ordering the absolute values of components for this particular $\mathbf{q}$ in non-increasing order. Note that $\sigma_{\mathbf{q}}(1) = 1$ for all $\mathbf{q}$. By a slight abuse of notation, in the following, we will write $c_i(\mathbf{q})$ instead of $c_{\sigma_{\mathbf{q}}(i)}(\mathbf{q})$ knowing that it represents the order of the best principal components.

The ordering of the principal components for each block is made by the absolute value of them. Notice that the first component has a quite different histogram than the other ones (Fig. 2.5), because it intuitively computes a mean value of the block. Indeed, the barycenter of all blocks is roughly a constant block whose average grey value is the image average grey level. The set of blocks is elongated in the direction of the average grey level and, therefore, the first component computes roughly an average grey level of the block. This explains why the first component histogram is similar to the image histogram.

2.3 The *A Contrario* Model for Image Blocks

Definition 1 (empirical probability) *Let $B_{\mathbf{q}}$ be a block in I. We call empirical probability that an observed block $B_{\mathbf{q}'}$ in I' be similar to $B_{\mathbf{q}}$ for the feature i,*

$$\widehat{p^i}_{\mathbf{q}\,\mathbf{q}'} = \begin{cases} H_i(\mathbf{q}') & \text{if} \quad H_i(\mathbf{q}) < |H_i(\mathbf{q}) - H_i(\mathbf{q}')| \\ 1 - H_i(\mathbf{q}') & \text{if} \quad 1 - H_i(\mathbf{q}) < |H_i(\mathbf{q}) - H_i(\mathbf{q}')| \\ 2 \cdot |H_i(\mathbf{q}) - H_i(\mathbf{q}')| & \text{otherwise} \end{cases}$$

where $H_i(\mathbf{q}) := H_i(c_i(\mathbf{q}))$ is the normalized cumulative histogram of $c_i(\mathbf{q})$ for the secondary image.

Fig. 2.2 illustrates how the empirical probability is computed.

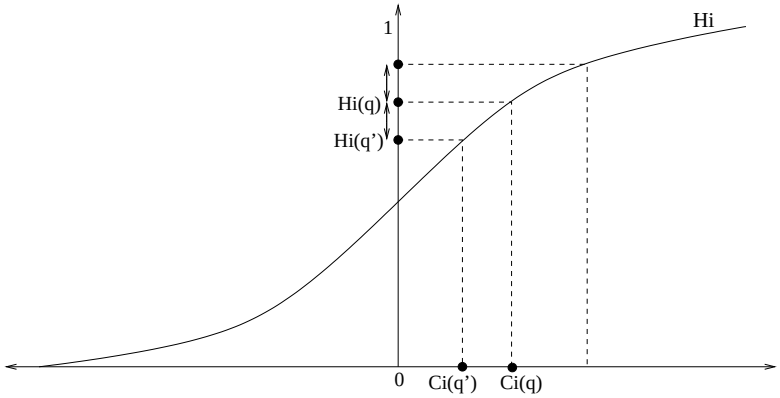


Figure 2.2: Normalized cumulative histogram of i-th PCA coordinates of the secondary image. $c_i(\mathbf{q})$ is the i-th PCA coordinate value in the first image. The empirical probability is twice the distance $|H_i(\mathbf{q}) - H_i(\mathbf{q}')|$ when $H_i(\mathbf{q})$ is not too close to the values 0 or 1.

The first principal components contain the more relevant information in the block. Thus, if two blocks are not similar for one of the first components, they should not be matched, even if their next components are similar. Due to this fact, the components will be compared with a nondecreasing exigency level. Furthermore, in our model, the number of tested correspondences shall be computed. In consequence, a quantized definition of the empirical probabilities is needed to limit the number of tests.

The last two remarks lead us to define the quantized probability as the smallest non-decreasing upper bound of $\widehat{p^i}_{\mathbf{q}\,\mathbf{q}'}$.

Definition 2 (quantized probability) *Let $B_{\mathbf{q}}$ be a block in I. Let $\Pi := \{\pi_j = 1/2^{j-1}\}_{j=1,\dots,Q}$ be a set of probability thresholds and let*

$$\Upsilon := \{\, p = (p_1, \dots, p_N) \mid p_i \in \Pi, \quad p_i \leqslant p_j \text{ if } i < j \},$$

be the family of non-decreasing N-tuples in Π^N.

The quantized empirical probability that $B_{\mathbf{q}'}$ be similar to $B_{\mathbf{q}}$ for the feature i, is defined as

$$p^i_{\mathbf{q}\,\mathbf{q}'} = \inf_{t\in\Pi}\{t \geqslant \sup_{j\leqslant i}(\widehat{p^j}_{\mathbf{q}\,\mathbf{q}'})\}\,.$$

Notice that $(p^1_{\mathbf{q}\,\mathbf{q}'},\ldots,p^N_{\mathbf{q}\,\mathbf{q}'}) \in \Upsilon$. Put another way the quantized probability vector $(p^1_{\mathbf{q}\,\mathbf{q}'},\ldots,p^N_{\mathbf{q}\,\mathbf{q}'})$ is the smallest upper bound of the empirical probabilities $(\widehat{p^1}_{\mathbf{q}\,\mathbf{q}'},\ldots,\widehat{p^N}_{\mathbf{q}\,\mathbf{q}'})$ that can be found in Υ.

Definition 3 (*a contrario* model) *We call* a contrario model *associated with a reference image a vectorial random field defined on the image domain, with values in* $\mathbb{R}^N$, $\mathbf{c}(\mathbf{q}) = (\mathbf{c}_1(\mathbf{q}),\ldots,\mathbf{c}_N(\mathbf{q}))$ *such that*

- *for each* $\mathbf{q}\in I$, *the components* $\mathbf{c}_i(\mathbf{q})$, $i=1,\ldots,N$ *are independent random variables;*
- *for each* i, *the law of* $\mathbf{c}_i(\mathbf{q})$ *is the empirical histogram of* $c_i(\cdot)$ *for the reference image.*

The *a contrario* model will be essentially used for computing a block resemblance probability as the product of the marginal resemblance probabilities of the $\mathbf{c}_i(\mathbf{q})$ in the *a contrario* model. This requires the independence of $\mathbf{c}_i(\mathbf{q})$ and $\mathbf{c}_j(\mathbf{q})$ for $i\neq j$. There is a strong adequacy to the empirical model, since the PCA transform ensures that $\mathbf{c}_i(\mathbf{q})$ and $\mathbf{c}_j(\mathbf{q})$ are decorrelated for $i\neq j$, a first approximation of the independence requirement.

Definition 4 (Number of false alarms) *Let* $B_{\mathbf{q}}\in I$ *and* $B_{\mathbf{q}'}\in I'$ *be two observed blocks. We define the Number of False Alarms of the event* "a random block $\mathbf{B}_{\mathbf{q}'}$ is as similar to $B_{\mathbf{q}}$ as $B_{\mathbf{q}'}$ is" *by*

$$NFA(B_{\mathbf{q}},B_{\mathbf{q}'}) = N_{test}\cdot Pr_{\mathbf{q}\,\mathbf{q}'},$$

where N_{test} *is the number of tested matches and* $Pr_{\mathbf{q}\,\mathbf{q}'}$ *the probability that* $\mathbf{B}_{\mathbf{q}'}$ *be as similar to* $B_{\mathbf{q}}$ *under the a contrario model as observed for* $B_{\mathbf{q}'}$.

We will write $NFA_{\mathbf{q}\,\mathbf{q}'}$ instead of $NFA(B_{\mathbf{q}},B_{\mathbf{q}'})$. Since by Def. 3, the principal components are independent under the *a contrario* model, the probability that $\mathbf{B}_{\mathbf{q}'}$ is that similar to $B_{\mathbf{q}}$ is equal $Pr_{\mathbf{q}\,\mathbf{q}'} = \prod_{i=1}^{N} p^i_{\mathbf{q}\,\mathbf{q}'}$. Therefore,

$$NFA_{\mathbf{q}\,\mathbf{q}'} = N_{test}\cdot\prod_{i=1}^{N} p^i_{\mathbf{q}\,\mathbf{q}'}\ .$$

Figure 2.3 illustrates the empirical and quantized probabilities in two cases.

Definition 5 (ϵ-meaningful match) *A pair of pixels* $\mathbf{q}$ *and* $\mathbf{q}'$ *in a stereo pair of images is an* ϵ*-meaningful match if*

$$NFA_{\mathbf{q}\,\mathbf{q}'}\leqslant\epsilon\ .$$

The last definition gives a tool to decide whether a match is meaningful or not. The NFA of a match actually gives a security level: the smaller the NFA, the more meaningful the match. The ϵ parameter can be fixed once and for ever to $\epsilon=1$ since the dependency on ϵ varies very slowly. Then, this decision rule can be seen as a parameterless method.

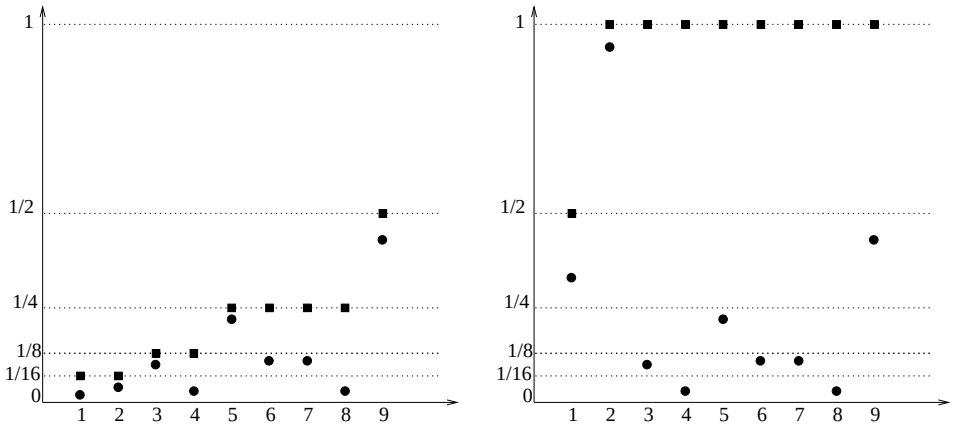


Figure 2.3: Two examples of probabilities with $Q = 5$ and $N = 9$. The probability thresholds are in abscissa and the features in ordinate. The empirical probabilities are represented with small circles and quantized probabilities with small squares. The example on the left has a final probability of $1/(16^2 \cdot 8^2 \cdot 4^4 \cdot 2)$. The right example has the same empirical probabilities excepting for features 1 and 2, but the final probability is $1/2$. Only the configuration on the left corresponds to a meaningful match.

2.3.1 Computing the Number of Tests

The number of performed tests for comparing all the blocks is the product of three terms. The first one is the image size $\#I$. The second one is the size of the search region which we denote by $S' \subset I'$. We mentioned before that the search is done on the epipolar line. In practice, a segment of this line is enough. If $\mathbf{q} = (q_1, q_2)$ is the point of reference we look for $\mathbf{q}' = (q'_1, q_2) \in I'$ such that $q'_1 \in [q_1 - R, q_1 + R]$ where R is a fixed integer larger than the maximal possible disparity. The third and most important factor is the number of tested non-decreasing probability distributions $FC_{N,Q}$. This number is a function of the number N of principal components and on the number Q of probability quanta and thus we have

$$N_{test} = \#I \cdot \#S' \cdot FC_{N,Q} \ = n\,(2R+1)\,\#\Upsilon.$$

Lemma 1 *With the above notations,*

$$FC_{N,Q} = \sum_{t=0}^{Q} (t+1) \cdot \binom{N+Q-t-3}{Q-t-1}.$$

We write formally

$$FC_{N,Q} = \#\{\, f : [1,N] \to [1,Q] \mid f(x) \leqslant f(y),\, \forall x \leq y \,\}$$
$$\overline{FC}_{N,Q} = \#\{\, f : [1,N] \to [1,Q] \mid f(1) = 1,\ f(N) = Q; \quad f(x) \leqslant f(y),\, \forall x \leqslant y \,\}.$$

Since $FC_{N,Q} = \sum_{t=0}^{Q} (t+1)\overline{FC}_{N,Q-t}$ and $\overline{FC}_{N,Q} = \binom{N+Q-3}{Q-1}$ the lemma is obvious.

Proposition 1 *Let* $\Gamma = \Sigma_{\mathbf{q},\mathbf{q}'}\chi_{B_{\mathbf{q}},\mathbf{B}_{\mathbf{q}'}}$ *be the random variable representing the number of occurrences of an* ϵ*-meaningful match between a deterministic patch in the first image and a random patch in the second image. Then the expectation of* Γ *is smaller than* ϵ*.*

Proof We have

$$\chi_{B_{\mathbf{q}},\mathbf{B}_{\mathbf{q}'}} = \begin{cases} 1, & \text{if } NFA(B_{\mathbf{q}},\mathbf{B}_{\mathbf{q}'}) \leqslant \epsilon; \\ 0, & \text{if } NFA(B_{\mathbf{q}},\mathbf{B}_{\mathbf{q}'}) > \epsilon. \end{cases}$$

Then, by the linearity of the expectation

$$\mathbb{E}[\Gamma] = \sum_{\mathbf{q},\mathbf{q}'} \mathbb{E}[\chi_{\mathbf{q},\mathbf{q}'}] = \sum_{\mathbf{q},\mathbf{q}'} \mathbb{P}\left[NFA(B_{\mathbf{q}},\mathbf{B}_{\mathbf{q}'}) \leqslant \epsilon\right].$$

The probability inside the expectation can be computed using definitions 4 and 1 as follows

$$\mathbb{E}[\chi_{\mathbf{q},\mathbf{q}'}] := \mathbb{P}\left[NFA(B_{\mathbf{q}},\mathbf{B}_{\mathbf{q}'}) \leqslant \epsilon\right] = \mathbb{P}\left[\prod_i^N p^i(B_{\mathbf{q}},\mathbf{B}_{\mathbf{q}'}) \leqslant \frac{\epsilon}{N_{test}}\right].$$

The probability of the non-disjoint union of events can be upper-bounded by their probability sum, and the intersection below involves only independent events according to our background model. Thus:

$$\begin{aligned}
&\mathbb{E}[\chi_{\mathbf{q},\mathbf{q}'}] = \\
&= \mathbb{P}\left[\bigcup_{\substack{p\in\Upsilon \\ \prod_i p_i \leqslant \epsilon/N_{test}}} \bigcap_i \left(2\cdot|H_i(c_i(\mathbf{q})) - H_i(\mathbf{c}_i(\mathbf{q}'))| \leqslant p_i\right)\right] \\
&\leqslant \sum_{\substack{p\in\Upsilon \\ \prod_i p_i \leqslant \epsilon/N_{test}}} \prod_i \mathbb{P}\left[2\cdot|H_i(c_i(\mathbf{q})) - H_i(\mathbf{c}_i(\mathbf{q}'))| \leqslant p_i\right] \\
&= \sum_{\substack{p\in\Upsilon \\ \prod_i p_i \leqslant \epsilon/N_{test}}} \prod_i p_i \leqslant \frac{\epsilon}{\#I\,\#S'}.
\end{aligned}$$

In the last line we used the fact that $H_i(\mathbf{c}_i(\mathbf{q}'))$ follows a Uni$[0,1]$ distribution, since the random variable $\mathbf{c}_i(\mathbf{q}')$ is drawn from the cumulative distribution H_i. Finally, recalling that $N_{tests} = \#I\,\#S'\,\#FC_{N,Q}$, this last sum can be upper bounded by $\frac{\epsilon}{\#I\,\#S'}$. So we have shown that

$$\mathbb{E}[\Gamma] = \sum_{\mathbf{q},\mathbf{q}'} \mathbb{E}\left[\chi_{B_{\mathbf{q}},\mathbf{B}_{\mathbf{q}'}}\right] \leqslant \sum_{\mathbf{q},\mathbf{q}'} \frac{\epsilon}{\#I\,\#S'} = \epsilon.$$

2.3.2 Local PCA

The lack of information in some regions of an image is a problem in the correspondence process. Blocks situated in flat zones with poor texture, for examples shadows, are difficult to match. Obviously, the matching thresholds must be adapted to the characteristics of each region. This means that we cannot be satisfied with a global *a contrario* model, but must adapt it to each region. This is the goal of the present section. If a pair of blocks are matched,

the mean and variance of the gray levels in the blocks should be similar. Hence, it is reasonable to make a rough regional partition based on block mean and variance only.

Let I be the set of pixels of the reference image. Denote for each $\mathbf{q}$ in I by $m_{\mathbf{q}}$ the mean and by $v_{\mathbf{q}}$ the variance of $B_{\mathbf{q}}$. Let $\tilde{m}_1 \leq \tilde{m}_2 \leq \ldots \leq \tilde{m}_n$ be the ordered mean values. We partition image pixels in T classes depending on the mean and variance of its corresponding blocks.

Definition 6 *Given a fixed $\alpha \in (0, 0.4)$, the j-th region of the partition is defined as*

$$I_j^M = \{\, q \in I \mid h_m\big((j-1)\frac{n}{T} - \alpha n\big) \leq m_{\mathbf{q}} \leq h_m\big(j\frac{n}{T} + \alpha n\big)\} \,,$$

where T is the number of regions and h_m is the function defined as:

$$h_m(t) = \begin{cases} \tilde{m}_1 & \text{if } t \leq 0 \\ \tilde{m}_{[t]} & \text{if } 0 < t < n \\ \tilde{m}_n & \text{if } t \geq n \end{cases}$$

Thus, $I = \bigcup_{j=1,\cdots,T} I_j^M$. Note that it is a non-disjoint partition of I since $\#\{I_j^M \cap I_{j+1}^M\} = 2\alpha n$. The smaller α is, the less pixels are in the intersection. Given α, T, and the sorted variance values $\tilde{v}_1 \leq \tilde{v}_2 \leq \ldots \leq \tilde{v}_n$ a variance partition of I can be defined $I = \bigcup_{i=j,\cdots,T} I_j^V$ like the mean partition.

Definition 7 *Given the mean partition and variance partition of I with a fixed α, the previous coarse partition of I can be defined as:*

$$I = \bigcup_{j,k=1,\cdots,T} G_{j\,k} \,,$$

where $G_{j\,k} = I_j^M \cap I_k^V$.

Given a couple of stereo images, the previous coarse partition of each image is computed:

$$I = \bigcup_{j,k} G_{j\,k} \,, \qquad I' = \bigcup_{j,k} G'_{j\,k}.$$

The aim of this partition is not to do a fine classification of the pixels, so fixing $T = 2$ (which means 4 regions in the coarse partition) will be enough for our purpose. Then, for each pixel in the reference image its matching pixel is searched for in the secondary image among the pixels in the same class. For example, a pixel in a shadow belongs to the region with low mean and variance. Its matching pixel is searched for in the region of the secondary image with low mean and variance. Whenever a pixel belongs to more than one region, the search is done in each region independently. The match is accepted if the candidates in every region coincide.

Fig. 2.4 shows the first principal components for each region and the last two columns of Fig. 2.5 show the histograms of each partition and the histograms of the coefficients with local PCA. Fig. 2.6 shows several blocks of the reference image for two different classes and random blocks following the corresponding law. It permits to assess how faithful the background model is to the original.

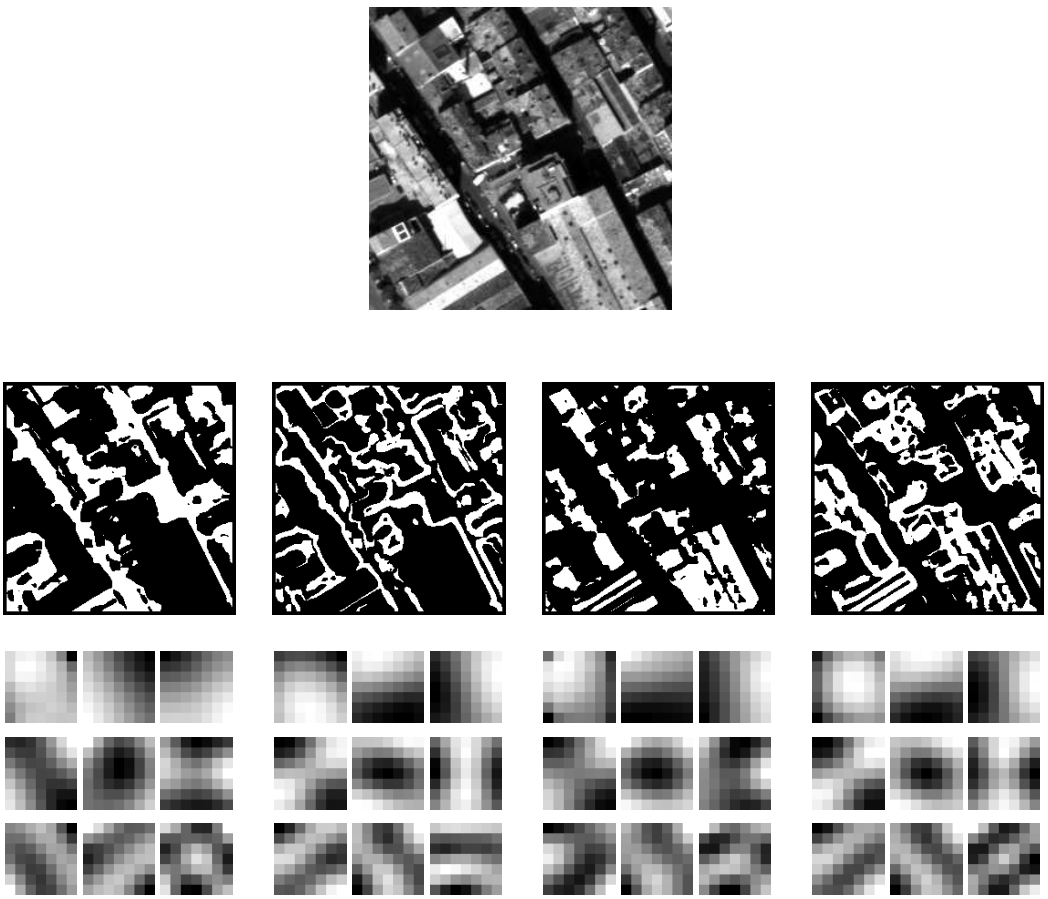


Figure 2.4: Top: Reference image of the stereo pair of images. Bottom: partition of the image in four classes and the respective nine first principal components. In each case, the PCA training is done on the white pixels. From left to right: low mean and variance, low mean and large variance, large mean and low variance and finally, large mean and variance.

Local PCA changes the considered number of pixels, since only one region is analyzed at the same time. Hence, every region will have a different number of tests. If a pixel $\mathbf{q}$ in the reference image has to be matched, the number of tests is:

$$N_{test} = \#G_{j\,k} \cdot \#S'_{j\,k} \cdot FC_{N,Q} \cdot T^2 \,,$$

where $G_{j\,k} \in I$ is the region of the coarse partition to which $\mathbf{q}$ belongs and $S'_{j\,k} = S' \cap G'_{j\,k} \in I'$.

2.3.3 Search of Meaningful Matches

In the following we give more details about the PCA training and the search of meaningful matches.

The computed coefficients c_i encode the information of each pixel in the regions of interests ($G_{j\,k}$ and $G'_{j\,k}$) and are the features used in the search step. The coefficients are computed independently in each region from a different basis.

For each pixel $\mathbf{q}=(q_1, q_2)$ in the region of interest $G_{j\,k}$ the quantized probabilities $p^i_{\mathbf{q}\,\mathbf{q}'}$ are computed for each feature i and $\mathbf{q}'$ in the epipolar segment $S' \cap G'_{j,k}$ of the second image

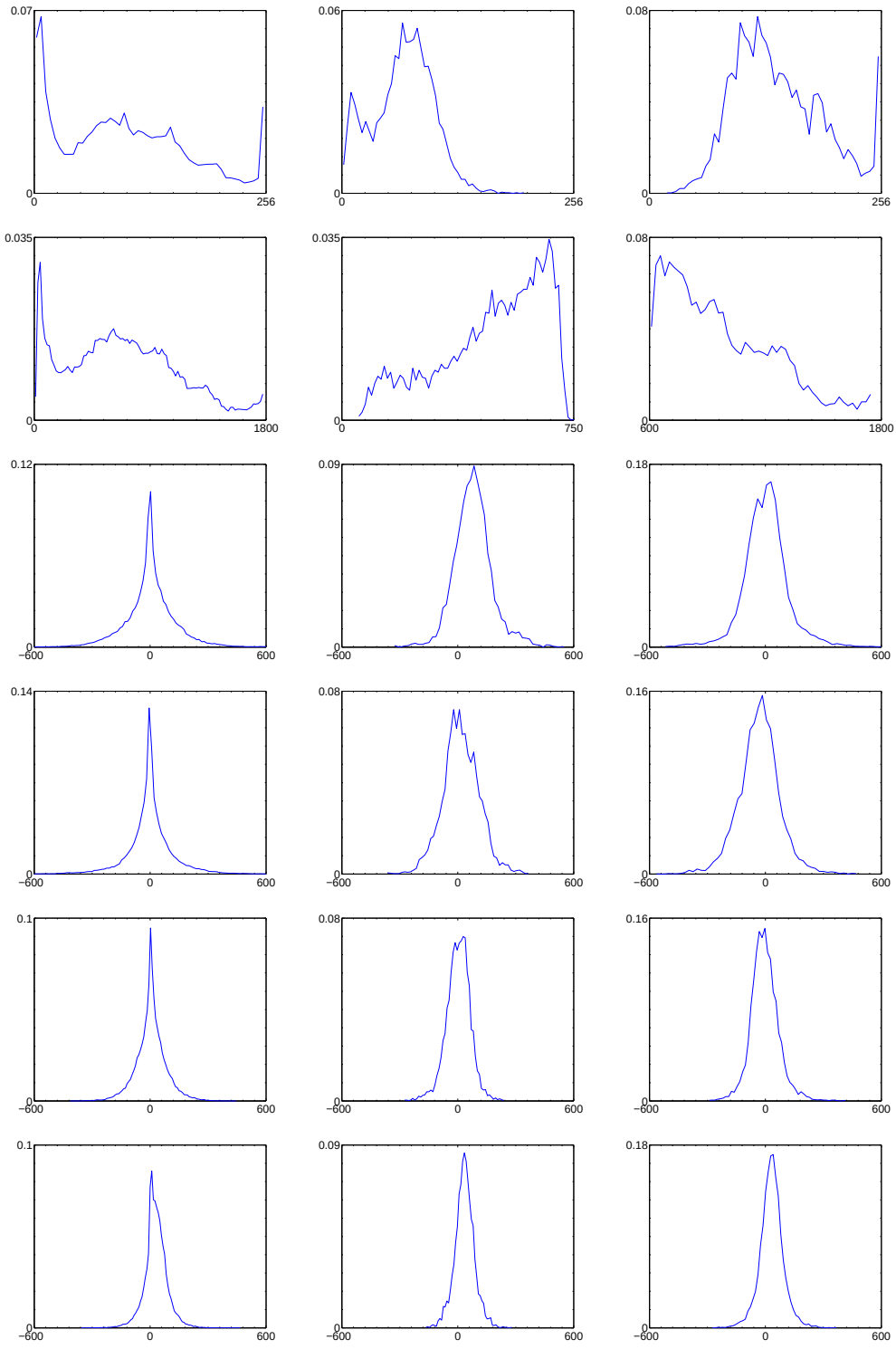


Figure 2.5: First column: Histograms for the global PCA. Second column: Histograms for the region of low mean and large variance of the local PCA. Third column: Histograms for the region of large mean and large variance of the local PCA. From top to bottom: Histogram of the reference image (histogram of the concerned region for the local PCA) and histograms of the coefficients for the first five principal components.

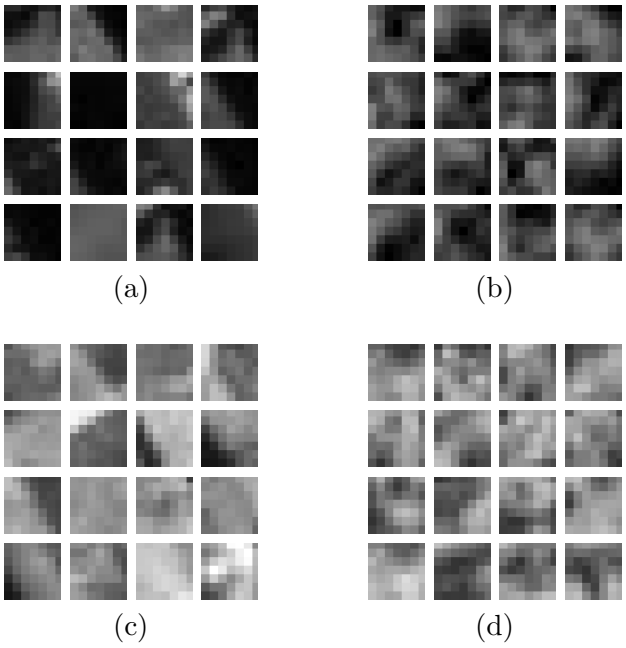


Figure 2.6: (a) Patches of the reference image of the class of low mean and large variance. (b) Random blocks following the law of the reference image of the class of low mean and large variance. (c) Patches of the reference image of the class of large mean and large variance. (d) Random blocks following the law of the reference image of the class of large mean and large variance.

where $S' = \{(q'_1, q'_2) \in I' \,|\, q_2 = q'_2,\ q'_1 \in [q_1 - R, q_1 + R]\}$. $p^i_{\mathbf{q}\,\mathbf{q}'}$ is the probability that the block centered on $\mathbf{q}$ $B_{\mathbf{q}}$ is similar to another block $B_{\mathbf{q}'}$ for the feature i.

For complexity reasons, prior to the probability computation, the coefficients of the second image I' should be sorted. More precisely, for each feature $i = 1, \cdots, s$ and each class $G'_{j\,k}$ of the partition of I' the coefficients are ordered $(o_i^{j\,k}(\mathbf{q}))_{i,\mathbf{q}}$. Then, for each feature i, to find the position of $c_i(\mathbf{q})$ in the sorted coefficients $o_i^{j\,k}(\cdot)$, a dyadic search is done.

2.4 The Self-Similarity Threshold

Quite often in urban scenes, a local structure (like windows on a roof) is repeated over and over again with little or no difference from one instance to another. Since, in general, the number of repetitions is insignificant with respect to the number of blocks that have been used to estimate the empirical *a contrario* probability distributions, the *a contrario* model does not learn this repetition, and can be fooled by such repetitions, thus signaling a significant match for each repetition of the same structure. Of course, one of those significant matches is the correct one, but chances are that the correct one is not the most significant one. In such a situation two choices are left: *(i)* trying to match the whole set of self-similar blocks of I as a single multi-block (typically, global methods such as graph-cuts do that implicitly); or *(ii)* remove any (probably wrong) response in the case where the stroboscopic effect is detected. The first alternative would lead to errors anyway, if the similar blocks have not the same height. This occurs in urban scenes where similar roofs can have different heights. Fortunately, stereo pair block-matching yields a straightforward adaptive threshold. A distance function d between blocks being defined, let $\mathbf{q}$ and $\mathbf{q}'$ be points in the reference and secondary images respectively that are candidates to match with each other. The match of $\mathbf{q}$ and $\mathbf{q}'$ will be accepted if the following conditions are satisfied:

- $(\mathbf{q}, \mathbf{q}')$ is a meaningful match;
- $d(B_{\mathbf{q}}, B_{\mathbf{q}'}) < min\{d(B_{\mathbf{q}}, B_{\mathbf{r}}) |\ r \in I \cap S(\mathbf{q})\}$,

where $S(\mathbf{q}) = [q_1 - R\,,\, q_1 + R] \setminus \{q_1,\, q_1 + 1,\, q_1 - 1\}$ and R is the search range. As noted earlier, the search for correspondences can be restricted to the epipolar line. This is why the automatic threshold is restricted to $S(\mathbf{q})$.

Computing the similarity of matches in one of the images is not a new idea in stereovision. In [Manduchi and Tomasi, 1999] the authors define the *distinctiveness* of an image point x as the perceptual distance to the most similar other point in the search window. In particular, they study the case of the auto-SSD function (Sum of Squared Differences computed in the same image). The flatness of the function contains the expected match accuracy and the height of the smallest minimum of the auto-SSD function beside the one in the origin gives the risk of mismatch. They are able to match correctly ambiguous points by matching intrinsic curves [Tomasi and Manduchi, 1998b]. However, the proposed algorithm only accepts matches when its quality is above a certain threshold. The obtained disparity maps are rather sparse and the accepted matches are completely concentrated on the edges of the image. Even if the SNR (Signal To Noise Ratio) in these pixels of the image is higher than the others, the accuracy of such disparities may be very low because of the fattening phenomenon affecting all block-matching methods and occlusions.

As [Sara, 2002], we think that ambiguous correspondences should be rejected. In this work a new *stability property* is defined as a condition a set of matches must satisfy to be

considered unambiguous at a given confidence level. The stability constraint and the tuning of two parameters allows to take care of flat or periodic autocorrelation functions. The comparison of this last algorithm with ours results will be done in section 3.5.

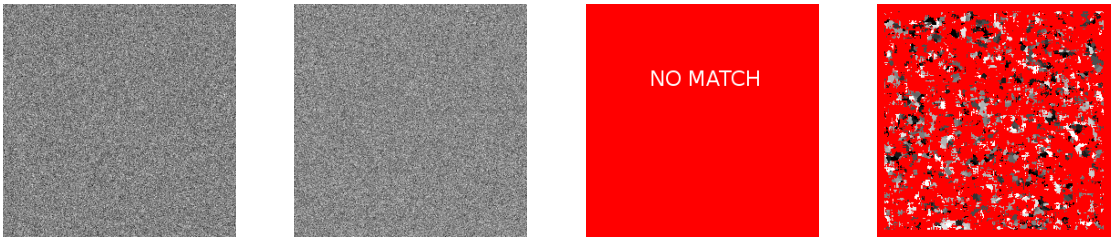
2.5 *A Contrario* vs Self-Similarity

The usefulness of the Self-Similarity (SS) threshold may be hazy when it is combined to the a contrario framework. One may be asked whether the *a contrario* decision rule to accept or reject correspondences between patches is sufficient. Likewise, we should clarify if the self-similarity threshold would be enough to reject false matches in a correlation algorithm.

In this section we are going to answer to these questions and we are going to analyze some simple examples to understand the need of both tests in our algorithm. More precisely, for each example we are going to compare the result of the a contrario test and the result of a classic correlation algorithm combined with the self-similarity threshold.

2.5.1 The Noise Case

Here, we consider two independent Gaussian noise images. It is obvious that we would like to reject any possible match between these two images. The *a contrario* test rejects all the possible patch matches as expected. On the other hand, the correlation algorithm combined with the self-similarity is not sufficient and lots of false matches are accepted.

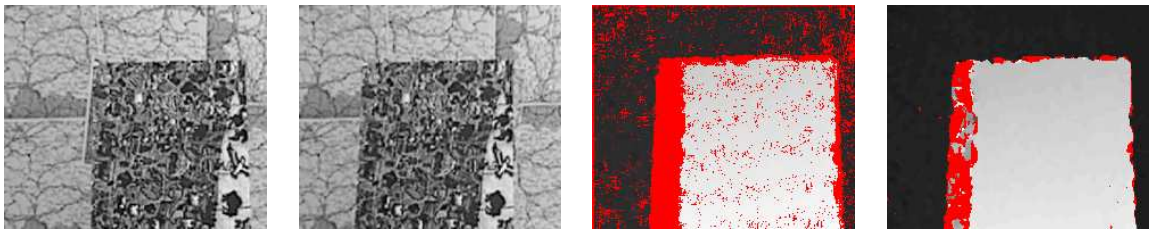


(a) (b) (c) (d)

Figure 2.7: (a) and (b) Noise images. (c) No match at all has been accepted by the *a contrario* test! (d) Many false correspondences have been accepted by the self-similarity threshold.

2.5.2 The Occlusion Case

If a point of the scene can be observed in only one of the images of the stereo pair, then an estimation of its disparity is simply impossible. The best decision is to reject any possible match. A good example to illustrate the performance of the two rejection tests AC and SS is the map image (Middlebury stereovision database) (Fig 2.8) which has a large baseline and therefore an important number of occluded pixels. As before, the computation of Number of False Alarms (NFA) gives the best result (see Table 2.1). The table indicates, however that the self-similarity test can remove a few points. Actually, even if the proportion of eliminated points is tiny, such mismatches can be very annoying and the gain is not negligible at all.



(a) (b) (c) (d)

Figure 2.8: (a) Reference image (b) Secondary image. The squared object occludes part of the background (c) The *a contrario* test does not accept any match for pixels in the occluded areas. (d) The disparity map is denser but spurious disparities remain in the occluded region with the self-similarity threshold.

	Bad matches	Total matches
SS	3.35%	85.86%
AC	0.37%	64.85%
AC+SS	0.36%	64.87%

Table 2.1: Quantitative results of the correlation algorithm with the self-similarity threshold (SS), the *a contrario* algorithm (AC) and the algorithm combining both (AC+SS). We have computed the percentage of matches for each algorithm in the whole image and among these the number of wrong matches. (A match is considered wrong if its disparity difference with the ground truth disparity is larger than one pixel).

2.5.3 Repetitive Patterns

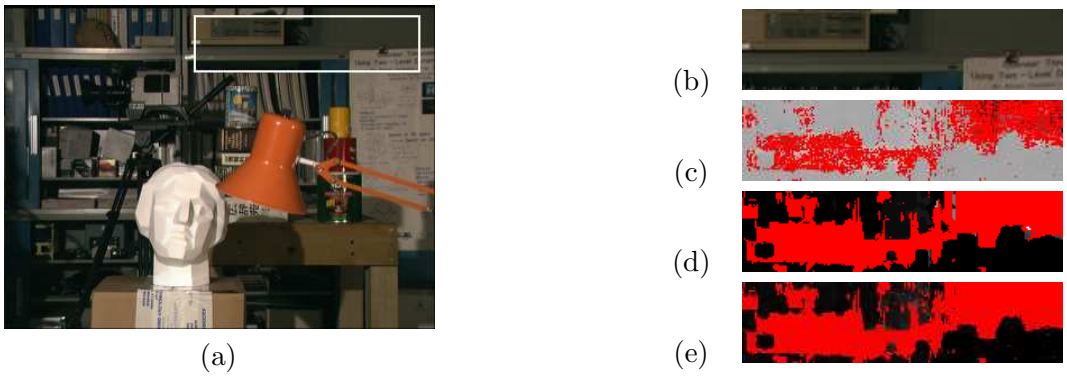
In natural images there are often repeated patterns that look locally the same. This ambiguity can lead the *a contrario* test to accept erroneous matches. In this situation, the self-similarity test is necessary. First, a synthetic case has been considered in Fig. 2.9, where the accepted correspondences are completely wrong in the *a contrario* test for the repeated lines. On the contrary, the self-similarity threshold is able to reject matches in this region of the image. Finally, the well known Tsukuba images are a real example where several patches in the image are strongly similar. In Fig. 2.10 one can compare the results of the two test, AC and SS, separately and together. The quantitative results are summarized in table 2.2 . We conclude that the *a contrario* test and the self-similarity threshold are both necessary and complementary. They will therefore always be applied in the sequel.

2.5.4 An Unsolved Case

There are a few cases where both rejection tests fail. AC + SS allowed some erroneous disparities in the Venus image (Fig. 3.5 in Chapter 3). The error appears in a planar surface with a repeated pattern (Fig. 2.11). Usually, such pixels are rejected by the self-similarity threshold. It turns out that the proposed meaningful match had a very small Number of False Alarms and was very similar in terms of quadratic distance as well. Only global optimization algorithms can resolve such cases.

(a) (b) (c)

Figure 2.9: (a) Reference image with a texture and a stripes periodic motif. The secondary image is a 2 pixels translation of the reference image. The obtained disparity map should be a constant image with value 2. (b) The *a contrario* test gives the right disparity 2 everywhere, except in the stripes region. (c) The repeated stripes are locally similar, so the self-similarity threshold rejects all the patches in this region.



(a) (b) (c) (d) (e)

Figure 2.10: (a) Tsukuba reference image. There are self-similar patches inside the white rectangle. (b) Image crop of the rectangle. The bookshelves and part of the wall look locally equal in the image. The ground truth is a constant image in this part of the image. (c) Disparity map of the cropped region with AC. There are several meaningful matches which are false correspondences. (d) Results of the correlation algorithm combined with the self-similarity threshold. Error due to the repeated texture disappear but other false correspondence are not rejected. (e) Result of our algorithm where the *a contrario* model is combined with the self-similarity threshold. In this disparity map all the accepted correspondences are correct.

	Bad matches	Total matches
SS	6.43%	73.7%
AC	5.02%	59.7%
AC+SS	4.07%	57.9%

Table 2.2: Quantitative results for the Tsukuba image. The percentage of bad matches of AC+SS remain high because no fattening correction has been performed in this study.

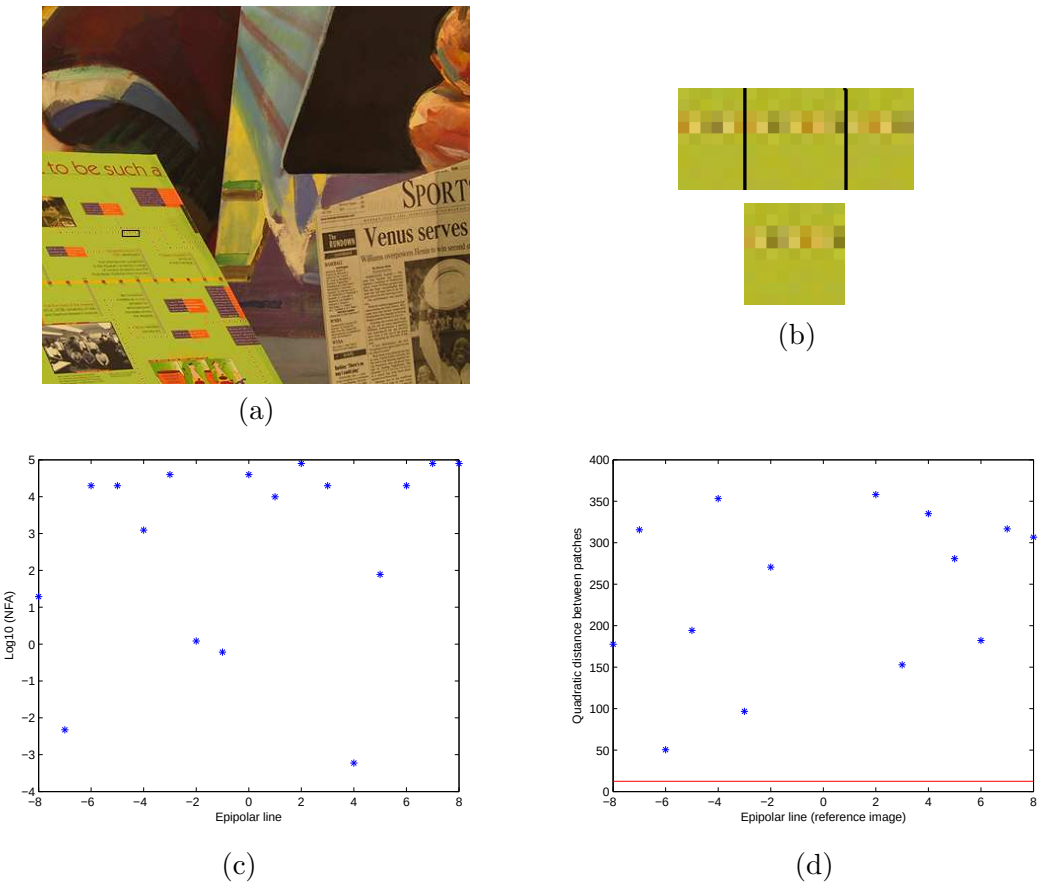


Figure 2.11: (a) Venus reference image. The region of interest has been surrounded with a black rectangle. (b) Top: 9×9 patch $B_{\mathbf{q}_0}$ of the reference image with its surrounding pixels. Bottom: accepted matching patch in the secondary image. This is a false match passing the AC and SS tests. (c) Given the fixed patch $B_{\mathbf{q}_0}$ in the reference image, the plot on graph of $log_{10}(NFA_{(\mathbf{q}_0,\mathbf{q}')})$ for each $\mathbf{q}'$ in the epipolar searching segment in the secondary image. The patch in the secondary image with a 4 pixels disparity is the one having the minimum NFA and it had been accepted by both tests. However, the correct patch is the one with -7 pixels disparity. Its associated NFA it's considerably bigger. (d) Plot of the quadratic distance between the fixed patch $B_{\mathbf{q}_0}$ with its neighboring patches. The red line is the quadratic distance of the proposed wrong match. This distance is considerably smaller than the other ones, so the self-similarity threshold accepts widely the match.

2.5.5 Application to Occlusions, Moving Objects and Poorly Textured Regions

In this section we shall see that the presented method which combines the *a contrario* test and the self-similarity threshold permits to blindly eliminate all wrong matches that might be caused by three different phenomena: occlusions, moving objects, and poorly textured regions. All of them are very common in satellite or aerial images of urban areas.

Discarding Occlusions

Clearly, disparities between two images can be only computed for the points that are visible in both images. The other ones are called occlusion points. Occlusions are always present in images pairs of scenes where the elevation map is not smooth. In urban areas, buildings cause many occlusions, whose area grows with their height.

Given that detecting occlusions is a key problem in the matching process, numerous authors have handled it. The first approach is to detect the occlusion, the second one is to reduce the sensitivity to occlusions, and the third one is to model the occlusions geometry.

Figure 2.8 is a clear example of the performance of the detector of meaningful matches presented above to discard occlusion points. The images had been taken with a large baseline, so the occlusions are more evident for this example.

Moving and Disappearing Objects and Shadows

One of the main drawbacks of block-matching algorithms is the appearance of false matches due to moving or disappearing objects. Essentially, this is the same problem as the occlusion problem but the occlusion is caused by camera motion in presence of a depth difference instead of object motion. We stress the importance of the *a contrario* model to manage moving objects such as cars or pedestrians in urban scenes. Figure 2.12 shows an aerial image where a car in the crossroad has changed its position before the second image was taken. On the right of this same image we can observe a pedestrian who has walked some meters between the two snapshots. We mostly see his/her big shadow because of the slanted position of the Sun (see red arrows). Remark than in both cases no match is present in the disparity map. In the disparity map, we can see other regions which have not been matched because the numerous shadows in this image. These regions are poor-textured regions and retain few meaningful matches.

Shadows are always present in images of urban areas. They are more or less important in the image depending of the position of the Sun. Due to the lack of texture, the matching of blocks inside shadows become ambiguous and several errors can appear. Thanks to the local PCA meaningful match method, however, some points in the shadow can be matched reliably. Changing the dynamics of the image (see Fig. 2.12-c), we realized why there are some matches inside the shadows. Notice, first, the matches due to the zebra crossing (red arrow). The ends of each band of the zebra crossing matched because they are a unique meaningful match. On the other hand, the self-similarity sanity check has rejected the rest of the points in the bands, thus avoiding any possible wrong match. Finally, in some parts of the image, we can observe that points in the street have a disparity larger than the buildings. This might seem to be an error, but it is not. Looking at Figure 2.12-c (green arrows) we have realized that there are lampposts.

Figures 2.13 and 2.14 show other examples for a stereo pair of images of the city of Marseille (France). In both cases, several cars have changed position between the two images. We see that our algorithm has not matched the points in these regions as MARC (Multiresolution Algorithm for Refined Correlation) did in this situation (see Fig. 2.14(d)). A deeper comparison between our algorithm and MARC has been done in Appendix D.

Figure 2.15 shows images from the archaeological site of Copan in Honduras. The scene does not change between the snapshots and no object has moved. However, some crosses have been added in a post-processing treatment. We have considered a smaller part of the image where there is one of these crosses and we have verified that no match had been found for pixels whose patch contained part of this cross.

In Figure 2.16 one can compare the disparity maps obtained in one of the shadows of the image with a classic PCA, and with a local PCA. Local PCA clearly gives better results. In the shadow, less points have found a match, but no errors appear in the matches. On the contrary, several mismatches appear with global PCA. Figure 2.17 shows the points in the image that has not been matched when using our local approach instead of the global one. It can be observed that no matches have been lost in textured regions and the set of lost matches are placed in non textured areas of the image where is preferable not to match these points.

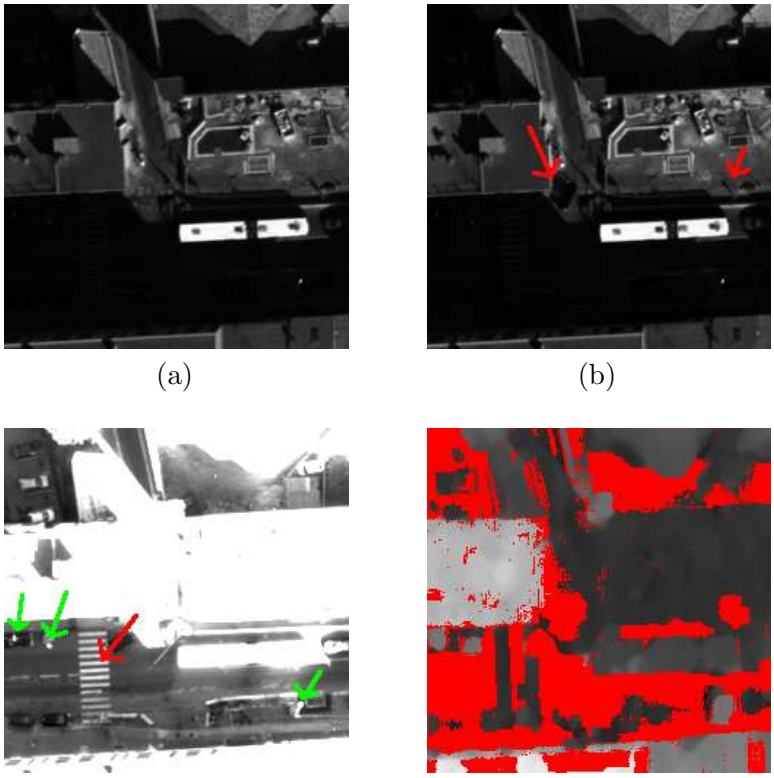
2.6 Choosing the Parameter ϵ

As we have said previously, the ϵ parameter can be fixed once and for ever to $\epsilon = 1$. The dependency of the reject decision rule (AC) to ϵ has been studied for a simulated pair of images of the St. Michel prison in Toulouse (see 2.16-(a)). Figure 2.18 shows the ROC curve where the percentage of true positives is plotted versus the percentage of false positives. We have only considered the points remaining after the self-similarity (SS) threshold and the fattening *a posteriori* correction (which will be explained in Chapter 3). This is why we have only a false positive percentage of 0.005% with $\epsilon = 1$.

Conclusion on Experiments

The *a contrario* block-matching thresholds, that were the principal object of the present chapter, combined with the self-similarity threshold is able to detect occlusion, moving objects and poor or periodic textured regions by performing a rigorous selection of meaningful, reliable matches. Wrong match thresholds and fattening are, in our opinion, the principal drawbacks for block-matching algorithms in stereovision. In Chapter 3 we deal with fattening where by adding an *a posteriori* rejection rule.

Block matching have led to the overall dominance of global methods such as graph cuts. However, it must not be forgotten that global methods have no validation procedure. The *a contrario* method must be viewed as a validation procedure, no matter what the stereo matching process was. Block matching, even with the multiple but careful thresholds established in this chapter, seems to give a fairly dense set of reliable matches. It may be objected that the obtained disparity map is not fully dense anyway. This objection is not crucial for two reasons. First, knowing which matches are reliable allows one to complete a given disparity map by fusing several stereo pairs. Since disposing of multiple observations of the same scene by several cameras and/or at several different times is by now an usual setting, it becomes more and more important to be able to fuse 3D information obtained from many stereo pairs.



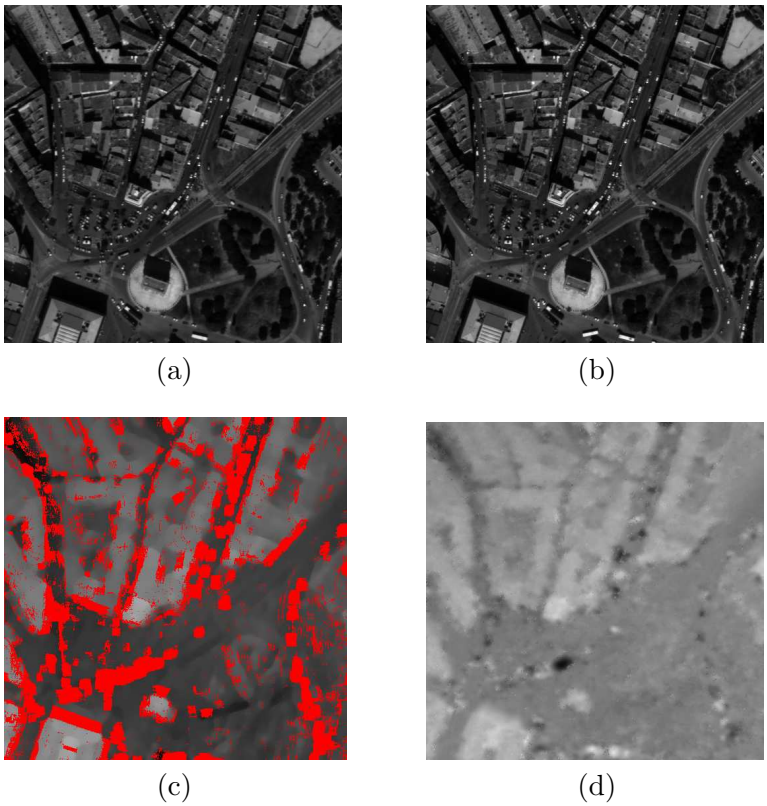
(a) (b)

(c) (d)

Figure 2.12: (a) Reference image. (b) Secondary image. (c) Reference image with a huge contrast change putting in evidence details inside the shadow. (d) Disparity map. Red points are points which haven't been matched. The car appearing only in the second image and the pedestrian who has moved into the two snapshots have not been matched.

(a) (b) (c)

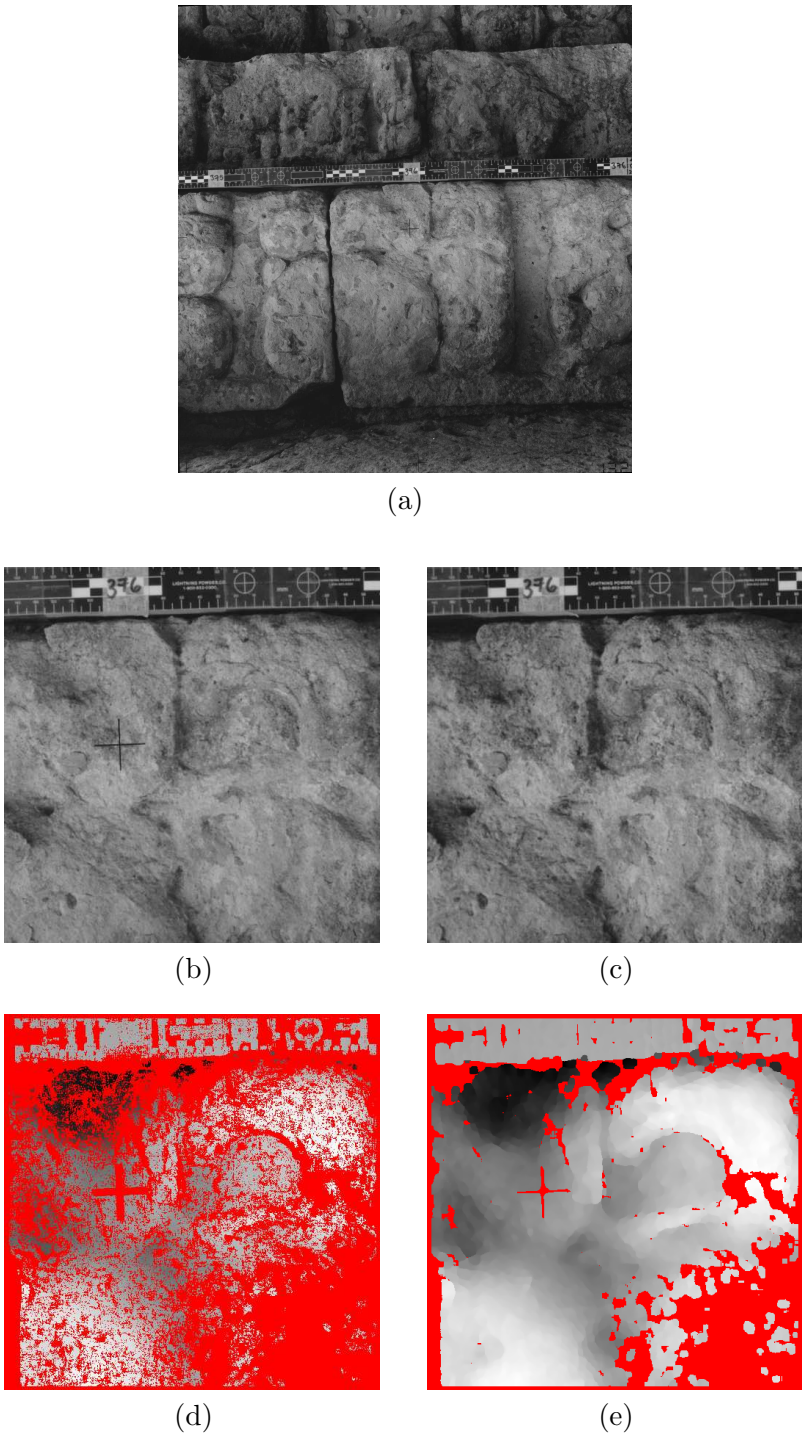
Figure 2.13: (a) Reference image. (b) Secondary image. (c) Disparity map. Red points are points which haven't been matched.



(a) (b)

(c) (d)

Figure 2.14: (a) Reference image. (b) Secondary image. (c) Disparity map. Red points are points which haven't been matched. (d) Disparity map obtained with a CNES classic multiscale block-matching algorithm (MARC) in the stereo pair of images of Marseille before any regularization.



(a)

(b)

(c)

(d)

(e)

Figure 2.15: (a) Copan archaeological ruins (Honduras). (b) Reference image. (c) Secondary image. The cross in the reference image has been added subsequently. (d) Refined disparity map. No meaningful match has been found for patches meeting the cross. Several pixels on the measure apparatus have been rejected because they are self-similar. The low density of disparities in the bottom right corner is due to the local lighting changes. (e) Disparity map after median filter.

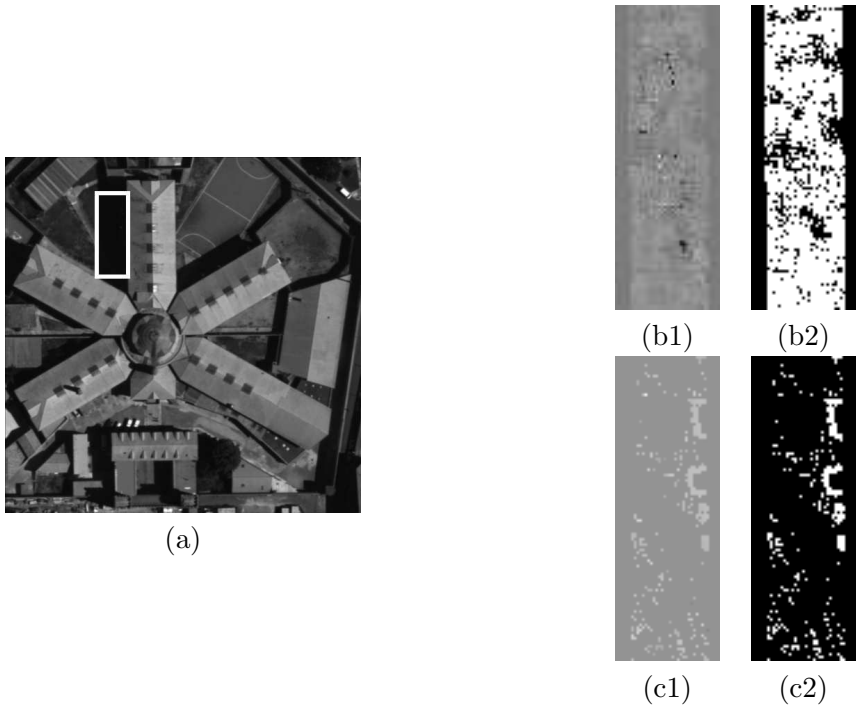


Figure 2.16: (a) Reference image of the simulated stereo pair. The rectangular box focuses on a shadowed region. (b1) Disparity map obtained in the shadow with global PCA. Several errors appear inside the shadow. Indeed pixels on the ground should have similar disparities. (b2) Valid points with global PCA. (c1) Disparity map with local PCA: Matches become coherent in the shadow. (c2) Map of valid points (in white). Many points have been discarded in the shadow as unreliable.

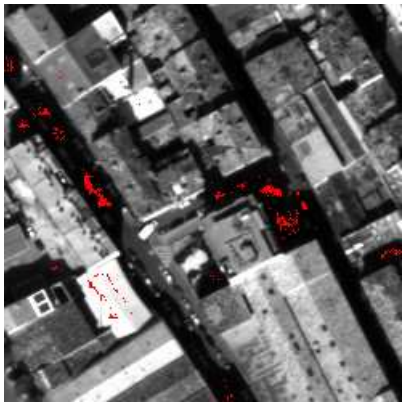


Figure 2.17: The red points are those whose match has been rejected. The majority are placed in the shadows and some of them in poorly textured regions of the roofs.

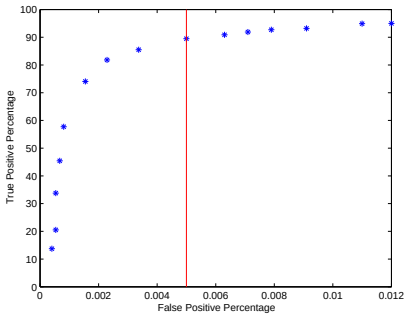


Figure 2.18: ROC curve. Percentage of true positives vs. percentage of false positives. Each plot corresponds to a different value of ϵ, from 10^{-8} to 10^{-6}. In particular, the (red) vertical line corresponds to $\epsilon = 1$, the previous plot corresponds to $\epsilon = 0.1$ and the next to $\epsilon = 10$.

Having almost only reliable matches in each pair makes the fusion job extremely easy. Second, having only validated matches permits to launch benchmarks based on precision, and to raise challenges about which precision can be ultimately attained (on *validated* matches only). The experiments performed so far show that algorithms mixing block comparison and interpolation have a poor precision performance. This precision issue is also a key to small baseline stereo maps, and small baseline stereo opens the way to obtaining dense urban maps obtained from nadir views.

Chapter 3

The Fattening Phenomenon

Contents

Résumé : La méthode de mise en correspondance stéréo présentée dans le chapitre 2 (AC + SS) souffre du phénomène d'adhérence comme toutes les autres méthodes de *block-matching*. Ce phénomène crée des erreurs de disparités à proximité des bords des objets de la scène. Cette distorsion se produit à cause de la comparaison de fenêtres et il est particulièrement aigu lorsque l'objet qui se trouve au premier plan, sur des bords (ou textures) importants par rapport à l'arrière-plan.

Pour démontrer l'efficacité de la règle de rejet AC + SS, nous avons mis en oeuvre une règle d'élimination d'adhérence (*a posteriori*). La méthode de rejet finale est testée sur des exemples de *benchmarks* classiques et sur des scènes urbaines avec faible B/H. Tous les tests confirment que l'algorithme en trois étapes proposé fournit des nappes de disparité assez denses $(40\% - 90\%)$ contenant moins de $0,4\%$ de mauvais appariements.

Abstract: The stereo matching method presented in Chapter 2 (AC+SS) as other block-matching methods suffers from the fattening phenomenon which is an error of disparities close to object borders. This distortion due to the windowing process is specially acute when foreground object have significant edges (or textures) with respect to the background.

To demonstrate the effectiveness of the AC+SS rule, we shall also implement an *a posteriori* fattening elimination rule. The final rejection method is tested on classic benchmark examples and urban aerial scenes with low baseline. All tests confirm that the proposed three steps rejection method yields a fairly dense disparity map (40%-90%) with less than 0.4% error matches.

3.1 Introduction

Fattening is probably the main drawback inherent to block-matching methods. In the stereovision literature it appears with different names: fattening, adhesion or border errors. This phenomenon is observed in the disparity map as an apparent foreground dilation. It is produced when one of the blocks contains a depth discontinuity, especially when this discontinuity coincides with a large gray level discontinuity. The size of the dilation depends on the window size: Every pixel in the image at a distance to the edge smaller than a half window risks fattening.

Fig. 3.1 shows a situation where fattening takes place. We study 3 points in the scene: Q on the roof of a building and, R and S two points on the ground. The projections of such points are $\mathbf{q}$, $\mathbf{r}$ and $\mathbf{s}$ in the reference image plane and $\mathbf{q}'$, $\mathbf{r}'$ and $\mathbf{s}'$ in the secondary image plane. In the matching process, squared blocks centered at these points are considered.

The corresponding blocks $B_{\mathbf{q}}$ in the reference image and $B_{\mathbf{q}'}$ in the secondary image lie entirely on the roof. The block-matching method presented in Chapter 2 permits to compute correctly the shift between $\mathbf{q}$ and $\mathbf{q}'$. In the same way, the matching blocks $B_{\mathbf{s}}$ and $B_{\mathbf{s}'}$ lie completely on the ground, and the disparity between their centers is correctly estimated whenever the region is textured enough. Our method is able to reject a match for $B_{\mathbf{s}}$ when it lies in a poor textured region as a shadow.

On the contrary, our *a contrario* method does not necessarily reduce the fattening effect. Indeed, the correct match of the point $\mathbf{r}$, lying on the ground, is the point $\mathbf{r}''$. But the match $B_{\mathbf{r}}/B_{\mathbf{r}'}$ is more meaningful than $B_{\mathbf{r}}/B_{\mathbf{r}''}$ is.

Indeed, the block $B_{\mathbf{r}}$ contains part of the roof, part of the ground and a contrasted edge separating both sides. Let us distinguish the two parts of $B_{\mathbf{r}}$: pixels belonging to the roof (including the edge) B_1, and pixels belonging to the ground B_2. The correct correspondence of $B_{\mathbf{r}}$ appears split in the secondary image. V_1 corresponds to B_1, W_2 corresponds to B_2, and V_2 and W_1 are occluded areas. Then, the two possible matches $B_{\mathbf{r}}/B_{\mathbf{r}'}$ and $B_{\mathbf{r}}/B_{\mathbf{r}''}$ contain occlusions. The choice of one or another depends on the most textured area, B_1 or B_2. Since B_1 and V_1 contain a contrast edge, which is the most relevant texture, $B_{\mathbf{r}}$ is (mis)matched to $B_{\mathbf{r}'}$. The edge steers the matching, especially when B_2 and W_2 lie on a shadow without texture.

Thus, the center of $B_{\mathbf{r}}$ inherits the roof disparity, and so do all points at a distance to the roof smaller than the half block side length. This results in an apparent dilation, or "fattening", of the building size. On the whole, the matching decision that associates a block meeting an edge with another is correct; but the disparity should be attributed to the edge points of the block, and not necessarily to its center. If the edge contrast dominates the texture, or if the roof texture dominates the ground roof, such block matches are meaningful, and are not necessarily rejected by the AC and SS thresholds.

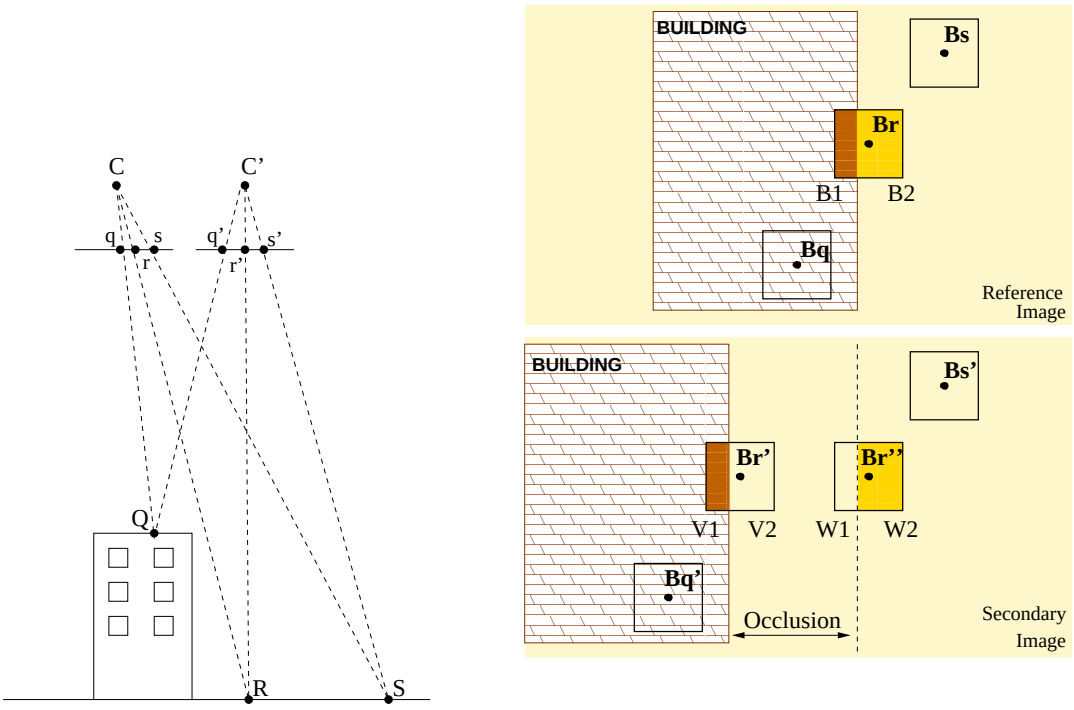


Figure 3.1: The fattening phenomenon. The correct disparity to $\mathbf{r}$ is $\mathbf{r}''$ but $B_{\mathbf{r}}$ is mismatched to $B_{\mathbf{r}'}$.

In urban images, dilation is observed as a dilation of buildings by the patch (but in fact there also is an internal fattening phenomenon, see Fig. 3.2). One may consider smaller comparison windows to reduce fattening but the dilation does not disappear completely. If the comparison window is too small the matching loses accuracy due to the noise. A line segment detector LSD [Grompone et al., 2008] (see Appendix B) can be used to eliminate all patches meeting line segments in the correlation process. This method avoids fattening quite

well in urban images, but this is anyway a limited improvement.

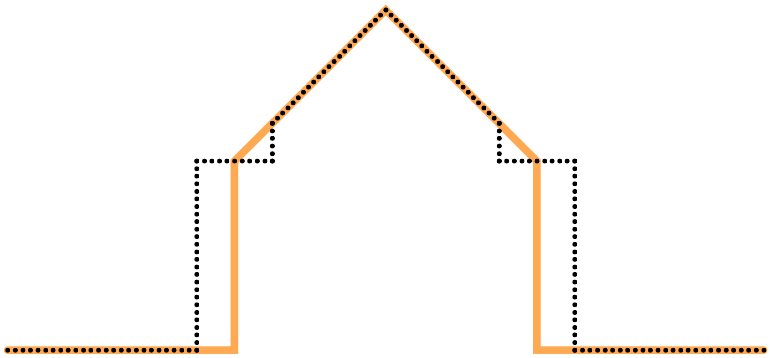


Figure 3.2: Continuous line: cut of the depth map of a house. Dotted line: the estimated depth affected with fattening when the border of the house coincides with a gray level discontinuity. Fattening provokes a dilation of the building, but also a flattening of the roof near the border.

Plan

This chapter is organized as follows: Section 3.2 gives the state-of-the-art of fattening correction. Section 3.3 gives a detailed explanation of our proposed fattening correction. Section 3.4 summarizes the presented algorithm and details the main steps. Section 3.5 details and discusses results of the resulting block-matching algorithm for various image pairs and evaluates the wrong match rates.

3.2 State-of-the-Art and Related Work

We mentioned that all block-matching methods suffer from fattening. The usual way to cope with it is to use adaptive windows that avoid image discontinuities. [Kanade and Okutomi, 1994] described a method to select an appropriate window by evaluating the local variation of the intensity and the disparity. An adaptive window (with varying size and shape) is chosen for each point in an iterative stereo matching algorithm. [Lotti and Giraudon, 1994] points out that this solution does not give good disparities without a good initial estimate of the discontinuities. This paper therefore proposes a contour constrained window correlation algorithm to obtain the initial disparity map with good discontinuity localization. However, thinning the window in one direction implies a lengthening in the orthogonal direction, if the window area has to be kept constant. [Boykov et al., 1998] presents a variable window approach, which chooses an arbitrarily shaped window that varies at each pixel. Results show improvements over classic correlation algorithms. However, the authors point out that a systematic error occurs when they propagate information from textured areas to nearby low-textured areas. Finally, [Veksler, 2002b] and [Veksler, 2003b] choose a range of window sizes and shapes for correlation evaluation but this method needs much parameter tuning for the window cost computation. [Hirschmuller et al., 2002] proposes a real-time correlation

algorithm with improved disparities at depth discontinuities by using multiple supporting windows. Since this treatment reduces slightly the fattening effect, a supplementary correction is done on the disparity discontinuities as a post-precessing step. In such discontinuities, the correlation is computed again with split windows and the disparity discontinuity placed at the new correlation minimum.

Other existing methods fix the size and shape of a local window and assign weights to each pixel on the window in order to improve the results on object borders [Prazdny, 1987], [Darrell, 1998], [Xu et al., 2002] and more recently, [Yoon, 2006].

Point feature matching methods overcome the fattening problem at the cost of a drastic reduction of the match density. Matched features can also be curvilinear, which also circumvents the fattening problem to some extent. For instance, [Schmid and Zisserman, 2000] describes a set of algorithms for automatically matching individual line segments and curves. [Robert and Faugeras, 1991] presents an edge-based stereovision algorithm, where the primitives to be matched are cubic B-splines approximations of the 2-D edges. [Musé et al., 2006a] and [Cao et al., 2007] discuss how to automatically match pieces of level lines and extract coherent groups of such matches. [Matas et al., 2004] solves the problem by matching stable and homogeneous image regions, but their match set is again sparse. Even if features may seem more local, they depend anyway on a broad neighborhood. It is true that the fine scale Laplacian extrema used (e.g.) in the SIFT method are very local, but their descriptor around involves anyway a 8×8 window. Thus, if this window contains some edge, the fattening problem can occur anyway.

Global methods do not suffer from fattening as block-matching does but they can propagate errors in homogeneous regions. In short, all methods either are rather sparse, or are more complete but make errors of a type or another: fattening errors in local methods and error propagation in global methods.

The *Barycentric Correction*

The solution proposed stems from [Delon and Rougé, 2007] and [Delon, 2004a]. In their analytic study of correlation, the authors deal with the fattening artifact and propose a new correction, the *barycentric correction*.

The *barycentric correction* consists in associating the estimated shift $\mu(\mathbf{q}) = q_1 - q_1'$ between $B_{\mathbf{q}}$ and its matched patch $B_{\mathbf{q}'}$ to the barycenter of the correlation window

$$G(\mathbf{q}) = \frac{\int_{\varphi_{\mathbf{q}}} d_{\mathbf{q}}(x) x \, dx}{\int_{\varphi_{\mathbf{q}}} d_{\mathbf{q}}(x) \, dx} ,$$

where φ is a spheroidal prolate function, $supp(\varphi_{\mathbf{q}}) \subseteq B_{\mathbf{q}}$ and $d_{\mathbf{q}}(x)$ is the density function depending on u and the derivative in the direction of the epipolar direction u_x

$$d_{\mathbf{q}}(x) := \frac{\|u\|^2_{\varphi_{\mathbf{q}}} u_x^2(x) - u(x)u_x(x) \int_{\varphi_{\mathbf{q}}} u(x)u_x(x) \, dx}{\|u\|^4_{\varphi_{\mathbf{q}}}} ,$$

with $\|u\|^2_{\varphi_{\mathbf{q}}} = \int_{\varphi_{\mathbf{q}}} u^2(x) \, dx = \int \varphi(\mathbf{q} - x) u^2(x) \, dx$ the weighted norm.

These authors justify by an optimization argument the choice of the density $d_{\mathbf{q}}$ as the most robust to noise. Yet, points with high $d_{\mathbf{q}}$ correspond mostly to image edges. Thus, assigning the computed disparity $\mu(\mathbf{q})$ to $G(\mathbf{q})$ instead of $\mathbf{q}$ concentrates disparities on the

border of contrasted objects. The values of $G(\mathbf{q})$ are not necessary integers, so the resulting disparity map is defined on an irregular sampling grid. An interpolation of the disparity map is needed to return to a regular grid.

The barycentric correction is optimal when the compared patch contains only one edge, with only one discontinuity in depth. Unfortunately, this is not always the case. The patch may well contain several gray level discontinuities corresponding to different depths. In that case, assigning the estimated disparity to the barycenter of the patch creates an erroneous disparity. The barycentric correction had been integrated to MARC (Multiresolution Algorithm for Refined Correlation) and some improvement was noticed between this and a classic correlation algorithm, but at the same time, MARC can be worse in some cases. Facciolo [Facciolo, 2005] noticed that the MARC interpolation performed at each scale to fill in the unknown values can propagate fattening errors. He proposed a variational framework as an alternative to the barycentric correction to reduce the undesirable results produced by MARC's interpolation. Facciolo concludes that his results are smoother than the barycentric correction, but the quantitative results in terms of quadratic error are not remarkably better.

In short, all the local methods, adaptive or weighted windows, feature matching or improved correlation can bring some improvement, but cannot completely eliminate the fattening effect. We have seen that the use of *a priori* gray level discontinuities for detecting zones of fattening is not either a good approach. This is why we have decided to detect *a posteriori* the pixels risking fattening and to eliminate them from the disparity map, thus avoiding any possible fattening error. In order to identify such pixels our disparity map is compared with a new disparity map inspired by the barycentric correction. The new disparity map is computed by assigning each estimated disparity to the pixels (and not just one pixel) in the patch that have most contributed to the similarity of the compared patches.

3.3 Avoiding the Fattening Phenomenon

In this section we describe a new fattening detection and elimination algorithm. Let u_1 and u_2 be a pair of stereo images. Let $\mu(\mathbf{q})$ be the disparity computed in $\mathbf{q}$ by the *a contrario* method complemented by the self-similarity threshold described in Chapter 2. Note that μ is not necessarily dense, and that it can be affected by fattening at some pixels. Let μ_m be the disparity map after a median filter

$$\mu_m(\mathbf{q}) = \underset{\mathbf{y}\in B_{\mathbf{q}}}{Med}\{\mu(\mathbf{y}) \,|\, \mu(\mathbf{y}) \neq \emptyset\}\,.$$

The median filter produces a denser disparity map. We start by defining a disparity map $\tilde{\mu}$ that will be more correct than μ at points suffering fattening. We interpret fattening as a wrong attribution of the disparity estimated in the patch. In Figure 3.1, $\mathbf{r}$ is mismatched with $\mathbf{r}'$ because the center pixel of $B_{\mathbf{r}}$ inherits the shift between the matched blocks. Instead, if the shift is attributed to pixels in the edge of $B_{\mathbf{r}}$ the estimated disparity is the correct one. Then, in order to find the pixels that should won the disparity in the patch, we will match their orientation gradients. Only pixels whose orientations match well inherit the computed disparity. Figure 3.3 shows the main difference between the new disparity assignment and the barycenter correction.

Definition 8 *For each* $\mathbf{x} \in B_{\mathbf{y}}$ *define*

$$\alpha^{\mathbf{y}}(\mathbf{x}) = Angle\left(\frac{\nabla u_1}{|\nabla u_1|}(\mathbf{x})\,,\, \frac{\nabla u_2}{|\nabla u_2|}(\mathbf{x}+\mu(\mathbf{y}))\right).$$

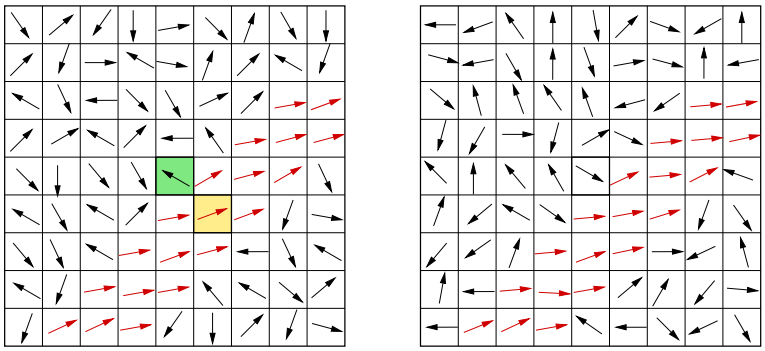


Figure 3.3: 9×9 corresponding patches in left and right images. The arrows represent the gradient direction at each pixel. Red arrows indicate the 25% patch pixels whose gradient orientation matches best. The proposed method assigns the block disparity to all them. On the other hand, the *barycentric correction* assigns the disparity to only one pixel, the barycenter of the patch, colored in yellow, and the classic Winner-Take-All (WTA) assigns the disparity to the center of the patch, colored in green.

Hence the corrected disparity $\tilde{\mu}$ *is defined as:*

$$\tilde{\mu}(\mathbf{q}) = \underset{\mathbf{y}\in B_\mathbf{q}}{Med}\Big\{\mu(\mathbf{y}) \;\Big|\; \alpha^\mathbf{y}(\mathbf{q}) < Q_1(B_\mathbf{y})\Big\}, \qquad \forall \mathbf{q} \tag{3.1}$$

where $Q_1(B_\mathbf{y})$ *is the lowest quartile of* $\{\alpha^\mathbf{y}(\mathbf{x}) \;|\; \mathbf{x} \in B_\mathbf{y},\ |\nabla u_1|(\mathbf{x}) > 3\sigma\}$ *and* σ *the noise standard deviation.*

We could keep the corrected disparity $\tilde{\mu}$ as the final disparity map but its density is rather low. Besides, $\tilde{\mu}$ is a corrected disparity map but fattening errors can remain. Instead, we will use $\tilde{\mu}$ (and μ) to detect pixels risking fattening in order to remove them from the set of reliable disparities.

Notice that neither μ nor $\tilde{\mu}$ are dense disparity maps, so the comparison between them cannot be done for all the pixels in the image. Then, in order to detect *all* the pixels risking fattening, three different situations must be considered, depending on the information available at a given pixel.

- First, consider pixels $\mathbf{q}$ where $\mu(\mathbf{q})$ and $\tilde{\mu}(\mathbf{q})$ are available and incoherent, that is, pixels with a difference between the initial disparity $\mu(\mathbf{q})$ and the corrected one $\tilde{\mu}(\mathbf{q})$ of more than a threshold θ:

$$\Omega_1 := \Big\{\mathbf{q} \in I \;\Big|\; |\mu(\mathbf{q}) - \tilde{\mu}(\mathbf{q})| > \theta, \quad \mu(\mathbf{q}), \tilde{\mu}(\mathbf{q}) \neq \emptyset\Big\}. \tag{3.2}$$

  In practice, θ can be fixed as the authorized error for μ.

- Second, consider pixels where the disparity has a jump in the horizontal or vertical direction of more than θ:

$$\Omega_2 := \Big\{\mathbf{q}=(q_1,q_2) \in I \;\Big|\; \exists \mathbf{r} \in \big\{(q_1 \pm 1, q_2), (q_1, q_2 \pm 1)\big\} \text{ s.t. } |\mu_m(\mathbf{q}) - \mu_m(\mathbf{r})| > \theta\Big\}. \tag{3.3}$$

  Indeed, fattening is present on the boundary of objects, so discontinuities in the disparity map are also candidates to suffer from fattening.

- Finally, consider the pixels with missing neighbors disparities

$$\Omega_3 := \Big\{ \mathbf{q}{=}(q_1,q_2) \in I,\ \mu_m(\mathbf{q}) \neq \emptyset \ \Big|\ \exists \mathbf{r} \in \big\{(q_1 \pm 1, q_2), (q_1, q_2 \pm 1)\big\} \text{ s.t. } \mu_m(\mathbf{r}) = \emptyset \Big\}. \tag{3.4}$$

The points belonging to Ω_3 are points delineating the holes of μ_m, so they are potentially pixels with fattening errors. If there is some missing information in a region of the image, neither a jump in μ_m nor a incoherence between μ and $\tilde{\mu}$ can be detected, whence the need of adding Ω_3 to the set of risking points. Notice that we use the disparity map μ_m instead of μ since the median filter fills small regions of missing disparities in μ and we are only interested in important regions of the image where no disparities have been computed. Consider for example, the following common situation: a building casting a shadow in the ground. On the one hand, the image patch containing part of the shadow and part of the building has a prominent edge. If the patch is not centered on the edge it risks fattening. On the other hand, shadows are poorly textured, so shadow pixels are rejected by the *a contrario* test and only disparities from the roof of the building are available. This creates a big hole in the disparity map.

The set of pixels in $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3$ are detected as pixels risking fattening. In fact, pixels around Ω risk fattening as well. Thus, we define the set of pixels with fattening risk as the dilated $D(\Omega)$ of Ω.

The size dilation is equal to the patch size but the dilation it is not done symmetrically in both directions. Assume that the reference and secondary images are sorted in such a manner that foreground points have larger disparities and background points smaller ones (white disparities in the roofs and black ones in the ground). With this convention a positive disparity jump means a left object border while a negative jump means a right object border. Then, the dilation is done in the direction of pixels with bigger disparities, corresponding to foreground objects. The case where there are only disparities in one side of the risking point the dilation is done in this direction. More precisely, if W is the pach size, the dilation in the horizontal direction are all the pixels in

$$\begin{aligned}
D_1 = \\
&= \Big\{ \mathbf{r} \in I \ \Big|\ \exists \mathbf{q} \in \Omega_1,\ \exists \mathbf{t}, \mathbf{s} \quad \text{s.t.} \quad 0 \leqslant (t_1 - q_1), (q_1 - s_1), (r_1 - q_1) \leqslant W, \quad \text{and} \quad \mu(\mathbf{t}) - \mu(\mathbf{s}) > \theta \Big\} \\
&\bigcup \Big\{ \mathbf{r} \in I \ \Big|\ \exists \mathbf{q} \in \Omega_1,\ \exists \mathbf{t}, \mathbf{s} \quad \text{s.t.} \quad 0 \leqslant (t_1 - q_1), (q_1 - s_1), (q_1 - r_1) \leqslant W, \quad \text{and} \quad \mu(\mathbf{s}) - \mu(\mathbf{t}) > \theta \Big\} \\
&\bigcup \Big\{ \mathbf{r} \in I \ \Big|\ \exists \mathbf{q} \in \Omega_2 \quad \text{s.t.} \quad 0 \leqslant (r_1 - q_1) \leqslant W \quad \text{and} \quad \mu_m(\mathbf{q}) - \mu_m(q_1 - 1, q_2) > \theta \Big\} \\
&\bigcup \Big\{ \mathbf{r} \in I \ \Big|\ \exists \mathbf{q} \in \Omega_2 \quad \text{s.t.} \quad 0 \leqslant (q_1 - r_1) \leqslant W \quad \text{and} \quad \mu_m(q_1 + 1, q_2) - \mu_m(\mathbf{q}) > \theta \Big\} \\
&\bigcup \Big\{ \mathbf{r} \in I \ \Big|\ \exists \mathbf{q} \in \Omega_3 \quad \text{s.t.} \quad 0 \leqslant (r_1 - q_1) \leqslant W \quad \text{and} \quad \mu_m(\mathbf{q}) \neq \emptyset, \quad \mu_m(q_1 - 1, q_2) = \emptyset \Big\} \\
&\bigcup \Big\{ \mathbf{r} \in I \ \Big|\ \exists \mathbf{q} \in \Omega_3 \quad \text{s.t.} \quad 0 \leqslant (q_1 - r_1) \leqslant W \quad \text{and} \quad \mu_m(\mathbf{q}) \neq \emptyset, \quad \mu_m(q_1 + 1, q_2) = \emptyset \Big\}.
\end{aligned} \tag{3.5}$$

In the same way, the vertical direction D_2 is defined taking q_2, r_2, s_2 and t_2 instead of q_1, r_1, s_1 and t_1. Finally $D(\Omega) = D_1 \cup D_2$ is the set of pixels risking fattening.

We call *fattening risk edges* the edges γ inside the risk zone $D(\Omega)$. They are detected as Canny-Deriche edges [Deriche, 1987] (with $\alpha = 1$ the width of the input response) which is

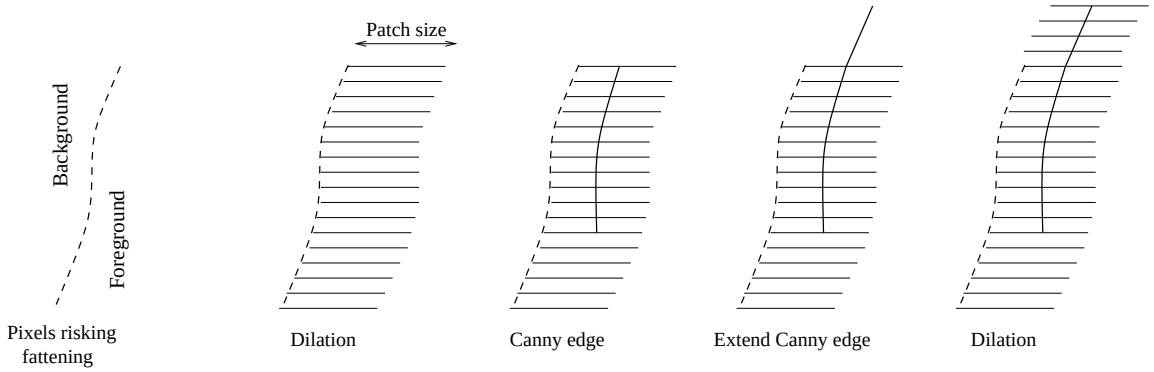


Figure 3.4: Steps of the algorithm to avoid fattening. First, the set of points risking fattening are considered ($\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3$). The values of disparities at each side determine the background/foreground positions. Then a dilation of the size of the patch size is done in the foreground direction. When no disparity is available on one side of the risking point, the other side is chosen for dilation. Fattening risk edges (Canny-Deriche edges) inside the dilation band are considered and extended, if needed. Finally, a dilation is done for the extended C.-D. edges.

based on Canny [Canny, 1986] criteria. These edges cause potentially fattening errors around them. If the C.-D. edges continue beyond the extreme of γ (and out of D(Ω)) they are probably causing fattening as well. For security, γ is extended with such pixels whenever the patch centered at them has disparities differing of more than θ.

Therefore, in our resulting disparity map we will remove from μ the set of pixels risking fattening:

Definition 9 *The final disparity map μ_F is defined as*

$$\mu_F(\mathbf{q}) = \begin{cases} \emptyset & \textit{if} \quad B_{\mathbf{q}} \cap \gamma \neq \emptyset \quad \textit{or} \quad \mathbf{q} \in D(\Omega)\,, \\ \mu(\mathbf{q}) & \textit{otherwise}\,. \end{cases} \tag{3.6}$$

Fig. 3.4 shows step by step this part of the algorithm (see more details in the section 3.4).

3.4 Algorithm Synopsis for Fattening Correction

1. Compute the median disparity map $\mu_m(\mathbf{q}) = Med\{\mu(\mathbf{y}) \mid \mathbf{y} \in B_{\mathbf{q}},\ \mu(\mathbf{y}) \neq \emptyset\}$.
2. Compute the corrected disparity map $\tilde{\mu}$ from def. (3.1).
3. Compute Ω_1, Ω_2 and Ω_3 from definitions (3.2), (3.3) and (3.4).
4. Sweep the image from left to right to compute D_1 (def. (3.5)), and sweep the image from top to bottom to compute D_2. Compute $D(\Omega) = \Omega_1 \cup \Omega_2$.
5. Compute Canny-Deriche edges and keep the parts of these edges γ within $D(\Omega)$ and mark the extremes of γ.
6. Extend Canny-Deriche edges: while $\exists \mathbf{r} \notin \gamma, \quad ||\mathbf{q} - \mathbf{r}|| \leqslant 1$, with $\mathbf{r}$ a pixel of the C.-D. edges and $\mathbf{q}$ an extreme of γ such that $|\max_{\mathbf{x} \in B_{\mathbf{r}}}\{\mu(\mathbf{x})\} - \min_{\mathbf{x} \in B_{\mathbf{r}}}\{\mu(\mathbf{x})\}| > \theta$, then $\mathbf{r}$ is added to γ and it will be the new extreme.

7. Compute the final disparity map (using definition (3.6)).

See algorithm 4 in chapter 5 for the pseudocode of the algorithm.

3.5 Experiments

In this section several results of the presented algorithm will be discussed. The algorithm parameters are fixed and the same for all experiments. The comparison window size is 9×9, the number of considered principal components is 9, the number of quantum probabilities is 5, and the number of regions for the local PCA is 4 (2×2).

More results will be discussed in chapter 6.

3.5.1 Comparison with Other Non-Dense Algorithms

Here we are going to compare our algorithm with the ones presented in [Sara, 2002], [Veksler, 2002a], [Veksler, 2003a] and [Mordohai and Medioni, 2006]. All of these papers have published experimental results on the first Middlebury dataset [Scharstein and Szeliski, 2002] (Tsukuba, Sawtooth, Venus and Map pair of images) on the non-occluded mask. All of these algorithms are characterized by computing sparse disparity maps and each of them proposes a different method to reject pixels. Table 3.1 summarizes the percentage of matched pixels (density) and the percentage of mismatches (the estimated disparity differs more than one pixel to the ground truth). In this table we report two results of our algorithm. First, the results of the original algorithm as it has been explained above. The error rate for this algorithm is very small and it yields larger densities than Sara's results. However the comparison is difficult when other algorithms propose denser disparity maps. Thus, we have made the results of our algorithm denser by the most straightforward interpolation, namely by a median filter on all the patches not meeting the risk-edges. Doing this, the density rises while keeping small error rates. Still, there are images with large regions containing poor or repeated textures on which reliable disparity maps cannot be very dense, even if the median filter is used. Figure 3.5 shows the resulting disparity maps for the four images.

The authors of [Mordohai and Medioni, 2006] compute an initial classic correlation disparity map and select correct matches based on the support they receive from their neighboring candidate matches in 3D after tensor voting. 3D points are grouped into smooth surfaces using color and geometric information and inconsistent points with the surface color distribution are removed in order to avoid the fattening phenomenon. The rejection of wrong pixels is not complete, because the algorithm fails when some objects appear only in one image, or when occluded surfaces change orientation. The choice of critical rejection parameters can lead to quite different results.

[Veksler, 2002a] detects and matches dense features which is a connected set of pixels in the left image and a corresponding set of pixels in the right image such that the intensity edges on the boundary of these sets are stronger than their matching error on the boundary (which is the absolute intensity difference between corresponding boundary pixels). They call this the "boundary condition". The idea is that even the boundary of an untextured region can give a correspondence. Then, each dense feature is associated with a disparity. Their main limitation is the way they extract dense features. They are extracted using a local algorithm which processes each scan line independently from the other. As a result, top and bottom boundaries are lost. On the contrary, [Veksler, 2003a] use graph cuts for dense feature extraction and enforce the boundary conditions. Veksler's results are rather

	Tsukuba		Sawtooth		Venus		Map	
	Error	Density	Error	Density	Error	Density	Error	Density
Our results 1	**0.31**	45.6	**0.09**	65.7	0.02	54.1	**0.0**	84.8
Our results 2	0.33	54.3	0.14	77.9	**0.0**	66.6	**0.0**	93.0
Sara	1.4	45	1.6	52	0.8	40	0.3	74
Veksler 02	0.38	66	1.62	76	1.83	68	0.22	87
Veksler 03	0.36	75	0.54	87	0.16	73	0.01	87
Mordohai and Medioni	1.18	74.5	0.27	78.4	0.20	74.1	0.08	94.2

Table 3.1: Quantitative results on the first Middlebury benchmark data set. The error statistics are computed on the mask of non occluded pixels and a mismatch is an error bigger than 1 pixel. Our algorithm obtains less mismatches in the four images.

dense and the error rate is one of the most competitive ones. However, its dense features can only overlap one displacement which is a very restrictive constraint and the algorithms should not be very performant in more complex images. Note that Sawtooth, Venus and Map are piecewise planar surfaces (almost fronto-parallel surfaces) and the ground truth of Tsukuba is piecewise constant with 6 different disparities.

Finally, [Szeliski and Scharstein, 2002] obtained an error rate of 2.1% with a density of 45% but semi-dense results on other images are not published.

3.5.2 The Simulated Stereo Pair

The aerial urban scene experiment has been performed with a simulated stereo pair, because this is the only way to have a completely reliable ground truth. The secondary image was simulated from the reference image and a real ground truth, that in fact had many errors. By the simulation the ground truth becomes really true. However, the simulation takes into account realistic acquisition parameters, namely an optical blur and strong enough independent white noise added to both images. Fig. 3.6 shows the resulting aerial stereo pair and its ground truth.

After the simulation of the secondary image from the reference image and the ground truth a white noise is added independently to each image. The more noise in the images, the less matches are found. This is coherent: There are less meaningful matches in presence of noise, but the established matches remain anyway weakly erroneous.

Table 3.2 compares the error committed after the four steps for various noise levels. The table gives the signal to noise ratio $SNR = \|u\|_2/\sigma$, where σ is the standard deviation of the noise, the percentage of matched pixels and the percentage of wrong matches. (We call wrong match any pixel at which computed disparity and ground truth differ by more than one pixel).

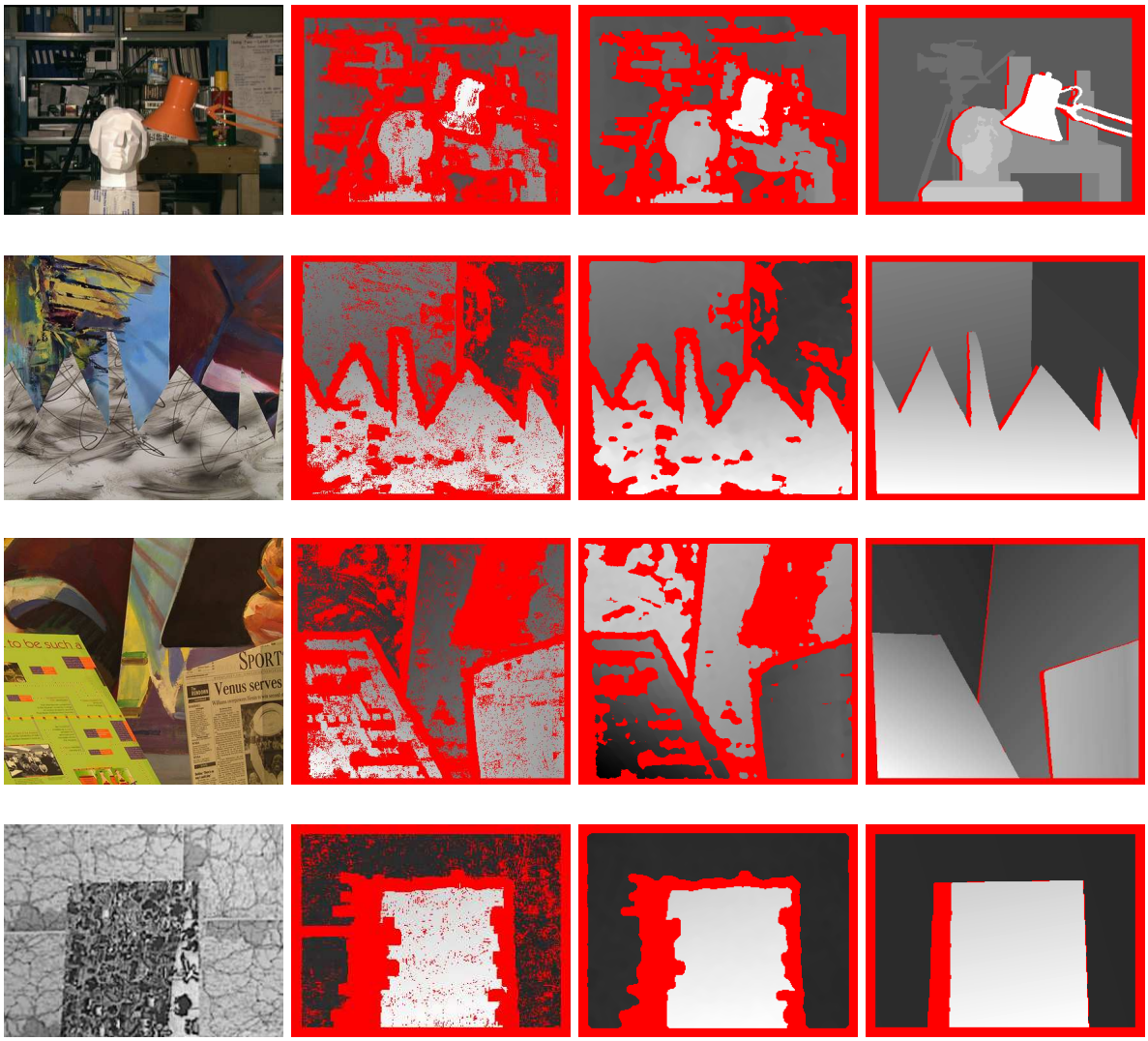


Figure 3.5: From top to bottom: Tsukuba, Sawtooth, Venus and Map experiment. From left to right: reference image, disparity map obtained with the three-step algorithm presented above (red pixels are not matched pixels), disparity map after median filter, ground truth (red pixels are not considered in the mask of non occluded points.)

SNR	Density	Error
∞	60.1	0.005
357.32	58.5	0.008
178.66	54.3	0.009
125.06	49.27	0.027

Table 3.2: From left to right: Signal to noise ratio. Percentage of matched points. Percentage of wrong matches.

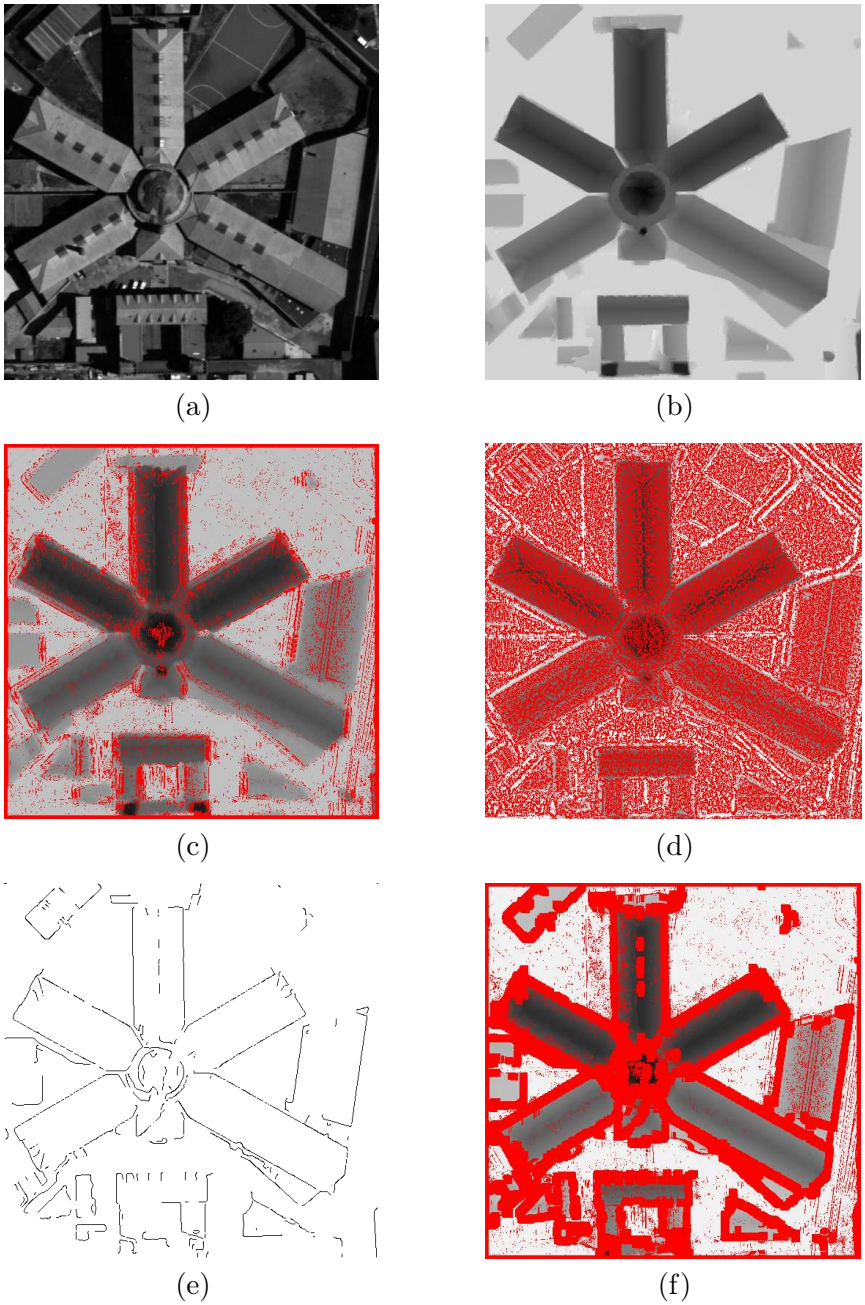


Figure 3.6: (a) Reference aerial image. (b) Ground truth. Notice that darker values correspond to higher points in the scene. (c) Subpixel disparity map with fattening errors (μ). (d) Corrected disparity map with an angle gradient matching ($\tilde{\mu}$). (e) Fattening risk edges. (f) Final disparity map.

Chapter 4

Optimal Stereo Matching Reaches Theoretical Accuracy Bounds

Contents

Résumé : La reconstruction 3D à partir de deux images nécessite la parfaite maîtrise d'une longue chaîne d'algorithmes : les calibrations interne et externe, la rectification épipolaire, la mise en correspondance par blocs et la reconstruction 3D. Ce chapitre porte sur l'étape cruciale de la mise en correspondance par blocs. Il montre que la carte des disparités d'images rectifiées épipolairement peut être calculée pour la majorité des points avec une précision de 1/20 pixels dans des conditions de bruit réalistes. Une prédiction théorique des erreurs dûes au bruit sera donnée. Il sera prouvé, sur plusieurs expériences simulées et réelles, que cette borne est atteinte par l'algorithme proposé. Les expériences sur le *benchmark* Middlebury montrent que la méthode, qui se base sur une interpolation et un échantillonnage précis, améliore la précision de la vérité du terrain.

Abstract: 3D reconstruction from two images requires the perfect control of a long chain of algorithms: internal and external calibration, stereo-rectification, block-matching, and 3D reconstruction. This chapter focuses on the crucial block-matching step. It shows that the disparity map in stereo-rectified images can be computed for a majority of image points up to a 1/20 pixel precision under realistic noise conditions. A theoretical prediction of the errors caused by noise will be given, and it will be proved on several simulated and real experiments that this bound is reached by the proposed algorithm. Experiments on the Middlebury benchmark even show that the method, relying on accurate interpolation and sampling, improves the precision of the ground truth.

4.1 Introduction

4.1.1 Small Baseline

Classic reviews on stereo vision [Scharstein and Szeliski, 2002], [Brown et al., 2003] distinguish local from global stereo methods. Local (block-matching) methods rely on a comparison of a small number of pixels surrounding a pixel of interest, and are sensitive to local ambiguities (occlusions, uniform textures, or simply lack of information). Blocks are usually compared by the normalized cross correlation (NCC) or the sum of squared differences (SSD). Block-matching methods can produce wrong disparities near the intensity discontinuities in the images. This phenomenon is called adhesion or fattening effect.

Several papers attempt to solve this problem by using adaptive windows [Kanade and Okutomi, 1994], [Lotti and Giraudon, 1994], [Kang et al., 2001], by a barycentric correction [Delon and Rougé, 2007], or by feature matching methods [Schmid and Zisserman, 2000]. Global methods such as graph cuts [Kolmogorov and Zabih, 2005] and dynamic programming [Ohta and Kanade, 1985], [Forstmann et al., 2004] are less sensitive to fattening but, because of the global nature of the optimization process, they are not prone to a precision analysis. Thus, even global methods could benefit from a previous highly accurate non dense block-matching.

Indeed, to the best of our knowledge, the block-matching precision and its noise dependence have never been properly quantified. The possibility of rigorous block-matching with sub-pixel accuracy by a factor 2 oversampling was actually noticed in [Szeliski and Scharstein, 2004]. Sub-pixel accurate matching is also sought in the MARC method [Giros et al., 2004] used by the French space agency (CNES). The first theoretical arguments towards high accuracy in stereo vision were given in [Delon and Rougé, 2007], who claimed that high precision matches can be obtained from a small baseline stereo pair of images if the Shannon-Whittaker conditions are met. However, this paper neither gave an accurate formula for the attainable

precision, nor demonstrated its practical feasibility. Small baselines in conjunction with larger ones had been considered in [Okutomi and Kanade, 1993], a pragmatic study where different baselines were used to eliminate errors. However, its final sub-pixel results were computed with the large baseline samples.

This chapter proposes sharp theoretical subpixel accuracy estimates depending on noise, and an algorithm to reach them. Simulated pairs and real examples including benchmark data will confirm that the theoretical error bounds are reached. Furthermore, on several realistic simulations and on benchmark examples, theory and practice will confirm a 1/20 pixel block-matching accuracy. This accuracy will be demonstrated at points which are predicted *a priori* on each image pair by a careful elimination of risky matches. This elimination is the object of Chapters 2 and 3.

Let us denote by $\mathbf{x} = (x, y)$ an image point in the continuous image domain, and by $u_1(\mathbf{x}) = u_1(x, y)$ and $u_2(\mathbf{x})$ the images of an ortho-rectified stereo pair. Assume that the epipolar direction is the x axis. The underlying depth map can be deduced from the disparity function $\varepsilon(\mathbf{x})$ giving the shift of an observed physical point $\mathbf{x}$ from the left image u_1 in the right image u_2. The physical disparity $\varepsilon(\mathbf{x})$ is not well-sampled. Therefore, it cannot be recovered at all points, but only essentially at points $\mathbf{x}$ around which the depth map is continuous. Around such points, a deformation model holds:

$$\begin{aligned} u_1(\mathbf{x}) &= u(x + \varepsilon(\mathbf{x}), y) + n_1(\mathbf{x}) \\ u_2(\mathbf{x}) &= u(\mathbf{x}) + n_2(\mathbf{x}). \end{aligned} \tag{4.1}$$

where n_i are Gaussian noises and $u(\mathbf{x})$ is the ideal image that would be observed instead of $u_2(\mathbf{x})$ if there were no noise. The deformation model (4.1) is *a priori* valid when the angle of the 3D surface at $\mathbf{x}$ with respect to the camera changes moderately, which is systematically true for small (0.02 to 0.15) baseline stereo systems. The restriction brought by (4.1) is moderate. Indeed, the trend in stereo vision is to have multiple views of the 3D object to be reconstructed and therefore many pairs with small base line.

4.1.2 The Causes of Error in Block-Matching Stereo

If the images u_1 and u_2 have little aliasing, [Delon and Rougé, 2007] showed that the recovered disparity map obtained by minimizing a continuous quadratic distance between u_1 and u_2 has two error terms: the fattening error, and the error due to noise. Fattening is a classic problem in block-matching methods. It occurs when a salient image feature lies *within* the comparison window φ but *away* from its center. This may produce a large error near points at which the disparity ε has a jump (see Chapter 3).

Measuring a high accuracy also requires eliminating all mismatches. Luckily, there are several techniques to avoid gross errors, including coarse-to-fine scale refinement [Giros et al., 2004], SIFT thresholds [Lowe, 2004] and *a contrario* methods [Musé et al., 2006a]. Here the *a contrario* rejection algorithm presented in previous chapters was used to eliminate *a priori* the unreliable pixels.

The unreliable pixels and the pixels risking fattening usually cover far less than half the image. In geographic information systems, where high accuracy is particularly relevant, these reliable points correspond in general to textured regions (roofs, lawn, terrains). The aim of this chapter is to study the noise error at all reliable pixels, the others being *a priori* detected as risking fattening and mismatch. The experiments will confirm that the predicted theoretical error at reliable pixels is essentially due to noise, and coincides strikingly with the observed

error. The formula for the main disparity error term due to noise given in this chapter is new, exact, and actually far more accurate than the upper bound proposed in [Delon and Rougé, 2007] for the same error, (which was more than 10 times larger).

Plan

This chapter is organized as follows: Section 4.2 describes the theoretical assumptions and the accurate interpolation techniques permitting high sub-pixel accuracy. Section 4.3 proves a formula for the theoretical noise error. Section 4.4 gives algorithms and a complexity analysis. Section 4.5 shows the obtained results for several simulated and real ground truths and demonstrates that the practical error meets its theoretical estimate.

4.2 Preliminaries on Sub-Pixel Interpolation

This section proves a discrete correlation formula which is faithful to the continuous image interpolates. Thanks to it, an accurate subpixel matching becomes possible. Without loss of generality, all considered images u, u_1, etc. are defined on a square $[0,a]^2$ and are supposed to be square integrable. Thus, the Fourier series decomposition applies

$$u(x,y)=\sum_{k,l\in\mathbb{Z}} \tilde{u}_{k,l} e^{\frac{2i\pi(kx+ly)}{a}}, \tag{4.2}$$

where the $\tilde{u}_{k,l}$ are the Fourier series coefficients (or shortly the Fourier coefficients) of u. By the classic Fourier series isometry, for any two square integrable functions $u(\mathbf{x})$ and $v(\mathbf{x})$ on $[0,a]^2$,

$$\int_{[0,a]^2} u(\mathbf{x})\overline{v}(\mathbf{x})d\mathbf{x} = a^2 \sum_{k,l\in\mathbb{Z}} \tilde{u}_{k,l}\overline{\tilde{v}_{k,l}}. \tag{4.3}$$

The digital images are usually given by their N^2 samples $u(\mathbf{m})$ for $\mathbf{m}$ in the grid

$$\mathbb{Z}^1_a = [0,a]^2 \cap \left((\frac{a}{2N}, \frac{a}{2N}) + \frac{a}{N}\mathbb{Z}^2 \right).$$

Similarly, the over-sampling grid with four times more samples is denoted by

$$\mathbb{Z}^{1/2}_a = [0,a]^2 \cap \left((\frac{a}{4N}, \frac{a}{4N}) + \frac{a}{2N}\mathbb{Z}^2 \right).$$

N is always an even integer. In all that follows we shall assume that the images obtained by a stereo vision system are band-limited. This assumption is classical and realistic, the aliasing in good quality CCD cameras being moderate. As classical in image processing, under the (forced) a-periodicity assumption a band-limited image becomes a trigonometric polynomial. This periodicity assumption is not natural, but it only entails a minor drawback, namely a small distortion near the boundary of the image domain $[0,a]^2$. The payoff for the *band-limited + periodic assumption* is that the image can be interpolated, and its Fourier coefficients computed from discrete samples. Indeed, given N^2 samples $u_{\mathbf{m}}$ for $\mathbf{m}$ in $\mathbb{Z}^1_a$, there is a unique trigonometric polynomial in the form

$$u(x,y)= \sum_{k,l=-N/2}^{N/2-1} \tilde{u}_{k,l} e^{\frac{2i\pi(kx+ly)}{a}} \tag{4.4}$$

such that $u(\mathbf{m}) = u_{\mathbf{m}}$. We shall call such polynomials *N-degree trigonometric polynomials*. The coefficients $\tilde{u}_{k,l}$ are the *Fourier coefficients* of u in the Fourier basis $e^{\frac{2i\pi(kx+ly)}{a}}$, $k, l \in \mathbb{Z}$. The map $u_{\mathbf{m}} \to u_{k,l}$ is nothing but the *2D Discrete Fourier Transform (DFT)*, and the map $(u_{\mathbf{m}}) \to N(\tilde{u}_{k,l})$ is an isometry from $\mathbb{C}^{N^2}$ to itself. The function $u(x, y)$ is therefore usually called the *DFT interpolate of the samples $u_{\mathbf{m}}$*. In consequence, there is an isometry between the set of N-degree trigonometric polynomials endowed with the $L^2([0, a]^2)$ norm, and $\mathbb{C}^{N^2}$ endowed with the usual Euclidean norm:

$$\int_{[0,a]^2} |u(x,y)|^2 = a^2 \sum_{k,l=-N/2}^{N/2-1} |\tilde{u}_{k,l}|^2 = \frac{a^2}{N^2} \sum_{\mathbf{m}\in\mathbb{Z}^1_a} |u(\mathbf{y}+\mathbf{m})|^2 \,, \tag{4.5}$$

where the N^2 samples grid can have an arbitrary origin $\mathbf{y}$. If $u(\mathbf{x})$ and $v(\mathbf{x})$ are two N-degree trigonometric polynomials, we therefore also have

$$\int_{[0,a]^2} u(\mathbf{x})\overline{v}(\mathbf{x}) = a^2 \sum_{k,l=-N/2}^{N/2-1} \tilde{u}_{k,l}\overline{\tilde{v}_{k,l}} = \frac{a^2}{N^2} \sum_{\mathbf{m}\in\mathbb{Z}^1_a} u(\mathbf{y}+\mathbf{m})\overline{v(\mathbf{y}+\mathbf{m})} \,, \tag{4.6}$$

where $\overline{v}$ is the complex conjugate of v. Taking four times more samples, it follows from (4.6) that

$$\int_{[0,a]^2} u(\mathbf{x})\overline{v}(\mathbf{x}) = a^2 \sum_{k,l=-N}^{N-1} \tilde{u}_{k,l}\overline{\tilde{v}_{k,l}} = \frac{a^2}{4N^2} \sum_{\mathbf{m}\in\mathbb{Z}^{1/2}_a} u(\mathbf{m})\overline{v(\mathbf{m})}. \tag{4.7}$$

which is also valid if $u(\mathbf{x})$ and $v(\mathbf{x})$ are up to $2N$-degree trigonometric polynomials in $\mathbf{x}$.

This last fact has a first important consequence in block-matching. Consider two images $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ on $[0, a]^2$ and a window function $\varphi(\mathbf{x})$. Block-matching is the search for a value of μ minimizing the continuous quadratic distance

$$e_{\mathbf{x}_0}(\mu) := \int_{[0,a]^2} \varphi(\mathbf{x}-\mathbf{x}_0)\big(u_1(\mathbf{x}) - u_2(\mathbf{x}+(\mu,0))\big)^2 d\mathbf{x}. \tag{4.8}$$

Proposition 2 (Equality of the discrete and the continuous quadratic distance) *Let $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ be two N-degree trigonometric polynomials on $[0, a]^2$ and let $\varphi(\mathbf{x})$ be a window function which we assume to be a $2N$-degree trigonometric polynomial. Then*

$$e_{\mathbf{x}_0}(\mu) = e^d_{\mathbf{x}_0}(\mu), \quad \textit{where} \tag{4.9}$$

$$e^d_{\mathbf{x}_0}(\mu) := \frac{a^2}{4N^2} \sum_{\mathbf{m}\in\mathbb{Z}^{1/2}_a} \varphi(\mathbf{m}-\mathbf{x}_0)\big(u_1(\mathbf{m}) - u_2(\mathbf{m}+(\mu,0))\big)^2. \tag{4.10}$$

The proof follows from (4.7). Indeed, $\big(u_1(\mathbf{x}) - u_2(\mathbf{x}+(\mu,0))\big)^2$ and $\varphi(\mathbf{x}-\mathbf{x}_0)$ are both $2N$-degree trigonometric polynomials in $\mathbf{x}$, so according to (4.7) the discrete scalar product defining $e^d_{\mathbf{x}_0}(\mu)$ equals the continues scalar product defining $e_{\mathbf{x}_0}(\mu)$. Thus *the continuous block distance is a finite sum of discrete samples!*

The block distance function $\mu \to e_{\mathbf{x}_0}(\mu)$, whose minimization is our main objective here, is also easily sampled. By (4.10) it is a $2N$-degree trigonometric polynomial with respect to μ. This proves:

Proposition 3 (Sub-pixel correlation requires $\times$2 zoom) *Let $u_1(\mathbf{x})$ and $u_2(\mathbf{x})$ be two N-degree trigonometric polynomials. Then the quadratic distance $e^d_{\mathbf{x}_0}(\mu)$ is well-sampled provided it has at least $2N$ successive samples. Thus the computation of $e^d_{\mathbf{x}_0}(\mu)$ at half samples $\mu \in \frac{a\mathbb{Z}}{2}$ (via zero-padding) allows the exact reconstruction of $e^d_{\mathbf{x}_0}(\mu)$ for any real μ by DFT interpolation.*

Remark that the last proposition does not require any assumption on the window function $\varphi(\mathbf{x})$. Prop. 3, which opens the way to rigorous block-matching with sub-pixel accuracy, has been noticed in [Szeliski and Scharstein, 2004]. It is also used in the MARC method [Giros et al., 2004] used by the French space agency (CNES). The above simple proof of Prop. 3 is new.

Sub-pixel block-matching will require to interpolate the noisy images. Thus, following Shannon's classical observation, the noise itself must also be interpolated as a band-limited function. In the periodic framework it therefore becomes a trigonometric polynomial. Assume that $(n_{\mathbf{m}}), \mathbf{m} \in \mathbb{Z}^1_a$ are N^2 independent $\mathcal{N}(0,\sigma^2)$ noise samples. This amounts to say that $(n_{\mathbf{m}})$ is a Gaussian vector. Since the DFT is an isometry, the noise Fourier coefficients $N(\tilde{n}_{\mathbf{k}})$ also form a Gaussian vector with diagonal covariance matrix $\sigma^2 Id$. By (4.6), the mapping $(n_{\mathbf{m}})_{\mathbf{m}\in\mathbb{Z}^1_a} \to (n(\mathbf{x}+\mathbf{m}))_{\mathbf{m}\in\mathbb{Z}^1_a}$ is an isometry from $\mathbb{C}^{N^2}$ to itself. It follows that $n(\mathbf{x})$ is $\mathcal{N}(0,\sigma^2)$ for every $\mathbf{x}$.

One can also estimate $\mathrm{Var}(n_x(\mathbf{x}))$, where $n_x(\mathbf{x}) = \frac{\partial n}{\partial x}(x,y)$.

$$\mathrm{Var}(n_x(\mathbf{x})) = \mathrm{Var}\left(\sum_{k,l=-N/2}^{N/2-1} \tilde{n}_{k,l}\frac{2ik\pi}{a}e^{2i\frac{k\pi x + l\pi y}{a}}\right) =$$

$$= \frac{4\pi^2\sigma^2 N}{N^2a^2}\sum_{k=-N/2}^{N/2-1} k^2 \simeq \frac{4\pi^2\sigma^2}{a^2N}\frac{N^3}{12} = \frac{\pi^2N^2}{3a^2}\sigma^2.$$

Since $n(\mathbf{x})$ is a normal law, $n(\mathbf{x})^2$ is a χ^2 law of order 1. Thus its variance is $2\sigma^4$. Finally we shall need to evaluate $\mathrm{Var}(n_1(\mathbf{x})n_2(\mathbf{x}))$, where n_i are two independent interpolated white noises of the above kind. Thus $n_1(\mathbf{x})n_2(\mathbf{x})$ is the product of two normal laws. The expectation of the product is zero and the variance is therefore $\mathrm{Var}(n_1n_2) = \mathbb{E}(n_1n_2)^2 = \mathbb{E}n_1^2\mathbb{E}n_2^2 = = (\mathbb{E}n^2)^2 = \mathrm{Var}(n)^2 = \sigma^4$. In summary:

Lemma 2 *Let $(n_{\mathbf{m}})_{\mathbf{m}\in\mathbb{Z}^1_a}$ be N^2 independent white Gaussian noise samples with variance σ^2. Then the DFT interpolate $n(\mathbf{x})$ on $[0,a]^2$ is $\mathcal{N}(0,\sigma^2)$ for every $\mathbf{x}$. If n_1 and n_2 are two independent noises like n, one has*

$$\mathrm{Var}(n^2(\mathbf{x}))=2\sigma^4, \tag{4.11}$$

$$\mathrm{Var}((n)_x(\mathbf{x}))\simeq\frac{\pi^2N^2}{3a^2}\sigma^2, \tag{4.12}$$

$$\mathrm{Var}(n_1(\mathbf{x})n_2(\mathbf{x}))=\sigma^4. \tag{4.13}$$

Lemma 3 *Take $a = N$ and let $n(\mathbf{x})$ be the DFT interpolate on $[0,N]^2$ of a white noise with variance σ^2 on $\mathbb{Z}^1_N$, as defined above. Let $\varphi(\mathbf{x})$ be a $2N$-degree trigonometric polynomial on $[0,N]^2$. Then*

$$\mathrm{Var}\left(\int_{[0,N]^2}\varphi(\mathbf{x})n(\mathbf{x})n_x(\mathbf{x})d\mathbf{x}\right) \leqslant \frac{\sigma^4}{2}\int_{[0,N]^2}\varphi_x(\mathbf{x})^2 d\mathbf{x}, \tag{4.14}$$

and the expectation of this random variable is null. Let $g(\mathbf{x})$ be any square integrable function on $[0, N]^2$ and let g_N be its least square approximation by a N-degree trigonometric polynomial. Then

$$\operatorname{Var}\left(\int g(\mathbf{x})n(\mathbf{x})d\mathbf{x}\right) = \sigma^2 \int_{[0,N]^2} g_N(\mathbf{x})^2 d\mathbf{x} \leqslant \sigma^2 \int_{[0,N]^2} g(\mathbf{x})^2 d\mathbf{x}. \tag{4.15}$$

Proof: Integrating by parts in x we have

$$V := \operatorname{Var}\left(\int \varphi(\mathbf{x})n(\mathbf{x})n_x(\mathbf{x})d\mathbf{x}\right) = \operatorname{Var}\left(\frac{1}{2}\int \varphi_x(\mathbf{x})n(\mathbf{x})^2 d\mathbf{x}\right).$$

Since $n(\mathbf{x})^2$ and $\varphi(\mathbf{x})$ are $2N$-degree trigonometric polynomials, (4.7) can be used with $a = N$:

$$V = \frac{1}{4}\operatorname{Var}\left(\frac{1}{4}\sum_{\mathbf{m}\in\mathbb{Z}_N^{1/2}} \varphi_x(\mathbf{m})n(\mathbf{m})^2\right).$$

Now, the sum can be split in

$$V = \frac{1}{4^3}\operatorname{Var}(S_1 + S_2 + S_3 + S_4) \leqslant \frac{1}{4^2}\sum_{i=1}^{4}\operatorname{Var}(S_i), \tag{4.16}$$

where $S_i = \sum_{\mathbf{m}\in A_i} \varphi_x(\mathbf{m})n(\mathbf{m})^2$, $A_i = [0, N]^2 \cap (a_i + \mathbb{Z}^2)$, $a_1 = (1/4, 1/4)$, $a_2 = (1/4, 3/4)$, $a_3 = (3/4, 1/4)$, and $a_4 = (3/4, 3/4)$. We shall evaluate for example

$$\operatorname{Var}(S_1) = \operatorname{Var}\left(\sum_{\mathbf{m}\in A_1} \varphi_x(\mathbf{m})n(\mathbf{m})^2\right).$$

The samples $n(\mathbf{m})$, $\mathbf{m} \in A_1$ being independent, $\operatorname{Var}(S_1) = \sum_{\mathbf{m}\in A_1} \varphi_x(\mathbf{m})^2 \operatorname{Var}(n(\mathbf{m})^2)$ which yields by Lemma 2 $\operatorname{Var}(S_1) = 2\sigma^4 \sum_{\mathbf{m}\in A_1} \varphi_x(\mathbf{m})^2$. Thus, from (4.16) follows that $V \leqslant \frac{2\sigma^4}{4^2}\sum_{\mathbf{m}\in\mathbb{Z}_N^{1/2}} \varphi_x(\mathbf{m})^2$ which, using again (4.7) with $a = N$, yields

$$V \leqslant \frac{4 \times 2\sigma^4}{4^2}\int \varphi_x^2(\mathbf{x}) = \frac{\sigma^4}{2}\int \varphi_x^2(\mathbf{x}).$$

Also,

$$\mathbb{E}\int \varphi(\mathbf{x})n(\mathbf{x})n_x(\mathbf{x})d\mathbf{x} = -\frac{1}{2}\mathbb{E}\int \varphi_x(\mathbf{x})n(\mathbf{x})^2 d\mathbf{x} =$$

$$= -\frac{1}{2}\int \varphi_x(\mathbf{x})\mathbb{E}n(\mathbf{x})^2 d\mathbf{x} = -\frac{\sigma^2}{2}\int \varphi_x(\mathbf{x})d\mathbf{x} = 0.$$

The second part of the lemma is easier. By the Fourier series isometry (4.3),

$$\int_{[0,N]^2} g(\mathbf{x})n(\mathbf{x})d\mathbf{x} = N^2 \sum_{k,l\in\mathbb{Z}} \tilde{g}_{k,l}\tilde{n}_{k,l} =$$

$$= N^2 \sum_{-\frac{N}{2}\leqslant k,l\leqslant\frac{N}{2}-1} \tilde{g}_{k,l}\tilde{n}_{k,l}.$$

Indeed, n being a degree N-trigonometric polynomial, $\tilde{n}_{k,l} = 0$ for $(k,l) \notin [-N/2, N/2-1]^2$. Since the $\tilde{n}_{k,l}$ are independent with variance $\frac{\sigma^2}{N^2}$, we obtain the announced result by taking the variance of the last finite sum:

$$\mathrm{Var}\left(\int_{[0,N]^2} g(\mathbf{x})n(\mathbf{x})d\mathbf{x}\right) = \sigma^2 N^2 \sum_{-\frac{N}{2}\leq k,l\leqslant \frac{N}{2}-1} |\tilde{g}_{k,l}|^2.$$

By (4.5), this yields

$$\mathrm{Var}\left(\int_{[0,N]^2} g(\mathbf{x})n(\mathbf{x})d\mathbf{x}\right) = \sigma^2 \int_{[0,N]^2} g_N(\mathbf{x})^2 d\mathbf{x},$$

where

$$g_N(\mathbf{x}) := \sum_{-N/2\leqslant k,l\leqslant N/2-1} \tilde{g}_{k,l} e^{\frac{2i\pi(kx+ly)}{a}}$$

is the degree N-trigonometric polynomial best approximating g for the quadratic distance.

4.3 Block-Matching Errors Due to Noise

Consider a stereo pair of digital images and their DFT interpolates $u_1(\mathbf{x})$, $u_2(\mathbf{x})$ satisfying (4.1). Block matching amounts to look for every $\mathbf{x}_0$ for the estimated disparity at $\mathbf{x}_0$ minimizing

$$e_{\mathbf{x}_0}(\mu) = \int_{[0,N]^2} \varphi(\mathbf{x}-\mathbf{x}_0)\big(u_1(\mathbf{x}) - u_2(\mathbf{x}+(\mu,0))\big)^2 d\mathbf{x}. \tag{4.17}$$

where $\varphi(\mathbf{x}-\mathbf{x}_0)$ is a soft window function centered at $\mathbf{x}_0$. For a sake of compactness in notation, $\varphi_{\mathbf{x}_0}(\mathbf{x})$ stands for $\varphi(\mathbf{x}-\mathbf{x}_0)$, $\int_{\varphi_{\mathbf{x}_0}} u(\mathbf{x})$ will be an abbreviation for $\int \varphi(\mathbf{x}-\mathbf{x}_0)u(\mathbf{x})d\mathbf{x}$; we will write $u(\mathbf{x}+\mu)$ for $u(\mathbf{x}+(\mu,0))$ and ε for $\varepsilon(\mathbf{x})$. The minimization problem (4.17) rewrites

$$\min_{\mu} \int_{\varphi_{\mathbf{x}_0}} \big(u(\mathbf{x}+\varepsilon(\mathbf{x})) + n_1(\mathbf{x}) - u(\mathbf{x}+\mu) - n_2(\mathbf{x}+\mu)\big)^2 d\mathbf{x}.$$

Differentiating this energy with respect to μ implies that any local minimum $\mu = \mu(\mathbf{x}_0)$ satisfies

$$\int_{\varphi_{\mathbf{x}_0}} \Big(u(\mathbf{x}+\varepsilon(\mathbf{x})) + n_1(\mathbf{x}) - u(\mathbf{x}+\mu) - n_2(\mathbf{x}+\mu)\Big) \times \Big(u_x(\mathbf{x}+\mu) + (n_2)_x(\mathbf{x}+\mu)\Big)\, d\mathbf{x} = 0. \tag{4.18}$$

One has by Taylor-Lagrange formula $u_x(\mathbf{x}+\mu) = (u(\mathbf{x}+\varepsilon))_x + O_1(\mu-\varepsilon)$, with

$$O_1(\mu-\varepsilon) \leq |\mu-\varepsilon| \max |u(\mathbf{x}+\varepsilon)_{xx}| \tag{4.19}$$

and $u(\mathbf{x}+\varepsilon(\mathbf{x})) - u(\mathbf{x}+\mu) = (u(\mathbf{x}+\varepsilon))_x(\varepsilon-\mu) + O_2((\varepsilon-\mu)^2)$, where

$$|O_2((\varepsilon-\mu)^2)| \leq \frac{1}{2} \max |(u(\mathbf{x}+\varepsilon))_{xx}| (\varepsilon-\mu)^2 .$$

Thus equation (4.18) yields

$$\int_{\varphi_{\mathbf{x}_0}} \Big((u(\mathbf{x}+\varepsilon))_x(\varepsilon-\mu) + O_2((\varepsilon-\mu)^2) + n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\Big)\times \Big((u(\mathbf{x}+\varepsilon))_x + O_1(\mu-\varepsilon) + (n_2)_x(\mathbf{x}+\mu)\Big)\, d\mathbf{x} = 0. \tag{4.20}$$

and therefore

$$\mu \int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon))_x^2 d\mathbf{x} = \int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon))_x^2\, \varepsilon(\mathbf{x})\, d\mathbf{x} + \tilde{\mathcal{A}} + \tilde{\mathcal{B}} + \mathcal{O}_1 + \mathcal{O}_2, \tag{4.21}$$

where

$$\begin{aligned}
\tilde{\mathcal{A}} &= \int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon))_x\big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\big)d\mathbf{x}; && (4.22)\\
\tilde{\mathcal{B}} &= \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\big)(n_2)_x(\mathbf{x}+\mu)d\mathbf{x}; && (4.23)\\
\mathcal{O}_1 &= \int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon))_x(\varepsilon-\mu)(n_2)_x(\mathbf{x}+\mu)d\mathbf{x} \\
&\quad + \int_{\varphi_{\mathbf{x}_0}} O_1(\mu-\varepsilon)\big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\big)d\mathbf{x}; && (4.24)\\
\mathcal{O}_2 &= \int_{\varphi_{\mathbf{x}_0}} O_2(\varepsilon-\mu)^2(u(\mathbf{x}+\varepsilon))_x d\mathbf{x} \\
&\quad + \int_{\varphi_{\mathbf{x}_0}} O_2(\varepsilon-\mu)^2[O_1(\mu-\varepsilon) + (n_2)_x(\mathbf{x}+\mu)]d\mathbf{x} \\
&\quad + \int_{\varphi_{\mathbf{x}_0}} O_1(\mu-\varepsilon)(u(\mathbf{x}+\varepsilon))_x(\varepsilon-\mu)d\mathbf{x}. && (4.25)
\end{aligned}$$

Denote by $\overline{\varepsilon}$ the average of ε on the support of $\varphi(\mathbf{x}-\mathbf{x}_0)$, denoted by $B_{\mathbf{x}_0}$. By the Taylor-Lagrange theorem we have

$$\tilde{\mathcal{A}} = \mathcal{A} + \mathcal{O}_{\mathcal{A}}\,,$$

where

$$\mathcal{A} = \int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon))_x\big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big)d\mathbf{x} \tag{4.26}$$

and

$$\mathcal{O}_{\mathcal{A}} = (\overline{\varepsilon}-\mu)\int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon))_x(n_2)_x(\mathbf{x}+\tilde{\varepsilon}(\mathbf{x}))d\mathbf{x}, \tag{4.27}$$

where $\tilde{\varepsilon}(\mathbf{x})$ satisfies $\tilde{\varepsilon}(\mathbf{x}) \in [\min(\mu,\overline{\varepsilon}), \max(\mu,\overline{\varepsilon})]$. In the same way,

$$\tilde{\mathcal{B}} = \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\mu)\big)(n_2)_x(\mathbf{x}+\mu)d\mathbf{x}\,.$$

so that $\tilde{\mathcal{B}} = \mathcal{B} + \mathcal{O}_{\mathcal{B}}$, where

$$\mathcal{B} = \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big)(n_2)_x(\mathbf{x}+\overline{\varepsilon})d\mathbf{x} \tag{4.28}$$

and

$$\mathcal{O}_{\mathcal{B}} = (\mu - \overline{\varepsilon}) \int_{\varphi_{\mathbf{x}_0}} n_1(\mathbf{x})(n_2)_{xx}(\mathbf{x} + \tilde{\varepsilon}(\mathbf{x})) - (n_2(n_2)_x)_x(\mathbf{x} + \tilde{\varepsilon}(\mathbf{x}))d\mathbf{x}. \tag{4.29}$$

The terms $\mathcal{A}$ and $\mathcal{B}$ are stochastic and we must estimate their expectation and variance. The terms $\mathcal{O}_1$, $\mathcal{O}_2$, $\mathcal{O}_{\mathcal{A}}$, $\mathcal{O}_{\mathcal{B}}$ are higher order terms with respect to $\varepsilon - \mu$ and are negligible if $\varepsilon - \mu$ is small, and the noise samples bounded.

Lemma 4 *Consider the main error terms*

$$\mathcal{A} = \int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x} + \varepsilon(\mathbf{x})))_x \big(n_1(\mathbf{x}) - n_2(\mathbf{x} + \overline{\varepsilon})\big) d\mathbf{x}$$

and

$$\mathcal{B} = \int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x} + \overline{\varepsilon})\big)(n_2)_x(\mathbf{x} + \overline{\varepsilon}) d\mathbf{x}$$

as defined above. One has $\mathbb{E}\mathcal{A} = \mathbb{E}\mathcal{B} = 0$ *and*

$$\begin{aligned}
\mathrm{Var}(\mathcal{A}) &= 2\sigma^2 \int [\varphi(\mathbf{x} - \mathbf{x}_0) u(\mathbf{x} + \varepsilon)_x]_N^2 \, d\mathbf{x} \\
&\leq 2\sigma^2 \int \varphi(\mathbf{x} - \mathbf{x}_0)^2 (u(\mathbf{x} + \varepsilon))_x^2; \\
\mathrm{Var}(\mathcal{B}) &\leq \frac{2\pi^2\sigma^4}{3} \int \varphi(\mathbf{x} - \mathbf{x}_0)^2 d\mathbf{x} + \sigma^4 \int \varphi_x(\mathbf{x} - \mathbf{x}_0)^2 d\mathbf{x}.
\end{aligned}$$

Proof: Notice that $n_1(\mathbf{x})$ and $n_2(\mathbf{x} + \overline{\varepsilon})$ are independent Gaussian noises with variance σ^2. Thus their difference is again a Gaussian noise with variance $2\sigma^2$. It therefore follows from (4.15) in the second part of Lemma 3 that

$$\mathrm{Var}(\mathcal{A}) = 2\sigma^2 \int [\varphi(\mathbf{x} - \mathbf{x}_0) u(\mathbf{x} + \varepsilon)_x)]_N^2 \, d\mathbf{x} \leq 2\sigma^2 \int \varphi(\mathbf{x} - \mathbf{x}_0)^2 (u(\mathbf{x} + \varepsilon)_x)^2 d\mathbf{x}.$$

The noises n_1 and n_2 being independent, by the second part of Lemma 3, by the second relation in Lemma 2 and by (4.14) in the first part of Lemma 3,

$$\begin{aligned}
\mathrm{Var}(\mathcal{B}) &\leq 2 \left[\mathrm{Var}\Big(\int_{\varphi_{\mathbf{x}_0}} n_1(\mathbf{x})(n_2)_x(\mathbf{x} + \overline{\varepsilon}) + \mathrm{Var}\Big(\int_{\varphi_{\mathbf{x}_0}} n_2(\mathbf{x} + \overline{\varepsilon})(n_2)_x(\mathbf{x} + \overline{\varepsilon})\Big) \right] \\
&\leq 2 \left[\sigma^2 \times \frac{\pi^2\sigma^2}{3} \int \varphi^2(\mathbf{x} - \mathbf{x}_0) + \frac{\sigma^4}{2} \int \varphi_x(\mathbf{x} - \mathbf{x}_0)^2 \right] \\
&= \frac{2\pi^2\sigma^4}{3} \int \varphi(\mathbf{x} - \mathbf{x}_0)^2 + \sigma^4 \int \varphi_x(\mathbf{x} - \mathbf{x}_0)^2.
\end{aligned}$$

Theorem 1 (Main disparity formula and exact noise error estimate) *Consider an optimal disparity* $\mu(\mathbf{x}_0)$ *obtained as any absolute minimizer of* $e_{\mathbf{x}_0}(\mu)$ *(defined by* (4.8)*). Then*

$$\mu(\mathbf{x}_0) = \frac{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x} + \varepsilon(\mathbf{x}))]_x^2 \varepsilon(\mathbf{x}) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x} + \varepsilon(\mathbf{x}))]_x^2 d\mathbf{x}} + \mathcal{E}_{\mathbf{x}_0} + \mathcal{F}_{\mathbf{x}_0} + \mathcal{O}_{\mathbf{x}_0}\,, \tag{4.30}$$

where

$$\mathcal{E}_{\mathbf{x}_0} = \frac{\int_{\varphi_{\mathbf{x}_0}} (u(\mathbf{x}+\varepsilon(\mathbf{x})))_x \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x}+\varepsilon(\mathbf{x}))]_x^2 d\mathbf{x}}$$

is the dominant noise term,

$$\mathcal{F}_{\mathbf{x}_0} = \frac{\int_{\varphi_{\mathbf{x}_0}} \big(n_1(\mathbf{x}) - n_2(\mathbf{x}+\overline{\varepsilon})\big)(n_2)_x(\mathbf{x}+\overline{\varepsilon}) d\mathbf{x}}{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x}+\varepsilon(\mathbf{x}))]_x^2 d\mathbf{x}}$$

and $\mathcal{O}_{\mathbf{x}_0}$ is made of smaller terms. In addition the variances of the main error terms due to noise satisfy

$$\mathrm{Var}(\mathcal{E}_{\mathbf{x}_0}) = 2\sigma^2 \frac{\int [\varphi(\mathbf{x}-\mathbf{x}_0) u(\mathbf{x}+\varepsilon)_x]_N^2 \, d\mathbf{x}}{\Big(\int \varphi(\mathbf{x}-\mathbf{x}_0) u(\mathbf{x}+\varepsilon)_x^2 d\mathbf{x}\Big)^2}; \tag{4.31}$$

$$\mathrm{Var}(\mathcal{F}_{\mathbf{x}_0}) \leqslant \frac{\frac{2\pi^2}{3}\sigma^4 \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x} + \sigma^4 \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2 d\mathbf{x}}{\Big(\int \varphi(\mathbf{x}-\mathbf{x}_0) u(\mathbf{x}+\varepsilon)_x^2 d\mathbf{x}\Big)^2}. \tag{4.32}$$

Finally,

$$\mathcal{O}_{\mathbf{x}_0} = \frac{\mathcal{O}_1 + \mathcal{O}_2 + \mathcal{O}_{\mathcal{A}} + \mathcal{O}_{\mathcal{B}}}{\int_{\varphi_{\mathbf{x}_0}} [u(\mathbf{x}+\varepsilon(\mathbf{x}))]_x^2 d\mathbf{x}},$$

and

$$\mathbb{E}\mathcal{O}_{\mathbf{x}_0} = O(\max_{\mathbf{x}\in B_{x_0}} |\varepsilon(\mathbf{x}) - \mu|),$$

$$\mathrm{Var}(\mathcal{O}_{\mathbf{x}_0}) = O(\max_{\mathbf{x}\in B_{x_0}} |\varepsilon(\mathbf{x}) - \mu|^2).$$

Proof: This result is an immediate consequence of (4.21) completed with the variance estimates in Lemma 4. The estimates for the higher order terms $\mathcal{O}$ are a straightforward application of Cauchy-Schwartz inequality.

Remark Theorem 1 makes sense only when the optimal disparity $\mu(\mathbf{x}_0)$ is consistent, namely satisfies for $\mathbf{x}$ in the support $B_{\mathbf{x}_0}$ of $\varphi(\mathbf{x}-\mathbf{x}_0)$,

$$|\varepsilon(\mathbf{x}) - \mu(\mathbf{x}_0)| << 1. \tag{4.33}$$

Thus, one of the main steps of block-matching must be to eliminate inconsistent matches.

Remark In all treated examples, it will be observed that $\mathrm{Var}(\mathcal{B}) \ll \mathrm{Var}(\mathcal{A})$, which by Lemma 4 directly follows from

$$\sigma^2 \left[\frac{2\pi^2}{3} \int \varphi(\mathbf{x}-\mathbf{x}_0)^2 + \int \varphi_x(\mathbf{x}-\mathbf{x}_0)^2\right] \ll 2 \int [\varphi(\mathbf{x}-\mathbf{x}_0) u(\mathbf{x}+\overline{\varepsilon})_x]_N^2. \tag{4.34}$$

Remark Theorem 1 gives us a theoretical prediction of the disparity and of its error due to noise at each point $\mathbf{x}_0$ satisfying (4.33). This requires in particular $\varepsilon(\mathbf{x})$ to be continuous at $\mathbf{x}_0$. The first term in (4.30) gives a deterministic estimate of ε, which is perfect at points around which ε is constant. The variance of the second main term $\mathcal{E}_{\mathbf{x}_0}$, given by (4.31) will be proved to be in practice almost equal to the empirically observed disparity. This fact will confirm that the formulas (4.30) and (4.31) give a full account of the block-matching disparity in stereo vision, and of its main error term.

4.3.1 Choice of the Function φ

Section 4.2 showed that the minimization of $e_{\mathbf{x}_0}(\mu)$ only requires its knowledge for $\mu \in a\mathbb{Z}/2$. The other values of $e_{\mathbf{x}_0}(\mu)$ are obtained by DFT interpolation. The 2-over-sampling of u_1 is easy by zero-padding. The one-dimensional interpolation of $e_{\mathbf{x}_0}$ is done by a numerical approximation to the DFT interpolation.
Concerning the window function φ we would like it to be both:

- *sufficiently regular* (a trigonometric polynomial of degree no larger than $2N$), in order to preserve the equality between the discrete and continuous quadratic distances (see proposition 2), so that we can make precise continuous computations on discrete samples; and
- *of small spatial support*, say a few pixels, in order to both reduce computations, and to make the distance as local as possible, thus avoiding fattening (*a.k.a.* adhesion) effects.

Since no function can have compact support both in the spatial and frequency domain both requirements are apparently contradictory, but there is a sensible solution in this case. Let us concentrate first on the small spatial support, then on the spectral support. At the end we explain how to construct a window function φ that conciliates both criteria.

φ with small spatial support. Here we relax slightly the band-limitedness assumption in favor of a small spatial support, to reduce computations and to better localize the result.

A prolate window function φ is optimal, in the sense that for a given spatial support $[-b, b]^2$ and the over-sampling factor 2, it concentrates its Fourier coefficients as much as possible in $[-N, N-1]^2$. For instance for a spatial support $b = 1.5\frac{a}{2N}$ of 3×3 half-pixels the prolate function concentrates more than 99.8% of its L^2 energy within its central $(2N)^2$ Fourier coefficients. Typical correlation window sizes are larger (at least 7×7 half-pixels), which leaves some more degrees of freedom. This parameter choice makes the discrete correlation e^d almost equal (up to a 0.2% error) to the continuous one e, in agreement with (4.9). The cost is just a 2 over-sampling, as specified in formula (4.10).

The fact that doubling the sampling rate was necessary to obtain accurate results had already been observed in [Szeliski and Scharstein, 2004], but their use of cubic interpolation and step window functions for φ limited the accuracy of their results. Exact interpolation and prolate functions have to be used to attain the twentieth of pixel. This is a crucial point: Otherwise, the resulting error is considerably higher, as shown in section 4.5.2.

φ with compact spectral support and small *discrete* spatial support. The previous choice of φ is computationally convenient but it has the disadvantage that it only approximately satisfies the hypothesis of proposition 2. Thus the equality between the computed e^d and the continuous e is only approximate (up to about 0.2% error), and so are all our error estimates, which are based on the continuous version e.

Alternatively we can take any spatial support $[-b, b]^2$, arbitrarily choose the values of φ at half-pixels

$$\varphi^d(x) = \begin{cases} f(x) & \text{if } x \in [-b, b]^2 \cap \mathbb{Z}_a^{1/2} \\ 0 & \text{otherwise} \end{cases}$$

and define φ as the $2N$-degree trigonometric polynomial interpolating those samples. Such a construction ensures the equality between discrete and continuous distances e^d and e, and

the small (discrete) spatial support allows to make computations fast. However it has the disadvantage that the continuous φ may have a large spatial support thus loosing localization of the result and potentially introducing fattening effects. This is especially true if φ^d is chosen as a box-function, thus leading to ringing artifacts when calculating the interpolated φ. However if we chose φ^d to decay smoothly to 0 near the borders of $[-b, b]^2$ then those ringing artifacts will be minimized. This is the idea of the combined solution explored next.

The final compromise In order to conciliate both criteria we shall choose the half-pixel samples $f(x)$ within the spatial support $[-b, b]^2$ of φ^d so as to minimize the L^2 energy of φ outside this spatial support. The construction is similar to the prolate window described in the previous paragraph, but inverting the roles of the Fourier and spatial domains. This way we can obtain a window function φ for which the equality $e = e^d$ is exactly true, and which has a small discrete spatial support (3×3 half-pixels), and a concentration of 99.8% of the L^2 energy of the continuous φ within this discrete spatial support, which is sufficient to avoid fattening effects beyond the size of the discrete window $[-b, b]^2$.

4.3.2 Numerical Error

In practice, the Shannon hypotheses are not completely satisfied in the interpolation of e^d. Indeed, not all of the $2N$ samples will be used for complexity reasons in this 1D interpolation. A slight accuracy loss in pixels close to edges of the image can therefore be observed in the toy example of a translated disk (see fig. 4.1). In this example, we have compared the committed error when interpolating the truncated e^d with some samples and the complete e^d with $2N$ samples. The small error committed with the truncated e^d will be neglected in the sequel, because it is much smaller than the noise error.

4.4 Discrete Correlation Algorithm

Since the quadratic distance $e_{\mathbf{x}_0}(\mu)$ may present several local minima, the algorithm for accurately finding the minimizing μ is composed of two steps:

1. Rough localization of the "correct" local minimum along the epipolar line of $\mathbf{x}_0$ within an interval of length less than one pixel.

2. Fine localization of the selected local minimum up to the desired or attainable accuracy.

The first step may is not the subject of this chapter. It uses the (AC+SS) method presented in Chapter 2. The second step is solved by an iterative quadratic fit which provides supralinear convergence with just one new interpolation of $g(\mu) = e_{\mathbf{x}_0}(\mu)$ per iteration. It consists of iteratively fitting a parabola to the current point and its two closest points among the previous iterations. The next point of the sequence is given by the analytical minimum of this parabola, and the initial iteration is performed with the endpoints and the midpoint of the input interval.

Now we turn to the critical aspect on how g is interpolated. The common approach, which consists in sampling g for integer disparities and interpolating these samples, provides a wrong result because of insufficient sampling rate. But DFT interpolation of a set of half-integer samples of g provides an *exact* interpolation, as shown in Section 4.2 (cf. proposition 3).

In practice, the spatial extent of the DFT has to be limited in order to save computational time. Here we used a DFT interpolation within an interval of length $L = 8$ around the

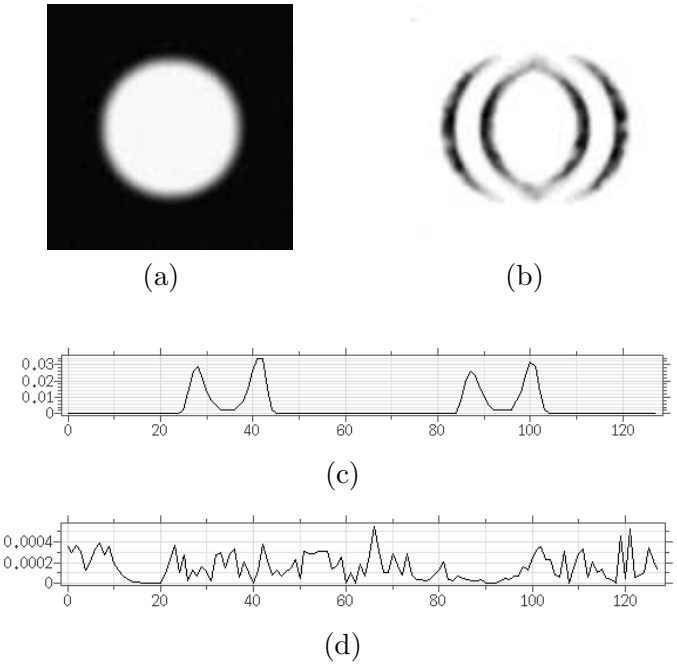


Figure 4.1: (a) Reference image. The secondary image is a trivial DFT translation of the reference. (b) Unsigned error image. Small errors appear close to the edges of the disk due to the lack of samples in the interpolation. (c) Plot of a horizontal line of the error image. The error peak is approximately of $3 * 10^{-2}$ pixels. (d) Plot of the same line error when using $2N$ samples for interpolating e^d. The error peaks close to the disk boundaries disappear and the error is smaller than $5 * 10^{-4}$ for all the pixels of the line image. However, this is more expensive computationally. The use of a function φ as explained in section 4.3.1 alleviates the numerical error due to the lack of samples. Indeed, the error doubles when a step window function φ is used, all other parameters being equal (peak of the order of $6 * 10^{-2}$).

initial search point μ_0. The computational cost of the algorithm is computed with a $W \times W$ half-pixels window size for φ ($W = 9$ in our case).

Initialization Computation of the half-integer samples $e(\mathbf{x}_0, \mu)$ for $\mu \in \mu_0 + \frac{1}{2}(\mathbb{Z} \cap [-\frac{L}{2}, \frac{L}{2}])$ requires

1. 2x zoom by zero-padding to obtain u_1 and $\tau_\mu u_2$ at the half integer scale for half-integer μ. This is done only once globally for the whole image ($(2 + 20 \log_2 N)$ flops/pixel).
2. Computation of the squared norm of the difference $\|\tau_\mu u_2 - u_1\|^2_{\varphi_\mathbf{q}}$ for each of the L samples: ($L \times 2W^2$ flops/pixel)

Total: $\underline{2LW^2 + \log_2 N^2 \text{ flops/pixel}}$

Evaluation of $e(\mathbf{x}_0, \mu)$ for a new value of $\mu \in \mu_0 + [-0.5, 0.5]$ requires a 1D Fourier translation of length L, *i.e.* $\underline{2L \log L \text{ flops/pixel/iteration}}$.

Note that we pay no penalty for each new interpolation and just a small initialization penalty with respect to the inexact version based on integer disparity sampling. On the other hand, an equally exact but brute-force solution based on image-interpolation instead of quadratic distance interpolation would transfer the burden of the initialization cost to each new evaluation of the distance function :

Initialization 2x zoom by zero-padding for u_1 and u_2 ($\underline{N^2(2 + 20 \log_2 N) \text{ flops}}$).

Evaluation of $e(\mathbf{x}_0, \mu)$ for a new value of $\mu \in \mu_0 + [-0.5, 0.5]$ requires:

1. Non-integer translation of a $W \times W$ patch of the zoomed u_1 by 1D sinc interpolation. (LW^2 flops/pixel)
2. Computation of the squared norm of the difference $\|\tau_\mu u_1 - u_2\|^2_{\varphi_\mathbf{q}}$ ($2W^2$ flops/pixel)

Total: $\underline{(2 + L)W^2 \text{ flops/pixel/iteration}}$

So, if the optimum search takes K iterations then the algorithm takes $2 + 20 \log_2 N + 2LW^2 + K \times [2L \log_2 L]$ whereas the brute force approach would take $2 + 20 \log_2 N + K \times [(2 + L)W^2]$ The previous mathematical analysis shows that the proposed method is as accurate as the brute force method, but for typical values of $W = 9$, $L = 8$ and $N = 1024$ it computes each iteration **10 times faster** at the cost of a longer initialization. For typical values of K (5 to 7) this still means a global speedup of a factor **3** which will become even larger for finer precision requirements.

4.5 Results and Evaluation

Three experiments were performed to evaluate the attainable disparity error under realistic noise conditions. The everlasting problem of such evaluations is the reference to a ground truth, that may be questionable. Two ways were found to go around this problem. The first sensible way is to simulate stereo pairs with realistic adhesion and noise features. This was done with a simulated pair of urban aerial images. Second, several simulated translations were applied to Brodatz textures, thus avoiding the adhesion problem and focusing on the

noise factor. Finally, images from the Middlebury dataset[1] were tested. In that case the noise was estimated, and the manual ground truth was actually improved by cross-validation. In all cases, the resulting performance is evaluated by the Root Mean Squared Error (RMSE) measured in pixels on all reliable points,

$$RMSE = \left(\frac{\sum_{\mathbf{q} \in M} \left(\mu(\mathbf{q}) - \varepsilon(\mathbf{q}) \right)^2}{\#M} \right)^{\frac{1}{2}},$$

where $\mu(\mathbf{q})$ is the computed disparity and $\varepsilon(\mathbf{q})$ is the ground truth value for the pixel $\mathbf{q}$ in the set of matched points M.

For the simulated cases the influence of noise in the matching process is studied with several signal to noise ratios $SNR = \frac{\| u \|_2}{\sigma_n}$, where σ_n is the standard deviation of the noise. In each case σ_n is known and the predicted noise error have been computed using the formula (4.31).

A main feature of the experimental setting is the use of a blind *a contrario* rejection method that *does not use the ground truth*. Thus, the percentage of wrong matches is also given, bad matches being those where the computed disparity differs by more than one pixel from the ground truth. As explained in the introduction, the accuracy of matches can only be evaluated on pixels that lie away from disparity edges. These being unknown, security zones were computed by dilating the strong grey level edges by the correlation window. The other pixels were matched only if they passed an *a contrario* test to ensure that the match is meaningful (see Chapter 2). These two safety filters usually keep more than half the pixels and ensure that the matched pixels are right with very high probability. For all experiments the sub-pixel refinement step goes up to $\frac{1}{64}$ pixel.

4.5.1 Simulated Stereo Pair

In order to provide the quantitative error when doing stereo sub-pixel matching, a secondary image has been simulated from a reference image and a ground truth provided by IGN (French National Geographic). In this case the resulting couple of images has a low baseline ($B/H = 0.045$) and a 25 cm/pixel resolution. Figure 4.2 shows the reference stereo image, its ground truth, the mask of matched points and the sparse disparity map. After the simulation of the secondary image a Gaussian noise has been added independently to both images. Table 4.1 gives the error committed with different noises with our algorithm. The table also gives the predicted noise error computed in the whole image, the percentage of matched pixels and the percentage of bad matches. It can be observed that the computed RMSE differs by not more than 0.01 pixel from the theoretically predicted noise error. The case without noise (SNR $= \infty$) shows the limit of the sub-pixel accuracy. These are the discretization errors.

In Appendix D a comparison between our algorithm and MARC (Multiresolution Algorithm for Refined Correlation) is done. In particular, qualitative results for this couple of simulated stereo pairs are given. MARC has been patented [Giros et al., 2004] by the French Space Agency (CNES).

[1] available on www.vision.middlebury.edu/stereo/

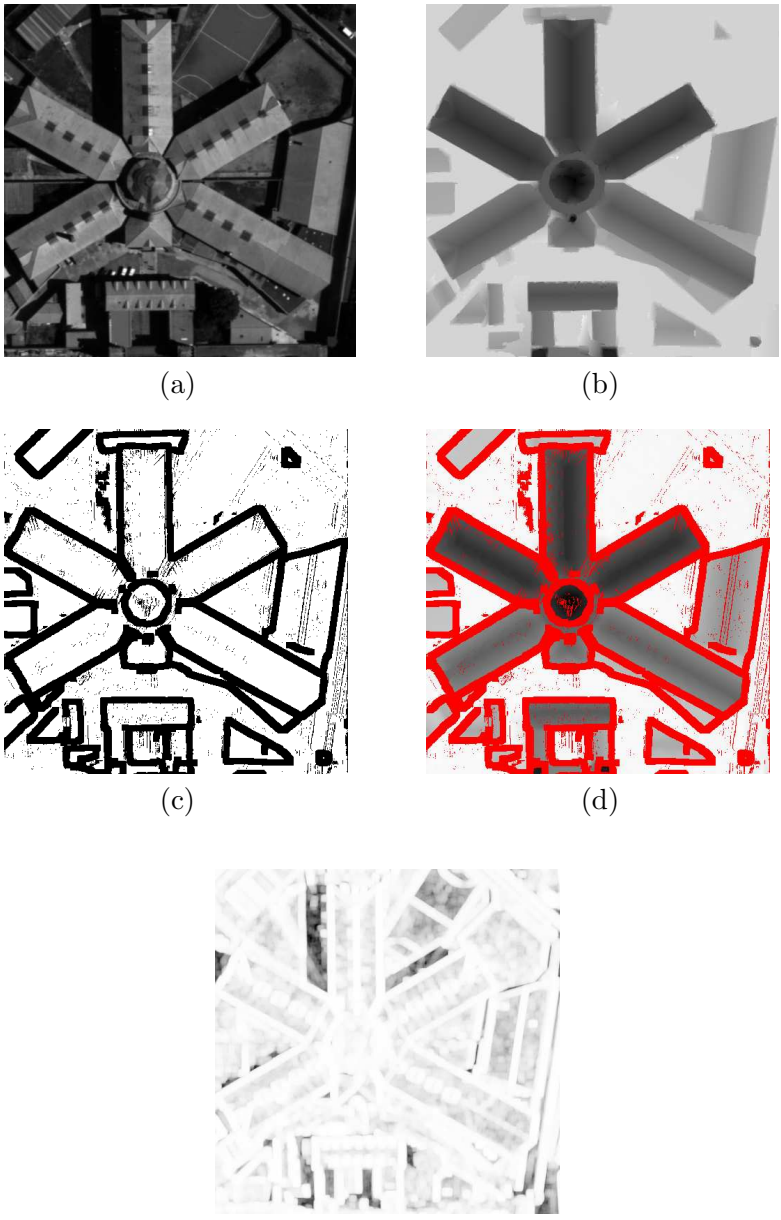


Figure 4.2: Results for the simulated stereo couple. (a) Reference aerial image. (b) Ground truth. (c) Mask of matched points (in white, 70.6% of matched points). Statistics are computed on the white points. (c) Obtained sparse disparity map. (e) Noise error prediction at each point. The darker the pixel, the higher the predicted error.

SNR	Predicted noise error	RMSE	Matched points (density)	Bad matches
∞	**0**	**0.023**	70.6%	0.00%
357.32	**0.029**	**0.033**	63.3%	0.00%
178.66	**0.041**	**0.049**	54.2%	0.01%
125.06	**0.052**	**0.058**	41.5%	0.02%

Table 4.1: Qualitative results for the simulated stereo couple (fig. 4.2). From left to right: signal to noise ratio; RMSE (in pixels) predicted by the theory; RMSE to ground truth (in pixels); percentage of matched points and percentage of bad matches.

4.5.2 Matching Textured Images

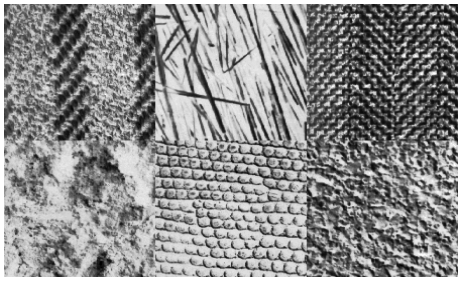
This experiment simulates the ideal case of two textured images (figure 4.3-(a)) obtained from each other by a 2.5 pixels translation using zero-padding. An independent Gaussian noise has been added independently to both images. Again, the observed RMSE turns out to be very close to the predicted noise error. The same study was also performed directly on a 1D signal (fig. 4.3-(b)), with a one-dimensional comparison window. For several textured images and signals the results were remarkably similar (see Table 4.2).

The very same test was led with cubic interpolation as proposed in [Szeliski and Scharstein, 2004] instead of the exact DFT interpolating method. The match of two textured images *without noise* had a RMSE of 0.24 instead of 0.0053. This test shows how badly a wrong interpolation decision can hem the stereo technology.

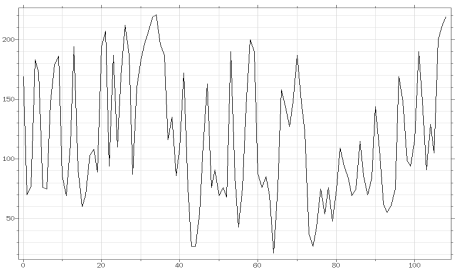
Table 4.3 summarizes the orders of magnitude of the terms in our main error formula (4.21) for the images in figures 4.2 and 4.3. For these figures we know exactly the ground truth and the standard deviation σ of the added noise. First, the standard deviation considering $\mathcal{E}_{\mathbf{x}_0}$ and $\mathcal{E}_{\mathbf{x}_0} + \mathcal{F}_{\mathbf{x}_0}$ has been computed (R_E and R_{E+F} respectively) where $\mathrm{Var}(\mathcal{E}_{\mathbf{x}_0} + \mathcal{F}_{\mathbf{x}_0})$ has been upper bounded by $\mathrm{Var}(\mathcal{E}_{\mathbf{x}_0}) + \mathrm{Var}(F_{\mathbf{x}_0}) + 2(\mathrm{Var}\,\mathcal{E}_{\mathbf{x}_0}\,\mathrm{Var}\,\mathcal{F}_{\mathbf{x}_0})^{1/2}$ in the computation of R_{E+F}. This table confirms that the formula (4.21) scales correctly the orders of magnitude, and that the main error term is due to noise.

4.5.3 Middlebury Images

The last experiments were done on the Middlebury classic dataset, which also publishes a hand-made ground truth. The first image used is Sawtooth which is one of the images of the initial Middlebury benchmark. This image is piecewise planar. Table 4.4 gives (column R_0) a 20/100 pixel distance to the ground truth. Dequantizing the Middlebury ground truth (column R_1) improves slightly this distance to ground truth to roughly 16/100. Still, with comparable noise level, this distance is thrice the error in the simulated experiments! *A closer analysis of the results, however, shows that the real error is close to* 9/100 *pixel.* Indeed, the manual ground truth in Middlebury is NOT sub-pixel accurate: As explained in the Middlebury web site, is in fact a quantized ground truth obtained from the estimation of the affine motion of each planar and hand labeled component of the image. *Yet, a more faithful ground truth can be actually recovered from the image pair itself.* Indeed, assuming that the data set was accurately piecewise planar permits to compute the error between the subpixel matching result and its



(a)



(b)

Figure 4.3: Brodatz texture and signal texture.

corresponding plane-fit. The standard deviation of this error goes down to 9/100 respectively (see column R_2). An independent error estimate of the obtained disparity map (not relying on the ground truth) can be obtained by cross-validating the disparity measurements applied to several different stereo pairs of the same 3D scene. Indeed, the Middlebury data set provides nine ortho-rectified views at uniform intervals, so that disparity maps taking the central view as reference image are related to one another by a scaling constant which depends on the baseline ratios. The RMSE error between the scaled disparity maps (see column R_3) turns out to be in full agreement with the piecewise planar check (column R_2). The predicted noise error was computed by using an estimation of the standard deviation of the noise of the image given by [Buades et al., 2008].

The above Sawtooth test demonstrates the (relatively) poor accuracy of the ground truth. In consequence, for the current four pairs of images in the benchmark, we have decided to cross-validate our results by using all the images of each scene in the dataset (5 images for Tsukuba and 9 images for Venus, Teddy and Cones). Table 4.5 compares the obtained RMSE by cross-validation with the predicted theoretical noise error. Figure 4.4 shows the Teddy and Cones results. (see Figure 3.5 in Chapter 3 for Sawtooth, Tsukuba and Venus results).

Comparison of existing algorithms in our mask

For the sake of evidence of the ground truth imperfection we have checked that classic stereo algorithms provide disparity maps that are closer to each other (and to our result) than to the ground truth. The tested algorithms are actually at the top of the Middlebury evaluation table: AdaptingBP [Klaus and Sormann, 2006], CoopRegion [Wang and Zheng, 2008], SubPixDoubleBP [Yang et al., 2007], CSemiGlob [Hirschmuller, 2006] and GC+SegmBorder [Chen et al., 2009].

SNR	Predicted noise error	RMSE	Matches	Bad matches
∞	**0**	**0.0053**	100%	0.0%
96.38	**0.0048**	**0.0073**	99.8%	0.0%
48.19	**0.0096**	**0.0109**	99.8%	0.0%
32.12	**0.0141**	**0.0160**	98.7%	0.0%
24.09	**0.0192**	**0.0203**	87.1%	0.0%

SNR	RMSE	Matches	Bad matches
∞	0.0113	100%	0.0%
96.38	0.0149	100%	0.0%
48.19	0.0241	100%	0.0%
32.12	0.0357	100%	0.0%
24.09	0.0422	100%	0.0%

Table 4.2: Qualitative results for textures (fig. 4.3). Top: Table of results for the textured images. Bottom: Table of results for the signal. From left to right: signal to noise ratio; RMSE (theoretical prediction); observed RMSE (in pixels); percentage of matched points, percentage of wrong matches.

SNR	R_E	R_{E+F}	R_{O_1}
∞	0	0	0
357.32	0.029	0.030	0.0005
178.66	0.041	0.044	0.0009
125.06	0.052	0.053	0.0011
∞	0	0	0
96.38	0.0048	0.0051	0.0008
48.19	0.0096	0.0103	0.0010
32.12	0.0141	0.0145	0.0013
24.09	0.0192	0.0193	0.0014

Table 4.3: Order of magnitude of the terms in formula (4.21). Top of the table: simulated stereo couple (fig. 4.2). Bottom of the table: Textures (Fig. 4.3). From left to right: Signal to noise ratio; R_E predicted noise error computed from $\mathcal{E}_{\mathbf{x}_0}$; R_{E+F} error from $\mathcal{E}_{\mathbf{x}_0} + \mathcal{F}_{\mathbf{x}_0}$. R_{O_1}: the explicit computation of O_1 using the ground truth ε. The contribution of $\mathcal{F}_{\mathbf{x}_0}$ in R_{E+F} is negligible and R_{O_1} is of the order of a thousandth of pixel.

Table 4.6 gives the quadratic errors when comparing two by two the considered algorithms. The values in the diagonal of the tables (gray) are the RMSE with respect to the ground truth. All of these error values have been computed in our reliable mask of valid points. The distance between any two solutions is for most of the cases smaller than the distance of these solutions to the ground truth. The only exception of the 6 tested algorithms is GC+SegmBorder.

	w.r.t. ground truth		cross-validation		Predicted noise error
	R_0	R_1	R_2	R_3	
Sawtooth	0.213	0.162	0.09	0.090	0.076

Table 4.4: From left to right: RMSE with "official" ground truth. RMSE with the plane-fit of the official ground truth. RMSE with the plane-fit of our results. RMSE of cross-validation with 7 additional views. Finally, predicted disparity error due to noise (using an accurate noise estimate on the pictures themselves).

	w.r.t. ground truth	cross-validation	Predicted Noise Error
Tsukuba	0.357	0.080	0.069
Venus	0.225	0.101	0.042
Teddy	0.424	0.093	0.072
Cones	0.319	0.082	0.066

Table 4.5: Quantitative results for the Tsukuba, Venus, Teddy and Cones images. The first column corresponds to the RMSE to the ground truth computed in the mask of valid points. The second column is the RMSE by cross-validation of the 5 or 9 images in the dataset. Finally, the noise error predicted by the theory appears in the last column.

Figure 4.4: From left to right: Reference and secondary images, estimated sparse disparity map and ground truth. On top: Teddy. On bottom: Cones.

4.5.4 Conclusion

The empirical sub-pixel accuracy in stereo vision can attain its predicted limit, which only depends on the noise at regular disparity points. The experiments on realistically simulated pairs and real benchmark images show a 1/20 pixel accuracy to be attained by block-matching, for more than half the image points. These image points are not found *a posteriori*, they are specified *a priori* by an autonomous algorithm. The two features, namely subpixel accuracy

IMAGES	Our Algorithm	AdaptingBP	CoopRegion	SubPixDoubleBP	CSemiGlob	GC+SegmBorder	ALGORITHMS
Tsukuba	0.357	0.281	0.258	0.245	0.216	0.223	Our Algorithm
		0.297	0.251	0.289	0.253	0.300	AdaptingBP
			0.337	0.241	0.214	0.241	CoopRegion
				0.264	0.253	0.272	SubPixDoubleBP
					0.275	0.272	CSemiGlob
						0.207	GC+SegmBorder
Venus	0.225	0.231	0.130	0.143	0.161	0.239	Our Algorithm
		0.215	0.146	0.163	0.176	0.223	AdaptingBP
			0.131	0.122	0.129	0.142	CoopRegion
				0.162	0.148	0.174	SubPixDoubleBP
					0.192	0.212	CSemiGlob
						0.142	GC+SegmBorder
Teddy	0.424	0.341	0.336	0.346	0.352	0.511	Our Algorithm
		0.421	0.330	0.312	0.303	0.531	AdaptingBP
			0.354	0.262	0.317	0.509	CoopRegion
				0.385	0.291	0.519	SubPixDoubleBP
					0.411	0.542	CSemiGlob
						0.481	GC+SegmBorder
Cones	0.319	0.281	0.267	0.245	0.253	0.421	Our Algorithm
		0.331	0.262	0.301	0.319	0.483	AdaptingBP
			0.272	0.252	0.262	0.446	CoopRegion
				0.349	0.234	0.487	SubPixDoubleBP
					0.365	0.510	CSemiGlob
						0.400	GC+SegmBorder

Table 4.6: Comparison of several stereo algorithms in the top classification. Values on the diagonals (gray) are the values with respect to the ground truth. In general the distance between two solutions is smaller than the distance to the ground truth, with the exception of GC+SegmBorder.

and wrong match control, make stereo-vision into a highly accurate 3D tool, potentially competitive with laser range scanners.

The above Middlebury experiments also showed that the ground truth provided as a benchmark reference is actually less accurate than the attainable level of 1/20 pixel (see [Yang et al., 2007] for similar conclusions). A rigorous methodology to create reliable ground truths is needed. Such ground truths should be built up by automatic devices used repeatedly on the same objects, so as to provide a cross-validated estimate of their own accuracy. Luckily enough, the attainable accuracy in the Middlebury data could be recovered indirectly through an additional set of views for cross-validation.

Chapter 5

Algorithm Synopsis

Contents

Résumé : Le but de ce chapitre est de présenter, étape par étape, l'algorithme entier de calcul de disparités entre deux images.

Abstract: The aim of this chapter is to present the algorithm step by step from which the disparity map between two stereo images is computed.

5.1 Major Parts of the Algorithm

Let us start by enumerate the major computation stages:

1. Building the *a contrario model*:
   (a) Create classes of similar blocks (same variance, same average grey level) among all admissible blocks (section 2.3.2).
   (b) Make PCA on each block class and obtain the empirical probability distributions of the PCA coefficients (section 2.2.1).
2. *A contrario* matching:
   (a) For each block in the left image, and each block on the corresponding epipolar line in the right image, find the meaningful matches and, if any, select the most meaningful one whenever it is unique (section 2.3).
   (b) Apply the self-similarity rejection step to avoid periodical structures (section 2.4).
3. Refinement:
   (a) Compute subpixel disparities for accepted reliable matches. (Chapter 4).
4. Fattening Correction: (Section 3.3)
   (a) Compute the corrected disparity map via gradient matching and compute fattening risk edges.
   (b) Compute the final disparity map: reject any patch meeting fattening risk edges.
5. Optional completion: (Appendix E)
   (a) For all pixels contained in at least one block that matched and no touching a fattening risk edge, compute the median value of all disparities of all such blocks.
   (b) Disparity map interpolation.

Steps 1-(b) and 2-(a) are the essential parts of the algorithm and contain the core ideas.

5.2 Pseudocode

This section contains the pseudocode of the main algorithms presented in this thesis. Let us recall some notations. Let $(G_{j\,k})_{j,k}$ and $(G'_{j\,k})_{j,k}$ be the previous coarse partition of the stereo pair of images defined in Def. 7 (Chapter 2) with $\alpha = 0.3$ and, let $T = 2$ be the number of regions of the mean and variance partitions. Let s be the number of pixels considered in each block. Recall that only $N < s$ features are considered and Q probability thresholds.

Algorithm 1 contains the training step of the main algorithm. Algorithm 2 computes the empiric probabilities. Algorithm 3 gives details about the search of meaningful matches. Finally, algorithm 4 corrects the fattening phenomenon.

Algorithm 1: Compute Coefficients

input : A couple of images I and I',
Partitions of each image $(G_{j\,k})_{j,k}$ and $(G'_{j\,k})_{j,k}$.
output: PCA Coefficients $(c_i(\mathbf{q}))_{i,\mathbf{q}}$.

foreach *class* $j, k = 1, \cdots, T$ *of the coarse partition* **do**
  Allocate two matrices $X_{G_{j\,k}}$ and $X_{G'_{j\,k}}$ of sizes $s \times \#G_{j\,k}$ and $s \times \#G'_{j\,k}$;
  foreach *pixel* $\mathbf{q} \in G = \{G_{j\,k}, G'_{j\,k}\}$ **do**
    Write the intensities of $B_{\mathbf{q}}$ as a column of X_G,
  end
  Compute the covariance matrix $C_{j\,k} = \mathrm{Cov}(X_{G_{j\,k}})$;
  Compute the eigenvectors $(v_1^{j\,k}, \ldots, v_s^{j\,k})$ of $C_{j\,k}$;
  Store eigenvectors in the rows of the $s \times s$ matrix $V_{j\,k}$;
  Compute the projections in the new basis of eigenvectors: $W_{G_{j\,k}} = V_{j\,k}^T \cdot X_{G_{j\,k}}$ and $W_{G'_{j\,k}} = V_{j\,k}^T \cdot X_{G'_{j\,k}}$;
  foreach *pixel* $\mathbf{q} \in G = \{G_{j\,k}, G'_{j\,k}\}$ *and* $i = 1$ *to* s **do**
    $c_i(\mathbf{q}) \leftarrow W_G(i, \mathbf{q})$;
  end
end
Return $(c_i(\mathbf{q}))_{i,\mathbf{q}}$;

Algorithm 2: COMPUTE PROBABILITIES

input : PCA sorted coefficients $(o_i^{j\,k}(\mathbf{q}))_{i,\mathbf{q}}$ of I' and coefficients $(c_i(\mathbf{q}))_{i,\mathbf{q}}$ of I;
Partitions of I and I': $G_{j\,k}$ and $G'_{j\,k}$.

output: Quantized non-decreasing probabilities $(p^i_{\mathbf{q}\,\mathbf{q}'})_{i,\mathbf{q},\mathbf{q}'}$

foreach *class* $j,k = 1,\cdots,T$ *of the coarse partition* **do**
  $NB_{pix} \leftarrow \#G'_{j\,k}$;
  foreach *pixel* $\mathbf{q} \in G_{j\,k}$ **do**
    Find the N best coefficients for $\mathbf{q}$: $|c_1(\mathbf{q})| > |c_2(\mathbf{q})| > \ldots > |c_N(\mathbf{q})|$;
    for $i = 1,\cdots,N$ **do**
      $a \leftarrow 0$, $b \leftarrow NB_{pix}$;
      while $b - a > 1$ **do**
        $ind = (a+b)/2$;
        if $o_i^{j\,k}(ind) < c_i(\mathbf{q})$ **then** $a \leftarrow ind$;
        else $b \leftarrow ind$;
      end
      foreach *pixel* $\mathbf{q}' \in S' \cap G'_{j\,k}$ **do**
        $ind' \leftarrow Ind$, such that $o_i^{j\,k}(Ind) = c_i(\mathbf{q}')$;
        if $(NB_{pix} - ind) < |ind - ind'|$ **then**
          $\widehat{p^i}_{\mathbf{q}\,\mathbf{q}'} \leftarrow (NB_{pix} - ind')/NB_{pix}$;
        else if $ind < |ind - ind'|$ **then**
          $\widehat{p^i}_{\mathbf{q}\,\mathbf{q}'} \leftarrow ind'/NB_{pix}$;
        else $\widehat{p^i}_{\mathbf{q}\,\mathbf{q}'} \leftarrow 2 \cdot |ind - ind'|/NB_{pix}$;
        $j \leftarrow Q$;
        while $\widehat{p^i}_{\mathbf{q}\,\mathbf{q}'} > \pi_j$ **do** $j--$;
        if $i > 1$ **then** $p^i_{\mathbf{q}\,\mathbf{q}'} \leftarrow MAX(\pi_j, p^{i-1}_{\mathbf{q}\,\mathbf{q}'})$;
        else $p^i_{\mathbf{q}\,\mathbf{q}'} \leftarrow \pi_j$;
      end
    end
  end
end
Return $(p^i_{\mathbf{q}\,\mathbf{q}'})_{i,\mathbf{q},\mathbf{q}'}$;

Algorithm 3: MEANINGFUL MATCHES

input : Quantized non-decreasing probabilities $(p^i_{\mathbf{q}\,\mathbf{q}'})_{i,\mathbf{q},\mathbf{q}'}$
output: Disparities for the meaningful matches.

foreach *pixel* $\mathbf{q} \in G_{j\,k}$, $j,k = 1,\cdots,T$ **do**
  forall $\mathbf{q}' \in S' \cap G'_{j\,k}$ **do** $NFA_{\mathbf{q}\,\mathbf{q}'} \leftarrow N_{test} \cdot \prod_{i=1}^{N} p^i_{\mathbf{q}\,\mathbf{q}'}$;
  $min \leftarrow \min_{\mathbf{q}'} NFA_{\mathbf{q},\mathbf{q}'}$;
  $s' \leftarrow \arg\min_{\mathbf{q}'} NFA_{\mathbf{q},\mathbf{q}'}$;
  if $min \leqslant \varepsilon$ **and** $\nexists\, \mathbf{r}' \in S' \cap G'_{j\,k}$ *such that* $NFA_{\mathbf{q},\mathbf{r}'} = min$ **then**
    $\mu(\mathbf{q}) = q_1 - s'_1$;
  else
    $\mu(\mathbf{q}) = \emptyset$;
  end
end
Return μ;

Algorithm 4: FATTENING CORRECTION

input : Refined disparity map μ.
output: Final disparity map μ_F.

foreach $\mathbf{q} \in I$ **do** Compute $\mu_m(\mathbf{q}) = Med\{\mu(\mathbf{y}) \mid \mathbf{y} \in B_{\mathbf{q}},\ \mu(\mathbf{y}) \neq \emptyset\}$;
Compute the disparity map $\tilde{\mu}$ and Ω_1 (Definition 3.1);
Compute $\Omega = \Omega_1 \cup \Omega_2 \cup \Omega_3$ (see definitions in Section (3.3));
Sweep the image from left to right: **foreach** $\mathbf{q} \in \Omega$ **do**
  compute the dilation direction;
  $D_1 \leftarrow$ (W= size patch) pixels in the dilation direction;
end
Sweep the image from top to bottom, and add the W pixels in the dilation direction to D_2;
Compute Canny-Deriche edges $\gamma \subseteq D(\Omega) = D_1 \cup D_2$;
Compute extreme points of γ;
foreach $\mathbf{q} \in \gamma$, $\mathbf{q}$ = extreme point of γ **do**
  while $\exists \mathbf{r} \notin \gamma, \quad ||\mathbf{q} - \mathbf{r}|| \leqslant 1, \quad \mathbf{r} \in$ *C.-D. edges* , $\quad |\max_{\mathbf{x}\in B_{\mathbf{r}}}\{\mu(\mathbf{x})\} - \min_{\mathbf{x}\in B_{\mathbf{r}}}\{\mu(\mathbf{x})\}| > \theta$ **do**
    $\gamma \leftarrow \mathbf{r}$;
    $\mathbf{r}$ = extreme point of γ;
  end
end
foreach $\mathbf{q} \in I$ **do**
  if $B_{\mathbf{q}} \cap \gamma \neq \emptyset$ **or** $\mathbf{q} \in D(\Omega)$ **then**
    $\mu_F(\mathbf{q}) = \emptyset$;
  else
    $\mu_F(\mathbf{q}) = \mu(\mathbf{q})$;
  end
end
Return μ_F;

Chapter 6

Experiments

Contents

Résumé : Dans ce chapitre, on donne plusieurs résultats de l'algorithme présenté dans cette thèse. D'abord, on a considéré des images publiques de la NASA de Mars. Ensuite, on a eu accès à des videos filmées depuis un hélicoptère à Los Angles et fournies par l'Office of Naval Research (USA), et enfin des couples d'images à faible B/H fournies par le CNES et l'IGN (PELICAN). Enfin, on a pris des images d'une statue avec un appareil réflex.

Abstract: In this chapter, we give some results obtained with the algorithm presented in this thesis. First, we have considered public NASA images of Mars. Then, we have used movies filmed from an helicopter in Los Angeles, courtesy of Office of Naval Research (USA) and low B/H pairs of images provided by the CNES-IGN(PELICAN). Finally, we have taken images of a statue with a reflex camera.

6.1 Mars' Images

The context of these experiments is a collaboration with a geophysical research team directed by A. Mangeney in the "Institut de Physique du Globe de Paris" (IPGP).

Mangeney's team is dedicated to the modeling of geophysical flows in general. In particular, it works on the planet Mars landslides, debris flows, and gullies. Mars has a weaker geological activity than the Earth. Its remote situation makes its study difficult. For the time being, the most reliable topographic data on Mars are due to a laser altimeter, the Mars Orbiter Laser Altimeter (MOLA) carried by a NASA spacecraft. This instrument is out-of-order and the collected data from all missions remain insufficient for the study. The resolution of MOLA is only $0.23 - 11.8$ Km. depending on the latitude.

In this context, the stereoscopy seems to be the best option to obtain Mars DEM.

Data

A couple of a NASA public stereo images from planet Mars obtained with the Context Imager (CTX) have been considered (see Fig. 6.1). From 400 kilometers above Mars, the Mars Reconnaissance Orbiter (MRO) Context Camera (CTX) is designed to obtain gray scale images of Mars at a 6 meters per pixel resolution over a 30 kilometers wide swath.

We have used images satisfying the epipolar constraints (see Fig. 6.2) after being rectified. Nevertheless, the images are in no case ortho-rectified, meaning that the images have not been re-sampled to simulate a nadir acquisition. Thence, there is a tilt between the images and, the resulting 3D relief information will correspond to the reality up to a projective transformation.

Furthermore, we know that the 8-bit images have been post-processed after acquisition (noise, vignetting and offset correction and contrast enhancing). This processing is a black box for us and could explain the presence of saturated pixels or the existing contrast change between the images.

Results

First of all, we have used the Midway equalization histogram algorithm [Delon, 2004b] in order to reduce the contrast change between the images. Then, the meaningful matches in the couple of stereo images have been computed in a symmetrical way and their coherence has been checked. This means that two disparity maps have been computed considering each image as reference and pixels having different matches have been rejected. Usually, the

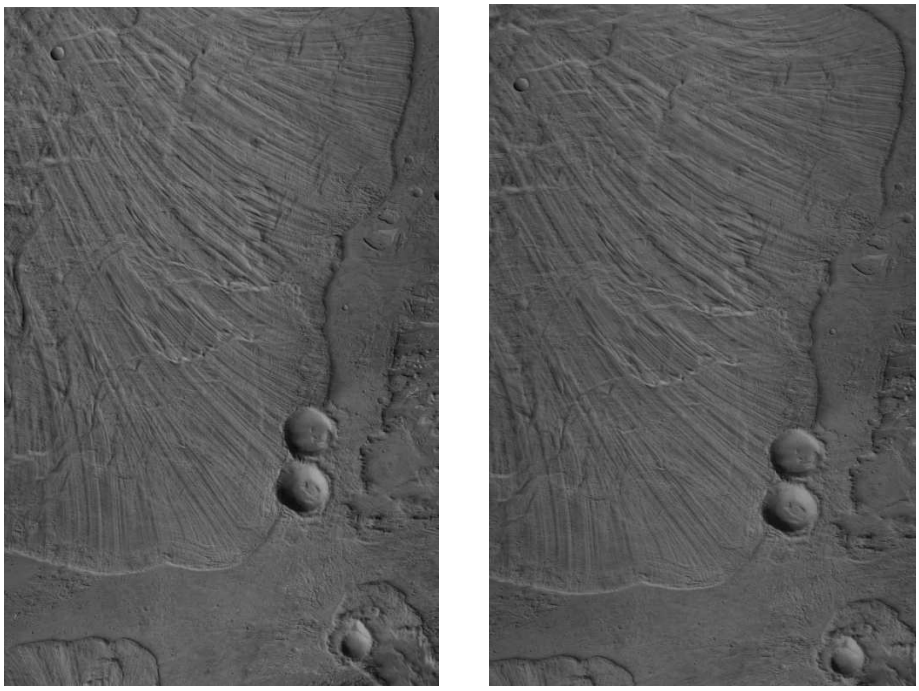


Figure 6.1: Original Images P22.levleo.png and P20.levleo.png. High-resolution images provided by NASA. The used camera CTX is able to do observations with a 6 meters resolution per pixel. The direction of lighthing has changed, thus altering the aspect of the details.

symmetry coherence is not checked because of the double complexity, but taking into account the characteristics of these images, this avoids false matches.

The resulting subpixel and sparse disparity map can be seen in Fig. 6.3. Our disparity map has a density of 61.2%. The rejected pixels (Fig. 6.4) stand mainly in the crater walls. The normal vectors of these surfaces are almost perpendicular to the camera view direction. In such a case, the big local tilt between the patches makes a meaningful match impossible. Besides tilts, there are other rejection causes such as the presence of poorly textured regions, and the fact that local radiometric changes persist after the Midway equalization.

After the application of a median filter the density rises to 79.7%. For a completely dense disparity map, an interpolation can be done (see appendix E for more details). Figure 6.5 shows the two obtained images.

As a matter of fact, the computed disparity map for this couple of stereo images has an important slope (Fig. 6.6). The obtained disparity map only provides the relative depth of the points in the scene. The images should be ortho-rectified in order to remove the tilt and obtain the absolute DEM.

Future Work

The resulting disparity map with our algorithm is a sparse disparity map, but for the purpose of this collaboration a completely dense map is necessary. We have seen that a median filter can densify the results and an interpolation of such a disparity map is possible. In a way, the goal has been achieved but the interpolation can be improved. This is why we have

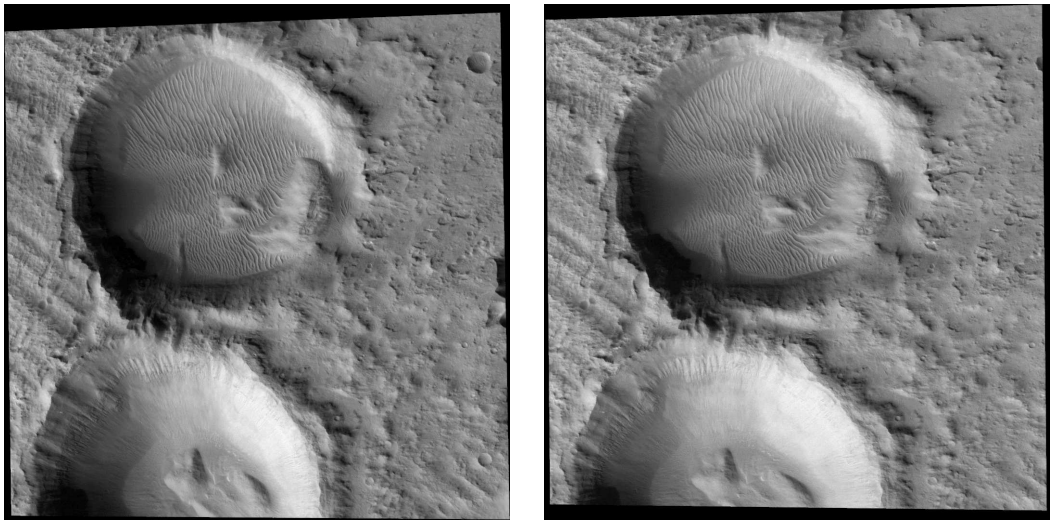


Figure 6.2: Rectified images. A 1024×1024 crop has been extracted from P20.levleo.png and P22.levleo.png and a contrast change has been applied. These images satisfy the epipolar constraint.

contemplated to combine stereoscopy and photometric stereo in a collaboration with J.-D. Durou.

Photometric stereo consists in estimating the surface normals by observing a scene under different lighting (see [Durou et al., 2009] and [Durou and Courteille, 2007]). Photometric stereo relies on several assumptions: invariant albedo, non deformable scenes, Lambertian reflectance of the surfaces, and constant internal and external camera parameters between the snapshots. This last hypothesis is not satisfied, since the available images are pairs of stereo images with different view angles. However, photometric stereo can be used for points having a correspondence via stereo matching. Thus, a new interpolation problem is considered: find the best surface through the 3D points and surface normals. Photometric stereo needs at least 3 images to estimate surface normals. Otherwise, there is an ambiguity in the normal direction. Luckily, more and more Mars NASA images are public, and having 3 images of the same area will not be a problem.

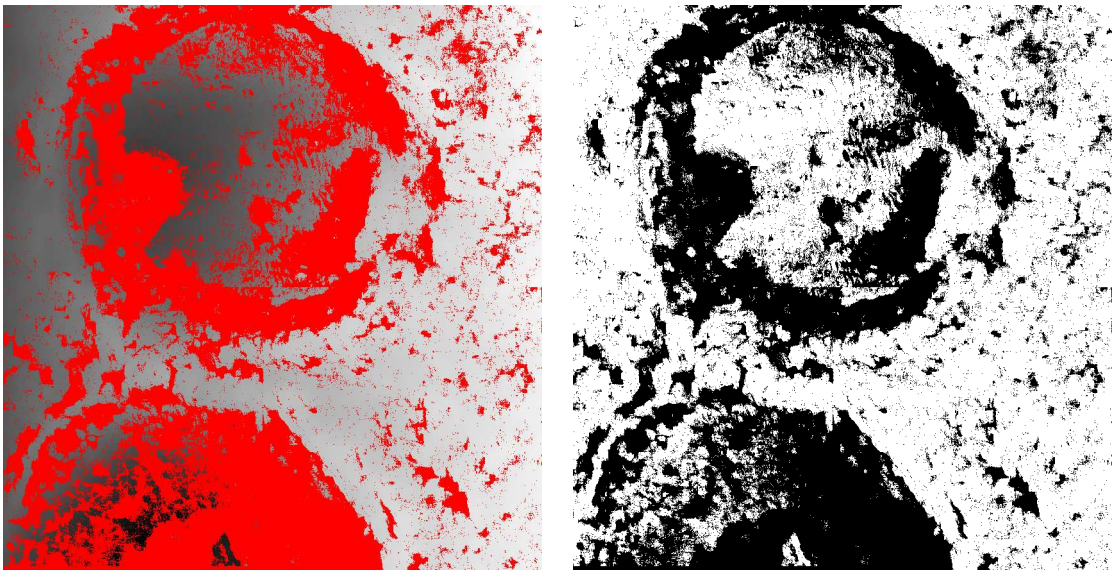
Finally, the use of shape-from-shading and shape-from-shadow techniques can densify the normal field and give reliable results in shadows. The use of both together, shape-from-shading and shape-from-shadow, has been studied by [Schlüns, 1997].

6.2 L.A. Videos

The context of these experiments is a work for Cognitech, Inc, a californian company specialized in video analysis.

Data

The two datasets are video sequences of Los Angeles filmed from an helicopter (Fig. 6.7 and Fig. 6.9) furnished by ONR. In the scenes there are several skyscrapers. Notice that in the first set of images there is an important blur and in the second set the skyscraper façade appears gradually (the images are not taken at nadir). Building façades are a tricky



(a) (b)

Figure 6.3: (a) Disparity map. The red pixels are rejected matches. For the other pixels, the darker the pixel the deeper the point in the scene. (b) Mask of meaningful matches associated to the disparity map. The white points are accepted as meaningful (61.2% density).

visual object in aerial stereovision, because, being roughly mirrors, they can change aspect completely with a viewpoint change.

Results

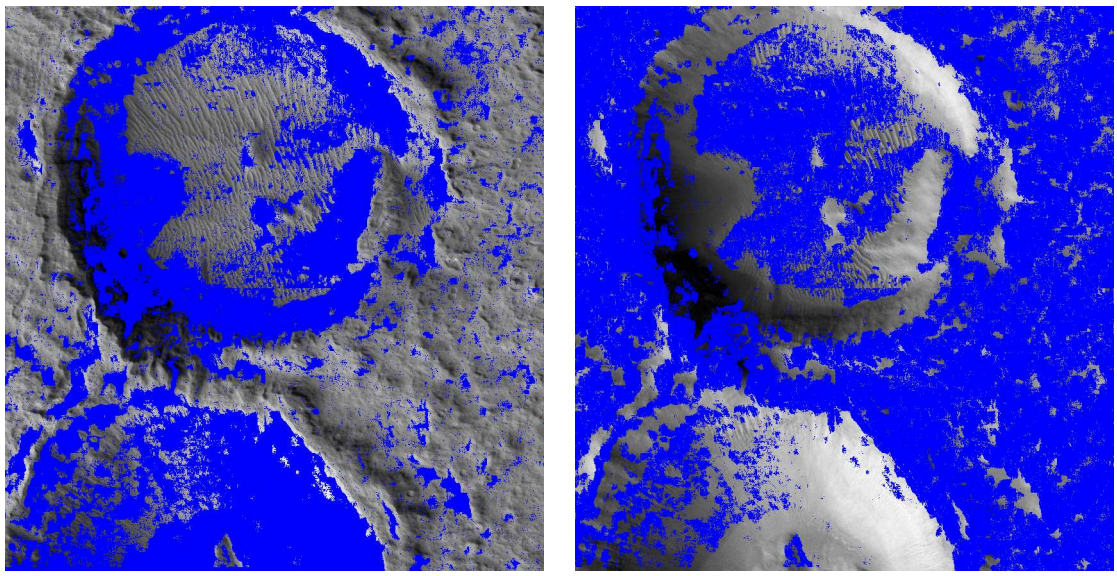
When several images of the same scene are available, as a sequence of frames of a movie, several disparity maps can be computed. Taking the central frame as reference image, a disparity map is computed taking as secondary image each one of the resting images. Then the set of disparity maps can be merged.

Unluckily, the epipolarity in this data set is not very accurate. In principle, the helicopter should have a straight trajectory, but in practice is serpentine. We do not even know if the velocity and the altitude of the helicopter are constant. All of these reasons make difficult the matching process. Regardless of the characteristics of such a data set we have sought to take advantage of the fact that several images of the same scene are available. Figures 6.8 and 6.10 show the result of our algorithm and the resulting disparity map after median filter and interpolation.

6.3 PELICAN Images

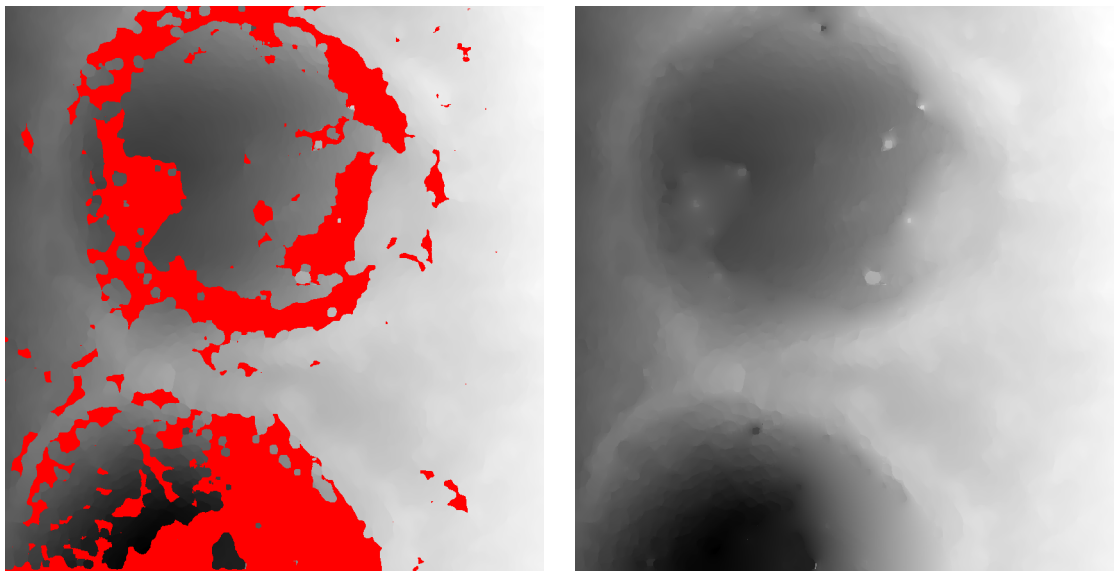
Data

PELICAN images are IGN images acquired by CNES. We got access to these data thanks to a collaboration agreement between the CMLA and the CNES (MISS Project). Several pairs of rectified stereo aerial images are available, with small baseline.



(a) (b)

Figure 6.4: (a) Accepted pixels are displayed with the gray level in the image and the rejected ones are in blue. Accepted pixels are inside textured regions. (b) Rejected pixels are displayed with the gray level in the image and the accepted ones are in blue. Rejected pixels are mainly in uniform and poor textured regions and crater walls where the surface is far from being fronto-parallel.



(a) (b)

Figure 6.5: (a) Disparity map obtained using a spatial median filter in the disparity map 6.3. The red pixels are the rejected matches. For the other pixels, the darker the pixel the deeper the point in the scene. (disparity 79.7%) (b) Interpolation of the spatial median disparity map.

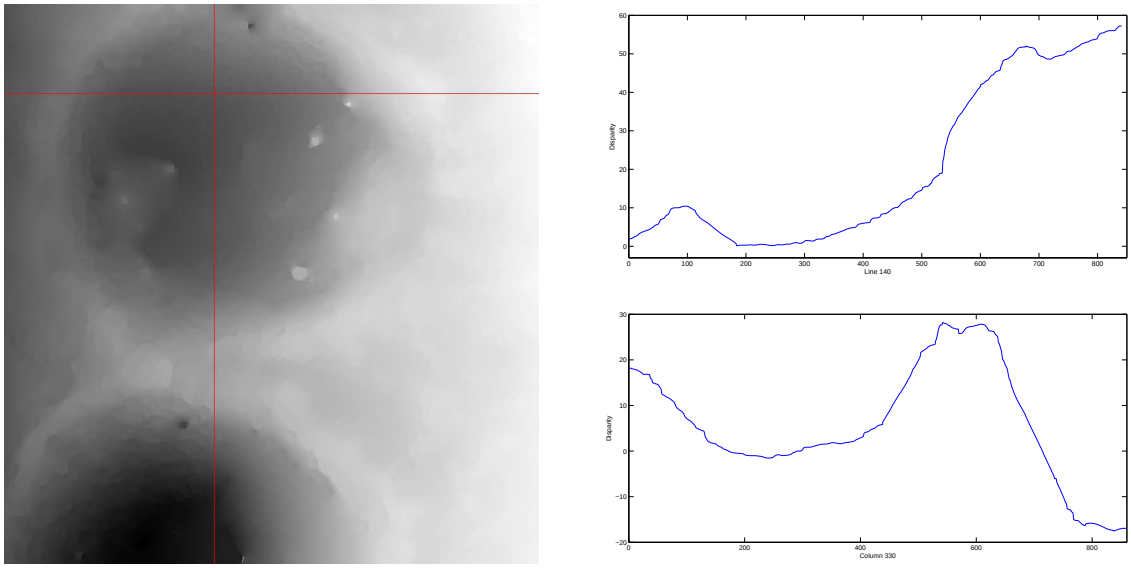


Figure 6.6: On the left, interpolated disparity map. On the top-right, plot of disparities of the red line. On the bottom-right, plot of disparities of the red column.

Results

Figure 6.11 shows a first stereo pair of images that we considered. The obtained disparity map is quite dense, but shadows have been mainly rejected. In this scene, many pedestrians and cars have advanced several meters in the seconds elapsed between the snapshots. Our algorithm has rejected all matches for these pixels (see Fig. 6.12). 0ther exemple is shown in Figures 6.13 and 6.14.

6.4 Lion Statue

Figure 6.15 shows the resulting disparity map of the "Lion" experiment. The stones of the statue are strongly textured, hence a dense final disparity map. No ground truth was available for this pair. The presented algorithm doesn't take into account the possible illuminance changes between the images of the stereo pair. To avoid this problem an image equalization was performed with the Midway equalization algorithm [Delon, 2004b].

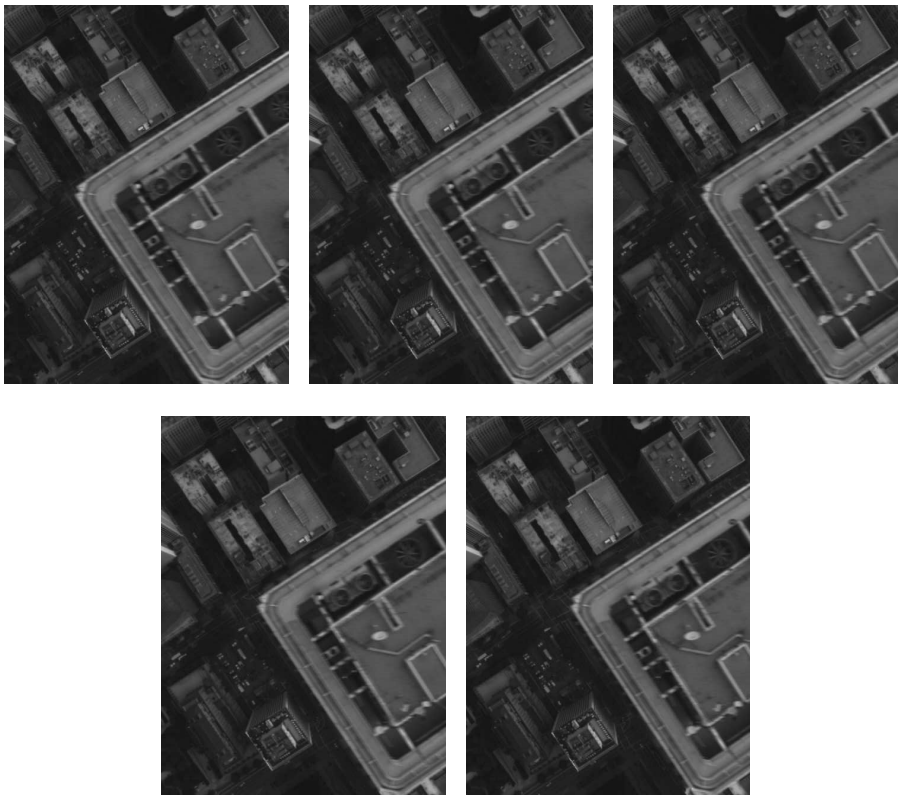
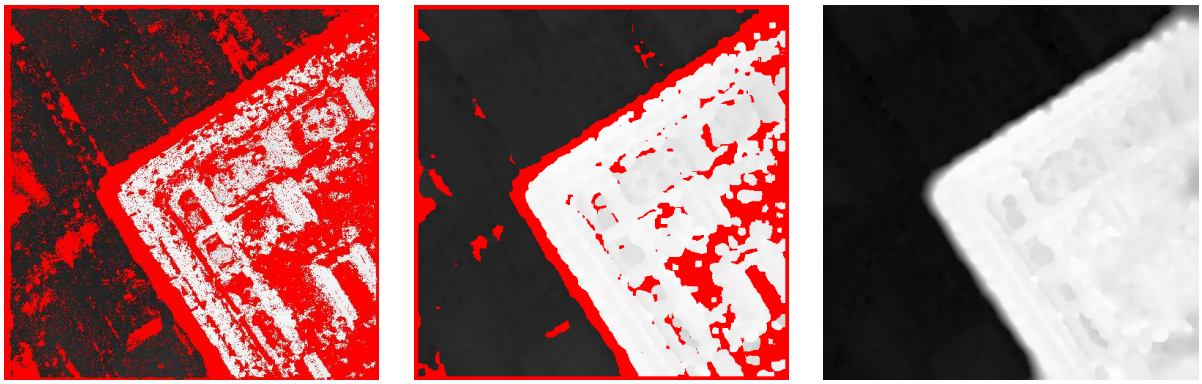


Figure 6.7: Five consecutive frames of a movie.



(a) (b) (c)

Figure 6.8: (a) Disparity map (56% density). Remark than several points are not matched on the building façades. The low density on the roof skyscraper is justified by the lost of texture in the blurred image. (b) Disparity map after median filter (82% density). (c) Interpolated disparity map.

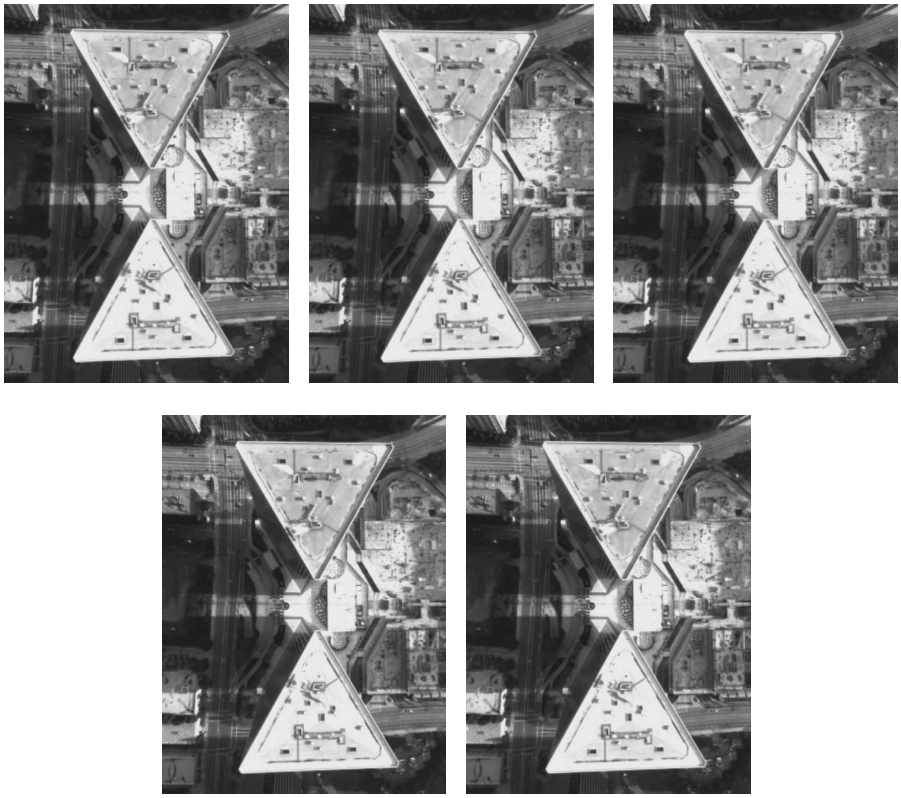
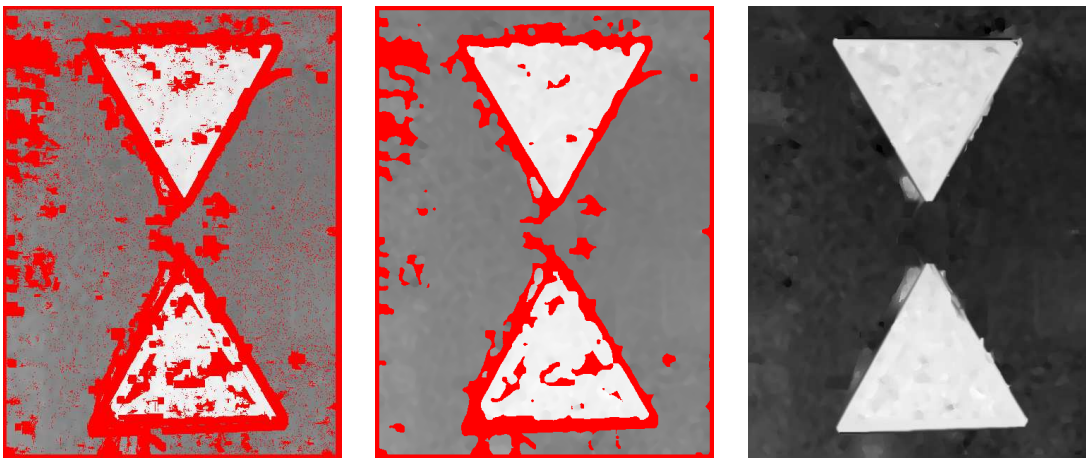


Figure 6.9: Five consecutive frames of the Century Plaza Towers (Century City, LA).



(a) (b) (c)

Figure 6.10: (a) Disparity map (70.5% density). Points on the top-left corner (on the road) have been rejected due to ambiguity in the epipolar direction. A patch in this regions is self-similar. The subpixel disparity should not be very accurate because the sequence has not been exactly rectified. (b) Disparity map after median filter (83% density). (c) Interpolated disparity map.

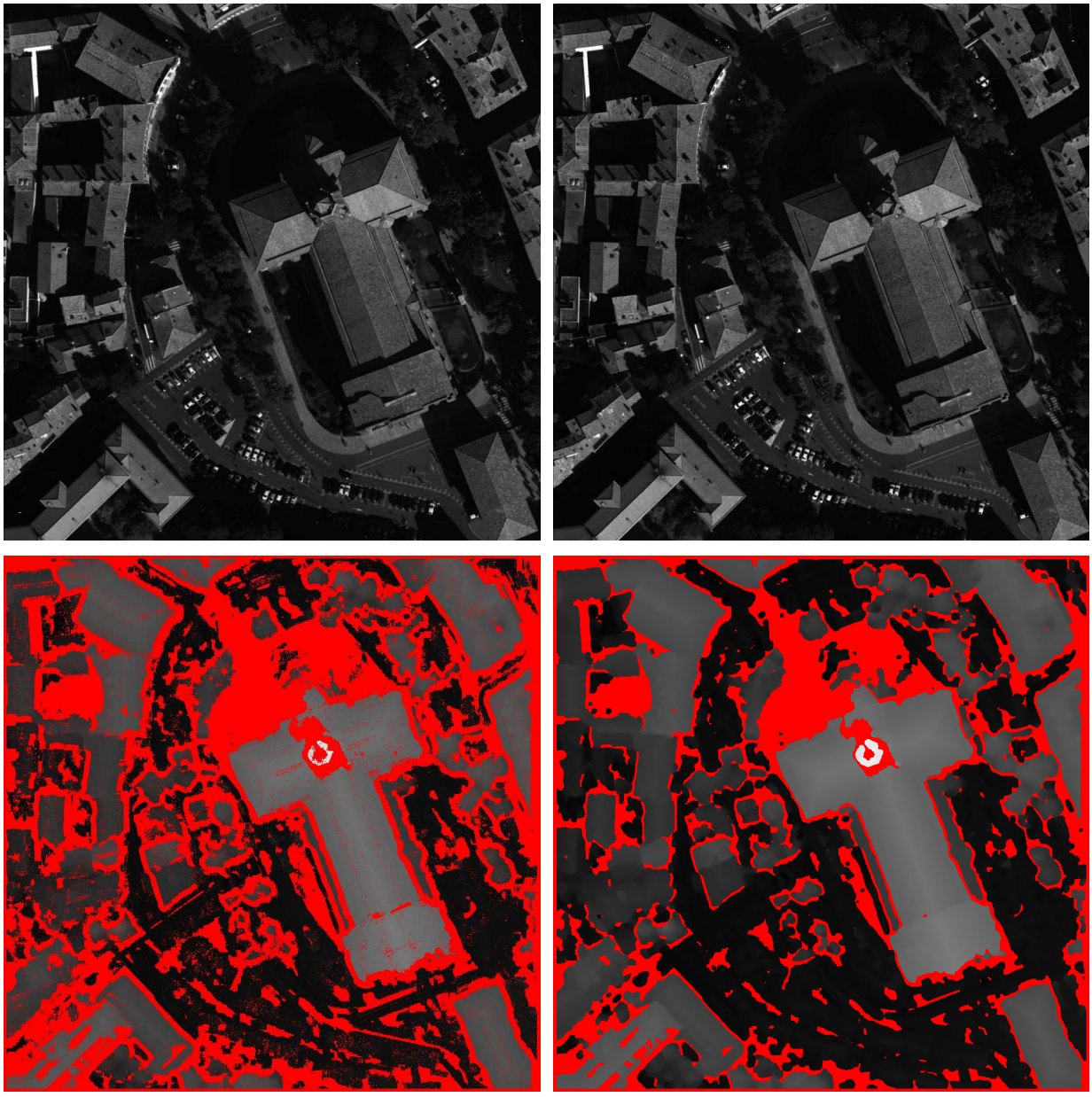


Figure 6.11: "Paroisse St. Sernin" (Toulouse). Top: pair of stereo images with 20cm resolution and $B/H = 0.08$. Bottom: on the left, disparity map (62% density). On the right, disparity map after median filter (78% density). A large region in the shadow has been rejected because there is no texture inside. Trees and vegetation have been matched successfully. The highest part of the church (second part of the tower) has been rejected in the fattening correction step.

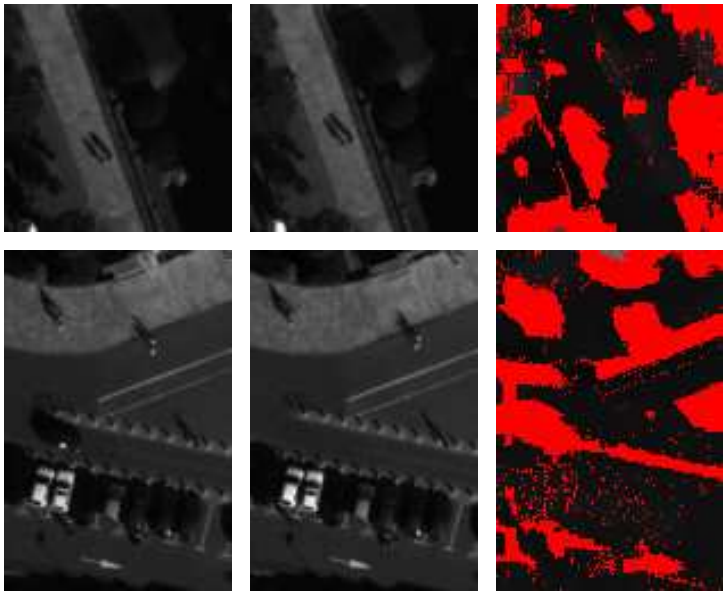


Figure 6.12: On the left, two crops of the reference and secondary images of images in Fig. 6.11. Several pedestrians and a car appear. On the right, the disparity map of the considered regions. There are no wrong matches.

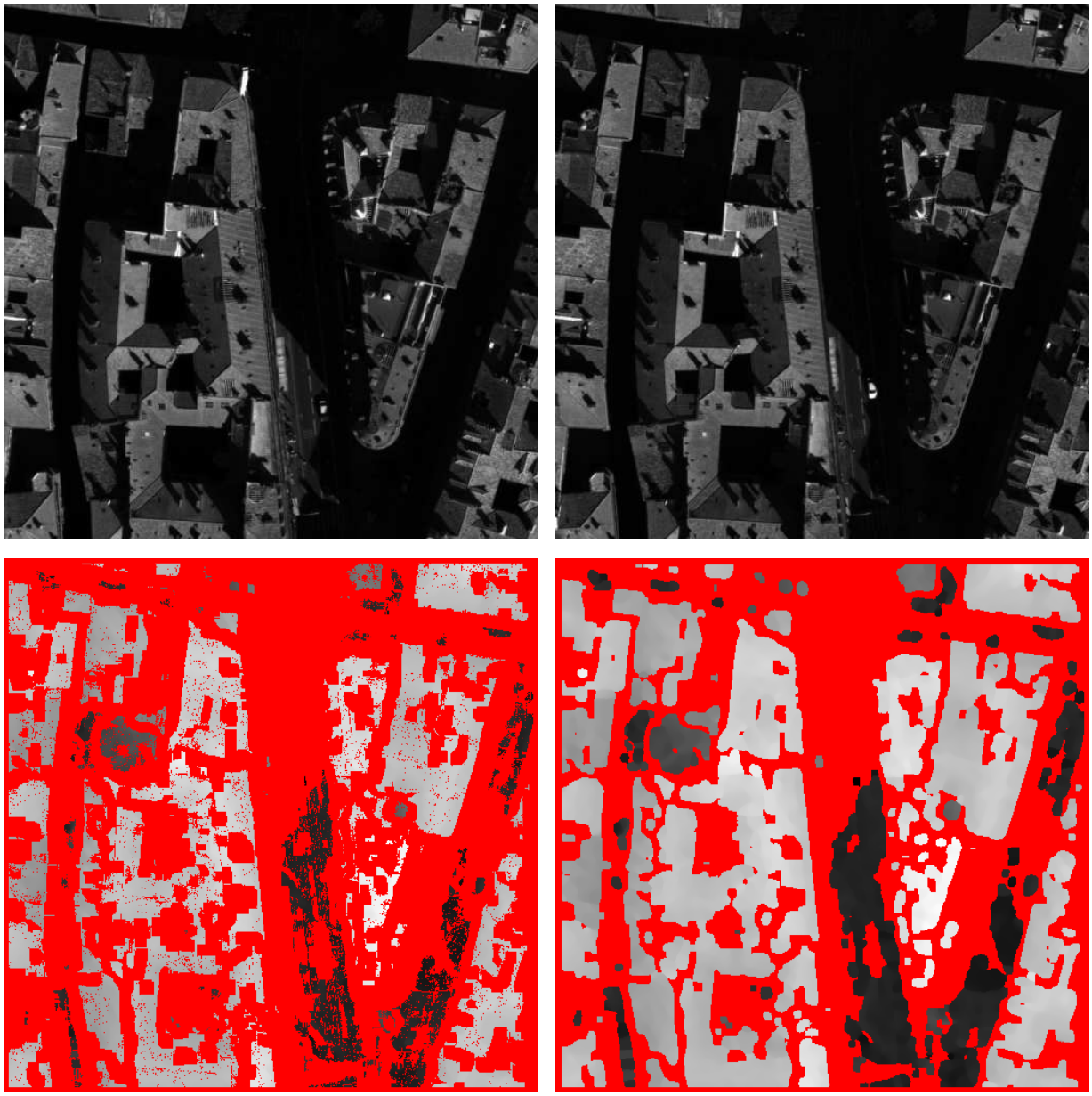


Figure 6.13: "Esquirol neighborhood" (Toulouse). Top: pair of stereo images with 20cm resolution and $B/H = 0.08$. Bottom: on the left, disparity map. The disparity map is not very dense (32%) because of the numerous shadows. Moreover, houses with different heights are right next to the others, and several patches are rejected because they risk fattening. The moving car has not been matched. On the right, disparity map after median filter (51% density).

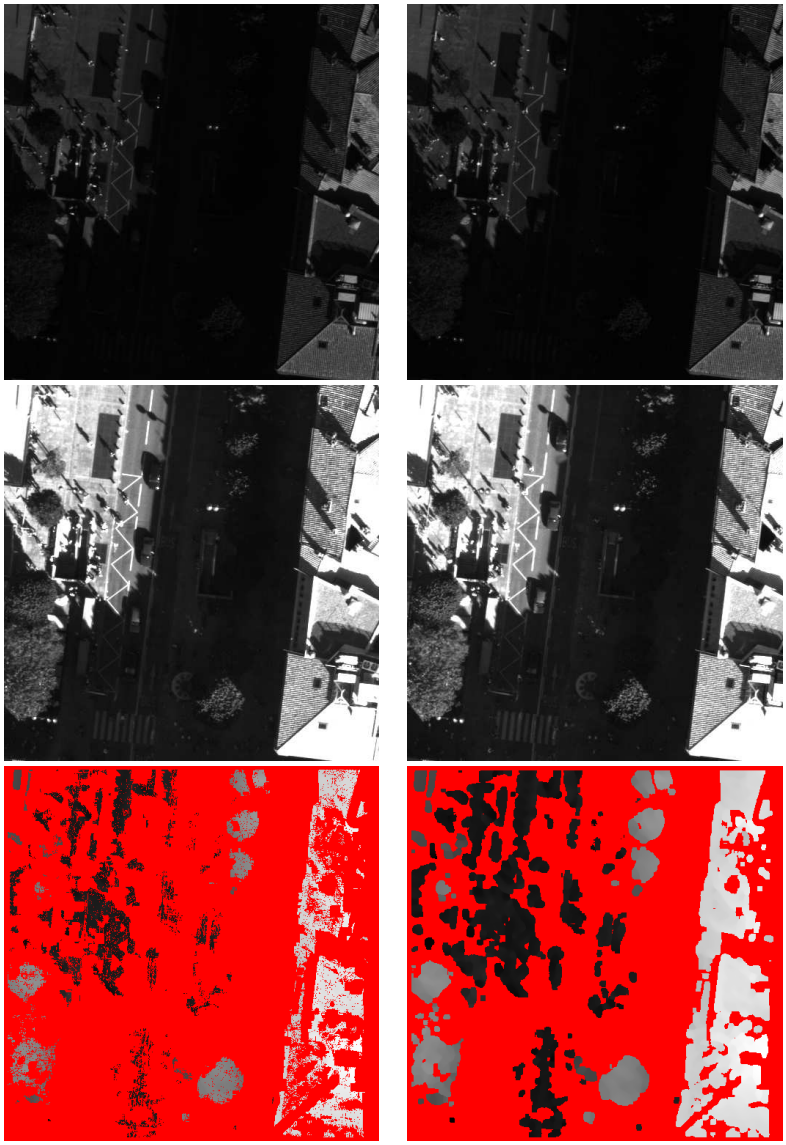
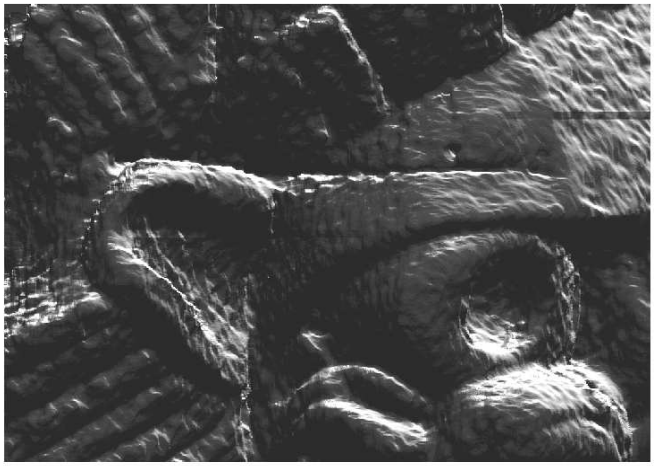


Figure 6.14: "Esquirol neighborhood 2". Pair of stereo images with 10cm resolution and $B/H = 0.08$. Top: reference and secondary images. Most of the pixels in the image are part of the shadows. Middle: reference and secondary image after a huge contrast change putting in evidence details inside the shadows. Pedestrians and cars have moved between the snapshots. Bottom: disparity map and median disparity map.

Figure 6.15: “Lion” experiment. Reference image. Secondary image. Disparity map (100% of points after median of the blocks).



(a)

(b)

(c)

Figure 6.16: 3D views of the disparity map obtained in figure 6.15. (a) Upper view of a 3D rendering of the computed surface. (b) Slanted view of a 3D rendering of the computed surface. (c) Slanted view of the computed surface with the reference image rendered as texture on the surface.

Chapter 7

Conclusion et Perspectives

Dans ce travail de thèse, nous avons étudié la mise en correspondance fiable et précise de paires d'images. En particulier, nous nous sommes intéressés aux images en milieu urbain à faible B/H.

L'approche que nous avons proposée pour la détection de points fiables repose sur un modèle *a contrario* combiné avec un seuil calculé sur les images elles-mêmes. La précision des appariements est obtenue grâce à un zoom ($\times 2$) par *zero-padding* initial de l'image, et une interpolation de Shannon de la corrélation. Finalement, un détecteur spécifique évite que le phénomène d'adhérence se produise. Cette approche permet d'aboutir à une carte de disparités fiable et précise pour plus de 40% de points de l'image.

La littérature en stéréo est très dense comme le montre le premier chapitre de cette thèse où nous avons mentionné les travaux les plus remarquables. Cependant, le nombre d'auteurs qui se sont intéressés à la mise en correspondance de points sûrs est très faible [Sara, 2002], [Veksler, 2002a, 2003a], [Mordohai and Medioni, 2006]. Nous avons vu que nos résultats étaient supérieurs à ceux-ci en termes de densité de points et de pourcentage de fausses correspondances. De plus, la plupart des auteurs en stéréo ne s'intéresse pas à la précision subpixélienne, ce qui peut s'expliquer par l'utilisation d'images à fort B/H. Dans l'esprit de [Delon and Rougé, 2007], nous avons étudié analytiquement le recalage subpixélien pour les méthodes locales et sa mise en place nous a permis d'atteindre des précisions inégalées de 1/20 pixels. La combinaison de fiabilité et précision a donné lieu au résultat clé de cette thèse : l'erreur empirique sur les points sûrs est pratiquement égale à l'erreur théorique (dûe au bruit) prédite.

Pour poursuivre ce travail, plusieurs améliorations sont possibles.

Correspondences significatives plus denses et plus précises

La méthode présentée est assez sensible aux fortes transformations géométriques de certains objets qui peuvent se produire lors de prises de vue différentes. Si on considère une fenêtre carrée sur une surface qui est loin d'être parallèle au plan image, sa fenêtre correspondante dans l'autre image apparaîtra déformée dans la direction épipolaire. La simulation locale de telles transformations géométriques des fenêtres donnerait lieu à des cartes de disparité plus denses.

La classification initiale de l'image selon moyenne et variance des fenêtres (ACP locale) a permis de mieux traiter des zones peu texturées comme les ombres. Une telle classification

n'a pas été faite pour les images couleur mais c'est encore une piste pour la continuation de ce travail.

Le choix entre les méthodes globales et locales

Le choix entre méthodes globales ou locales n'est pas simple. Il semblerait que la tendance aille dans la direction des méthodes globales mais leur plus grand défaut, à notre sens, est leur manque de validation. La méthode *a contrario* (AC) présentée dans cette thèse est une validation en elle-même et elle peut être intégrée à tout autre méthode de mise en correspondance (locale ou globale).

On pourrait reprocher aux méthodes locales de ne pas fournir de façon fiable la disparité pour tous les points de l'image, celles donnant lieu à des cartes de disparités incomplètes. Comment peut-on compléter la carte de disparités ? Nous avons proposé un remplissage possible (cf. annexe E) mais ce n'est qu'un essai. Il nous semble raisonnable de penser que la carte de disparités pourra se densifier avec une méthode globale qui tiendrait compte des points déjà appariés et de la géométrie globale de l'image.

Le besoin d'une référence

Comme on l'a déjà dit, la validation des résultats devient un point crucial quand l'on veut obtenir de la très haute précision. Cependant, le manque de vérités terrain précises rend presque impossible cette validation. Le *benchmark* de Middlebury, qui est apparu il y a moins de dix ans, a déclenché une course pour arriver aux premières positions de leur tableau d'évaluation. Mais, nous avons observé des inexactitudes dans les vérités terrain de Middlebury. De fait, la méthode par laquelle cette vérité terrain a été obtenue est discutable. A une vérité terrain extrinsègue, nous avons opposé la vérité terrain intrinsèque obtenue par la validation croisée à partir de jeux de plusieurs images de la même scène.

Appendix A

Choosing an Adequate *A Contrario* Model for Patch Comparison.

The goal of this appendix is to discuss the possible alternatives to the probabilistic block model that we have considered in Chapter 2.

In recent years, patch models and patch spaces are becoming increasingly popular. We refer to [Mairal et al., 2008] and references therein for algorithms generating sparse bases of patches. Here, our goal can be formulated in one single question, that clearly depends on the observed set of patches in our particular images, and not on the probability space of *all* patches. The question is: "*What is the probability that given two images and two similar patches in these images, the similarity arises just by chance?* ".

The "just by chance" implies the existence of a *background model*, or an *a contrario* model. There is an interesting simplification in *a contrario* models with respect to classic Bayesian ones. In the Bayes model, a model of the set of patches (the background model) would be required, but also a model of the patch itself. The $H1$ alternative would be that patch 1 and patch 2 arise from the same patch model, and the $H0$ or null alternative would be that patches such as patch 1 and patch 2 are likely to happen and to be similar in the background model. The algorithm would then choose for each patch which probability is higher: $H0$ or $H1$. In the *a contrario* framework, a background model is enough to gain a strict control of the number of wrong matches, as Prop. 1 will show, and experimental evidence confirm.

When trying to define a well suited model for image blocks, many possibilities open. Simple arguments show, however, that over-simplified models do not work. Let H be the gray-level histogram of the second image I'. The simplest *a contrario model* of all might simply assume that the observed values $I'(\mathbf{x})$ are instances of i.i.d. random variables $\mathcal{I}'(\mathbf{x})$ with cumulative distribution H. This would lead to declare that pixels $\mathbf{q}$ in image I and $\mathbf{q}'$ in image I' are a meaningful match if their gray level difference is unlikely small,

$$\mathbb{P}\Big[|I(\mathbf{q}) - \mathcal{I}'(\mathbf{q}')| \leqslant |I(\mathbf{q}) - I'(\mathbf{q}')| := \theta\Big] \leqslant \frac{1}{N_{tests}}.$$

As we shall see later, the number of tests N_{tests} is quite large in this case ($N_{tests} \approx 10^7$ for typical image sizes), since it must consider all possible pairs of pixels $(\mathbf{q}, \mathbf{q}')$ that may match. But such a small probability can be achieved (assuming for simplicity that H is uniform over $[0, 256]$) only if the threshold $\theta = |I(\mathbf{q}) - I'(\mathbf{q}')| < 128 \cdot 10^{-7}$. On the other hand, $|I(\mathbf{q}) - I'(\mathbf{q}')|$ cannot be expected to be very small because both images are corrupted by noise, among other distortions. Even in a very optimistic setting, where there are only

noise distortions between both images (of about 1 gray level standard deviation), such a small difference will only happen for about a tiny proportion ($3.2 * 10^{-5}$) of the correct matches.

This means that a pixel-wise comparison would require an extremely strict detection threshold to ensure the absence of false matches, but this leads to an extremely sparse detection (about thirty meaningful matches per mega-pixel image). This suggests that the use of local information around the pixel is unavoidable. The next simplest way to do that would be to compare blocks of a certain size $W \times W$ with the usual ℓ^2 norm, and with the same background model as before. Thus, we could declare blocks $B_{\mathbf{q}}$ and $B_{\mathbf{q}'}$ as meaningfully similar if

$$\mathbb{P}\left[\frac{1}{|B_0|}\sum_{\mathbf{x}\in B_0}|I(\mathbf{q}+\mathbf{x})-\mathcal{I}'(\mathbf{q}'+\mathbf{x})|^2 \leqslant \frac{1}{|B_0|}\sum_{x\in B_0}|I(\mathbf{q}+\mathbf{x})-I'(\mathbf{q}'+\mathbf{x})|^2 := \theta\right] \leqslant \frac{1}{N_{tests}}.$$

Now the test would be passed for a more reasonable threshold ($\theta = 6$, 28, 47 for blocks of size 3×3, 5×5, 7×7 respectively), which would ensure a much denser response. However, this *a contrario* model is by far too naive, and produces many false matches. Indeed, blocks stemming from natural images are much more regular than the white noise generated by the background model. Considering all pixels in a block as independent leads to overestimate the similarity probability of two observed similar blocks. It therefore leads to an over-detection.

In order to fix this problem, we need a background model that actually reflects the statistics of natural image blocks. But directly learning such a probability distribution from a single image in dimension 49 (for 9×9 blocks) is hopeless.

Fortunately, as pointed out in [Musé et al., 2006a], shape high-dimensional distributions can be approximated by the tensor product of their adequately chosen marginal distributions. Such marginal laws, being one-dimensional, are more easily learned from a single image. Ideally, ICA should be used to learn which marginal laws are the most independent, but the simpler PCA analysis will show accurate enough for our purposes. Indeed, it ensures that the principal components are decorrelated, a first approximation to independence. Fig. 2.6 permits a visual assessment of how well a local PCA model simulates image patches in a class.

Appendix B

Avoiding Fattening with the Line Segment Detector (LSD)

In Chapter 2 we have dealt with the fattening phenomenon which is one of the main drawbacks in block-matching methods. Although the fattening correction presented in that chapter is the best solution in a general case, here we are going to discuss the approach we had in our first researches about the subject. It is interesting to known which are the limits of this approach and why we have decided to adopt another solution.

In urban images the borders of the structures are generally straight lines, so it makes sense to consider line segments in the image as the possible border objects to be avoided. Assuming straight lines in urban areas was also done in [Jakubowicz, 2007] where a classification between urban and not urban areas is done with a carefully detection of the right angles formed by the line segments. Here we use line segments in a different way, fattening is avoided by comparing only blocks that do not meet the image line segments. Thus, they must be eliminated by forbidding blocks to meet any conspicuous straight segment in the image. As a consequence, no disparity will be at first available in the image regions obtained by dilating all segments by the block window. The pixels inside these regions are also ignored in the training phase of the local PCA.

Removing blocks in the image containing edges is exactly the opposite than other algorithms do. The explanation of such a different approach is in a conflictive fact. Edges contain important geometrical information about the images but, nonetheless, edges are the main cause of fattening errors because they often border depth discontinuities. [Delon and Rougé, 2007] studied the second derivative of the correlation coefficient (correlation curvature) in order to detect points where correlation can be computed accurately. The flatness of the correlation coefficient near the maximum gives an *a priori* information about the obtainable precision at each point. The bigger the curvature, the more reliable the correlation is. Most pixels in the image with higher correlation curvature are in fact edge pixels. Thus, the accepted as reliable pixels in this approach are concentrated in border objects.

Line segment disparities can be computed by matching each line segment of the first image to the second image. Indeed, even if avoiding disparities close to the edges is necessary, pixels lying exactly on the line segments will not affected by fattening. An automatic line matching has already been studied in [Schmid and Zisserman, 1997]. These authors presented a method

for matching individual line segments based on grey level information and geometric relations between images. [Baillard and Zisserman, 2000] obtained good results in urban areas by matching segments between both images to find the height of the 3D edges, and then matching each half-plane on both sides of the segment to find the tilt of the corresponding 3D planes. However, this approach cannot be applied in the urban low baseline setting because the upper and lower part of a vertical wall are then so close to each other that they are fused by the instrument's PSF into a single segment. Finally, [Bay et al., 2005] matches line segments between two uncalibrated wide-baseline images. The authors generate an initial set of line segment correspondences and then add matches consistent with the topological structure of the current ones. A coplanar grouping stage allows them to estimate the fundamental matrix.

The Selected Line Segment Detector: LSD [Grompone et al., 2008]

The LSD Algorithm defines a line segment as a straight region whose points overwhelmingly share the same image gradient direction. LSD proceeds by first partitioning the image into *line-support regions*: groups of connected pixels that share the same gradient direction up to a certain tolerance. When large enough, these regions are approximated by rectangles. Figure B.1 extracted from [Grompone et al., 2008] shows the growing process for the computation of the *line-support regions*. Figure B.2 shows the regions computed in an aerial image. Figure B.3 shows an example of the obtained segments of another aerial image.

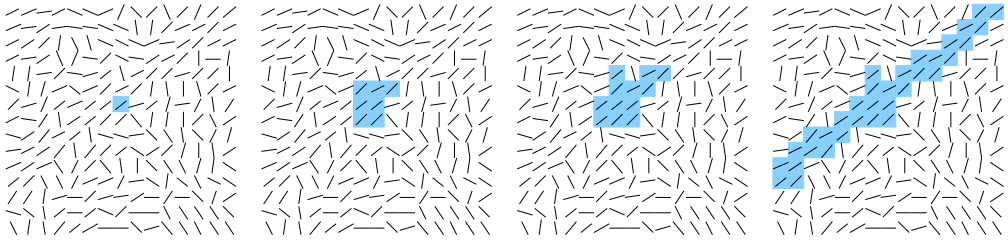
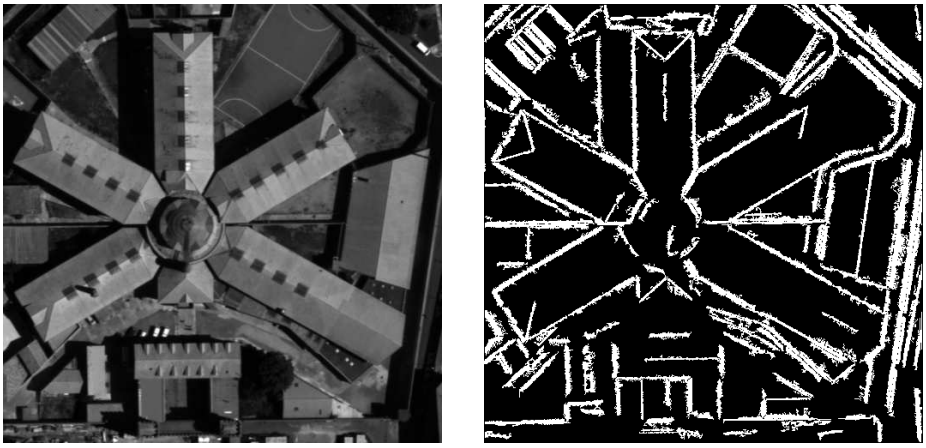


Figure B.1: Growing process of a region of aligned points. The level-line orientation field (orthogonal to the gradient orientation field) is represented by dashes. Colored pixels are the ones forming the region. From left to right: first step, second step, third step, and final result.

In order to match segments, the LSD *line-support regions* become the correlation blocks and the correlation maximum gives the segment's region disparity (see Fig. B.4). All in all, this approach gives satisfactory results since line segments not corresponding to a depth discontinuity match very exactly. Comparing the disparity on the line segment with the left and right block disparities gives a reliable test ensuring that the segment is or not a depth discontinuity. When no depth discontinuity is detected on both sides of a segment, the adjacent pixels can be removed from the list of pixels risking fattening. This permits to reintegrate in the block-matching process points close to harmless segments, in particular those due outline shadows. Figure B.7 shows the obtained results for a simulated pair of images of the Toulouse's prison (France). The algorithm was run with and without a previous line segment detection, and the resulting disparity maps are compared. Fig. B.6 shows the



(a) (b)

Figure B.2: (a) Reference image. (b) *line-support regions* computed.

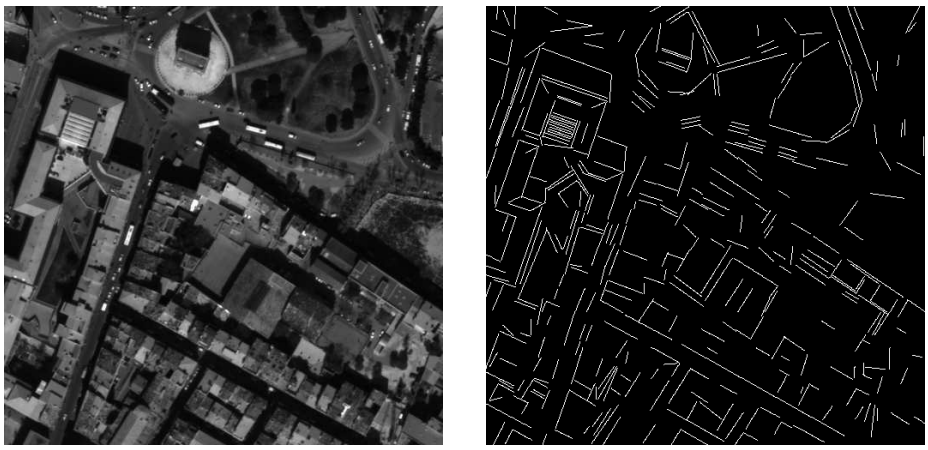
difference between the computed disparity map and the ground truth. The improvement when using the line segments is obvious.

For a sake of honesty, some failure cases must be pointed out:

- Detected segments in the image are sometimes part of slanted surfaces of the scene and should have different height at each point. However, the proposed method attributes the computed disparity to the whole segment.
- Line segments are not well adapted to curved structures. In urban areas, not only straight borders occur (e.g. domes as in the image in Fig. B.2).
- LSD can fail to detect a line segment which corresponds to a real depth discontinuity.
- The stroboscopic phenomenon also takes place in the segment matching.

Figure B.8 and Figure B.5 illustrate these failure cases.

These last remarks lead us to believe that another approach available in a more general case has to be considered. This is why in the version proposed in Chapter 2 we correct *a posteriori* the patches risking fattening according to the computed disparity map and we use the Canny-Deriche edge detector [Canny, 1986; Deriche, 1987].



(a) (b)

Figure B.3: (a) Aerial reference image. (b) Detected segments.

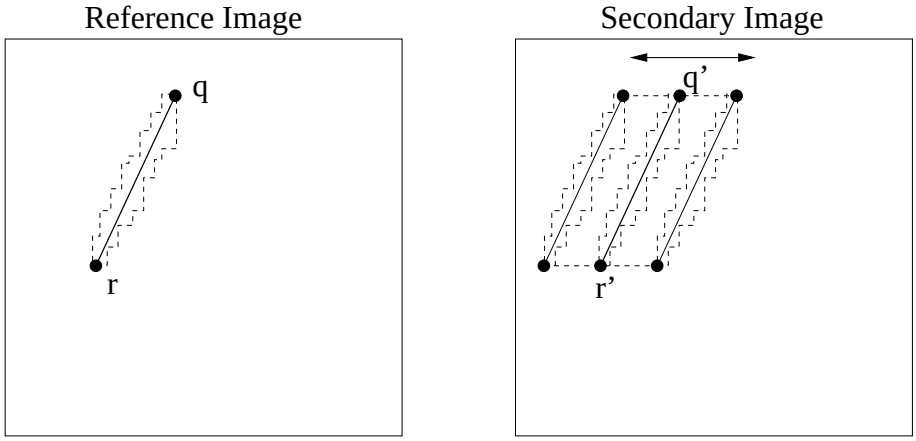
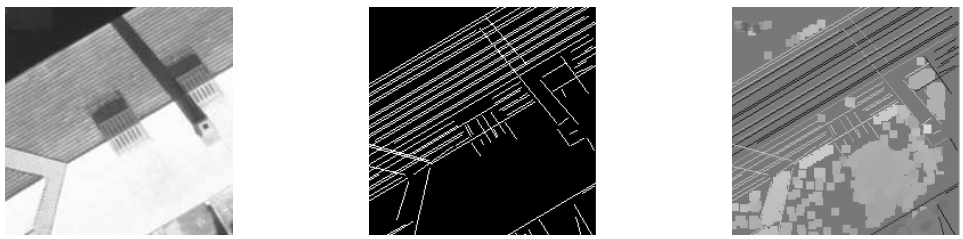


Figure B.4: On the left reference image with a line segment $\overline{qr}$. Around the segment, the line-support computed by LSD. In order to find the disparity for the pixels $s \in \overline{qr}$ correlation is computed shifting the correlation window in the epipolar direction. The correlation window has the shape and size of the line-support regions.



(a) (b) (c)

Figure B.5: (a) Reference image. (b) Mask of detected line segments. Several segments have been found in one side of the roof due to its lined texture. (c) Disparity map. Several line segments have been mismatched.

(a) (b)

Figure B.6: (a) Absolute value of the difference image between the ground truth and the disparity map obtained without detecting previously line segments. (b) Difference image between ground truth and disparity obtained using the line segments. There are more errors in pixels belonging to line segments than on the other ones. If a line segment is missed by the detector, fattening occurs anyway.

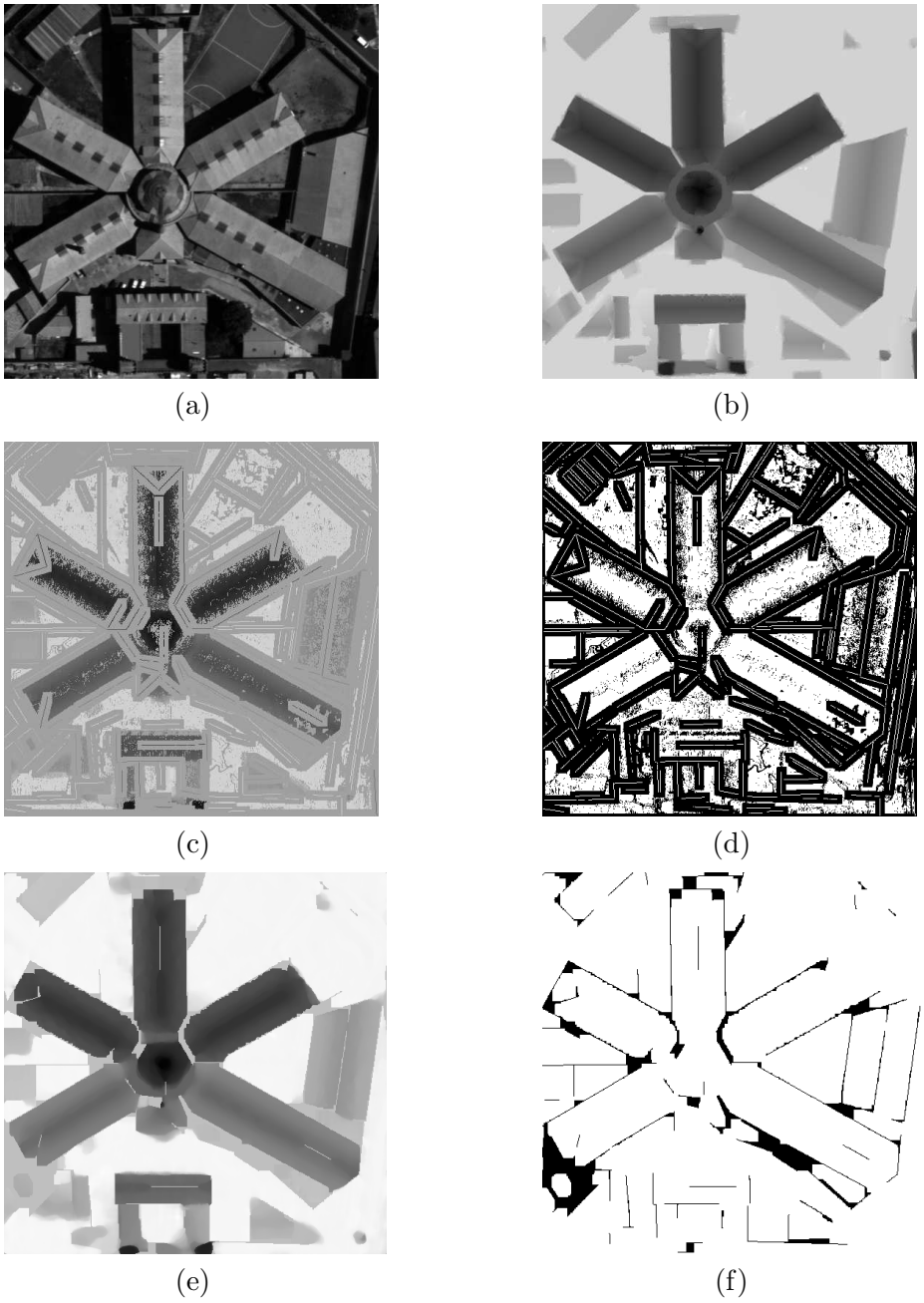


Figure B.7: (a) Reference image. (b) Ground truth. The secondary image has been simulated from the reference image and the ground truth. In this case darker pixels in the disparity map correspond to higher points in the scene. (c) Resulting disparity map. (d) Valid points of the algorithm. (e) Disparity map after completion (median). Segments with the same block estimation disparity on both sides had been removed from the set of risk segments. The disparity in these points has been computed to obtain denser disparity maps. (f) Mask of valid points (91%).

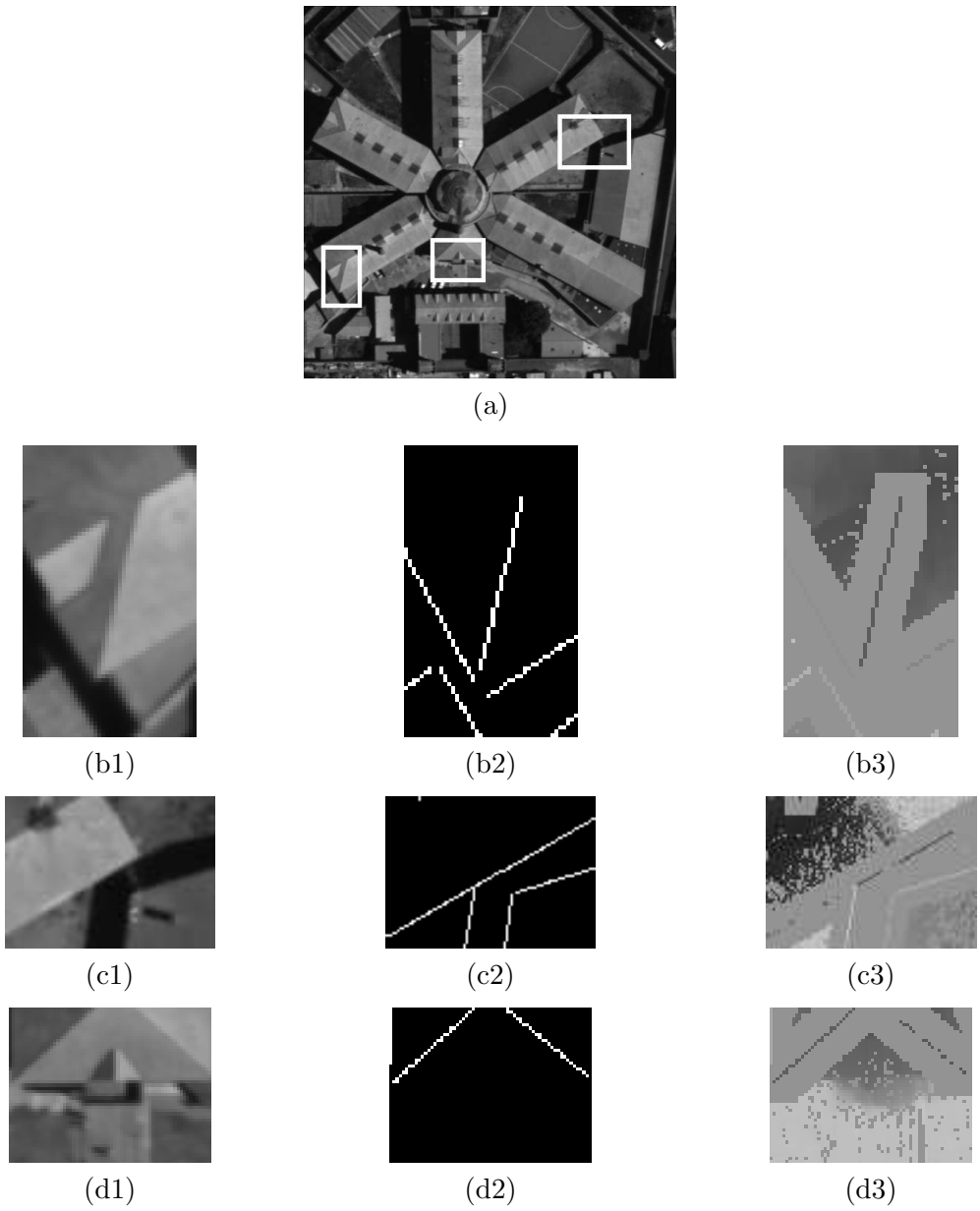


Figure B.8: This figure explains why line segments cannot always be matched efficiently, and illustrates the fattening problem. (a) Reference image with three boxes containing failures. (b1) First extract of the reference image. (b2) A segment is detected on a pitch of the roof. (b3) Extract of the disparity map on the segment obtained by matching it globally to the other image. As a consequence, the computed disparity is the same on the whole segment, and therefore wrong. (c1) Second extract of the reference image. (c2) A long segment is detected on the edge of the building. Because of a shadow, this segment is longer than the building. (c3) Matching the segment in the other image leads to an error in the disparity map: points in the ground have the same disparity as the points on the edge. This last problem must be alleviated by matching pieces of segments when they are too long. (d1) Third extract of the image. (d2) A segment on the edge of the building is not detected. (d3) As a consequence of the missed segment, fattening occurs.

Appendix C

Generalization to Color Images

The generalization of our algorithm for color images is quite easy. In this appendix we give some details about it. Mainly there are two differences: the computation of the principal component analysis and the computation of the quadratic distance.

Color PCA

If u_1 and u_2 are color images the three channels of each patch have to be stored in the matrix X. Fig. C.1 shows how the patch information is stored in X. Then, the eigenvectors and eigenvalues of $\mathrm{Cov}(X)$ are computed. Notice than the eigenvectors are three times bigger than the gray case because they contain the three channels. Then, they are recomposed in patches of size 9×9. The patches corresponding to the first principal components are in Fig. C.2.

In the color version we have not considered a local partition of the pixels in the image, but it is a clue for further research.

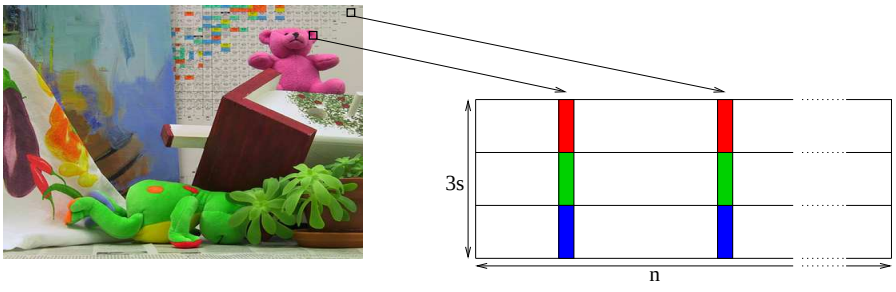


Figure C.1: Color PCA. The RGB channels of each patch are stored as a column of X.

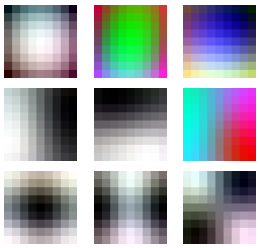


Figure C.2: First color principal components ordered from left to right and top to bottom.

Quadratic Distance

The quadratic distance is computed in the subpixel refinement but also in the computation of the self-similarity threshold. The quadratic distance for color images can be written as

$$\begin{aligned} e^d_{\mathbf{q}}(\mu) = &\sum_{\mathbf{m}\in B_{\mathbf{q}}} \varphi_{\mathbf{q}}(\mathbf{m}) \left[(R_{u_1}(\mathbf{m}) - R_{u_2}(\mathbf{m}+\tau_\mu))^2\right] \\ &+ \sum_{\mathbf{m}\in B_{\mathbf{q}}} \varphi_{\mathbf{q}}(\mathbf{m}) \left[(G_{u_1}(\mathbf{m}) - G_{u_2}(\mathbf{m}+\tau_\mu))^2\right] \\ &+ \sum_{\mathbf{m}\in B_{\mathbf{q}}} \varphi_{\mathbf{q}}(\mathbf{m}) \left[(B_{u_1}(\mathbf{m}) - B_{u_2}(\mathbf{m}+\tau_\mu))^2\right] . \end{aligned}$$

where $\tau_\mu = (\mu, 0)$ and R_u, G_u and B_u are the RGB channels of u.

Appendix D

Comparison with MARC

In this appendix we detail the Multiresolution Algorithm for Refined Correlation (MARC) coded by N. Camlong and V. Muron [Camlong, 2001][Muron, 2001] and patented by CNES (French patent by A. Giros, B. Rougé and H. Vadon [Giros et al., 2004]). Our algorithm has been compared with MARC and a considerable improvement has been noticed from a quantitative and a qualitative point of view.

MARC Description

The main goal of MARC is the computation of Digital Elevation Models from a stereoscopic pair of images with small baseline (small ratio B/H, of the order of several hundredths). MARC is a local algorithm and computes correlation with squared patches of different sizes in order to estimate the disparity at each point. Furthermore, it is assumed that the couple of images satisfies the epipolar constraints.

The analytical study of correlation considered for MARC interpolation is done in [Delon and Rougé, 2007].

Multi-Scale Treatment

The main particularity of MARC is the multi-scale treatment of the image. Indeed, it starts by computing a disparity map at the lowest resolution and at each step the resulting disparity map is used as an initial solution for the computation of the disparity map at the next scale. At each iteration the refined disparity is sought in a one pixel interval left and right of its current quantized estimate. The main difficulty is to avoid the convergence to an erroneous second correlation peak. Nevertheless, The multi-scale treatment in the implementation often allows to obtain reliable results in parts of the image where there is a lack of texture at a certain resolution.

The Barycentric Correction

As all local methods, MARC suffers from fattening. In order to correct this error, the barycentric correction is performed. Roughly, it consists on approximating the correlation density by a delta function placed at the barycenter of the correlation window. Then, the computed disparity is assigned to the barycenter and no to the center of the window. Chapter 3 gives a full description of fattening and more detail about the barycentric correction.

Correlation Curvature

The flatness of the correlation function gives an idea of the reliability one can expect from the resulting estimated disparity. The flatter the correlation, the less reliable the disparity. This is equivalent to study the second derivative of correlation ρ called correlation curvature,

$$\rho''(\mu(\mathbf{x}_0)) = -(d_{\mathbf{x}_0}(u, u_x) * \varphi)(\mathbf{x}_0) + O(\|\varepsilon\|_\infty)\,,$$

where φ is the correlation window and $d_{\mathbf{x}_0}$ the correlation density around $\mathbf{x}_0$. Given the standard deviation σ of the noise in the images, MARC evaluates the precision one can get with such images. The used error upper bound is

$$N(u, \varphi, \mathbf{x}_0) = \frac{\sigma}{\|u\|_{\varphi_{\mathbf{x}_0}} \sqrt{(d_{\mathbf{x}_0}(u, u_x) * \varphi)(\mathbf{x}_0)}}\,.$$

Since the acquisition system is known, the value of σ is also known. Then, given a precision λ, at each point $\mathbf{x}_0$ MARC takes the smaller correlation window φ such that

$$N(u, \varphi, \mathbf{x}_0) < \lambda.$$

Based on this, MARC only keeps the points where this inequality can be achieved. As a result of this threshold and of the barycentric correction, MARC computes a non-dense disparity map at each scale. Thus, an interpolation is performed before passing to the next scale in order to fill in the unknown values. At the end, a completely dense disparity maps is available even if only a percentage of points is validated as accurate enough.

Subpixel Correlation

MARC has been created specially to the small baseline model. It therefore computes very precise disparities at high subpixel accuracy. These subpixel disparities are obtained with a Shannon interpolation detailed and analyzed in Chapter 4.

Indeed, MARC respects the Shannon principle when interpolating the correlation, meaning that the images are previously over-sampled by 2 factor , and the used correlation windows are spheroidal prolate functions.

Comparison with our Algorithm

We are going to compare our algorithm with the disparity map obtained at the last scale by MARC. More precisely, we are going to compare it with the MARC disparity map just before the last smoothing step.

The criterion to accept or reject a point is very different in both algorithms. Fig. D.1 shows the different validated masks of points for the simulated pair of images of St. Michel prison (Toulouse).

A quantitative comparison can be done by using a simulated image pair, for which the ground truth is perfectly known. We have evaluated the density and the error in terms of RMSE[1] for different levels of noise.

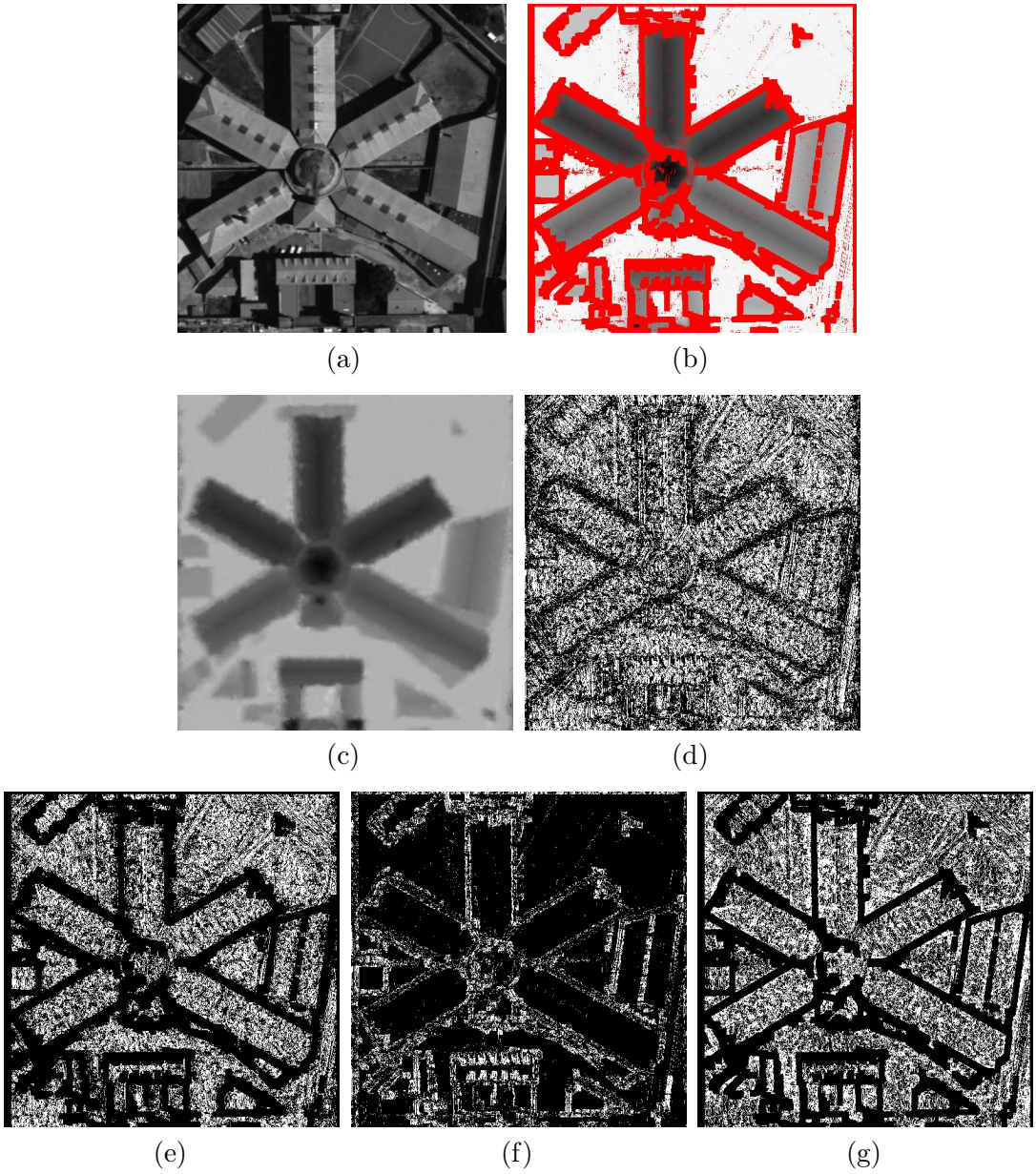


Figure D.1: (a) Reference image. (b) Our disparity map (63.9% density). (c) MARC resulting disparity map (without the last scale smoothing). (d) MARC valid points mask, (42.3%). White points are validated. (e) Points validated by both algorithms (30.6%). (f) White points have been validated by MARC but rejected by our algorithm (11.6%). These points are mainly close to the building contours. Our algorithm rejects them because they risk fattening. (g) White points (30.6%) are points with a reliable match in our algorithm but rejected by MARC. Most of these points are situated in textured areas where there is actually almost no error chance.

	$\sigma = 0$		$\sigma = 3.5$		$\sigma = 7$		$\sigma = 10$	
	RMSE	density	RMSE	density	RMSE	density	RMSE	density
MARC	0.122	42.3%	0.126	41.0%	0.181	27.4%	0.187	24.7%
Our algorithm	0.029	66.1%	0.052	56.2%	0.090	47.5%	0.108	33.9%

Table D.1: Quantitative results for MARC and our algorithm.

See table D.1 for a summary of such results. We conclude that our algorithm has higher densities with a considerably lower error.

In Chapters 2 and 6 we have seen that our algorithm rejects any possible match where there are moving objects in the scene. Figures D.2 and D.3 show the disparity maps for boths algorithms and the mask of validated points. In spite of the low densities of MARC validated points (9.8% and 16.6% respectively), there are false correspondences due to the changing car positions.

[1]We recall that $RMSE = \left(\frac{\Sigma_{\mathbf{q} \in M} (\mu(\mathbf{q}) - \varepsilon(\mathbf{q}))^2}{\#M} \right)^{\frac{1}{2}}$, where ε is the ground truth, μ the estimated disparity and $M \subseteq I$ the set of accepted points.

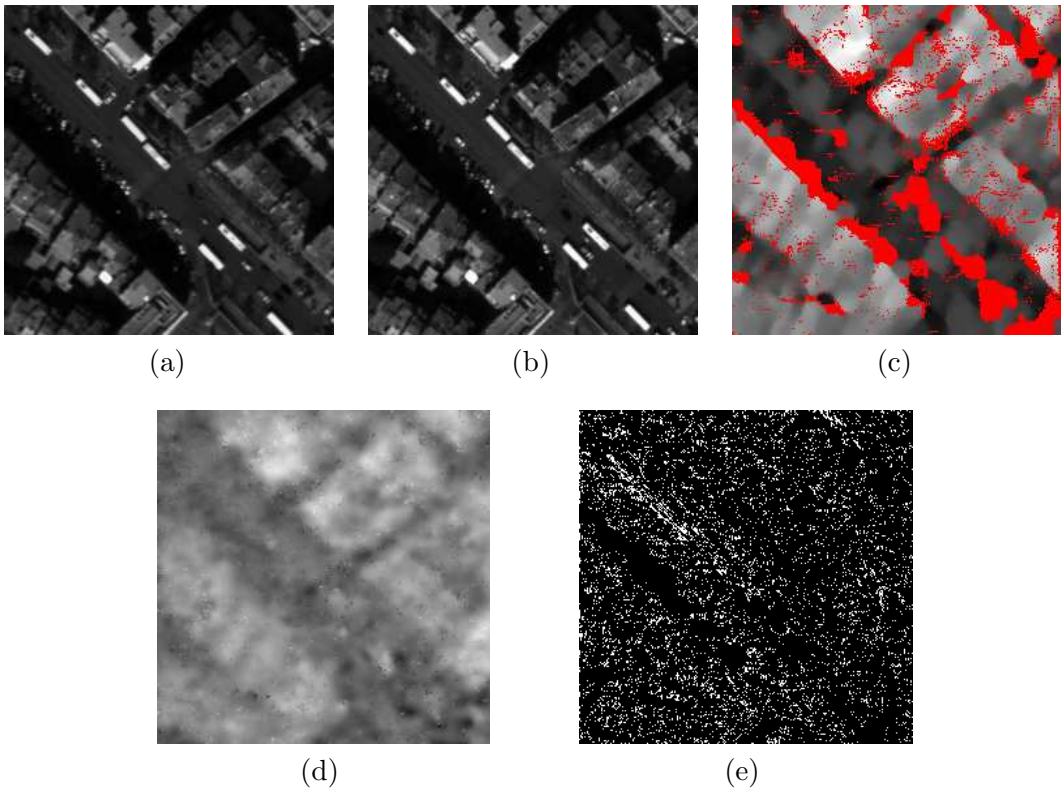


Figure D.2: (a) Reference image. (b) Secondary image. (c) Our disparity map (60.1% density). Moving objects and shadows have been rejected. (d) MARC disparity map (without smoothing at the last scale). (e) Mask of MARC valid points (9.8% density).

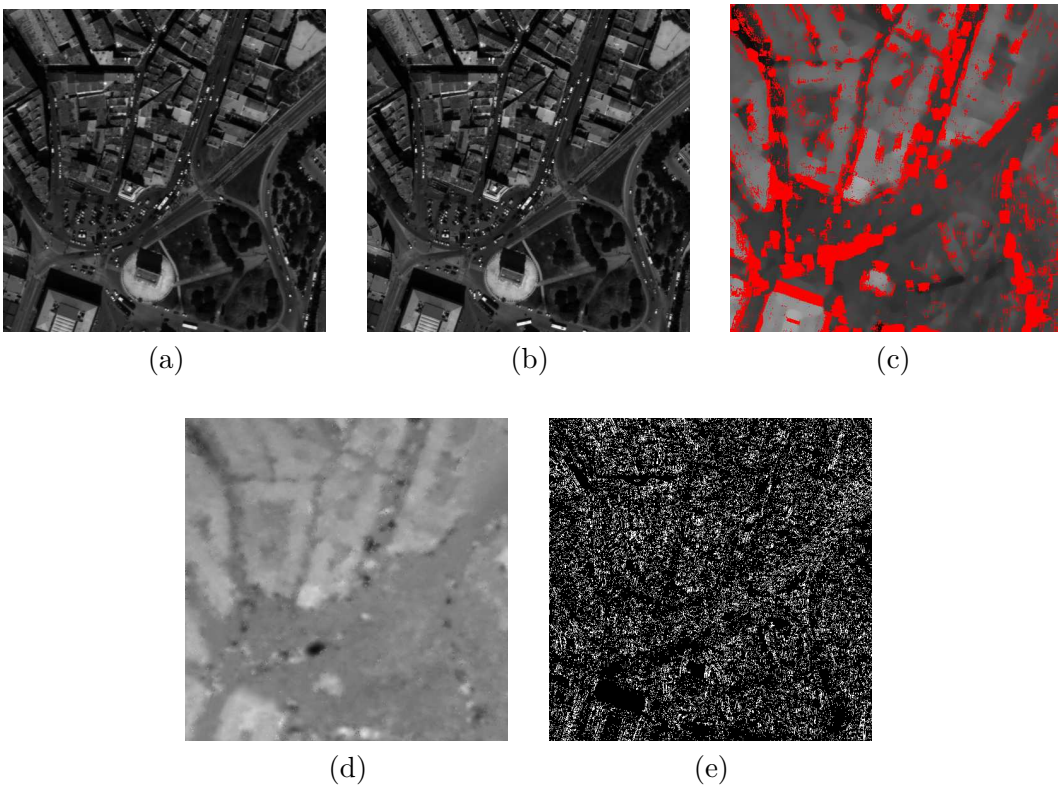


Figure D.3: (a) Reference image. (b) Secondary image. (c) Our disparity map (78.6% density). (d) MARC disparity map. (e) Mask of MARC valid points (16.6% density).

Appendix E

Disparity Map Completion

The main goal of the presented method in this thesis is not the computation of dense disparity maps, but in some cases a partially or a full completion of the disparity map can be necessary. In this appendix we explain the completion we have realized in our disparity maps. Two options are possible. First, the completion with a median filter which increases the disparity map density but does not yield a completely dense one. Second, the bilateral filter can be used iteratively to get a 100% disparity map. However, this interpolation is far from being as accurate as the disparity in the validated pixels.

The interpolation of disparity maps and in particular of Digital Elevation Models (DEMs) has been considered in several recent works. [Facciolo et al., 2006] proposes to interpolate unknown areas by constraining a diffusion anisotropic process to the geometry imposed by a reference image, and coupling the process with a data fitting term which tries to adjust the reconstructed surface to the known data. More recently, [Facciolo and Caselles, 2009] has proposed a new interpolation method which defines a geodesic neighborhood and fits an affine model at each point. The geodesic distance is used to find the set of points that are used to interpolate a piecewise affine model in the current sample. This interpolation is refined by merging the obtained affine patches with a Mumford-Shah like algorithm. Finally, [Bughin and Almansa, 2009] finds the optimal grouping of 3D point clouds representing the underlying surface into planar surfaces, whenever possible. The *a contrario* methodology has been used in a merging procedure. These three works are strongly related to ours. In the urban context, [Lafarge et al., 2008] uses a dictionary of complex building models to fit the disparity map. However, the applicability of such a method is less evident because of the initial delineation of buildings by a rectangle fitting.

Median filter

The median filter is the easiest way to extend the disparity map. The disparity map after a median filtering is

$$\mu_M(\mathbf{q}) = \underset{\mathbf{r}\in B_\mathbf{q}}{\text{Med}}\{\mu_F(\mathbf{r}) \mid \mu(\mathbf{r}) \neq \emptyset\},$$

where Med is the median operator, $B_\mathbf{q}$ is the patch centered at $\mathbf{q}$, $\mu_F(\mathbf{r})$ is the final computed disparity in $\mathbf{r}$, defined in Chapter 3. Apart from points risking fattening γ, it is a sound guess to assume than the disparity map is smooth. Thus, it is reasonable to fill in small regions in the image with the median filter.

Bilateral filter

The bilateral filter averages the pixel colors, based on both their geometric closeness and their photometric similarity, preferring of course near values to distant values in space and color. [Tomasi and Manduchi, 1998a] have used it to smooth the images while preserving edges and [Yoon, 2006] have used it to weight the correlation windows before the stereo correspondence search. Here is the proposed adaptation of the bilateral filter to a disparity map interpolation. Let $\mathbf{q}$ be a point in I. Consider $L_{\mathbf{q}} \subset I$ the subimage where the weight is learned. For each $\mathbf{r} \in L_{\mathbf{q}}$ the weight due to color similarity and proximity are computed:

color similarity: We consider the color distance

$$d_c(u_{\mathbf{q}}, u_{\mathbf{r}}) = \left((R_u(\mathbf{q}) - R_u(\mathbf{r}))^2 + (G_u(\mathbf{q}) - G_u(\mathbf{r}))^2 + (B_u(\mathbf{q}) - B_u(\mathbf{r}))^2\right)^{1/2} ,$$

where R_u,G_u and B_u are the red, green and blue channels of u. Then the weight corresponding to the color similarity between $\mathbf{r}$ and $\mathbf{q}$ is

$$w_c(\mathbf{r}, \mathbf{q}) = exp\left(-\frac{d_c(u_{\mathbf{q}}, u_{\mathbf{r}})^2}{h_1^2}\right) .$$

proximity: We consider the Euclidean distance between the points positions in the image plane

$$d(\mathbf{q}, \mathbf{r}) = \left((q_1 - r_1)^2 + (q_2 - r_2)^2\right)^{1/2} ,$$

where $\mathbf{r} = (r_1, r_2)$ and $q = (q_1, q_2)$. Then the weight corresponding to proximity is

$$w_d(\mathbf{r}, \mathbf{q}) = exp\left(-\frac{d(\mathbf{q}, \mathbf{r})^2}{h_2^2}\right) .$$

Therefore, the total associated weight between the two points q and p is

$$W(\mathbf{r}, \mathbf{q}) = \frac{1}{Z_{\mathbf{q}}} w_c(\mathbf{r}, \mathbf{q}) w_d(\mathbf{r}, \mathbf{q}) = \frac{1}{Z_{\mathbf{q}}} exp\left(-\left(\frac{d_c(u_{\mathbf{q}}, u_{\mathbf{r}})^2}{h_1^2} + \frac{d(\mathbf{q}, \mathbf{r})^2}{h_2^2}\right)\right) ,$$

where $Z_{\mathbf{q}}$ is the normalizing factor $Z_{\mathbf{q}} = \sum_{\mathbf{r} \in L_{\mathbf{q}}} w_c(\mathbf{r}, \mathbf{q}) w_d(\mathbf{r}, \mathbf{q})$. The interpolated disparity map μ_I is computed via an iterative schema

$$\mu_I(\mathbf{q}, k) = \sum_{\mathbf{r} \in L_{\mathbf{q}}} W(\mathbf{r}, \mathbf{q}) \mu_I(\mathbf{r}, k-1) ,$$

where k is the current iteration and the initialization $\mu_I(\cdot, 0) = \mu_M(\cdot)$.

Interpolation of Middlebury results

Figures E.1 and E.2 show the interpolated Middlebury results (100% density). Let us analyze where the bigger errors appear.

- Objects thinner than the patch size. Indeed, the fattening correction removes patches near disparity discontinuities. If the object is too thin, all the disparities are removed from the object and the interpolation can do poorly. This is the case of the Tsukuba lamp or the sticks inside the cup in cones.
- Slanted surfaces. When the normal vector to an observed surface points out in a very different direction than the optical center of the camera, the surface appears strongly different in the second image. Locally, a transformation exists between them (but with strong tilts and shears) and therefore a fixed squared patch is not adapted. This is the case for the floor scene in Teddy or for the box below the bust in Tsukuba.
- Occluded objects. If the baseline between the images is too large, some objects or parts of the scene can appear in only one of the images. The interpolation of such parts is false due to the lack of disparities. This is the case of the left part of the Teddy and Cones scenes, where an important part of the scene is missing in the second image.
- Border objects. Near the object border the interpolation can be erroneous depending of the color border. However, the ground truth is not precise, specially in the border pixels.

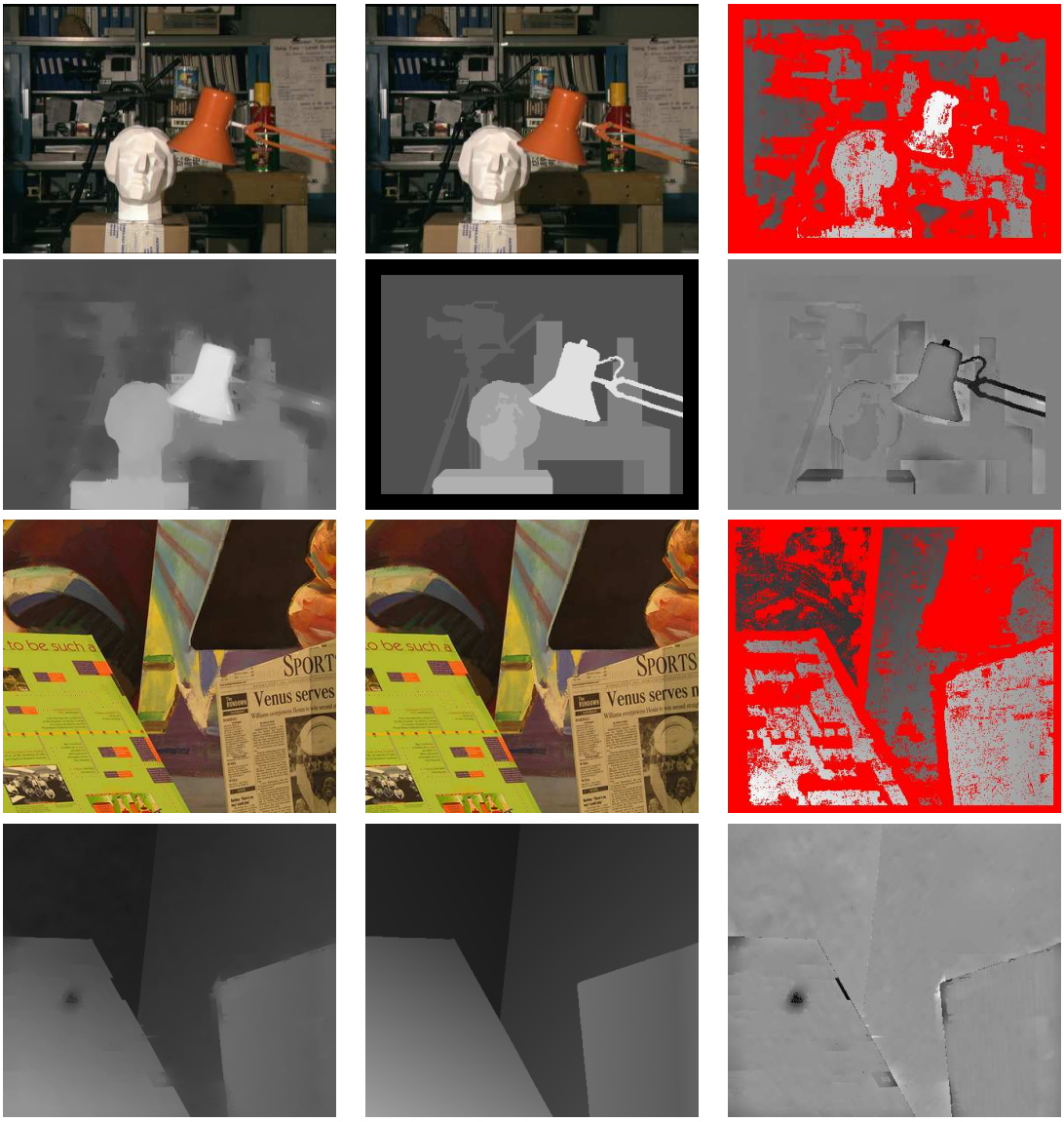


Figure E.1: Tsukuba and Venus results. For each couple of images: stereo pair of images, output of our algorithm (red points are the rejected correspondences), interpolated version of our results, ground truth and signed error between the interpolated image and the ground truth.

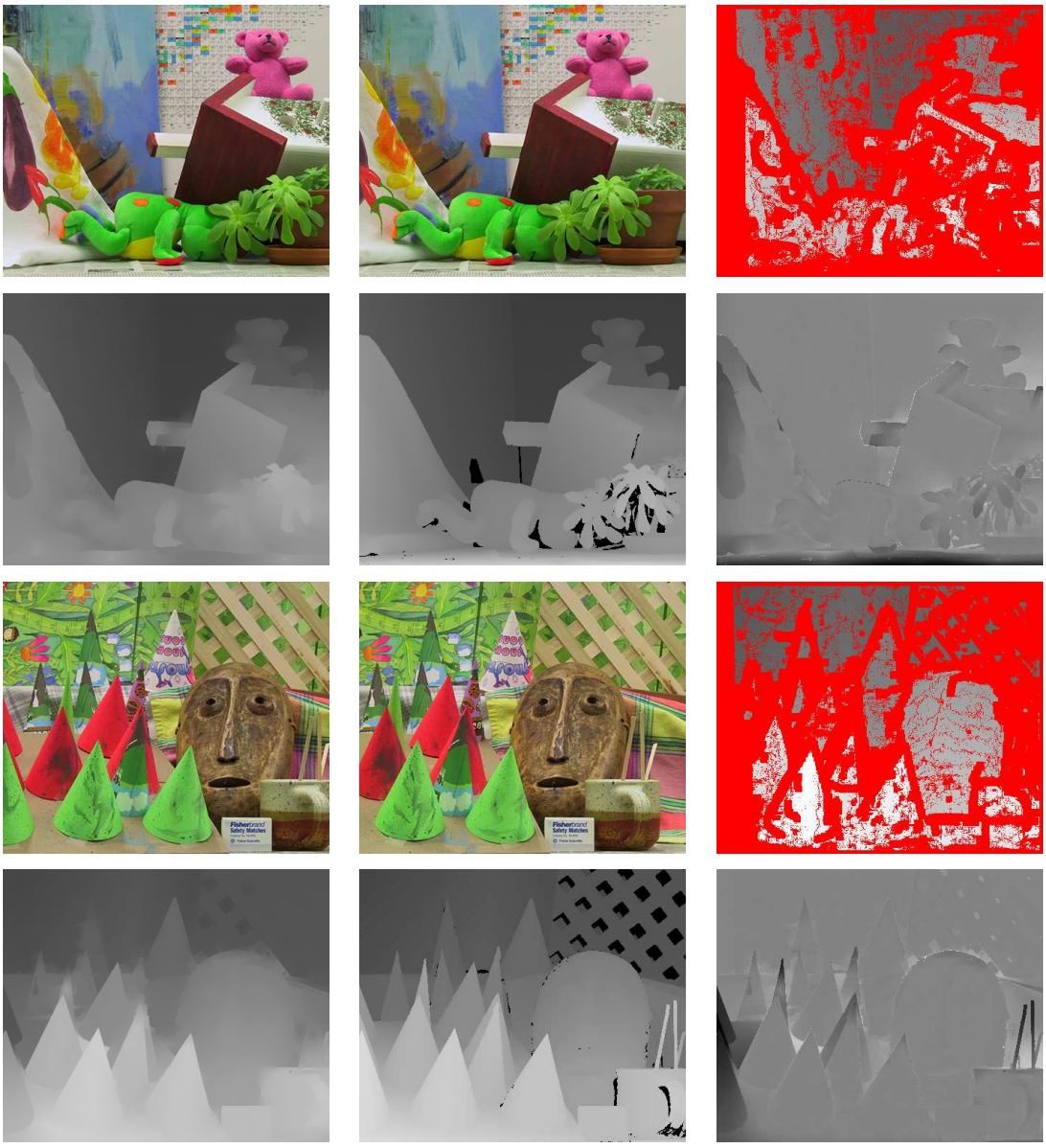


Figure E.2: Teddy and Cones results.

Bibliography

Aschwanden, P. and Guggenbuhl, W. (1993). Experimental results from a comparative study on correlation type registration algorithms. In Forstner, W. and Ruwiedel, S., editors, *Robust computer vision: Quality of Vision Algorithms*, pages 268–282. Wichmann.

Baillard, C. and Zisserman, A. (2000). A plane-sweep strategy for the 3D reconstruction of buildings from multiple images. *19th ISPRS Congress and Exhibition*.

Bay, H., Ferrari, V., and Van Gool, L. (2005). Wide-baseline stereo matching with line segments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 329–336.

Bhat, D. N. and Nayar, S. K. (1998). Ordinal measures for image correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):415–423.

Bignone, F., Henricsson, O., Fua, P., and Stricker, M. A. (1996). Automatic extraction of generic house roofs from high resolution aerial imagery. In *Proceedings of the European Conference on Computer Vision.*, volume I, pages 85–96. Springer-Verlag.

Birchfield, S. and Tomasi, C. (1998a). Depth discontinuities by pixel-to-pixel stereo. In *IEEE Conference Proceedings of International Conference on Computer Vision*, pages 1073–1080.

Birchfield, S. and Tomasi, C. (1998b). A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):401–406.

Birchfield, S. and Tomasi, C. (1999). Multiway cut for stereo and motion with slanted surfaces. *IEEE International Conference on Computer Vision*, 1:489.

Black, M. J. and Rangarajan, A. (1996). On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19:57–91.

Boykov, Y. and Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 26(9):1124–1137.

Boykov, Y., Veksler, O., and Zabih, R. (1998). A variable window approach to early vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1283–1294.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239.

Brown, M., Burschka, D., and Hager, G. (2003). Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008.

Buades, T., Lou, Y., Morel, J.-M., and Tang, Z. (2008). A note on multi-image denoising. In *Proceedings of International Workshop on Local and Non-Local Approximation in Image Processing (LNLA)*. (To appear).

Bughin, E. and Almansa, A. (2009). Planar patches detection in disparity maps. SMAI oral communication.

Burrus, N., Bernard, T. M., and Jolion, J.-M. (2009). Image segmentation by a contrario simulation. *Pattern Recognition*, 42(7):1520–1532.

Camlong, N. (2001). Manuel utilisateur de la chaîne de calcul de décalages entre images par l'algorithme marc. Technical Report cssi/lll-l/cor-et-marc-2, CNES.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698.

Cao, F. (2004). Application of the Gestalt principles to the detection of good continuations and corners in image level lines. *Computing and Visualisation in Science*, 7:3–13.

Cao, F., Delon, J., Desolneux, A., Muse, P., and Sur, F. (2007). A unified framework for detecting groups and application to shape recognition. *Journal of Mathematical Imaging and Vision*, 27(2):91–119.

Cao, F., Lisani, J.-L., Morel, J.-M., and Sur, P. M. F. (2008). *A Theory of Shape Identification*. Springer.

Carfantan, E. and Rougé, B. (2001). Estimation non biaisée de décalages subpixellaires sur les images spot. *GRETSI'01 on Signal and Image Processing*.

Caselles, V., Garrido, L., and Igual, L. (2005). A contrast invariant approach to motion estimation. *Scale Space and PDE Methods in Computer Vision*, 3459:242–253.

Caselles, V., Garrido, L., and Igual, L. (2006). A contrast invariant approach to motion estimation: Validation and application to motion estimation improvement. *Progress in Industrial Mathematics at ECMI 2006*, pages 863–867.

Chambon, S. and Crouzil, A. (2003). Dense matching using correlation: new measures that are robust near occlusions. *In Proc. British Machine Vision Conference (BMVC 2003)*, 1:143–152.

Chang, C., Chatterjee, S., and Kube, P. (1991). On an analysis of static occlusion in stereo vision. *Computer Vision and Pattern Recognition*, pages 722–723.

Chen, W., Zhang, M., and Xiong, Z. (2009). Segmentation-based stereo matching with occlusion handling via region border constrains. Submitted to CVIU.

Darrell, T. (1998). A radial cumulative similarity transform for robust image correspondence. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*.

Delon, J. (2004a). *Fine comparison of images and other problems*. PhD thesis, Ecole Normale Supérieure de Cachan.

Delon, J. (2004b). Midway image equalization. *Journal of Mathematical Imaging and Vision*, 21(2):119–134.

Delon, J. and B. Rougé, B. (2001). Le phénomène d'adhérence en stéréoscopie dépend du critère de corrélation. *GRETSI'01 on Signal and Image Processing*.

Delon, J. and Rougé, B. (2007). Small baseline stereovision. *Journal of Mathematical Imaging and Vision*, 28(3):209–223.

Delvit, J.-M. and Latry, C. (2006). Accurate geometric references from low b/h stereoscopic airbone acquisition. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 2840–2843.

Deng, Y. and Lin, X. (2006). A fast line segment based dense stereo algorithm using tree dynamic programming. In *Proceedings of the European Conference on Computer Vision*, pages 201–212.

Deriche, R. (1987). Using canny's criteria to derive a recursively implemented optimal edge detector. *International Journal of Computer Vision*, pages 167–187.

Desolneux, A., Moisan, L., and Morel, J. (2007). *From Gestalt Theory to Image Analysis. A probabilistic Approach*. Springer.

Dhond, U. and Aggarwal, J. (1989). Structure from stereo-a review. *IEEE Transactions on Systems, Man and Cybernetics*, 19:1489–1510.

Durou, J.-D. (2007). Shape from shading, eclairages, réflexions et persectives. Habilitation à Diriger des Recherches. Université Paul Sabatier (Toulouse).

Durou, J.-D., Aujol, J.-F., and Courteille, F. (2009). Integrating the normal field of a surface in the presence of discontinuities. In *Proceedings of the Internation Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 261–273.

Durou, J.-D. and Courteille, F. (2007). Integration of a normal field without boundary condition. In *Proceedings of the First International Workshop on Photometric Analysis For Computer Vision (PACV)*.

El-Etriby, S., Al-Hamadi, A. K., and Michaelis, B. (2006). Dense depth map reconstruction by phase difference-based algorithm under influence of perspective distortion. *Machine Graphics and Vision International Journal*, 15(3):349–361.

Facciolo, G. (2005). Variational adhesion correction with image based regularization for digital elevation models. Master's thesis, Facultad de ingenieria. Universidad de la Republica. Uruguay.

Facciolo, G. and Caselles, V. (2009). Geodesic neighborhoods for piecewise affine interpolation of sparse data. In *International Conference on Image Processing*.

Facciolo, G., Lecumberry, F., Almansa, A., Pardo, A., Caselles, V., and Rougé, B. (2006). Constrained anisotropic diffusion and some applications. In *British Machine Vision Conference*.

Farid, H. and Simoncelli, E. (2004). Differentiation of discrete multidimensional signals. *IEEE Transactions on Image Processing*, 13:496–508.

Faugeras, O., Hotz, B., Metthieu, H., Vieville, T., Zhangand, Z., Fua, P., Theron, E., Moll, L., Berry, G., Vuillemin, J., Bertin, P., and C.Proy (1993). Real-time correlation based stereo: algorithm, implementations and applications. Technical Report 2013.

Faugeras, O. and Keriven, R. (1998). Variational principles, surface evolution, pdes, level set methods, and the stereo problem. *IEEE Transactions on Image Processing*, 7:336–344.

Faugeras, O. and Luong, Q.-T. (2001). *The Geometry of Multiple Images: The Laws That Govern The Formation of Images of A Scene and Some of Their Applications*. The MIT Press, Cambridge, MA, USA.

Fleet, D. J., Jepson, A. D., and Jenkin, M. R. M. (1991). Phase-based disparity measurement. *Computer Vision, Graphics and Image Processing. Image Understanding*, 53:198–210.

Forstmann, S., Kanou, Y., Ohya, J., Thuering, S., and Schmitt, A. (2004). Real-time stereo by using dynamic programming. *In Conference on Computer Vision and Pattern Recognition Workshop*, 3:29–36.

Frohlinghaus, T. and Buhmann, J. (1996). Regularizing phased-based stereo. *International Conference on Pattern Recognition*, 1:451–455.

Fua, P. (1993). A parallel stereo algorithm that produces dense depth maps and preserves image features. *Machine vision and applications*, 6(1):35–49.

Fusiello, A., Roberto, V., and Trucco, E. (1997). Efficient stereo with multiple windowing. *International Journal of Pattern Recognitiond and Artificial Intelligence*, 14(8):858–863.

Garrido, L. and Salembier, P. (1998). Region based analysis of video sequences with a general merging algorithm. In *Eusipco: European signal processing conference*.

Gasquet, C. and Witomski, P. (1999). *Fourier Analysis and Applications: Filtering, Numerical Computation, Wavelets*. Springer.

Giros, A., Rougé, B., and Vadon, H. (2004). Appariement fin d'images sétréoscopiques et instrument dédié avec un faible coefficient stéréoscopique. French Patent N.0403143.

Gong, M. and Yang, R. (2005). Image-gradient-guided real-time stereo on graphics hardware. In *Proceedings of the Fifth International Conference on 3-D Digital Imaging and Modeling*, pages 548–555.

Gong, M., Yang, R., Wang, L., and Gong, M. (2007). A performance study on different cost aggregation approaches used in real-time stereo matching. *International Journal of Computer Vision*, 75(2):283–296.

Grimson, W. and Huttenlocher, D. (1991). On the verification of hypothesized matches in model-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(12).

Grompone, R. (2009). Detection of micro-vibrations on satellite stereo images. *SMAI*.

Grompone, R., Jakubowicz, J., Morel, J.-M., and Randall, G. (2008). Lsd: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (To appear).

Harris, C. G. and Stephens, M. (1988). A combined corner and edge detector. *4th Alvey Vision Conference*, pages 147–151.

Hartley, R. and Zisserman, A. (2000). *Multiple view geometry in computer vision*. Cambridge University Press, New York, NY, USA.

Hirschmuller, H. (2006). Stereo vision in structured environments by consistent semi-global matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2386–2393.

Hirschmuller, H., Innocent, P., and Garibaldi, J. (2002). Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47(1-3):229–246.

Hirschmuller, H. and Scharstein, D. (2007). Evaluation of cost functions for stereo matching. *Computer Vision and Pattern Recocgition*, pages 1–8.

Igual, L., Preciozzi, J., Garrido, L., Almansa, A., Caselles, V., and Rougé, B. (2007). Automatic low baseline stereo in urban areas. *Inverse Problems and Imaging*, 1(2):319–348.

Jakubowicz, J. (2007). *La recherche des alignements dans les images digitales et ses applications à l'imagerie satellitaire*. PhD thesis, Ecole normale supérieure de Cachan.

Julesz, B. (1960). Binocular depth perception of computer-generated patterns. *Bell System Technical Journal*, 39:1125–1162.

Kanade, T., Kato, H., Kimura, S., Yoshida, A., and Oda, K. (1995). Development of a video-rate stereo machine. In *Proc. of International Robotics and Systems Conference (IROS '95), Human Robot Interaction and Cooperative Robots*, volume 3, pages 95 – 100.

Kanade, T. and Okutomi, M. (1994). A stereo matching algorithm with an adaptive window: Theory and experiment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(9):920–932.

Kang, S. B., Szeliski, R., and Chai, J. (2001). Handling occlusions in dense multi-view stereo. *Computer Vision Pattern Recocgition*, 1:103–110.

Klaus, A. and Sormann, M.and Karner, K. (2006). Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proceedings of the International Conference on Pattern Recognition*, pages 15–18.

Kolmogorov, V. and Zabih, R. (2001). Computing visual correspondence with occlusions using graph cuts. *IEEE International Conference on Computer Vision*, 2:508–515.

Kolmogorov, V. and Zabih, R. (2005). *Graph Cut Algorithms for Binocular Stereo with Occlusions*. Mathematical Models in Computer Vision: The Handbook, Springer-Verlag.

Lafarge, F., Descombes, X., Zerubia, J., and Pierrot-Deseilligny, M. (2008). Automatic building extraction from dems using an object approach and application to the 3d-city modeling. *Journal of Photogrammetry and Remote Sensing*, 63(3):365–381.

Lavest, J.-M., Viala, M., and Dhome, M. (1998). Do we really need an accurate calibration pattern to achieve a reliable camera calibration? In *Proceedings of the 5th European Conference on Computer Vision-Volume I*, pages 158–174.

Lei, C., Selzer, J., and Yang, Y.-H. (2006). Region-tree based stereo using dynamic programming optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Loop, C. and Zhang, Z. (1999). Computing rectifying homographies for stereo vision. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:1125.

Lotti, J. and Giraudon, G. (1994). Correlation algorithm with adaptive window for aerial image in stereo vision. *In Image and Signal Processing for Remote Sensing*, 1:2315–10.

Lowe, D. (1985). *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *In Seventh International Joint Conference on Artificial Intelligence (IJCAI-81)*, pages 674–679.

Mairal, J., Elad, M., and Sapiro, G. (2008). Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69.

Manduchi, R. and Tomasi, C. (1999). Distinctiveness maps for image matching. *Proceedings of the International Conference on Image Analysis and Processing*, pages 26–31.

Marr, D. and Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, 194:283–287.

Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. In *Proceedings of the Royal Society of London B*, volume 204, pages 301–328.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767.

Mikolajczyk, K. and Schmid, C. (2003). A performance evaluation of local descriptors. *IEEE Conference on Computer Vision and Pattern Recognition*, 2:257–263.

Mittal, A. and Paragios, N. (2004). Motion-based background subtraction using adaptive kernel density estimation. *Computer Vision and Pattern Recognition*, 2:302–309.

Moisan, L. and Stival, B. (2004). A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *International Journal of Computer Vision*, 57(3):201–218.

Mordohai, P. and Medioni, G. (2006). Stereo using monocular cues within the tensor voting framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:968–982.

Morel, J.-M. and Yu, G. (2009). Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2.

Mühlmann, K., Maier, D., Männer, R., and Hesser, J. (2001). Calculating dense disparity maps from color stereo images, an efficient implementation. In *Proceedings of the IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV'01)*, page 30.

Muron, V. (2001). Manuel utilisateur de la chaîne de calcul de décalages entre images par l'algorithme marc. Technical Report cssi/lll-l/cor-et-marc-5, CNES.

Musé, P., Sur, F., Cao, F., Gousseau, Y., and Morel, J.-M. (2006a). An a contrario decision method for shape element recognition. *International Journal of Computer Vision*, 69(3):295–315.

Musé, P., Sur, F., Cao, F., Lisani, J.-L., and Morel, J.-M. (2006b). Theory of shape identification. Lectur Notes in Mathematics. Springer.

Musé, P., Sur, F., and Morel, J.-M. (2003). Sur les seuils de reconnaissance des formes. *Traitement du Signal*, 20(3):279–294.

Née, G., Jehan-Besson, S., Brun, L., and Revenu, M. (2008). Significance tests and statistical inequalities for region matching. *Structural, Syntactic, and Statistical Pattern Recognition*, pages 350–360.

Ohta, Y. and Kanade, T. (1985). Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(2):139–154.

Okutomi, M. and Kanade, T. (1993). A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):353–363.

Patwardhan, K. A., Sapiro, G., and Morellas, V. (2008). Robust foreground detection in video using pixel layers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):746–751.

Pollefeys, M. (2002). Visual 3d modeling from images. Technical report, University of North Carolina-Chapel Hill USA.

Prados, E. (2004). *Application of the theory of the viscosity solutions to the Shape From Shading problem.* PhD thesis, Uniersité de Nice-Sophia Antipolis (France).

Prazdny, K. (1987). Detection of binocular disparities. *Readings in computer vision: issues, problems, principles, and paradigms*, pages 73–79.

Rabin, J., Delon, J., and Y.Gousseau (2008). A contrario matching of sift-like descriptors. In *International Conference on Pattern Recognition*.

Randriamasy, S. Gagalowicz, A. (1991). Region based stereo matching oriented image processing. *Conference on Computer Vision and Pattern Recognition.*, pages 736–737.

Robert, L. and Faugeras, O. (1991). Curve-based stereo: figural continuity and curvature. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Robin, A., Moisan, L., and Le Hégarat-Mascle, S. (2009). An a-contrario approach for subpixel change detection in satellite imagery. (Tech. Repport MAP5 Nro. 2009-15).

Roy, S. and Cox, I. (1998). A maximum-flow formulation of the n-camera stereo correspondence problem. In *Proceedings of the Sixth International Conference on Computer Vision*, page 492.

Sabater, N., Morel, J.-M., and Almansa, A. (CMLA Preprint 2008-28. 2008). Rejecting wrong matches in stereovision.

Sara, R. (2002). Finding the largest unambiguous component of stereo matching. In *Proceedings of the European Conference on Computer Vision-Part III*, pages 900–914. Springer-Verlag.

Sara, R. and Bajcsy, R. (1997). On occluding contour artifacts in stereo vision. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 852–857.

Scharstein, D. and Szeliski, R. (1998). Stereo matching with nonlinear diffusion. *International Journal of Computer Vision*, 28(2):155–174.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(47(1/2/3)):7–42.

Schlüns, K. (1997). Shading based 3d shape recovery in the presence of shadows. In *Proceedings International Conference on Digital Image and Vision Computing*, pages 10–12.

Schmid, C. and Zisserman, A. (1997). Automatic line matching across views. *In Conference on Computer Vision and Pattern Recognition*, pages 666–671.

Schmid, C. and Zisserman, A. (2000). The geometry and matching of lines and curves over multiple views. *International Journal of Computer Vision*, 40(3):199–234.

Shannon, C. and Weaver, W. (1963). *The Mathematical Theory of Communication*. University of Illinois Press.

Shimizu, M. and Okutomi, M. (2001). Precise sub-pixel estimation on area-based matching. *International Conference on Computer Vision*, 1:90–97.

Stewart, C. (1995). Minpran: A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):925–938.

Szeliski, R. and Scharstein, D. (2002). Symmetric sub-pixel stereo matching. *European Conference on Computer Vision*, 2:525–540.

Szeliski, R. and Scharstein, D. (2004). Sampling the disparity space image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):419–425.

Szeliski, R. and Zabih, R. (1999). An experimental comparison of stereo algorithms. In *Proceedings of the International Workshop on Vision Algorithms*, pages 1–19.

Tian, Q. and Huhns, M. (1986). Algorithms for subpixel registration. *Computer Vision, Graphics and Image Processing*, 35(2):220–233.

Tomasi, C. and Manduchi, R. (1998a). Bilateral filtering for gray and color images. In *International Conference on Computer Vision*.

Tomasi, C. and Manduchi, R. (1998b). Stereo matching as a nearest-neighbor problem. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 20(3):333–340.

Tombari, F., Mattoccia, S., Di Stefano, L., and Addimanda, E. (2008). Classification and evaluation of cost aggregation methods for stereo correspondence. In *Computer Vision and Pattern Recognition*, pages 1–8.

Veksler, O. (2001). Stereo matching by compact windows via minimum ratio cycle. *International Conference on Computer Vision*, 1:540–547.

Veksler, O. (2002a). Dense features for semi-dense stereo correspondence. *International Journal of Computer Vision*, 47(1-3):247–260.

Veksler, O. (2002b). Stereo correspondence with compact windows via minimum ratio cycle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1654–1660.

Veksler, O. (2003a). Extracting dense features for visual correspondence with graph cuts. In *Computer Vision and Pattern Recognition*, volume 1, pages 689–694.

Veksler, O. (2003b). Fast variable window for stereo correspondence using integral images. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:556–561.

Venkateswar, V. and Chellappa, R. (1995). Hierarchical stereo and motion correspondence using feature groupings. *Internationl Journal of Computer Vision*, 15(3):245–269.

Wang, L., Liao, M., Gong, M., Yang, R., and Nister, D. (2006). High-quality real-time stereo using adaptive cost aggregation and dynamic programming. In *Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 798–805.

Wang, Z.-F. and Zheng, Z.-G. (2008). A region based stereo matching algorithm using cooperative optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Weng, J. J. (1993). Image matching using the windowed fourier phase. *International Journal of Computer Vision*, 11(3):211–236.

Wu, Y. and Maître, H. (1989). A Dynamic Programming Algorithm for Stereo Matching Using Inter-line Coherence. In *12ème Colloque GRETSI*, pages 751–754, Juan les Pins, France.

Xu, Y., Wang, D., Feng, T., and Shum, H.-Y. (2002). Stereo computation using radial adaptive windows. *International Conference on Pattern Recognition*, 3:595–598.

Yang, Q., Wang, L., Yang, R., Stewenius, H., and Nister, D. (2006). Stereo matching with color-weighted correlation, hierachical belief propagation and occlusion handling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2354.

Yang, Q., Yang, R., Davis, J., and Nister, D. (2007). Spatial-depth super resolution for range images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Yoon, K.-J.and Kweon, S. (2006). Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650–656.

Zabih, R. and Woodfill, J. (1994). Non-parametric local transforms for computing visual correspondence. *Proceedings of the European conference on Computer Vision*, 2:151–158.

Zhang, Z. (1998). Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, 27(2):161–195.