

THESE

En vue de l'obtention du : **DOCTORAT**

Structure de Recherche : Laboratoire de Recherche en Informatique et
Télécommunications (LRIT)

Discipline : Sciences de l'ingénieur

Spécialité : Informatique

Présentée et soutenue le 11/12/2021 par :

Oussama ZEGLAZI

**Mise en correspondance stéréoscopique pour les systèmes de
transport intelligents**

JURY

Ahmed TAMTAOUI	PES, Institut National des Postes et Télécommunications, Rabat, Maroc	et Président
Hicham LAANAYA	PH, Faculté des Sciences, Université Mohammed V, Rabat, Maroc	Rapporteur / Examineur
Amina RADGUI	PH, Institut National des Postes et Télécommunications, Rabat, Maroc	Rapporteuse/ Examinatrice
Rachid ALAOUI	PH, École Supérieure de Technologie, Salé, Maroc	Rapporteur / Examineur
Cédric DEMONCEAUX	Professeur des Universités, Université de Bourgogne Franche-Comté, France	Examineur
Aouatif AMINE	PES, École Nationale des Science Appliqués, Université Ibn Tofail, Kenitra, Maroc	Co-directrice de thèse
Mohammed RZIZA	PES, Faculté des Sciences, Université Mohammed V, Rabat, Maroc	Directeur de thèse

Année Universitaire : 2020 - 2021



Université Mohammed V
Rabat



Faculté des Sciences
Rabat

Thèse de doctorat

Présentée par :

ZEGLAZI Oussama

Discipline: Sciences de l'ingénieur

Spécialité: Réseaux et Télécommunications

Mise en correspondance stéréoscopique pour les systèmes de transport intelligents

Soutenue le 11/12/2021 devant la commission d'examen

Président :

PR. AHMED TAMTAOUI

PES, Institut National des Postes et Télécommunications,
Rabat, Maroc

Examineurs :

PR. HICHAM LAANAYA

PH, Faculté des Sciences, Université Mohammed V, Rabat,
Maroc

PR. AMINA RADGUI

PH, Institut National des Postes et Télécommunications,
Rabat, Maroc

PR. RACHID ALAOUI

PH, École Supérieure de Technologie, Salé, Maroc

PR. CÉDRIC DEMON-
CEAUX

Professeur des Universités, Université de Bourgogne
Franche-Comté, France

PR. AOUATIF AMINE

PES, École Nationale des Science Appliqués, Université Ibn
Tofail, Kenitra, Maroc

PR. MOHAMMED RZIZA

PES, Faculté des Sciences, Université Mohammed V, Rabat,
Maroc

Travaux préparés au laboratoire : LRIT-CNRST (URAC-29) de la Faculté des Sciences de Rabat.



REMERCIEMENTS

Les travaux présentés dans cette thèse ont été réalisés au sein du Laboratoire de Recherche en Informatique et Télécommunications (Rabat-IT-Center) de la Faculté des Sciences de Rabat (FSR), Université Mohammed V au Maroc sous la direction du Mr. **Mohammed RZIZA**, professeur d'enseignement supérieur, à la faculté des Sciences de Rabat, et le co-encadrement de Mme. **Aouatif AMINE**, professeur d'enseignement supérieur, à l'université Ibn Tofail de Kénitra.

Tout d'abord, je tiens à adresser toute ma gratitude à mon directeur de thèse, Monsieur **Mohammed RZIZA**, qui m'a ouvert les portes de la recherche et m'a donné la chance de faire une thèse. Je le remercie pour sa disponibilité absolue, son accompagnement, ses conseils avisés ainsi que son soutien scientifique. Ses qualités humaines et professionnelles ont contribué à la réalisation de ce travail de recherche dans les meilleures conditions possibles.

Mes sincères remerciements vont également à Mme. **Aouatif AMINE** qui m'a soutenu avec ses précieux conseils scientifiques et pour ses qualités humaines. Sans son aide et le soutien qu'elle m'a apporté, cette thèse n'aurait jamais vu le jour.

Je tiens également à adresser mes remerciements à tous les membres du jury pour leur attention à la lecture du manuscrit, leurs questions pointues ainsi que leurs remarques constructives. Je remercie vivement Mr. **Ahmed TAMTAOUI**, professeur d'enseignement supérieur à l'Institut National des Postes et Télécommunications de Rabat, pour avoir accepté de présider le jury de thèse, ce qui représente un grand honneur pour moi.

J'adresse tous mes remerciements à Mr. **Hicham LAANAYA**, professeur habilité à Faculté des Sciences de Rabat, pour avoir accepté de prendre en charge le rôle de rapporteur de mon travail.

J'adresse mes plus vifs remerciements à Mme. **Amina RADGUI**, professeur habilité à l'Institut National des Postes et Télécommunications de Rabat, pour avoir assumé la responsabilité de rapporter cette thèse.

Je tiens à exprimer mes remerciements à Mr. **Rachid ALAOUI**, professeur habilité à l'École Supérieur de Technologie de Salé, pour avoir accepté d'examiner ma thèse.

Je remercie sincèrement Mr. **Cédric DEMONCEAUX**, professeur des universités à l'université de Bourgogne Franche-Comté de France, pour sa précieuse expertise, sa disponibilité et la pertinence de ses observations. Je leur suis également reconnaissant pour l'intérêt qu'ils ont accordé à ce travail en acceptant de faire partie du jury.

Je tiens à remercier chaleureusement tous les membres du laboratoire LRIT que j'ai eu le plaisir de côtoyer quotidiennement pendant ces années de thèse. Je remercie particulièrement mes amis qui m'ont toujours apporté leur aide, et leurs encouragements. Je remercie ma famille pour l'amour et le soutien qu'elle m'a toujours témoigné. Un grand merci à mes parents pour tout l'amour et l'éducation qu'ils m'ont donné. Merci également pour leurs encouragements dans les moments difficiles et leur disponibilité.



RÉSUMÉ

La sécurité des usagers du trafic routier constitue la préoccupation centrale des Systèmes de Transport Intelligents (STI). A cet effet, la détection des obstacles et notamment celle des piétons, est un élément fondamental des STI. Bien que ce thème ait été étudié depuis près de cinquante ans, une solution idéale à ce problème se fait toujours attendre. Les méthodes courantes utilisées pour la détection des piétons reposent sur l'utilisation du LIDAR (Détection et estimation de la distance par la lumière) et de la vision par ordinateur. Dans cette thèse, on opte pour la vision par ordinateur, notamment pour la vision stéréoscopique, qui fournit des moyens de réduction de l'espace de recherche des objets dans les scènes routières grâce à l'utilisation des informations de profondeur afin d'obtenir des estimations fiables de l'emplacement des piétons. Ainsi, il est essentiel de disposer d'une carte de disparité robuste afin de dégager des régions de recherche fiables. Plusieurs approches sont présentées dans cette thèse en vue de garantir une meilleure qualité des cartes de disparité en tenant compte des principaux problèmes observés dans le domaine de la vision stéréoscopique pour des scènes réelles, tels que les changements de métrique radio et les régions non texturées.

Mots-clés : Mise en correspondance stéréoscopique ; Systèmes de transport intelligents ; Carte de Disparité ; KITTI ; Middlebury.



ABSTRACT

The safety of road users is the central concern of Intelligent Transport Systems (ITS). For this purpose, the detection of obstacles, especially pedestrians, is a fundamental component of ITS. Although this topic has been studied for almost fifty years, an ideal solution to this problem is still pending. The current methods used for pedestrian detection are based on the use of LIDAR and computer vision. In this thesis, computer vision is opted for, especially stereovision, which provides ways to reduce the search space of objects in road scenes through the use of depth information to obtain reliable estimates of pedestrian locations. Thus, a robust disparity map is essential to derive reliable search regions. Several approaches are presented in this thesis to ensure better quality disparity maps by addressing the main problems observed in stereo vision for real scenes, such as radiometric changes and non-textured regions.

Key Words : Stereo matching, Intelligent Transportation Systems, Disparity Map, KITTI, Middlebury.



TABLE DES MATIÈRES

Remerciements	i
Résumé	iii
Abstract	v
Liste des notations et abréviations	xi
Liste des figures	xvi
Introduction	1
0.0.1 Les systèmes multi-capteurs	1
0.0.2 Les systèmes de vision	2
0.0.3 La stéréovision	3
0.0.4 Problématique	3
0.0.5 Organisation du document	4
Liste des tableaux	1
Chapitre 1 : État de l'art : Stéréovision	5
1.1 Introduction	5
1.2 Les principes de la vision stéréoscopique	6
1.2.1 Modèle géométrique de la caméra	6
1.2.2 Modèle du capteur stéréoscopique	7
1.2.3 La mise en correspondance stéréoscopique	8
1.2.4 Les difficultés de la mise en correspondance	8
1.3 Les algorithmes de mise en correspondance stéréoscopique	14
1.3.1 Méthodes locales	15
1.3.2 Méthodes globales	18
1.4 Les bases de données de la stéréovision	20
1.5 Fonctions de coûts	21

1.5.1	Les coûts mixtes	25
1.6	Conclusion	26
Chapitre 2 : Mise en correspondance stéréoscopique local pour les scènes routières		27
2.1	Introduction	27
2.2	La méthode proposée	28
2.2.1	Fonction du coût proposée	28
2.2.2	Agrégation des coûts	29
2.2.3	Post-traitement	31
2.3	Les résultats expérimentaux	32
2.3.1	Évaluation de la fonction de coût proposée	32
2.3.2	Évaluation de la méthode d'agrégation croisée modifiée	33
2.3.3	Évaluation de l'algorithme proposé sur la plateforme KITTI	33
2.4	Conclusion	34
Chapitre 3 : Approche hiérarchique		35
3.1	Introduction	35
3.2	Fonction de coût proposée	36
3.3	Approche hiérarchique	38
3.3.1	Méthode d'agrégation adaptative	38
3.3.2	Schéma hiérarchique de traitement	41
3.4	Résultats expérimentaux	42
3.4.1	Évaluation de la fonction du coût proposée	43
3.4.2	Évaluation de l'algorithme hiérarchique proposé	44
3.4.3	Évaluation sur la plateforme en ligne de la base KITTI	45
3.4.4	Comparaison avec d'autres algorithmes hiérarchiques	45
3.4.5	Complexité algorithmique	45
3.5	Conclusion	47
Chapitre 4 : Coût basé sur les informations structurelles pour les algorithmes de correspondance des scènes routières		49
4.1	Introduction	49
4.2	Fonctions de coût proposées	50
4.2.1	Fonction de coût basée sur le SSIM (C_{SSIM})	50
4.2.2	SSIM gradient variant (C_{GSSIM})	51
4.3	Les résultats expérimentaux	52
4.3.1	Évaluation de la capacité discriminatoire des coûts proposés	52
4.3.2	Évaluation des coûts proposés avec une technique d'agrégation adaptative	55
4.3.3	Sensibilité des fonctions de coût face aux distorsions radiométriques	56
4.4	Discussion	56

4.5	Conclusion	58
Chapitre 5 : Génération de cartes de disparité denses à partir des mesures éparses		59
5.1	Introduction	59
5.2	Méthode proposée	60
5.2.1	Échantillonnage de données	60
5.2.2	Méthode de propagation	61
5.2.3	Filtre médian vertical adaptatif	62
5.3	Résultats expérimentaux	63
5.3.1	Résultats de reconstruction en utilisant les mesures de la carte de disparité vérité du terrain	63
5.3.2	Résultats de reconstruction en utilisant les autres cartes de disparités	67
5.4	Conclusion	68
Conclusion		69
Liste des publications		71
Bibliographie		73

LISTE DES FIGURES

1.1	Modèle géométrique de la caméra	6
1.2	Modèle géométrique du capteur stéréoscopique binoculaire.	7
1.3	Géométrie épipolaire.	11
1.4	Exemple de violation de la contrainte d'unicité : tous les pixels du segment de droite $[M_1, M_2]$ sont projetés sur le pixel m de l'image de droite.	13
1.5	La contrainte d'ordre.	13
1.6	Mise en correspondance par corrélation.	15
1.7	Exemple d'un pixel représenté en rouge et sa fenêtre d'agrégation associée ayant des valeurs de disparité différentes.	16
1.8	Illustration de la Transformée Census	24
1.9	Illustration de la Transformée Rank	24
2.1	Construction de la chaîne de bits pour CCC : "000000" (a) et pour notre variante : "00010011111011110" (b)	28
2.2	De haut en bas et de gauche à droite : l'image de gauche, la carte de disparité de la vérité du sol, la carte de disparité calculée sur la base de C_{CCC} et la carte de disparité utilisant le coût proposé	30
2.3	De gauche à droite : image de gauche, carte de disparité de la vérité du terrain, image de contour, carte de disparité calculée en fonction de la région adaptative croisée, et carte de disparité obtenue par la méthode modifiée.	31
2.4	Erreurs de disparité moyennes par rapport aux valeurs de $step_2$ pour les 10 premières images des ensembles d'entraînement KITTI	32
3.1	Quelques résultats des cartes de disparités produites via la fonction du coût classique CT et la fonction du coût proposée pour les deux images stéréo #7 et #131 issues de la base de données KITTI 2012. de haut en bas : L'image de référence, la carte de disparité calculée en utilisant la fonction du coût CT , la carte de disparité calculée à partir de la fonction du coût proposée et finalement les vérités du terrain. Les cartes de disparités sont calculées après l'utilisation de la méthode d'agrégation [76]. Les régions cadrées représentent des zones d'amélioration.	39

3.2	méthode adaptative pour le pixel p : (a) un squelette est construit pour chaque pixel. La région adaptative pour p est effectuée en regroupant les branches horizontales des pixels (q par exemple). (b) le coût du pixel p est rassemblé en deux passages sur la direction horizontale et verticale. . .	40
3.3	L'organigramme de l'approche hiérarchique	42
3.4	Les cartes de disparité pour les deux images stéréo #0 et #2 de la base d'entraînement de KITTI. Du haut en bas : L'image de référence, carte de disparité produite sans l'utilisation du schéma hiérarchique, Carte de disparité produite en utilisant l'approche hiérarchique et les cartes de disparité vérité du terrain, les zones encadrées sont des régions non texturées	44
4.1	Résultats de disparité. La première colonne correspond à deux paires stéréoscopiques à gauche (#0 et #7) de la base de données KITTI 2012 avec les cartes de disparité de la vérité du terrain correspondantes. Les cartes de disparité calculées sont répertoriées dans les lignes suivantes, correspondant aux fonctions de coût $C_{DiffCensus}$, C_{GCCC} , C_{SSIM} et C_{GSSIM} , respectivement. La première et la troisième colonne correspondent à la solution obtenue sans l'utilisation d'une méthode d'agrégation, tandis que la deuxième et la quatrième colonne correspondent à la solution basée sur la méthode d'agrégation adaptative.	54
4.2	Les résultats de disparité obtenus par rapport à la taille de la fenêtre de support pour les deux fonctions de coût, pour les ensembles de formation KITTI 2012.	55
4.3	Les résultats de disparité obtenus par rapport aux distorsions radiométriques pour les fonctions de coût présentées, dans le cas de la base de donnée KITTI 2012.	57
5.1	Diagramme de la méthode proposée en utilisant comme exemple l'image stéréo "Tsukuba"	60
5.2	Exemple de vue rapprochée de la segmentation en super pixel de la paire stéréo Tsukuba, De gauche à droite : Image de référence, La vérité de terrain, graphe de superpixel sur la carte de disparité avec 2000, 2200, 2400 and 2600 nombre de super pixels, respectivement.	61
5.3	Résultats du processus de propagation pour les quatre paires stéréoscopiques de Middlebury (Tsukuba, Venus, Teddy, Cones). Du haut vers le bas : Les cartes de disparité produites, les erreurs de disparité (marquées en noir) dans toutes les régions avec un seuil d'erreur d'un pixel.	62
5.4	Résultats de disparité obtenus par rapport aux nombre de super pixels avec les images Tsukuba, Venus, Teddy et Cones	64
5.5	Résultats de disparité obtenus par rapport aux nombre de super pixels avec les images Tsukuba, Venus, Teddy et Cones	65
5.6	Pourcentage des échantillons pour chaque pair pour la nouvelle base Middlebury avec un échantillonnage de 2600 super pixels	66

5.7	Résultats sur la nouvelle base Middlebury pour (a) ArtL. (b) Piano. (c) Playtable. (d) Shelves. Du gauche à droite : l'image de référence, cartes de disparité produites par [51], [39], notre approche, vérité du terrain, respectivement	67
-----	--	----



LISTE DES TABLEAUX

2.1	Pourcentage de disparités incorrectes dans les régions occultées et non-occultées pour l'ensemble d'entraînement KITTI.	33
2.2	Pourcentage de disparités incorrectes dans les régions non occultées pour la base d'entraînement de Middlebury.	33
2.3	Résultats de la plateforme d'évaluation KITTI pour le seuil par défaut de 3 pixels. Colonnes de gauche à droite : méthode ; pourcentage de pixels erronés dans les régions non-occultées et au total ; erreur moyenne de disparité dans les régions non-occultées ; erreur moyenne de disparité au total. Date de l'évaluation : 17 janvier 2017	34
3.1	Pourcentage de disparités erronées dans les régions non occultées (Nocc) et les autres régions (toutes) pour la base d'entraînement KITTI	43
3.2	Pourcentage de disparités incorrectes dans la base de données de d'entraînement KITTI	43
3.3	Les résultats obtenus par la plateforme d'évaluation KITTI pour le seuil par défaut de 3 pixels. Les colonnes de gauche à droite : méthode ; pourcentage de pixels incorrects dans les régions non-occultées et au total ; erreur moyenne de disparité dans les régions non-occultées ; erreur moyenne de disparité au total. Date de l'évaluation : 06 mai 2017	46
3.4	Temps d'exécution moyen (en s) nécessaire au calcul des cartes de disparité pour les paires d'images KITTI	46
4.1	Pourcentage de disparités incorrectes de la correspondance stéréoscopique sans l'utilisation d'une méthode d'agrégation pour la base d'entraînement KITTI 2012	53
4.2	les résultats sur les bases de données de formation KITTI 2015	53
4.3	Pourcentage de disparités incorrectes dans les régions non occultées et les autres régions pour l'ensemble de formation KITTI 2012	55
5.1	Rapports d'échantillonnage pour chaque valeur de superpixel pour les paires stéréo Tsukuba, Venus, Teddy et Cones.	63

5.2	Pourcentage d'erreurs dans différentes régions (nonocc, all , disc) avec un seuil de 1 pixel	66
5.3	Pourcentage des erreurs de disparité dans les régions non occultés pour la base Middlebury 2014	68



INTRODUCTION

Introduction

La sûreté des moyens de transport a toujours été une préoccupation particulière, notamment la voiture, statistiquement le moyen de transport le moins sûr. En effet, conformément à de nombreuses enquêtes en accidentologie, une collision se présente comme une défaillance du système associant le conducteur, le véhicule et l'environnement. En effet, le facteur humain est la cause de l'accident dans la grande majorité des cas, et ce, notamment, en raison de sa perception réduite de l'environnement routier ou de son mauvais contrôle du véhicule.

Pour pallier les conséquences des erreurs humaines, des scientifiques se penchent sur la mise en place de dispositifs embarqués pour l'aide à la conduite ayant pour objectif de prévenir le conducteur du danger que représente la situation dans laquelle il se trouve, ou même d'agir de façon préventive sur le véhicule.

Ces systèmes embarqués permettent en effet la reconstitution à tout moment de la situation routière entourant le véhicule grâce à une perception multi-capteurs et une étape de fusion de données. Des indicateurs sont tirés de cette représentation dynamique de l'environnement, et utilisés pour prévenir le conducteur. Le moyen d'alerte peut être constitué d'un ensemble de signaux visuels ou sonores adressés au conducteur. Par ailleurs, ces mêmes indicateurs peuvent être également considérés comme des entrées d'un module de contrôle destiné à agir sur le véhicule afin de réduire le niveau de danger de la situation ou d'éviter un accident.

Les premières études ont abouti au développement de systèmes composés d'un capteur et d'une station de calcul chargés de contrôler les alentours du véhicule. Leur objectif consiste en effet à reconstruire l'environnement en respectant certaines règles : la détection des objets présents dans la scène, leur localisation vis-à-vis du véhicule, ainsi que le suivi et le contrôle de la trajectoire de ces objets.

0.0.1 Les systèmes multi-capteurs

Les systèmes multi-capteurs sont caractérisés par différents types de capteurs qui peuvent être actifs ou passifs [38]. Les capteurs employés relèvent de différentes classes, à savoir des technologies leur permettant, notamment, de percevoir une scène dans plu-

sieurs domaines fréquentiels, allant du domaine des micro-ondes et des ondes RADAR (longueurs d'onde de quelques millimètres) à la gamme infrarouge (longueurs d'onde de quelques centaines de nanomètres). Ainsi, par le biais d'un traitement particulier des signaux émis par chaque capteur, l'environnement observé est ainsi représenté par des données à la fois complémentaires et répétitives. Le mode de fonctionnement de chacun de ces capteurs varie en fonction, notamment, des conditions d'éclairage ou climatiques. La complémentarité entre ces capteurs est envisageable [60] : par exemple, si un objet est invisible dans le noir pour un capteur donné, il est possible de le repérer pour un autre. De plus, la combinaison d'informations redondantes provenant de plusieurs capteurs augmente la fiabilité des résultats obtenus, notamment en cas de défaillance de l'un des capteurs.

Les capteurs actifs, comme le RADAR et le LIDAR, produisent leurs propres signaux pour calculer la portée des objets. Cette mesure est effectuée en fonction de la durée du trajet d'une onde électromagnétique envoyée et ensuite reflétée par un objet. Le RADAR opère dans le spectre des hyperfréquences. Il offre la possibilité de mesurer des portées à grande distance, indépendamment des conditions climatiques. Toutefois, ce capteur reste particulièrement vulnérable face aux contraintes liées à la scène étudiée, comme la pente ou les virages. De plus, en raison de sa faible résolution angulaire, le RADAR est incapable de repérer les petits obstacles. Le LIDAR agit dans le spectre des ondes LASER. Cette technologie offre la possibilité de maîtriser la navigation en milieu ouvert, et ce, particulièrement dans le cadre de l'application de robotique mobile. Pourtant, sa taille, sa fragilité et son coût élevé représentent des contraintes importantes pour sa mise en œuvre.

À la différence des capteurs actifs, les capteurs passifs, notamment les caméras, ne produisent pas de signaux. La caméra peut en effet percevoir une scène par la réflexion de la lumière des objets qui la constitue. Elle produit donc une image de la scène en la projetant sur un capteur photosensible 1D (caméra linéaire) ou 2D (caméra matricielle). L'image ainsi obtenue comporte une information très complète que celle véhiculée par le signal capté par un RADAR ou un LIDAR. Certains systèmes de vision peuvent être actifs lorsqu'ils comportent un système d'éclairage tel que le LED infrarouges ou une source de lumière structurée.

0.0.2 Les systèmes de vision

Le système de vision est considéré comme mono-caméra si une seule caméra est employée. Dans le cas où il comporte deux caméras, le système est nommé système stéréoscopique, on parle dans ce cas de la stéréovision. Les modèles utilisant la vision monoculaire peuvent, dans la plupart des cas, détecter les obstacles routiers. Cependant leurs distances restent impossibles à mesurer précisément sans l'ajout d'un autre capteur ou de connaissances a priori sur la scène. Les informations de distance sont indispensables pour repérer les objets qui représentent un obstacle éventuel pour le véhicule. Les approches basées sur la vision stéréoscopique permettent de retrouver cette information par un principe de triangulation, qui consiste à calculer les coordonnées 3D d'un point

à partir de ses projections dans les images.

0.0.3 La stéréovision

La stéréovision permet de déterminer les coordonnées $3D$ de chaque point de la scène grâce aux coordonnées de leur projection sur les deux images issues de deux caméras. La tâche cruciale de tout algorithme de stéréovision consiste à la mise en correspondance des primitives entre les deux images. Ces primitives peuvent être des points de contour, des segments ou simplement des pixels. La mise en correspondance repose sur la recherche du correspondant de chaque primitive d'une image dans l'autre image. Peu importe la position respective des deux caméras, le correspondant d'une primitive de la première image se trouve sur une droite dans la seconde image. Cette ligne est appelée ligne épipolaire et représente la zone de l'image dans laquelle la recherche peut être menée. Ainsi, le problème de l'étape de mise en correspondance devient un problème de type $1D$. De plus, en présence de la phase de calibration préalable, les lignes épipolaires deviennent les lignes de l'image. Ainsi, la position des primitives varie uniquement en fonction de leurs abscisses le long de ces lignes épipolaires, on appelle ce déplacement *disparité*. Le résultat final est une carte de disparité dont la densité dépend de la primitive utilisée. Si la primitive est l'ensemble des pixels de l'image, la carte de disparité est dense. Dans le cas contraire, la carte de disparité obtenue est éparse. La valeur de disparité du pixel permet de calculer la profondeur du point correspondant dans la scène réelle.

Cette thèse porte sur la mise en correspondance stéréoscopique dense afin de détecter les obstacles routiers, à partir d'un système de vision stéréoscopique embarqué dans un véhicule en mouvement.

0.0.4 Problématique

Les techniques de mise en correspondance par corrélation se basent sur la similarité de deux groupes de pixels, cette similarité est calculée grâce à une mesure de similarité. Dans la littérature, plusieurs mesures ont été proposées afin de répondre aux différentes difficultés de la mise en correspondance, telles que les changements radio métriques, les régions non texturées ou les occultations. Le choix d'une mesure de similarité est compliqué et dépend notamment des situations étudiées. Par conséquent, nous avons jugé indispensable, dans un premier temps, de recenser, de décrire et d'évaluer les mesures de similarité existantes.

Un des problèmes relativement peu abordé dans la littérature sur la mise en correspondance concerne la prise en considération des variations d'illumination provoquées entre les images stéréoscopiques. Ce problème est primordial, car en présence de ces variations, l'hypothèse de conservation de la luminance sur laquelle la plupart des méthodes de mise en correspondance sont fondées, se voit défectueuse. Notre étude dans ce contexte indique que ces variations peuvent être palliées en proposant des mesures de similarité convenables à ce problème.

0.0.5 Organisation du document

Cette thèse s’articule autour de six chapitres : bien que la finalité de cette étude porte sur la détection d’obstacles dans un environnement routier par voie de vision stéréoscopique, notre démarche consiste à traiter le problème de mise en correspondance. En effet, les méthodes de mise en correspondance présentées dans ce travail sont tout à fait génériques et sont donc applicables à tous les types d’images stéréoscopiques acquises par deux caméras.

Dans le chapitre 1, nous étudions certains éléments généraux de la mise en correspondance, qui nous seront utiles ultérieurement. D’abord, nous discutons des contraintes fréquemment utilisées pour surmonter les difficultés de ce problème. Ensuite, nous faisons un bref état de l’art des méthodes d’estimation de la disparité, ce qui nous amène à adopter des méthodes locales pour traiter ce problème.

Le chapitre 2, présente notre première contribution pour l’estimation de la carte de disparité. Nous proposons deux améliorations à l’algorithme classique qui concernent notamment le calcul et l’agrégation des coûts afin de traiter deux problèmes majeurs de la correspondance stéréoscopique, à savoir les changements radiométriques et les zones non texturées.

Dans le chapitre 3, nous abordons un autre type d’algorithmes de mise en correspondance stéréoscopique, à savoir les méthodes hiérarchiques. Dans ce contexte, nous proposons une nouvelle méthode locale dans un schéma hiérarchique, ce qui permet d’améliorer significativement les disparités tout en conservant un temps de calcul proche du temps réel.

Le chapitre 4, se penche sur le problème des changements d’illumination dans la scène observée. Nous proposons un nouveau coût basé sur l’information pour traiter ce problème, nous comparons le coût proposé aux meilleurs coûts de la littérature, sur des bases de données représentant des scènes routières réelles.

Dans le chapitre 5, nous proposons un nouvel algorithme de mise en correspondance stéréoscopique qui permet de reconstruire efficacement une carte de disparité dense à partir de quelques mesures de disparité éparses. L’algorithme procède d’abord au calcul des disparités pour des points spécifiques dans l’image, puis estime les disparités restantes pour la totalité des pixels à l’aide d’une méthode de propagation particulière.

Dans la conclusion générale, nous présentons une synthèse de notre étude et nous traçons quelques perspectives dans la continuité des travaux menés.

ÉTAT DE L'ART : STÉRÉOVISION

Sommaire

1.1	Introduction	5
1.2	Les principes de la vision stéréoscopique	6
1.2.1	Modèle géométrique de la caméra	6
1.2.2	Modèle du capteur stéréoscopique	7
1.2.3	La mise en correspondance stéréoscopique	8
1.2.4	Les difficultés de la mise en correspondance	8
1.2.4.1	Les primitives à mettre en correspondance	9
1.2.4.2	Contraintes de la mise en correspondance	10
1.3	Les algorithmes de mise en correspondance stéréoscopique	14
1.3.1	Méthodes locales	15
1.3.2	Méthodes globales	18
1.3.2.1	Programmation dynamique	18
1.3.2.2	Coupures de graphes	19
1.3.2.3	Propagation de croyance	19
1.4	Les bases de données de la stéréo vision	20
1.5	Fonctions de coûts	21
1.5.0.1	Les coûts paramétriques	22
1.5.0.2	Les coûts non paramétriques	23
1.5.1	Les coûts mixtes	25
1.6	Conclusion	26

1.1 Introduction

L'objectif de ce chapitre est d'introduire les concepts nécessaires à la compréhension des problèmes de la cartographie stéréo. Bien que ce domaine ait été le sujet de nombreuses études ces dernières années, la génération de cartes de profondeur en temps réel avec de faibles ressources constitue toujours un défi ouvert. Sur le plan biologique, les êtres humains sont tout à fait capables de repérer les informations de profondeur et de détecter les obstacles à travers leurs yeux. Une tâche qui est si facilement acquise par

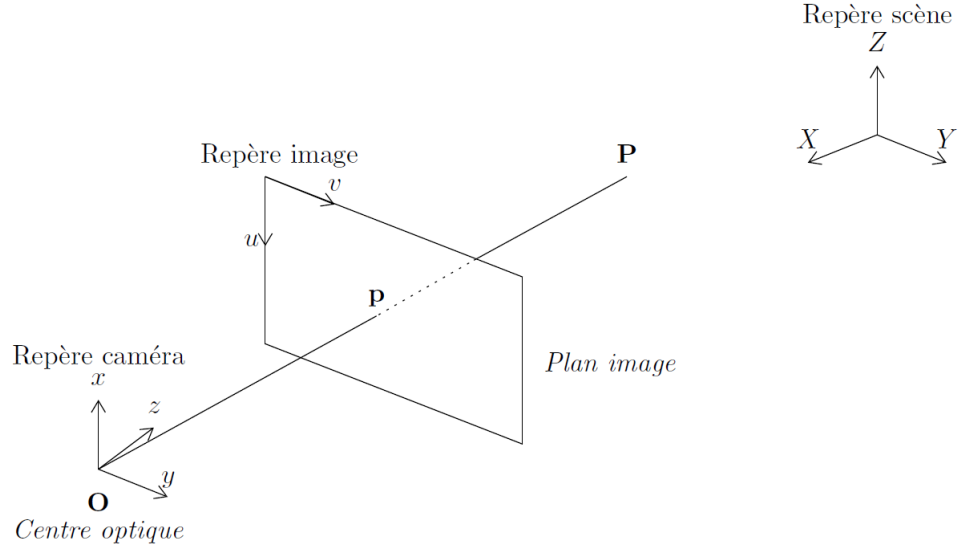


FIGURE 1.1 – Modèle géométrique de la caméra

le cerveau humain et exécutée inconsciemment s'est avérée compliquée pour les ordinateurs. Par conséquent, la simulation de ce comportement par une machine, en utilisant la vision stéréoscopique, implique l'extraction d'informations de profondeur d'une scène 3D en utilisant un système à deux caméras. Dans cette thèse, nous nous intéressons à la vision stéréoscopique binoculaire, qui fait référence à deux caméras (gauche et droite) placées horizontalement à une certaine distance afin de pouvoir observer la scène sous différents angles. Les caméras sont calibrées et les images capturées sont rectifiées. Considérons que la capture par le système stéréo est lancée simultanément, ce qui produit une paire d'images synchronisées, aussi appelée paire d'images stéréoscopiques. Les images stéréoscopiques sont exploitées pour déterminer la distance des objets par rapport à la caméra stéréoscopique et pour générer la carte de profondeur.

Nous exposons d'abord le modèle projectif de la caméra et la géométrie épi-polaire d'un système stéréoscopique. Nous abordons ensuite les primitives et les contraintes les plus couramment utilisées pour pallier les difficultés de la correspondance stéréoscopique. Enfin, nous terminons par un bref état de l'art des approches d'estimation de la disparité.

1.2 Les principes de la vision stéréoscopique

1.2.1 Modèle géométrique de la caméra

Le modèle du sténopé (voir la figure 1.1) est composé du plan image et du centre optique O . Un point P de la scène, dont les coordonnées sont (X, Y, Z) , est projeté sur le plan image en un point p , dont les coordonnées sont (u, v) , qui correspond à l'intersection de la droite OP avec le plan image.

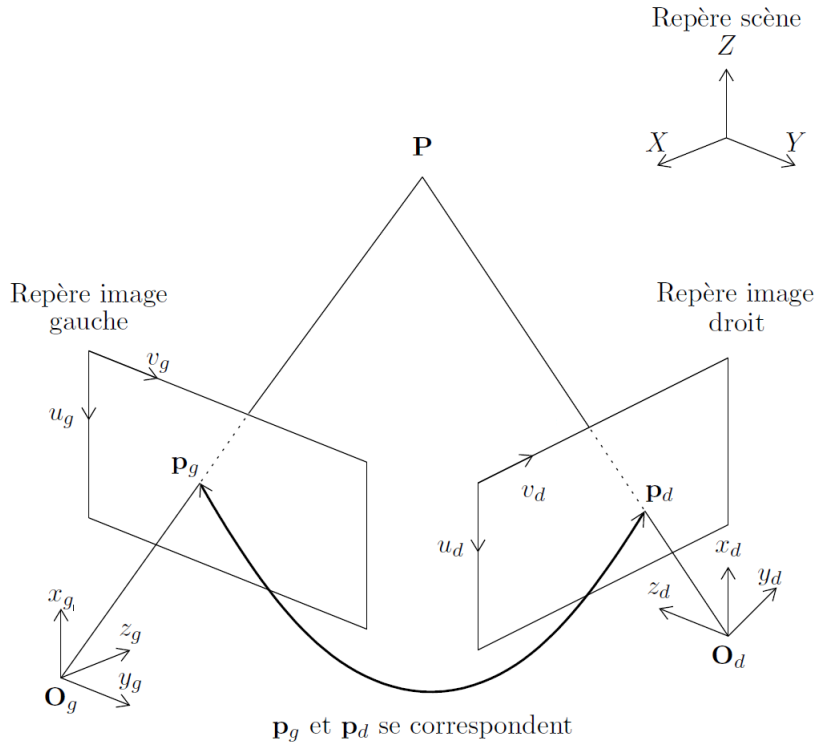


FIGURE 1.2 – Modèle géométrique du capteur stéréoscopique binoculaire.

Cette projection peut être décrite par la formule suivante :

$$\lambda (u \ v \ 1)^T = M (X \ Y \ Z \ 1)^T, \lambda \in \mathbb{R}^*, \quad (1.1)$$

où M est la matrice de projection perspective qui comporte les paramètres du modèle. Le calibrage d'une caméra consiste donc à calculer cette matrice.

1.2.2 Modèle du capteur stéréoscopique

Dans stéréo vision binoculaire, deux images sont obtenues à partir de deux points de vue différents. Pour cette fin, on fait appel à deux caméras. La position relative des deux prises de vue doit être déterminée de façon à rendre visible une grande partie de la scène sur les deux images. La position de la deuxième prise de vue ne peut être choisie par rotation autour du centre optique de la caméra, sinon la reconstruction 3D devient alors impossible. Les paramètres du modèle géométrique du capteur stéréoscopique binoculaire (voir figure 1.3) peuvent être représentés par les deux matrices de projection en perspective liées aux deux images.

1.2.3 La mise en correspondance stéréoscopique

La figure 1.3 nous indique qu'un point P de la scène est représenté dans l'image gauche par le point p_g dont les coordonnées sont (u_g, v_g) et dans l'image droite par le point p_d dont les coordonnées sont (u_d, v_d) . Dans ce cas, on considère que les points p_g et p_d sont homologues, autrement dit qu'ils correspondent, car ils représentent le même point de la scène.

La correspondance stéréoscopique est le processus qui permet de reconstituer une scène 3D à partir de deux ou plusieurs images de la même scène, capturées sous des angles différents. La qualité de la reconstruction 3D est directement liée à la qualité des résultats obtenus par les algorithmes d'appariement stéréoscopique. Cela explique l'importance de ce processus, qui s'avère être un problème difficile à résoudre, notamment en raison de la présence de changements de luminosité entre les deux images, d'occultations (zones contenant des pixels sans correspondance) et de surfaces non texturées.

Le fruit des algorithmes de correspondance stéréoscopique est une carte de disparité. La disparité est inversement proportionnelle à la profondeur. Cette dernière est la différence en coordonnées Z entre un point 3D dans l'espace autour du système stéréoscopique, mesurée en coordonnées du monde réel. Les deux caméras d'un système stéréo sont séparées par une distance de base b . Cette séparation entraîne un angle de vue légèrement différent pour chaque caméra. En conséquence, les objets capturés dans les images ont des emplacements différents dans les images stéréoscopiques.

1.2.4 Les difficultés de la mise en correspondance

La conception d'un algorithme de mise en correspondance stéréoscopique demande la connaissance de nombreux challenges auxquels les chercheurs sont confrontés. Parmi ces problèmes, on peut mentionner :

Distorsions radiométriques : Les différences ou distorsions radiométriques concernent les cas où des pixels homologues possèdent des valeurs d'intensité différentes. Le postulat selon lequel les pixels des deux images stéréoscopiques correspondant au même point de la scène présenteront la même intensité[56] est uniquement vrai pour des surfaces lambertiennes qui offrent une réflexion identique et multidirectionnelle de la lumière, quel que soit l'angle de vue. En pratique, les surfaces spéculaires sont toutefois très fréquentes. En outre, les distorsions radiométriques sont aussi engendrées par la présence de bruit ou simplement par les différences de sensibilité entre les caméras employées. Par ailleurs, l'une des principales limitations des algorithmes de mise en correspondance, utilisant l'hypothèse de conservation de la luminance, réside dans leur sensibilité aux distorsions radiométriques.

Ambiguïté : Les régions homogènes posent des conditions difficiles aux algorithmes de mise en correspondance. En effet, pour distinguer deux pixels homologues, il faut faire appel à une mesure qui permet de discriminer la paire correspondante par rapport aux

autres appariements potentiels. Hélas, il est difficile de découvrir une telle mesure car les intensités des pixels dans ces régions ne représentent pas des primitives discriminantes et par conséquent plusieurs correspondances potentielles avec la même valeur d'intensité sont présentes dans l'image cible. En outre, les régions présentant des textures répétitives posent également un problème parce que plusieurs points sont susceptibles de devenir des candidats potentiels à la correspondance, ce qui engendre une ambiguïté.

Occultations : Le phénomène d'occultation constitue un autre défi pour les algorithmes de mise en correspondance. Il se manifeste dans des situations où un pixel est visible dans une des images mais pas dans l'autre. Ces régions n'ont pas de correspondants dans l'image cible et génèrent des ambiguïtés de correspondance. Elles apparaissent fréquemment aux discontinuités de profondeurs.

1.2.4.1 Les primitives à mettre en correspondance

La mise en correspondance consiste à appairer des pixels homologues à partir des indices visuels qui les caractérisent. Ces indices visuels sont appelés *primitives* et peuvent être des pixels, des points d'intérêt, des points de contours, etc. La conception des algorithmes de mise en correspondance dépend essentiellement de la nature des primitives à appairer.

Les pixels de l'image Les pixels sont les primitives les plus simples à choisir, pour lesquelles les valeurs d'intensité en niveau de gris ou en couleur sont analysées. Cette primitive permet d'obtenir des cartes de disparités dites denses pour lesquelles une valeur de disparité est attribuée pour chaque pixel de l'image.

Les points d'intérêt Ces pixels sont définis par des propriétés qui les rendent plus discriminants par rapport à leurs voisins pour la mise en correspondance. Parmi Les détecteurs de points d'intérêt les plus populaires, on peut citer Harris [18] et Canny [7]. De nombreuses études de mise en correspondance sont basées sur des points d'intérêt [33, 67].

Les primitives structurées Les auteurs dans [29] ont proposé une étude sur des algorithmes de mise en correspondance par primitives structurées. Parmi ces primitives, on peut distinguer :

- Les segments : Ces attributs peuvent être de types géométriques ou photométriques : la position, l'orientation, la longueur, le niveau de gris, la surface, le contraste sur le segment et les relations avec les segments voisins. Néanmoins, l'appariement utilisant les segments comme primitives est très compliqué à résoudre puisque les formes et les tailles des segments peuvent différer d'une image à l'autre à cause du phénomène de distorsion projective. De nombreux travaux sur la mise en correspondance de segments ont été proposés dans la littérature, parmi lesquels on peut citer [25, 4, 64].

- Les contours : Les attributs utilisés généralement sont les niveaux de gris, les informations de couleur si existent ou les gradients des pixels qui forment le contour [53].
- Les régions : Les attributs utilisés sont les niveaux de gris ou les couleurs de la région [43], une mesure globale sur la région comme la variance des niveaux de gris proposée dans [17].

Les approches de mise en correspondance sont généralement réparties en deux grandes classes en fonction des primitives à mettre en correspondance. Les algorithmes de mise en correspondance basés sur les pixels, dites *pixel-based matching*, sont souvent très sensibles aux occlusions et aux changements radiométriques. Ils sont également souvent coûteux en termes de temps de calcul. Toutefois, l'avantage majeur de ces méthodes est qu'elles permettent d'obtenir directement un appariement dense. En contrepartie, les méthodes reposant sur des primitives éparées, telles que des points d'intérêt ou des régions, dites *feature-based matching*, requièrent la détection a priori de ces primitives qui doivent être marquées par des attributs à la fois invariants et discriminants, ce qui est souvent paradoxal. De plus, l'appariement de primitives ne permet pas d'obtenir directement une correspondance très dense.

Dans le contexte de notre application où obtenir une reconstruction dense de la scène est une exigence, les primitives que nous étudions sont les pixels de l'image dans la plupart de nos travaux. Nous tirons également profit de la puissance de l'appariement des points d'intérêt pour une reconstruction éparse robuste, puis nous appliquons une interpolation afin d'estimer les disparités manquantes pour une reconstruction dense.

1.2.4.2 Contraintes de la mise en correspondance

Afin de réduire le nombre de correspondants potentiels et de supprimer les faux appariements, un certain nombre de contraintes sont utilisées. Ces contraintes ont découlé d'hypothèses faites sur :

- *La géométrie du capteur stéréoscopique* : la contrainte la plus couramment utilisée dans vision stéréoscopique est la contrainte épipolaire.
- *La géométrie de la scène* : les contraintes visent plus précisément les objets constituant la scène. Il s'agit des contraintes d'unicité, d'ordre et de symétrie. Elles sont aussi appelées contraintes de compatibilité [31].
- *La réflectance de la surface des objets et le type et la position de la source lumineuse* : cela concerne la contrainte de dissimilarité. Si l'on admet que les sources lumineuses sont figées et que les objets sont idéalement lambertiens¹, un point sur un objet de la scène est toujours représenté avec la même intensité dans les deux images.

La contrainte épipolaire Cette contrainte est à la fois géométrique et unaire, peut servir à simplifier l'appariement en diminuant l'espace de recherche à une seule dimension

1. Les surfaces lambertiennes sont les surfaces qui reflètent la même lumière quel que soit l'angle de vue de l'observateur.

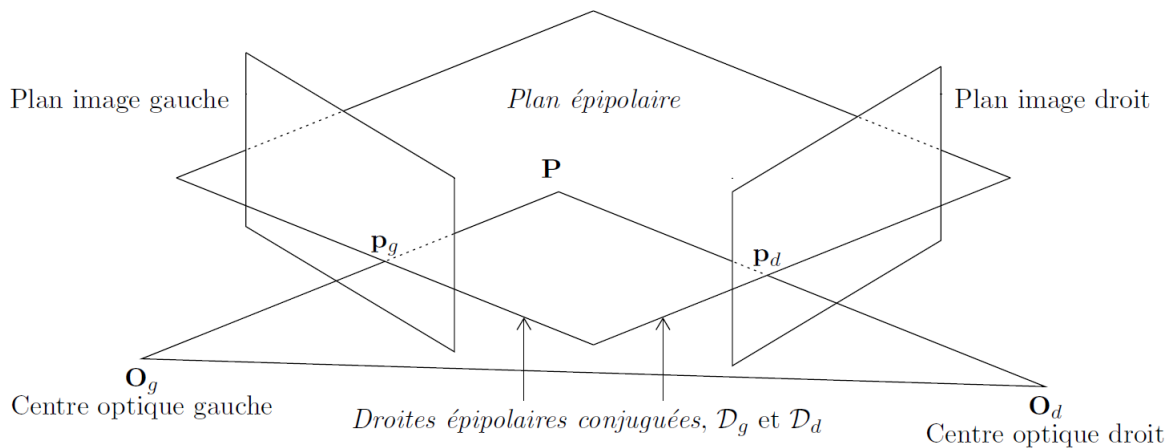


FIGURE 1.3 – Géométrie épipolaire.

en considérant que les points homologues sont toujours localisés sur des lignes épipolaires conjuguées. La figure 1.3 récapitule le principe de la géométrie épipolaire. Le point P de la scène et les centres de projection des deux caméras, O_g et O_d , forment le plan épipolaire. Les intersections de ce plan avec les deux plans images sont deux droites dites épipolaires conjuguées. Ainsi, la contrainte épipolaire se traduit par : le correspondant p_d d'un point p_g , qui se trouve sur la droite épipolaire gauche D_g , se trouve forcément sur la droite D_d , image de D_g dans le plan droit. La ligne D_d est donc dite ligne épipolaire droite liée au point p_g . Les droites épipolaires possèdent la caractéristique suivante : toutes les lignes épipolaires se croisent en un point appelé épipole. La contrainte épipolaire permet de définir entre les points de l'image gauche et les points de l'image droite une relation qui dépend uniquement des paramètres des caméras et des coordonnées de l'image et qui est indépendante des coordonnées du point dans la scène. La matrice représentant cette relation est nommée matrice fondamentale.

Rectification épipolaire Il existe une configuration géométrique particulière du capteur stéréoscopique, appelée configuration parallèle, selon laquelle : les lignes épipolaires sont toutes parallèles et horizontales dans les deux images. Deux pixels qui correspondent se situent sur la même ligne. Ainsi, le processus de mise en correspondance des images se voit considérablement simplifié du fait que le déplacement entre un pixel de l'image de gauche et son homologue de l'image de droite se résume à une simple translation.

Cependant, il est assez peu fréquent dans la pratique que les caméras soient correctement alignées selon le modèle projectif parallèle. Une transformation géométrique des images sera à réaliser pour que les lignes épipolaires soient parallèles et horizontales dans les deux images. Cette transformation, dite rectification épipolaire, consiste à reprojeter les images constituées sur les plans images de chaque caméra sur un nouveau plan de projection commun aux deux caméras. Il est indispensable, pour que les lignes épipolaires puissent être parallèles, que ce nouveau plan de projection soit parallèle à

l'axe reliant les deux centres de projection. Après rectification, on obtient une géométrie épipolaire simplifiée où le changement de référentiel entre les deux caméras se ramène à une simple translation.

Le but de la rectification consiste à faciliter grandement l'appariement des points entre les deux images en réduisant l'espace de recherche des correspondants à des paires de lignes horizontales. La difficulté vient du fait qu'il faut estimer les paramètres de la transformation à mettre en œuvre pour rectifier les images stéréoscopiques. De nombreuses approches ont été envisagées, en fonction de l'information disponible :

- Si les paramètres intrinsèques et extrinsèques du capteur sont disponibles, alors les matrices de projection en perspective liées aux deux images peuvent être calculées. Ceci permet de déterminer la transformation à effectuer pour obtenir deux plans d'image coplanaires [12] et parallèles à la ligne passant par les deux centres optiques.
- Dans le cas où les paramètres de calibration sont inconnus, la méthode de rectification est basée sur la connaissance de la matrice fondamentale obtenue par la mise en correspondance des points entre les deux images. La connaissance au préalable de cette matrice donne la possibilité, par la suite, de calculer les paramètres de rectification [19].

A partir de ce paragraphe, nous considérons toujours que nous traitons des images rectifiées. Tous les exemples que nous citons sont dans le cadre des lignes épipolaires conjuguées qui correspondent à la même ligne dans les deux images analysées.

La contrainte d'unicité La contrainte d'unicité signifie qu'une primitive dans une image ne peut correspondre qu'à une seule et unique primitive dans l'autre image. Cette règle prend pour hypothèse le fait qu'une primitive représente la projection d'un seul élément 3D. Dans le cas d'images rectifiées, cette contrainte s'exprime de la manière suivante :

$$\text{Si } x_1 - u(x_1, y) = x'_1, \quad \text{alors } \forall x_2 \neq x_1 \quad x_2 - u(x_2, y) = x'_1, \quad (1.2)$$

où (x_1, y) et (x_2, y) sont deux points de l'image de gauche.

La contrainte d'unicité est violée uniquement si le plan de la scène est très incliné par rapport à l'une des deux caméras. Ainsi, plusieurs points 3D peuvent être projetés sur la même primitive dans une image. La figure 1.5 illustre ce phénomène. En effet, les points M_1 et M_2 se projettent en un même point dans l'image de droite et en deux points différents dans l'image de gauche.

La contrainte d'ordre La contrainte d'ordre suppose la préservation de l'ordre des pixels homologues sur les deux lignes épipolaires. Elle est respectée si la distance entre deux pixels sur une ligne épipolaire est identique à celle de leurs homologues sur la ligne épipolaire conjuguée. Or, les objets transparents violent cette contrainte. Par conséquent, la majorité des algorithmes d'appariement considèrent que la scène étudiée est exempte d'objets transparents. De même, dans le cas où la distance entre les deux caméras est

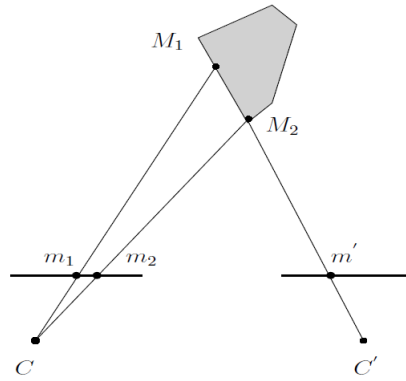


FIGURE 1.4 – Exemple de violation de la contrainte d'unicité : tous les pixels du segment de droite $[M_1, M_2]$ sont projetés sur le pixel m de l'image de droite.

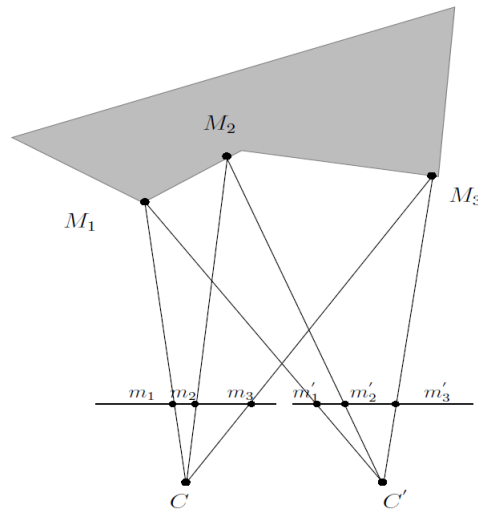


FIGURE 1.5 – La contrainte d'ordre.

importante et que la scène est composée de deux objets placés l'un derrière l'autre, l'ordre entre ces deux objets risque de ne plus être préservé. Pour les images rectifiées, la contrainte est représentée par :

$$\text{Si } x_1 - u(x_1, y) = x'_1 \text{ et } x_2 - u(x_2, y) = x'_2 \quad \text{alors } (x_1 - x_2)(x'_1 - x'_2) \geq 0, \quad (1.3)$$

où (x_1, y) et (x_2, y) sont deux points de l'image de gauche et (x'_1, y) et (x'_2, y) leurs homologues dans l'image de droite.

La contrainte de continuité de disparité La contrainte se base sur le constat que les surfaces des objets formant la scène étudiée sont de type continu et que la variation

locale de la profondeur est réduite. Elle implique donc que la disparité change de manière presque constante dans toute l'image. Cette contrainte de cohérence spatiale peut être respectée dans le cas de variations lentes de la profondeur, mais reste invérifiée en présence de discontinuités de profondeur qui constituent les frontières des objets.

La contrainte de symétrie La contrainte de symétrie est la conséquence directe de la contrainte d'unicité. Celle-ci exige qu'un point m de l'image de droite soit l'homologue d'un point m' de l'image de gauche, si et seulement si m est l'homologue de m' . Elle impose une correspondance à la fois de gauche à droite et de droite à gauche. Cette correspondance est nommée *correspondance bidirectionnelle* et se traduit par :

$$\text{Si } x - u(x_1, y) = x' \text{ alors } x' + u_d(x', y) = x, \quad (1.4)$$

Où (x, y) est un pixel de l'image de gauche, et u_d est obtenue par mise en correspondance de la droite vers la gauche.

1.3 Les algorithmes de mise en correspondance stéréoscopique

La littérature contient un grand nombre d'algorithmes de mise en correspondance. En effet, ce domaine connaît d'ailleurs une augmentation du nombre de publications qui remonte aux années 80. Une étude approfondie des techniques disponibles est menée par Scharstein et Szeliski dans [56]. Ils définissent quatre étapes indispensables pour décrire les algorithmes de stéréovision :

1. le coût de correspondance
2. la zone d'agrégation
3. la méthode d'optimisation/minimisation
4. l'affinement des disparités

Dans le cadre de notre étude, nous tenons compte de la méthode d'optimisation des algorithmes de mise en correspondance, on peut les diviser en quatre catégories : les approches locales, les approches progressives, les approches coopératives et les approches globales.

Les approches locales se caractérisent par la mise en correspondance des éléments les plus similaires en utilisant une mesure de similarité. Ces méthodes consistent généralement à déterminer une fenêtre appropriée qui contient l'élément à mettre en correspondance et à définir un facteur de similarité. Ces méthodes offrent l'avantage de pouvoir être appliquées de manière simple et efficace avec un temps de calcul relativement faible. Toutefois, cette catégorie d'algorithmes reste très vulnérable aux problèmes engendrés par les occultations, la faible texture ou les différences d'illumination entre les différentes vues. Par ailleurs, les approches globales semblent pallier ces défauts en optimisant une fonction de coût globale qui inclut la globalité des pixels des images, mais cela entraîne une complexité de calcul plus importante. Les méthodes progressives et coopératives consistent en une combinaison d'une méthode locale et d'une méthode globale, pour tirer avantage de leurs qualités mutuelles.

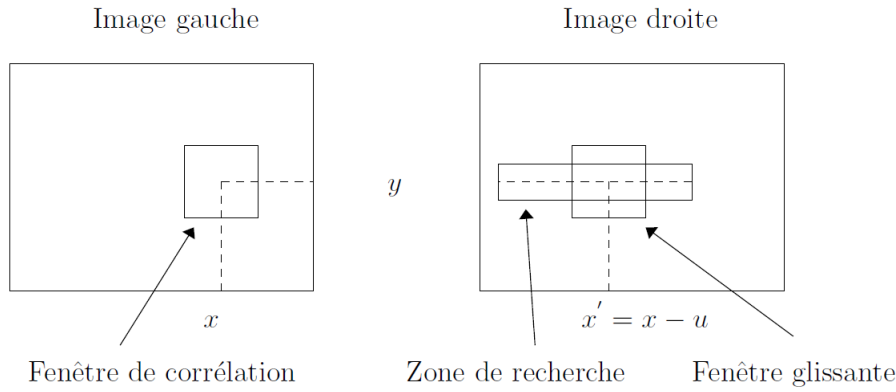


FIGURE 1.6 – Mise en correspondance par corrélation.

1.3.1 Méthodes locales

Les méthodes locales se basent sur la similarité entre deux ensembles de pixels et sont aussi appelées méthodes de corrélation. Elles ont été largement employées en vision par ordinateur pour diverses applications comme l'estimation de mouvement, l'indexation vidéo, la détection des obstacles, etc.

Le principe de ces méthodes consiste à effectuer une comparaison entre le voisinage du point à mettre en correspondance et le voisinage des correspondants éventuels, situés sur la ligne épipolaire correspondante dans l'autre image. Si l'on cherche le correspondant d'un point (x, y) dans l'image de gauche, la première phase consiste à cerner une région de recherche et une fenêtre glissante qui balayera cette région dans l'image de droite (voir figure 1.6). Un score de corrélation est alors calculé, au point (x, y) et pour une disparité d , en comparant les intensités des pixels d'une fenêtre centrée sur le point $p = (x, y)$ dans l'image de gauche et d'une fenêtre centrée sur le point $q = (x - u, y)$ dans l'image de droite. L'ensemble des scores de correspondance calculés sur la région de recherche constitue l'image de l'espace de disparité (DSI). Dans la plupart du temps, le point sélectionné est obtenu en utilisant la stratégie winner take all, notée (WTA), c'est-à-dire le point qui obtient le meilleur score de corrélation dans cet espace.

Cette stratégie est décrite comme suit :

$$d_p = \min_{d_{min} \leq d \leq d_{max}} \sum_{q \in N_p} c(q, q - d) \quad (1.5)$$

Où d_p est la disparité finale attribuée au pixel p , d_{min} et d_{max} représentent respectivement la disparité minimale et maximale possible. N_p représente le voisinage du pixel p qui est considéré comme zone d'agrégation et $c(q, q - d)$ représente un coût entre le pixel q dans l'image de gauche et le pixel correspondant à la disparité d dans l'image de droite.

Les approches locales se caractérisent notamment par la zone d'agrégation employée. Dans la suite, les différents types de zones d'agrégation utilisés sont mentionnés, ainsi

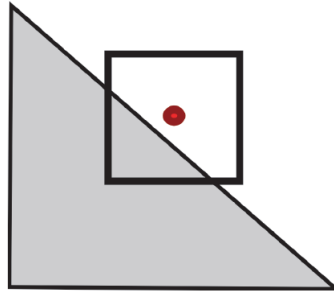


FIGURE 1.7 – Exemple d’un pixel représenté en rouge et sa fenêtre d’agrégation associée ayant des valeurs de disparité différentes.

que les avantages et inconvénients de chacun d’entre eux.

Zone d’agrégation par corrélation Dans l’approche d’agrégation par fenêtre, la zone de voisinage est généralement représentée par une fenêtre carrée de taille prédéfinie. Le principal avantage de cette approche est sa rapidité de calcul. Cependant, elle présente de nombreux inconvénients. En effet, le premier problème apparaît dans les régions de discontinuité de disparité, comme illustré dans la figure 1.7. Dans ces régions, l’hypothèse qui consiste à considérer que tous les pixels de la zone d’agrégation définie possèdent les mêmes disparités échoue et conduit à des erreurs dans la carte de disparité.

Un autre problème concerne le choix d’une bonne taille de fenêtre. En effet, une grande fenêtre augmente le temps de calcul, mais permet de capturer plus de texture. Par contre, une petite taille de fenêtre fournit un temps d’exécution rapide, mais réduit les possibilités de saisir des éléments discriminants.

Comme l’indique Nishihara dans [52], la qualité de la correspondance dépend de la cohérence entre la taille du voisinage et l’information présente dans cette région de l’image. Certaines méthodes proposent de définir, dans chaque pixel des images gauche et droite, la taille et la géométrie optimales de la zone d’agrégation, d’autres méthodes suggèrent des supports adaptatifs pour enlever les ambiguïtés qui peuvent exister en utilisant une fenêtre de taille fixe. En voici quelques-unes, certaines de ces méthodes sont présentées dans les paragraphes suivants.

Zone d’agrégation par fenêtres adaptatives Certains auteurs estiment des scores de similarité pour plusieurs fenêtres où le pixel à correspondre est situé à différentes positions. *Fusiello et al* ont proposé une approche permettant d’améliorer la corrélation classique basée sur les fenêtres en utilisant neuf fenêtres différentes, appelée “Symmetric Multi-Window” (SMW) dans[11]. Le pixel, dont la disparité est estimée, ne se situe plus au centre de la fenêtre d’agrégation, mais occupe des positions différentes. Le but est de repérer la meilleure fenêtre possible qui n’enfreindra pas une discontinuité de disparité. Par conséquent, plus le score est faible, plus la fenêtre trouvée a de chances de couvrir

une région de profondeur constante. La disparité ayant le score minimal par fenêtre est retenue.

Hirschmüller et al [20] proposent une autre approche qui consiste à diviser une fenêtre d'agrégation centrée en neuf parties. Leur hypothèse consiste à considérer que les parties d'une fenêtre d'agrégation ne possèdent pas toutes la même valeur. Par conséquent, le score de correspondance est calculé en gardant seulement les cinq meilleurs coûts des sous-fenêtres.

L'inconvénient de ces approches est toujours le choix de la bonne taille de fenêtre. De plus, il est souvent impossible de trouver une fenêtre qui maintienne la discontinuité de la disparité ou de déterminer les cinq sous-fenêtres pertinentes dans une fenêtre d'agrégation.

Zone d'agrégation par pondération adaptative En prolongeant le concept proposé par *Hirschmüller et al* [20] indiquant que toutes les sous-parties d'une fenêtre d'agrégation ne doivent pas être prises en compte dans le calcul du score final, des techniques associant un poids à chaque sous-partie ont été développées. Par conséquent, si dans l'agrégation classique basée sur la fenêtre, tous les pixels ont la même influence sur le coût de la correspondance (équation 1.6), dans l'approche de poids de support adaptatif, un poids $w(p, q)$ est utilisé pour déterminer la probabilité que deux pixels, p et q , possèdent la même disparité (équation 1.7).

$$C_{(p,q)} = \sum_{q \in N_p} c(q, q - d) \quad (1.6)$$

$$C_{(p,q)} = \sum_{q \in N_p} w(p, q) * c(q, q - d) \quad (1.7)$$

Yoon et Kweon [72] ont proposé une fonction pour le calcul du poids qui prend en compte à la fois la similarité des couleurs et la distance spatiale entre deux pixels (équation 1.8). Cette méthode présente l'avantage de produire de bons résultats dans les régions de discontinuité de disparité. Cependant, son coût de calcul est élevé.

$$w(p, q) = \exp\left(-\left(\frac{\delta c_{pq}}{\gamma c} + \frac{\delta g_{pq}}{\gamma g}\right)\right) \quad (1.8)$$

où δc_{pq} calcule la dissimilarité des couleurs, δg_{pq} calcule la distance euclidienne et $(\gamma c, \gamma g)$ sont des paramètres prédéfinis.

La distance euclidienne entre deux pixels ne force pas deux pixels à se retrouver sur la même surface, une solution a été proposée dans [26] en considérant les distances géodésiques. Une distance géodésique représente le chemin le plus court qui relie deux pixels (équation 1.9)

$$w(p, q) = \exp\left(-\frac{D(p, q)}{\gamma}\right) \quad (1.9)$$

où $D(p, q)$ désigne la distance géodésique et γ est un paramètre prédéfini.

Ces méthodes apportent certes des améliorations dans les résultats obtenus. En revanche, elles présentent un inconvénient majeur, à savoir un temps de calcul considérable.

Zone d'agrégation croisée *Zhang et al* [76] ont proposé une technique performante qui repose sur l'agrégation de zones croisées afin de calculer une région d'agrégation de pixels, en tenant compte à la fois de la couleur et des distances euclidiennes.

Le principe consiste à créer une région transversale pour chaque pixel. Pour cela, il suffit de trouver quatre pixels, correspondant à l'extrémité des quatre axes : haut, bas, gauche et droite. Puis, afin de construire une région de différentes formes, pour chaque pixel qui se trouve sur le bras vertical, le bras horizontal fournit les limites de la région pour la ligne spécifique.

Afin de choisir un point d'extrémité du bras pour un pixel p donné, deux règles sont appliquées qui imposent des limites à la similarité des couleurs et à la longueur maximale de l'axe.

1.3.2 Méthodes globales

Une méthode de mise en correspondance est qualifiée comme étant globale si sa fonction de coût est estimée sur tous les pixels de l'image. Ce coût global présente la forme générale suivante :

$$E(D) = E_{data}(D) + E_{smooth}(D) \quad (1.10)$$

où $E_{data}(D)$ représente le terme de liaison de données, également appelé terme d'observation, $E_{smooth}(D)$ est un terme de régularisation.

La résolution du problème d'appariement consiste donc à déterminer la fonction de disparité, u , qui minimise ce coût global. De nombreuses techniques d'optimisation ont été décrites dans la littérature. On peut différencier les méthodes à variables discrètes, qui procèdent à un appariement par paire de pixels, comme la programmation dynamique ou les coupures de graphes, aux méthodes à variables continues, dans lesquelles la disparité calculée est réelle, comme dans les approches variationnelles.

1.3.2.1 Programmation dynamique

La méthode d'optimisation la plus utilisée pour l'appariement est la programmation dynamique [61, 77, 28, 3, 66]. Elle est appliquée entre deux séquences de pixels sur les lignes épipolaires conjuguées en tenant compte de l'ordre des pixels dans chaque séquence. Les éléments des séquences à appairer forment les deux dimensions d'une matrice qui comprend les coûts des appariements éventuels entre les pixels. L'objectif consiste à déterminer le chemin de coût optimal dans cette matrice en fonction d'une mesure de similarité (SAD , SSD , etc.) entre deux éléments. Le coût d'un chemin est la somme des coûts de ses éléments. La figure 2.13 présente la matrice de recherche, dont les axes sont constitués par les lignes de recherche des images de gauche et de droite.

Cette approche offre la possibilité de mettre en place un lissage le long d'une ligne

épipolaire donnée. En revanche, son principal inconvénient réside dans l'absence de lissage entre les lignes épipolaires voisines. Plusieurs variantes de cette méthode ont été proposées afin de considérer les lignes épipolaires adjacentes qui se trouvent au-dessous de la ligne épipolaire en question [34, 8, 40]

1.3.2.2 Coupures de graphes

Le recours aux coupes de graphes en vision par ordinateur a été initié par *Roy et Cox* [55] pour résoudre le problème de l'incohérence des lignes inter-épipolaires en programmation dynamique. Par la suite, cette technique a été formalisée par *Veksler* [65] à partir de laquelle de nombreux travaux fondés sur les techniques de coupures de graphes ont été menés [2, 32, 24]. Ces derniers définissent le problème de mise en correspondance sous la forme d'un problème consistant à trouver le flux maximal dans un graphe en minimisant une fonction de coût liée à ce graphe. Ce flux maximal est obtenu pour la coupure minimale du graphe.

Le problème de flot dans un graphe peut être considéré comme un problème de flux d'eau dans un réseau de tuyaux, autrement dit, sachant qu'il existe une source d'eau s et un puits p , on cherche le flux maximal capable de traverser le réseau de tuyaux reliant la source au puits. Afin de symboliser ce réseau, les méthodes de découpage de graphes emploient un graphe, désigné par G , constitué par un ensemble de nœuds N , qui représentent les connexions, et un ensemble d'arcs orientés A qui correspondent aux tuyaux. En utilisant ce genre de méthode pour effectuer une correspondance de pixels, la coupure de graphe est réalisée à la base d'un coût de correspondance global. Les nœuds du graphe représentent les pixels et le coût attribué à chaque arc est défini par les termes de la fonction de coût 1.10. Ainsi, on peut distinguer deux types d'arcs : les arcs reliant un pixel (x, y) de l'image gauche et un pixel $(x - u, y)$ de l'image droite pour une disparité u avec l'attribut du premier terme de la fonction de coût et les arcs connectant les pixels voisins avec l'attribut du second terme de la fonction d'énergie. La source et le puits sont respectivement les pixels qui sont les projections des points les plus proches et les plus éloignés du plan de l'image. La capacité d'un arc traduit le coût d'appariement lié aux deux nœuds connectés et le coût d'une coupe correspond à la somme des capacités de ses arcs.

L'avantage de cette méthode réside dans le fait que son utilisation sur l'ensemble de l'image est réalisable avec des temps de calcul raisonnables, contrairement à la programmation dynamique qui est très coûteuse. En fait, cette technique a été utilisée assez fréquemment, comme la programmation dynamique, sur des lignes épipolaires conjuguées [28, 2].

1.3.2.3 Propagation de croyance

Dans [62], les auteurs ont prouvé que la propagation de croyance permet de trouver une solution approximative pour le problème de minimisation d'énergie de la mise en correspondance stéréoscopique. La propagation de croyance est un algorithme itératif de minimisation de l'énergie qui agit en passant des messages à travers des pixels voisins connectés de manière directionnelle. Par conséquent, le terme de coût des données de la

fonction d'énergie est combiné à quatre ensembles de valeurs de message correspondant à chaque disparité possible pour chaque pixel. À chaque itération, une valeur de message actualisée est envoyée à chacun des quatre pixels voisins. Une fois les itérations effectuées, la valeur de la disparité est estimée pour chaque pixel. Étant donné que les coûts des données et les valeurs des messages doivent être stockés pour chaque disparité possible à chaque pixel, les exigences de mémoire sont considérables. En outre, étant donné qu'il s'agit d'un processus itératif, il peut être assez lent. Par conséquent, diverses méthodes ont été développées afin d'accélérer cet algorithme.

Une des variantes est la propagation hiérarchique des croyances [9]. Un schéma pyramidal est construit dans lequel la largeur et la hauteur sont divisées par deux à chaque niveau de la pyramide. Les valeurs des messages des niveaux inférieurs sont initialisées par les niveaux supérieurs de la pyramide. D'autres accélérations de l'algorithme sont proposées par l'implémentation sur *GPU* [5, 16]. D'autres accélérations et réductions de l'espace de recherche ont été proposées par [15].

1.4 Les bases de données de la stéréo vision

Plusieurs bases de données sont disponibles pour évaluer les algorithmes d'appariement stéréoscopique. Il s'agit notamment de bases de données de scènes routières réelles avec des vérités de terrain, comme KITTI 2012 [13] et KITTI 2015 [46]. En outre, il existe des bases de données bien connues conçues pour évaluer les algorithmes de mise en correspondance stéréoscopique, comme la base de données Middlebury [56] et la nouvelle base de données Middlebury v3 [57].

La base de données Middlebury [56] est l'un des ensembles de données les plus utilisés pour comparer les algorithmes de mise en correspondance stéréoscopique. Bien que cette base de données offre de nombreux défis en termes de situations rencontrées, les images sont capturées à l'intérieur d'un laboratoire et dans des conditions contrôlées. La base de données Middlebury 2014 est divisée en deux parties, les ensembles d'entraînement et de test qui contiennent chacun 15 couples stéréoscopiques. L'ensemble d'entraînement est disponible en trois résolutions, pour lesquelles des cartes de disparité de vérité-terrain sont fournies. L'ensemble de test utilise une plateforme d'évaluation pour enregistrer les résultats. Dans nos expériences, nous avons utilisé 15 images à quart de résolution de l'ensemble d'entraînement. Les deux bases de données KITTI fournissent des images de routes réelles avec des vérités de base pour environ 50 de pixels, ce qui constitue une excellente source de données pour évaluer les différents algorithmes de correspondance stéréoscopique. La base de données KITTI 2012 contient 389 paires d'images stéréo divisées en 194 images pour l'entraînement et 195 pour le test. Les auteurs fournissent la vérité-terrain uniquement pour les séquences d'entraînement, tandis que pour les séquences de test, un serveur d'évaluation doit être utilisé pour obtenir les résultats. La carte de disparité de vérité terrain a été obtenue à l'aide d'un scanner laser Velodyne, la vérité terrain n'est donc disponible que pour environ 50% des pixels de l'image. Selon le même principe, le jeu de données KITTI 2015 contient 200 paires stéréo d'entraînement et 200 paires de test. L'évaluation pour les ensembles de données KITTI 2012 est me-

surée en calculant le pourcentage d’erreurs de disparité par rapport à la vérité de base. Alors que pour KITTI 2015, la mesure d’erreur $D1 - all$ est calculée, elle représente le pourcentage de pixels pour lesquels l’erreur d’estimation est supérieure à trois pixels et supérieure à 5% de la disparité de la vérité au sol à chaque pixel.

Les principaux défis des ensembles de données KITTI sont les distorsions radiométriques causées par les éruptions solaires, les réflexions et les images bruitées (causées par de grandes différences d’intensité entre la lumière et l’ombre). Pour nos expériences, nous avons choisi de travailler avec les quatre jeux de données présentés : les deux jeux de données Middlebury, en raison du nombre considérable d’algorithmes de correspondance stéréoscopique qui ont été comparés sur ces images, et les deux bases de données KITTI, compte tenu de notre champ d’application.

1.5 Fonctions de coûts

Les fonctions de coût permettent de mesurer la qualité d’une correspondance. On distingue trois types de coûts de correspondance : les coûts paramétriques, non paramétriques et les coûts basés sur l’information mutuelle [23].

Dans notre étude, nous examinerons ces catégories en fonction de leur résistance aux distorsions radiométriques. Les pixels radiométriquement similaires font référence aux pixels situés sur deux images différentes, mais qui correspondent au même point 3D. Par conséquent, ils devraient avoir des valeurs d’intensité similaires ou, dans un cas plus idéal, identiques dans les deux images [21]. Le terme de distorsions ou de différences radiométriques s’applique quand des pixels homologues ont des valeurs d’intensité différentes. Ces différences peuvent provenir de plusieurs facteurs tels que les paramètres de la caméra, les surfaces non lumineuses ou le décalage temporel de la prise de vue (comme c’est le cas pour l’imagerie satellitaire).

Les coûts paramétriques sont basés sur la valeur de l’intensité du pixel. Le principal inconvénient des coûts consiste dans leur vulnérabilité aux changements radiométriques. Les coûts non-paramétriques prennent en compte seulement un ordre local des intensités, ce qui permet de déduire que ces derniers sont plus résistants aux distorsions radiométriques. Les coûts de l’information mutuelle (IM) sont calculés à partir d’une carte de disparité initiale. L’information mutuelle traite efficacement les changements radiométriques [10], mais elle ne peut traiter que les distorsions radiométriques qui se produisent globalement, sans tenir compte des changements radiométriques locaux (qui sont plus courants dans la pratique).

La qualité de la carte de disparité dépend principalement de la fonction de coût utilisée. La comparaison des fonctions de coût a été effectuée dans plusieurs travaux de recherche, dont les plus célèbres sont [21] et [23]. Contrairement à l’étude de 2007, dans laquelle six fonctions de coût ont été évaluées, auteurs dans [23] ont pu examiner quinze coûts de correspondance stéréoscopique différents pour des images présentant des différences radiométriques. La comparaison de ces coûts a été effectuée à l’aide de trois algorithmes de correspondance stéréoscopique différents : un algorithme basé sur l’optimisation globale de l’énergie (Graph Cuts), un second utilisant la correspondance

semi-globale [22] et un troisième basé sur une fenêtre d'agrégation locale. Les auteurs ont conclu que le coût basé sur la transformée de Census offre les meilleures performances sur tous les algorithmes. Dans l'étude [21], les auteurs ont évalué la performance des coûts retenus sur des changements radiométriques simulés et réels dans un environnement de laboratoire (ensemble de données Middlebury²). Dans la présente étude, on a opté pour des expériences sur des bases de données plus compétitives et bien connues, telles que les images de routes réelles issues des bases de données KITTI 2012 et KITTI 2015, ainsi que celles extraites de la nouvelle base de données Middlebury v3³. En outre, nous proposons trois nouvelles fonctions de coût, deux non paramétriques basées sur la transformée de Census et une basée sur l'information structurelle. Nous montrons que nos fonctions de coût proposées offrent une bonne résistance aux changements radiométriques.

Dans ce qui suit, nous décrivons brièvement les fonctions de coût courantes. Nous les répartissons en trois catégories, à savoir les coûts paramétriques, les coûts non paramétriques et les coûts mixtes. Nous qualifions de coûts paramétriques mixtes les coûts qui visent à améliorer le pouvoir discriminant d'un coût non paramétrique en incorporant des informations supplémentaires généralement fournies par un coût paramétrique.

1.5.0.1 Les coûts paramétriques

Parmi les fonctions de correspondance des coûts les plus courantes, on peut citer les différences d'intensité quadratiques (SD) (voir l'équation 1.11), comme celles utilisées par *Kolmogorov et Zabih* [37], ou les différences d'intensité absolues (AD) (voir l'équation 1.12), qui sont généralement combinées à d'autres informations, comme celles utilisées par *Mei et al* [44] et *Klaus et al* [50]. Les coûts SD et AD sont basés sur l'hypothèse de la similarité des valeurs de couleur des pixels et sont donc très sensibles aux distorsions radiométriques.

Soit p un pixel de l'image de gauche de coordonnées (x, y) et d la valeur de disparité pour laquelle on souhaite déterminer le coût de p . Aussi $I_g(x, y)_i$ est la valeur d'intensité du pixel p dans l'image de gauche sur le canal de couleur i , tandis que $I_d(x, y - d)_i$ est la valeur d'intensité du pixel donné par les coordonnées $(x, y - d)$ dans l'image de droite. Nous considérons que n est le nombre de canaux de couleur utilisés ($n = 1$ pour les images en échelle de gris et $n = 3$ pour les images en couleur).

$$C_{SD}(x, y, d) = \frac{1}{n} \sum_{i=1, n} (I_g(x, y)_i - I_d(x, y - d)_i)^2. \quad (1.11)$$

$$C_{AD}(x, y, d) = \frac{1}{n} \sum_{i=1, n} |I_g(x, y)_i - I_d(x, y - d)_i|. \quad (1.12)$$

Si on considère que $N(x, y)$ est le voisinage du pixel de coordonnées (x, y) , alors le coût

2. <http://vision.middlebury.edu/stereo/>

3. <http://vision.middlebury.edu/stereo/eval3/>

AD sur ce voisinage est par l'équation 1.13

$$C_{SAD}(x, y, d) = \sum_{(a,b) \in N(x,y)} C_{AD}(a, b, d). \quad (1.13)$$

Il existe également d'autres coûts paramétriques calculés à l'intérieur d'une fenêtre de support, notamment la somme moyenne zéro des différences absolues ($ZSAD$), la corrélation croisée normalisée (NCC) et la somme moyenne zéro de la corrélation croisée normalisée ($ZNCC$). Le coût $ZSAD$ soustrait l'intensité moyenne d'une fenêtre de support de chaque intensité à l'intérieur de cette fenêtre avant de calculer la somme des différences absolues. Le NCC est un coût paramétrique qui peut compenser de changement de gain, tandis que le $ZNCC$ (voir l'équation 1.14) est une variante qui compense à la fois le gain et le décalage dans la fenêtre de corrélation [21]. Étant donné que $ZNCC$ est une fonction de corrélation dont ses valeurs sont comprises entre $]0, 1]$, pour obtenir le coût, nous la soustrairons d'une valeur de 1.

$$C_{ZNCC}(x, y, d) = 1 - ZNCC(x, y, d). \quad (1.14)$$

$$ZNCC(x, y, d) = \frac{\sum_{(a,b) \in N_{x,y}} ZV(I_g, a, b) ZV(I_d, a, b - d)}{\sqrt{\sum_{(a,b) \in N_{x,y}} (ZV(I_g, a, b))^2 (ZV(I_d, a, b - d))^2}}. \quad (1.15)$$

$$ZV(I, x, y) = 1 - I(x, y) - \overline{I_{N_{x,y}}}(x, y). \quad (1.16)$$

où $\overline{I_{N_{x,y}}}$ est la valeur moyenne calculée dans le voisinage $N(x, y)$.

1.5.0.2 Les coûts non paramétriques

Cette section présente les fonctions de coûts non paramétriques les plus courantes dans la littérature. Dans [73], les auteurs ont proposé une méthode de mise en correspondance basée sur une transformée non paramétrique de l'image afin de diminuer la vulnérabilité aux imperfections liées à la luminosité. Comme défini par Zabih et Woodfill [73] pour calculer la transformée de Census (CT) d'un pixel p , une fenêtre appelée voisinage de support ($n \times m$), doit être centrée sur chaque pixel. Une chaîne de bits est calculée en transformant les valeurs d'intensité à l'intérieur de la fenêtre en une valeur de 1, si le pixel correspondant a une valeur d'intensité supérieure à la valeur d'intensité du pixel central, ou en une valeur de 0 autrement. La relation d'intensité locale est donnée par l'équation 1.18, où p et q sont des pixels de l'image. La transformée de Census est donnée par l'équation 1.18, où $\otimes$ désigne une concaténation au sens des bits et $n \times m$ est la taille de la fenêtre de voisinage. Le coût de la CT est donné par la distance de Hamming (DH) entre les deux chaînes de bits (équation 1.17).

$$C_{CT}(x, y, d) = D_H(CT(x, y), CT(x, y - d)). \quad (1.17)$$

12	14	26	32	31	0	0	1	1	1	001110011100110001100111
13	16	24	30	33	0	0	1	1	1	
11	15	23	31	33	0	0	×	1	1	
10	14	22	29	32	0	0	0	1	1	
12	17	25	34	36	0	0	1	1	1	

FIGURE 1.8 – Illustration de la Transformée Census

12	14	26	32	31	...	...	...	...	...
13	16	24	30	33	...	...	...	...	...
11	15	23	31	33	...	...	11	...	...
10	14	22	29	32	...	...	...	...	...
12	17	25	34	36	...	...	...	...	...

FIGURE 1.9 – Illustration de la Transformée Rank

où CT est la chaîne de bits construite comme dans suit.

$$CT_I(u, v) = \bigotimes_{i=1, n} \bigotimes_{j=1, m} (\xi(I(u, v), I(u + i, v + j))). \quad (1.18)$$

où $n \times m$ désigne la fenêtre de support, $\otimes$ désigne l'opération de concaténation, et la fonction ξ est définie comme suit :

$$\xi(p, q) = \begin{cases} 1 & p \leq q \\ 0 & \text{otherwise} \end{cases} \quad (1.19)$$

La figure 1.8 illustre un exemple de calcul de la transformée de Census pour le pixel central, à gauche les valeurs d'intensité des pixels d'une fenêtre de taille (5×5) . Chaque pixel dont la valeur d'intensité est supérieure à celui du pixel central est codé avec la valeur 1, sinon on met la valeur à 0. Une chaîne de binaire (à droite) est ainsi déterminée pour caractériser les pixels voisins du pixel à apparier.

Les auteurs [73] ont également proposé une autre transformée nommée Rank peu utilisée dans la littérature pour sa sensibilité aux variations radiométriques. Cette dernière permet de déterminer le nombre de pixels ayant une valeur d'intensité inférieure à celle du pixel à apparier et de l'attribuer à ce dernier. La figure 1.9 présente à gauche les valeurs d'intensité des pixels d'une fenêtre de taille (5×5) et à droite la transformée de rang correspondant au pixel central. Dans cet exemple, 11 pixels de la fenêtre ont une valeur d'intensité plus faible que le pixel central. Ensuite, une mesure de similarité de type SSD (Sum of Squared Differences) est utilisée pour établir les correspondances. Les coûts basés sur le gradient ou les mesures non-paramétriques sont plus résistants aux

changements de gain et de biais de la caméra ou aux surfaces non-lambertiennes tout en restant moins performants [36].

1.5.1 Les coûts mixtes

Les coûts non paramétriques offrent une bonne résistance aux distorsions radiométriques, mais leur performance reste faible. Pour cette raison, plusieurs études récentes proposent des combinaisons de coûts paramétriques et non paramétriques afin d'améliorer la discrimination des coûts utilisés. Dans la suite, nous exposerons les coûts mixtes les plus couramment utilisés dans la littérature.

Dans [36], les auteurs proposent un coût nommé C_{klaus} (voir équation 1.20) qui combine à la fois le coût C_{SAD} (Sum of Absolute Differences) et une mesure basée sur le gradient C_{GRAD} (voir équation 1.21). Cette approche est adoptée par l'un des algorithmes les mieux classés dans l'évaluation de la base de données Middlebury. Ces deux coûts sont calculés dans un voisinage $N(x, y)$ de (3×3) pixels et sont pondérés par w .

$$C_{klaus}(x, y, d) = (1 - w) * C_{SAD}(x, y, d) + w * C_{GRAD}(x, y, d). \quad (1.20)$$

Où

$$\begin{aligned} C_{GRAD}(x, y, d) = & \sum_{(a,b) \in N(x,y)} \left| \Delta_x I_g(a, b) - \Delta_x I_d(a, b - d) \right| \\ & + \sum_{(a,b) \in N(x,y)} \left| \Delta_y I_g(a, b) - \Delta_y I_d(a, b - d) \right| \end{aligned} \quad (1.21)$$

où $|\Delta_x$ et $|\Delta_y$ sont respectivement les gradients horizontal et vertical de l'image.

Le coût basé sur la transformée de Census C_{CT} et la différence absolue de C_{AD} a remporté un grand succès grâce à ses résultats satisfaisants obtenus sur l'ensemble de données Middlebury [44]. Ce coût est définie comme suit :

$$\begin{aligned} C_{ADCT}(x, y, d) = & \rho(C_{CT}(x, y, d), \lambda_{CT}) \\ & + (C_{AD}(x, y, d), \lambda_{AD}) \end{aligned} \quad (1.22)$$

où λ_{CT} et λ_{AD} contrôlent le poids de chaque coût, et ρ est défini dans l'équation 3.13.

$$\rho(c, \lambda) = 1 - \exp\left(-\frac{c}{\lambda}\right) \quad (1.23)$$

Les auteurs dans [59] proposent un autre coût basé sur une combinaison d'un C_{CT} et d'un C_{AD} (voir équation 1.24), où les deux sont calculés sur les images de gradient.

$$\begin{aligned} C_{cstent}(x, y, d) = & \rho(C_{\Delta CT}(x, y, d), \lambda_{CT}) \\ & + \rho(C_{AD}(x, y, d), \lambda_{AD}) \\ & + \rho(C_{\Delta AD}(x, y, d), \lambda_{\Delta CT}) \end{aligned} \quad (1.24)$$

où ΔCT et ΔAD sont les coûts, respectivement CT et AD , calculés sur les images de gradient.

1.6 Conclusion

Dans ce chapitre, on a présenté des concepts généraux, couramment exploités dans le cadre de la mise en correspondance stéréoscopique. À la suite des considérations purement géométriques, les approches d'appariement ont été réparties en fonction de la méthode d'optimisation utilisée. Ainsi, nous avons distingué deux catégories : les approches locales et les approches globales. Les approches locales exploitent le voisinage d'un point dans une image afin de localiser son point homologue dans l'autre image via une mesure de similarité. Ces approches reposent sur la définition d'une fenêtre de taille et de forme appropriées et sur le critère de similarité sélectionné. Elles sont assez performantes, néanmoins elles restent très sensibles aux occultations, aux zones homogènes, etc. Enfin, nous avons listé un très grand nombre d'articles utilisant des approches globales. La programmation dynamique représente la technique la plus répandue. Elle est appliquée entre deux séquences de pixels sur les lignes épipolaires. Une autre approche visant à résoudre le problème de correspondance de manière globale fait appel à la théorie des graphes. Contrairement à la programmation dynamique, cette approche permet de prendre en compte une cohérence bidimensionnelle par rapport à la scène observée. Dans cette thèse, nous nous intéressons aux méthodes locales dans la mesure où nous souhaitons estimer un champ de disparité aussi fidèle que possible en un temps réduit. Ces approches nous apparaissent comme étant les plus adaptées à notre domaine d'application. Ainsi, nous avons privilégié ce type de méthodes afin de concevoir une nouvelle approche d'appariement dense, thème qui sera abordé dans les chapitres suivants.

MISE EN CORRESPONDANCE STÉRÉOSCOPIQUE LOCAL POUR LES SCÈNES ROUTIÈRES

Sommaire

2.1	Introduction	27
2.2	La méthode proposée	28
2.2.1	Fonction du coût proposée	28
2.2.2	Agrégation des coûts	29
2.2.3	Post-traitement	31
2.3	Les résultats expérimentaux	32
2.3.1	Évaluation de la fonction de coût proposée	32
2.3.2	Évaluation de la méthode d'agrégation croisée modifiée	33
2.3.3	Évaluation de l'algorithme proposé sur la plateforme KITTI	33
2.4	Conclusion	34

2.1 Introduction

Dans ce chapitre, nous présentons une nouvelle variante de la fonction de transformation de recensement (CT) qui est plus robuste contre les changements radiométriques dans les scènes de route réelles. Nous démontrons que la fonction de coût proposée surpasse les fonctions de coût classiques en utilisant le repère KITTI. La méthode d'agrégation des coûts est également améliorée pour tenir compte de l'information sur les contours. Cela permet d'améliorer considérablement les coûts agrégés, en particulier dans les régions homogènes. La stratégie "Winner-Takes-All" (WTA) est utilisée pour calculer les valeurs de disparité. Une étape de post-traitement est effectuée pour éliminer les ambiguïtés de correspondance restantes. Des expériences ont été menées sur le nouveau set de données Middlebury ainsi que sur les scènes réelles de trafic routier issues de la base de données KITTI. Les résultats de disparité obtenus ont démontré que la méthode proposée est prometteuse.

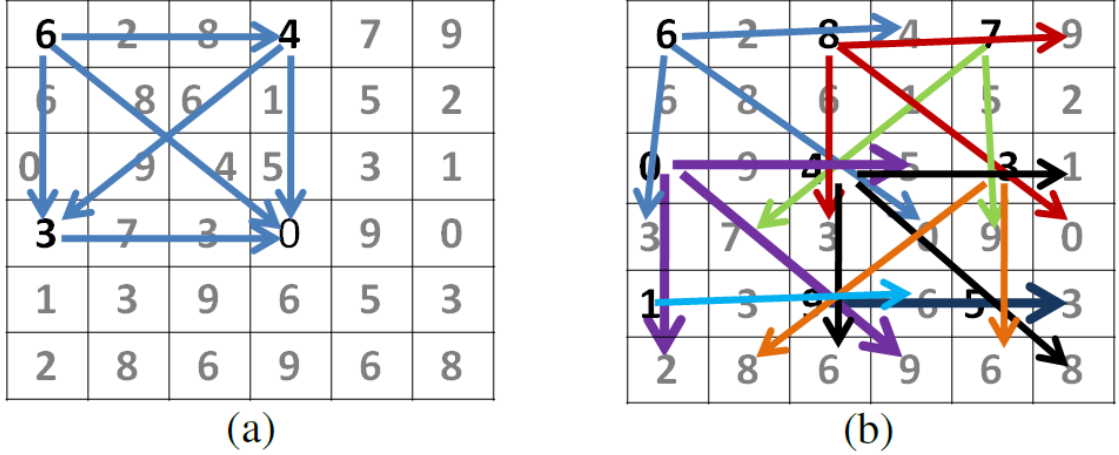


FIGURE 2.1 – Construction de la chaîne de bits pour CCC : "000000" (a) et pour notre variante : "000100111110111110" (b)

2.2 La méthode proposée

2.2.1 Fonction du coût proposée

La transformation de recensement [73] a été largement utilisée dans les algorithmes de correspondance stéréo en raison de sa robustesse face aux différences radiométriques. Les auteurs de [48] ont récemment développé une fonction de coût C_{CCC} comme variante de la fonction C_{CT} . La chaîne de bits CCC est obtenue en comparant chaque pixel de la fenêtre de support avec ceux de son voisinage direct dans le même sens que les aiguilles d'une montre.

Le C_{CCC} est calculé par la distance de Hamming (D_H) entre un pixel $p = (x, y)$ dans l'image de référence (I_1) et son correspondant théorique $q = (x, y - d)$ dans l'image cible (I_2) à une disparité d comme suit :

$$C_{CCC}(x, y, d) = D_H(CCC_{I_1}(x, y), CCC_{I_2}(x, y - d)), \quad (2.1)$$

où la chaîne de bits, CCC, est définie par :

$$CCC_I(u, v) = \bigotimes_{\substack{i=1:step:n \\ j=1:step:m}} (\xi(I(u + i, v + j), I(u + i', v + j'))), \quad (2.2)$$

où $n \times m$ désigne la fenêtre de support, $\bigotimes$ désigne l'opération de concaténation, et la fonction ξ est définie comme suit :

$$\xi(p, q) = \begin{cases} 1 & p \leq q \\ 0 & otherwise \end{cases} \quad (2.3)$$

Dans l'équation (2), $(i', j') = \{(i, j + step); (i + step, j + step); (i + step, j); (i + step, j -$

$step\}$, qui représente le voisinage immédiat dans le sens des aiguilles d’une montre d’un pixel de coordonnées (u, v) . Notez que $(j + pas) < m$, $(i + pas) < n$ et $(j - pas) \geq 0$. $step$ est une valeur empirique, choisie pour sauter quelques pixels dans la fenêtre de support. Bien que le coût CCC permette de prendre en compte l’information des pixels voisins, il arrive que dans certains cas l’information la plus significative soit située ailleurs que dans les pixels fournis par le pas choisi. En outre, la valeur de pas doit être prise de manière à ne pas dépasser la moitié de la taille de la fenêtre de support, sinon la chaîne de bits résultante pourrait ne pas fournir une chaîne de bits cohérente pour décrire le patch considéré. La figure 2.1(a) présente un cas typique où la valeur $step$ est fixée à 3 pour une fenêtre de support (5×5) . Par conséquent, dans le présent document, nous proposons de prendre deux valeurs de pas différentes. La première ($step_1$) est utilisée pour prendre un pixel et sauter son voisin tandis que la seconde ($step_2$) est utilisée pour la procédure de comparaison. Cette dernière est choisie pour être plus flexible dans la mesure où elle recherche le pixel véhiculant l’information la plus significative de manière plus large.

Pour mieux illustrer le coût proposé, la figure 2.1(b) présente la chaîne de bits proposée avec $step_1$ fixé à 2, et $step_2$ fixé à 3.

De plus, les gradients d’image étant moins sensibles aux changements radiométriques et aux motifs répétitifs, la fonction de coût proposée est mise en œuvre sur les principales directions du gradient $\delta I/\delta x$ et $\delta I/\delta y$ pour l’image d’entrée. Ainsi, la chaîne de bits proposée est définie comme suit :

$$GCCC_I(u, v) = \otimes_{\frac{\delta I}{\delta x}, \frac{\delta I}{\delta y}} \otimes_{i=1:step_1:n} \otimes_{j=1:step_1:m} (\xi(\delta I(u + i, v + j), \delta I(u + i', v + j'))) \quad (2.4)$$

où $(i', j') = \{(i, j + step_2); (i + step_2, j + step_2); (i + step_2, j); (i + step_2, j - step_2)\}$. Comme la fonction de correspondance des coûts C_{CCC} , la fonction de coût proposée G_{CCC} est calculée par la distance de Hamming comme suit :

$$G_{CCC}(x, y, d) = D_H(G_{CCC_{I1}}(x, y), G_{CCC_{I2}}(x, y - d)) \quad (2.5)$$

Les performances de la fonction de coût proposée, appliquée au couple stéréo #0 de l’ensemble de données d’entraînement KITTI, sont illustrées à la figure 2.2, qui montre l’amélioration des résultats de disparité, zones indiquées, par rapport à la fonction C_{CCC} originale.

2.2.2 Agrégation des coûts

Nous portons notre attention sur la méthode d’agrégation basée sur les croisements telle que décrite par [76]. Ainsi, nous proposons d’ajouter l’information sur les contours comme critère supplémentaire dans la stratégie de construction des régions. Selon l’étude présentée dans [54], la méthode de détection des contours de Canny est la technique la plus performante. Par conséquent, cet algorithme de détection est retenu dans notre étude. L’exécution de ce détecteur de contour pour une image donnée fournit une image binaire I_b . Ensuite, pour chaque pixel $p = (x, y)$, on détermine ses quatre

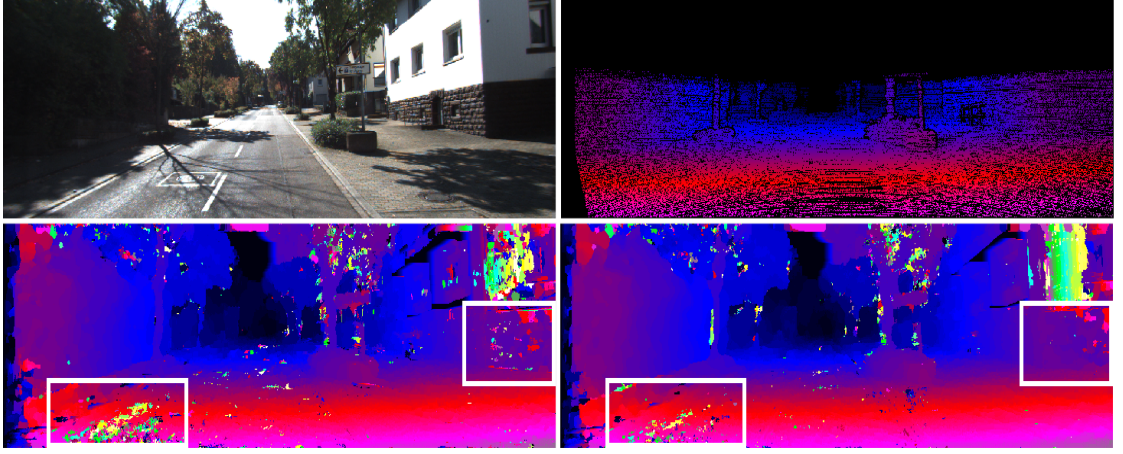


FIGURE 2.2 – De haut en bas et de gauche à droite : l'image de gauche, la carte de disparité de la vérité du sol, la carte de disparité calculée sur la base de C_{CCC} et la carte de disparité utilisant le coût proposé

longueurs de bras (c'est-à-dire la gauche, la droite, le haut et le bas) représentées par $\{hp^+, hp^-, vp^+, vp^-\}$, séparément. L'algorithme procède en deux étapes, dans la première étape, pour un pixel p son bras gauche respectivement (droite, haut, bas) s'arrête lorsqu'il trouve un point final p_l qui dépasse une des règles suivantes :

- (i) $D_c(p_l, p) < \tau$, où $D_c(p_l, p)$ représente la différence de couleur entre p_l et p , et τ est une valeur seuil prédéfinie. $D_c(p_l, p) = \max_{i=R,G,B} |I_i(p_l) - I_i(p)|$
- ii) $D_s(p_l, p) < L$, où $D_s(p_l, p)$ représente la distance spatiale entre p_l et p , et L est une longueur maximale prédéfinie. $D_s(p_l, p) = |p_l - p|$.
- iii) $I_b(p_l) = I_b(p)$, qui représente l'information sur le contour entre p et p_l .

Les deux premières règles décrites ci-dessus fixent des limites à la fois sur la similarité des couleurs et la longueur des bras avec deux seuils τ et L . Dans le cas de régions homogènes, où la similarité de couleur est plus élevée, les deux règles ci-dessus sont insuffisantes pour décrire la structure de la région locale. Dans ce but, le critère d'information de contour est employé pour définir la région adaptative. Dans la deuxième étape, sur la base de la longueur des bras hp^+, hp^-, vp^+, vp^- détectée pour un pixel p , deux segments transversaux orthogonaux, le segment horizontal $H(p)$ et le segment vertical $V(p)$ sont considérés. Par conséquent, la région de support de forme adaptative $U(p)$ pour le pixel p est donnée par :

$$U(p) = \bigcup_{q \in V(p)} H(q), \quad (2.6)$$

où q est le pixel de voisinage situé sur le segment vertical $V(p)$. Le même processus est effectué pour déterminer la région de support $U'(q)$ du pixel hypothétique correspondant $q = (x, y - d)$ dans la cible. Ensuite, la fenêtre combinée implique l'intersection des deux



FIGURE 2.3 – De gauche à droite : image de gauche, carte de disparité de la vérité du terrain, image de contour, carte de disparité calculée en fonction de la région adaptative croisée, et carte de disparité obtenue par la méthode modifiée.

régions de support comme suit :

$$C(p, d) = \frac{1}{\|U_d(p)\|} \sum_{q \in U_d(p)} C(q, d), \quad (2.7)$$

où $U_d(p) = \{(x, y) \mid (x, y) \in U(p), (x, y - d) \in U'(q)\}$, $\|U_d(p)\|$ est le nombre de pixels dans $U_d(p)$ utilisé pour normaliser le coût agrégé $C(p, d)$. En général, pour réduire le bruit et lisser les images d'entrée, un filtre médian (3×3) est appliqué. En outre, la longueur minimale du bras requise pour toutes les directions est fixée à 1 pour garantir une région de support minimale de 9 pixels.

2.2.3 Post-traitement

Les résultats de disparité obtenus à l'aide des étapes précédentes contiennent des valeurs aberrantes dans les régions occultées et dans les discontinuités de profondeur. Afin de détecter ces valeurs aberrantes, la méthode de vérification de la cohérence gauche-droite est appliquée entre les cartes de disparité gauche et droite. Puis, pour remplir les aberrations détectées, la méthode de remplissage de trous [27] est employée. Elle consiste à les remplir avec les plus petites valeurs de disparité des disparités fiables les plus proches situées sur la même ligne de balayage (ligne). Cette dernière méthode peut générer des artefacts de type stries dans la carte de disparité. Pour résoudre ce problème, nous proposons d'appliquer un filtre médian via une région croisée [59]. Enfin, un filtre médian, d'une taille de (5×5) est effectué afin de lisser davantage la carte de disparité finale en éliminant les bruits restants.

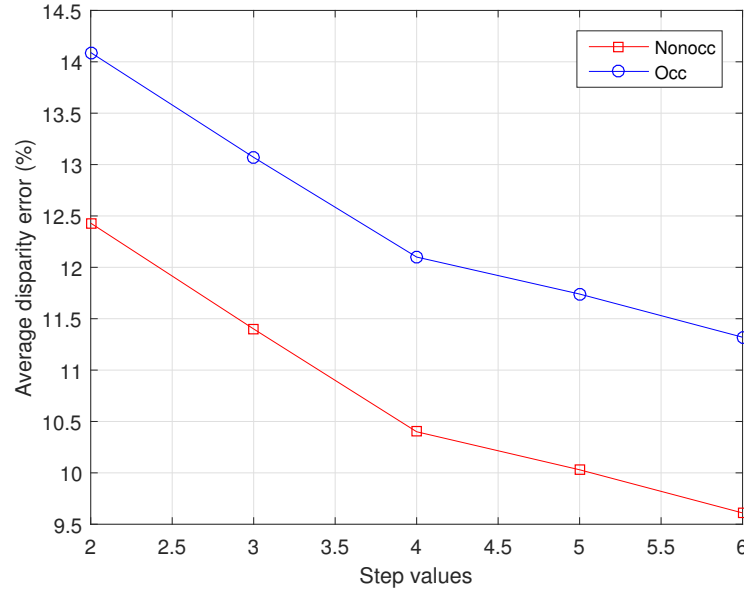


FIGURE 2.4 – Erreurs de disparité moyennes par rapport aux valeurs de $step_2$ pour les 10 premières images des ensembles d’entraînement KITTI

2.3 Les résultats expérimentaux

2.3.1 Évaluation de la fonction de coût proposée

Afin d’illustrer l’efficacité de la fonction de coût proposée, nous l’avons évaluée par rapport aux meilleures fonctions de coût définies dans l’enquête [49], $C_{DIFFCensus}$ et $C_{DIFFCCC}$ et ceci sur le benchmark KITTI, où des changements radiométriques réels apparaissent. Aucun processus de raffinement n’a été effectué dans cette comparaison. La méthode d’agrégation adaptative au coût a été utilisée. La stratégie WTA locale a été adoptée afin de générer les cartes de disparité. Les valeurs optimales des paramètres définis dans [49] ont été considérées. Nous avons effectué des tests approfondis sur les 10 premières images de KITTI afin de trouver les valeurs de pas optimales qui minimisent le taux d’erreur. La figure 2.4 présente les erreurs de disparité en fonction des valeurs de $step_2$. Par conséquent, $step_2$ a été fixé à 6 (chaîne de bits de 64 éléments). Le tableau 2.1 présente le pourcentage moyen de pixels incorrects dans les régions occultées et non occultées, calculé au seuil d’erreur de 3 pixels par défaut pour l’ensemble d’entraînement KITTI. Les résultats obtenus en utilisant notre coût sont les plus bas dans les deux régions. En effet, l’amélioration est de l’ordre de 1,04, 1,60 pour la région non occultée et de 1,02, 1,58 pour les autres zones, par rapport aux coûts $C_{DIFFCensus}$ et $C_{DIFFCCC}$, respectivement. Cette évaluation démontre que le coût proposé est plus robuste contre les changements radiométriques extérieurs réels que le coût le plus performant en matière de changements radiométriques.

TABLEAU 2.1 – Pourcentage de disparités incorrectes dans les régions occultées et non-occultées pour l’ensemble d’entraînement KITTI.

	seuil 3-px	
	<i>Non – occ</i>	<i>Occ</i>
$C_{DIFFCensus}$ [49]	12.97	14.91
$C_{DIFFCCC}$ [49]	13.53	15.47
Coût proposé	11.93	13.89

TABLEAU 2.2 – Pourcentage de disparités incorrectes dans les régions non occultées pour la base d’entraînement de Middlebury.

Méthode Agrégation	Seuil 1 px	Seuil 2 px	Seuil 3 px
	Nocc	Nocc	Nocc
Original method	19.69	14.93	13.41
Proposed one	18.56	14.11	12.61

2.3.2 Évaluation de la méthode d’agrégation croisée modifiée

Dans cette section, nous avons effectué des tests pour évaluer l’effet de la prise en compte d’information sur les contours dans la méthode d’agrégation croisée par rapport à la méthode classique. La figure 2.3 illustre un cas de performance de la méthode d’agrégation proposée appliquée à la paire Playtable de la base de données Middlebury, qui présente des améliorations dans les résultats de disparité lors de l’incorporation des informations de contour, en particulier dans les régions homogènes. De plus, nous avons réalisé des expériences sur toute la base d’entraînement Middlebury. Aucun processus de raffinement n’a été utilisé dans cette évaluation. Le tableau 2.2 présente la moyenne des pixels d’erreur incorrects dans la région non occultée à résolution du quart. Les erreurs sont calculées à 3 seuils d’erreur de pixels différents. Les résultats obtenus démontrent clairement une amélioration des résultats de disparité en utilisant l’agrégation de coûts modifiée.

2.3.3 Évaluation de l’algorithme proposé sur la plateforme KITTI

Finalement, nous avons réalisé nos expériences en évaluant la performance de l’algorithme présenté par rapport à l’état actuel de l’art sur le dataset de test KITTI [13]. Les paramètres de l’algorithme proposé sont les suivants : la taille des fenêtres de support a été fixée à $(m \times n) = (9 \times 9)$. Les seuils de similarité spatiale et de couleur ont été fixés à $L = 9$ et $\tau = 20$, respectivement. De plus, un léger filtre gaussien de taille $((3 \times 3), r = 0)$ a été effectuée sur les images en échelle de gris avant de calculer les dérivées partielles, dans l’étape de calcul des coûts, afin de réduire le bruit et de lisser les contours des images. Nos expériences ont été réalisées sur un ordinateur de bureau équipé d’un processeur In-

TABLEAU 2.3 – Résultats de la plateforme d’évaluation KITTI pour le seuil par défaut de 3 pixels. Colonnes de gauche à droite : méthode ; pourcentage de pixels erronés dans les régions non-occultées et au total ; erreur moyenne de disparité dans les régions non-occultées ; erreur moyenne de disparité au total. Date de l’évaluation : 17 janvier 2017

Method	Out-Noc	Out-All	Avg-Noc	Avg-All
Our Method	8.71 %	10.05 %	2.1 px	2.7 px
Deep-Raw	8.93 %	11.07 %	3.9 px	4.9 px
S+GF (Cen)	9.03 %	11.21 %	2.1 px	3.4 px
CrossCensus	9.46 %	10.86 %	2.3 px	2.7 px
SymST-GP	9.79 %	11.66 %	2.5 px	3.3 px
SM_GPTM	9.79 %	11.38 %	2.1 px	2.6 px
LAMC-DSI	9.82 %	11.49 %	2.1 px	2.7 px
IIW	10.78 %	12.62 %	3.3 px	4.3 px
SDM	10.95 %	12.14 %	2.0 px	2.3 px
HLSC_mesh	11.22 %	12.82 %	2.3 px	2.9 px
GF (Census)	11.65 %	13.76 %	4.5 px	5.6 px

tel core i5 à 3,0 GHz et de 4 Go de mémoire. L’implémentation de l’algorithme proposé sur le CPU a conduit à un temps de calcul moyen de 125s pour KITTI dataset. Le tableau 2.3 présente le résultat de l’évaluation de la plateforme KITTI. Notre algorithme est classé 64ème parmi plus de 84 algorithmes de correspondance stéréo et est l’une des meilleures méthodes locales. En effet, il surpasse de nombreux algorithmes locaux connus, tels que Cross-Census[76], GF(Census)[27], LAMC-DSI[59], S+GF(Cen)[75].

2.4 Conclusion

Dans ce chapitre, nous avons présenté un nouvel algorithme d’appariement stéréoscopique utilisant une nouvelle variante de la fonction de coût de Census pour l’étape de calcul des coûts. Des résultats expérimentaux, utilisant des scènes routières réelles de la base de données KITTI, révèlent que la variante proposée mène aux erreurs moyennes de disparité les plus faibles par rapport à l’algorithme le plus performant dans cette base de données. De plus, une méthode locale basée sur l’agrégation croisée est améliorée afin d’incorporer les informations sur les contours. Les coûts agrégés modifiés ont conduit à une amélioration des résultats de disparité. Une étape de post-traitement est effectuée pour éliminer tout bruit restant. Les résultats de disparité obtenus sont considérés comme prometteurs.

APPROCHE HIÉRARCHIQUE

Sommaire

3.1	Introduction	35
3.2	Fonction de coût proposée	36
3.3	Approche hiérarchique	38
3.3.1	Méthode d'agrégation adaptative	38
3.3.2	Schéma hiérarchique de traitement	41
3.4	Résultats expérimentaux	42
3.4.1	Évaluation de la fonction du coût proposée	43
3.4.2	Évaluation de l'algorithme hiérarchique proposé	44
3.4.3	Évaluation sur la plateforme en ligne de la base KITTI	45
3.4.4	Comparaison avec d'autres algorithmes hiérarchiques	45
3.4.5	Complexité algorithmique	45
3.5	Conclusion	47

3.1 Introduction

Le système visuel humain traite la correspondance stéréoscopique à différentes échelles [41]. Conformément à cette l'étude, les informations des échelles approximatives et des échelles finales sont traitées de manière interactive au cours de la recherche de correspondances par le système visuel humain. Cette bio-inspiration justifie le recours à plusieurs échelles au lieu de se limiter uniquement à l'échelle de l'image. Pour bénéficier de ce principe, une approche hiérarchique (grossière à fine) est adoptée pour la mise en correspondance stéréoscopique [42]. Cette approche a été principalement adoptée pour exploiter les représentations multi-échelles des images [58, 63] pour réduire l'espace de recherche [71, 30], ainsi que pour représenter l'agrégation des coûts, soit au moyen d'un modèle d'agrégation des coûts, soit au moyen d'une approche multi-échelle utilisant la diffusion anisotropique [47] ou par le problème d'optimisation des moindres carrés pondérés [75]. Une nouvelle approche hiérarchique basée sur une modification du noyau gaussien 3D, proposée par [59]. Ce chapitre présente nos contributions axées sur deux vallées. Tout d'abord, afin de traiter le problème des changements radio-métriques, nous

proposons une nouvelle variante de la fonction de coût CT . Cette variante intègre la vérification du contour latéral entre le pixel étudié et les autres pixels voisins. Grâce à cette information, le coût proposé devient plus robuste aux variations radio-métriques par rapport à la fonction CT classique. Par la suite, nous introduisons une nouvelle approche hiérarchique exploitant la méthode d'agrégation adaptative [76]. Pour cela, notre démarche consiste à construire un volume de coûts multi-échelle de $3D$ (également appelé espace de disparité d'image (ISD)), et ensuite à agréger les coûts en utilisant une fonction exponentielle [44] pour créer un ISD à échelle fine. Le rôle de celle-ci consiste à contrôler la cohérence inter-échelle entre les différentes échelles d'image.

3.2 Fonction de coût proposée

La transformée de Census (CT) est une forme de transformée locale non-paramétrique largement utilisée en vue de définir des fonctions de coût de correspondance au sein des algorithmes de correspondance stéréoscopique, proposée par [73]. Comme le montre l'étude réalisée par [23], cette fonction de coût fait partie de celles les plus robustes vis-à-vis des changements radiométriques. En effet, pour un voisinage de support $N(m \times n)$ d'un pixel donné $\mathbf{p} = (\mathbf{x}, \mathbf{y})$, une chaîne de bits est construite pour $\mathbf{p}$ avec une longueur de $m \times n$ en comparant sa valeur d'intensité I_p aux intensités des pixels voisins du support N . Les pixels dont l'intensité est supérieure à $\mathbf{p}$ sont donc mis à 1, tandis que les autres sont mis à 0. La fonction de correspondance de coût CT C_{CT} se calcule à travers la distance de Hamming (D_H) entre un pixel $\mathbf{p}$ dans l'image de référence (I_1) et son correspondant théorique $q = (x, y - d)$ dans l'image cible (I_2) à une disparité d comme suit :

$$C_{CT}(x, y, d) = D_H(CT_{I_1}(x, y), CT_{I_2}(x, y - d)) \quad (3.1)$$

Dans l'équation 3.1 CT désigne la chaîne de bits définie comme suit :

$$CT_I(u, v) = \otimes_{\substack{i=1, n \\ j=1, m}} \xi(I(u, v), I(u + i, v + j)), \quad (3.2)$$

Où $n \times m$ représente la fenêtre de corrélation, $\otimes$ désigne l'opération de concaténation et ξ une fonction est définie par :

$$\xi(p_1, p_2) = \begin{cases} 1 & I_{p_1} \leq I_{p_2} \\ 0 & otherwise \end{cases} \quad (3.3)$$

Les distorsions radiométriques se réfèrent aux pixels qui correspondent au même point $3D$ et qui ont des valeurs d'intensité différentes dans les images stéréoscopiques [21]. La transformée de Census peut traiter efficacement ce problème, vu que celle-ci ne traite pas les valeurs d'intensité; elle décrit seulement l'ordre relatif des valeurs d'intensité dans la fenêtre de corrélation. Par conséquent, la variation du gain et du biais entre les deux images stéréoscopiques ne modifiera pas l'ordre des pixels dans la fenêtre de corrélation. Dans notre travail, on a remarqué que la séquence relative des intensités des

pixels produite par la transformée de Census ne décrit pas suffisamment la structure de la fenêtre de corrélation concernée, afin de combler ce manque d'information, on propose d'ajouter l'information sur les contours pour avoir une structure complète où la position, concernant les contours, et la séquence relative des pixels sont introduites conjointement. En effet, nous introduisons l'information sur les côtés du contour dans la chaîne de bits construite par CT . Cette idée est inspirée des filtres préservant les contours, dont l'information sur les contours est utilisée pour rassembler les coûts sous une forme pondérée.

Plus précisément, pour une fenêtre de corrélation $(n \times n)$, le terme suivant est calculé entre deux pixels p_1 et p_2 dans la fenêtre considérée :

$$E(p_1, p_2) = (I_{p_1} - \mu_k)(I_{p_2} - \mu_k) \quad (3.4)$$

Où μ_k est la valeur moyenne de I dans la fenêtre corrélation $(n \times n)$.

Le terme $E(p_1, p_2)$ a un signe positif si I_{p_1} et I_{p_2} se trouve du même côté du contour, et a une valeur négative dans le cas échéant [27]. Dans notre travail, on calcule le terme ci-dessus dans les domaines de gradient $\delta I/\delta x$ et $\delta I/\delta y$. La raison derrière l'intégration des gradients est d'avoir plus de détails sur les contours dans les deux directions x et y . Ainsi, le terme proposé est défini comme suit :

$$\delta E(p_1, p_2) = (\delta I_{p_1} - \delta \mu_k)(\delta I_{p_2} - \delta \mu_k) \quad (3.5)$$

Où δI représente le gradient de l'image dans les directions x ou y . $\delta \mu_k$ est la valeur moyenne de δI dans la direction x ou y dans la fenêtre de corrélation $(n \times n)$. Les gradients de l'image caractérisent les propriétés structurelles de l'image (telles que les points et les contours). De la même manière, $\delta E(p_1, p_2)$ a un signe positif si δI_{p_1} et δI_{p_2} se trouvent dans le même coté du contour, et a une valeur négative dans le cas échéant. Cette information est incluse dans l'initiale chaîne de bits CT , en calculant $\delta E(p_1, p_2)$ entre le pixel central p_1 et son pixel p_2 dans la fenêtre de corrélation N . Puis, si $\delta E(p_1, p_2)$ est positif, dans ce cas on met 0 comme valeur correspondante dans la chaîne de bits, sinon on met 1 dans le cas échéant.

Cette approche produit une chaîne de bits étendue :

$$ECT_I(u, v) = \bigotimes_{i=1, n} \bigotimes_{j=1, m} \xi_{edge}(I(u, v), I(u + i, v + j)), \quad (3.6)$$

Où ECT est la nouvelle chaîne de bits avec une taille double $2(m \times n)$ par rapport à la

chaîne originale, et la fonction ξ_{edge} est définie comme suit :

$$\xi_{edge}(p_1, p_2) = \begin{cases} 00 & I_{p_1} > I_{p_2}, \delta E(p_1, p_2) \geq 0 \\ 01 & I_{p_1} > I_{p_2}, \delta E(p_1, p_2) < 0 \\ 10 & I_{p_1} \leq I_{p_2}, \delta E(p_1, p_2) \geq 0 \\ 11 & I_{p_1} \leq I_{p_2}, \delta E(p_1, p_2) < 0 \end{cases} \quad (3.7)$$

Cette nouvelle formule décrit non seulement la différence d'intensité entre le pixel centre p_1 et son pixel voisin p_2 , mais aussi l'emplacement de ces pixels par rapport aux côtés des contours dans les deux directions x et y .

Similaire à la fonction du coût C_{CT} , notre fonction du coût proposée C_{ECT} est calculée par la distance de Hamming comme suit :

$$C_{ECT}(x, y, d) = D_H(ECT_{I_1}(x, y), ECT_{I_2}(x, y - d)) \quad (3.8)$$

Les performances de la fonction du coût proposée par rapport au coût original C_{CT} , sont présentées dans la figure 3.1 sur deux images de la base de données d'entraînement de KITTI numéros #7 (gauche) et #131 (droite), qui représente des améliorations au niveau des erreurs de disparités grâce à l'utilisation de l'information du contour.

3.3 Approche hiérarchique

Dans cette section, nous présentons notre schéma de traitement hiérarchique pour la correspondance stéréoscopique. Les scènes réelles présentent des difficultés complexes pour tout algorithme de mise en correspondance stéréoscopique, à savoir de vastes régions non texturées et du bruit dans l'image. Pour remédier à ces problèmes, nous proposons de calculer les disparités sous forme hiérarchique.

3.3.1 Méthode d'agrégation adaptative

Les méthodes de correspondance stéréoscopique locale reposent sur l'hypothèse que les pixels appartenant au même ensemble de voisinage possèdent les mêmes disparités avec une forte probabilité. Ainsi, chaque méthode locale établit son propre système de voisinage afin de regrouper autant de pixels ayant les mêmes disparités que possible. L'approche adaptative basée sur la forme pour chaque pixel de l'image a gagné en intérêt vu qu'elle se limite pas aux fenêtres rectangulaires, qui contiennent des pixels aberrants, surtout dans les régions proches des discontinuités. Dans cette section, notre attention est portée sur la méthode d'agrégation adaptative proposée par [76]. Sa stratégie consiste à construire un squelette adaptatif basé sur la similarité de couleur entre le pixel étudié et les autres pixels voisins.

Concrètement, pour chaque pixel $p = (x, y)$, on détermine les quarts branches gauche, droite, haut et bas représentés respectivement par hp^+ , hp^- , vp^+ , vp^- .

L'algorithme procède en deux étapes, comme indiqué dans la figure 3.2. Dans la

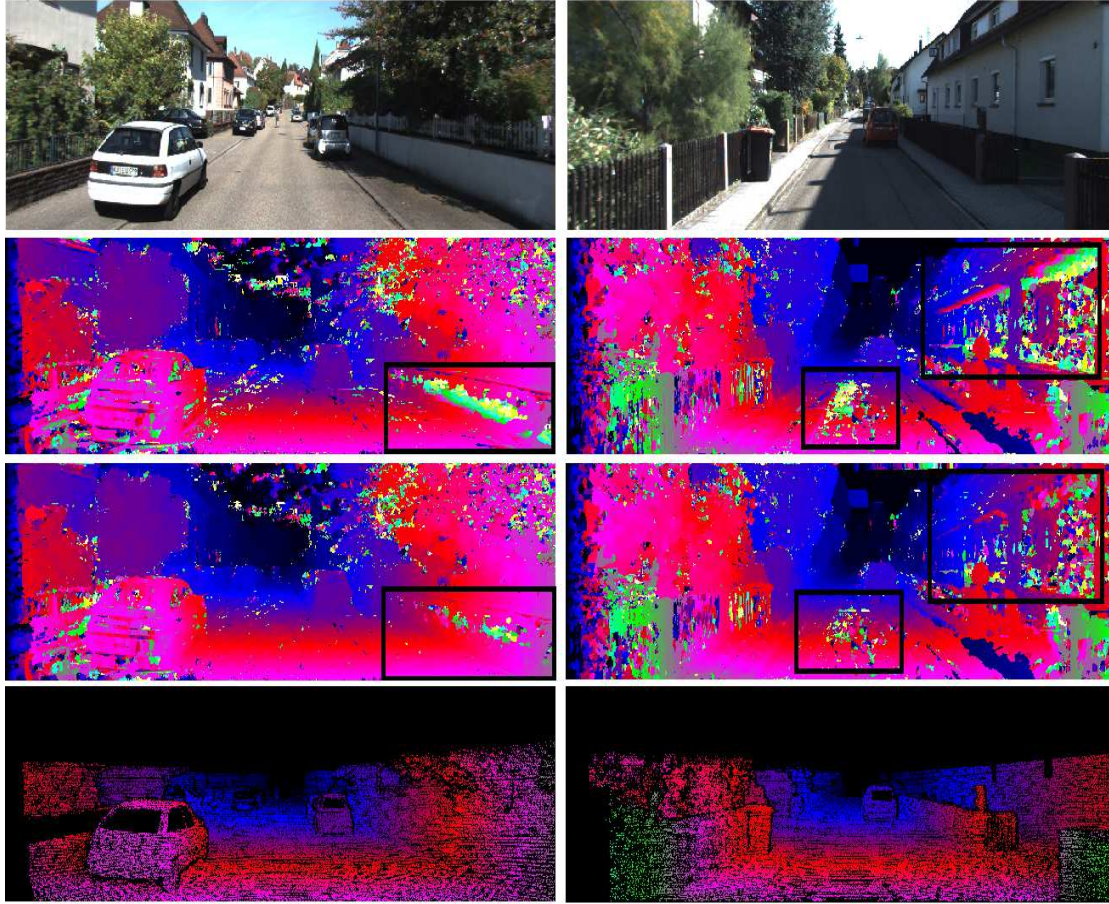


FIGURE 3.1 – Quelques résultats des cartes de disparités produites via la fonction du coût classique CT et la fonction du coût proposée pour les deux images stéréo #7 et #131 issues de la base de données KITTI 2012. de haut en bas : L'image de référence, la carte de disparité calculée en utilisant la fonction du coût CT , la carte de disparité calculée à partir de la fonction du coût proposée et finalement les vérités du terrain. Les cartes de disparités sont calculées après l'utilisation de la méthode d'agrégation [76]. Les régions cadrées représentent des zones d'amélioration.

première étape, pour un pixel p la longueur de sa branche gauche respectivement (droite, haut, bas) se fixe dès qu'on détecte un pixel p_l qui viole l'une des règles suivantes :

- (i) $D_c(p_l, p) < \tau$, Où $D_c(p_l, p)$ représente la différence de couleur entre le pixel p_l et le pixel p , τ est une valeur seuil prédéfinie. La différence de couleur est définie par $D_c(p_l, p) = \max_{i=R,G,B} |I_i(p_l) - I_i(p)|$
- ii) $D_s(p_l, p) < L$, Où $D_s(p_l, p)$ représente la distance spatiale entre p_l et p , et L est une longueur maximale prédéterminée. La distance spatiale est définie par $D_s(p_l, p) = |p_l - p|$.

Les deux règles évoquées ci-dessus imposent des limitations sur la similarité de couleur

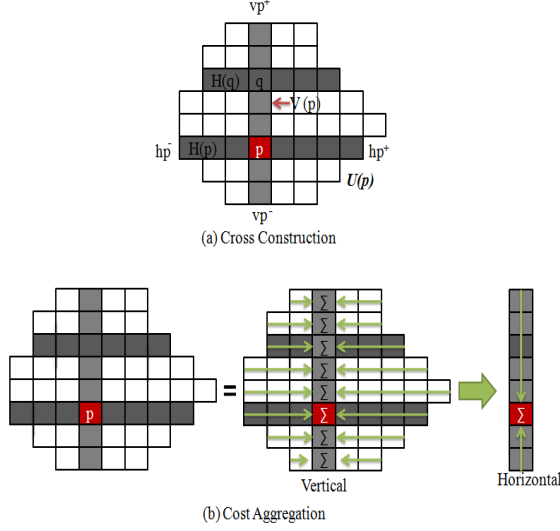


FIGURE 3.2 – méthode adaptative pour le pixel p : (a) un squelette est construit pour chaque pixel. La région adaptative pour p est effectuée en regroupant les branches horizontales des pixels (q par exemple). (b) le coût du pixel p est rassemblé en deux passages sur la direction horizontale et verticale.

et les longueurs des quarts de branches avec les seuils τ et L . Dans la deuxième étape, on construit deux segments orthogonaux croisés, qui sont le segment horizontal $H(p)$ le segment vertical $V(p)$ comme le montre la figure 3.2. La forme adaptative de la région $U(p)$ pour le pixel p peut s'exprimer comme suit :

$$U(p) = \bigcup_{q \in V(p)} H(q) \quad (3.9)$$

Où q est un pixel voisin situé sur le segment vertical $V(p)$.

La particularité de cette méthode est due au fait qu'on n'utilise pas seulement de la région $U(p)$ du pixel p dans l'image de référence pour construire la région d'agrégation adaptative mais aussi il faut considérer l'autre région $U'(q)$ du pixel candidat correspondant $q = (x, y - d)$ dans l'image cible à une disparité d . Ainsi, la région d'agrégation adaptative pour un pixel p et est son candidat correspondant q est le résultat de l'intersection des deux régions $U(p)$ et $U'(q)$. Finalement, Le coût de correspondance total entre le pixel p et q à une disparité d est définie comme suit :

$$C(p, d) = \frac{1}{\|U_d(p)\|} \sum_{q \in U_d(p)} C(q, d) \quad (3.10)$$

Où $U_d(p) = \{(x, y) \mid (x, y) \in U(p), (x, y - d) \in U'(q)\}$, $\|U_d(p)\|$ est le nombre total de

pixels dans la région $U_d(p)$, utilisé pour normaliser le coût obtenu $C(p, d)$.

Généralement, pour réduire le bruit et lisser les contours des images un filtre médian de taille (3×3) est appliqué aux deux images d'entrée. De plus, une longueur minimale de 1 des branches est demandée dans toutes les directions afin d'assurer une région d'agrégation minimale de 9 pixels.

3.3.2 Schéma hiérarchique de traitement

Les méthodes locales de mise en correspondance stéréoscopique prévoient des hypothèses implicites de lissage dans la phase d'agrégation des coûts, qui s'avèrent souvent insuffisantes pour atténuer les ambiguïtés de correspondance en raison à plusieurs facteurs tels que l'existence de grandes régions non texturées.

Afin d'améliorer la précision de correspondance dans ces régions, le processus de correspondance doit être effectué à des échelles inférieures à l'échelle originale de l'image d'entrée [45]. Par rapport aux autres algorithmes hiérarchiques, notre approche se base sur l'agrégation des informations entre les différentes échelles en utilisant une fonction exponentielle [44] simple, robuste et efficace. Ainsi, l'intégration d'informations multi-échelles en utilisant une fonction exponentielle est l'originalité de notre travail. La figure 3.3 désigne l'approche hiérarchique. L'espace multi-échelle est obtenu en appliquant successivement la pyramide gaussienne ($\eta = 2$) [6] sur les images sous-échantillonnées. D'abord, le volume de coût initial est généré pour chaque échelle de l'image, en utilisant la fonction de coût introduite dans la section 3.2. Puis, on applique la méthode d'agrégation adaptative sur le volume initial pour réduire le bruit et construire le volume raffiné des coûts (référéncé par l'agrégation intra-échelle dans la figure 3.3). En effet, le coût raffiné entre le pixel p^s et son candidat correspondant q^s pour une disparité d^s à une échelle s , où $s \in \{0, 1, \dots, S\}$ représente l'échelle, est défini par :

$$C^s(p^s, d^s) = \frac{1}{\|U_d(p^s)\|} \sum_{q^s \in U_{d^s}(p^s)} C^s(q^s, d^s), \quad (3.11)$$

Dans la mesure de contrôler l'uniformité des coûts entre les différentes échelles de l'image, on applique des fonctions exponentielles. Ainsi, le nouveau coût à une échelle données est défini comme suit :

$$\tilde{C}(p^s, d^s) = \sum_{s=0}^S \rho(C^s, \lambda^s), \quad (3.12)$$

Où $\rho(C^s, \lambda^s)$ est une fonction définie par :

$$\rho(C^s, \lambda^s) = 1 - \exp\left(-\frac{C^s(p^s, d^s)}{\lambda^s}\right) \quad (3.13)$$

L'objectif de cette fonction est double : tout d'abord, elle normalise les coûts dans l'intervalle $[0, 1]$, de sorte que l'équation 3.13 ne sera pas considérablement biaisée par l'un des coûts ; Ensuite, elle permet de contrôler facilement l'influence du coût correspondant

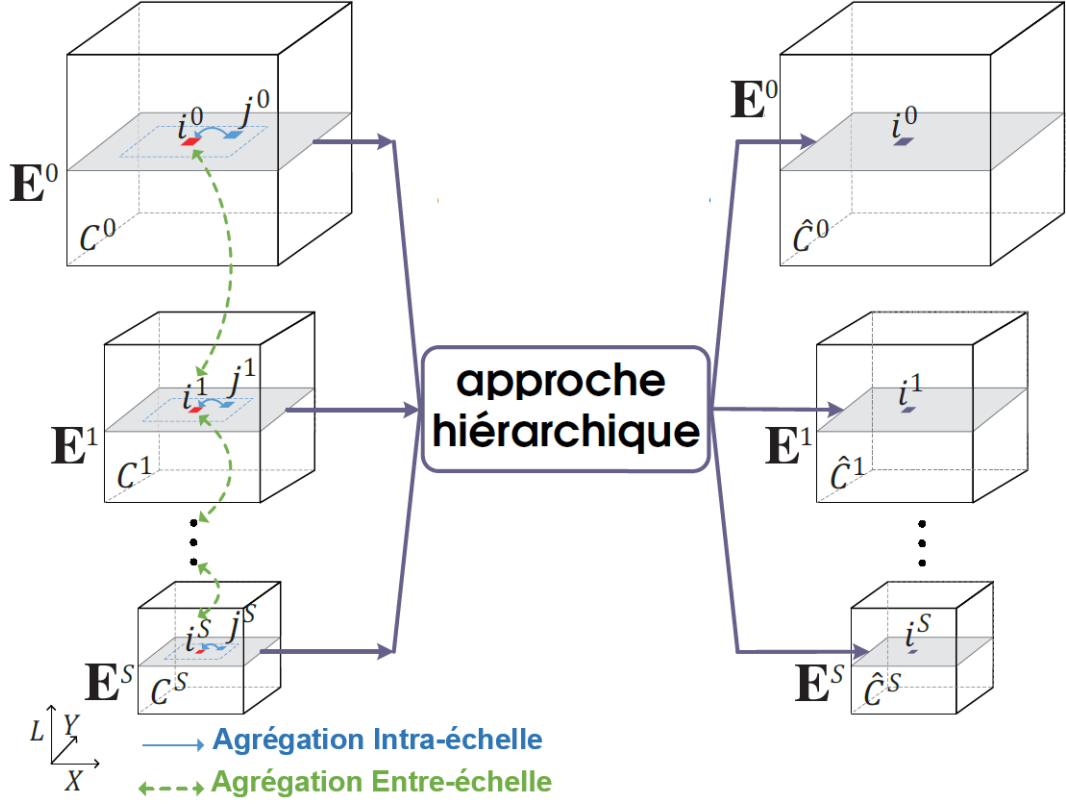


FIGURE 3.3 – L'organigramme de l'approche hiérarchique

à l'échelle s sur le coût final par le paramètre λ^s .

3.4 Résultats expérimentaux

L'évaluation de l'algorithme proposé est effectuée sur la base de donnée KITTI pour la vision stéréoscopique de la navigation autonome. Le jeu de paramètres de notre algorithme est : la taille de la fenêtre de corrélation est fixée à (9×9) , les seuils $L = 7$ et $\tau = 20$ sont utilisés dans la construction de la région adaptative. Nous effectuons une recherche minutieuse des valeurs λ^s optimales qui réduisent l'erreur de moyenne des disparités défectueuses sur les 3 images de la base d'entraînement KIITI. Ainsi, nous fixons $\lambda^0 = 50$ et $\lambda^s = 3\lambda^{s-1}$. Les valeurs élevées de λ^s des échelles inférieures sont considérées pour éviter la domination des informations à basse fréquence sur les coûts finales. L'évaluation se fait en calculant le pourcentage de valeurs défectueuses des disparités par rapport à la vérité du terrain. La stratégie du "gagnant emporte tout" (WTA) est utilisée afin de générer les résultats des disparités.

TABLEAU 3.1 – Pourcentage de disparités erronées dans les régions non occultées (Nocc) et les autres régions (toutes) pour la base d’entraînement KITTI

	seuil de 2 pixels		seuil de 3 pixels		seuil de 5 pixels	
	Nocc	toutes	Nocc	toutes	Nocc	toutes
Coût de base C_{CT}	22.10	23.85	18.27	20.09	15.74	17.57
C_{Cstent}	23.60	25.30	20.15	21.90	17.90	19.65
C_{CCC}	21.94	23.70	17.78	19.62	15.13	16.99
Coût Proposé C_{ECT}	19.62	21.43	15.50	17.38	12.95	14.85

TABLEAU 3.2 – Pourcentage de disparités incorrectes dans la base de données de d’entraînement KITTI

	Seuil de 3 pixel	
	<i>toutes</i>	<i>Non – occ</i>
Approche originale	17.38	15.50
Approche hiérarchique	10.81	8.74

3.4.1 Évaluation de la fonction du coût proposée

Dans cette section, la robustesse de la fonction de coût proposée est évaluée sur les bases de données de d’entraînement KITTI. L’évaluation inclut la fonction de coût par défaut basée sur la transformée de Census C_{CT} [73] ainsi que deux variantes de ce même coût C_{Cstent} [59] et C_{CCC} [48]. Pour comparer objectivement les fonctions de coût, aucun processus de raffinement, ou schéma hiérarchique n’est effectué. Les valeurs optimales des paramètres définies dans [59] et [48] ont été retenues. Le tableau 3.1 présente l’erreur moyenne dans toutes les zones (occultations comprises) ainsi que dans les régions non occultées. Cette erreur est calculée à 3 seuils d’erreur différents pour la base d’entraînement KITTI. Les résultats obtenus montrent clairement que le coût proposé conduit au taux d’erreur le plus faible dans toutes les zones et dans les régions non occultées et ce, pour tous les seuils d’erreur. En effet, pour le seuil d’erreur par défaut de 3, les améliorations obtenues par le coût proposé de C_{ECT} sont de l’ordre de 2, 77, 4, 65, 2, 28 pour la région non-occultée et de 2, 71, 4, 52, 2, 24 pour toutes les zones, par rapport aux fonctions de coût C_{CT} , C_{Cstent} et C_{CCC} . Ces résultats montrent que la fonction de coût proposée est plus robuste aux changements radiométriques extérieurs réels par rapport aux autres fonctions de coût. À la lumière de ces résultats, la fonction de coût utilisée dans la suite de cet article est celle proposée.

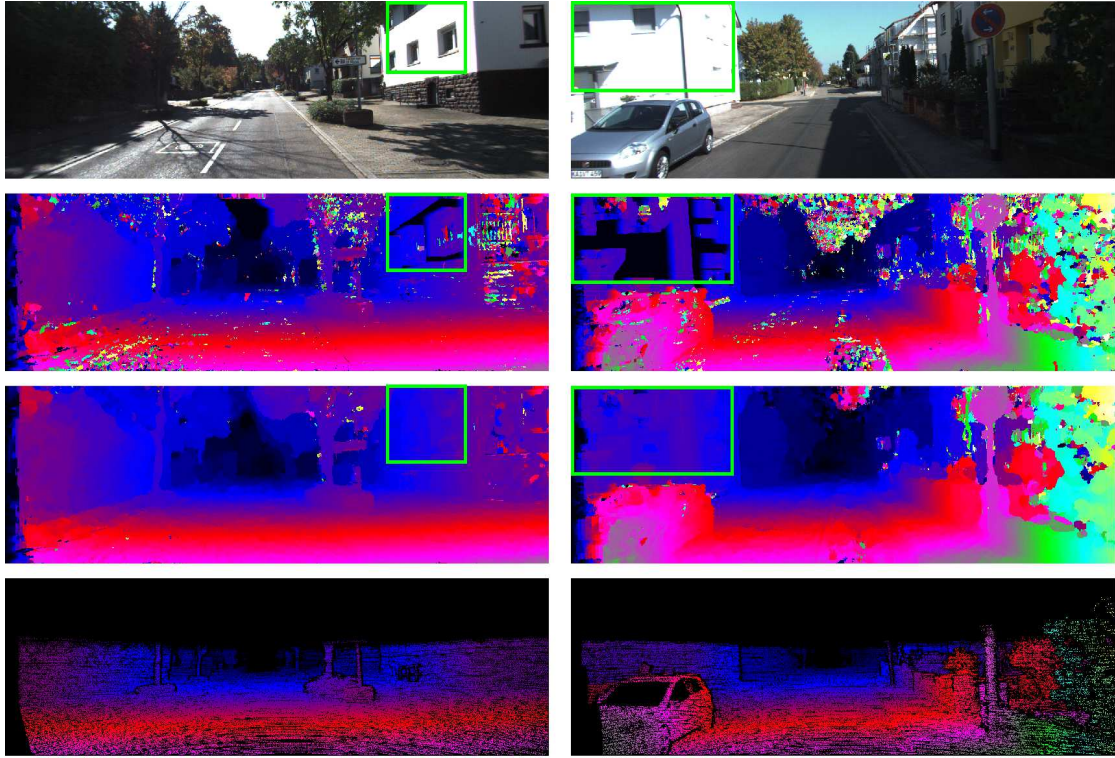


FIGURE 3.4 – Les cartes de disparité pour les deux images stéréo #0 et #2 de la base d’entraînement de KITTI. Du haut en bas : L’image de référence, carte de disparité produite sans l’utilisation du schéma hiérarchique, Carte de disparité produite en utilisant l’approche hiérarchique et les cartes de disparité vérité du terrain, les zones encadrées sont des régions non texturées

3.4.2 Évaluation de l’algorithme hiérarchique proposé

La présente section vise à évaluer les avantages de l’utilisation de l’algorithme hiérarchique proposé par rapport à l’approche traditionnelle. La figure 3.4 représente l’amélioration des résultats de disparité lors de l’exploitation du schéma hiérarchique sur deux paires #0 et #2 de l’ensemble d’entraînement KITTI. Comme prévu, les améliorations sont particulièrement prononcées dans les régions sans texture (régions colorées). Pour les deux paires, la figure montre un pourcentage d’erreur plus élevé (dans les régions non occultées) en utilisant la méthode d’agrégation croisée dans l’image à l’échelle la plus fine, sans l’utilisation du schéma hiérarchique proposé : pour la paire #0, le pourcentage d’erreur est d’environ 10,98%, et pour l’autre paire #2, il est de 17,11%, alors qu’en exploitant le schéma hiérarchique, nous obtenons une réduction significative du pourcentage d’erreur pour les deux paires, qui tombe à 3,80% pour la première paire et à 10,02% pour la seconde. première paire et 10,02 pour la seconde.

Dans le tableau 3.2, les résultats l'ensemble de la base KITTI sont rapportés et comparés en termes de précision obtenue. Les taux d'erreur pour le processus de mise en correspondance de base sont de 15,50%, 17,38%, dans toutes les régions non occultées, lorsque les images sont mises en correspondance uniquement sur la résolution d'origine et de 10,81%, 8,74% dans toutes les régions non occultées lorsqu'elles sont mises en correspondance à travers les échelles en utilisant le schéma hiérarchique. L'utilisation du schéma hiérarchique améliore les résultats de 6,57% dans toutes les zones et de 6,76% dans les zones non occultées.

3.4.3 Évaluation sur la plateforme en ligne de la base KITTI

Dans cette section, nous comparons les performances de l'algorithme proposé par rapport aux techniques actuelles de pointe. Le tableau 3.3 présente les résultats obtenus à partir de la plateforme d'évaluation en ligne KITTI. Le rang de l'algorithme proposé est (68ème), la position obtenue est prometteuse, cependant des considérations supplémentaires sont nécessaires pour améliorer les performances de notre algorithme. Dans l'ensemble, le schéma hiérarchique proposé améliore les performances de l'algorithme conventionnel CrossCensus [76], il surpasse également certains algorithmes locaux bien connus tels que ELAS [14], GF (Census) [27].

3.4.4 Comparaison avec d'autres algorithmes hiérarchiques

3.4.5 Complexité algorithmique

Dans cette section, nous présentons une synthèse de l'efficacité de calcul de l'algorithme proposé. Ce dernier a été implémenté en code C++ sur un ordinateur de bureau équipé d'un CPU Intel Core i5 à 3.0GHz et d'une mémoire de 4GB. Considérons la complexité de calcul de l'agrégation de coûts croisée [76] comme $O(S_I D N_s)$, où S_I présente la taille de l'image d'entrée I, D les niveaux de disparité, s une échelle pyramidale et N_s le nombre de pixels dans la région de support adaptatif. La complexité théorique de l'algorithme hiérarchique est décrite comme suit :

$$\begin{aligned} &O(S_I D / 8^0 N_0) + O(S_I D / 8^1 N_1) + O(S_I D / 8^2 N_2) + \dots \\ &+ O(S_I D / 8^S N_S) \simeq 1 + \sum_{s=1}^S \frac{1}{8^s} O(S_I D N_0) \end{aligned} \quad (3.14)$$

Le nombre de pixels et la plage de disparité à une échelle s sont respectivement $[\frac{S_I}{4^s}]$, $[\frac{D}{2^s}]$. Ainsi, la complexité de calcul de l'algorithme hiérarchique augmente d'une petite constante par rapport à la méthode d'agrégation classique. En fixant S à 4 par exemple, on obtient une augmentation d'un facteur de 0,35.

Le temps de calcul moyen de la méthode proposée, ainsi que des autres méthodes comparées, est indiqué dans le tableau 3.4. Le temps d'exécution de la méthode LAMC-DSM inclut les étapes de post-traitement. Une implémentation CPU de l'algorithme proposé a conduit à un temps de calcul moyen de 44s pour les ensembles de données

TABLEAU 3.3 – Les résultats obtenus par la plateforme d'évaluation KITTI pour le seuil par défaut de 3 pixels. Les colonnes de gauche à droite : méthode ; pourcentage de pixels incorrects dans les régions non-occultées et au total ; erreur moyenne de disparité dans les régions non-occultées ; erreur moyenne de disparité au total. Date de l'évaluation : 06 mai 2017

Méthode	Out-Noc	Out-All	Avg-Noc	Avg-All
SGM	5.76 %	7.00 %	1.2 px	1.3 px
mSGM-LDE	6.01 %	8.22 %	1.4 px	2.4 px
Toast2	6.16 %	7.42 %	1.2 px	1.4 px
ITGV	6.20 %	7.30 %	1.3 px	1.5 px
TFS	6.28 %	6.59 %	1.4 px	1.5 px
WlinBPM	6.43 %	8.60 %	1.2 px	2.0 px
ELSEP	6.84 %	8.02 %	1.3 px	1.5 px
OCV-SGBM	7.64 %	9.13 %	1.8 px	2.0 px
SSMW	7.83 %	8.95 %	1.6 px	1.8 px
MSMW	8.01 %	9.24 %	1.6 px	1.7 px
Notre méthode	8.15 %	10.33 %	1.9 px	2.9 px
ELAS	8.24 %	9.96 %	1.4 px	1.6 px
linBP	8.56 %	10.70 %	1.7 px	2.7 px
GSS	8.69 %	10.82 %	2.4 px	3.5 px
ADSM	8.71 %	10.05 %	2.1 px	2.7 px
Deep-Raw	8.93 %	11.07 %	3.9 px	4.9 px
S+GF (Cen)	9.03 %	11.21 %	2.1 px	3.4 px
CrossCensus	9.46 %	10.86 %	2.3 px	2.7 px
SymST-GP	9.79 %	11.66 %	2.5 px	3.3 px
SM_GPTM	9.79 %	11.38 %	2.1 px	2.6 px
LAMC-DSM	9.82 %	11.49 %	2.1 px	2.7 px
IIW	10.78 %	12.62 %	3.3 px	4.3 px
SDM	10.95 %	12.14 %	2.0 px	2.3 px
HLSC_mesh	11.22 %	12.82 %	2.3 px	2.9 px
GF (Census)	11.65 %	13.76 %	4.5 px	5.6 px

TABLEAU 3.4 – Temps d'exécution moyen (en s) nécessaire au calcul des cartes de disparité pour les paires d'images KITTI

Méthodes	LAMC-DSM [59]	$S + GF$ [75]	Notre approche
Temps d'exécution	684(s)	140(s)	44(s)

KITTI, qui est le plus faible comparé aux autres méthodes de traitement hiérarchique.

3.5 Conclusion

Dans la littérature, les fonctions de coût proposées ne traitent que les pixels voisins. Cependant, l'information relative aux bords est négligée dans la phase de calcul des coûts. C'est la raison pour laquelle, dans le présent chapitre, une nouvelle variante de la fonction de coût de Census basée sur les contours est proposée. Cette nouvelle variante considère l'information sur les contours, ce qui fournit une structure complète où la position, concernant les contours, et l'ordre relatif de l'intensité du pixel dans une fenêtre de support, sont conjointement incorporés. Cela permet d'améliorer considérablement la capacité de la fonction de coût à traiter à la fois les changements radiométriques et le bruit de l'image. Cela améliore considérablement la capacité de la fonction de coût à gérer à la fois les changements radiométriques et le bruit de l'image. Les résultats présentés dans le tableau 3.1 montrent que la fonction de coût proposée surpasse les autres fonctions de coût basées sur la tomographie par ordinateur en termes de résultats de disparité. La fonction de coût proposée a atteint un taux de 15,50 pour les régions non occultées alors que les fonctions de coût C_{CT} , C_{CCC} et C_{Cstent} ont obtenu respectivement 18,27, 17,78 et 20,15. La région de support adaptative a été utilisée dans l'étape d'agrégation, en raison de son temps de calcul rapide et de sa précision. La dernière étape a été mise en œuvre dans le schéma de traitement hiérarchique, qui a été introduit en incorporant les informations de basse fréquence dans celles de haute fréquence à l'aide d'une fonction exponentielle robuste. Cette nouvelle approche s'est avérée prometteuse et a permis d'améliorer considérablement les résultats de la disparité. Cette amélioration est particulièrement notable dans les régions sans texture, comme le montre le tableau 2, où le pourcentage d'erreur pour les régions sans occlusion passe de 15,50 à 8,74 avec le schéma hiérarchique proposé au lieu de la méthode originale. La complexité de calcul du schéma hiérarchique est supérieure d'un petit facteur constant, par rapport à la méthode conventionnelle d'agrégation adaptative des coûts.

COÛT BASÉ SUR LES INFORMATIONS STRUCTURELLES POUR LES ALGORITHMES DE CORRESPONDANCE DES SCÈNES ROUTIÈRES

Sommaire

4.1	Introduction	49
4.2	Fonctions de coût proposées	50
4.2.1	Fonction de coût basée sur le SSIM (C_{SSIM})	50
4.2.2	SSIM gradient variant (C_{GSSIM})	51
4.3	Les résultats expérimentaux	52
4.3.1	Évaluation de la capacité discriminatoire des coûts proposés .	52
4.3.2	Évaluation des coûts proposés avec une technique d'agrégation adaptative	55
4.3.3	Sensibilité des fonctions de coût face aux distorsions radiomé- triques	56
4.4	Discussion	56
4.5	Conclusion	58

4.1 Introduction

La perception visuelle humaine utilise des informations structurelles pour reconnaître les correspondances stéréo dans les scènes naturelles. L'information structurelle peut donc s'avérer importante pour élaborer un algorithme de mise en correspondance robuste. Dans ce chapitre, nous démontrons que la prise en compte de la similarité de l'information structurelle, extraite soit directement à partir des valeurs d'intensité de l'image (SSIM), soit à partir des valeurs de gradient de l'image (GSSIM), qui existe entre deux zones, permet notamment de décrire précisément les caractéristiques structurelles de celles-ci et de fournir des valeurs de coût initial plus fiables. L'un des principaux phénomènes rencontrés dans la correspondance stéréo pour les scènes réelles, les changements radiométriques, est efficacement traité dans ce chapitre. La performance des fonctions de coût proposées a été évaluée en deux étapes : Dans la première, ces coûts sont étudiés sans processus d'agrégation, et dans la seconde, la technique d'agrégation adaptative rapide est employée. Les expérimentations ont été réalisées sur des scènes réelles de trafic routier, avec les base de données KITTI 2012 et KITTI 2015. Les résultats obtenus ont démontré les avantages potentiels de la mesure de similarité proposée.

Le reste du chapitre est organisé comme suit : Dans la section 2, nous présentons la fonction de coût proposée. Les résultats expérimentaux et la discussion sont donnés dans la section 3. Enfin, nous tirons des conclusions dans la section 4.

4.2 Fonctions de coût proposées

4.2.1 Fonction de coût basée sur le SSIM (C_{SSIM})

Le problème de la correspondance stéréo est considéré comme un problème visuel. L'extraction des informations les plus importantes capturées par le système visuel humain (HVS) peut fournir des informations pertinentes afin de décrire précisément la zone considérée, et ainsi faciliter le processus de correspondance. Dans ce contexte, nous proposons une nouvelle fonction de coût basée sur l'information structurelle [24]. Plus précisément, soit $p(x, y)$ un pixel de l'image de référence (I_1), I_p la valeur d'intensité du pixel p et $q(x, y - d)$ son homologue théorique, avec la valeur d'intensité I_q dans l'image cible (I_2) à une disparité d . Le C_{SSIM} entre p et q est défini par la formule suivante :

$$C_{SSIM}(p, q, d) = [l(p, q, d)]^\alpha \cdot [c(p, q, d)]^\beta \cdot [s(p, q, d)]^\gamma \quad (4.1)$$

Où, $l(p, q, d)$ est la luminance, $c(p, q, d)$ est le contraste et $s(p, q, d)$ sont les mesures de structure entre $\mathbf{p}$ et $\mathbf{q}$, définies dans eq(2), eq(3) et eq(4), respectivement.

$$l(p, q, d) = \frac{2\mu_p\mu_q + C}{\mu_p^2 + \mu_q^2 + C} \quad (4.2)$$

$$c(p, q, d) = \frac{2\sigma_p\sigma_q + C}{\sigma_p^2 + \sigma_q^2 + C} \quad (4.3)$$

$$s(p, q, d) = \frac{\sigma_{(p,q)} + C}{\sigma_p\sigma_q + C} \quad (4.4)$$

C est une faible constante permettant d'éviter que la valeur du dénominateur soit nulle. μ_p et μ_q sont les valeurs moyennes calculées dans le voisinage N_p et N_q de p dans I_1 et q dans I_2 , respectivement.

σ_p et σ_q sont les écarts types de p et q , respectivement. Les écarts types de p dans la fenêtre de support N_p sont décrits comme suit :

$$\sigma_p = \left(\frac{1}{(|N_p| - 1)} \sum_{p' \in N_p} (I_{p'} - \mu_p)^2 \right)^{1/2} \quad (4.5)$$

Où $\|N_p\|$ est le nombre de pixels dans la fenêtre de support N_p . Le $\sigma_{(p,q)}$ est la covariance entre p et q , et peut être estimé comme suit :

$$\sigma_{(p,q)} = \frac{1}{(\|N_p\| - 1)} \sum_{p' \in N_p, q' \in N_q} (I_{p'} - \mu_p)(I_{q'} - \mu_q) \quad (4.6)$$

Enfin, $\alpha > 0$, $\beta > 0$ et $\gamma > 0$ sont des paramètres permettant de contrôler l'influence de chacune des trois composantes.

4.2.2 SSIM gradient variant (C_{GSSIM})

En complément de l'information structurelle, le système visuel humain est capable d'extraire les caractéristiques structurelles basées sur les gradients de l'image, telles que (les contours et les points). Ainsi, afin de prendre en compte cette hypothèse, l'information structurelle est extraite des dérivées de l'image $\partial I / \partial x$, $\partial I / \partial y$ plutôt que des intensités de l'image. Pour cela, la luminance (l), le contraste (c) et la mesure de la structure (s), dans l'équation (1) seront modifiés en intégrant le gradient. Par conséquent, la fonction de coût de l'information structurelle basée sur le gradient C_{GSSIM} est définie comme suit :

$$C_{GSSIM}(p, q, d) = [l_g(p, q, d)]^\alpha \cdot [c_g(p, q, d)]^\beta \cdot [s_g(p, q, d)]^\gamma \quad (4.7)$$

où l_g , c_g et s_g sont des informations structurelles définies comme suit :

$$l_g(p, q, d) = \sum_{p \in \{\frac{\partial I_1}{\partial x}, \frac{\partial I_1}{\partial y}\}, q \in \{\frac{\partial I_2}{\partial x}, \frac{\partial I_2}{\partial y}\}} \frac{2\partial\mu_p\partial\mu_q + C}{\partial\mu_p^2 + \partial\mu_q^2 + C} \quad (4.8)$$

$$c_g(p, q, d) = \sum_{p \in \{\frac{\partial I_1}{\partial x}, \frac{\partial I_1}{\partial y}\}, q \in \{\frac{\partial I_2}{\partial x}, \frac{\partial I_2}{\partial y}\}} \frac{2\partial\sigma_p\partial\sigma_q + C}{\partial\sigma_p^2 + \partial\sigma_q^2 + C} \quad (4.9)$$

$$s_g(p, q, d) = \sum_{p \in \{\frac{\partial I_1}{\partial x}, \frac{\partial I_1}{\partial y}\}, q \in \{\frac{\partial I_2}{\partial x}, \frac{\partial I_2}{\partial y}\}} \frac{\partial\sigma_{(p,q)} + C}{\partial\sigma_p\sigma_q + C} \quad (4.10)$$

$\partial\mu_p$ et $\partial\mu_q$ sont les valeurs moyennes calculées pour le voisinage ∂N_p dans ∂I_1 pour p et ∂N_q dans ∂I_2 pour q et q dans ∂I_2 . ∂I_1 et ∂I_2 sont les gradients selon les directions x et y , respectivement. $\partial\sigma_p$ et $\partial\sigma_q$ sont les écarts types de p dans ∂I_1 et de q dans ∂I_2 . L'écart-type de p $\partial\sigma_p$ est défini ainsi :

$$\partial\sigma_p = \sum_{p \in \{\frac{\partial I_1}{\partial x}, \frac{\partial I_1}{\partial y}\}} \left(\frac{1}{(\|N_p\| - 1)} \sum_{p' \in \partial N_p} (\partial I_{p'} - \partial\mu_p)^2 \right)^{1/2} \quad (4.11)$$

$\partial\sigma_{(p,q)}$ est défini comme suit :

$$\begin{aligned} \partial\sigma_{(p,q)} = & \sum_{p \in \{\frac{\partial I_1}{\partial x}, \frac{\partial I_1}{\partial y}\}} \frac{1}{(||N_p|| - 1)} \\ & * \sum_{p' \in \partial N_p, q' \in \partial N_q} (\partial I_{p'} - \partial\mu_p)(\partial I_{q'} - \partial\mu_q) \end{aligned} \quad (4.12)$$

A la différence de l'équation (1), cette formule permet de calculer les nouvelles caractéristiques structurales sur les dérivées principales de l'image par rapport aux coordonnées x et y .

4.3 Les résultats expérimentaux

Dans cette section, nous évaluons la capacité des fonctions de coût proposées à discriminer les correspondances stéréo. Nous étudions les coûts proposés pour la correspondance stéréo au moyen de deux algorithmes différents : un algorithme de correspondance stéréo sans phase d'agrégation ainsi qu'une technique d'agrégation locale adaptative rapide. Ces fonctions de coût sont ensuite comparées aux fonctions de coût les plus importantes : $C_{DIFFCensus}$, et C_{GCCC} [74]. Les valeurs optimales des paramètres proposées dans [49, 74] ont été retenues. Des expériences ont été menées sur les bases de données d'entraînement KITTI 2012 [13] et KITTI 2015 [46], afin d'évaluer l'approche proposée dans le contexte des applications de voitures intelligentes.

L'évaluation pour les jeux de données KITTI 2012 est effectuée en calculant le pourcentage d'erreurs de disparité par rapport à la vérité terrain. Tandis que, pour les KITTI 2015, la mesure d'erreur $D1 - all$ est calculée, elle représente le pourcentage de pixels pour lesquels l'erreur d'estimation est supérieure à 3 pixels et supérieure à 5% de la disparité de la vérité terrain pour chaque pixel. Les ensembles de paramètres des deux fonctions de coût, C_{SSIM} et C_{GSSIM} , ont été définis expérimentalement par les valeurs suivantes : $\alpha = 0,9$, $\beta = 0,1$ et $\gamma = 0,2$ pour minimiser le taux d'erreur global. Le paramètre C est fixé à une valeur minimale pour éviter la division par zéro. Dans la phase d'agrégation, les seuils de similarité spatiale et de couleur ont été fixés à $L = 9$ et $\tau = 20$, respectivement.

La stratégie locale *WTA* a été adoptée pour générer les résultats de disparité. Nous avons utilisé le coût de mise en correspondance le plus grand, au lieu du plus petit, puisque les coûts proposés sont basés sur une mesure de la similarité.

4.3.1 Évaluation de la capacité discriminatoire des coûts proposés

Dans cette section, l'efficacité des fonctions de coût proposées est étudiée sur les deux ensembles de données KITTI sans utiliser de méthode d'agrégation des coûts.

La figure 1 présente une représentation des résultats de disparité pour chaque fonc-

TABLEAU 4.1 – Pourcentage de disparités incorrectes de la correspondance stéréoscopique sans l'utilisation d'une méthode d'agrégation pour la base d'entraînement KITTI 2012

	Seuil de 3 pixel	
	<i>Non – occ</i>	<i>toutes</i>
$C_{DIFFCensus}$ [49]	54.25	55.30
C_{GCCC} [74]	30.78	32.34
CSSIM	20.94	22.73
C_{GSSIM}	18.00	19.86

TABLEAU 4.2 – les résultats sur les bases de données de formation KITTI 2015

	sans méthode d'agrégation		a vec méthode d'agrégation	
	D1–all (Non-occulté)	D1–all(Toutes)	D1–all (Non-occultés)	D1–all (Toutes)
$C_{DIFFCensus}$ [49]	50.74	49.87	18.63	20.05
C_{GCCC} [74]	27.52	28.76	14.63	16.05
CSSIM	15.23	16.69	11.08	13.06
C_{GSSIM}	15.38	16.83	10.06	12.07

tion de coût utilisant les deux algorithmes stéréoscopiques. Les colonnes un et trois présentent les résultats obtenus sans recours à une méthode d'agrégation, tandis que les colonnes deux et quatre présentent les résultats obtenus avec une méthode d'agrégation adaptative. Les résultats de sortie pour deux paires #0 (gauche) et #7 (droite) de la base de données d'entraînement KITTI 2012 sont présentés. La figure 4.1 indique que, dans les deux cas, les fonctions de coût proposées donnent des résultats intéressants, contrairement aux coûts courants qui fournissent des résultats de disparité très bruités.

Les tableaux 4.1 et 4.2 présentent le taux d'erreur moyen pour toutes les paires stéréoscopiques dans différentes régions (non–occultés et toutes les régions) sur les deux bases de données, KITTI 2012 et KITTI 2015, respectivement. Les taux d'erreur les plus faibles ont été obtenus par les fonctions de coût proposées sur les deux ensembles de données dans différentes régions (non–occultés et toutes les régions). Selon ces résultats, les deux fonctions de coût proposées atteignent des résultats remarquables. En particulier, la fonction C_{GSSIM} fournit le plus faible taux d'erreur dans les deux ensembles de données.

Les résultats présentés montrent la capacité discriminatoire des coûts proposés indépendamment des coûts d'agrégation, ce qui prouve l'efficacité de l'information $SSIM$ pour capturer des informations locales fiables pour la mise en correspondance stéréoscopique. La section suivante étudie l'efficacité de ces coûts en utilisant des techniques d'agrégation.

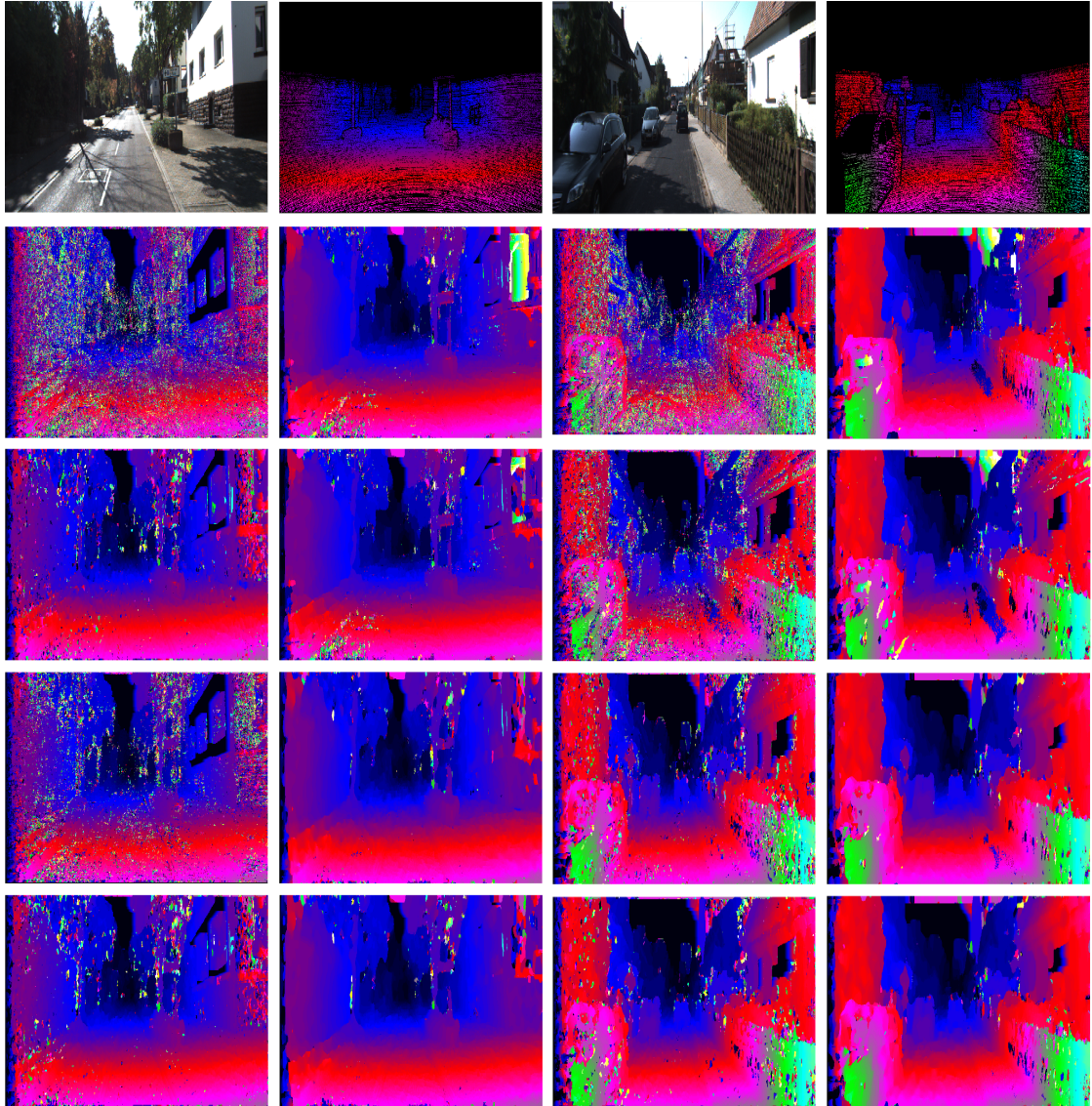


FIGURE 4.1 – Résultats de disparité. La première colonne correspond à deux paires stéréoscopiques à gauche (#0 et #7) de la base de données KITTI 2012 avec les cartes de disparité de la vérité du terrain correspondantes. Les cartes de disparité calculées sont répertoriées dans les lignes suivantes, correspondant aux fonctions de coût $C_{DiffCensus}$, C_{GCCC} , C_{SSIM} et C_{GSSIM} , respectivement. La première et la troisième colonne correspondent à la solution obtenue sans l'utilisation d'une méthode d'agrégation, tandis que la deuxième et la quatrième colonne correspondent à la solution basée sur la méthode d'agrégation adaptative.

TABLEAU 4.3 – Pourcentage de disparités incorrectes dans les régions non occultées et les autres régions pour l'ensemble de formation KITTI 2012

	seuil 2 px		seuil 3 px		seuil 5 px	
	non-occultés	Toutes	non-occultés	Toutes	non-occultés	Toutes
$C_{DIFFCensus}$ [49]	20.08	21.88	12.97	14.91	9.70	11.63
C_{GCCC} [74]	20.47	22.27	11.93	13.89	10.19	12.18
C_{SSIM}	16.39	18.29	11.08	13.06	8.18	10.17
C_{GSSIM}	14.06	16.00	10.06	12.07	7.83	9.83

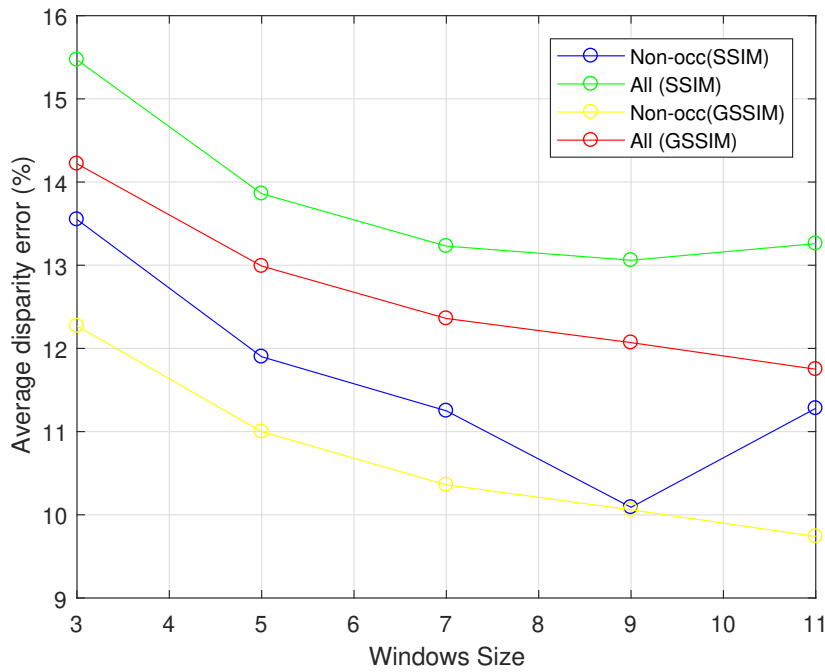


FIGURE 4.2 – Les résultats de disparité obtenus par rapport à la taille de la fenêtre de support pour les deux fonctions de coût, pour les ensembles de formation KITTI 2012.

4.3.2 Évaluation des coûts proposés avec une technique d'agrégation adaptative

Afin de réduire davantage le bruit et de construire des fonctions de coût plus raffinées, la méthode d'agrégation adaptative [76] a été appliquée. Ce choix est motivé par le fait que cette méthode est rapide et précise, ce qui est adapté aux applications en temps réel. L'efficacité de notre méthode a d'abord été évaluée en fonction de la taille de la fenêtre de support sur les bases de données d'entraînement KITTI 2012. La figure 4.2 présente le taux d'erreur moyen, dans les régions non occultées et toutes les régions, calculé au seuil par défaut de 3 pixels pour toutes les images de l'ensemble d'entraînement. On

peut constater que la taille de la fenêtre de support a un impact important sur les performances de l'algorithme pour les deux fonctions de coût. En effet, une amélioration significative des performances de l'algorithme de correspondance locale peut être obtenue lorsque la taille de la fenêtre de support augmente. Plus précisément, l'amélioration est d'un facteur de 1,65% pour les zones non occultées et de 1,61% pour les zones occultées, pour la fonction de coût C_{SSIM} lorsque la taille de la fenêtre passe de 3 à 5 par exemple.

Dans la suite, nous évaluons la fiabilité des fonctions de coût proposées basées sur la méthode d'agrégation adaptative par rapport à la fonction de coût de l'état de l'art.

Les tableaux 4.2 et 5.2 présentent le pourcentage moyen de pixels incorrects dans les régions non-occultées et dans toutes les régions. Dans le tableau 5.2, les erreurs ont été calculées à 3 seuils d'erreur différents, tandis que dans le tableau 4.2, seule l'erreur D1-all a été calculée. Les résultats obtenus indiquent que les fonctions de coût C_{GSSIM} proposées surpassent les autres par une marge significative. En effet, la fonction C_{GSSIM} fournit les erreurs de disparité moyennes les plus faibles sur les deux ensembles de données, suivie par la fonction de coût C_{SSIM} proposée dans différents scénarios. En effet, dans le tableau 5.2 au seuil de 3 pixels par défaut, l'amélioration obtenue par C_{GSSIM} est de l'ordre de 2,23, 3,47 pour la région non-occluse et de 2,84, 3,4 pour les autres zones, par rapport aux coûts $C_{DIFFCensus}$ et C_{GCCC} . De plus, à partir du tableau 4.2 nous pouvons voir clairement que les performances de nos méthodes sont significativement meilleures que toutes les autres fonctions de coût dans les deux régions. Par exemple, l'amélioration obtenue par C_{GSSIM} est de l'ordre de 1.87, 2.91 pour la région non-occluse et de 1.82, 2.84 pour les autres zones, par rapport aux coûts C_{GCCC} et $C_{DIFFCensus}$.

Cette évaluation montre que la fonction de coût C_{GSSIM} proposée est plus appropriée pour le calcul de la disparité réelle en extérieur que les plus performants $C_{DiffCensus}$ et C_{GCCC} .

4.3.3 Sensibilité des fonctions de coût face aux distorsions radiométriques

Dans cette section, nous étudions l'impact des distorsions radiométriques sur différentes fonctions de coût. Ces distorsions sont générées en utilisant la différence absolue de couleur entre les pixels correspondants [49]. En effet, à chaque niveau de distorsion radiométrique, nous calculons les erreurs moyennes de disparité pour tous les ensembles d'entraînement KITTI pour les fonctions de coût C_{SSIM} , C_{GSSIM} , $C_{DIFFCensus}$ [49] et C_{GCCC} [74]. On peut visualiser sur la figure 4.3 que le coût proposé C_{GSSIM} donne le plus faible taux d'erreur à tous les niveaux de distorsion radiométrique.

4.4 Discussion

Dans la littérature, il a été démontré que les fonctions de coût basées sur l'intensité des pixels sont très sensibles aux changements radiométriques. Dans cet article, de nouvelles fonctions de coût basées sur l'intensité ont été proposées. Elles prennent en compte l'intensité, la luminance et le contraste locaux, qui fournissent une information locale si-

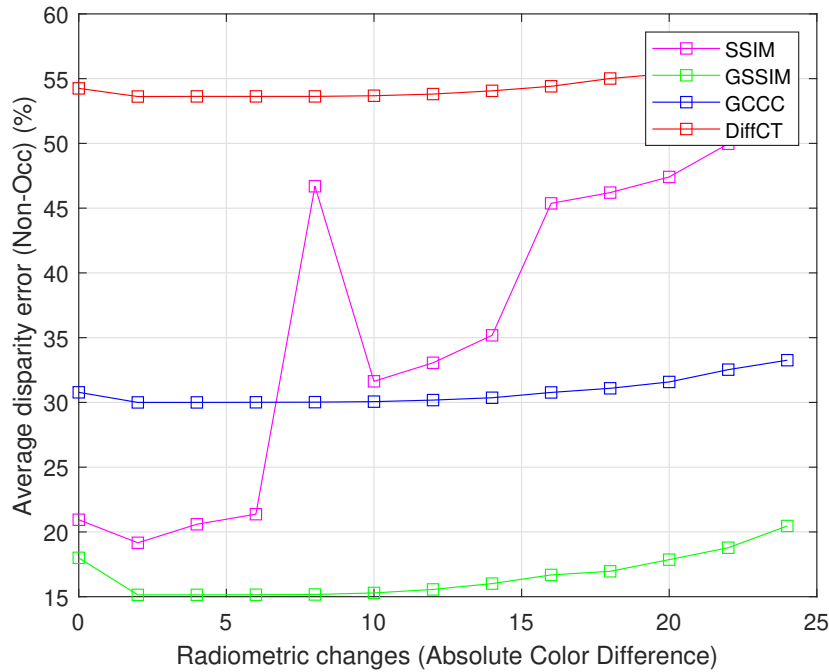


FIGURE 4.3 – Les résultats de disparité obtenus par rapport aux distorsions radiométriques pour les fonctions de coût présentées, dans le cas de la base de donnée KITTI 2012.

gnificative pour décrire le pixel considéré dans une fenêtre de support. Cette nouvelle considération donne à la fonction de coût proposée la capacité de traiter les changements radiométriques (voir figure 4.3). Les résultats décrits dans les tableaux 2 et 3 montrent que les fonctions de coût proposées donnent de meilleurs résultats que les fonctions de coût les plus performantes dans les deux ensembles de données KITTI 2012 et KITTI 2015, par rapport aux fonctions de coût $C_{DiffCensus}$ et C_{GCCC} . Bien que ces derniers promettent de meilleurs résultats avec les techniques d'agrégation, les coûts d'agrégation proposés ont conduit aux meilleurs résultats (voir tableau 5.2 et 4.2). Il faut noter que la performance globale des fonctions de coût proposées dépend de la taille de la région de support, comme le montre la figure 4.2, on constate que les deux fonctions de coût sont performantes lorsque la taille de la région de support augmente. Ceci est trivial puisque les grandes régions de support détiennent suffisamment d'informations pour décrire plus précisément le patch considéré, et conduisent alors à de bonnes fonctions de coût initiales précises.

4.5 Conclusion

Dans ce chapitre, nous avons présenté un nouvel algorithme d'appariement stéréo avec une nouvelle fonction de coût basée sur l'information structurelle pour l'étape de calcul des coûts. Ainsi, deux fonctions de coût ont été proposées et évaluées en utilisant des scènes de route réelles provenant des ensembles de données d'entraînement KITTI 2012 et KITTI 2015. Les résultats obtenus ont démontré que les deux fonctions de coût conduisent aux erreurs moyennes de disparité les plus faibles par rapport à la fonction la plus performante dans cet ensemble de données sous différents scénarios, ce qui a prouvé que nos fonctions de coût sont plus robustes aux distorsions radiométriques que les fonctions de coût conventionnelles. L'évaluation de l'algorithme d'appariement stéréo local proposé utilisant la fonction de coût la plus performante par rapport aux algorithmes de pointe actuels a démontré les mérites potentiels de notre mesure de similarité stéréoscopique.

GÉNÉRATION DE CARTES DE DISPARITÉ DENSES À PARTIR DES MESURES ÉPARSES

Sommaire

5.1	Introduction	59
5.2	Méthode proposée	60
5.2.1	Échantillonnage de données	60
5.2.2	Méthode de propagation	61
5.2.3	Filtre médian vertical adaptatif	62
5.3	Résultats expérimentaux	63
5.3.1	Résultats de reconstruction en utilisant les mesures de la carte de disparité vérité du terrain	63
5.3.2	Résultats de reconstruction en utilisant les autres cartes de disparités	67
5.4	Conclusion	68

5.1 Introduction

Ce chapitre traite de la reconstruction de la carte des disparités à partir des disparités déjà connues en des points bien répartis. C'est dans ce contexte que nous présenterons notre contribution. L'algorithme proposé offre une reconstruction dense de l'ensemble des disparités calculées pour une image sans aucune perte d'information et en présentant des résultats équivalents aux disparités initiales. Tout d'abord, une méthode de segmentation des super-pixels avec la méthode SLIC [1] (Simple Linear Iterative Clustering) est appliquée. Cette segmentation permet de générer la disparité sur les pixels appartenant aux contours de chaque région, plutôt que sur toute l'image. Puis la reconstruction complète de la carte est effectuée en utilisant une méthode de propagation linéaire. Enfin, une méthode de post-traitement est utilisée pour éliminer les aberrations produites par la propagation linéaire. Les expériences réalisées sur la nouvelle base de données Middlebury montrent clairement l'efficacité de notre approche par rapport aux méthodes alternatives.

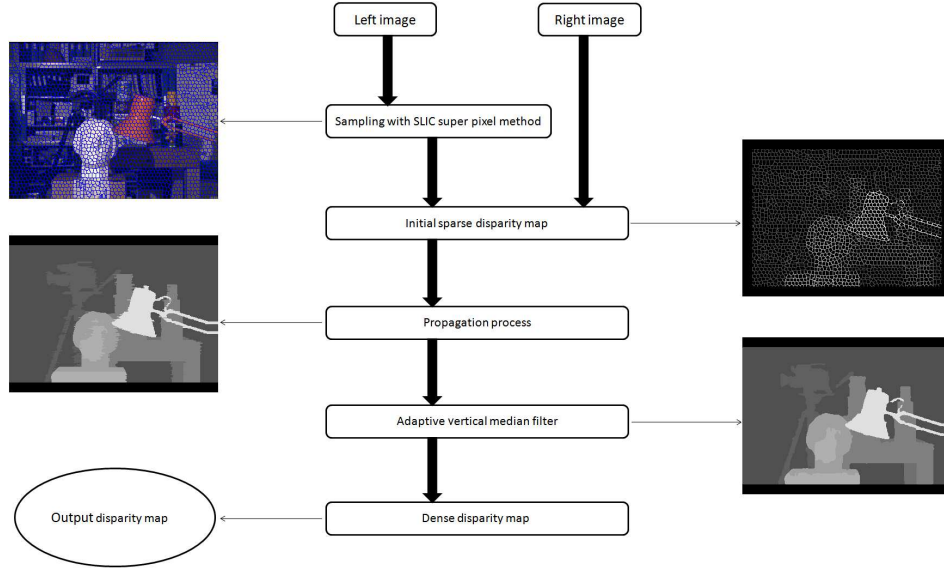


FIGURE 5.1 – Diagramme de la méthode proposée en utilisant comme exemple l’image stéréo ”Tsukuba”

5.2 Méthode proposée

Notre algorithme calcule la carte de disparité conformément au plan général illustré dans la figure 5.1. Nous détaillons chaque étape dans les sections suivantes.

5.2.1 Échantillonnage de données

Les méthodes de superpixels ont été largement utilisées dans les algorithmes de mise en correspondance stéréo afin de fournir un lissage préalable associé à l’appartenance de tous les autres pixels à la même surface 3D [68], [69], [35]. Contrairement à ces méthodes, notre approche considère les pixels frontières de chaque région de superpixels comme des points sources. Ces points de départ sont transmis aux régions de superpixels par le biais des lignes de scan en fonction des similarités de couleur.

En effet, l’image de référence est préalablement traitée comme un ensemble de superpixels quasi réguliers en utilisant la méthode SLIC. Les superpixels obtenus se composent donc de pixels ayant des attributs similaires, ce qui permet de préserver les contours de l’image. Nous nous concentrons donc sur les pixels de frontière situés sur les segments de frontière obtenus par le processus de segmentation. La carte de disparité éparse peut être obtenue à l’aide de tout algorithme de correspondance stéréo. Cependant, une valeur de disparité initiale incorrecte conduit à des erreurs de disparité considérables au cours du processus de propagation. L’étude de la précision des mesures de disparité initiales ne faisant pas l’objet de ce chapitre, nous générons les mesures de disparité initiales en utilisant des mesures de disparité optimales tirées des cartes de disparité de la vérité du terrain. De plus, pour mieux évaluer notre méthode, nous avons également utilisé d’autres

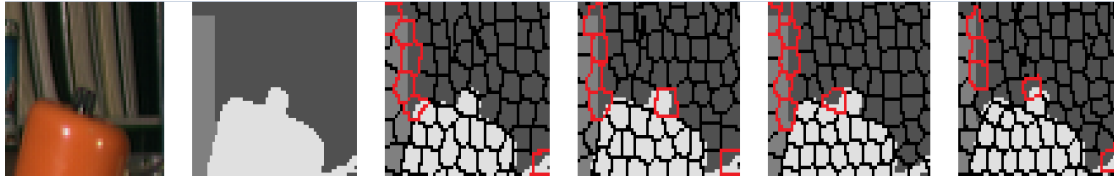


FIGURE 5.2 – Exemple de vue rapprochée de la segmentation en super pixel de la paire stéréo Tsukuba, De gauche à droite : Image de référence, La vérité de terrain, graphe de superpixel sur la carte de disparité avec 2000, 2200, 2400 and 2600 nombre de super pixels, respectivement.

mesures de disparité initiales, comme la méthode de correspondance semi-globale (SGM) [22].

5.2.2 Méthode de propagation

Les pixels d'une même région de superpixels sont supposés faire partie du même objet 3D étudié, à condition que la taille de ce superpixel reste suffisamment réduite [70]. Cependant, cette hypothèse risque d'être violée, notamment dans les régions proches des discontinuités de profondeur. La figure 5.2 présente un cas typique de la paire stéréoscopique de Tsukuba, dans lequel les pixels au sein du même superpixel (régions colorées) possèdent des informations de profondeur différentes, même si l'échelle du superpixel est très faible. En conséquence, il est inefficace d'attribuer la même valeur de disparité à tous les pixels d'un superpixel. Afin de résoudre ce problème, la valeur de disparité a été obtenue à partir du pixel frontière de chaque région de superpixel, puis la carte de disparité a été balayée par lignes, chaque fois que nous rencontrons un pixel non analysé p , nous recherchons les pixels les plus proches pour lesquels la valeur de disparité est connue, et situés sur la même ligne de balayage du superpixel (ligne de pixels). Puis, si un seul pixel-semence est identifié, sa disparité est attribuée au pixel non traité p . Dans le cas contraire, s'il trouve des pixels-semences gauche et droit (p_{s_1}, p_{s_2}) avec des valeurs de disparité connues d_1 et d_2 , respectivement. Nous attribuons au pixel nonensemencé la disparité d en appliquant la règle suivante :

$$d = \begin{cases} d_1 & D_c(p_{s_1}, p) < D_c(p_{s_2}, p) \\ d_2 & otherwise \end{cases} \quad (5.1)$$

Où $D_c(p_{s_i}, p) = \max_{i=R,G,B} |I_i(p_{s_i}) - I_i(p)|$ représente la différence de couleur entre $p_{s_i}, i \in \{1, 2\}$ et p . Les cartes de disparité obtenues après le processus de propagation sont présentées dans la figure 5.3, ce qui illustre bien l'efficacité de la méthode proposée, puisqu'elle produit de légères erreurs. Afin de corriger les erreurs restantes, nous utilisons le filtre médian vertical, détaillé dans la prochaine section.

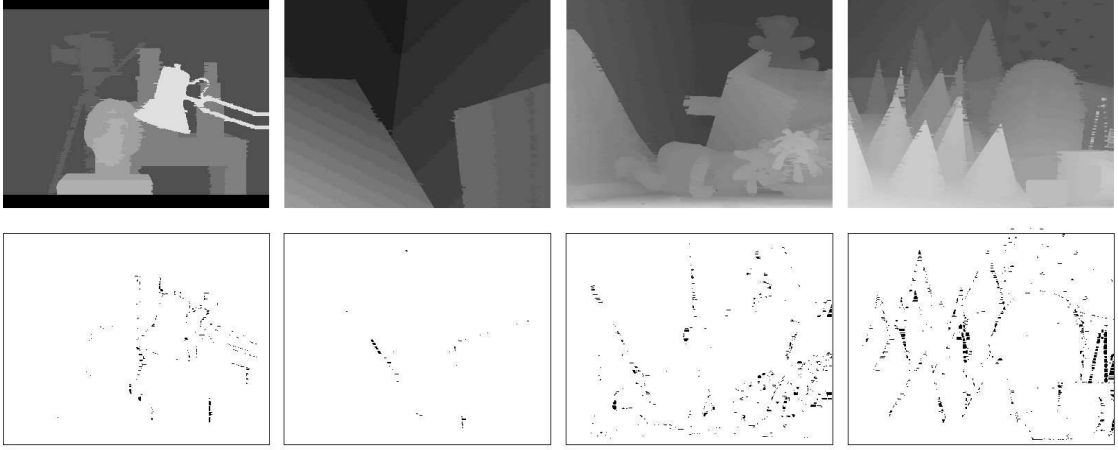


FIGURE 5.3 – Résultats du processus de propagation pour les quatre paires stéréoscopiques de Middlebury (Tsukuba, Venus, Teddy, Cones). Du haut vers le bas : Les cartes de disparité produites, les erreurs de disparité (marquées en noir) dans toutes les régions avec un seuil d'erreur d'un pixel.

5.2.3 Filtre médian vertical adaptatif

Puisque le processus de propagation ci-dessus est effectué en lignes, des artéfacts de type stries horizontales risquent de se produire dans la carte de disparité. Ainsi, l'inclusion d'informations de disparité dans la direction verticale peut les réduire efficacement. À cette fin, nous avons introduit l'information de disparité verticale en attribuant la valeur médiane verticale au pixel considéré. Les artéfacts de stries se situent généralement aux limites des objets. Ainsi, la limitation de la taille du filtre vertical permet de construire un filtre de support avec des valeurs aberrantes provenant d'autres structures de l'image, ce qui conduit à des résultats de disparité incorrects. Pour résoudre ce problème, nous avons opté pour la construction d'un filtre de support vertical adaptatif, contenant uniquement des pixels de la même structure d'image. Dans ce contexte, notre attention s'est portée sur le filtre médian croisé présenté dans [59], qui a été utilisé comme méthode de post-traitement pour traiter les valeurs aberrantes produites par la méthode de vérification de la cohérence gauche-droite. Nous avons également suivi l'hypothèse de la couleur afin de construire des filtres de support vertical. La ligne verticale de chaque pixel p , est définie à l'aide de ses deux axes directionnels (haut ou bas) selon les règles suivantes :

- (1) $D_c(p_l, p) < \tau$, où p_l est un pixel situé sur le bras vertical de p . $D_c(p_l, p) = \max_{i=R,G,B} |I_i(p_l) - I_i(p)|$, qui désigne la différence de couleur entre p_l et p , et τ est la valeur seuil prédéfinie.
- (2) $D_s(p_l, p) < L$, où $D_s(p_l, p) = |p_l - p|$, qui représente la distance spatiale entre p_l et p , et L est la longueur maximale prédéfinie.

La première règle garantit l'hypothèse de similarité des couleurs, tandis que la seconde impose une limite à la longueur de la ligne verticale pour éviter que les résultats de

TABLEAU 5.1 – Rapports d'échantillonnage pour chaque valeur de superpixel pour les paires stéréo Tsukuba, Venus, Teddy et Cones.

Nombre de superpixels	Rapports d'échantillonnage (%)			
	Tsukuba	Venus	Teddy	Cones
600	15.44	12.98	12.73	12.75
800	17.87	15.06	14.81	15.08
1000	19.67	16.53	16.38	16.57
1200	21.51	18.33	17.95	18.03
1400	22.93	19.86	19.13	19.21
1600	24.50	20.86	20.55	20.71
1800	25.80	21.82	21.71	21.82
2000	27.22	23.09	22.90	23.02
2200	28.39	23.92	23.67	23.84
2400	29.06	25.00	24.74	24.97
2600	30.37	25.60	25.88	26.01

disparité soient surlissés.

5.3 Résultats expérimentaux

Dans cette section, nous effectuons une évaluation pour mesurer les performances de l'algorithme stéréoscopique proposé. Les expériences ont été réalisées sur les bases de données Middlebury classique et récente. Les mesures de disparité éparses ont été extraites à partir de la vérité terrain et aussi d'autres disparités telles que les résultats de disparité SGM [22]. Dans la construction de la médiane verticale adaptative, les seuils de similarité spatiale et de couleur ont été fixés expérimentalement à $L = 7$ et $\tau = 8$, respectivement. Dans notre travail, le nombre de superpixels définit les ratios d'échantillonnage des données. Nous représentons d'abord les rapports d'échantillonnage des données correspondants pour les ensembles de données standard de Middlebury en fonction du nombre de superpixels dans le tableau 5.1.

5.3.1 Résultats de reconstruction en utilisant les mesures de la carte de disparité vérité du terrain

Nous examinons la précision de la reconstruction en fonction des différents nombres de superpixels en utilisant les disparités de la vérité du terrain. La figure 5.4 montre les erreurs de disparité obtenues avec différents nombres de superpixels pour les quatre paires

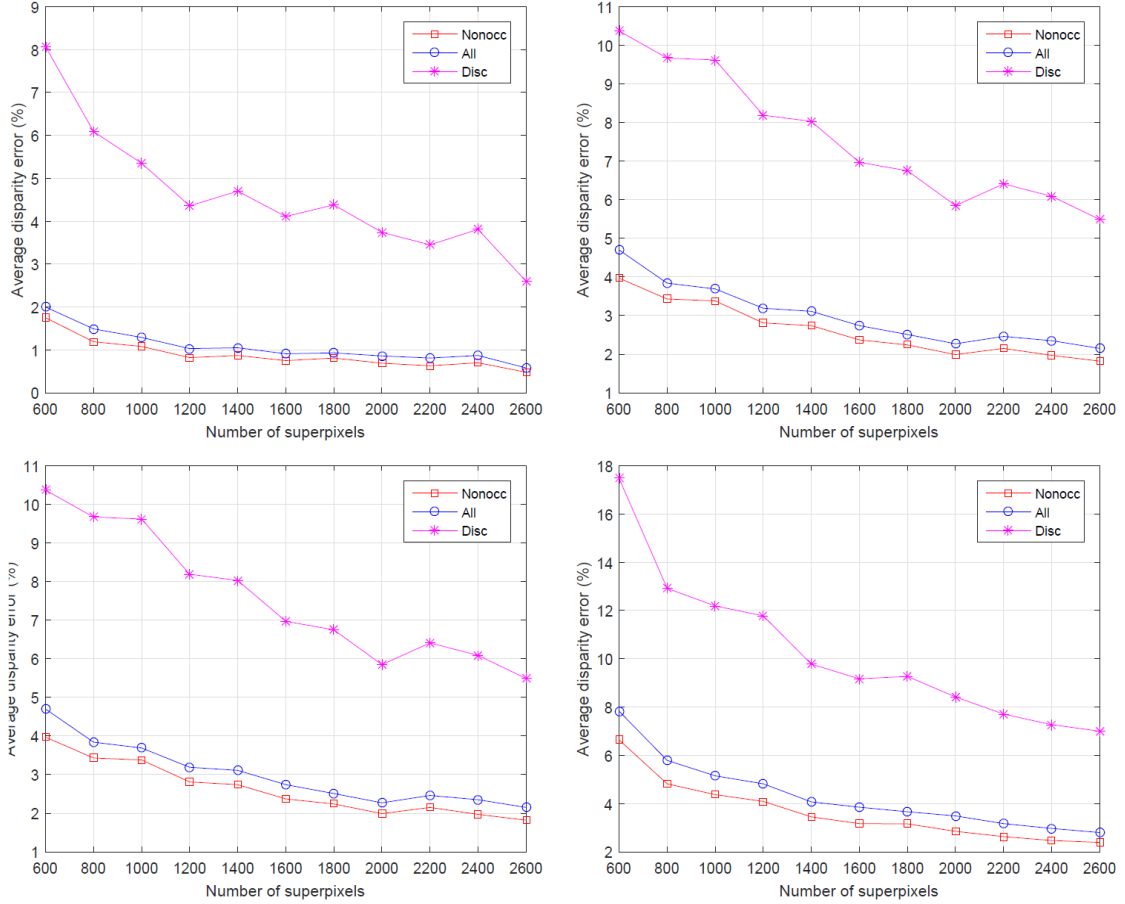


FIGURE 5.4 – Résultats de disparité obtenus par rapport au nombre de super pixels avec les images Tsukuba, Venus, Teddy et Cones

stéréo. Les erreurs sont données dans les régions non occluses, toutes les discontinuités et la profondeur proche, sont calculées à 1 seuil d'erreur de pixel par défaut. On peut noter que les performances de l'approche proposée augmentent avec le nombre de superpixels. Cela peut s'expliquer par le fait qu'un petit nombre de superpixels (c'est-à-dire une grande échelle de superpixels) peut contenir plusieurs objets ou une petite partie de ceux-ci. Ensuite, dans le processus de propagation, la même valeur de disparité peut être attribuée aux pixels de différents objets 3D.

nous avons également évalué notre algorithme contre certains des algorithmes de reconstruction de disparité de pointe. En effet, nous avons comparé notre méthode avec la méthode récemment introduite [39] en utilisant ses deux variantes Alternating Direction Method of Multipliers (ADMM) on Wavelet (WT) and Contourlet (CT), *ADMMWT+CTGrid* et *ADMMWT+CT2-Stage*, où Grid et 2-stage se réfèrent respectivement aux données utilisées pour l'échantillonnage et la méthode [39]. Pour mener à bien les expériences, nous avons utilisé les mêmes échantillons de données pour les deux

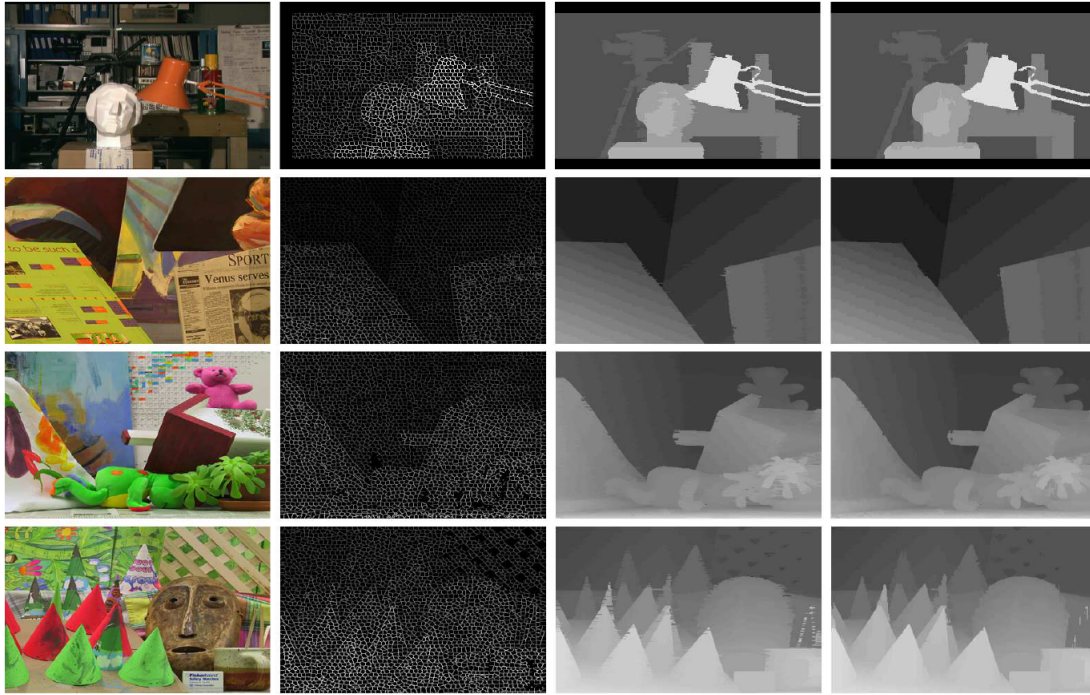


FIGURE 5.5 – Résultats de disparité obtenus par rapport aux nombre de super pixels avec les images Tsukuba, Venus, Teddy et Cones

méthodes [39] et la nôtre. Ainsi, nous avons utilisé 2600 superpixels qui représentent un taux d'échantillonnage de données de 30%, 25%, 25% et 26% pour les paires stéréo Tsukuba, Venus, Teddy et Cones, respectivement. Étant donné que pour la méthode [51], l'échantillonnage est effectué en utilisant la méthode des k moyennes, si elle est définie, nous gardons la même valeur du paramètre k [51]. Sinon, nous fixons $K = 12$, car cela donne expérimentalement les erreurs de disparité les plus faibles.

Le tableau 5.2 présente l'erreur de disparité dans les discontinuités non occultées (nonocc), toutes et à proximité de la profondeur. Les erreurs ont été calculées au seuil d'erreur par défaut de 1 pixel pour la base de données Middlebury. Les résultats obtenus démontrent l'efficacité de la méthode proposée. En effet, notre méthode permet une reconstruction en profondeur avec un taux d'erreur plus faible pour les paires stéréo Teddy et Cones dans toutes les régions. Dans le cas de la paire stéréo Venus, la méthode "ADMM WT + CT2- Stage" [39] et notre approche donnent presque exactement les mêmes taux d'erreur. Dans l'ensemble, des résultats de disparité satisfaisants ont été obtenus pour toutes les paires stéréo. De plus, en termes de temps d'exécution, la méthode proposée fonctionne mieux. En effet, le temps de calcul pour les quatre paires stéréo (Tsukuba, Venus, Teddy et Cones) est de 13 : 1 secondes, 13 : 2 secondes, 11 : 6 secondes et 13 : 4 secondes, respectivement. Les résultats de la disparité intermédiaires et finaux pour les quatre paires stéréo de l'ensemble de données standard de Middlebury sont présentés à la figure 5.5.

TABLEAU 5.2 – Pourcentage d’erreurs dans différentes régions (nonocc, all , disc) avec un seuil de 1 pixel

Algorithmes	Tsukuba			Venus			Teddy			Cones		
	nonocc	toutes	disc	nonocc	toutes	disc	nonocc	toutes	disc	nonocc	toutes	disc
Notre Méthode	0.48	0.58	2.60	0.13	0.16	1.86	1.82	2.15	5.49	2.39	2.80	7.00
ADMM WT+CT2-Stage	0.51	0.62	2.01	0.08	0.11	1.19	2.06	2.56	7.37	2.68	3.16	8.09
ADMM WT+CT Grid	1.14	1.43	5.19	0.26	0.39	3.73	2.57	3.36	8.98	3.39	4.05	10.21
Method	2.57	2.77	10.82	2.92	3.00	20.05	11.36	12.17	18.39	13.20	14.15	21.03

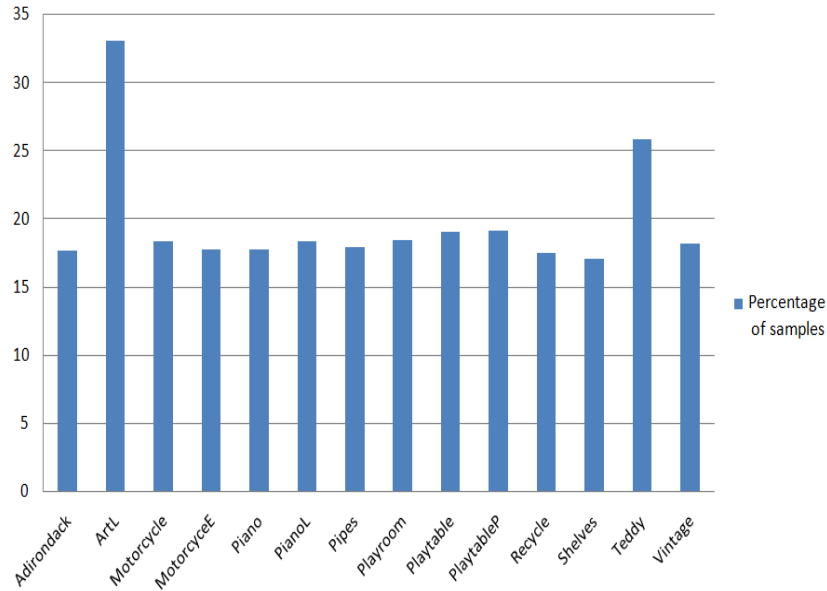


FIGURE 5.6 – Pourcentage des échantillons pour chaque pair pour la nouvelle base Middlebury avec un échantillonnage de 2600 super pixels

Nous avons effectué une autre expérience sur l’un des ensembles de données les plus utilisés pour la mise en correspondance stéréo, les datasets 2014 de Middlebury. Cette référence est divisée en deux parties, les ensembles d’entraînement et de test qui contiennent chacun 15 paires stéréo. L’ensemble d’entraînement est disponible en trois résolutions, pour lesquelles les cartes de disparité de la vérité du terrain sont fournies. L’ensemble de test utilise une plateforme d’évaluation pour enregistrer les résultats. Dans nos expériences, nous avons utilisé 15 images en résolution d’un quart de l’ensemble d’entraînement. Nous avons également utilisé le même nombre de superpixels, qui est fixé à 2600 comme première expérience. La figure 5.6 donne le pourcentage d’échantillons correspondants pour chaque paire.

Le tableau 5.3 présente les erreurs moyennes dans les régions non occluses calculées à 3 seuils d’erreur de pixels différents. Les résultats obtenus démontrent que la méthode proposée donne le taux d’erreur de disparité le plus bas dans tous les seuils d’erreur de pixel. Enfin, la figure 5.7 montre des exemples de résultats de disparité des nouveaux

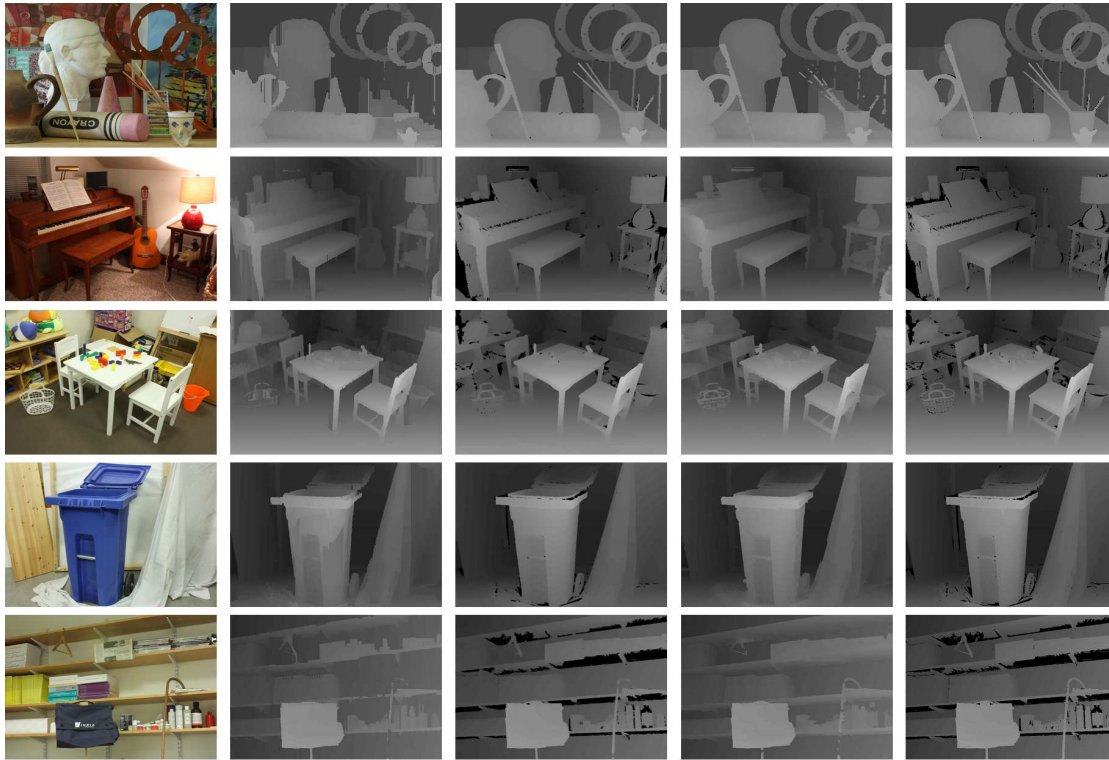


FIGURE 5.7 – Résultats sur la nouvelle base Middlebury pour (a) ArtL. (b) Piano. (c) Playtable. (d) Shelves. Du gauche à droite : l'image de référence, cartes de disparité produites par [51], [39], notre approche, vérité du terrain, respectivement

ensembles de données de Middlebury 2014 pour la méthode [51], ADMM WT + CT2-Stage [39] et les nôtres, respectivement.

5.3.2 Résultats de reconstruction en utilisant les autres cartes de disparités

Pour mieux évaluer l'efficacité de la méthode proposée, nous avons étudié les performances de reconstruction à l'aide de l'algorithme SGM [22]. Nous avons opté pour la méthode SGM car c'est l'approche la mieux étudiée entre l'appariement local et global. Nous avons conservé le même ensemble de configurations que celui utilisé précédemment pour les jeux de données de Middlebury 2014, et nous avons modifié les graines de disparité initiales par celles dérivées de l'algorithme SGM en utilisant les mêmes ratios d'échantillonnage de la figure 5.6.

Le tableau 5.3 présente les résultats de disparité obtenus en utilisant l'algorithme SGM et ceux de reconstruction. Selon les résultats rapportés dans le tableau 5.3, on constate que l'approche proposée permet de reconstruire des images basées sur la méthode SGM sans perdre aucune information. Par exemple, en utilisant un seuil d'erreur de 2 pixels, l'erreur moyenne a légèrement changé de 9h44 à 9h77 seulement. Cela sug-

TABLEAU 5.3 – Pourcentage des erreurs de disparité dans les régions non occultés pour la base Middlebury 2014

Algorithmes	seuil de 1 pixel nonocc	seuil de 2 pixels nonocc
SGM [22]	12.47	9.44
Notre méthode	13.40	9.77

gère l'efficacité de la méthode proposée pour traiter d'autres cartes de disparité initiales plutôt que la vérité du terrain.

5.4 Conclusion

Dans ce chapitre, nous avons présenté un nouvel algorithme de correspondance stéréo basé sur l'échantillonnage de superpixels SLIC. L'estimation des cartes de disparité denses est basée sur des disparités de frontières de superpixels. La reconstruction de la carte de disparité dense à partir des disparités de limites a été réalisée en utilisant la technique de propagation de la ligne de balayage. Les artefacts ont été traités efficacement à l'aide d'un filtre médian vertical adaptatif. Les résultats expérimentaux menés sur les ensembles de données de Middlebury ont démontré la précision et l'efficacité de la méthode proposée.



CONCLUSION GÉNÉRALE ET PERSPECTIVES

Conclusion

La recherche dans le domaine du transport intelligent, comprenant le développement de véhicules avec des systèmes d'aide à la conduite, vise à augmenter la fiabilité du transport routier. Dans cette thèse, nous avons tenté de contribuer à cet effort en proposant des approches de la vision stéréoscopique qui peuvent être utilisées pour la détection d'obstacles routiers dans un véhicule en mouvement.

L'utilisation de la vision stéréoscopique afin de mesurer la profondeur des objets ainsi que de reconstruire la structure tridimensionnelle de la scène observée nous a amené à aborder le problème de la correspondance stéréoscopique. La résolution de ce problème, qui se résume à l'estimation d'un champ de disparité, est rendue difficile par la présence inévitable de changements radio métriques de la scène, de bruit et de régions non texturées. Dans ce contexte, notre travail a consisté à améliorer les algorithmes de mise en correspondance stéréoscopique existants en proposant de nouvelles fonctions de coût robustes aux distorsions radiométriques. Ces travaux ont été présentés dans les chapitres 2, 3 et 4. Dans le cadre des méthodes locales, la méthode d'agrégation des coûts est également améliorée pour prendre en compte l'information sur les contours. Cette amélioration permet d'améliorer considérablement les coûts agrégés, notamment dans les régions homogènes, comme le montre le chapitre 2. Dans le chapitre 4, un autre type d'algorithmes de mise en correspondance stéréoscopique, à savoir les méthodes hiérarchiques, a été abordé. Dans ce contexte, nous avons proposé une nouvelle méthode locale dans un schéma hiérarchique, qui a permis de diminuer de manière significative les disparités tout en gardant un temps de calcul proche du temps réel. Le chapitre 6 était consacré à la reconstruction de cartes de disparité à partir de mesures de disparité éparses, au cours duquel nous avons proposé un nouvel algorithme d'appariement stéréoscopique qui permet de reconstruire efficacement une carte de disparité dense à partir d'une carte de disparité. L'algorithme calcule d'abord les disparités pour des points spécifiques de l'image, puis estime les disparités restantes pour la totalité des pixels en utilisant une méthode de propagation particulière.

Perspectives

Plusieurs perspectives et directions de recherche peuvent être envisagées dans ce travail. Nous prévoyons utiliser les cartes de disparité obtenues dans nos travaux dans le domaine des véhicules intelligents, notamment pour la détection d'obstacles routiers tels que les piétons. Nous voulons également déterminer le coût le plus approprié pour chaque pixel de l'image en utilisant des mesures de confiance telles que la mesure du score de correspondance (MSM), la courbure (CUR) qui peuvent prédire la performance des coûts.



LISTE DES PUBLICATIONS

O. Zeglazi, M. Rziza, A. Amine and C. Demonceaux, « Structural Similarity Measurement Based Cost Function for Stereo Matching of Automotive Applications », *Journal of Imaging*, vol 6(8), 2020.

O. Zeglazi, M. Rziza, A. Amine and C. Demonceaux, « A hierarchical stereo matching algorithm based on adaptive support region aggregation method », *Pattern Recognition Letters*, vol 112, pp 205-211, 2018.

O. Zeglazi, M. Rziza, A. Amine and C. Demonceaux, « Efficient Dense Disparity Map Reconstruction using Sparse Measurements », *13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, pp 534-540, 2018.

O. Zeglazi, M. Rziza, A. Amine and C. Demonceaux, « Accurate dense stereo matching for road scenes », *IEEE International Conference on Image Processing (ICIP)*, pp 720-724, 2017.

O. Zeglazi, M. Rziza and A. Amine, « An enhanced cross scale adaptive cost aggregation for stereo matching », *International Conference on Wireless Networks and Mobile Communications (WINCOM)*, pp. 1-5, 2017.



BIBLIOGRAPHIE

- [1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11) :2274–2282, Nov 2012.
- [2] M. Agrawal and L.S. Davis. Window-based, discontinuity preserving stereo. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I, 2004.
- [3] S. Birchfield and C. Tomasi. Depth discontinuities by pixel-to-pixel stereo. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 1073–1080, 1998.
- [4] S. Boukir, P. Bouthemy, F. Chaumette, and D. Juvin. Mise en correspondance de segments dans une squence d’images par une approche locale. Technical Report 1792, INRIA, October 1992.
- [5] A. Brunton, Chang Shu, and G. Roth. Belief propagation on the gpu for stereo vision. In *The 3rd Canadian Conference on Computer and Robot Vision (CRV’06)*, pages 76–76, 2006.
- [6] Peter J Burt. Fast filter transform for image processing. *Computer Graphics and Image Processing*, 16(1) :20–51, 1981.
- [7] J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-8(6) :679–698, 1986.
- [8] Ingemar J. Cox, Sunita L. Hingorani, Satish B. Rao, and Bruce M. Maggs. A maximum likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63(3) :542–567, 1996.
- [9] P.F. Felzenszwalb and D.R. Huttenlocher. Efficient belief propagation for early vision. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I, 2004.
- [10] C. Fookes, A. Maeder, S. Sridharan, and J. Cook. Multi-spectral stereo image matching using mutual information. In *Proceedings. 2nd International Symposium*

- on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pages 961–968, 2004.
- [11] Andrea Fusiello, Vito Roberto, and Emanuele Trucco. Efficient stereo with multiple windowing. In *CVPR*, pages 858–863, 1997.
 - [12] Andrea Fusiello, Emanuele Trucco, and Alessandro Verri. A compact algorithm for rectification of stereo pairs. *Mach. Vis. Appl.*, 12(1) :16–22, 2000.
 - [13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
 - [14] Andreas Geiger, Martin Roser, and Raquel Urtasun. Efficient large-scale stereo matching. In *Computer Vision - ACCV 2010 - 10th Asian Conference on Computer Vision*, volume 6492 of *Lecture Notes in Computer Science*, pages 25–38, 2010.
 - [15] Scott Grauer-Gray and Chandra Kambhamettu. Hierarchical belief propagation to reduce search space using cuda for stereo and motion estimation. In *2009 Workshop on Applications of Computer Vision (WACV)*, pages 1–8, 2009.
 - [16] Scott Grauer-Gray, Chandra Kambhamettu, and Kannappan Palaniappan. Gpu implementation of belief propagation using cuda for cloud tracking and reconstruction. In *2008 IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS 2008)*, pages 1–4, 2008.
 - [17] Kyu-Phil Han, Kun-Woen Song, Eui-Yoon Chung, Seok-Je Cho, and Yeong-Ho Ha. Stereo matching using genetic algorithm with adaptive chromosomes. *Pattern Recognit.*, 34(9) :1729–1740, 2001.
 - [18] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.
 - [19] Richard I. Hartley and Rajiv Gupta. Computing matched-epipolar projections. In *CVPR*, pages 549–555. IEEE, 1993.
 - [20] Heiko Hirschmüller, Peter R. Innocent, and Jonathan M. Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *Int. J. Comput. Vis.*, 47(1-3) :229–246, 2002.
 - [21] Heiko Hirschmüller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *CVPR*, 2007.
 - [22] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2) :328–341, Feb 2008.

- [23] Heiko Hirschmuller and Daniel Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9) :1582–1599, 2009.
- [24] Li Hong and G. Chen. Segment-based stereo matching using graph cuts. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I, 2004.
- [25] R. Horaud and T. Skordas. Stereo correspondence through feature grouping and maximal cliques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(11) :1168–1180, 1989.
- [26] Asmaa Hosni, Michael Bleyer, Margrit Gelautz, and Christoph Rhemann. Local stereo matching using geodesic support weights. In *ICIP*, pages 2093–2096, 2009.
- [27] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2) :504–511, 2013.
- [28] Hiroshi Ishikawa and Davi Geiger. Occlusions, discontinuities, and epipolar lines in stereo. In *Computer Vision — ECCV’98*, pages 232–248, 1998.
- [29] C. V. Jawahar and P. J. Narayanan. Generalised correlation for multi-feature correspondence. *Pattern Recognit.*, 35(6) :1303–1313, 2002.
- [30] Yi-Hung Jen, Enrique Dunn, Pierre Fite Georgel, and Jan-Michael Frahm. Adaptive scale selection for hierarchical stereo. In *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*, pages 1–10. BMVA Press, 2011.
- [31] G. A. Jones. Constraint, optimization, and hierarchy : Reviewing stereoscopic correspondence of complex features. *International Journal of Computer Vision, IJCV*, 65(1) :57–58, janvier 1997.
- [32] Sing Bing Kang, Richard Szeliski, and Jinxiang Chai. Handling occlusions in dense multi-view stereo. In *CVPR (1)*, pages 103–110. IEEE Computer Society, 2001.
- [33] Jae-Chul Kim, Kyoung Mu Lee, Byoung-Tae Choi, and Sang Uk Lee. A dense stereo matching using two-pass dynamic programming with generalized ground control points. In *CVPR (2)*, pages 1075–1082, 2005.
- [34] Jae Chul Kim, Kyoung Mu Lee, Byoung Tae Choi, and Sang Uk Lee. A dense stereo matching using two-pass dynamic programming with generalized ground control points. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 2, pages 1075–1082 vol. 2, 2005.
- [35] Seungryong Kim, Bumsub Ham, Seungchul Ryu, Seon Joo Kim, and Kwanghoon Sohn. Robust stereo matching using probabilistic laplacian surface propagation. In *Computer Vision – ACCV 2014*, pages 368–383, 2015.

- [36] A. Klaus, M. Sormann, and K. Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 15–18, 2006.
- [37] V. Kolmogorov and R. Zabih. Computing visual correspondence with occlusions using graph cuts. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 508–515 vol.2, 2001.
- [38] Vincent Lemonde. *Stéréovision embarquée sur véhicule : de l'auto-calibrage à la détection d'obstacles*. Theses, INSA de Toulouse, November 2005.
- [39] L. K. Liu, S. H. Chan, and T. Q. Nguyen. Depth reconstruction from sparse samples : Representation, algorithm, and sampling. *IEEE Transactions on Image Processing*, 24(6) :1983–1996, June 2015.
- [40] L. Morin et A. Tamtaoui M. Rziza, D. Aboutajdine. Schéma multirésolution d'estimation d'un champ de disparités dense sous contrainte épipolaire pour les images bruitées. In *Dans actes du colloque GRETSI sur le traitement du signal et des images, Toulouse, France, septembre, 2001*.
- [41] Hanspeter A. Mallot, Sabine Gillner, and Petra A. Arndt. Is correspondence search in human stereo vision a coarse-to-fine process? *Biol. Cybern.*, 74(2) :95–106, 1996.
- [42] D. Marr and T. Poggio. A Computational Theory of Human Stereo Vision. *Proceedings of the Royal Society of London Series B*, 204(1156) :301–328, 1979.
- [43] Jiri Matas, Ondrej Chum, Martin Urban, and Tomáš Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis. Comput.*, 22(10) :761–767, 2004.
- [44] Xing Mei, Xun Sun, Mingcai Zhou, Shaohui Jiao, Haitao Wang, and Xiaopeng Zhang. On building an accurate stereo matching system on graphics hardware. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 467–474, 2011.
- [45] M. D. Menz and R. D. Freeman. Stereoscopic depth processing in the visual cortex : a coarse-to-fine mechanism. *Nature Neuroscience*, 6 :59–65, 2003.
- [46] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015.
- [47] Dong Bo Min and Kwanghoon Sohn. Cost aggregation and occlusion handling with WLS in stereo matching. *IEEE Trans. Image Process.*, 17(8) :1431–1442, 2008.
- [48] A. Miron, A. Bensrhair, A. Rogozan, and S. Ainouz. Cross-comparison census for colour stereo matching applied to intelligent vehicle. *Electronics Letters*, 48(24) :1530–1532, November 2012.

- [49] Alina Miron, Samia Ainouz, Alexandrina Rogozan, and Abdelaziz Bensrhair. A robust cost function for stereo matching of road scenes. *Pattern Recognition Letters*, 38 :70–77, 2014.
- [50] K. Muhlmann, D. Maier, R. Hesser, and R. Manner. Calculating dense disparity maps from color stereo images, an efficient implementation. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pages 30–36, 2001.
- [51] S. Mukherjee and R. M. R. Guddeti. A hybrid algorithm for disparity calculation from sparse disparity estimates based on stereo vision. In *2014 International Conference on Signal Processing and Communications (SPCOM)*, pages 1–6, July 2014.
- [52] H. K. Nishihara. Prism, a practical real-time imaging stereo matcher. *Rapport technique A. I. Memo 780, Institut de Technologie du Massachusetts, MIT, MA, USA*, 1984.
- [53] Euripides G. M. Petrakis, Aristeidis Diplaros, and Evangelos E. Milios. Matching and retrieval of distorted and occluded shapes using dynamic programming. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(11) :1501–1516, 2002.
- [54] R. Maini Raman and H. Aggarwal. Study and comparison of various image edge detection techniques. *International Journal of Image-processing*, 3(1) :1–12, 2009.
- [55] S. Roy and I.J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 492–499, 1998.
- [56] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47 :7–42, 2002.
- [57] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, pages 31–42. Springer International Publishing, 2014.
- [58] Mikhail Sizintsev. Hierarchical stereo with thin structures and transparency. In *2008 Canadian Conference on Computer and Robot Vision*, pages 97–104, 2008.
- [59] C. Stentoumis, L. Grammatikopoulos, I. Kalisperakis, and G. Karras. On accurate dense stereo-matching using a local adaptive multi-cost approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 91 :29–49, 2014.
- [60] Christoph Stiller, J. Hipp, C. Rössig, and A. Ewald. Multisensor obstacle detection and tracking. *Image and Vision Computing*, 18 :389–396, 04 2000.

- [61] C. Sun. Fast optical flow using 3D shortest path techniques. *ivc*, 20(13/14) :981–991, Dec 2002.
- [62] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7) :787–800, 2003.
- [63] Li Tang, Mona K. Garvin, Kyungmoo Lee, Wallace L.W. Alward, Young H. Kwon, and Michael D. Abramoff. Robust multiscale stereo matching from fundus images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11) :2245–2258, 2011.
- [64] Hai Tao, Harpreet S. Sawhney, and Rakesh Kumar. A global matching framework for stereo computation. In *ICCV*, pages 532–539, 2001.
- [65] O. Veksler. Efficient graph-based energy minimization methods in computer vision. In *PHD. thesis Universie de Cornell, Ithaca, United stat*, August 1999.
- [66] O. Veksler. Fast variable window for stereo correspondence using integral images. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I, 2003.
- [67] Yichen Wei and Long Quan. Region-based progressive stereo matching. In *CVPR (1)*, pages 106–113, 2004.
- [68] Koichiro Yamaguchi, Tamir Hazan, David McAllester, and Raquel Urtasun. Continuous markov random fields for robust stereo estimation. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part V, ECCV’12*, pages 45–58, 2012.
- [69] Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *Computer Vision – ECCV 2014 : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 756–771, 2014.
- [70] J. Yan, Y. Yu, X. Zhu, Z. Lei, and S. Z. Li. Object detection by labeling superpixels. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5107–5116, June 2015.
- [71] Ruigang Yang and M. Pollefeys. Multi-resolution real-time stereo on commodity graphics hardware. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I, 2003.
- [72] Kuk-Jin Yoon and In-So Kweon. Locally adaptive support-weight approach for visual correspondence search. In *CVPR (2)*, pages 924–931, 2005.

- [73] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *Computer Vision - ECCV'94, Third European Conference on Computer Vision, Stockholm, Sweden, May 2-6, 1994, Proceedings, Volume II*, volume 801 of *Lecture Notes in Computer Science*, pages 151–158, 1994.
- [74] Oussama Zeglazi, Mohammed Rziza, Aouatif Amine, and Cédric Demonceaux. Accurate dense stereo matching for road scenes. In *ICIP*, pages 720–724. IEEE, 2017.
- [75] Kang Zhang, Yuqiang Fang, Dongbo Min, Lifeng Sun, Shiqiang Yang, Shuicheng Yan, and Qi Tian. Cross-scale cost aggregation for stereo matching. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1590–1597. IEEE Computer Society, 2014.
- [76] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit. Cross-based local stereo matching using orthogonal integral images. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(7):1073–1079, 2009.
- [77] Ye Zhang and Chandra Kambhamettu. Stereo matching with segmentation-based cooperation. In *Computer Vision — ECCV 2002*, pages 556–571, 2002.

Résumé

La sécurité des usagers du trafic routier constitue la préoccupation centrale des Systèmes de Transport Intelligents (STI). A cet effet, la détection des obstacles et notamment celle des piétons, est un élément fondamental des STI. Bien que ce thème a été étudié depuis près de cinquante ans, une solution idéale à ce problème se fait toujours attendre. Les méthodes courantes utilisées pour la détection des piétons reposent sur l'utilisation du LIDAR et de la vision par ordinateur. Dans cette thèse, on opte pour la vision par ordinateur, notamment pour la vision stéréoscopique, qui fournit des moyens de réduction de l'espace de recherche des objets dans les scènes routières grâce à l'utilisation des informations de profondeur afin d'obtenir des estimations fiables de l'emplacement des piétons. Ainsi, il est essentiel de disposer d'une carte de disparité robuste afin de dégager des régions de recherche fiables.

Plusieurs approches sont présentées dans cette thèse en vue de garantir une meilleure qualité des cartes de disparité en tenant compte des principaux problèmes observés dans le domaine de la vision stéréoscopique pour des scènes réelles, tels que les changements de métrique radio et les régions non texturées.

Mots-clefs : *Mise en correspondance stéréoscopique, Systèmes de transport intelligents, Carte de Disparité, KITTI, Middlebury*

Abstract

The safety of road users is the central concern of Intelligent Transport Systems (ITS). For this purpose, the detection of obstacles, especially pedestrians, is a fundamental component of ITS. Although this topic has been studied for almost fifty years, an ideal solution to this problem is still pending. The current methods used for pedestrian detection are based on the use of LIDAR and computer vision. In this thesis, computer vision is opted for, especially stereovision, which provides ways to reduce the search space of objects in road scenes through the use of depth information to obtain reliable estimates of pedestrian locations. Thus, a robust disparity map is essential to derive reliable search regions.

Several approaches are presented in this thesis to ensure better quality disparity maps by addressing the main problems observed in stereo vision for real scenes, such as radio metric changes and non-textured regions

Key Words: *Stereo matching, Intelligent Transportation Systems, Disparity Map, KITTI, Middlebury*