Détection et localisation tridimensionnelle par stéréovision d'objets en mouvement dans des environnements complexes - Application aux passages à niveau

Nizar Fakhfakh

Numéro d'ordre : 156

École Centrale de Lille

THÈSE

pour l'obtention du

Doctorat de l'École Centrale de Lille

Discipline : Automatique, Génie informatique, Traitement du signal et Images

Détection et localisation tridimensionnelle par stéréovision d'objets en mouvement dans des environnements complexes

Application aux passages à niveau

présentée par

Nizar FAKHFAKH

Soutenue publiquement le 14 Juin 2011 devant le jury composé de :

Pr. Jacques JACOT	Professeur, EPFL-LPM/Suisse	Président
Pr. Michel DEVY	Directeur de recherche, LAAS-CNRS/France	Rapporteur
Pr. Yassine RUICHEK	Professeur des Universités, UTBM/France	Rapporteur
Dr. Alain CROUZIL	MCF, IRIT/France	Examinateur
Dr. Dominique BERTRAND	Directeur d'études, CERTU/France	Examinateur
Dr. Simon COLLART-DUTILLEUL	MCF, HDR, ECL/France	Examinateur
Pr. El-Miloudi EL-KOURSI	Directeur de recherche, IFSTTAR-ESTAS	Directeur de thèse
Dr. Louahdi KHOUDOUR	CR1, HDR, IFSTTAR-LEOST	Co-directeur de thèse

Thèse préparée aux laboratoires ESTAS et LEOST de l'Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux (IFSTTAR), et le laboratoire LPM de l'Ecole Polytechnique Fédérale de Lausanne (EPFL).
Laboratoire LAGIS de l'Ecole Centrale de Lille.

Je dédie ce mémoire à mon père Tijani, ma mère Chafia, mes frères Lassaad, Walid, Mohammed et Jacques, et ma femme Amenie qui par leur amour et leur soutien permanent m'ont permis de faire aboutir ce travail.

Remerciements

Les travaux présentés dans ce mémoire ont été menés à l'Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux (IFSTTAR) (issu de la fusion entre l'Institut National de Recherche sur les Transports et leur Sécurité (INRETS) et le Laboratoire Central des Ponts et Chaussées (LCPC)) au sein des laboratoires "Évaluation des Systèmes de Transport Automatisés et de leur Sécurité (ESTAS)" dirigé par El-Miloudi EL-KOURSI, et le laboratoire "Électronique, Ondes et Signaux pour les Transports (LEOST)" en la personne de Louahdi KHOUDOUR. La thèse est en collaboration avec l'École Polytechnique Fédérale de Lausanne (EPFL), au sein du "Laboratoire de Production Microtechnique (LPM)" dirigé par le professeur Jacques Jacot. Tout d'abord, je remercie l'IFSTTAR et le Conseil Régional du Nord-Pas de Calais de m'avoir octroyé une bourse de recherche pour mener à bien les travaux présentés dans ce mémoire.

Je tiens à exprimer ma profonde reconnaissance à mes directeurs de thèse, El-Miloudi El-KOURSI qui m'a accompagné des années durant, partageant mes difficultés, mes inquiétudes, m'encourageant à persévérer sans relâche et auprès duquel j'ai beaucoup appris, ainsi que Louahdi KHOUDOUR qui a consacré à l'encadrement de ma thèse un temps et une disponibilité d'esprit considérables, sa modestie dût-elle en souffrir un peu, de dire publiquement combien j'ai apprécié sa passion pour la recherche et son encadrement efficace. Je peux le dire : la pertinence de ses remarques et critiques, ses conseils avisés m'ont été d'un grand secours pour l'achèvement de cette thèse. Je le remercie également pour son humour, sa gentillesse. J'ai ainsi largement pu profiter de leur grande acuité scientifique et de leur enthousiasme indéfectible et communicatif. Je leur suis donc redevable d'avoir pu faire une thèse dans des conditions exceptionnelles.

J'ai été très sensible aussi à l'ouverture d'esprit du professeur Jacques JACOT, qui m'a accompagné tout au long de ce travail de thèse, au cours de longues et fructueuses discussions lors de mes visites à Lausanne en Suisse ou ses visites à Villeneuve d'Ascq en France. Je le remercie pour sa simplicité, sa bonhomie, sa générosité et spécialement son don particulier de remonter le moral quand on est désemparé.

Je remercie sincèrement Monsieur Yassine RUICHEK, professeur à l'Université de Belfort, et Monsieur Michel DEVY, directeur de recherche au LAAS qui ont eu la gentillesse d'avoir accepté de rapporter sur ma thèse, et enfin Messieurs Jacques JACOT, Alain CROUZIL, Simon COLLART-DUTILLEUL, et Dominique BERTRAND pour avoir accepté d'être examinateurs.

Je tiens par ailleurs à remercier Dr. Jean-Luc BRUYELLE, un collègue du LEOST à l'IFSTTAR, et Jacques MALMAISON, mon meilleur ami, qui ont manifesté un grand intérêt pour l'ensemble de ma thèse en prenant la peine de lire en détail la thèse et de me faire part de leurs commentaires. Je leur en suis très reconnaissant.

Un grand merci à mes collègues aux laboratoires LPM de l'EPFL et LEOST de l'IFSTTAR, en particulier Alain DUFAUX, Amaury FLANCQUART, Jean-Luc BRUYELLE pour leur disponibilité, conseils et leur précieuse aide technique.

Un immense merci à mon père Tijani et ma mère Chafia ABID, à mes frères Lassaad, Walid, et Mohamed, à mon ancien propriétaire et meilleur ami Jacques MALMAISON pour leur patience, amour, et leur soutien continuel, qui sans eux, ne m'aurait pas permis d'arriver jusque là.

Un très grand merci à ma femme Amenie FAKHFAKH CHERIF pour m'avoir toujours soutenu et encouragé, pour son grand amour, et pour tout ce qu'elle a fait pour moi et pour tout ce que qu'elle fera encore.

Et si après tout ce petit monde, il se trouve que j'ai oublié certaines personnes, qu'ils me pardonnent et trouvent ici toute ma sympathie.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Le développement considérable des outils informatiques a permis une large expansion de la vision artificielle. Elle demeure une des techniques émergentes introduites dans de très nombreuses applications telles que le contrôle qualité, la médecine, et la vidéosurveillance. Les techniques de vision artificielle s'adaptent aux applications envisagées et aux contraintes qui y sont associées. Certaines applications nécessitent un traitement en temps réel, telles la détection d'obstacles routiers par vision embarquée. D'autres applications nécessitent de la précision dans le traitement et l'analyse des données visuelles telles que la vision industrielle pour la détection de défauts de fabrication. En vidéosurveillance, l'humain est devenu de plus en plus incapable de gérer manuellement la multitude de moniteurs filmant les zones surveillées. Les objectifs d'un tel système de vision varient de l'analyse statistique à l'identification automatique de situations dangereuses. L'analyse statistique est introduite comme un outil d'aide à l'exploitation et d'aménagement de l'infrastructure permettant la réduction des coûts. Nous citons à titre d'exemple des systèmes basés sur la vision artificielle permettant le comptage du nombre de passagers entrant et sortant d'un bus, et l'estimation de la longueur des files d'attente dans un aéroport. Récemment, le besoin incessant en terme de sécurité des personnes et des biens ne cesse d'augmenter. L'aspect sécuritaire est récemment intégré comme contrainte dans diverses applications de vidéosurveillance. La détection d'intrusion, de colis abondonnés, et de situations dangereuses sont des exemples d'applications dont la contrainte de sécurité est fortement imposée.

Contexte applicatif

L'introduction progressive des nouvelles technologies dans différents modes de transport a augmenté le besoin en termes de sécurité des usagers et des infrastructures. Depuis quelques années, les passages à niveau (PN) ont connu un regain d'attention suite à des études statistiques ayant identifié ces lieux comme particulièrement dangereux. Chaque année en Europe, les accidents aux PN font 1200 victimes corporelles, dont 400 tués [KGB+09]. La même source a identifié, que parmi le nombre de décès survenant sur le réseau ferroviaire, 29% surviennent aux PN. Les procès verbaux d'accidents révèlent que 90% de ces accidents sont dus aux mauvais comportements humains plutôt qu'à un dysfonctionnement des dispositifs techniques. Une action de coordination du

6ème programme intitulé "*Safer European Level Crossing Appraisal and Technology (SELCAT)*" a donné des recommandations afin d'accroitre la sécurité aux passages à niveau. Ces recommandations sont développées autour des deux idées suivantes : l'utilisation des technologies émergeantes permettant de réduire l'impact des mauvais comportements humains, et la coordination des actions des opérateurs routiers et ferroviaires dans le contrôle et la réduction des risques aux PN.

Les passages à niveau sont équipés par des systèmes conventionnels de signalisation tels que les feux d'autorisation/interdiction de passage, les demi/double barrières, les panneaux d'informations, etc. Ces dispositifs ne sont pas conçus pour éviter des comportements dangereux, mais plutôt pour informer les usagers sur l'état du PN. À l'heure actuelle, très peu d'informations est disponible sur la position des usagers sur et autour du passage à niveau. La plupart des accidents aux PN sont dûs aux collisions entre un train et un obstacle présent dans la zone de croisement. En effet, la détection et la localisation d'obstacles sont devenues des informations cruciales dans l'évaluation des risques d'accidents et l'anticipation des situations dangereuses. Un tel système de détection d'obstacles semble être une solution permettant la réduction du nombre et de la gravité des accidents. Il devrait répondre aux exigences suivantes :

- Amélioration de la sécurité des usagers aux passages à niveau (piétons, voitures, camions, cyclistes, etc.).
- Minimisation des retards, aussi bien pour les trains que pour les usagers routiers.
- Évaluation des situations et de leur impact sur l'exploitation.
- Réduction des coûts et de la complexité d'installation et de maintenance.

Différentes techniques permettent la détection d'obstacles sur les PN, telles que l'utilisation de capteurs optiques ou sonores, de boucles électromagnétiques, de radars et la vision artificielle. Le choix du capteur est fortement dépendant des facteurs extérieurs tels que les conditions environnementales, la nature et la taille des obstacles à détecter. Les détecteurs d'obstacles peuvent être répartis en deux catégories : techniques conventionnelles, et techniques nouvelles.

Parmi les techniques conventionnelles existantes nous citons le détecteur optique d'obstacle appelé Faisceau optique ou "*Optical beam*". Malgré sa simplicité de mise en œuvre, cette technique présente certains inconvénients tels que son coût élevé d'installation, le besoin d'avoir plusieurs détecteurs optiques autour de la zone de croisement, l'arrêt du trafic ferroviaire et routier pendant son installation. Une deuxième technique conventionnelle est basée sur des détecteurs ultra-soniques qui ont l'avantage de détecter à la fois des voitures en mouvement ou à l'arrêt. L'inconvénient de

cette technique est que les coûts d'achat et d'installation sont trop élevés ; elle manque aussi de précision en cas d'embouteillage, et elle n'est pas adaptée à la détection de piétons. Nous pouvons citer une troisième technique, appelé Boucle Inductive, basée sur un flux magnétique. Les avantages de la Boucle Inductive résident dans sa facilité d'installation, et son indépendance aux conditions environnementales. Malgré ces avantages, le coût d'installation et de maintenance est trop élevé puisqu'elle nécessite plusieurs boucles inductives pour augmenter son efficacité. Par ailleurs, les piétons ne peuvent pas être détectés, et les trafics routier et ferroviaire doivent être interrompus lors de l'installation.

D'autres techniques ont été introduites pour la détection d'obstacles sur les passages à niveau, comme les radars. Cette technique émergeante permet de détecter à la fois des voitures et des piétons, n'interrompt pas le trafic lors de l'installation, et est peu sensible aux interférences électromagnétiques. L'inconvénient réside dans la difficulté de maintenance.

La vidéosurveillance a récemment été considérée comme une des méthodes prometteuses permettant la détection d'obstacles sur les passages à niveau. L'avantage de la vision artificielle est qu'elle est précise, et permet de détecter n'importe quel type d'objet (voiture, piéton, cycliste, etc.). Le système est facilement installé sans interruption du trafic. Le coût d'achat dépend de la qualité des caméras à installer. À ce jour très peu de systèmes à base de vision artificielle ont été installés en exploitation à des passages à niveau. Nous n'avons recensé que deux recherches à ce sujet :

- **Un premier système basé sur la vision monoculaire**, proposé par [For98]. Une seule caméra est installée sur un coin du passage à niveau supervisant la zone de croisement. Les obstacles sont initialement détectés par une différence entre une image du fond et l'image courante de la séquence. Les obstacles sont localisés en se basant sur les paramètres intrinsèques de la caméra, et en supposant le plan continu du sol. Les obstacles sont suivis à partir d'un filtre de Kalman. La détection et la localisation 3D ne sont pas très précises, surtout en présence d'occultations.

- **Un deuxième système basé sur la vision stéréoscopique**, a été proposé par [Oht05]. L'auteur propose d'utiliser deux caméras supervisant la zone de croisement afin de gérer les problèmes d'ombres et de minimiser les taux de fausses alarmes. La carte de disparité, obtenue à partir des deux images gauche et droite, est traduite en une carte des distances 3D en utilisant les paramètres intrinsèques et extrinsèques du système de vision. La carte de disparité est estimée à partir d'un simple algorithme de mise en correspondance stéréoscopique basé sur des méthodes locales. Elle contient des erreurs de localisation 3D, pour plusieurs raisons que nous abordons brièvement dans le paragraphe suivant.

Problématique scientifique

Le nombre de caméras de vidéosurveillance ne cesse d'augmenter. De ce fait, le nombre de moniteurs dans les centres de contrôles se sont multipliés, rendant la tâche de surveillance des écrans inéfficace. Des recherches ont été menées afin d'automatiser le processus de surveillance pour aider l'exploitant en lui fournissant des informations pertinentes sur les scènes supervisées. Une des contraintes forte qu'un système de détection d'obstacles doit satisfaire est la précision en termes de détection et de localisation d'obstacles. La qualité de la détection et de la localisation tridimensionnelle est d'une importance majeure, puisqu'une fausse détection génère une fausse alerte dont les conséquences peuvent être catastrophiques. Une surestimation de la distance d'un obstacle peut conduire à interrompre le trafic et à des pertes d'exploitation injustifiés. Une sous-estimation de la position d'un obstacle peut conduire à une mauvaise évaluation de la dangerosité de la situation et engendrer un accident. La détection des régions d'intérêt, qui correspondent aux objets en mouvement ou en stationnement temporaire, est réalisée à partir d'une seule caméra fixe. La détection consiste à extraire les régions affectées par un mouvement dans une séquence d'images. Une grande partie des méthodes existantes se base sur la différence entre une image de référence qui correspond au fond, et une image quelconque de la séquence. Les méthodes de détection diffèrent dans leurs temps de traitement et leurs capacité d'extraction dans des environnements complexes. Compte tenu de la variabilité des conditions environnementale, améliorer la qualité d'extraction et la rendre aussi robuste que possible est une de nos prioritées.

Une fois les objets extraits, nous cherchons à les localiser dans un espace tridimensionnel avec la plus grande précision possible compte tenu du contexte sécuritaire de l'application. La vision stéréoscopique, effectuée à partir de deux images bidimensionnelles, permet d'éviter des problèmes. D'une manière similaire à la vision humaine, la vision tridimensionnelle de l'environnement est possible en utilisant deux caméras supervisant la même scène. La connaissance de la géométrie du système stéréoscopique est indispensable. Les distances 3D sont estimées avec la connaissance préalable de la position et de l'orientation relative des caméras, des paramètres intrinsèques de chaque caméra, et du décalage de chaque pixel dans les deux images gauche et droite acquises simultanément. L'estimation des positions de chaque pixel dans les deux images gauche et droite, acquises simultanément, est un problème difficile à cause de plusieurs contraintes. Pour chaque pixel de l'image gauche, considérée comme image de référence, il s'agit de chercher le pixel homologue dans l'image droite. Afin de faciliter cette tâche, le système est calibré de sorte que l'homologue d'un pixel de l'image de référence se trouve sur la même ligne image. Le décalage horizontal entre les positions de deux pixels homologues est appelé disparité. L'ensemble des décalages forme une carte appelée "carte de disparités".

Un algorithme de mise en correspondance stéréoscopique, appelé aussi appariement stéréo-

scopique, est appliqué sur les deux images gauche et droite. L'appariement consiste à trouver les pixels ou primitives homologues dans les images afin de calculer leur disparité. Diverses *méthodes locales* existent : ces dernières ne tiennent compte que d'un voisinage réduit autour du pixel à apparier. L'idée est de trouver dans l'image droite un pixel dont le voisinage est le plus semblable à celui de l'image de référence. Comment se comportent-elles les méthodes locales dans les cas des pixels ou des primitives :

– Appartenant à des régions de couleur ou d'intensité uniforme ou homogène ?

– Appartenant à des régions de texture répétitive ?

– Occultés ?

Malgré un temps d'appariement raisonnable, les méthodes locales ne sont pas en mesure de gérer ces cas d'occultations, de textures répétitives, de zones de couleur uniforme. Ces problèmes sont en partie résolus par les méthodes d'appariement globals. Cette famille de méthodes introduit des contraintes globales, telles que la contrainte de symétrie. La limite majeure des méthodes globales réside dans leur temps de traitement important. La question que nous nous posons est alors la suivante : Comment exploiter conjointement les méthodes locales et globales de façon à obtenir des appariements aussi précis que possible en un temps de traitement raisonnable ?

Organisation du mémoire

Le présent mémoire s'articule autour de quatre chapitres, illustrés par la figure 1 :

- **Introduction générale** : Nous introduisons dans ce chapitre deux points principaux. Le premier point décrit le contexte applicatif au sein duquel notre travail s'intègre. Le deuxième point détaille la problématique scientifique que nous développons tout au long de la thèse.

- **Chapitre 1 – Estimation du relief par vision Stéréoscopique : État de l'art** : Ce chapitre commence par identifier les techniques permettant la vision en relief, notament en mettant l'accent sur la vision stéréoscopique. La deuxième partie détaille l'aspect géométrique d'un système de vision stéréoscopique, principalement les disparités calculées pour chaque pixel. La troisième partie est consacrée à un état de l'art sur les méthodes existantes permettant l'établissement des cartes de disparités.

- **Chapitre 2 – Développement d'un algorithme d'appariement stéréoscopique sélectif** : L'état de l'art détaillé dans le premier chapitre nous a permis d'identifier les avantages et les

limites des différentes méthodes d'appariement stéréoscopique. Compte tenu des limites à surmonter et des champs applicatifs, nous proposons dans ce chapitre une nouvelle méthode d'appariement stéréoscopique tirant profit des méthodes locales et globales, et permettant une amélioration significative de la qualité de l'appariement. L'algorithme proposé est composé de trois parties. La première partie est une nouvelle méthode locale d'appariement. La deuxième partie permet d'identifier automatiquement les faux appariements à partir d'une mesure de confiance calculée sur l'ensemble des appariements. La dernière partie consiste à ré-estimer les disparités erronées en se basant sur une méthode globale permettant la propagation sélective des croyances.

- **Chapitre 3 – Estimation du mouvement par analyse en composante Indépendante (ACI)** : L'estimation du mouvement est introduite dans notre contexte en tant que contrainte permettant l'accélération du processus d'appariement stéréoscopique. Nous faisons l'hypothèse que l'homologue d'un pixel affecté d'un mouvement, est lui aussi affecté par ce mouvement. Ce chapitre débute par un état de l'art sur les méthodes de soustraction du fond à partir d'une séquence d'images. Compte tenu des limites de ces dernières ainsi que du cadre applicatif, nous proposons une nouvelle méthode d'extraction des régions affectées par un mouvement. L'algorithme proposé se compose de deux parties. La première partie consiste à détecter les régions en mouvement par une méthode rapide, beaucoup moins sensible aux variations d'illumination, et tenant compte de la stationnarité des objets. La deuxième partie consiste à filtrer le résultat de détection issu de la première partie, afin de bien extraire les vraies régions affectées par un mouvement.

- **Chapitre 4 – Localisation tridimensionnelle d'obstacles sur les passages à niveau** : Nous proposons dans ce chapitre d'appliquer les algorithmes développés dans le deuxième et troisième chapitre, à la détection et la localisation 3D d'obstacles en mouvement ou en stationnement temporaire dans le cadre de l'application visant à accroitre la sécurité aux PN. Ce chapitre commence par une analyse préliminaire des risques afin d'établir un cahier des charges. Cette étape nous permet d'analyser et d'identifier les causes possibles affectant la sécurité des usagers aux PN. Le principal objectif est de mettre en valeur les performances et l'apport de nos algorithmes de localisation 3D d'obstacles dans un environnement nécessitant un niveau de précision important.

- **Conclusion et perspectives** : Nous terminons ce rapport par un résumé des apports et des démarches suivies tout au long de la thèse. Les perspectives que nous proposons dans cette partie portent sur des aspects scientifiques et applicatifs.

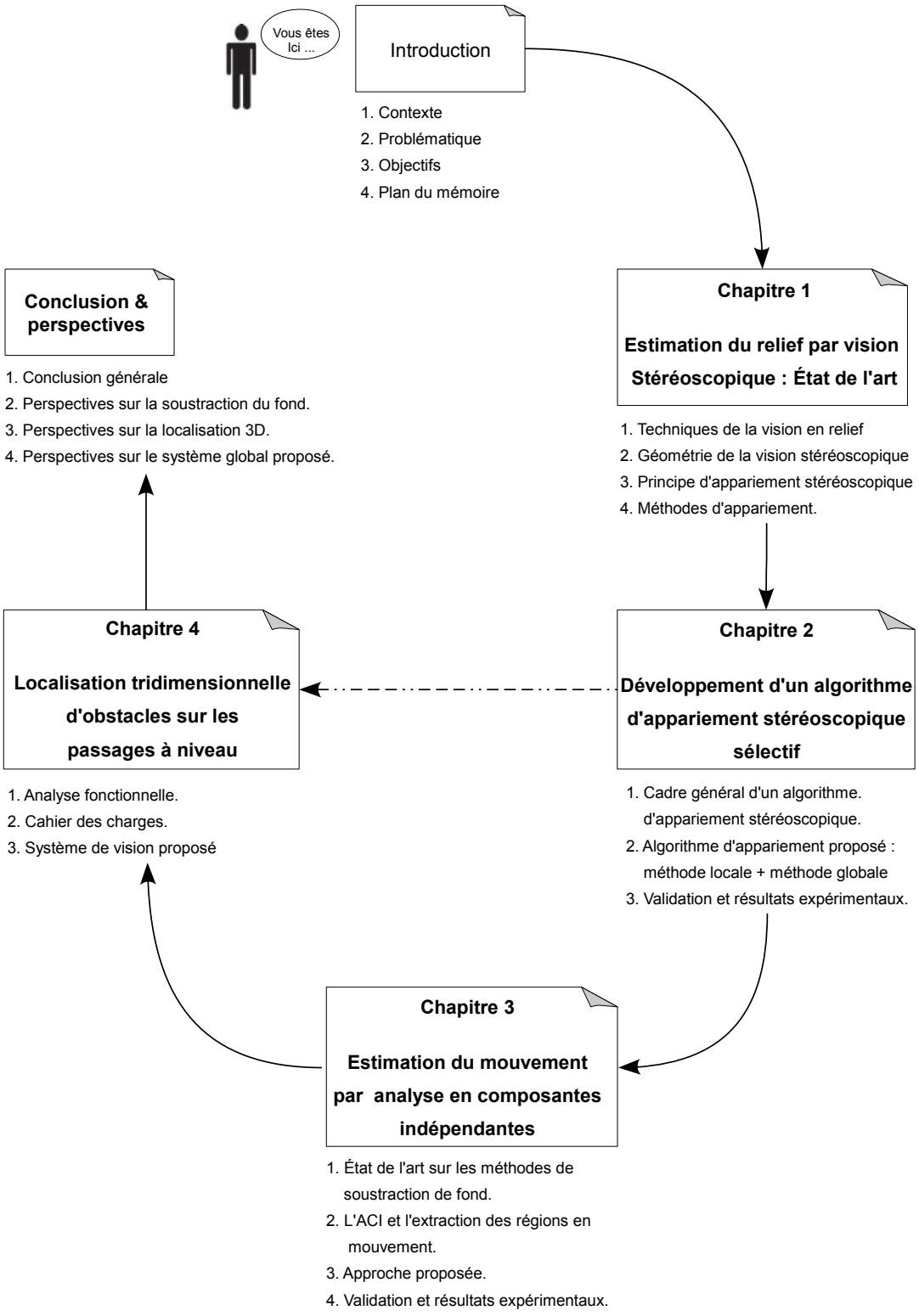


FIGURE 1 – Plan du mémoire.

Chapitre 1

Estimation du relief par vision stéréoscopique : État de l'art

1.1 Introduction

L'avancée des techniques informatiques et méthodologiques a permis à plusieurs applications de s'intégrer dans notre vie sous différentes formes. Certaines pour automatiser certaines tâches et d'autres pour nous protéger des dangers, présents dans notre environnement, qui sont de plus en plus nombreux. Durant ces dernières décennies, diverses applications de la vision artificielle ont vu le jour, notamment avec le développement considérable des capacités de calcul, et peuvent être réparties selon ces deux catégories : vision fixe et vision en mouvement. L'aide à la conduite automobile, les robots mobiles, sont quelques exemples d'applications utilisant des capteurs visuels embarqués. Ce type d'applications est généralement implémenté sur des architectures matérielles dédiées et exigent souvent du traitement en temps réel. Une scène réelle peut être perçue par un ou plusieurs capteurs vidéo. Chaque objet tridimensionnel de la scène est projeté sur le plan image bidimensionnel de chaque caméra. La perception bidimensionnelle ne permet qu'une analyse superficielle qui, par analogie au Système Visuel Humain, correspond aux traitements de la rétine [TTJ+00]. Une perception aussi réaliste que possible de l'environnement conduit à une meilleure compréhension et interprétation de ce que nous observons. D'où l'intérêt de percevoir la troisième dimension, la distance, et d'analyser le comportement d'objets tridimensionnels en introduisant la quatrième dimension, le temps. La figure 1.1 illustre un schéma simplifié des différentes étapes de la vision naturelle :

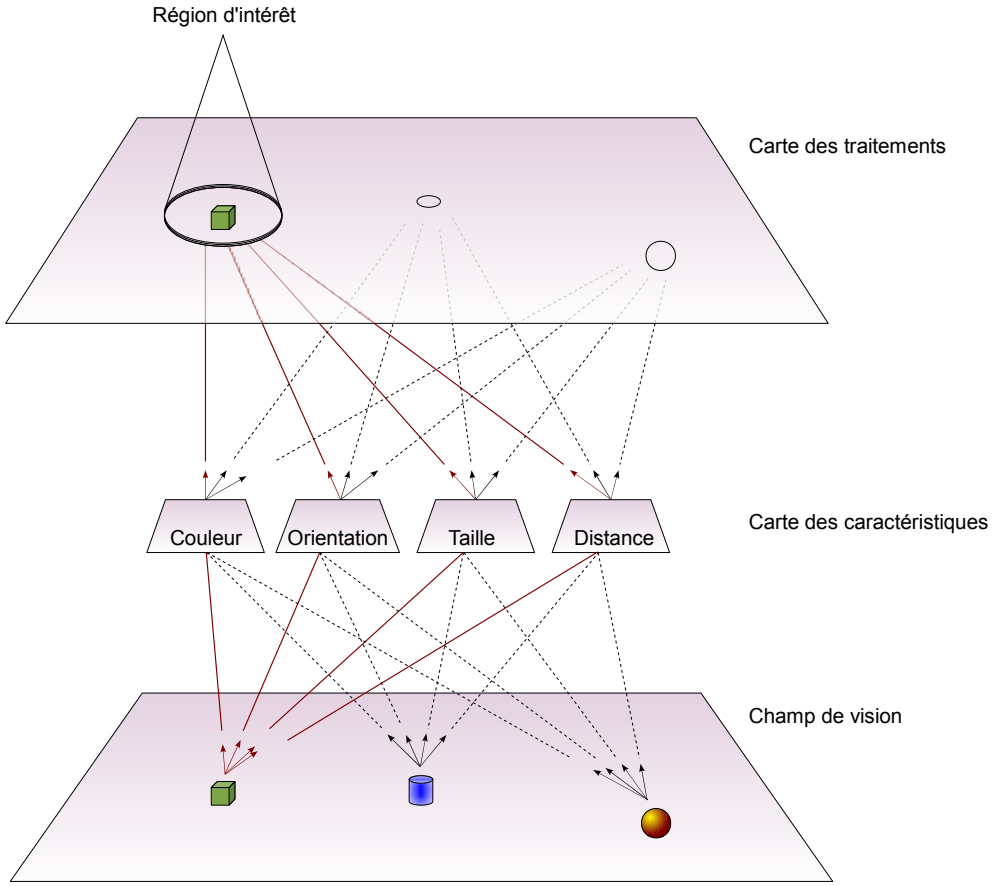


FIGURE 1.1 – Principe du processus de perception de l'environnement par le système visuel humain.

1.2 La perception tridimensionnelle

En vision naturelle, l'homme et de nombreux espèces animales sont capables de reconstituer l'espace à partir d'une ou de plusieurs vues de la scène [FCF10]. Pour ce faire, de nombreuses informations sont prises en compte, telles que :

- **L'accommodation**, qui est la mise au point dynamique de notre cristallin. Le phénomène de la mise au point rend la région d'intérêt très nette tout en rendant flous les détails des autres régions.

- **La parallaxe monoculaire**, qui est le déplacement relatif des objets observés.

- **Les perspectives géométriques**, qui sont des facteurs fondés sur les différentes comparaisons entre des objets visibles et qui font appel à des éléments cognitifs du sujet, en l'occurrence les expériences passées et la reconnaissance des objets.

En se basant sur des repères cognitifs, le cerveau humain est capable de reproduire les reliefs d'une scène même en présence d'occultations (couverture partielle d'un objet par un autre). Cela est réalisé par la connaissance a priori de la taille et de la forme de chaque objet de la scène observée. Ce principe nécessite une importante étape d'apprentissage qui rend le processus de reconstruction tridimensionnelle précis et efficace. Malgré l'importance capitale de la connaissance préalable de la scène observée, ceci n'est pas toujours facile à réaliser à cause de l'indisponibilité d'une base d'apprentissage telle que le projet de l'exploration de la planète Mars [mis].

La vision humaine est une source importante d'inspiration des algorithmes de perception artificielle de l'environnement. En se basant sur la vision binoculaire, M. Taira [TTJ+00] a étudié le rôle de certains neurones dans la perception de l'orientation des surfaces tridimensionnelles. L'auteur a montré que l'orientation des surfaces 3D est obtenue en estimant le gradient de disparité tout au long des contours. Farell a examiné dans son article [FLM04] l'idée originalement proposé par Marr et Poggio [MP76] pour l'analyse des interactions unidirectionnelle et multi-échelle pour l'estimation de la disparité. Dans [BKR+08], l'auteur examine les différentes méthodes permettant l'estimation de la profondeur à partir des mesures de disparité. Il suggère l'hypothèse que la profondeur d'un object réel est estimée à partir de l'orientation des yeux par rapport au mouvement, l'orientation et la gravité du reste du corps humain. L'auteur a montré que la disparité rétinale ne peut pas donner une unique estimation de la profondeur des objets. Y. Liu [LBC08] propose une analyse statistique sur la distribution des disparités dans des environnements intérieurs et extérieurs. L'échelle de disparité varie en fonction de la scène observée : les expérimentations montrent une forte variation de disparité dans le cas d'une scène intérieure, et une faible variation dans le cas d'une scène extérieure. L'auteur montre que, dans le cas des deux scènes testées, les disparités sont centrées sur la valeur zéro, ayant aussi des pics importants et varient de 5 degrés autour de la moyenne.

Nous proposons aux lecteurs intéressés d'autres travaux récents sur le fonctionnement et l'interaction des différentes zones du cerveau impliquées dans la perception tridimensionnelle [BES05], [HC05], [MVF05], [Ner05], [HL06], [SO06], [RSG06], [CCD+07], [RPBD07], [TKWB08], [CF09], [JOK09], [WB09], [Wes09], [AGV10], [HB10], [LY10], étude sur la précision de l'estimation des profondeurs de la vision monoculaire et binoculaire dans des configurations naturelles [MT10], la perception stéréoscopique d'objets lointains [PGG+10], détection des occlusions [TWA10], [WHVW10], [WEEW10], et [YIC10].

En vision monoculaire, la troisième dimension n'est pas facile à estimer. La figure 1.2 illustre le problème d'avoir une seule projection, sur le plan image de la caméra, de plusieurs points de l'espace réel. Il manque d'autres informations pour pouvoir estimer les coordonnées de chaque

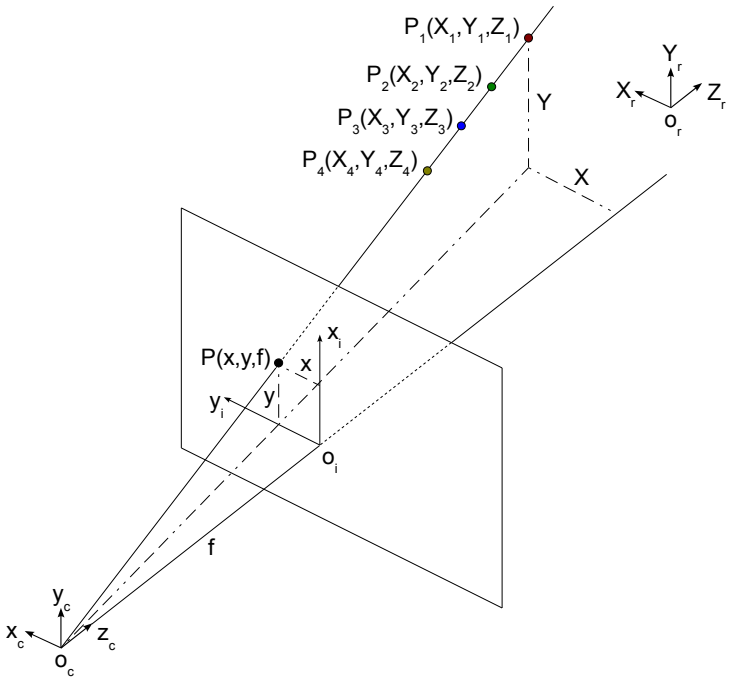


FIGURE 1.2 – La géométrie d'un système de vision monoculaire.

point de l'espace. En vision par ordinateur, la composante tridimensionnelle peut être obtenue de différentes façons. S.T. Barnard [BF82] a distingué trois techniques permettant la vision des reliefs : la vision active (laser, radar, lumière structurée), la stéréophotométrie et la vision binoculaire. En nous basant sur la *vision passive* [1], nous détaillons ci-après les trois principales techniques permettant la vision tridimensionnelle : la stéréophotométrie, la vision monoculaire en présence du mouvement, et la stéréovision.

1.2.1 Stéréophotométrie

En stéréophotométrie, la scène et le capteur sont fixes, seule la source lumineuse se déplace. La stéréophotométrie consiste à estimer la structure tridimensionnelle d'une scène supervisée par une caméra et soumise successivement à différentes sources d'illumination. La contrainte d'invariance colorimétrique n'est pas exigée puisque deux images de la scène ne peuvent être prises qu'à deux instants différents. Cette technique est généralement utilisée pour une reconstruction tridimensionnelle des formes des objets dans des environnements contrôlés [HMJI09] ou non contrôlés [CC06] [BJK07]. Les sources lumineuses peuvent être colorées [HVB+07] ou non [KS06]. Cette technique a l'avantage de reconstruire parfaitement des surfaces non texturée. N. Alldrin [AZK08] propose un algorithme de reconstruction de la forme d'un objet basée sur des approximations bi-variées

1. En vision passive, une scène est simplement observée par une ou plusieurs caméras sans interaction entre les capteurs utilisés et l'environnement.

des fonctions de réflectances isotropes. H.-S. Chung [CJ08] propose d'estimer les paramètres de la fonction de distribution de réflectance bidirectionnelle par analyse de l'ombre d'un objet. Ceci, avec un processus d'optimisation itérative, permet de calculer les normales d'une surface de façon robuste. Une des applications de la stéréophotométrie est la restauration d'une scène réelle extérieure à partir de plusieurs vues de cette dernières prise sous différentes conditions météorologiques telles que le brouillard et la brume [KN09]. D'autres travaux ont été proposés récemment pour l'estimation de la structure tridimensionnelle des objets dans des environnements intérieurs contrôlés [DC07] [DC08] [HVC08b] [HVC08a], [VG08] [MS09]. L'inconvénient majeur de cette technique est que dans la réalité les scènes sont généralement dynamiques.

1.2.2 Structure 3D basée sur le mouvement

Estimer la structure 3D à partir du mouvement "*Structure from Motion*" est un domaine qui attire de plus en plus l'attention des chercheurs en vision artificielle. Le principe consiste à estimer les paramètres du mouvement d'une ou des caméras ainsi que la structure 3D d'une scène observée en ne se basant que sur une séquence d'images. Les images sont généralement prises à partir d'une scène dynamique, mais cela ne rajoute pas beaucoup d'information pour la reconnaissance du relief. L'analyse de scène, basée sur le mouvement, présente cependant un point commun avec la vision stéréoscopique, puisqu'il faut réaliser la mise en correspondance d'un même point physique sur une séquence d'images.

[KC06] propose un algorithme basé sur un facteur d'échelle absolu pour l'estimation des paramètres du mouvement et de la structure d'une scène observée à partir de l'appariement basé sur le mouvement. [SD10] propose une reconstruction 3D d'une zone urbaine à partir d'une séquence d'images issues d'une seule caméra. [TM08], propose de représenter le problème d'estimation de la structure 3D d'une scène par une propagation de croyance particulaire combiné avec une méthode de *Monte Carlo par Chaine de Markov*. L'estimation de la structure 3D d'une scène dynamique à partir de plusieurs vues monoculaires a pris de plus en plus d'importance durant la dernière décennie. Plusieurs travaux récents traitent le problème d'estimation des paramètres des caméras, et de la structure 3D d'une scène en se basant seulement sur une séquence d'images [FPC+09] [FCSS09] [WKI09] [RN10]. Le problème de la vision des reliefs basée sur le mouvement est que le rendu tridimensionnel n'est pas toujours précis pour des environnements extérieurs non contrôlés [BKP09].

1.2.3 Vision binoculaire : la stéréoscopie

La stéréoscopie est une méthode de vision passive inspirée de la vision humaine permettant d'obtenir l'information de relief d'une scène à partir de deux projections bidimensionnelles de la même scène (*cf.* figure 1.3). La position tridimensionnelle des points objet est déduite à partir

de deux images d'une même scène prises de deux points de vue légèrement différents. Les deux images stéréoscopiques doivent être prises en même temps dans le cas où la scène observée est dynamique. Le rendu 3D d'une scène est obtenu en se référant aux trois étapes fondamentales suivantes : le calibrage, la mise en correspondance ou l'appariement et la reconstruction 3D par triangulation. Il est crucial de bien choisir les paramètres du modèle géométrique (entraxe, focales, etc.) du capteur stéréoscopique parce qu'on se heurte au dilemme suivant :

– Plus l'entraxe, défini par la distance entre les deux centres optiques des deux caméras, est faible et les axes optiques parallèles, plus le champ visuel couvert communément par les deux caméras est grand, et plus les images sont proches au sens de l'appariement des indices.

– Plus les caméras sont écartées, meilleure est la reconstruction tridimensionnelle obtenue.

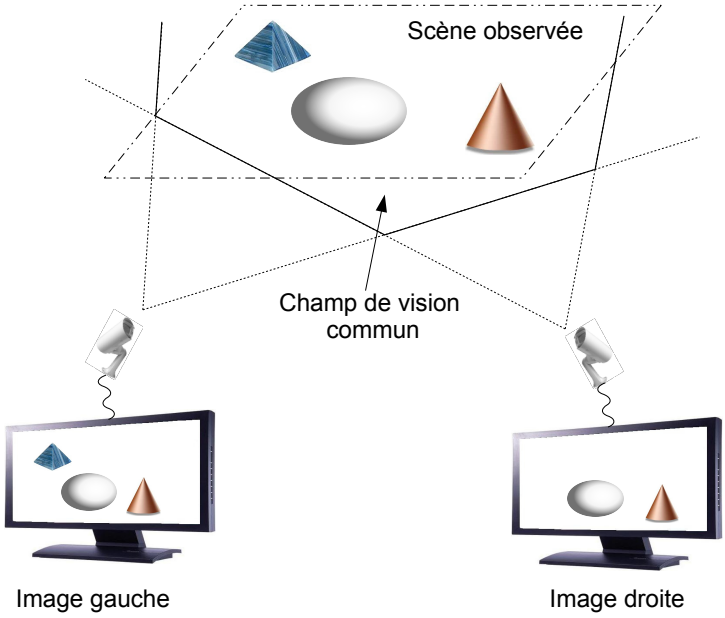


FIGURE 1.3 – Principe d'un système de vision stéréoscopique.

En plus de la vision binoculaire, d'autres architectures existent pour l'estimation des distances 3D, telles que les systèmes n-oculaire. Le choix du nombre et de la position des capteurs dépend des performances et de la précision souhaitées. L'augmentation du nombre de caméras permet de lever les ambiguïtés d'appariement et d'accroître la précision de localisation tridimensionnelle en fournissant plusieurs positions différentes pour chaque point de la scène observée. Calculer la position 3D d'un point réel à partir d'un capteur stéréoscopique consiste à passer par les étapes suivantes :

– Définition de l'architecture géométrique du capteur stéréoscopique (calibrage).

– Appariement des primitives des images stéréoscopiques.

– Reconstruction tridimensionnelle par triangulation géométrique.

L'utilisation de la vision stéréoscopique pour la perception de l'environnement est largement répandue dans différents domaines. Une des premières applications basée sur la stéréovision était la création d'un *oeil virtuel de contact* pour la téléconférence. [OLC93] a proposé un système de synthèse de vues qui consiste en la création de nouvelles vues à partir de deux images de la même scène. D'autres applications des systèmes stéréoscopiques ont trouvé place pour la construction de cartes urbaines et en reconnaissance aérienne [CW81], [WMK$^+$08], [LG94], navigation pour la robotique [KOY00], [FJLV07], reconstruction d'objets [FL95], reconstruction dense dans des environnements intérieures [FCSS09] et extérieures [CRF$^+$08], sécurité routière [HRK03], [BCFG05], [DGL08], [SAHK09], vision sous marine [SSHN09].

Pour des raisons méthodologiques, les deux processus indissociables suivants seront étudiés : la géométrie du capteur/scène et le processus de mise en correspondance des images stéréoscopiques. Afin d'aborder l'aspect algorithmique du traitement des données visuelles, la connaissance de la configuration géométrique des capteurs et de la scène réelle s'avère indispensable.

1.3 Géométrie stéréoscopique

1.3.1 Géométrie épipolaire

Un système de vision stéréoscopique permet de mesurer les distances métriques et angulaires dans un espace tridimensionnel. Étant donné un système de vision composé de deux caméras observant la même scène de deux points de vue différents : supposant le *modèle sténopé* [2] des caméras, la scène observée est projetée linéairement sur les plans image des deux caméras pour former deux vues bidimensionnelles. En outre, la connaissance préalable du modèle géométrique des capteurs de vision est d'une importance capitale pour la réussite du processus d'interprétation des scènes et de l'interaction du système avec son environnement. Le principe du système d'acquisition est illustré sur la figure 1.4. L'image de chaque point de l'espace est obtenue par une projection perspective linéaire sur le plan image du capteur gauche, ou droit.

La projection du point P_r d'un objet réel sur les plans image (dits aussi plans de projection) des

2. Le modèle sténopé est une représentation linéaire de la projection perspective, les rayons lumineux convergeant en un seul point, correspondant au centre optique de l'objectif, avant de se projeter sur le plan image.

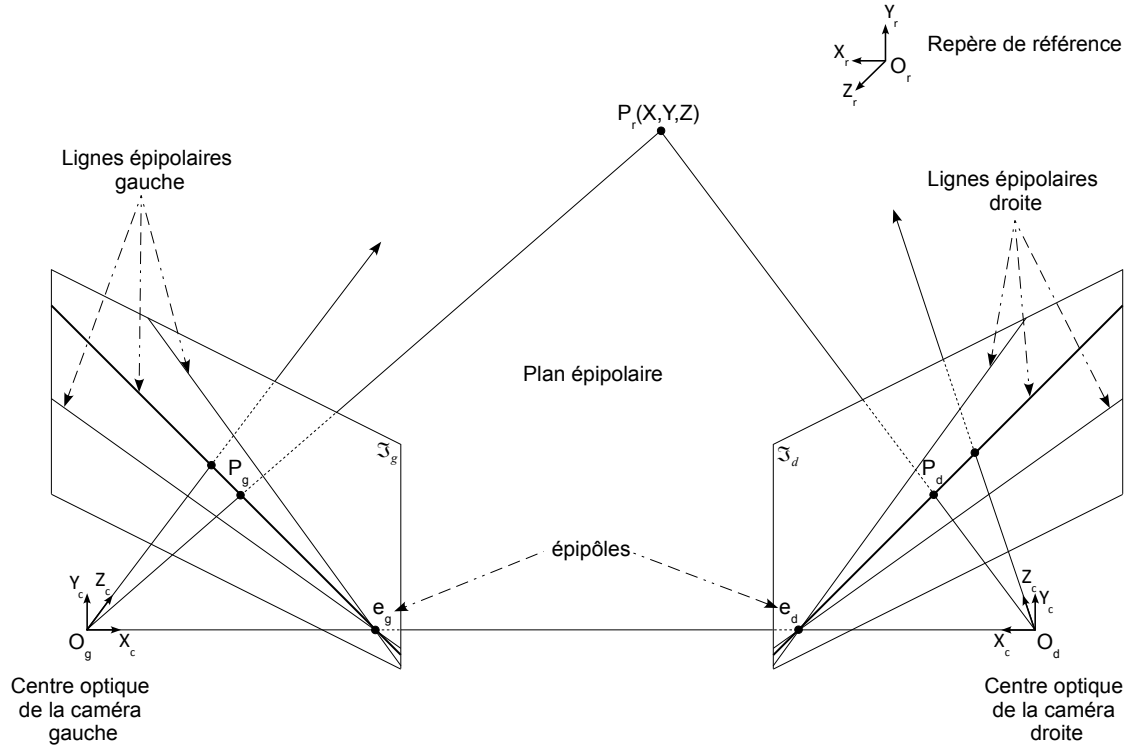


FIGURE 1.4 – Configuration d'un système de vision stéréoscopique.

caméras gauche et droite, est l'intersection du faiscau lumineux réfléchi par le point P_r et passant par les centres de projection (dits aussi centres optiques ou points focaux) des caméras gauche et droite, avec les plans image des caméras gauche et droite respectivement. La projection du point P_r sur les deux plans image gauche et droite donne naissance à deux points P_g et P_d, dite *points homologues*. Les points P_r, P_g et P_d forment un plan appelé *plan épipolaire*, qui passe aussi par les centres de projection, O_g et O_d, des deux caméras. Les deux points homologues ont des coordonnées relatives différentes, et l'écart de leurs positions dépend de la distance du point P_r par rapport au capteur (selon l'axe de profondeur). Plus un point de la scène est loin du capteur, plus la disparité, défini par l'écart entre la position des points homologues (§1.4.3), est réduit, et vice versa.

Le plan épipolaire coupe les plans image des caméras gauche et droite en une droite dite *droite épipolaire* gauche et droite respectivement. La notion de *droites épipolaires* est très importante dans la géométrie stéréoscopique, puisqu'elle introduit une première contrainte géométrique appelée contrainte épipolaire. Cette contrainte réduit la recherche des correspondants potentiels du point P_g de l'image gauche à ceux situés sur la droite épipolaire correspondante sur l'image droite, ce qui décroit considérablement l'espace de recherche et par conséquent réduit le temps de mise en correspondance. Chaque point de la scène appartient à une droite épipolaire sur les plans image de chaque caméra.

L'intersection de l'ensemble des droites épipolaires d'un plan image s'appelle *épipôle*, ou *point épipolaire*, soit e_g et e_d pour les épipôles gauche et droit respectivement. L'épipôle peut être défini aussi par la projection du centre optique d'une caméra sur le plan image de l'autre caméra. Par ailleurs, les épipôles e_g et e_d sont obtenus aussi par l'intersection du segment de droite $[O_g, O_d]$ avec les plans image des caméras gauche et droite respectivement. La distance entre les deux centres optiques, défini par le segment de droite $[O_g, O_d]$, est appelée *entraxe*.

1.3.2 Calibrage

Le calibrage est une étape cruciale pour la réussite du processus de reconstruction tridimensionnelle. Le calibrage consiste à modéliser l'ensemble du processus de vision du côté mécanique et optique. Le côté mécanique consiste à établir la relation entre la disposition et l'orientation des capteurs par rapport à la scène observée, alors que le côté optique consiste à modéliser le processus permettant la projection de la scène réelle sur les plans de projection de chaque caméra.

Pour illustrer l'importance du calibrage, nous prenons l'exemple d'une ligne droite dans l'espace objet qui peut être représentée par une courbe sur le plan de projection. Cette transformation est dûe aux déformations projectives lors du processus d'acquisition. Sachant que nous ne disposons que des plans de projection (deux courbes pour l'exemple précédent), il s'agit d'estimer les paramètres des transformations géométriques, éventuellement les matrices de translation, de rotation et de changement d'échelle, permettant de retrouver la forme et la position réelle de l'objet réel. Ceci est réalisé en faisant le passage du repère objet dans $\mathbb{R}^3$ aux repères des plans de projections des caméras gauche et droite, exprimé dans $\mathbb{R}^2$. L'établissement des relations de passage entre les différents repères permet de retrouver les coordonnées de chaque point de la scène observée à partir de ses projections. Le passage entre les différents repères est caractérisé par deux matrices :

- **La matrice essentielle**, décrivant les paramètres de passage du repère objet (*cf.* figure 1.4) vers le repère caméra, ceci en se basant sur les paramètres extrinsèques qui sont la disposition et l'orientation des capteurs l'un par rapport à l'autre selon le repère de référence,

- **La matrice fondamentale**, décrivant les paramètres permettant le passage des coordonnées des points exprimés selon le repère caméra vers ceux exprimés selon le repère image. Ceci ne dépend que des paramètres intrinsèques tels que la focale, la taille des photosites, etc. Ces paramètres ne dépendent pas de la configuration du stéréoscope mais uniquement des caractéristiques propres à chaque caméra.

Avec un système calibré, il serait possible, à partir de deux points homologues, d'estimer

l'angle obtenu par l'intersection des deux rayons lumineux émis par le point correspondant (point de la scène), et par conséquent d'estimer la position de ce point selon le repère objet (repère de référence). De plus, la rectification des aberrations optiques des objectifs des caméras s'avère une étape importante. Différentes techniques ont été proposées afin de rectifier les paires d'images stéréoscopiques [PD96], [FTVV00]. La plupart des méthodes utilisent des mires : un ensemble de points formant un damier, dont les positions relatives sont connues a priori. Le damier est placé à une certaine distance des capteurs. Il s'agit alors d'estimer les transformations géométriques inverses qui permettent de prendre en compte la distorsion pour redresser les lignes de l'image [Tsa86].

1.3.3 Configuration particulière

Nous allons voir dans la section suivante que la position tridimensionnelle d'un point objet peut être estimée à partir de ses projections sur les plans image de chaque caméra. Le problème consiste à trouver les couples de points homologues dans les deux images gauche et droite. Sans la connaissance de la géométrie du capteur stéréoscopique, la recherche des points homologues serait un processus lent à cause de la complexité algorithmique du problème. Etant donné un pixel de l'image gauche, son homologue peut être un pixel quelconque de l'image droite. D'où l'importance de la géométrie épipolaire telle que, dans le cas d'une configuration quelconque, un point sur le plan de projection de la caméra gauche appartient systématiquement à une ligne épipolaire. Son point homologue sur le plan de projection de la caméra droite doit appartenir à la ligne épipolaire correspondante. Les équations des lignes épipolaires conjuguées peuvent être calculées lors du processus de calibrage. Ceci réduit le temps de recherche des points homologues et simplifie, entre autres, la tâche de la mise en correspondance.

La géométrie stéréoscopique présente l'avantage de pouvoir passer d'une configuration quelconque à une configuration particulière appelée *configuration parallèle*. Une transformation linéaire en coordonnées projectives, dite aussi rectification épipolaire, permet ainsi d'avoir des lignes épipolaires parallèles et confondues avec les lignes de l'image, et des plans de projections coplanaires et parallèles à la droite O_gO_d passant par les centres de projection des caméras. Les épipôles sont alors projetés à l'infini. L'avantage de cette configuration est que l'homologue d'un pixel de l'image gauche se trouve sur la même ligne dans l'image droite, ce qui réduit considérablement le temps de mise en correspondance.

1.3.4 Stéréo-triangulation

La stéréo-triangulation consiste à retrouver les coordonnées tridimensionnelles d'un point objet à partir de ses projections sur les plans image d'une part, et des paramètres intrinsèques et extrinsèques des caméras d'autre part. Nous détaillons dans cette section les équations permettant

d'établir la relation entre ces différents paramètres. La position et l'orientation de chaque caméra peuvent être identifiées selon un repère de référence noté (O_r, X, Y, Z). Dans notre illustration, le centre de projection de la caméra gauche O_g est choisi comme origine du repère de référence. Nous allons détailler le cas où les caméras sont pivotées autour de leur axe Y, ceci représente un système réaliste puisque la rotation des caméras autour de l'axe X et Z n'a pas d'influence significative sur la précision de l'estimation de la distance réelle. Ainsi, la rotation d'une caméra autour de l'axe X peut se traduire par un décalage dans le sens vertical de l'un des plans de projection. La figure 1.5 illustre un capteur stéréoscopique dont les caméras gauche et droite forment un angle α et θ autour de leur axe Y respectivement. Les axes optiques se coupent en un point V dit *point de fixation*. Les axes optiques forment un angle appelé *angle de vergence*.

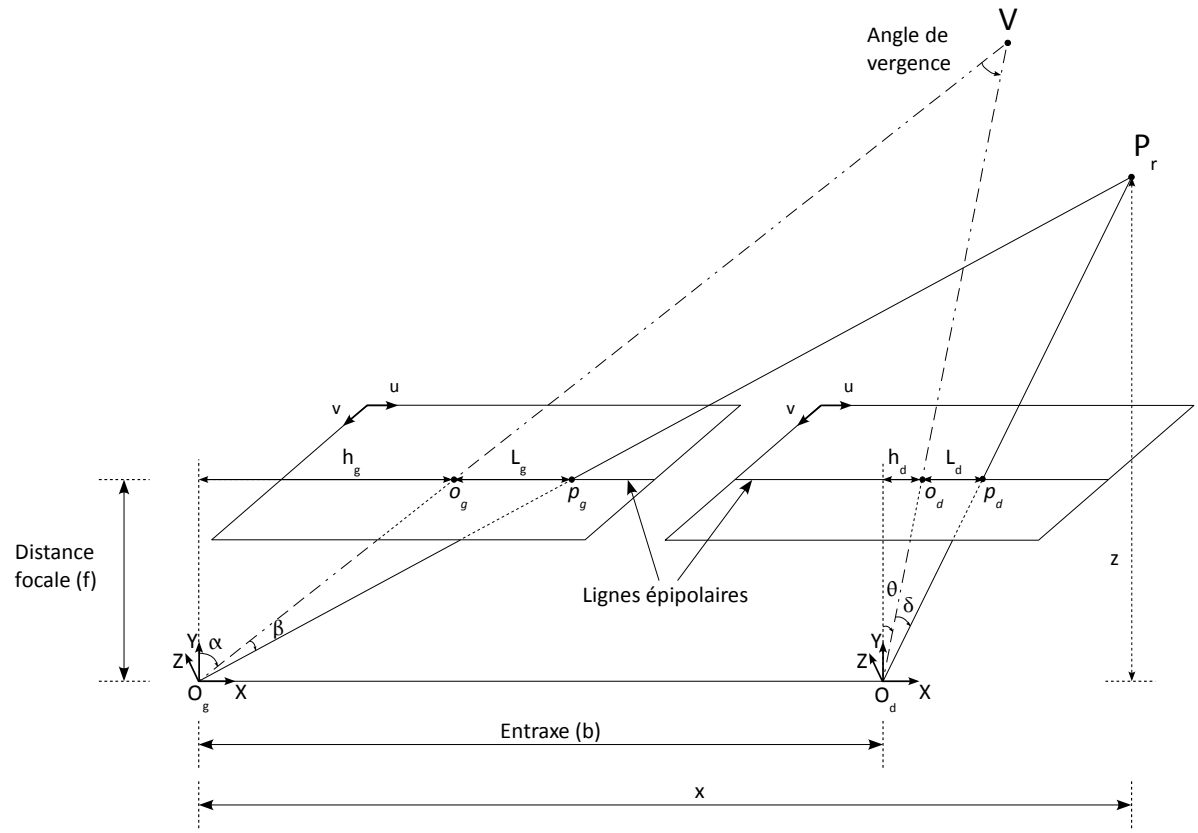


FIGURE 1.5 – La géométrie épipolaire et la stéréo-triangulation.

Soit A le segment $[O_g P_r]$, nous pouvons écrire les équations 1.1 et 1.2 comme suit :

$$\cos\left(\frac{\pi}{2} - (\alpha + \beta)\right) = \frac{x}{A} \tag{1.1}$$

$$\sin\left(\frac{\pi}{2} - (\alpha + \beta)\right) = \frac{z}{O_g P_r} \tag{1.2}$$

$$z\cos\left(\frac{\pi}{2} - (\alpha + \beta)\right) - x\sin\left(\frac{\pi}{2} - (\alpha + \beta)\right) = 0 \tag{1.3}$$

$$z\sin(\alpha + \beta) - x\cos(\alpha + \beta) = 0 \tag{1.4}$$

De la même façon nous obtenons :

$$z\cos\left(\frac{\pi}{2} - (\theta + \delta)\right) - (x - b)\sin\left(\frac{\pi}{2} - (\theta + \delta)\right) = 0 \tag{1.5}$$

$$z\sin(\theta + \delta) - (x - b)\cos(\theta + \delta) = 0 \tag{1.6}$$

En se référant à l'équation 1.4, la composante z s'écrit :

$$z = \frac{x}{\tan(\alpha + \beta)} \tag{1.7}$$

De même pour l'équation 1.6 :

$$z = \frac{(x - b)}{\tan(\theta + \delta)} \tag{1.8}$$

A partir des équations 1.7 et 1.8, nous obtenons la composante x :

$$x = \frac{b.\tan(\alpha + \beta)}{\tan(\alpha + \beta) - \tan(\theta + \delta)} \tag{1.9}$$

En remplacant 1.9 dans 1.7, nous obtenons la composante z :

$$z = \frac{b}{\tan(\alpha + \beta) - \tan(\theta + \delta)} \tag{1.10}$$

D'un autre coté, $\tan(\alpha + \beta)$ peut s'exprimer comme suit :

$$\tan(\alpha + \beta) = \frac{h_g + L_g}{f} \tag{1.11}$$

De même pour $\tan(\theta + \delta)$:

$$\tan(\theta + \delta) = \frac{h_d + L_d}{f} \tag{1.12}$$

Les équations 1.9 et 1.10 peuvent être simplifiées en considérant la configuration particulière des capteurs stéréoscopiques avec des axes optiques parallèles. Les angles α et θ deviennent nuls, de sorte que $\alpha = \theta = 0$ et $h_g = h_d = 0$. La position du point objet P_r ne peut être estimée à partir des angles β et δ puisqu'ils sont inconnus a priori. L'équation 1.11 peut être exprimée en fonction des coordonnées des pixels homologues p_g et p_d, les points de projection du point P_r sur les plans de projection des caméras :

$$\tan(\beta) = \frac{L_g}{f} \tag{1.13}$$

L'équation 1.12 devient :

$$\tan(\delta) = \frac{L_d}{f} \tag{1.14}$$

La composante z devient :

$$z = \frac{b}{\tan(\beta) - \tan(\delta)} \tag{1.15}$$

En remplaçant les équations 1.11, 1.12 dans 1.13, nous obtenons :

$$z = \frac{b.f}{L_g - L_d} \tag{1.16}$$

A partir des équations 1.9, 1.10 et 1.13, la composante x s'écrit ainsi comme suit :

$$x = L_g.\frac{z}{f} \tag{1.17}$$

La composante y est donnée par l'équation suivante :

$$y = y_g.\frac{z}{f} \tag{1.18}$$

Le paramètre y_g représente la position de la projection du point P_r sur le plan image du capteur gauche selon l'axe Y.

Dans ce qui précède, nous avons vu que la connaissance de la géométrie d'un tel système stéréoscopique est un élément clef de la vision tridimensionnelle. Le calibrage, la rectification épipolaire et le passage vers une configuration particulière sont des processus qui ne se font qu'une seule fois durant la chaîne de traitement. Les équations 1.16, 1.17 et 1.18 montrent que les coordonnées d'un point réel peuvent être retrouvées à partir de l'entraxe, la focale, et les coordonnées de la projection du point P_r sur les plans image des caméras gauche et droite. Etant donné que l'entraxe et la focale restent les mêmes pour une configuration donnée, le seul paramètre à estimer est la position des projections du point P_r sur le plan de projection de chaque caméra. À ce niveau, le problème peut se reformuler alors comme la recherche des points de projection homologues d'un même point de la scène observée, ceci en se référant seulement aux deux images gauche et droite qui correspondent aux plans de projection des caméras gauche et droite respectivement.

1.4 Appariement stéréoscopique

1.4.1 Principe d'appariement stéréoscopique

La mise en correspondance stéréoscopique, dite aussi appariement stéréoscopique, est l'étape la plus délicate du processus de reconstruction tridimensionnelle. Il s'agit de chercher les points de projection homologues d'un même point de la scène observée dans les images gauche et droite. La recherche des points homologues est un processus naturellement lent à cause de l'espace de recherche important. Ce problème est toujours posé en vision stéréoscopique malgré les progrès technologiques considérables. Nous commençons par définir la notion d'image, qui est une représentation numérique du plan de projection d'une caméra. Le signal électrique contenu dans chaque photosite du capteurs est transformé en un signal numérique pouvant être visualisé et manipulé.

Chaque pixel de l'image contient une information visuelle, l'intensité et la couleur. Un pixel se caractérise par sa position dans l'image et sa valeur photométrique. L'appariement consiste à évaluer le degré de ressemblance entre deux pixels à mettre en correspondance. Faire correspondre deux pixels revient à évaluer le degré de ressemblance entre les vecteurs d'attribut de chacun, tels que les caractéristiques spatiales, photométriques, et colorimétriques. Chaque pixel à apparier de l'image gauche a un ensemble de pixels candidats, noté $\mathcal{P}_d$, dans l'image droite. Seul le pixel candidat qui optimise une certaine métrique est retenu comme homologue. Le choix de la métrique, dite aussi fonction de vraisemblance, dépend de l'algorithme d'appariement utilisé. Elle est définie comme suit :

$$\begin{aligned} &f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \\ &\Phi\left(I_g(u,v), I_d(u-i, v-j)\right) \mapsto E_{app} \end{aligned} \tag{1.19}$$

Òu Φ désigne la fonction de vraisemblance utilisée pour l'évaluation de la qualité d'appariement de deux pixels candidats. La fonction de vraisemblance évalue la qualité de l'appariement des deux pixels $I_g(u,v)$ et $I_d(u-i, v-j)$. Chaque pixel correspond à un vecteur d'attributs de taille n. Le résultat de l'évaluation est un scalaire noté, E_{app} dans $\mathbb{R}$, mesurant le degré de ressemblance des données d'entrée. E_{app} correspond au coût local de la mise en correspondance obtenue avec une fonction de vraisemblance donnée, ou une énergie obtenue dans le cadre d'une fonction d'optimisation globale. Dans le cas de la configuration à axes parallèles, les pixels candidats se trouvent sur la même ligne de l'image que le pixel à faire correspondre, ce qui réduit l'espace de recherche et accélère le processus d'appariement.

1.4.2 Notations

Pour des raisons de clarté, nous décrivons ci-après les notations qui seront utilisées dans le reste du mémoire. L'image gauche est notée par I_g et l'image droite par I_d. Comme précisé sur la figure 1.6, I_g et I_d comprennent N lignes et M colonnes. Pour un pixel p de I_g à apparier, l'ensemble des pixels candidats dans l'image droite forme ce que nous appelons *Support* $\mathcal{S}_p$. Nous avons vu dans la section 1.4.1 que la fonction de vraisemblance ne tient compte que des caractéristiques photométriques ou colorimétrique du pixel à apparier. Le voisinage du pixel à apparier n'était pas pris en compte lors de l'appariement. Le fait de se limiter au niveau pixel pour établir la corrélation entre deux pixels donnés, constitue une limite du fait qu'un pixel à apparier appartenant à une région de couleur uniforme : les candidats ont des caractéristiques photométrique ou colorimétrique très similaires. Par ailleurs, le voisinage du pixel à apparier est souvent utilisé par la fonction de vraisemblance. C'est le principe des méthodes locales d'appariement détaillées dans la section 1.6.3. L'ensemble des pixels $\mathcal{N}_q$ voisins du pixel $q = I_d(u, v')$ forme une *zone d'agrégation*, notée ZA. Dans le cas d'une zone d'agrégation de forme carrée, de coté C et centrée sur le pixel candidat, la zone d'agrégation est défini par 1.20 :

$$\mathcal{N}_q = \{I_d(u+i, v'+j)/i, j \in [-\frac{C}{2}, +\frac{C}{2}]\} \tag{1.20}$$

Comme mentionné sur la figure 1.6, l'origine du repère de l'image est supposée placée au coin supérieur gauche de l'image, l'axe x pour les colonnes et l'axe y pour les lignes.

- $I_{g\vee d}(u, v)$: La représentation de base d'un pixel situé à la ligne u et la colonne v de l'image gauche I_g ou droite I_d. Cette représentation désigne l'intensité ou le niveau de gris du pixel.

- $I_{g\vee d}^{c\in RVB}(u, v)$: Il s'agit d'une des composantes de l'espace couleur RVB du pixel de coordonnées (u, v) de l'image gauche ou droite. À titre d'exemple, $I_g^{R\in RVB}(u, v)$ représente la composante rouge du pixel de coordonnées (u, v) de l'image gauche I_g. En plus de cette notation, nous introduisons dans les chapitres suivants les notations p, q, ou s pour désigner un pixel.

- $\mathcal{P}_{g\vee d} \equiv \mathcal{P}_g$ ou $\mathcal{P}_d$: Désigne l'ensemble des pixels de l'image gauche ou droite.

- $\#\mathcal{P}_{g\vee d}$: Désigne le nombre de pixels dans l'ensemble $\mathcal{P}_{g\vee d}$ de l'image gauche ou droite.

- $\mathcal{N}_q$: Désigne l'ensemble des pixels contenus dans la Zone d'Agrégation ZA centré sur le

pixel q. La ZA correspond à l'ensemble de pixels pris en compte par une fonction de vraisemblance donnée lors de l'établissement d'un score de corrélation entre deux pixels.

– $\#\mathcal{N}_q$: Désigne le nombre de pixels de l'ensemble $\mathcal{N}_q$.

– $\mathcal{S}_p$: Désigne l'ensemble des pixels contenus dans le Support $\mathcal{S}$ regroupant les pixels candidats dans l'image droite pour un pixel p à apparier de l'image gauche.

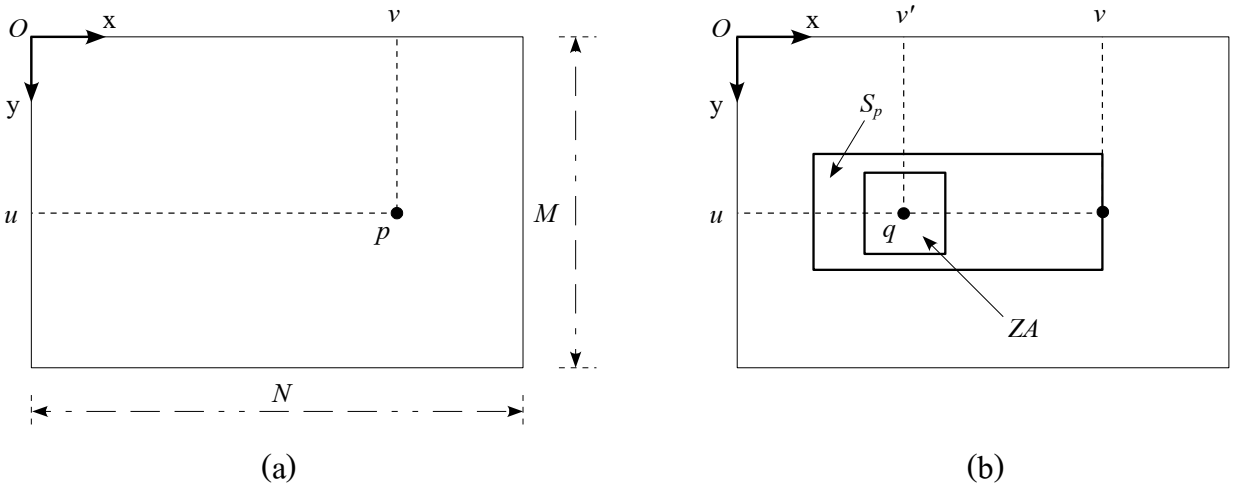


FIGURE 1.6 – Principe de recherche de pixels homologues (a) image gauche de taille $N \times M$. Le pixel à apparier est noté par p dans l'image de référence (image gauche) (b) image droite : les pixels candidats sont regroupés dans la région $Support\ S_q$. Le voisinage d'un pixel candidat q est regroupé dans la région $\mathcal{N}_{\mathrm{II}}$.

1.4.3 La disparité

La quantité $(L_g - L_d)$ introduite dans l'équation 1.16 est appelée *disparité*. Le terme "*disparité*" a été introduit initialement dans le domaine de la vision humaine pour décrire la différence de position de la projection d'un même point objet sur la rétine de chaque œil. La disparité est inversement proportionnelle à la distance de l'objet, de sorte qu'un point proche du capteur correspond à une grande valeur de disparité et inversement.

Définition : *La disparité, notée "disp", est la différence de position entre la projection d'un même point de la scène observée sur les plans image des caméras gauche et droite.*

$$\begin{aligned} & d : \mathbb{N} \rightarrow \mathbb{R}^{+} \\ & I_g(u,v), I_d(u,v') \mapsto disp\left(I_g(u,v)\right) = |v - v'| \end{aligned} \tag{1.21}$$

où v' est l'abscisse du pixel $I_d(u,v')$ homologue au pixel $I_g(u,v)$. Dans la configuration à axes parallèles, deux pixels homologues se trouvent sur la même ligne dans chaque image, mais à des positions horizontales différentes. Dans ce cas, la disparité correspond à la valeur absolue de la différence des positions horizontales de deux pixels donnés.

Une erreur sur cette mesure conduit à une mauvaise estimation de la composante z. Ce point sera abordé dans le chapitre suivant, dans la partie dédiée à l'étude de l'impact de l'imprécision de quelques paramètres dans la localisation tridimensionnelle. Une carte de disparité est une représentation 3D de l'image de référence. À chaque pixel $I_g(u,v)$ est ajoutée une troisième composante qui correspond à la distance 3D : d'où la nouvelle notation $I_g(u,v,disp)$. Dans le cas général, le déplacement vertical de la disparité est ajouté et la notation sera $I_g(u,v,disp_h,disp_v)$, où $disp_h$ et $disp_v$ correspondent respectivement aux disparité horizontale et verticale. Une autre représentation *multi-valeurs* de la disparité consiste à attribuer un vecteur de valeurs à chaque pixel $I_g(u,v,disp_1,\ldots,disp_n)$. Il s'agit des configurations multi-vues : un appariement stéréoscopique est appliqué sur chaque paire d'images. Pour un pixel de l'image de référence, chaque disparité $disp_i$, où $i \in [1,\ldots,n]$, correspond au décalage entre le pixel dans l'image de référence et son homologue dans une image donnée.

1.4.4 Contexte et enjeux

Comme mentionné précédemment, l'appariement stéréoscopique reste le problème majeur pour la réussite de la vision en relief. Malgré des avancées qualitatives significatives, l'appariement stéréoscopique reste une tâche fastidieuse. La compréhension des contraintes que peuvent rencontrer les chercheurs en abordant le problème de la mise en correspondance est d'une importance capitale pour le développement d'algorithmes efficaces. Les problèmes liés à l'appariement stéréoscopique peuvent être résumés dans les points suivants :

– ***Variation d'illumination*** : En vision stéréoscopique, un point objet est représenté par deux pixels différents, un dans l'image gauche et l'autre dans l'image droite. Les caractéristiques photométriques peuvent varier d'une image à l'autre. La variation d'illumination locale ou globale dépend des conditions de prise de vue. Ceci peut être dû à la variabilité des paramètres intrinsèques des deux caméras, dans le cas d'une prise de vue instantanée, ou des variations d'illumination de la scène, pour une prise de vue différée. Les variations locales

d'intensité peuvent être dûes à la présence d'un éclairage variable à la présence des réflexions spéculaires, ou à cause du vignetage. Les variations globales d'intensité sont généralement dues à la différence de gain ou du temps d'exposition du capteur CCD des caméras [HS07]. La cohérence photométrique et colorimétrique entre les deux vues reste la principale contrainte utilisée pour la mise en correspondance. La vérification de la cohérence photométrique globale fait l'objet de plusieurs travaux [MPP09].

– ***Occultation*** : un pixel est dit occulté, s'il n'est visible que par une seule caméra. La différence entre les positions de la prise de vue d'une scène est l'un des facteurs d'occultations. Les pixels occultés situés dans des régions de discontinuité de profondeur peuvent posé de sérieux problèmes pour certaines applications, telles que la segmentation basée sur les disparités [MG04], et la génération de nouvelles vues à partir de deux vues [Sch99], [ZKU+04]. W. Xing [XCJ09] propose un cadre assez général pour la prise en compte des occultations. Il modélise à la fois les pixels totalement occultés et non occultés.

– ***Texture répétitive*** : il s'agit d'une autre source d'ambigüité. Chaque pixel à apparier a plusieurs candidats sur l'autre image. Plus le vecteur caractéristique d'un pixel est discriminant, plus le nombre de pixels candidats dans l'autre image est restreint, et inversement. L'extraction des pixels discriminants à apparier n'est pas toujours évidente. Le cas particulier d'une texture répétitive monodirectionnelle est connu sous le nom de problème d'ouverture (la texture est répétée tout au long des lignes épipolaires). C'est le cas, par exemple, d'une scène contenant un mur en briques.

– ***Région uniforme peu texturée*** : Dans certaines applications, l'estimation des reliefs semble très contraignante par la nature de la scène observée. À titre d'exemple, la reconstruction tridimensionnelle d'une route n'est pas toujours évidente à cause de l'ambigüité d'appariement. L'homologue d'un pixel de l'image gauche appartenant à une région uniforme non texturés, appartient lui aussi à une région non texturée de l'image droite. L'ambigüité d'appariement est due à la multitude des primitives potentiellement candidates. Les techniques d'appariement locales sont souvent insuffisantes et sont couplées avec des techniques globales.

1.5 Les contraintes

Avant de détailler l'aspect algorithmique du processus de la mise en correspondance, nous adordons maintenant un autre aspect important : la notion de contrainte.

Définition : *Une contrainte est une hypothèse introduite lors du processus de la mise en correspondance afin de restreindre l'espace de recherche des pixels candidats ou de réduire les ambiguïtés d'appariement.*

L'introduction des contraintes permet de lever certaines ambigüité d'appariement. Le choix de la contrainte dépend de l'application envisagée. Plusieurs travaux ont utilisé une ou plusieurs contraintes à la fois afin d'éliminer les mauvaises correspondances. Nous proposons de répartir les contraintes en trois classes en fonction de la nature des hypothèses prises. Ces hypothèses concernent la géométrie de la scène observée, la géométrie du capteur stéréoscopique, et d'autres hypothèses d'ordre local, telles que la vraisemblance spatiale et colorimétrique.

1.5.1 Contraintes sur la géométrie du capteur stéréoscopique

Il s'agit de la contrainte épipolaire : Cette contrainte exploite un aspect important de la géométrie particulière du capteur stéréoscopique qui, après quelques transformations géométriques, permet d'obtenir une configuration consistant à avoir des lignes épipolaires parallèles. En terme de mise en correspondance, ceci se traduit par le fait que le correspondant d'un pixel situé sur la ligne u de l'image gauche, est situé aussi sur la même ligne u de l'image droite. Au lieu de chercher partout dans l'image, cette contrainte réduit la recherche en une seule ligne. L'application de cette contrainte est capitale puisque le temps ainsi gagné est considérable comparé aux autres contraintes. Cette contrainte est largement utilisée dans la littérature [SR98], [SMM08], [ZJWB08].

1.5.2 Contraintes liées à la géométrie de la scène

Il n'existe pas une liste exhaustive des contraintes utilisées en stéréovision. Le choix et la définition d'une contrainte ou d'une autre dépend du contexte et de l'algorithme d'appariement choisi. Le but ici est de sensibiliser le lecteur à l'importance de la notion de contrainte dans le processus de mise en correspondance. Nous décrivons ci-après quelques unes des contraintes les plus utilisées en vision stéréoscopique.

1.5.2.1 Contrainte d'unicité

La contrainte d'unicité impose que la fonction disparité soit une bijection ou une injection. Dans le cas d'une bijection, un pixel de l'image gauche ne peut avoir qu'un et un seul pixel homologue sur l'image droite. La fonction de disparité est supposé injective s'il existe, pour un pixel de l'image droite, plus d'un pixel homologue dans l'image gauche, ceci peut être dû à la présence des régions occultées : C'est le cas d'un pixel de l'image gauche qui n'a pas d'homologue dans l'image droite. Cette contrainte est introduite dans plusieurs travaux [MP79], [STG03] et [BG07].

La contrainte d'unicité peut être formulée comme suit :

$$\forall\ I_d(u,j) \in \mathcal{P}_d, \exists!\ I_g(u,v) \in \mathcal{P}_g\ /\ disp\,(I_g(u,v)) = |v-j| \tag{1.22}$$

Cette contrainte peut ne pas être respectée dans le cas d'objets fortement inclinés, il s'agit d'une injection dans ce cas. Ceci est illustré sur la figure 1.7. Deux points de la scène observée, P_1 et P_2, ont la même projection sur l'image droite.

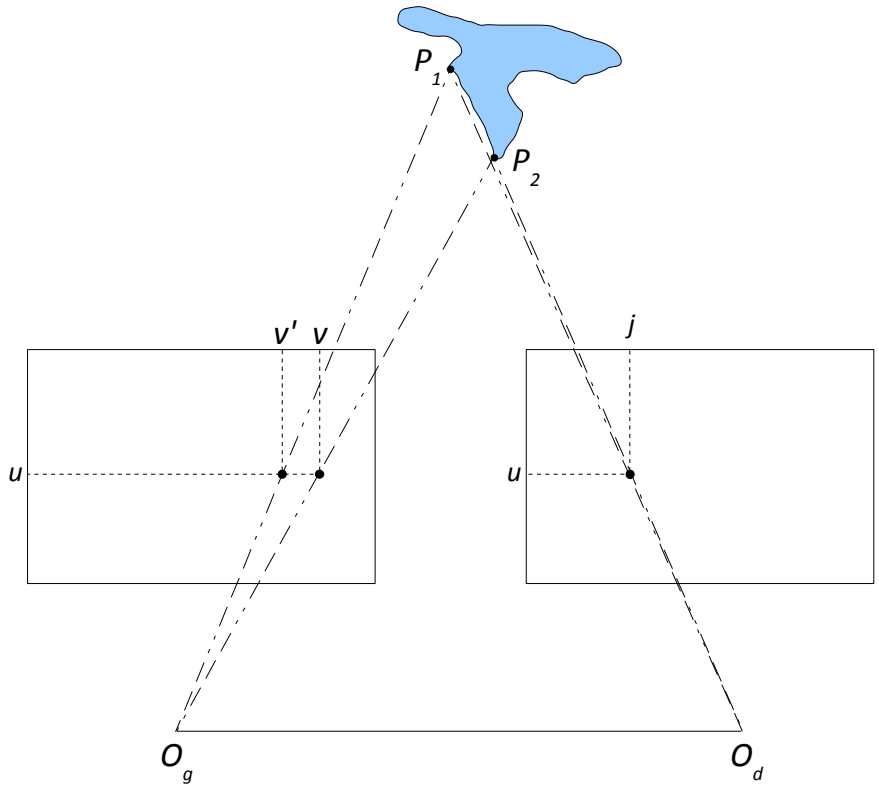


FIGURE 1.7 – Exemple de violation de la contrainte d'unicité. Il s'agit de deux points appartenant à un même objet dont une face est fortement inclinée par rapport au plan image de chaque caméra.

Dans le cas où la contrainte d'unicité n'est pas respectée, la formulation devient :

$$\begin{gathered}\forall\ I_d(u,j) \in \mathcal{P}_d, \exists\ I_g(u,v), I_g(u,v') \in \mathcal{P}_g\ /\ disp\,(I_g(u,v)) = |v-j|\ \ ; \\ disp\,(I_g(u,v')) = |v'-j|\ \ et\ \ v \neq v'\end{gathered} \tag{1.23}$$

1.5.2.2 Contrainte d'ordre

La contrainte d'ordre signifie que l'ordre d'appariement des pixels sur une ligne épipolaire doit être respecté. Cette contrainte n'est valide que pour les pixels appartenant à une même surface convexe.

$$\forall\ I_g(u,v), I_g(u,v') \in \mathcal{P}_g\ \ et\ \ v > v', \exists\ I_d(u,j), I_d(u,j') \in \mathcal{P}_d\ \ /\ \ disp\left(I_g(u,v)\right) = |v-j| \\ , disp\left(I_g(u,v')\right) = |v'-j'|\ ,\ \ et\ \ j > j'. \tag{1.24}$$

La figure 1.8 illustre un exemple de violation de la contrainte d'ordre. Il s'agit de deux points appartenant à deux objets différents, placés l'un derrière l'autre (à des distances différentes par rapport aux capteurs). La contrainte d'ordre peut aussi être violée dans le cas où la surface d'un même objet est concave. L'intégration de cette contrainte dépend de l'objectif ; elle peut être intégrée au processus d'appariement d'objets pour une reconstruction tridimensionnelle dense. La violation de cette contrainte permet de détecter la présence ou non d'occultations. Elle est intégrée dans certain travaux tels que [SR98], [LCCB01] et [SMM08].

1.5.2.3 Contrainte de symétrie ou test de vérification croisée

Cette contrainte est introduite pour la détection des occultations partielles. Un point de la scène observé est occulté s'il est visible par une caméra et invisible par l'autre. La vérification croisée consiste à établir deux cartes de disparités, en considérant l'image gauche puis l'image droite comme image de référence. Pour un point visible des deux caméras, la différence de disparités dans les deux cartes de disparité est égale à zéro. La contrainte de symétrie peut être formulée comme suit :

$$\forall\ I_g(u,v) \in \mathcal{P}_g, \exists\ I_d(u,j) \in \mathcal{P}_d\ \ / \left|disp\left(I_g(u,v)\right) - disp\left(I_d(u,j)\right)\right| = 0 \tag{1.25}$$

La figure 1.9 illustre un cas d'occultation : le point P_1 n'est visible que par la caméra gauche. La projection de ce point sur le plan de l'image gauche, soit le pixel $I_g(u,v)$, n'a pas d'homologue sur l'image droite. Son pixel homologue peut être estimé par une interpolation dans l'espace des disparités. La disparité obtenue dans ce pixel est supposée $disp_1$. Le point P_2, visible par les deux caméras, a une projection sur les deux plans de l'image gauche et droite, soit les pixels $I_g(u,v')$

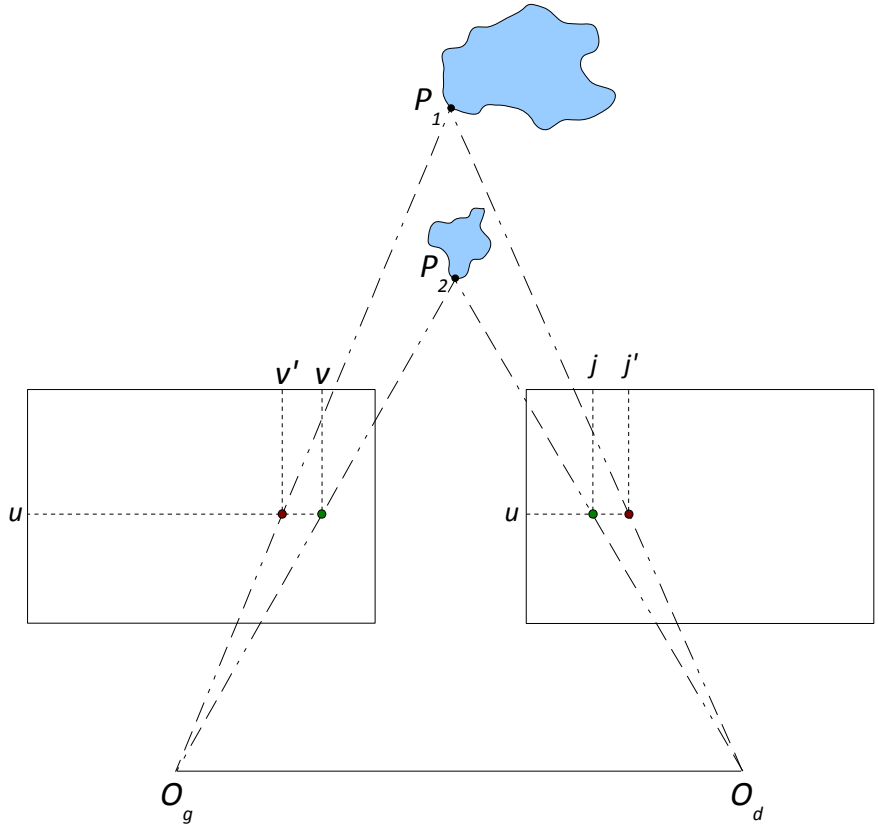


FIGURE 1.8 – Exemple d'une configuration dont la contrainte d'ordre est violée. Il s'agit de deux objets appartenant chacun à un objet de différent distance par rapport aux caméras.

et $I_d(u, j')$ respectivement. Soit $disp_2$ la disparité obtenue pour le point P_2. En appliquant le test de vérification croisée, les mesures de disparité ne sont pas les mêmes ($disp_1 \neq disp_2$). Un des points, P_1 ou P_2, est alors considéré comme occulté. La contrainte de symétrie est introduite dans plusieurs travaux tels que dans [LG94], [BT98], [LP06a], [LSY06], [ZGY08], [WWLHG08], [HBGR09], [LHYJ09], et [YWY+09].

1.5.2.4 Contrainte de visibilité

Cette contrainte permet de pénaliser les zones occultées. Elle se base sur la méthode Z-Buffer [BG04], qui exploite les deux vues afin d'estimer les régions occultées. Si une cellule du Z-buffer contient plus d'un pixel, seul le pixel avec la plus grande disparité est visible et les autres sont occultés dans le second point de vue. Cette contrainte est introduite dans [JSLKS05], [BG05], [MYS06], [MAW+07], [KKB07a], [KKB+07b], [FP07], [FP08].

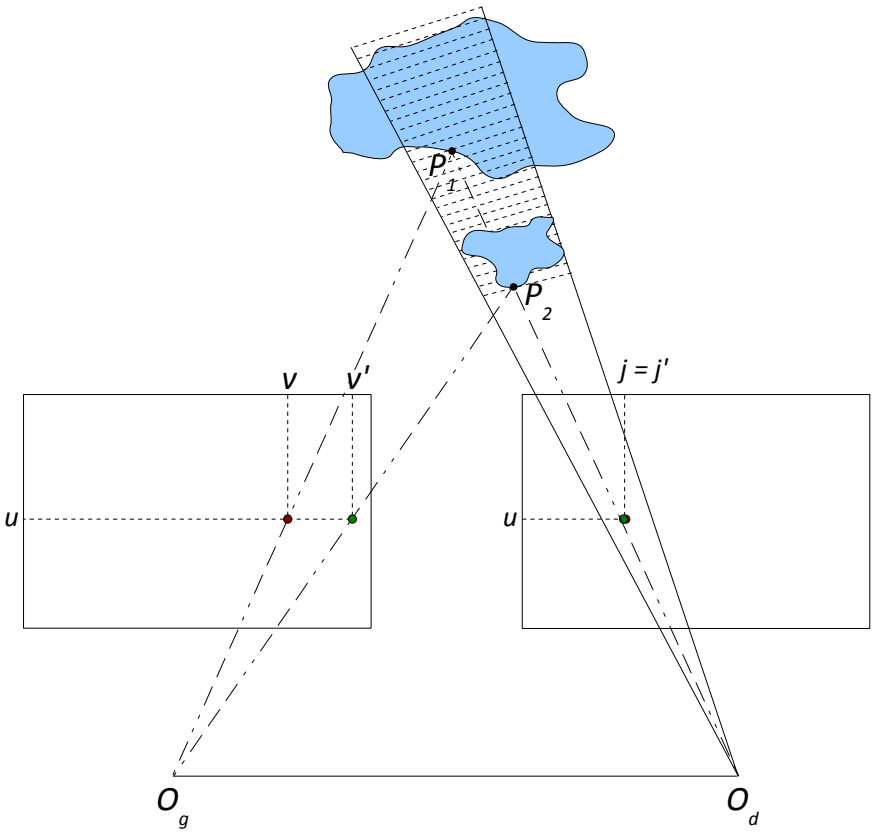


FIGURE 1.9 – Exemple de violation de la contrainte de vérification croisée, dûe à la présence d'occultation partielle : certains pixels ne sont vus que par une seule caméra.

1.5.2.5 Contrainte d'invariance photométrique ou colorimétrique

Il s'agit d'une des contraintes les plus utilisées dans les algorithmes de mise en correspondance. La mise en correspondance de deux pixels à partir de leur niveau de gris s'appuie sur l'hypothèse que les projections d'un même point M de l'espace tridimentionnel ont des intensités lumineuses similaires [RSC04], [MPP09]. Techniquement parlant, ceci est la propriété des surface lambertiennes : l'intensité lumineuse d'un point reste invariable au changement de vues. Si cette contrainte n'est pas vérifiée, les mesures de corrélations de points ou de motifs sur les intensités sont perturbées par des variations non ergodiques provoquées par des phénomènes tels que le changement de point de vue, les occultations partielles, l'échantillonnage, la numérisation, la transparence, etc. qui peuvent difficilement être modélisés par de simples lois normales. Nous citons quelques travaux récents intégrant la contrainte d'invariance photométrique [FP07], [HVC07], [SMP07], [VHTC07], [CVHC08], [FP08], [HKLP09], [MS09], [SSS09], [SZJ09], [KPC10]. D'autres contraintes existent dans la littérature, telles que la contrainte de continuité figurale [MMN89] [BHM05], la contrainte de disparité telle que la continuité et la disparité maximale [MPP09], la contrainte de rang [BB01] [RGS03] [ZHK+03], et la contrainte de limite de gradient [May03].

1.6 Diverses classifications des techniques d'appariement

La perception de la troisième dimension par vision stéréoscopique est l'un des domaines les plus explorés en vision artificielle. La diversité des méthodes d'appariement ne permet pas d'avoir une classification unique. L'appariement stéréoscopique n'est pas un processus stable à cause des problèmes discutés au §1.4.4 : à cause du problème d'occultation, la mise en correspondance n'est pas souvent une fonction non bijective ; un pixel d'une image peut ne pas avoir d'homologue dans l'autre image. [SS02] et [BBH03] ont proposé chacun une taxonomie des méthodes d'appariement. La classification dépend généralement de critères tels que la complexité algorithmique des méthodes d'appariement et la nature de la carte de disparités obtenus. En se basant sur la carte de disparités résultat, nous distinguons les *méthodes denses* des *méthodes éparses*. Les techniques d'appariement dense font correspondre tous les pixels de l'image de référence, ce qui permet une reconstruction tridimensionnelle de toute la scène observée. La synthèse d'images et la réalité virtuelle sont des exemples d'applications nécessitant l'appariement de l'ensemble des pixels de l'image. L'appariement épars consiste à apparier une partie des pixels, des primitives, de l'image de référence. Si nécessaire, la reconstruction de la scène nécessite dans ce cas un post-traitement pour la disparité des pixels ou des primitives non appariés, généralement par interpolation. Les méthodes éparses sont généralement utilisées pour des applications exigeant un traitement en temps réel, telles que la détection des obstacles routiers [HRK03], [BCFG05].

Une autre classification a été largement proposée dans la littérature, celle qui classe les méthodes d'appariement en des *méthodes locales* et *globales*. Les méthodes locales exploitent le voisinage du pixel à apparier sans tenir compte de l'information contenue dans le reste des images utilisées pour l'appariement. La notion de voisinage sera détaillée dans la section 1.6.3 des méthodes locales. Les méthodes globales considèrent l'ensemble des pixels de l'image pour l'estimation récursive de la disparité. Les différentes méthodes d'appariement globales seront détaillées dans la section 1.6.4. Nous proposons de classifier les méthodes d'appariement comme donné par la figure 1.10.

1.6.1 Méthodes d'appariement denses

Il s'agit de l'appariement de l'ensemble des pixels formant une image. Etant donné le peu d'information que contient un pixel, la mise en correspondance exploite l'information contenue dans son voisinage proche [BT80]. La vraisemblance de l'appariement de deux pixels est établie en mesurant la corrélation existante entre les voisins de chacun. La vraisembance peut être estimée par l'écart des valeurs photométrique ou colorimétriques des pixels voisinages correspondants. Cette famille de techniques suppose la continuité des surfaces dans la scène réelle, traduite par une continuité des disparités. Cependant, la réduction des ambigüités d'appariement peut répondre à l'un des deux objectifs suivant : l'élimination des grosses erreurs d'appariement pour l'obtention

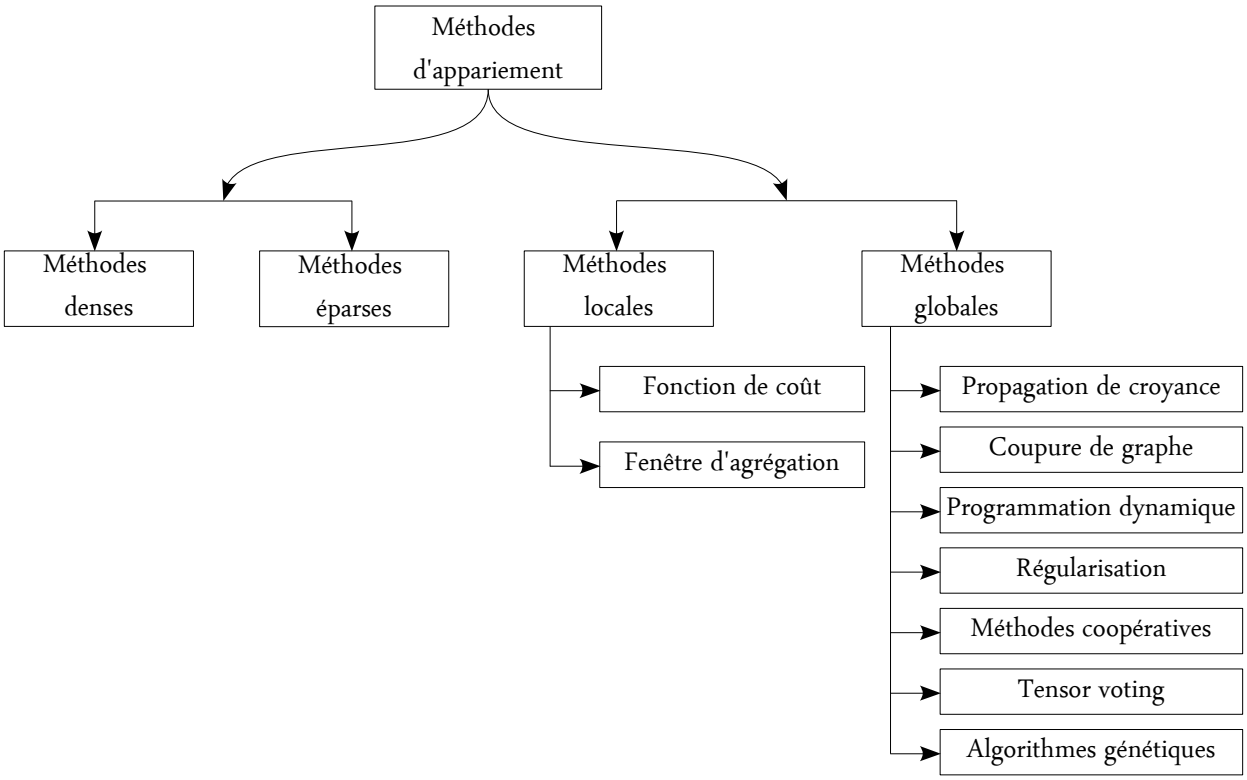


FIGURE 1.10 – Classification possible des méthodes d'appariement.

d'un rendu tridimensionnel macroscopique de la scène observée, ou l'amélioration de la précision de localisation donnant lieu à une reconstruction tridimensionnelle microscopique et très précise de la scène observée. Ceci est assuré par l'introduction de contraintes d'ordre global, qui conduit à accroître le temps d'appariement. L'amélioration de la précision de localisation est obtenue de différentes façons : l'estimation des disparités en "sub-pixel", la variation itérative de l'entraxe des caméras, ou l'exploitation de plusieurs vues de la scène. Voici quelques critères d'évaluation pour la comparaison des algorithmes de mise en correspondance :

- ***Précision*** : dont l'évaluation est effectuée en sub-pixel. L'évaluation est généralement obtenue à partir d'une carte de disparité de référence, dite vérité terrain. La précision concerne l'évaluation du potentiel des algorithmes en terme de lissage des surfaces continues, des zones occutées, et des régions de discontinuité de profondeur.

- ***Fiabilité*** : résistance d'un algorithme de mise en correspondance aux erreurs d'appariement.

- ***Adaptabilité*** : l'applicabilité d'un algorithme dans différentes scènes plus ou moins complexes.

– ***Prévisibilité*** : disponibilité des modèles de performances.

– ***Complexité algorithmique*** : regroupe le coût de l'implémentation sur différents plateformes, les exigences en ressources matérielles.

1.6.2 Méthodes d'appariement éparses

Les points d'intérêt ont la particularité d'être invariants aux changements d'échelle des images dûs à des transformations géométriques. L'appariement épars est utile dans certaines applications, telles que le calcul de la matrice fondamentale pour le calibrage du capteur stéréoscopique (coins d'une mire en damiers utilisée pour sa rapidité), ou pour la détection d'obstacles routiers par un stéréoscope embarqué. L'inconvénient des méthodes éparses réside dans l'impossibilité d'avoir une reconstruction 3D complète de la scène. L'appariement de certains points d'intérêt, considérés comme non ambigus, est adopté dans plusieurs algorithmes de mise en correspondance. Ces points correspondent aux primitives qui peuvent avoir un certain nombre d'attributs. Plus les primitives sont discriminantes, plus l'appariement est précis, robuste et fiable. Les facteurs suivants sont d'importance capitale dans la réussite de l'appariement des primitives :

– Dimensionnalité (pixel, contour, etc.).
– Contraste.
– Contenu sémantique.
– Densité de l'occurrence.
– Facilité de mesure des propriétés.
– Unicité.

Nous présentons ci-après les primitives les plus utilisées par les algorithmes d'appariement épars.

– *Les contours* : Dans le cas où le modèle des caméras est connu, un contour semble être une primitive intéressante pour l'appariement [Arn78], [GM79], [Bak80], [OK85a], [HA89], [SK07], [NB09], [ZZW09], [LMC10]. Un algorithme de détection de contours ne fournit pas une décision binaire sur l'existence ou non d'un élément de contour, mais plutôt une décision floue. Les attributs des contours peuvent être le contraste et la direction. Cette famille de techniques cherche généralement à estimer le passage par zéro d'une dérivée seconde (qui est équivalent à la recherche du maximum local d'une dérivée première) des intensités de l'image.

- Les mofifs [BN98]
- Les coins [BT99]
- Les courbes [SZ98], [FK10]
- *Autres points d'intérêts* : H. Moravec [Mor79] a proposé un opérateur permettant la détection des points d'intérêt portant son nom. L'opérateur de Moravec détecte les points dont la variance entre les pixels adjacents est forte dans les quatre directions. Cet opérateur a été amélioré par [Han80] qui a considéré le rapport des variances dans les quatre directions. Cette amélioration a permis de détecter des points caractéristiques plus complexes.

L'appariement de primitives ne fournit pas une carte de profondeur dense : le rendu tridimensionnel présente des trous qui ne sont pas toujours facilement remplis par de simples interpolations. Ceci dépend de la densité des primitives appariées. Le contenu sémantique des primitives pourrait être considéré comme un moyen d'interpolation. À titre d'exemple, les contours des immeubles, dans le cas d'une scène urbaine, semblent être une des primitives ayant un sens sémantique. La reconstruction tridimensionnelle dense peut être obtenue en exploitant la sémantique des primitives, ceci en supposant que l'espace intermédiaire entre les contours est occupé par des murs dont la surface est continue. Un autre exemple est celui des primitives qui correspondent aux marquages au sol d'une scène routière. La reconstruction de la surface de la route est obtenue en se basant sur l'hypothèse que la route et les marquages au sol ont des disparités similaires et continues.

1.6.3 Méthodes d'appariement locales

Les méthodes locales se basent sur le voisinage des pixels à apparier pour établir leur degré de corrélation. Pour des raisons de simplicité, les pixels à apparier sont limités aux pixels. Les attributs de chacun peuvent être interprétés en tant que scalaires, vecteurs, matrices ou tenseurs. Comme défini dans la section 1.4.2, le voisinage d'un pixel $p = I_{g \vee d}(u, v)$ est noté par $\mathcal{N}_p$. Chaque pixel voisin est obtenu par un simple déplacement, selon les coordonnées de référence du plan image, par $(u + d_u, v + d_v)$. Par abus de langage, on appelle "*fonction de vraisemblance*" toute fonction d'appariement permettant d'évaluer le degré de corrélation, de similarité ou de dissimilarité entre deux ensembles de données.

1.6.3.1 Choix de la fonction de vraisemblance

Le but de ce paragraphe n'est pas de décrire toutes les fonctions de vraisemblance existantes. Nous ne présenterons dans ce chapitre que quelques exemples de fonctions d'appariement afin

d'expliquer le principe des fonctions locales. Cet aspect sera détaillé dans le chapitre 2. Nous décrivons ci-après quelques fonctions de vraisemblance parmi les plus populaires permettant une évaluation de la corrélation entre deux régions : ϕ_{SAD} "***Sum of Absolute Differences***", ϕ_{SSD} "***Sum of Squared Differences***" et ϕ_{NCC} "***Normalized Cross Correlation***".

$$\phi_{SAD}(I_g(u,v), I_d(u,v')) = \sum_{(i,j)\in\mathcal{N}} |I_g(u+i, v+j) - I_d(u+i, v'+j)| \tag{1.26}$$

$$\phi_{SSD}(I_g(u,v), I_d(u,v')) = \sum_{(i,j)\in\mathcal{N}} \left(I_g(u+i, v+j) - I_d(u+i, v'+j)\right)^2 \tag{1.27}$$

$$\phi_{NCC}(I_g(u,v), I_d(u,v')) = \sum_{(i,j)\in\mathcal{N}} \frac{I_g(u+i, v+j).I_d(u+i, v'+j)}{\sqrt{\sum_{(i,j)\in\mathcal{N}} |I_g(u+i, v+j)|^2}.\sqrt{\sum_{(i,j)\in\mathcal{N}} |I_d(u+i, v'+j)|^2}} \tag{1.28}$$

Où $\mathcal{N}$ correspond au voisinage des pixels $I_g(u,v)$ et $I_d(u,v')$. Bien que ces fonctions soient de complexité algorithmique nettement inférieure à celles d'autres fonctions de vraisemblance, il reste à choisir entre les critères de temps d'exécution et de qualité d'appariement. Ces fonctions sont très sensibles aux valeurs aberrantes [BN97], en l'occurrence dans le cas des images réelles. Pour traiter cette situation, [Sebe00] a proposé d'introduire des propriétés statistiques pour décrire le bruit présent dans les images. Il a remplacé l'hypothèse du bruit Gaussien par la métrique de $Cauchy$, montrée plus performante que la mesure de $Kullback$. Pour deux pixels $p = I_g(u,v)$ et $q = I_d(u,v')$, la fonction de $Cauchy$ est donnée par :

$$\phi_{Cauchy}(I_g(u,v), I_d(u,v')) = \log\left[1 + \left(\frac{D(\mathcal{N}_p, \mathcal{N}_q)}{\chi}\right)\right] \tag{1.29}$$

Le terme χ contrôle les paramètres de la distibution de $Cauchy$, et D est la norme de la différence entre $\mathcal{N}_p$ et $\mathcal{N}_q$. Afin de gérer l'incertitude causée par le bruit, d'autres variantes des fonctions SAD et SSD ont été proposées : il s'agit des versions centrées et normalisées détaillées dans [Cha05].

Toutefois, un prétraitement peut être appliqué sur les données de la fonction de vraisemblance. En l'occurrence, l'intensité du signal est prétraitée par une transformation de Census non-paramétrique. Dans ce cas, seule une comparaison binaire est effectuée entre l'intensité de chaque pixel voisin et l'intensité d'un pixel de référence. Une distance est alors calculée entre deux séries binaires b_g et b_d. Diverses mesures de distance ont été proposées, telles que la distance de Tanimoto donnée par l'équation 1.30, et la distance de Hamming donnée par l'équation 1.31 [Cyg04] qui est, dans la pratique, rapide et offre les meilleures performances en terme de qualité d'appariement. Le symbole $\otimes$ utilisé dans l'équation 1.31 désigne l'opérateur ou-exclusif (XOR).

$$D_T(b_g, b_d) = \begin{cases} 1 & \text{si } b_g = b_d = 0 \\ \frac{b_g^T \; b_d}{b_g^T \; b_g \; + \; b_d^T \; b_d \; - \; b_g^T \; b_d} & \text{si non} \end{cases} \tag{1.30}$$

$$D_H(b_g, b_d) = \frac{1}{N} \sum_{i=1}^{N} b_{gi} \otimes b_{di} \tag{1.31}$$

[HS07] propose une évaluation de la sensibilité de quelques fonctions de vraisemblance sur les changements radiométriques ou en présence de bruit. L'auteur compare les fonctions ϕ_{NCC}, transformation des rangs et Census [ZW94], ainsi que d'autre fonctions plus complexes basées sur l'information mutuelle [Egn00], [Hir05], et [KKZ03]. Les expérimentations montrent qu'avec des changements radiométriques simulés et réels, la transformation des rangs donne les meilleurs résultats.

1.6.3.2 Choix de la forme de la fenêtre d'agrégation

La sortie d'une fonction de vraisemblance est un scalaire décrivant le degré de corrélation entre deux ensembles de données. La stratégie la plus souvent utilisée pour la sélection du meilleur candidat est intitulée "*Winner-Take-All*". Le candidat qui optimise la fonction de vraisemblance, est choisi comme le meilleur candidat. La disparité est alors calculée comme l'écart entre les positions du pixel à apparier et le pixel homologue ayant obtenu le meilleur score.

Le choix de la fonction de corrélation est un élément important pour la réussite du processus d'appariement : plus la fonction choisie est discriminante, plus les ambigüités d'appariement seront réduites. La forme de la fenêtre d'agrégation des fonctions décrites précédemment est limitée

à une fenêtre carrée de taille fixe, centrée sur le pixel à apparier [OK93]. Plusieurs travaux ont analysé le comportement des fonctions de vraisemblance face au changement de la taille et de la forme de la fenêtre d'agrégation. Pour une petite fenêtre, les erreurs d'appariement sont dues à l'ambigüité et au bruit. Alors que pour une grande fenêtre, les disparités des pixels dans la fenêtre ont plus de chances d'être différentes. Ceci à cause des occlusions et des discontinuités de disparité.

Les méthodes locales supposent une faible variation de profondeur des pixels appartenant à une même zone d'agrégation. Autrement dit, ces pixels appartiennent à une région homogène qui correspond à une même surface. Toutefois, la disparité du pixel candidat est fortement influencée par la disparité des pixels voisins entrant en jeu lors du processus d'appariement. La présence de valeurs aberrantes dans le support fait converger la fonction de vraisemblance vers une solution erronée. Nous présentons ci-après une classification des différentes méthodes locales d'appariement dont le critère est la forme des fenêtres considérées :

- **Fenêtre adaptative** : Il s'agit de choisir la taille convenable de la fenêtre utilisée pour l'appariement de deux ensembles de données [OK92], [KSC01], [YP05], [Cyg05], [YK06], [WZ08], [ZGY08], [Bro09], [LHYJ09], [HBG10] et [GC10]. Kanade et Okutomi [KO94] ont présenté une méthode qui analyse la variation locale d'intensité et de disparité afin de choisir la fenêtre appropriée. La forme de la fenêtre est toujours rectangulaire, et reste invariable pour tous les pixels à apparier de l'image. Cette technique est très coûteuse, et dépend de la première estimation de la disparité. [BVZ98] propose de choisir une forme arbitraire de fenêtres connectées pour chaque pixel à apparier. Dans [Vek02] et [Vek03], les auteurs cherchent un ensemble de fenêtres de différentes tailles et formes pour chaque pixel à apparier. Toutefois, la forme de la fenêtre n'est pas la même pour tous les pixels. La fonction de vraisembance proposée nécessite l'initialisation et la gestion de plusieurs paramètres.

- **Fenêtres multiples ou fenêtre déplaçable** : Le principe des fenêtres multiples consiste à établir un ensemble de fenêtres prédéfinies de différentes tailles, ayant toutes la même forme. La fenêtre avec laquelle la fonction de vraisemblance donne le coût optimal est choisie comme la plus appropriée. [Arn83], [GLY92], [FRT97], [BI99], [FR00], [KSC01], [OKO02], [HIG02], [CWD03], [AKN07]. Le problème majeur de cette technique est que la forme des fenêtres est choisie arbitrairement. Ceci pose des problèmes dans le cas où le pixel à apparier appartient à une zone de discontinuité de profondeur. Pour résoudre ce problème, l'image de référence est segmentée en régions d'intensité ou de couleur homogènes [TS00] et [WKS04]. La région qui contient le pixel à apparier est alors choisie comme support, dont la taille et la forme sont arbitraires. La segmentation n'est pas toujours facile à cause des régions fortement texturées que peut contenir une image. Les fenêtres multiples permettent de résoudre les probèmes liés aux occultations géométriques, [BI99]. La combinaison des

fenêtre multiples et des fenêtres adaptatives améliore la qualité d'appariement des pixels spéculaires, et des pixels dans la frontière entre les régions diffuses et spéculaires [LLL+02].

– **Fenêtre pondérée** : Le principe est de pondérer les pixels appartenant à une fenêtre de taille et de forme fixes [Dar98], [YK06]. K. Prazdny [Pra85] a proposé une nouvelle fonction de vraisemblance permettant, pour un support S donné, d'assigner itérativement un poids aux pixels voisins. Ce principe est connu sous l'appelation de *diffusion itérative* [SS02]. [XWFS02] a ainsi proposé une fonction permettant d'estimer un poids adaptatif par des calculs radiaux. Cette technique se base sur la distribution des disparités initialement estimées par une autre méthode locale. L'inconvénient de cette technique est qu'elle est sensible à la carte de disparités initiale. [YK06] a présenté une technique de pondération des pixels contenus dans une fenêtre donnée.

1.6.4 Méthodes d'appariement globales

Les méthodes locales ne sont pas capable de lever les ambiguïtés d'appariement des pixels appartenant à des régions de couleur homogène ou de texture répétitive. Ces ambiguïtés peuvent être gérer par les méthodes globales. Dans cette section, nous proposons un état de l'art sur les différentes méthodes globales d'appariement stéréoscopique. Les méthodes ne seront pas décrites en détail, mais un bref rappel du principe sera proposé pour chacune. D. Scharstein a proposé dans [SS02] un cadre général permettant la classification de la plupart des méthodes d'appariement existantes. L'auteur décompose le processus d'appariement en quatre étapes :

1. Calcul du coût d'appariement.
2. Agrégation d'une zone.
3. Calcul/Optimisation des disparités.
4. Raffinement des disparités.

La majeure partie des méthodes globales est formulée comme un problème de minimisation d'énergie [Dem86]. Le principe est d'estimer une fonction de disparité d permettant d'optimiser une énergie globale E. L'étape d'agrégation est souvent ignorée par les méthodes globales. Souvent, l'espace de disparités est directement estimé lors du processus d'optimisation. Rappelons que les méthodes globales se basent sur l'optimisation d'une certaine fonction d'énergie. Les deux étapes majeures de ce processus sont la définition d'une fonction d'énergie, et l'application d'une technique d'optimisation. L'énergie à optimiser dépend de la précision souhaitée et des contraintes à intégrer. Elle prend généralement la forme suivante :

$$E(d) = E_{correspondance}(d) + E_{lissage}(d) \tag{1.32}$$

Le terme $E_{correspondance}(d)$ mesure le coût d'appariement obtenu avec une fonction de vraisemblance donnée. Le terme $E_{lissage}(d)$ décrit les contraintes de lissage permettant le contrôle de la convergence de l'algorithme utilisé. La forme la plus usuelle du premier terme est donnée par l'équation 1.33 :

$$E_{correspondance}(d) = \sum_{\substack{u,v' \in \mathcal{P}_d \\ u,v \in \mathcal{P}_g}} \Phi(I_g(u,v), I_d(u,v')) \tag{1.33}$$

Le terme de lissage peut prendre différentes formes. Les différents termes d'une fonction d'énergie seront détaillés dans le chapitre 2. Une fois la fonction d'énergie définie, divers algorithmes existent pour la recherche de la configuration de disparité optimale permettant d'atteindre un minimum local : ceci revient à estimer un étiquetage dans un graphe non orienté. Une formulation du problème, largement proposée dans la littérature, consiste à modéliser l'image de référence par un graphe en utilisant la formulation des *Champs de Markov Aléatoires*. Les trois méthodes les plus utilisées sont la *Programmation Dynamique*, la *Propagation de Croyance*, et la *Coupure de Graphe*.

1.6.4.1 Programmation dynamique

La programmation dynamique est une méthode permettant de réduire la complexité d'un problème d'optimisation en le décomposant en plusieurs sous-problèmes de complexité similaire [CLR90]. Cette méthode consiste en la recherche de la solution optimale, qui correspond au chemin optimal d'un réseau bidimensionnel. Chaque nœud de ce graphe correspond à un coût obtenu pour un couple de pixels candidats. Le coût correspond à une mesure de vraisemblance obtenue avec une fonction de corrélation. Le chemin optimal est celui qui optimise le coût global, défini comme la somme des coûts locaux. Dans le cas où les pixels à apparier se réduisent aux pixels, les axes du graphe sont les deux droites épipolaires conjuguées [BT98] [BMD96][BI99]. D'autres travaux ont proposé d'apparier des pixels de contour [BB81] [Ben84] [OK85b][SMM08] ou des segments de droite [LR94].

Afin de réduire les ambigüités, [BB81] propose une nouvelle contrainte permettant de tenir compte des dépendances verticales des disparités. Cette contrainte est introduite afin de rectifier les erreurs de disparité obtenues avec les lignes épipolaires horizontales conjuguées. [OK85a] a intégré cette contrainte dans le processus d'optimisation, en minimisant la somme des coûts dans une région bidimensionnelle définie par les contours horizontaux et verticaux. [MVPG02]

propose un algorithme symétrique indépendant de la fonction de coût, utilisé pour l'initialisation des coûts dans le graphe. L'auteur démontre la possibilité d'une implémentation hiérarchique permettant une réduction de la complexité et un gain considérable en temps d'exécution et en mémoire. [GY03] introduit le principe de la logique floue dans le calcul du coût des nœuds : une mesure de disparité n'est assignée qu'aux noeuds ayant une mesure de confiance supérieure à un seuil.[SMM08] introduit la couleur lors de l'appariement des pixels de contour par programmation dynamique. [FYO+04] propose un algorithme de programmation dynamique hiérarchique pouvant être exécuté en temps réel. Une synthèse est proposée par [FZ10] sur les algorithmes d'optimisation dans des graphes, en l'occurrence les algorithmes de programmation dynamique. Nous citons d'autres travaux intéressants basés sur la programmation dynamique [LAC04][KLCL05][LSY06][WLGRY06][ACRB06][SSE+09].

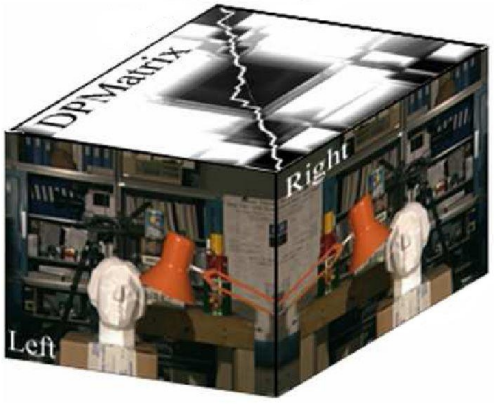


FIGURE 1.11 – Représentation tridimensionnelle du principe de la programmation dynamique d'après [FYO+04].

1.6.4.2 Régularisation

Cette méthode utilise des informations supplémentaires pour résoudre les problèmes mal posés, tels que le problème d'appariement stéréoscopique, et pour éviter un sur-apprentissage. Elle est utilisée généralement dans le cadre des problèmes inverses. L'appariement stéréoscopique étant un problème mal posé, la régularisation assure la continuité en présence de valeurs bruitées. Elle intègre explicitement un terme de lissage dans la fonction globale à optimiser. En stéréovision, il s'agit de définir une fonction mesurant le lissage "*smoothness*" de l'espace des disparités, tout en maximisant la cohérence colorimétrique. Le terme de lissage encode généralement l'hypothèse de continuité de disparité des régions homogènes.

[Dem86] a démontré que l'application d'une contrainte globale de lissage, basée sur la théorie standard de régularisation, échoue. Il a proposé un cadre général basé sur la théorie de régulari-

sation pour la prise en compte simultanée des contraintes de lissage et de continuité. [RM02] a proposé une nouvelle formulation basée sur la régularisation pour résoudre les problèmes inverses en vision artificielle. La régularisation se base sur une fonction de lissage du second ordre permettant de préserver les contours, ceci en se basant sur la constance de disparité, de pente et de courbature des régions homogènes. [MYS06] propose une fonction d'énergie basée sur la régularisation en introduisant la contrainte de visibilité et la segmentation couleur de l'image de référence. L'optimisation par régularisation considère la continuité de la profondeur entre les régions segmentées, et pénalise les occultations à travers la contrainte de visibilité.

Plus récemment, [PCBC10] propose un algorithme "primal-dual" permettant l'estimation des solutions globales des modèles variationnels de régularité convexe. L'auteur propose une nouvelle fonction d'énergie composée de certains termes de régularité convexes. L'inconvénient majeur des méthodes basées sur la régularisation est la difficulté d'intégrer la contraintes d'occultation dans la fonction d'énergie à optimiser. Ce modèle exige que chaque pixel de l'image de référence ait un homologue dans l'autre image.

1.6.4.3 Propagation de croyance

[ZWGY10] propose l'introduction de la composante temporelle dans la formulation du cadre général des champs de Markov aléatoire spatiaux. Le nouveau modèle proposé établit un lien spatial et temporel entre les messages transmis dans un voisinage local. Le maximum a posteriori est estimé par un algorithme de propagation de croyance itératif. L'auteur montre que l'introduction de l'aspect temporel lors de la propagation de croyance améliore la robustesse et la précision des estimations de la profondeur des scènes dynamiques.

[Zhe10] propose un algorithme de propagation de croyance hiérarchique couplé avec une sursegmentation multi-échelle des images d'entrée pour l'estimation d'une carte de disparité dense. La composante temporelle est introduite en tant que contrainte dans le modèle pour raffiner le rendu tridimensionnel. La cohérence spatio-temporelle des intensités et des disparités permet la réduction du bruit et des pixels aberrants.

Malgré leur capacité à fournir une carte de disparités précise, les algorithmes d'appariement basés sur la propagation de croyance dans le cadre des champs de Markov aléatoires sont connus pour leur lenteur. [YWA10] propose un algorithme de propagation de croyance hiérarchique dont le temps d'exécution est constant. L'espace de recherche des disparités est réduit à chaque niveau hiérarchique. L'auteur montre que son algorithme est trente fois plus rapide que la version standard des algorithmes de propagation de croyance.

1.6.4.4 Méthodes coopératives

Les algorithmes coopératifs ont été initialement développés pour l'appariement des primitives dans le cadre de la stéréovision éparse, afin de développer une solution permettant la résolution d'un problème mal posé. Les algorithmes coopératifs prennent en compte les erreurs d'estimation des disparités lors du processus d'optimisation. [ZK00] a proposé d'introduire les contraintes d'unicité et de continuité initialement proposées dans [MP76], pour développer un nouvel algorithme d'appariement coopératif. Ces contraintes sont incorporées dans une structure tridimensionnelle dans l'espace de disparités. Chaque point de la structure correspond à un pixel apparié de l'image de référence ayant comme attribut le triplet *(x, y, disp)*. Cette formulation a permis d'améliorer les erreurs d'appariements dûes à la présence d'occultations. Concernant les régions non occultées, l'auteur avait montré dans [ZK00] des résultats satisfaisants. La disparité reste imprécise au niveau des frontières des objets à cause de l'utilisation d'une fenêtre fixe.

Des améliorations ont été apportées par [ZK02] qui utilise une fenêtre 3D adaptative basée sur une segmentation couleur de l'image de référence. [Hua07] propose un système multi-agent coopératif pour une optimisation distribuée. L'objectif est de trouver la solution optimale pour chaque agent, sachant que chaque agent a sa propre fonction d'énergie à optimiser. [WZ08] utilise un algorithme coopératif inter-régions pour l'appariement de primitives de type région. La fonction d'énergie proposée dépend à la fois de la couleur de chaque région, de la contrainte de continuité de disparité et de la présence d'éventuelles occultations entre deux régions adjacentes. Plus récemment, [Bro09] a introduit la contrainte colorimétrique dans la formulation de la fonction d'énergie. La carte de disparités est initialement obtenue par une méthode locale utilisant un support adaptatif. Une optimisation coopérative est ensuite appliquée pour attribuer une probabilité à chaque mesure de similarité obtenue pour chaque couple. L'optimisation des algorithmes coopératifs est de nature locale, elle exige donc une bonne initialisation de la carte des disparités pour garantir une bonne convergence vers une solution optimale. Au contraire des méthodes de régularisation, les algorithmes coopératifs ont la capacité d'intégrer des contraintes de discontinuités non-convexes.

1.6.4.5 "*Tensor voting*"

Cette méthode est originellement proposée par [MM06] qui se base sur le concept d'organisation perceptuelle postulé par la théorie de la Gestalt, connu sous le nom de la théorie de psychologie de la forme. La méthode *Tensor voting* vient du principe de fonctionnement de notre système visuel, sensible aux changements locaux d'intensité, d'orientation et de changement d'échelle des structures locales. Elle permet d'estimer la structure géométrique d'un objet ou d'une scène à partir des données manquantes, irrégulières ou bruitées. Cette méthode a été proposée pour résoudre le problème d'appariement de petites structures locales entre deux images. Une structure locale

peut être caractérisée par certains paramètres tels que l'amplitude du signal correspondant, sa cohérence et son orientation locale. Chaque point d'un voisinage donné est alors représenté sous la forme d'un tenseur. L'ensemble des tenseurs forme un champ de vote, à travers duquel les votes se propagent. [MM06] propose un algorithme d'appariement composé des quatre étapes suivantes :

1. Initialisation de la carte de disparités en utilisant une méthode locale.

2. Détection d'appariements corrects.

3. Regroupement des pixels appariés en surfaces homogènes.

4. Raffinement de l'appariement.

[BMB08] propose une technique basée sur le Tensor voting pour la reconstruction tridimensionnelle de surfaces à partir de données 3D bruitées ou manquantes. [WYJT10] propose un algorithme de reconstruction 3D dense à partir de caméras non calibrées. L'auteur utilise un tenseur 3D et combine dans le modèle les avantages de la cohérence photométrique, des contraintes géométrique et de visibilité. L'auteur propose un cadre général pour l'appariement, la propagation et le filtrage des cartes de disparités.

1.6.4.6 Algorithmes génétiques

John Holland est considéré comme le père des Algorithmes Génétiques [Hol75]. Cette branche des algorithmes évolutionnaires suscite un intérêt croissant, en particulier pour la résolution des systèmes adaptatifs complexes. Basés sur la théorie de Darwin [Dar59], les algorithmes génétiques reposent sur le principe de l'évolution de la structure des organismes, permettant leur survie face aux changements continuels de l'environnement auquel ils sont confrontés. La coadaptation et la coévolution entre les individus d'une population évolutive conduit à l'émergence de chaque individu vers un nouvel état rendant la population plus stable compte tenu de son environnement. Chaque état de la population comporte des innovations et des améliorations par rapport aux états précédents. Ce processus naturel a été modélisé par des relations mathématiques afin d'être adapté à des systèmes artificiels complexes. L'algorithme génétique tente de trouver une solution approchée à un problème d'optimisation avec des contraintes. Il repose sur le principe suivant :

1. Initialisation aléatoire ou non de la population.

2. Définition d'une *fonction d'ajustement*.

3. Sélection des meilleurs individus.

4. Variation par croisement et mutation.

5. Vérification de la stabilité de la population. Autrement dit, il s'agit de vérifier la solution obtenue par rapport à la fonction objective choisie.

Ce principe a été appliqué aux problèmes d'appariement stéréoscopique. Une population correspond à une carte de disparités dont les individus sont les différents points qui la compose. Le point d'entrée de l'algorithme consiste à initialiser, aléatoirement ou non, chaque individu. Ceci revient à donner une mesure de disparité à chaque pixel de la carte. L'algorithme étant itératif, un croisement et une mutation des individus sont ainsi appliqués à chaque itération, ceci en optimisant une fonction s'appuyant sur un coût global, et soumise à certaines contraintes telles que la contrainte de continuité des disparités, comme proposé dans [SM95]. Dans [LCCB01], l'auteur propose une technique basée sur des pyramides d'ondelettes complexes conjuguées. La fonction de coût correspond à la différence des valeurs des coefficients d'ondelettes, et utilise la contrainte d'ordre et la contrainte de lissage des disparités dans le cadre d'un algorithme génétique. D'autres travaux ont été proposés [HSC$^+$01], [GY02], [GL03], et [IRP05]. Le lecteur intéressé par cette technique est invité de voir l'article [PU07] qui propose un aperçu sur les algorithmes génétiques appliqués à la segmentation et à l'amélioration d'images.

1.6.4.7 La "*Stereo Matting*"

Une contrainte importante en stéréo vision est la constance de couleur entre deux images de la même scène. C'est le cas des pixels appartenant à des surfaces mattes. Deux pixels homologues doivent avoir la même apparence. Cette hypothèse n'est pas toujours valide et peut être violée en présence de transparence. Par exemple au flou des lentilles ou de la discrétisation : deux pixels homologues peuvent alors avoir deux couleurs différentes. Cependant, la couleur apparente d'un pixel n'est que la combinaison linéaire de la couleur de deux pixels, un correspondant au premier plan et l'autre à l'arrière plan. Les premiers travaux ayant traité ce problème sont [BSA98] et [SG98]. [ZKU$^+$04] [XJ07], [TWZ08] proposent d'appliquer ce principe comme post-traitement, [XJ07], [TWZ08].

1.7 Conclusion

Nous avons présenté dans ce chapitre les différents aspects de la vision tridimensionnelle à partir de deux capteurs stéréoscopiques passifs. Nous avons vu qu'à partir d'un système calibré, la profondeur d'un objet ne dépend que d'un seul paramètre qui est la disparité. Diverses méthodes

d'appariement locales et globales ont été présentées, permettant l'estimation d'une carte de disparités. Le choix d'une méthode d'appariement ou d'une autre dépend principalement de la complexité de la méthode utilisée, de la précision et du temps d'exécution souhaités. Il est important de noter qu'une classification exhaustive des méthodes d'appariement semble une tâche difficile à cause de la variabilité des modèles proposés pour la modélisation du problème. Par ailleurs, le problème d'appariement stéréoscopique est considéré comme un problème mal posé. Les critères d'évaluation des algorithmes d'appariement sont loin d'être déterministes malgré la disponibilité d'une base stéréoscopique avec des vraies cartes de disparité [mid]. Cette base reste toutefois limitée, puisque les paires d'images proposées sont obtenues dans des environnements bien contrôlés.

Chapitre 2

Développement d'un Algorithme d'Appariement Stéréoscopique Sélectif

2.1 Introduction

Une carte de disparités peut être obtenue en appliquant une méthode d'appariement locale ou globale sur une paire d'images stéréoscopiques. Les méthodes locales sont connues pour leur rapidité et leur manque de précision, alors que les méthodes globales sont relativement lentes mais assurent une bonne qualité d'appariement. Le choix de la stratégie d'appariement est un compromis à trouver en fonction de la variabilité des applications et des contraintes associées. La localisation d'obstacles routiers par stéréovision embarquée est un exemple d'application exigeant un traitement en temps réel [HPHH08] : les obstacles doivent être localisés à temps afin de réagir de façon sûre et efficace pour les éviter. En revanche, ce genre d'application ne nécessite pas une reconstruction 3D parfaite de l'environnement observé. À l'autre extrême, la reconstruction tridimensionnelle de bâtiments dans une scène urbaine est un exemple d'application nécessitant un rendu 3D très précis de l'ensemble des pixels de l'image, mais où le temps de traitement n'est pas considéré comme une contrainte forte [WMK$^+$08]. Une carte de disparités dense nécessite l'appariement de primitives de type pixel. L'initialisation de la carte de disparités est habituellement faite pour une méthode locale. Les méthodes globales sont introduites pour prendre en compte les contraintes, telles que des contraintes de lissage des surfaces homogènes, des régions de discontinuité et les occultations. Une région de discontinuité correspond à la frontière entre deux ou plusieurs régions de disparités différentes. En terme de distance, deux régions de disparités différentes se situent à des profondeurs différentes par rapport aux caméras.

Comme vu au §1.6.3 du chapitre 1, les méthodes d'appariement locales dépendent du choix de la fonction de vraisemblance, et du choix de la taille et la forme de la fenêtre d'agrégation. Ces deux derniers attributs ont été pris en compte dans plusieurs travaux afin d'accélérer le temps de traite-

ment, et pour améliorer la qualité d'appariement. Nous détaillons ces deux facteurs dans la section 2.5.1. En se référant à la littérature, de nombreuses recherches récentes se sont orientées vers la combinaison des méthodes locales et globales. Ce choix est justifié par l'avancée technologique considérable permettant l'exécution dans des temps raisonnables des algorithmes de complexité importante. Malgré leur diversité, les méthodes globales introduisent toutes des contraintes d'ordre global dans le processus d'appariement ; c'est la façon d'encoder et de modéliser ces contraintes qui fait la différence entre elles.

2.2 Objectifs

Comme vu au §2.1, le choix de l'architecture d'un algorithme de mise en correspondance stéréoscopique dépend fortement de l'application envisagée. Certains travaux visent à obtenir un rendu tridimensionnel très précis, sans tenir compte du temps d'appariement. Les applications exigeant un traitement en temps réel se basent généralement sur l'appariement de primitives plutôt que de l'ensemble des pixels. Un algorithme d'appariement doit tenir compte des trois aspects suivants :

- ***La robustesse :*** un algorithme d'appariement robuste est capable de gérer les problèmes d'occultation et de discontinuité de profondeur. Pour un même objet, les disparités varient de façon continue et homogène. La variation des disparités n'est significative qu'à la frontière des objets.

- ***La précision :*** ce paramètre signifie que l'écart entre les disparités réelles et estimées ne doit pas être élevé. L'imprécision d'appariement concerne souvent les pixels appartenant à des régions de couleur uniforme, de texture répétitive, et des régions partiellement occultées.

- ***Le temps d'appariement :*** l'application d'un algorithme robuste, permettant l'obtention d'une carte de disparités précise, nécessite la prise en compte de contraintes supplémentaires, ce qui augmente naturellement le temps d'estimation des disparités. La manière d'introduire ces contraintes, tout en diminuant le temps des traitements, est un élément clé pour la réussite du processus d'appariement.

Malgré leur diversité, les objectifs des méthodes d'appariement restent toujours les mêmes. Le besoin d'une méthode d'appariement tenant compte à la fois de la robustesse, de la précision, et du temps de traitement, nous a motivé à réfléchir sur la façon dont les contraintes d'occultation, d'ordre, et de discontinuité, en particulier, sont exploitées. Nous proposons dans ce chapitre un nouvel algorithme d'appariement stéréoscopique, tirant profit simultanément des méthodes locales et globales, et qui s'intègre dans le cadre général basé sur le principe d'optimisation d'énergie

largement exploité dans la littérature [BVI98], [TF03], [SZS+06], [AKT08], et [PCBC10]. L'algorithme que nous développons a pour finalité la résolution des problèmes d'occultations et de discontinuités. Il tente ainsi d'assurer une bonne précision d'appariement tout en diminuant le temps de traitement, un compromis non encore résolu, et toujours posé en vision stéréoscopique.

2.3 Cadre général d'un algorithme d'appariement global

La plupart des algorithmes d'appariement proposés durant cette dernière décennie sont basés sur la formulation proposée dans [SS02]. Dans le cas des méthodes locales, l'appariement consiste à définir en premier lieu la fonction de vraisemblance à utiliser pour l'évaluation du degré de corrélation entre chaque paire de pixels. La deuxième étape consiste à définir la zone d'agrégation, dans laquelle la fonction de vraisemblance va s'exécuter. Une disparité est ainsi obtenue pour chaque pixel en appliquant la stratégie "*Winner-Take-All*" [Gro73] qui consiste à choisir le pixel candidat qui optimise la fonction de vraisemblance. Dans le cas des méthodes globales, le problème d'appariement est vu comme un problème d'optimisation globale dans laquelle certaines contraintes, telles que les contraintes de lissage, de discontinuité et d'occlusion, sont prises en compte explicitement par le modèle. La définition de la zone d'agrégation n'est pas explicitement abordée par les méthodes globales puisque l'optimisation concerne l'ensemble de l'image. Les principales étapes d'un algorithme d'appariement global sont les suivantes :

1. ***Modélisation du problème :*** la première étape consiste à choisir une architecture permettant la modélisation du problème d'appariement. Une modélisation unidimensionnelle du problème est proposée par la programmation dynamique. A partir de deux lignes à apparier, la programmation dynamique consiste à trouver le chemin minimisant un coût global. Une des modélisations les plus utilisées consiste à représenter le problème dans un cadre markovien : il s'agit d'une modélisation par des Champs de Markov Aléatoires. L'ensemble des pixels à apparier forme un graphe de nœuds interconnectés. Les arêtes matérialisent les dépendances spatiales entre les nœuds. Le principe des méthodes globales est de trouver une solution optimisant une énergie globale.

2. ***Initialisation des disparités :*** l'initialisation fournit une première estimation des disparités de l'ensemble des pixels de l'image de référence. Les méthodes locales sont souvent utilisées pour l'initialisation en raison de leur rapidité.

3. ***Optimisation selon des contraintes :*** en partant des disparités obtenues lors de la deuxième étape, l'optimisation consiste à ré-estimer la disparité de chaque pixel, en se basant sur des contraintes d'ordre global. Plusieurs types de contraintes peuvent être utilisés, telles que la

cohérence colorimétrique entre des pixels voisins, et la cohérence spatiale des disparités. Cette étape permet l'élimination des disparités aberrantes, et résoud les problèmes d'occultation, de discontinuité de profondeur, et des régions de couleur ou de textures uniformes.

4. ***Raffinement des disparités :*** le raffinement est une étape supplémentaire permettant d'obtenir un rendu tridimensionnel de haute qualité. Ce processus permet l'élimination des "trous" de disparités par interpolation, et le lissage des surfaces homogènes par estimation de l'orientation et de l'inclinaison des plans, ou segments, de disparités.

2.3.1 Modélisation par Champs de Markov Aléatoires

Les modèles graphiques probabilistes sont souvent introduits pour la modélisation des problèmes liés à la vision artificielle. L'utilisation des graphes permet de mieux représenter la distribution des probabilités. Les modèles graphiques dirigés, tels que les réseaux Bayésiens, sont souvent utilisés pour l'inférence des systèmes de causes à effet. D'autres modèles sont introduits en contrôle et en traitement de signal, tels que les champs de Markov cachés et les modèles de représentation d'états continus. Une autre alternative est d'utiliser des modèles graphiques dont les graphes sont non dirigés. Les Champs de Markov Aléatoires, plus connus sous l'appellation anglaise "*Markov Random Fields*" (en abrégé *MRF*), sont devenus durant cette dernière décennie un outil populaire et puissant de modélisation d'images. Il s'agit d'un modèle graphique non dirigé permettant la résolution des problèmes inverses relatifs à la vision de bas niveau (traitement au nveau pixel), tels que la restauration, la segmentation, la reconstruction de surface 3D, et la mise en correspondance stéréoscopique. Le choix des modèles basés sur une représentation graphique est justifié par leur capacité à gérer les occultations et à intégrer des contraintes telles que les contraintes d'ordre et de symétrie. Nous nous sommes basés sur les *MRF* pour la modélisation du problème d'estimation des disparités.

Soit $\mathcal{G} = \langle \nu, \varepsilon \rangle$ un graphe non orienté tel que ν est l'ensemble des nœuds, et ε l'ensemble des arêtes liant les nœuds. L'ensemble ν des nœuds correspond aux pixels de l'image de référence I_g. Les arêtes modélisent les dépendances spatiales entre les pixels. A chaque nœud $p \in \nu$ est associée une variable aléatoire x dont les valeurs possibles forment un ensemble discret, noté $\mathcal{X}$. Les variables aléatoires correspondent aux observations qui sont, dans le cas de l'appariement stéréoscopique, les intensités ou les composantes couleur des pixels des images gauche et droite. L'inférence consiste à estimer les variables cachées, qui sont les disparités, à partir des observations. Un label $l \in \mathcal{L}$ est attribué à chaque pixel, tel que $\mathcal{L}$ est un ensemble fini, discret ou continu, de disparités possibles. En se basant sur le cadre général des *MRF*, l'optimisation des problèmes est connue pour être NP-difficile. Une approximation de la solution optimale est possible par d'autres méthodes telles que la méthode du *recuit simulé* [GG84]. Cette méthode propose une solution pour n'importe quelle fonction d'énergie en un temps exponentiel, ce qui la rend trés lente en pratique.

Plus récentes, les méthodes de coupure des graphes [PCF06] (pages 79-95), et la propagation de croyance [FH06] donnent une estimation précise avec des temps raisonnables. Ces deux méthodes seront décrites au §2.3.2. Etant donné l'ensemble des observations o qui correspondent aux composantes couleur des pixels des images gauche et droite, l'estimation d'une solution $f \in \mathcal{F}$ peut être vue comme la probabilité conditionnelle étant donnée la probabilité jointe $P(f, o)$. Une solution f correspond à l'ensemble des labels attribués à tous les pixels, et $\mathcal{F}$ correspond à l'ensemble des solutions possibles. La probabilité à posteriori est donnée par :

$$P(f/o) = \frac{P(f,o)}{P(o)} \propto P(f,o) \tag{2.1}$$

La notation $\propto$ signifie que la quantité $\frac{P(f,o)}{P(o)}$ est *proportionnelle à* $P(f, o)$. Il a été démontré dans [KS80] que la distribution de probabilités dans un ***MRF*** suit une distribution de Gibbs qui dépend d'un voisinage restreint pour une fonction de coût donnée. Cette propriété permet de gérer les problèmes de discontinuité, largement rencontrés en vision stéréoscopique. En se basant sur le théorème de Hammersley-Clifford, la distribution jointe $P(f, o)$ peut s'écrire de la manière suivante :

$$P(f,o) \propto \frac{1}{Z} \prod_{p \in \mathcal{P}} \phi(l_p, o) \prod_{\substack{p,q \\ q \in \mathcal{N}_p}} \varphi(l_p, l_q) \tag{2.2}$$

Dans l'équation 2.2, le paramètre Z est une constante de normatisation globale. $\mathcal{N}_p$ représente le voisinage du pixel p. La fonction $\phi(l_p, o)$, appelée *fonction potentiel*, permet d'encoder la vraisemblance locale. Elle est souvent mesurée à partir d'une fonction de vraisemblance évaluant le coût d'attribuer le label l_p au pixel p sachant l'observation o. La fonction $\varphi(l_p, l_q)$, appelée *fonction de compatibilité*, permet de mesurer le degré de lissage entre deux pixels voisins. Les deux quantités évaluées sont les deux labels l_p et l_q des pixels voisins p et q respectivement. Une solution peut être obtenue par un algorithme intitulé *produit-maximal* permettant d'approximer le maximum a posteriori du MRF. Il est défini en terme de distribution de probabilité que nous cherchons à maximiser. L'équation 2.2 peut s'écrire comme suit :

$$\begin{aligned} -log\left(P(f,o)\right) &= -log\left(\prod_{p\in\mathcal{P}}\phi(l_p,o)\prod_{\substack{p,q\\q\in\mathcal{N}_p}}\varphi(l_p,l_q)\right)+log(Z) \\ &\propto \sum_{p\in\mathcal{P}} -log(\phi(l_p,o)) + \sum_{\substack{p,q\\q\in\mathcal{N}_p}} -log(\varphi(l_p,l_q)) \end{aligned} \tag{2.3}$$

Le terme $logZ$ est omis dans l'équation 2.3 puisqu'il s'agit d'une constante. Le choix de cette formulation est justifié par le fait qu'elle est moins sensible aux valeurs aberrantes. Le problème de maximisation de probabilité se traduit alors par un problème de minimisation, donné par l'équation 2.4 :

$$max(P(f,o)) = min\left(\sum_{p\in\mathcal{P}} -log(\phi(l_p,o)) + \sum_{\substack{p,q\\q\in\mathcal{N}_p}} -log(\varphi(l_p,l_q))\right) \tag{2.4}$$

2.3.2 Optimisation d'une fonction d'énergie

Le principe d'optimisation consiste à trouver une solution optimale à partir d'un ensemble de solutions candidates $\mathcal{F}$. Il s'agit de définir une fonction d'énergie $E : \mathcal{F} \rightarrow \mathbb{R}$ permettant de mesurer la qualité d'une solution candidate. Une solution $f \in \mathcal{F}$ correspond à une configuration donnée de labels. Dans notre cadre, un label correspond à une disparité possible qu'un pixel du graphe peut avoir. Dans le cas où la fonction d'énergie mesure la *"mauvaiseté"* (*badness*) d'une solution, le problème d'optimisation se transforme en un problème de minimisation d'énergie. Une faible valeur d'énergie désigne une bonne solution, tandis qu'une valeur élevé correspond à une mauvaise solution. Le problème revient alors à trouver l'étiquetage optimal f permettant de minimiser la fonction d'énergie donnée par l'équation 2.5 :

$$E(f) = E_{etiquetage}(f) + E_{lissage}(f) \tag{2.5}$$

Le premier terme de la fonction d'énergie, $E_{etiquetage}(f)$, mesure le coût d'une solution f étant données les observations. Ce terme correspond au coût d'appariement local permettant d'obtenir des couples de la forme $\langle p, l\rangle$, avec $p \in \mathcal{P}$ et $l \in \mathcal{L}$. Le deuxième terme, $E_{lissage}(f)$, évalue dans quelle mesure une solution f est lisse. Une solution est dite lisse si les disparités d'une même ré-

gion varient d'une manière homogène. Le terme de lissage tient compte des interactions spatiales des pixels dans le graphe. Le *terme de lissage*, dit aussi *terme de discontinuité*, correspond à la différence de labels entre deux pixels voisins. Plutôt que d'établir la différence entre deux observations, telles que les intensités, le terme de lissage mesure la différence entre les disparités possibles de deux pixels voisins. Etant donné deux pixels voisins p et q tels que $p \in \mathcal{P}$, $q \in \mathcal{N}_p$ où $\mathcal{N}_p$ est le voisinage du pixel p, le terme de lissage peut s'écrire $V_{p,q}(l_p, l_q) = |l_p - l_q|$, où l_p et l_q sont les labels attribués aux pixels p et q respectivement. La forme la plus courante de la fonction d'énergie est donnée dans l'équation 2.6 telle qu'elle a été proposée dans [FH06], où $D_p(l_p)$ correspond au terme de données ou d'étiquetage :

$$E(f) = \sum_{p \in \mathcal{P}} D_p(l_p) + \sum_{\substack{p,q \\ q \in \mathcal{N}_p}} V_{p,q}(l_p, l_q) \tag{2.6}$$

2.4 Algorithme d'appariement stéréoscopique sélectif proposé

Le présent paragraphe décrit les différentes parties de l'algorithme d'appariement proposé. Nous proposons au §2.5 une nouvelle fonction de vraisemblance permettant l'obtention d'une première carte de disparités, une étape indispensable pour la réussite du processus d'optimisation. Cette première étape donne en sortie une carte de disparités dense, de telle sorte qu'un appariement est obligatoirement effectué pour chaque pixel de l'image de référence. Le pixel de l'image droite, pour lequel le score est optimal, est retenu comme le meilleur candidat, et est affecté au pixel à apparier de l'image gauche. À ce niveau, aucune information n'est disponible sur la certitude et la qualité de l'appariement. Nous introduisons au §2.6.2 une nouvelle fonction permettant d'évaluer chaque appariement. Cette fonction calcule une mesure de confiance pour chaque appariement, en se basant sur des critères d'ordre local. Seuls les appariements possédant une mesure de confiance élevée sont retenus. Cette étape réduit le nombre d'appariements et fournit une carte de disparités éparse. La sélection automatique des bons appariements est une étape importante pour la réussite de la troisième étape qui effectue la ré-estimation itérative des disparités des appariements, classés comme non pertinents dans la deuxième étape. Ceci conduit à une carte de disparités dense. Nous introduisons les deux contraintes suivantes :

1. ***La contrainte de cohérence colorimétrique***, qui répond aux problèmes d'occultations et de discontinuité des profondeurs.
2. ***La contrainte d'incertitude***, développée dans le cadre de nos travaux. Elle consiste à renforcer la contrainte de lissage des disparités et permet ainsi d'accélérer le processus de ré-estimation des disparités.

Nous adoptons le principe de propagation de croyance, en tant que méthode globale, pour la correction des disparités erronées. Les contraintes d'incertitude et de cohérence colorimétrique sont utilisées dans le processus d'inférence. Une vue générale de l'algorithme proposé est fournie dans la figure 2.1.

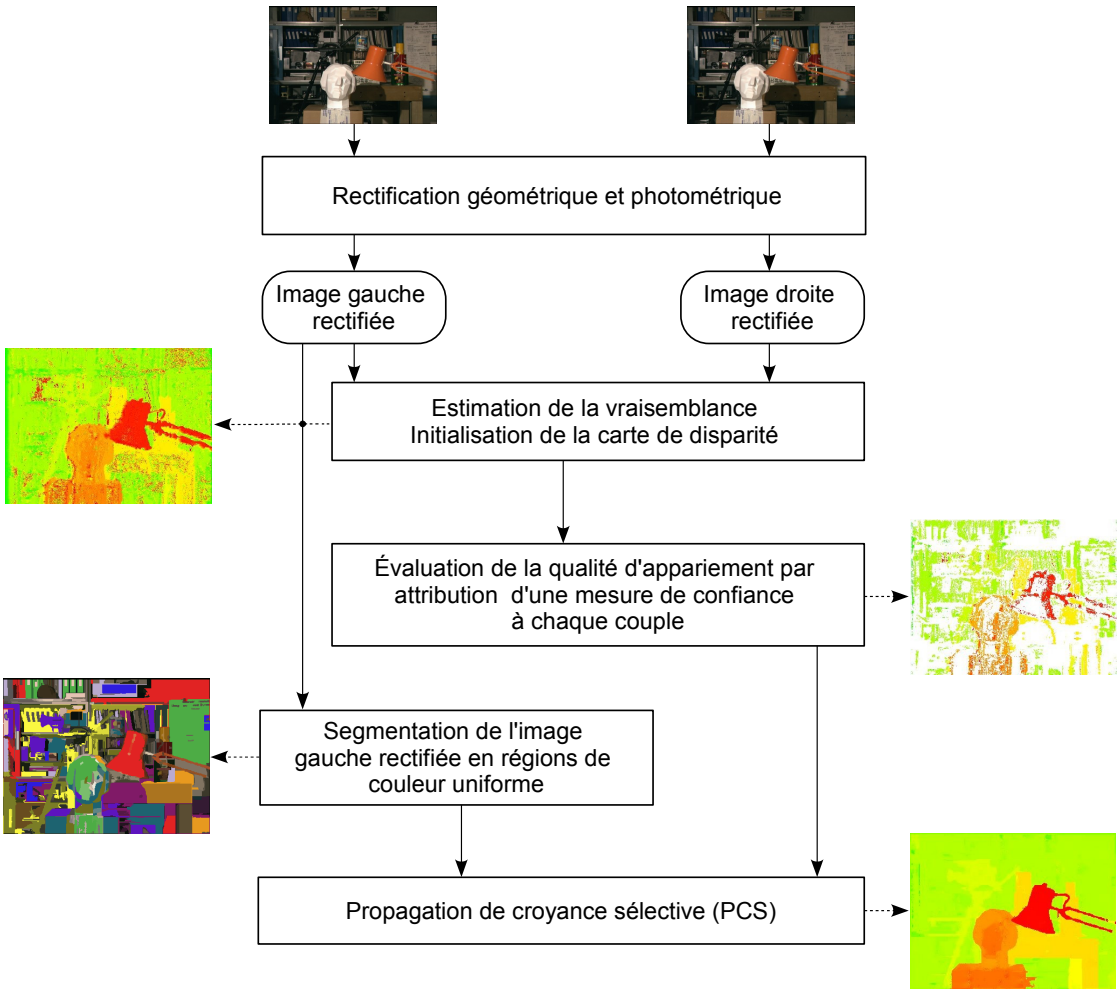


FIGURE 2.1 – Vue d'ensemble de l'algorithme d'appariement proposé.

La première étape de l'algorithme d'appariement proposé est la rectification des images stéréoscopiques d'entrée. Il s'agit de corriger les aberrations optiques et d'estimer la géométrie épipolaire, ce qui permet la réduction de l'espace de recherche. Cette étape est ignorée dans ce chapitre puisque les paires d'images évaluées sont prétraitées avant l'application de l'algorithme de mise en correspondance. Elle fait l'objet d'une discussion dans le dernier chapitre.

L'algorithme d'appariement proposé s'intègre dans le cadre général décrit au §2.3. Le principe général consiste à estimer une solution d'étiquetage minimisant une certaine énergie selon certaines contraintes. Une solution est une distribution de disparités sur l'ensemble des pixels de l'image

de référence. Le point de départ consiste à initialiser le graphe par attribution d'une disparité à chaque pixel de l'image de référence, en appliquant une fonction de vraisemblance sur chaque paire de pixels à apparier. Nous proposons une nouvelle fonction de vraisemblance discutée au §2.5. L'étape suivante consiste à ré-estimer itérativement les disparités à partir d'un algorithme de propagation de croyance sélectif, détaillé au §2.8. La propagation dépend des paires de pixels bien appariées. Cet ensemble des pixels bien appariés est sélectionné à partir d'une mesure de confiance, ou d'incertitude, calculée pour chaque paire. La fonction permettant le calcul d'une mesure de confiance est discutée au §2.6. La contrainte de cohérence colorimétrique est introduite dans le processus de rectification des disparités ; elle est détaillé au §2.7.

2.5 Initialisation des disparités : une méthode locale

2.5.1 Diversité des fenêtres d'agrégation et des fonctions de vraisemblance

L'idée de base des méthodes locales d'appariement est d'estimer le degré de corrélation entre deux pixels, issus de deux images différentes. Estimer le degré de similarité, ou de dissimilarité, de deux pixels par différence des intensités était une des premières méthodes. L'intensité étant un attribut peu discriminant et sensible à la présence de bruit, la comparaison pixel à pixel a dû être améliorée par la prise en compte du voisinage des pixels à apparier. Le voisinage d'un pixel forme une zone d'agrégation, qui peut être une fenêtre fixe unidimensionnelle [Yah07], [Lef08], bidimensionnelle [HBGR09], ou tridimensionnelle [ZGY08]. Il est aussi possible d'ajuster la taille et la forme de la zone d'agrégation en tenant compte d'autres informations supplémentaires contenues au voisinage du pixel central à apparier §1.6.3.2. L'appariement revient alors à faire correspondre deux zones centrées sur les pixels à apparier. La position relative des pixels voisins dans la zone d'agrégation est une information supplémentaire qui s'ajoute à l'intensité lors de l'estimation de la vraisemblance. Le principe d'appariement est le suivant : nous commencons par définir une fenêtre centrée sur le pixel à apparier dans l'image gauche. Une fenêtre de mêmes taille et forme est définie autour de chaque pixel candidat dans la zone de recherche de l'image droite. L'application d'une fonction de vraisemblance permet d'évaluer le score de corrélation de deux fenêtres. Le candidat avec qui la fonction de vraisemblance donne le meilleur score, est choisi comme le meilleur candidat. À ce niveau, deux questions se posent :

1. Quelle taille et forme de fenêtre d'agrégation faut-il utiliser ?

2. Quelle fonction de vraisemblance faut-il appliquer pour minimiser les ambiguïtés d'appariement ?

Le fait d'augmenter la taille de la fenêtre d'agrégation permet de rendre la comparaison plus discriminante, et donc permet de réduire le nombre d'appariement erronés. L'inconvénient est que l'utilisation d'une fenêtre d'agrégation large conduit à une carte de disparités dont les objets sont sur-segmentés. Le choix de la taille de la fenêtre est donc un paramètre important. Pour la réussite du processus d'appariement, seuls les pixels voisins appartenant à un même plan de disparité doivent être considérés lors de l'appariement. Plusieurs travaux ont proposé de faire varier la taille et/ou la forme des fenêtres d'agrégation [Pra85][XWFS02]. Ces méthodes sont itératives et donc très sensibles à l'estimation initiale des disparités, et exigent un temps de traitement important. Les pixels voisins du pixel à apparier n'ont pas le même degré d'importance dans le calcul du score de corrélation. [YK06] propose d'attribuer un poids à chaque pixel voisin selon sa similarité et sa proximité avec le pixel à apparier. La similarité consiste à établir la distance colorimétrique entre deux pixels dans l'espace couleur CIE Lab (un modèle de représentation des couleurs développé par la Commission Internationale de l'Éclairage) [YK06]. La proximité est mesurée en appliquant un noyau laplacien.

2.5.2 Nouvelle zone d'agrégation et nouvelle fonction de vraisemblance

Nous proposons dans ce paragraphe une nouvelle fonction permettant l'évaluation de la vraisemblance entre deux pixels à apparier. La fenêtre d'agrégation n'est plus considérée comme l'ensemble des pixels centrés sur le pixel à apparier : seules les lignes passant par le pixel à apparier sont prises en compte. Notons $L_{n,g\vee d}$, $n = \{1, 2, 3, 4\}$, l'ensemble des lignes verticales, horizontales et diagonales passant par le pixel à apparier de l'image gauche, ou droite. Nous divisons chaque ligne en deux segments symétriques centrés sur le pixel à apparier. Notons par $L^{+}_{n,g\vee d}$ et $L^{-}_{n,g\vee d}$ les segments correspondant à la ligne n passant par le pixel central. L'ensemble des segments sont regroupés dans un ensemble noté $\mathcal{H} = \bigcup_{n\in\{1,2,3,4\}} L^{+\wedge -}_{n,g\vee d}$. La figure 2.2 illustre ces notations.

Nous avons choisi d'utiliser une fenêtre d'agrégation fixe et non adaptative. Ceci réduit considérablement le temps d'appariement. Contrairement aux méthodes permettant l'évaluation des différences pixel à pixel, nous proposons de ne considérer que la moyenne de chaque composante couleur dans chaque segment. Pour un segment donné d'une fenêtre centrée sur le pixel de coordonnées (u, v), les moyennes des composantes couleur sont représentées sous la forme d'un vecteur noté $v_{g\vee d,s\in\mathcal{H}}(u, v) = \langle \overline{R}_s, \overline{V}_s, \overline{B}_s \rangle$, où s représente un segment de l'ensemble $\mathcal{H}$. Le terme $v^{k}_{g\vee d,s\in\mathcal{H}}(u, v)$ désigne la moyenne de la composante k du segment s appartenant à une fenêtre centrée sur le pixel de coordonnées (u, v) de l'image gauche ou droite, respectivement. La fonction de vraisemblance, donnée pour les deux pixels à apparier $\mathcal{P}_1 = I_g(u, v)$ et $\mathcal{P}_2 = I_d(u, v')$, peut être vue comme *la Différence de Couleur Moyenne Pondérée* $(DCMP)$. Elle est définie comme suit :

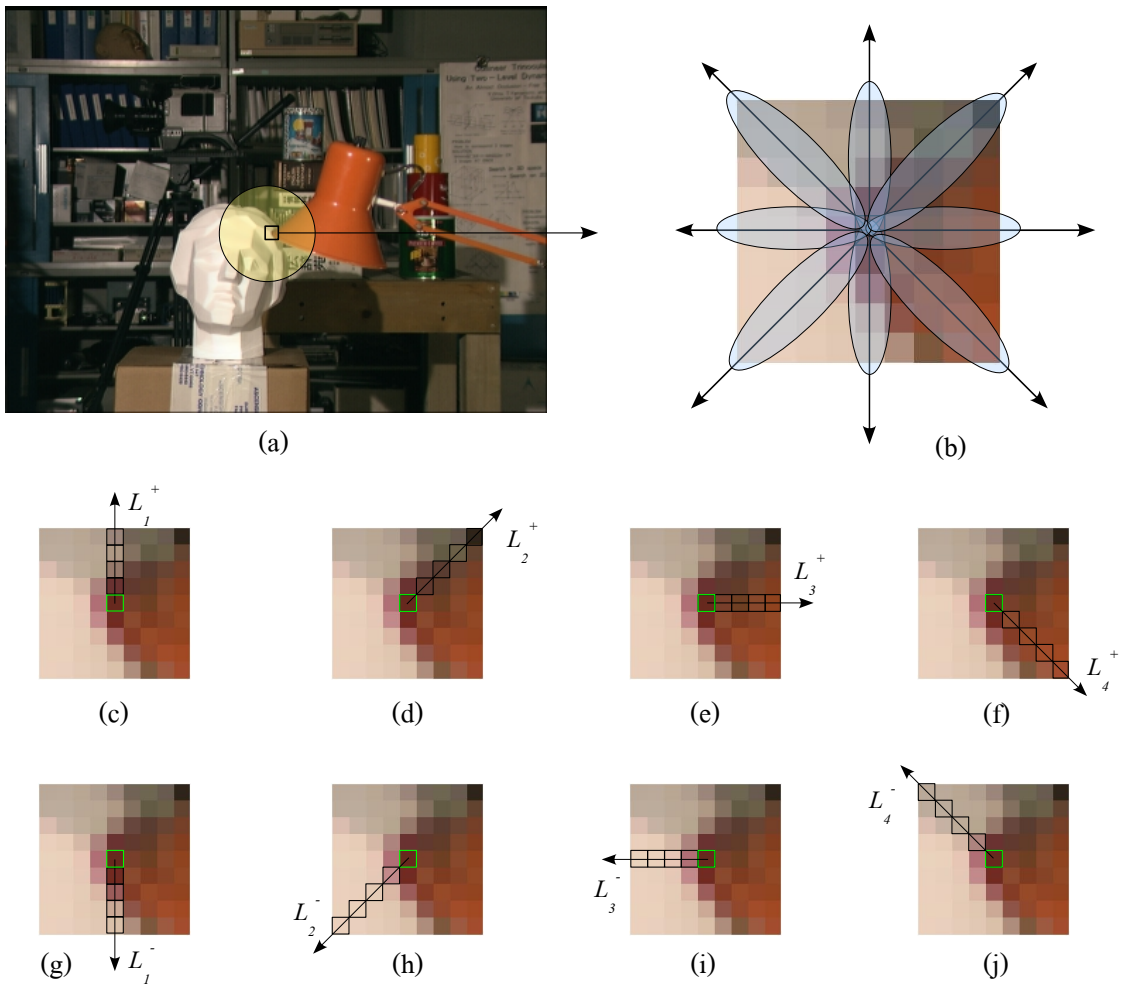


FIGURE 2.2 – Les segments pris en compte par la fonction de vraisemblance pour une fenêtre de taille 9×9, figure (b), centrée sur le pixel de coordonnées $(160, 189)$ de l'image Tsukuba, figure (a), de la base stéréoscopique [mid]. Les figures (c) à (j) illustrent les 8 segments.

$$DCMP\left(\mathcal{P}_1, \mathcal{P}_2\right) = \Delta_2\left(\mathcal{P}_1, \mathcal{P}_2\right) \times \sum_{\substack{s \in \mathcal{H} \\ k \in \{R,V,B\}}} \left(D_{k,s} \times W_{k,s}\right) \tag{2.7}$$

Dans la fonction de vraisemblance proposée, le score obtenu pour deux fenêtres données est pondéré par la distance colorimétrique entre les deux pixels centraux. Pour tous les espaces colorimétrique, il a souvent été proposé d'utiliser la distance euclidienne pour le calcul de la distance d'ordre q entre deux couleur [Cro97]. Elle est définie comme suit pour $q = 2$:

$$\Delta_2\left(\mathcal{P}_1,\mathcal{P}_2\right)=\left(\sum_{k\in\{R,V,B\}}\left(I_g^k(u,v)-I_d^k(u,v')\right)^2\right)^{\frac{1}{2}} \tag{2.8}$$

Pour un segment $s\in\mathcal{H}$ et une composante couleur $k\in\{R,V,B\}$, le terme $D_{k,s}$ mesure l'écart de la différence entre la composante k du pixel central, et la moyenne de la même composante couleur du segment s des images gauche et droite :

$$D_{k,s}=\begin{cases}\mathcal{Q}_1=\left|\left(I_g^k(u,v)-v_{g,s}^k(u,v)\right)-\left(I_d^k(u,v')-v_{d,s}^k(u,v')\right)\right| & \text{si } \mathcal{Q}_1\geq 1\\ 1 & \text{sinon}\end{cases} \tag{2.9}$$

Où $I_g^k(u,v)$ et $I_d^k(u,v')$ représentent respectivement la composante k du pixel central de coordonnées (u,v) de l'image gauche, et du pixel candidat de coordonnées (u,v') de l'image droite. Les termes $v_{g,s}^k(u,v)$ et $v_{d,s}^k(u,v')$ représentent la moyenne de la composante couleur k du segment s d'une fenêtre centrée sur le pixel de coordonnées (u,v) de l'image gauche, et d'une autre fenêtre centrée sur le pixel de coordonnées (u,v') de l'image droite. Le deuxième terme, noté $W_{k,s}$, peut être vu comme une pondération du terme $D_{k,s}$. Il permet de tenir compte des variations locales d'illumination entre les deux fenêtres. Le terme $W_{k,s}$ est donné par l'équation 2.10 :

$$W_{k,s}=\begin{cases}\mathcal{Q}_2=\left|\left(\frac{I_g^k(u,v)+v_{g,s}^k}{2}\right)-\left(\frac{I_d^k(u,v')+v_{d,s}^k}{2}\right)\right| & \text{si } \mathcal{Q}_2\geq 1\\ 1 & \text{sinon}\end{cases} \tag{2.10}$$

La figure 2.3 illustre l'impact du terme $W_{k,s}$ sur l'estimation de la vraisemblance entre deux pixels. La quantité $D_{k,s}$ qui est l'écart de la différence d'une composante couleur donnée entre le pixel central et un segment s, reste constante pour l'appariement des couples de segments (L_4^+(a), L_4^+(b)), (L_4^+(a), L_4^+(c)) et (L_4^+(b), L_4^+(c)) de la figure 2.2. Afin de tenir compte des variations globales d'illumination, le terme $W_{k,s}$ pénalise la quantité $D_{k,s}$ par la prise en compte de la moyenne locale de chaque composante couleur sur l'ensemble des segments.

L'application de la stratégie "*Winner-take-all*" permet d'associer au pixel à apparier le candidat ayant le score optimal. Cependant, le score est loin d'être un critère discriminant. La figure 2.4 illustre deux courbes de variation des scores obtenus avec une fonction de vraisemblance. En

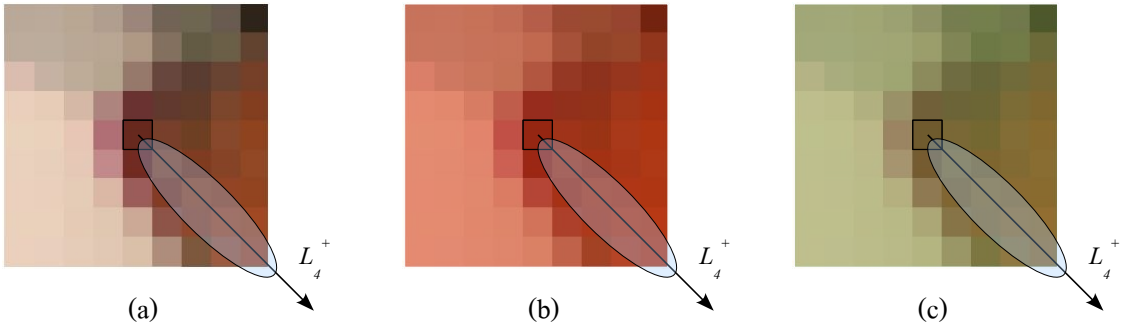


FIGURE 2.3 – Prise en compte des variations locales d'illumination par l'introduction d'un terme de pénalité $W_{k,s}$. Les images (a), (b) et (c) représentent des fenêtres de tailles 9×9 centrées sur des pixels à faire correspondre. Plus la variation d'illumination entre deux fenêtres est grande, plus le score de vraisemblance est grand.

se référant à la première courbe (a), le pixel de rang 12 est choisi comme le meilleur candidat puisqu'il possède le meilleur score. L'appariement n'est pas ambigu puisqu'il existe un saut significatif entre le score du meilleur candidat de rang 12 et ceux des candidats voisins. L'exemple (b) illustre le cas d'un appariement ambigu. Le candidat ayant le meilleur score est celui de rang 12, mais les candidats voisins ont des scores proches, d'où l'idée d'évaluer la qualité d'appariement en intégrant d'autres critères plus discriminants. Le §2.6 détaille les différentes méthodes existantes d'évaluation de la qualité d'appariement, ainsi que notre nouvelle fonction mesurant le degré de confiance des appariements.

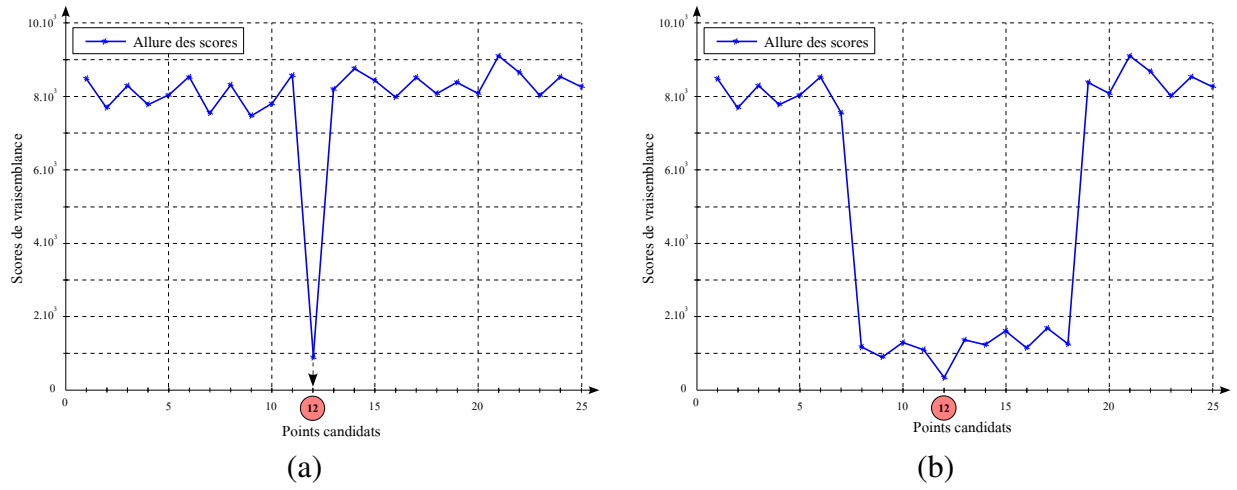


FIGURE 2.4 – Allures des courbes illustrant le cas d'un appariement non ambigu, courbe (a), et d'un appariement ambigu, courbe (b).

2.6 Evaluation de la qualité d'appariement

2.6.1 Différentes méthodes d'évaluation

La détection automatique d'erreurs d'appariement est une étape primordiale pour la réussite de la vision en relief. Une des techniques permettant la détection de pixels occultés est basée sur la vérification de la contrainte de symétrie, en appliquant le processus d'appariement deux fois : de droite à gauche puis de gauche à droite. D'autres proposent d'analyser la carte de disparité afin de détecter des discontinuités correspondant aux zones occultées. Une autre technique se base sur le principe de la bimodalité (par analyse d'histogramme des disparités) : les disparités des pixels voisins d'un pixel occulté appartiennent à deux régions de disparités différentes ; il faut alors détecter les deux extrema les plus proches dans l'histogramme des disparités de la région concernée afin de lever l'ambiguïté d'appariement.

Nous allons examiner ci-après d'autres types de techniques plus complexes permettant d'évaluer la qualité d'appariement. A partir d'une fonction de vraisemblance, un score de corrélation est attribué à chaque couple de pixels formé par le pixel à apparier de l'image gauche, et un pixel candidat dans la zone de recherche de l'image droite. L'ensemble des scores obtenus pour tous les pixels candidats forme une courbe (*cf. figure* 2.5). Si la fonction de vraisemblance est une fonction de similarité, plus le score de corrélation est élevé, plus les pixels sont similaires. [Cha05], fait un état de l'art sur les différentes fonctions de vraisemblance, en précisant l'intervalle des scores possible pour chacune des fonctions. Une des stratégies les plus courantes consiste à choisir le candidat ayant le meilleur score. Elle est connue sous l'appellation de "*Winner-take-all*" dans la littérature anglaise. À ce stade, aucune conclusion n'est tirée sur la qualité des appariements.

En se référant à la littérature, [DJMMR01] propose d'évaluer la qualité d'appariement d'un couple de pixels à partir de l'allure des scores de corrélation. L'auteur mesure deux critères : l'imprécision et l'ambiguïté. Ces deux critères évaluent l'impact du choix du meilleur candidat en fonction des scores des autres pixels candidats spatialement voisins. Notons que les deux critères suivants sont estimés sur une courbe de disparités, de sorte qu'un score minimum correspond au meilleur candidat :

- ***L'imprécision*** quantifie l'erreur probable de localisation du meilleur pixel candidat. Un appariement est considéré comme imprécis s'il existe plusieurs pixels candidats dans le voisinage du meilleur candidat, tels que la moyenne de leurs scores ne dépasse pas un certain seuil S_{imp} (figure 2.5).
- ***L'ambiguïté*** mesure l'erreur que peut générer un appariement lors du choix du candidat ayant le score minimum. Un appariement est ambigu si la différence entre les scores du meilleur et du deuxième pixel candidat ne dépasse pas un certain seuil S_{amb}. L'ambiguïté est

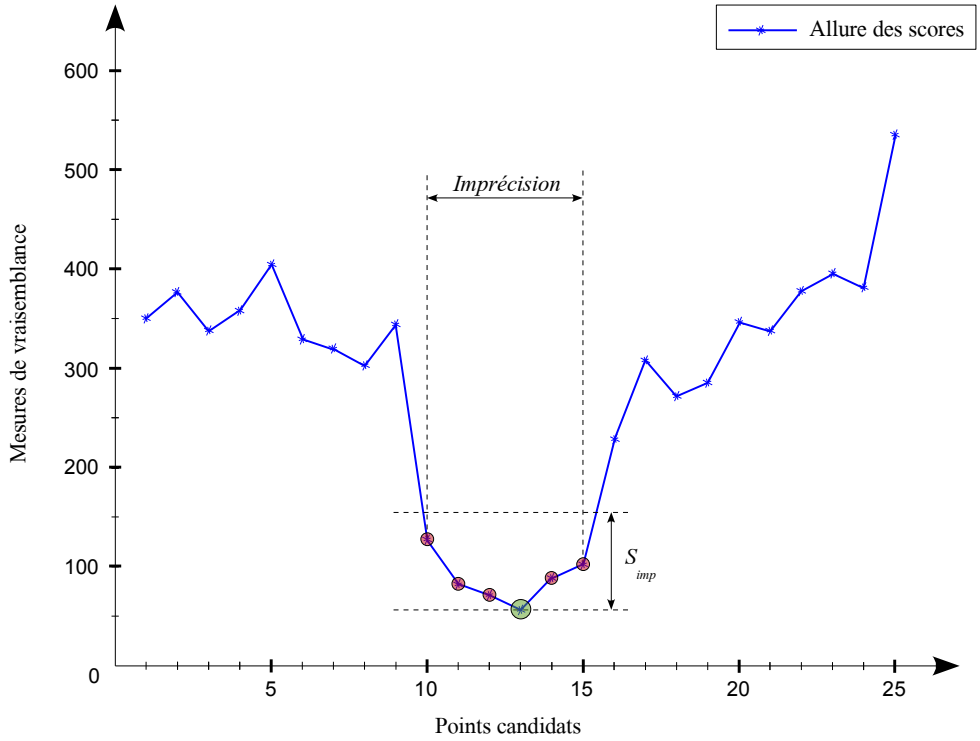


FIGURE 2.5 – Principe de mesure de l'imprécision d'appariement selon [DJMMR01].

mesurée par la distance entre le meilleur et le deuxième pixel candidat (figure 2.6).

[ZK02] part de l'hypothèse qu'un pixel bien apparié doit nécessairement avoir un score élevé. L'auteur se base sur le principe de la vérification croisée (contrainte de symétrie), la segmentation de l'image de référence en régions de couleur homogène, et le principe de la multi-résolution pour ajuster les scores initialement obtenus par une fonction de vraisemblance. Le niveau de résolution des disparités varie en fonction du niveau de segmentation. Chaque région est étiqueté en fonction du niveau de confiance relatif aux disparités initialement calculées. Une classification floue est effectuée comme suit :

$$L(s) = \begin{cases} VALID & \text{si } r \geq \alpha_2 \\ SEMIVALID & \text{si } \alpha_1 \leq r < \alpha_2 \\ INVALID & \text{si } r < \alpha_1 \end{cases} \tag{2.11}$$

Où L est la fonction d'étiquetage, r la moyenne des disparités dans le segment s, et α_1 et α_2 deux seuils positifs. $VALID$, $SEMIVALID$, et $INVALID$ qualifient l'étiquetage L, et corres-

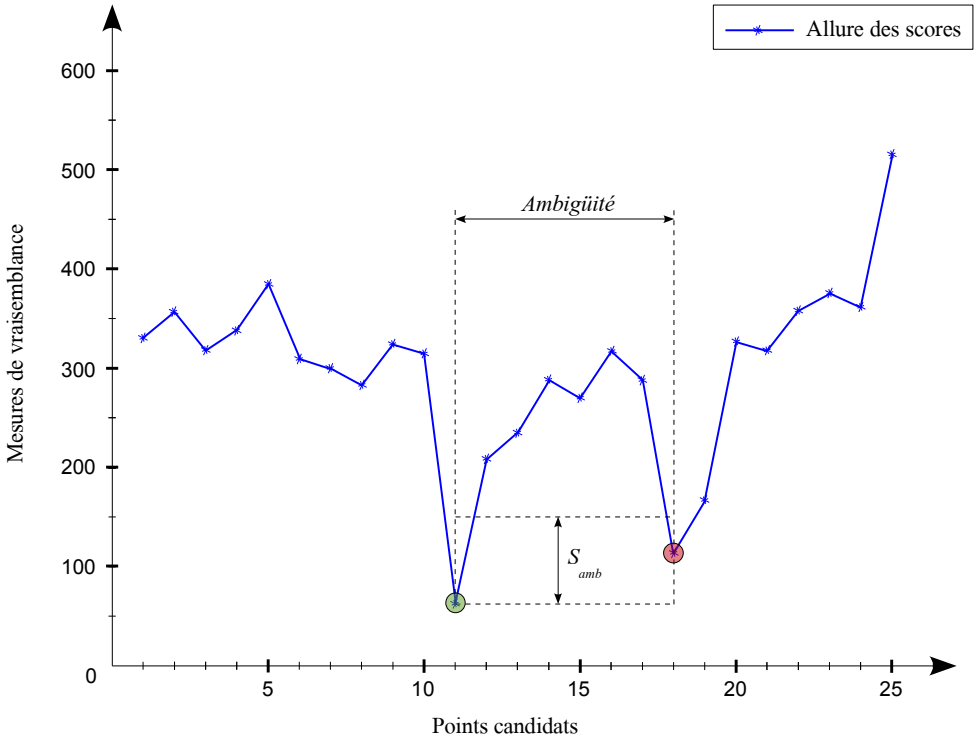


FIGURE 2.6 – Principe de mesure de l'ambiguïté d'appariement selon [DJMMR01].

pondent respectivement à une mesure de confiance élevée, moyenne, et basse.

En se basant sur la programmation dynamique, [GY03] propose de n'affecter une disparité qu'aux pixels ayant une mesure de confiance supérieure à un seuil donné. L'auteur utilise le terme "*reliability*" pour mesurer la certitude d'attribuer une disparité d au pixel p. Cette mesure est donnée par une fonction notée $R(pixel, disparit) \equiv R(p, d)$. La mesure de confiance pour un couple *(pixel, disparité)* est définie comme la différence entre le meilleur chemin ne passant pas par le pixel (p, d), et le meilleur chemin passant par le même pixel (p, d) dans l'espace de recherche. La mesure $R(p, d)$ représente le degré de confiance de la disparité d associée au pixel p. Avec la mesure d'incertitude, l'auteur applique un test de cohérence, fort et faible, afin de retenir ou non un appariement.

[EMW04] propose d'évaluer la qualité de la mise en correspondance afin d'identifier les erreurs d'appariement dues aux pixels appartenant à des régions peu texturées. Une fonction d'étiquetage est ainsi appliquée en classifiant les appariements en deux catégories : valide ou invalide. Un appariement est considéré comme invalide si la disparité obtenue est en dehors d'un intervalle de disparités délimité par un seuil, et valide sinon. Cette méthode est justifiée par le fait qu'un échec d'appariement, dû à un bruit lié au capteur, entre deux vues successives d'une même caméra, en-

traîne systématiquement un échec dans le processus d'appariement stéréoscopique. L'auteur propose d'évaluer l'appariement en analysant la courbure C de la courbe des scores :

$$C \triangleq 2.S_{opt} - S_{Left} - S_{Right} \tag{2.12}$$

Où S_{opt} est le score optimal obtenu par une fonction de vraisemblance donnée pour le meilleur candidat de rang (opt). S_{Left} et S_{Right} correspondent respectivement aux scores des pixels candidats à gauche et à droite du meilleur score. Une faible courbure signifie un mauvais appariement, probablement dû à l'appartenance du pixel à apparier à une région de couleur ou de texture uniforme.

[PNF+08] propose un système en temps réel de reconstruction 3D d'une scène urbaine à partir d'une vidéo stéréoscopique. L'auteur estime le plan de balayage optimal $\prod_m$ pour chaque pixel de l'image de référence. La précision d'appariement d'un couple de pixel (x, y) pour un plan $\prod_m$, est mesurée par une méthode heuristique. L'auteur suppose que la fonction de vraisemblance utilisée pour l'évaluation du degré de corrélation de l'ensemble $(x, y, \prod_m)$ peut être perturbée par un bruit de distribution gaussienne. L'idée est d'estimer la probabilité que la disparité correspondant au coût optimal ne varie pas même après avoir perturbé la fonction par un bruit gaussien. La probabilité P est estimée comme suit :

$$P = e^{-\left(C\left(x,y,\prod_m\right) - C\left(x,y,\widetilde{\prod}\right)\right)^2/\sigma^2} \tag{2.13}$$

Où le paramètre σ, défini empiriquement, dépend de l'amplitude du bruit. $\widetilde{\prod}$ représente le plan dans la direction du balayage qui minimise une certaine fonction de coût. $C\left(x, y, \prod_m\right) = min\{C_L\left(x, y, \prod_m\right), C_R\left(x, y, \prod_m\right)\}$ est le coût obtenu pour le couple de pixels (x, y) et un plan $\prod_m$. Les deux quantités $C_L\left(x, y, \prod_m\right)$ et $C_R\left(x, y, \prod_m\right)$ désignent les coûts associés au triplet $(x, y, \prod_m)$ dans les deux cas où l'image gauche et l'image droite sont considérées comme images de référence pour l'appariement. La mesure de confiance $c(x, y)$ est définie comme l'inverse de la somme des probabilités obtenues pour toutes les disparités possibles :

$$c(x,y) = \left(\sum_{\prod_m \neq \widetilde{\prod}} e^{-\left(C\left(x,y,\prod_m\right) - C\left(x,y,\widetilde{\prod}\right)\right)^2/\sigma^2} \right)^{-1} \tag{2.14}$$

[XJ08] se base sur une carte de confiance pour détecter les pixels appartenant à des régions partiellement occultées. La carte de confiance est calculée à partir d'une première carte de disparités. Chaque mesure de confiance est une valeur comprise entre 0 et 1 : plus la valeur de confiance est proche de 1, plus l'appariement est considéré comme bon. La carte de confiance des appariements, notée $U_l(x)$, est donnée par :

$$U_l(x) = \begin{cases} 1 & \text{si } |d_l(x) - d_r(x - d_l(x))| \geq 1 \\ T\left(\frac{b_x(d^*) - b_{min}}{\|b_0 - b_{min}\|}\right) & \text{si } b_x(d^*) > t \wedge |d_l(x) - d_r(x - d_l(x))| = 0 \\ 0 & \text{sinon} \end{cases} \tag{2.15}$$

Etant donné un pixel x, la mesure de confiance correspondante est mise à 1 si la contrainte de symétrie est violée. Ceci est exprimé par la condition suivante : $|d_l(x) - d_r(x - d_l(x))| \geq 1$, où $d_l(x)$ est la disparité du pixel x de l'image gauche, et $d_r(x)$ la disparité du pixel correspondant dans l'image droite. D'une façon générale, le degré de confiance dépend du coût d'appariement : un appariement ayant un coût élevé désigne probablement un pixel occulté ou appartenant à une région de couleur uniforme. Une mesure de confiance proche de 1 désigne un mauais appariement et inversement. L'auteur initialise les disparités à partir d'un algorithme de propagation de croyance. Pour le pixel x, le coût d'appariement est codé dans le message $b_x(d)$ qui correspond à la disparité d. $b_x(d^*)$ représente le message qui correspond au coût minimum pour le pixel x. Le paramètre t est un seuil fixe, b_0 est considéré comme la moyenne des croyances obtenues pour l'ensemble des pixels qui correspondent à la première condition de l'équation 2.15, et b_{min} est la moyenne des meilleures croyances retenues sur l'ensemble des pixels. La fonction T assure que les mesures de confiance sont comprises entre 0 et 1 :

$$T(y) = \begin{cases} 0 & \text{si } y < 0 \\ 1 & \text{si } y > 1 \\ y & \text{sinon} \end{cases} \tag{2.16}$$

2.6.2 Estimation de l'incertitude des appariements

2.6.2.1 Motivations

L'analyse des scores de corrélation obtenus sur l'ensemble des pixels candidats du pixel à apparier s'avère une démarche intéressante pour identifier les mauvais appariements. Un indice de

confiance est attribué à chaque appariement. Cet indice appartient à un intervalle qui varie selon la méthode d'évaluation utilisée. À titre d'exemple, la mesure de confiance proposée dans [XJ08] varie dans l'intervalle [0,1]. Une valeur proche de 1 signifie un bon appariement, alors qu'une valeur proche de 0 signifie un mauvais appariement. Différentes façons d'évaluer la qualité d'appariement ont été proposées dans la littérature. Nous proposons dans ce paragraphe une nouvelle méthode permettant une meilleure interprétation des scores obtenus par une fonction de vraisemblance donnée. Toutefois, un pixel appartenant à une zone occultée ou à une zone de couleur ou de texture uniforme peut engendrer un appariement ambigu. En examinant l'allure de la courbe des scores de corrélation, nous avons constaté que le comportement de la fonction de vraisemblance dépend du pixel à apparier. Les scores obtenus pour un pixel à apparier occulté ne sont pas distribués de la même façon que pour un pixel à apparier appartenant à une région uniforme. Nous sommes partis des postulats suivants :

– ***Postulat 1 :*** *Le voisinage d'un point caractéristique non occulté à apparier de l'image gauche est très semblable au voisinage de son homologue dans l'image droite.*

Le caractère discriminant d'un point caractéristique de l'image gauche influe sur le comportement de la fonction de vraisemblance. L'homologue d'un point caractéristique de l'image gauche, est souvent unique dans l'image droite. Un point caractéristique peut être un coin, un pixel d'un contour, un pixel dont les caractéristiques spatio-colorimétriques sont uniques. Le voisinage d'un point caractéristique marque ainsi l'unicité de ce point : pour un point caractéristique donné dans l'image gauche, il n'existe qu'un seul pixel dans l'image droite dont le voisinage est similaire. L'application d'une fonction de vraisemblance locale est sensée fournir, pour un pixel caractéristique, un score discriminant distinct des scores des autres candidats.

– ***Postulat 2 :*** *Pour un pixel à apparier appartenant à une région de couleur homogène ou de textures répétitives, il existe un grand nombre de pixels candidats qualifiés comme les meilleurs.*

Un pixel appartenant à une région de couleur uniforme ou de textures répétitives est considéré comme un pixel ambigu et non discriminant. Le nombre de pixels candidats qualifiés comme les meilleurs parmi l'ensemble des candidats possibles, est donc élevé. Un pixel appartenant à une région de couleur uniforme est un exemple d'appariement ambigu. Dans ce cas, le voisinage du pixel à apparier est homogène, et il existe une multitude de pixels candidats dans l'image droite dont le voisinage est très semblable à celui du pixel à apparier. Les scores ainsi obtenus sont très proches. Ceci montre l'importance du nombre des meilleurs candidats parmi l'ensemble des candidats possibles. Plus le nombre des meilleurs candidats est élevé, plus l'appariement est ambigu, et vice-versa.

– ***Postulat 3 :*** *La variation des disparités des meilleurs pixels candidats donne une idée de la région dans laquelle le pixel à apparier appartient.*

Un pixel appartenant à une région de texture uniforme engendre probablement un appariement ambigu. Les meilleurs candidats sont bien espacés et sont répartis uniformément, et les scores obtenus sont très proches. Les disparités correspondantes sont ainsi bien différentes. En effet, l'analyse de la moyenne des disparités des pixels potentiellement candidats semble une caractéristique intéressante.

– ***Postulat 4 :*** *Le bruit présent dans les images d'entrées agit uniformément sur les scores obtenus pour l'ensemble des candidats.*

Les sources de bruit peuvent être, selon [EMW04], le bruit du capteur CCD, la quantification, et l'échantillonnage. Le bruit peut être local ou global. Le score optimal obtenu avec une fonction de vraisemblance donnant des mesures de dissimilarité, se rapproche de 0 si le bruit diminue, et augmente avec le bruit. La distribution des scores reste inchangée sur l'ensemble des candidats pour une fonction de vraisemblance donnée.

Ces différentes constatations ont permis d'identifier certains paramètres jouant un rôle plus ou moins discriminant dans l'évaluation de la qualité d'appariement. Ceci fait l'objet du paragraphe suivant.

2.6.2.2 Paramètres pris en compte

Compte tenu des postulats précédents, la fonction de confiance que nous proposons dépend des quatre paramètres suivants :

−*Score optimal (S_{opt}) :* il s'agit du meilleur score obtenu avec la fonction de vraisemblance utilisée. Dans le cas de la fonction de dissimilarité proposée, le candidat retenu est celui qui correspond au score minimum, selon la stratégie *"Winner-take-all"*. L'importance de ce paramètre est justifiée par le postulat 4 : la présence de bruit remonte le score minimal, et les scores du reste des candidats sont aussi élevés. Plus le score minimum est grand, plus l'appariement est considéré comme imprécis. Un score élevé peut être interprété de deux façons : le pixel à apparier n'a pas d'homologue à cause d'une occultation, ou le pixel à apparier appartient à une région bruitée, et peut correspondre à plusieurs candidats. Un score faible ne correspond pas forcément à un bon appariement puisque, à ce stade, l'évaluation de la qualité d'appariement n'est pas discriminante.

−*Nombre des meilleurs candidats (τ) :* nous avons mentionné ci-dessus qu'un score minimal

peut ne pas correspondre à un vrai appariement. Un bon appariement est caractérisé par son caractère discriminant. La discrimination est évaluée selon le nombre des meilleurs candidats. Plus τ est élevé, plus l'appariement est imprécis, et vice-versa. Ce paramètre s'avère d'une importance majeure puisqu'il représente un bon indice d'évaluation. L'estimation de ce paramètre est détaillée plus loin dans ce paragraphe.

−*Variance des disparités des τ meilleurs candidats (σ) :* comme mentionné dans les postulats 2 et 3, la moyenne des disparités des meilleurs candidats donne une idée de la région dans laquelle le pixel à apparier appartient. Une faible variation des disparités des τ candidats signifie que la disparité du meilleur candidat a de fortes chances d'être la vraie disparité. Le pixel à apparier appartient a priori à une région de discontinuité en profondeur. Une grande variation des disparités des τ candidats signifie que le pixel à apparier appartient à une région de couleur ou de texture uniforme. Le paramètre σ est considéré comme l'écart type des disparités des τ meilleurs candidats.

−*Ecart significatif entre scores successifs (ω) :* il s'agit de la différence entre les τ^{eme} et $(\tau+1)^{eme}$ meilleurs candidats. Ce paramètre mesure la distinguabilité des meilleurs scores. Seul, ce paramètre n'a pas un effet majeur sur la mesure de confiance. À titre d'exemple, un petit nombre des meilleurs candidats et une valeur élevée de ω, signifie que l'appariement a une confiance élevée et que le candidat choisi est probablement le bon.

2.6.2.3 Estimation du nombre des meilleurs candidats

Nous détaillons dans ce paragraphe la manière dont le nombre des meilleurs candidats est estimé. Dans un souci de clarté, nous allons nous baser sur trois exemples d'appariement afin de mieux comprendre le principe. Les exemples que nous allons développer concernent trois pixels issus de trois images différentes obtenues dans la base stéréoscopique [mid]. L'algorithme 1 décrit l'ensemble des étapes permettant l'estimation du nombre de candidats potentiels étant donné l'ensemble des scores des candidats.

Les pixels candidats appartenant au support S sont considérés lors de l'appariement du pixel de l'image gauche de coordonnées (u, v). L'ensemble des candidats sont tout d'abord classés par ordre croissant en fonction de leurs scores obtenus par une fonction de vraisemblance donnée. Le candidat ayant le meilleur score est classé au premier rang. Dans ce qui suit, nous n'avons retenu qu'un sous ensemble S_τ de candidats ayant les τ meilleurs scores. Le paramètre τ dépend de l'intervalle des disparités retenu pour chaque couple d'images. Nous l'avons plafonné empiriquement à 25 dans ce travail : ce choix est justifié par les deux constatations suivantes :

– Dans le cas d'un appariement correct, le rang du bon candidat est généralement petit, et

Algorithme 1 Estimation du nombre des meilleurs candidats (τ).

1 - Effectuer un appariement stéréoscopique sur l'ensemble des pixels de l'image gauche en utilisant une fonction de vraisemblance.

2 - Appliquer la stratégie "*Winner-take-all*" en faisant correspondre chaque pixel à apparier avec le candidat ayant le meilleur score.

3 - Estimer la fonction du taux d'accroissement moyen, noté η, de l'ensemble des scores déjà triés par ordre croissant.

4 - Caractériser les sauts de scores en appliquant une fonction notée ξ.

5 - Déduire le nombre des meilleurs candidats. Ceci correspond au rang du candidat maximisant la fonction ξ.

le nombre des meilleurs candidats est réduit. Un grand nombre de candidats n'a plus d'influence dans l'estimation de la confiance d'un appariement.

– Dans le cas d'un appariement ambigu, le nombre de candidats ayant des scores proches, est élevé. Ceci augmente le nombre des meilleurs candidats. Plus le nombre des meilleurs candidats est grand, plus l'appariement est considéré comme imprécis. Expérimentalement, nous avons constaté que les candidats au-delà du rang 25 n'ont pas d'influence pour juger de l'incertitude d'un appariement.

Nous rappelons que le score obtenu par une fonction de vraisemblance pour le couple de pixels $I_g^(u, v)$ de l'image gauche et $I_d^(u, v')$ de l'image droite, est noté $\phi\left(I_g(u, v), I_d(u, v')\right)$. Afin de simplifier les notations, le score est représenté par $\phi(u, v')$, où u et v' représentent les coordonnées du candidat $I_d(u, v')$. En partant de la courbe des candidats triés par ordre croissant, nous introduisons la notion de *rang* des candidats. Un candidat ayant un score $\phi(u, v')$, et situé au rang i, sera représenté par $\phi(u, v')_i$. La fonction η donnée par l'équation 2.17 permet d'estimer le taux d'accroissement moyen de la fonction ϕ.

$$\eta(\phi) = \frac{\phi(u, v')_m - \phi_1(u, v')}{m - 1} \qquad m \in [1, \ldots, k] \tag{2.17}$$

L'indice m désigne le rang du candidat $\mathcal{P}_d^S(u, v')$ parmi l'ensemble des candidats triés et

$\phi(u,v')_m$ désigne le score obtenu pour le candidat de rang m. La fonction suivante, notée ξ, permet de mettre en valeur les sauts significatifs entre les scores. La fonction ξ correspond à la différence entre les pixels successifs de la fonction η, réduite par le carré du rang de chaque candidat :

$$\xi(\phi_m) = \frac{\boldsymbol{\nabla}\eta_m}{m^2} \tag{2.18}$$

Où $\boldsymbol{\nabla}\eta_m = (\eta_m - \eta_{m-1})$ est la différence entre les scores successifs de la fonction η. La quantité $\boldsymbol{\nabla}\eta_m$ permet de caractériser l'écart entre les scores. L'introduction d'un terme de pénalité, m^2, permet de mettre en valeur les sauts significatifs en fonction de l'importance des rangs des scores. Le nombre des meilleurs candidats correspond au rang m qui maximise la fonction ξ :

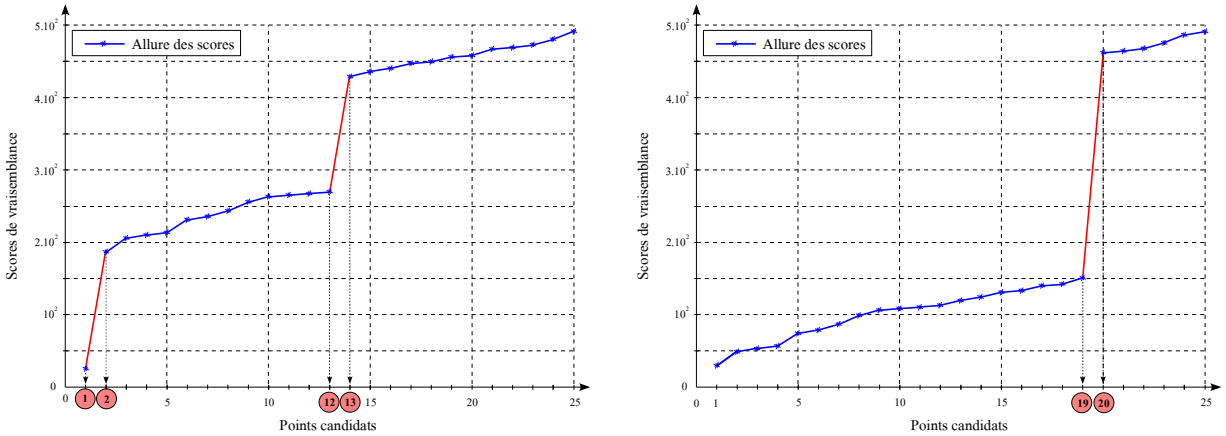
$$\tau = \arg\max_m \xi(\phi_m) \tag{2.19}$$

La figure 2.7 illustre l'importance de ce paramètre pour l'évaluation de la qualité d'appariement. À titre d'exemple, un saut entre les scores des candidats ayant des rangs faibles, est plus significatif qu'un saut entre les scores des candidats ayant des rangs élevés. Sur la figure 2.7, la courbe (a) montre l'importance du terme m de pénalité. Malgré l'existence de deux sauts remarquables, l'appariement est considéré comme non ambigu. Le premier saut entre les candidats de rang 1 et 2, est plus important que le deuxième saut entre les candidats 12 et 13. Plus le rang des sauts augmente, plus l'appariement est ambigu. La courbe (b) illustre le cas d'un appariement ambigu. Le saut significatif entre les scores figure entre les candidats de rang 19 et 20.

2.6.2.4 Nouvelle fonction de mesure de confiance

La fonction permettant l'établissement d'une mesure de confiance dépend des quatre paramètres précédemment décrits. Elle est notée ψ, et est détaillée dans l'équation 2.20. Une mesure de confiance est attribuée à chaque couple de pixels appariés. Elle permet d'évaluer la décision d'attribuer le candidat ayant le score optimal au pixel à apparier. La fonction de mesure de confiance peut être interprétée comme la probabilité d'associer le candidat $I_d(u,v')$ au pixel à apparier $I_g(u,v)$, étant donné un certain nombre de paramètres.

$$\psi\left(I_g(u,v), I_d(u,v')\right) \equiv P\left(I_d(u,v')/I_g(u,v), \tau, S_{opt}, \sigma, \omega\right) \tag{2.20}$$



(a) (b)

FIGURE 2.7 – Deux exemples correspondant aux scores obtenus par une fonction de vraisemblance pour deux pixels à apparier quelconques. Cette figure illustre l'importance des rangs des saut significatifs dans l'établissement de la qualité d'appariement. (a) appariement non ambigu malgré l'existence de deux sauts significatif : ceci est dû au rang du premier saut. (b) exemple d'un appariement ambigu : le saut significatif se situe entre deux candidats de rang élevé.

Ci-après la fonction de mesure de confiance exprimée selon les quatre paramètres :

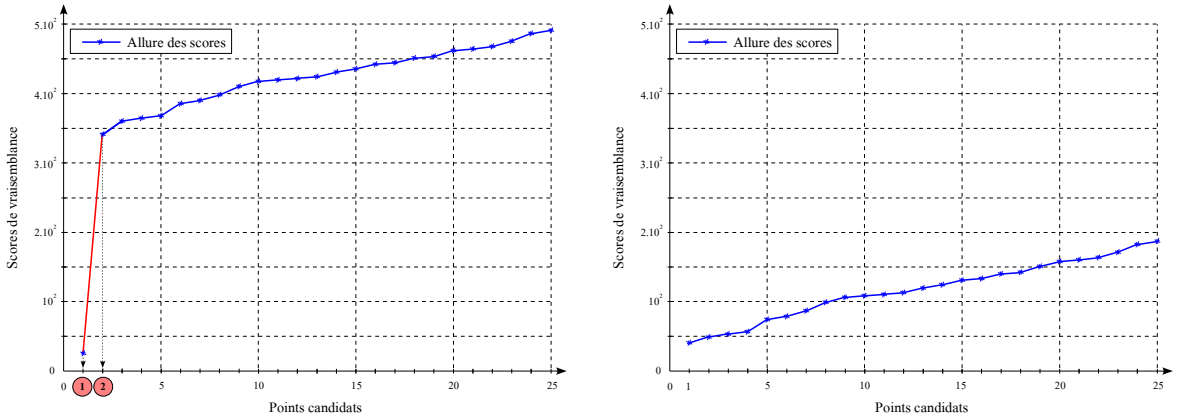
$$\psi(I_g(u,v), I_d(u,v')) = \left(1 - \frac{S_{opt}}{\omega}\right)^{\tau^2.log(\sigma)} \tag{2.21}$$

Le paramètre σ est remplacé par $log(\sigma)$ afin de diminuer l'impact des grandes valeurs de σ sur la mesure de confiance. De plus, certaines contraintes sont ajoutées afin de s'assurer que les mesures de confiance donnent des valeurs comprises entre 0 et 1. L'écart entre le score du τ^{eme} et $(\tau+1)^{eme}$ candidats, ω, doit être inférieur au score minimal S_{opt} :

$$\omega = \begin{cases} \omega & \text{si } S_{opt} \leq \omega \\ S_{opt} + 1 & \text{sinon} \end{cases} \tag{2.22}$$

La courbe (a) de la figure 2.8 montre l'allure des scores triés obtenus pour un pixel à apparier quelconque. Il s'agit d'un appariement non ambigu du fait qu'un saut significatif de scores existe entre le premier et le deuxième candidat. La courbe (b) correspond à l'allure de la courbe des

scores d'un appariement ambigu. Tous les candidats possibles sont considérés comme les meilleurs candidats puisqu'il n'existe pas un saut significatif entre deux scores successifs.



(a) (b)

FIGURE 2.8 – (a) Courbe des scores triés obtenus pour un pixel à apparier dont l'appariement est non ambigu du fait qu'un saut significatif de scores existe entre le premier et le deuxième candidat. (b) Courbe des scores d'un appariement ambigu. Tous les candidats possibles sont considérés comme les meilleurs candidats puisqu'il n'existe pas un saut important entre deux scores successifs.

2.6.2.5 Exemple d'un appariement non ambigu

La figure 2.9 illustre le cas d'un appariement non ambigu. Le pixel à apparier, de coordonnées $(135, 274)$, est un point caractéristique (coin) de l'image gauche Teddy [mid].

En se référant aux figures 2.10, 2.11 et 2.12, l'appariement d'un point caractéristique est considéré comme non ambigu puisqu'il existe un seul candidat ayant un score optimal. Ceci se traduit par le saut significatif qui existe entre le meilleur candidat et le candidat suivant. D'après la figure 2.10, le candidat de rang 20 correspond au score optimal. Le pixel candidat ayant un saut significatif est celui qui maximise la fonction ξ. Le rang de la valeur qui maximise la fonction ξ correspond au nombre des meilleurs candidats. Nous rappelons que les meilleurs candidats constituent un sous ensemble de l'ensemble des candidats.

2.6.2.6 Exemples d'appariements ambigus

Le premier exemple, figure 2.13, illustre le cas d'un pixel ambigu de coordonnées $(85, 310)$ appartenant à une région de couleur et de texture uniforme dans l'image gauche Sawtooth [mid]. Il existe plusieurs pixels candidats ayant des caractéristiques similaires à celles du pixel à apparier.

Il s'agit d'un appariement ambigu du fait que les attributs du pixel à apparier ne possèdent pas des caractéristiques discriminantes. D'après les figures 2.14, 2.15 et 2.16, il existe plusieurs

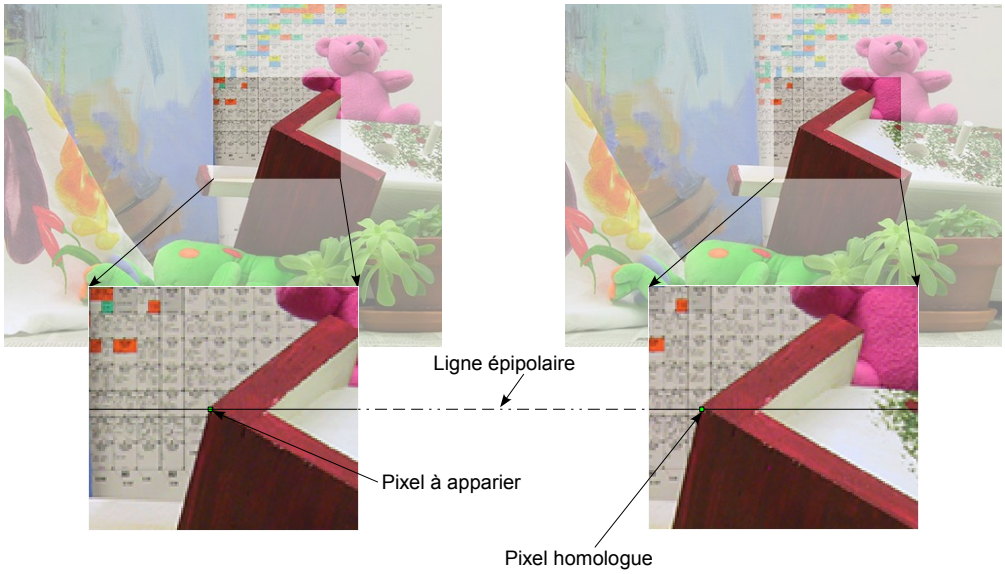


FIGURE 2.9 – Exemple illustrant le cas d'un point caractéristique de coordonnées $(135, 274)$ de l'image gauche Teddy de la base stéréoscopique [mid]. L'appariement n'est pas ambigu puisqu'il existe un seul pixel candidat dans l'autre image ayant des caractéristiques similaires.

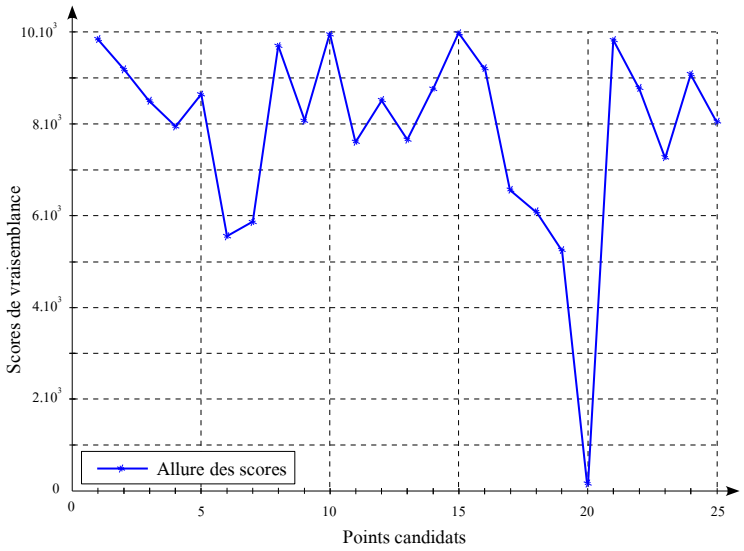


FIGURE 2.10 – Courbe correspondant aux scores non ordonnés, obtenus pour le pixel à apparier de coordonnées $(135, 274)$ de l'image gauche Teddy.

candidats ayant des scores proches. La fonction ξ de la figure 2.16 donne le nombre de candidats potentiels. La figure 2.16 montre que le pixel à apparier possède quatre meilleurs candidats parmi l'ensemble des candidats.

Le deuxième exemple, figure 2.17, illustre le cas d'un pixel occulté de l'image gauche Cones

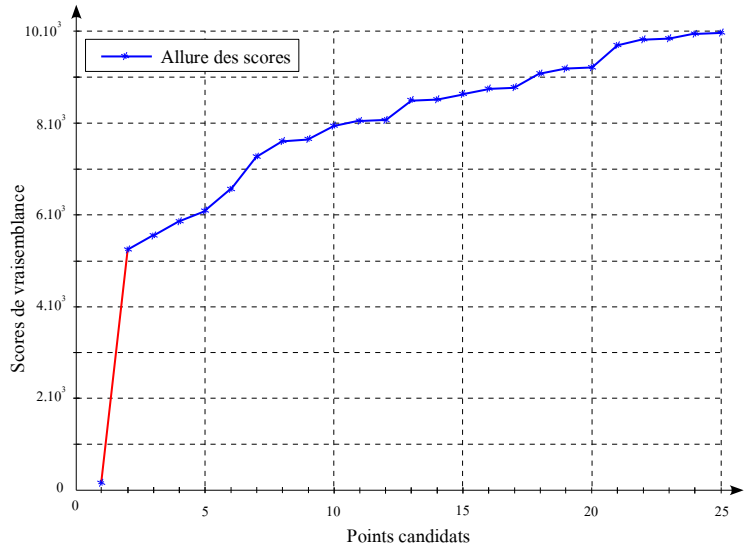


FIGURE 2.11 – Courbe des scores ordonnés, obtenus pour le pixel à apparier de coordonnées $(135, 274)$ de l'image gauche Teddy.

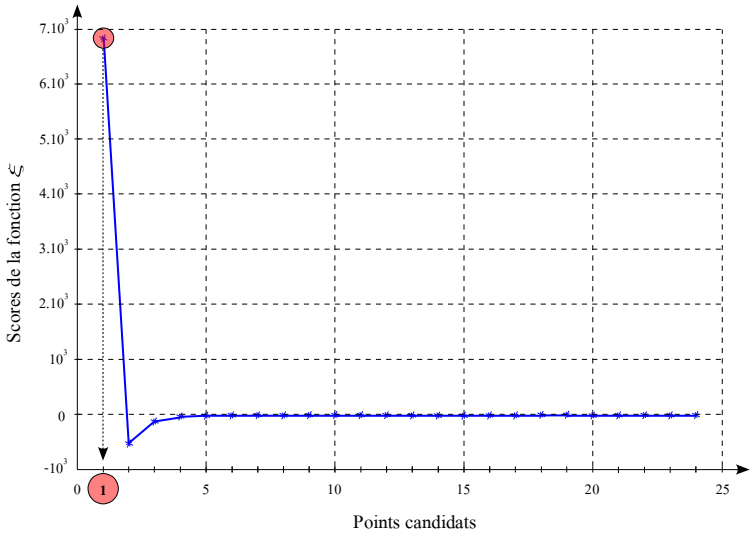


FIGURE 2.12 – La fonction ξ obtenus pour le pixel à apparier de coordonnées $(135, 274)$ de l'image gauche Teddy. Il existe un seul meilleur candidat, qui correspond au rang de la valeur maximale de la fonction ξ.

[mid]. Le pixel à apparier de l'image gauche n'a pas d'homologue sur l'image droite. Le pixel candidat le plus similaire présente un score très élevé. Il n'existe pas de saut significatif entre des scores successifs. L'appariement est alors considéré comme ambigu.

D'aprés les 2.18, 2.19 et 2.20, les scores des pixels candidats ne permettent pas d'identifier d'une manière unique et sûre l'homologue du pixel à apparier. Dans ce cas, le candidat ayant le

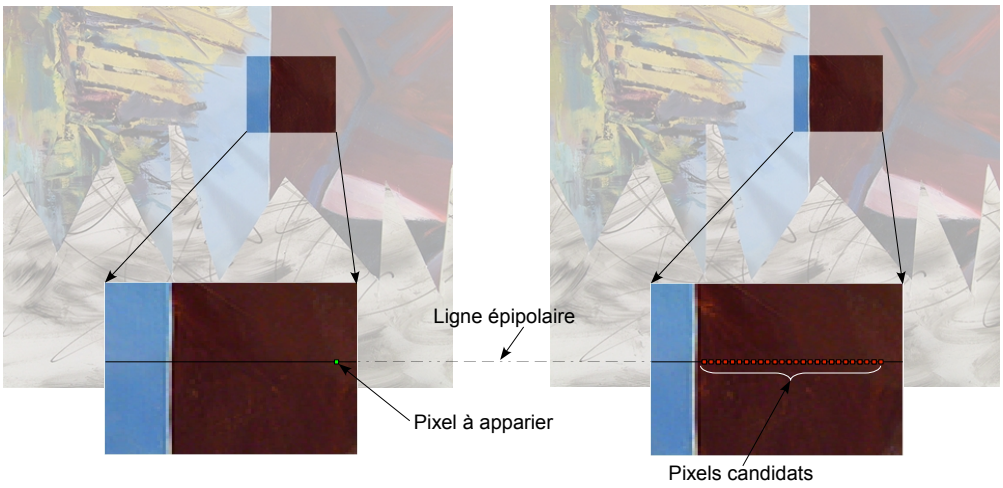


FIGURE 2.13 – Exemple illustrant le cas d'un pixel de coordonnées $(85, 310)$ de l'image gauche Sawtooth de la base stéréoscopique [mid] appartenant à une région de couleur et de texture uniforme. Il existe plusieurs meilleurs candidats dont leurs scores sont très proches.

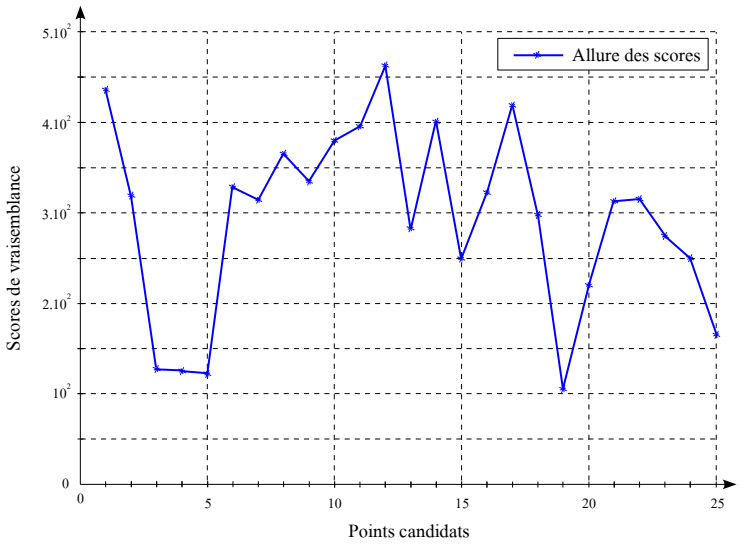


FIGURE 2.14 – Courbe correspondant aux scores non ordonnés, obtenus pour le pixel à apparier de coordonnées $(85, 310)$ de l'image gauche Sawtooth.

meilleur score ne correspond pas au vrai homologue. Aucun autre candidat n'est apte à être le correspondant du pixel à apparier. D'après la figure 2.20, il existe 12 meilleurs candidats. Nous allons voir dans la section suivante l'importance de ce paramètre dans l'évaluation de l'appariement.

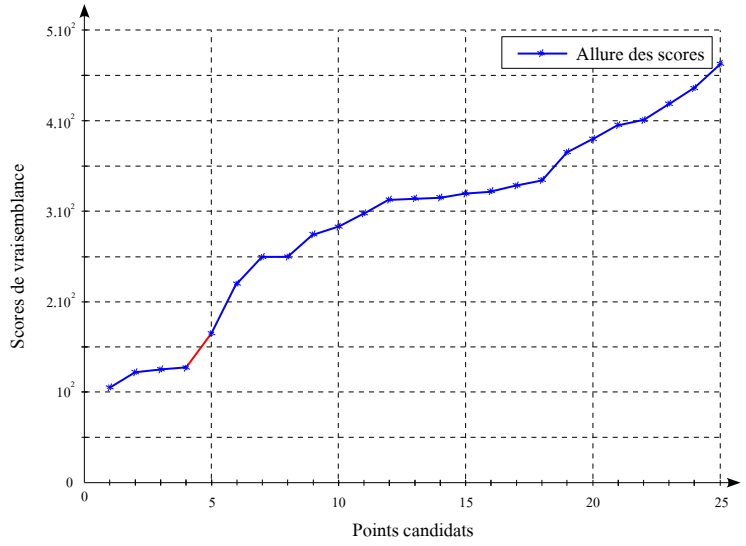


FIGURE 2.15 – Courbe des scores ordonnés, obtenus pour le pixel à apparier de coordonnées $(85, 310)$ de l'image gauche Sawtooth.

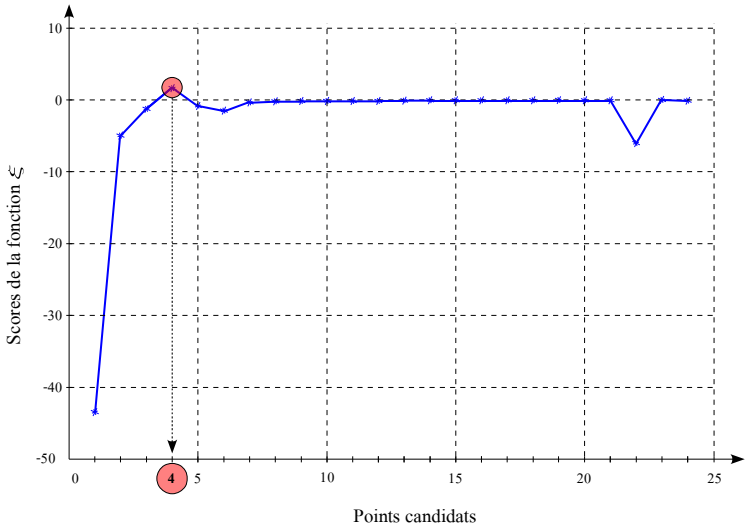


FIGURE 2.16 – La fonction ξ obtenus pour le pixel à apparier de coordonnées $(85, 310)$ de l'image gauche Sawtooth. Il existe quatre meilleurs candidats puisque la valeur maximale est au quatrième rang.

2.7 Cohérence spatio-colorimétrique par segmentation couleur

2.7.1 Concept de base

Comme nous l'avons vu au chapitre 1, il existe différentes méthodes d'appariement stéréoscopique pour résoudre le problème d'estimation des disparités dans des régions particulières de

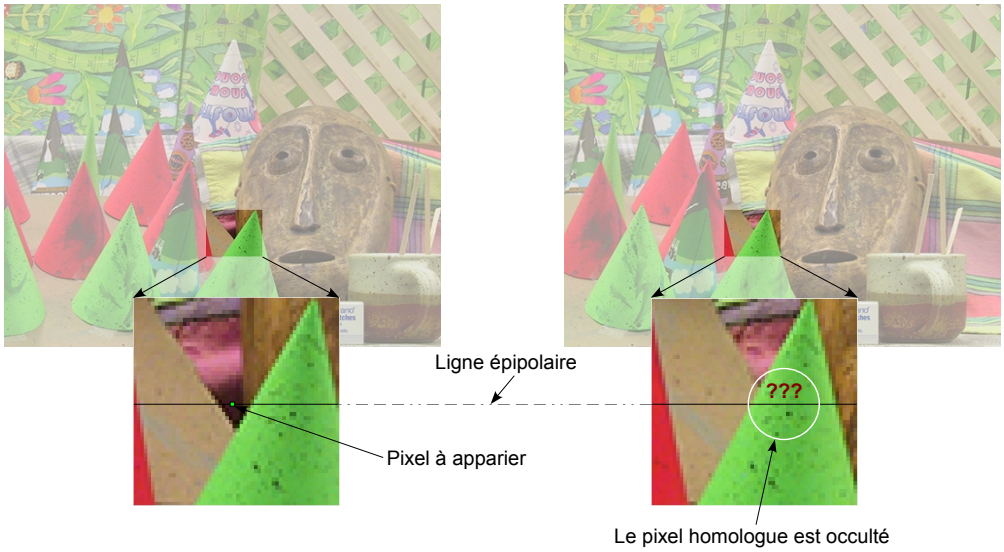


FIGURE 2.17 – Exemple illustrant le cas d'un pixel de coordonnées $(250, 235)$ de l'image gauche Cones, de la base stéréoscopique [mid], appartenant à une région occultée. Il existe plusieurs meilleurs candidats dont les scores sont très proches.

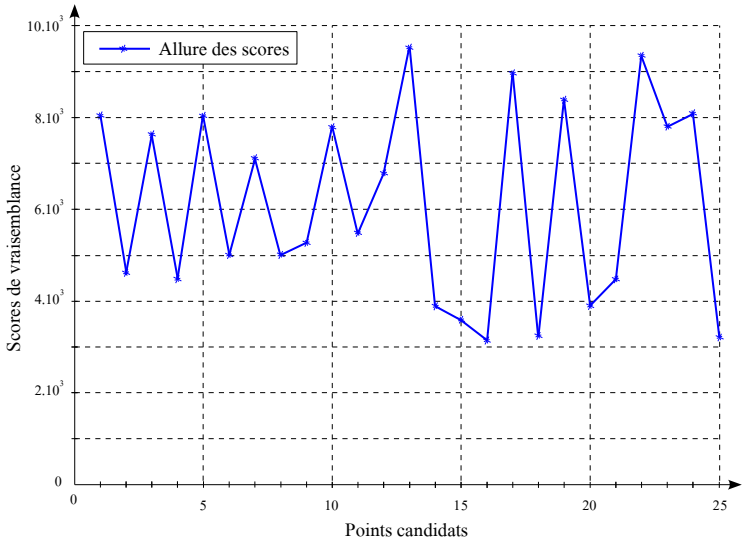


FIGURE 2.18 – Courbe correspondant aux scores non ordonnés obtenus pour le pixel à apparier de coordonnées $(250, 235)$ de l'image gauche Cones de la base stéréoscopique [mid].

l'image. Les zones occultées, les zones de texture répétitive, de couleur ou d'intensité uniforme, sont des régions particulièrement problématiques en stéréovision en raison des ambiguïtés d'appariement. En se référant aux travaux liés à la vision humaine, les chercheurs confirment l'importance de prendre en compte l'information spatio-colorimétrique dans le processus de vision tridimensionnelle. Cette observation a conduit les chercheurs à utiliser les contours pour la détection des

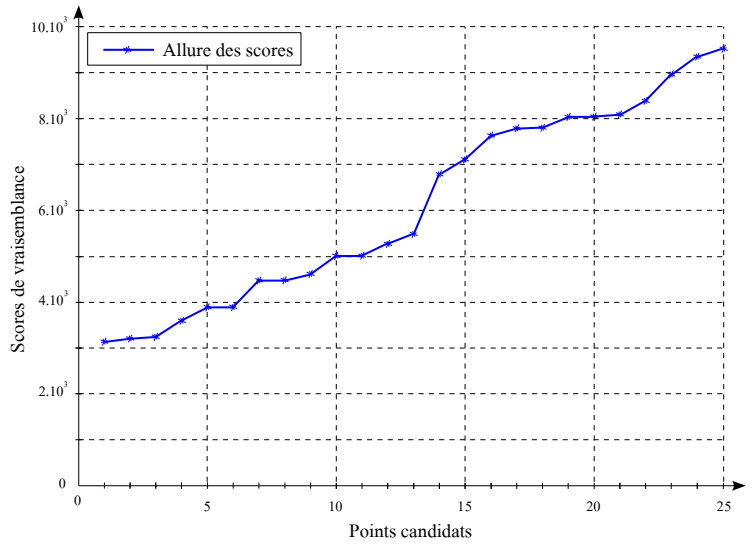


FIGURE 2.19 – Scores ordonnés obtenus pour le pixel à apparier de coordonnées $(250, 235)$ de l'image gauche Cones de la base stéréoscopique [mid].

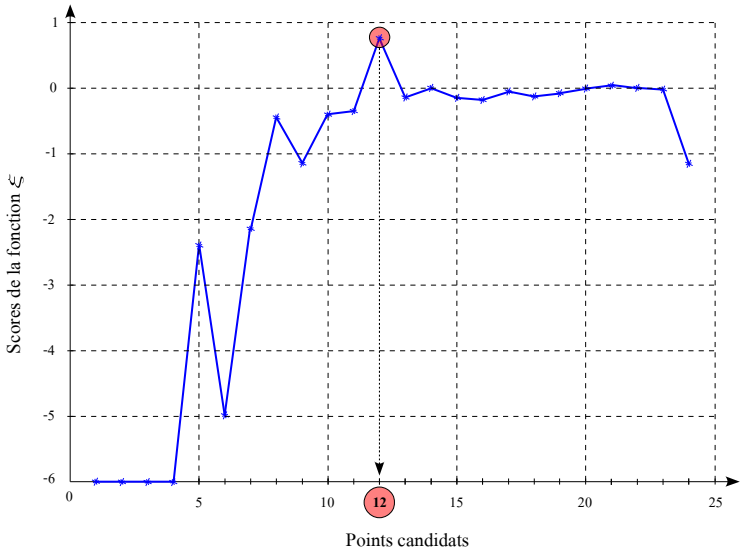


FIGURE 2.20 – La fonction ξ obtenus pour le pixel à apparier de coordonnées $(250, 235)$ de l'image gauche Cones. Il existe 12 candidats potentiels puisque le candidats qui maximise la fonction ξ est au douzième rang.

discontinuités de disparités. Par ailleurs, plusieurs travaux ont proposé d'introduire une nouvelle contrainte liée à la cohérence spatiale et colorimétrique par segmentation en régions homogènes. La segmentation peut être appliquée sur l'image de référence ou sur la carte de disparités. Les méthodes basées sur la segmentation partent des constatations suivantes :

– La segmentation basée sur la discontinuité de profondeur dans l'espace des disparités per-

met de regrouper les pixels selon leurs positions tridimensionnelles. Une région segmentée peut avoir des intensités ou des couleurs différentes. Deux régions voisines de disparités non homogènes, séparées par une discontinuité de profondeur, ont des disparités distinctes.

– Les régions de couleur homogène correspondant à des surfaces dont les disparités sont homogènes, appartiennent probablement à une même entité (ou objet). Une région donnée peut avoir une seule ou un ensemble de disparités qui varient d'une manière homogène (Une seule disparité dans le cas où le plan correspondant à la région est perpendiculaire aux axes optiques des caméras).

2.7.2 Avantages de la segmentation couleur

La continuité de disparité dans les régions homogènes est souvent introduite comme une contrainte de continuité dans les algorithmes d'appariement :

– La contrainte de lissage des disparités est explicitement introduite par l'hypothèse de la cohérence colorimétrique. Les disparités varient légèrement dans une même région, ce qui permet malgré tout d'estimer les disparités dans des segments de couleur uniforme.

– La plupart des algorithmes d'appariement permettent une reconstruction des reliefs d'un objet réel. Cette reconstruction est cependant loin d'être parfaite à cause des problèmes dûs aux plans fortement inclinés dans un même objet. La segmentation permet d'accroître la qualité des reliefs, et de repérer correctement les bordures des objets.

– Les algorithmes basés sur l'appariement des segments présentent l'avantage d'être plus rapides que ceux basés sur l'appariement des pixels, puisque le nombre de segments à apparier est plus réduit que le nombre de pixels de l'image.

– En réalité, l'estimation des disparités dans des régions occultées est possible en se basant sur l'image de référence segmentée. La disparité de la partie occultée d'une région peut être estimée à partir des disparités obtenues dans la partie visible de la même région.

– Une mauvaise segmentation entraîne une mauvaise localisation des bordures des objets. Une sur-segmentation semble alors utile pour résoudre ce problème [ZK07].

– La modélisation des objets de la scène observée permet d'avoir une estimation des surfaces. Les disparités ainsi obtenues sont plus précises que les disparités obtenues par segmentation. A titre d'exemple, les disparités d'un objet de forme cylindrique peuvent être précisément

estimées puisque l'évolution des disparités est décrite par les paramètres du modèle.

2.8 Ré-estimation des disparités par optimisation

2.8.1 Principe de la propagation de croyance

La propagation de croyance consiste à passer itérativement et parallèlement des messages entre pixels voisins dans un graphe non dirigé. La figure 2.21 illustre le principe du passage de message pour un voisinage 4-connexe. Cette méthode est itérative et parallèle, de telle sorte qu'un message $m^t_{p\to q}$ transmis à l'itération t du pixel p vers le pixel q dépend des messages $m^{t-1}_{s_1\to p}$, $m^{t-1}_{s_2\to p}$, et $m^{t-1}_{s_3\to p}$ reçus à l'itération $t-1$ par le pixel p de ses pixels voisins s_1, s_2, et s_3 respectivement. Les voisins d'un pixel p forment un ensemble noté $\mathcal{N}_p$. Un message est considéré comme un vecteur de labels possibles $\langle l^1, \ldots, l^n \rangle$ tels que $l \in \mathcal{L}$ et n est le nombre de labels qu'un pixel peut avoir. Un message $m^t_{p\to q}$ est calculé de la façon suivante :

$$m^t_{p\to q} = min_{l_p}\left(D_p(l_p) + V_{p,q}(l_p - l_q) + \sum_{s\in\mathcal{N}_p\setminus q} m^{t-1}_{s\to p}(l_p) \right) \tag{2.23}$$

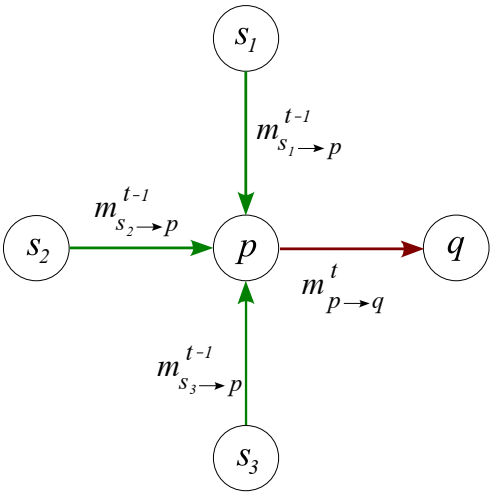


FIGURE 2.21 – Principe de passage de messages entre des pixels voisins (4-connexes). Le message transmis à l'itération t du pixel p vers le pixel q dépend des messages reçus par le pixel p à l'itération précédente $t-1$ de la part de ses voisins s_1, s_2, et s_3.

2.8.2 Propagation de Croyance Sélective (PCS)

Nous proposons dans cette section une nouvelle variante de la méthode de propagation de croyance. Notre apport consiste en l'introduction d'une mesure de confiance à chaque pixel du graphe, et l'extension du voisinage pris en compte lors du processus de passage de messages. La configuration standard, consistant à ne considérer qu'un voisinage 4-connexe pour le passage des messages dans le graphe, présente certains inconvénients. Le premier inconvénient est que, tous les messages ayant le même poids, aucune information n'est disponible sur le degré d'importance des message. L'effet direct est la mauvaise décision d'appariement. Un mauvais message reçu par un pixel p peut affecter le processus de mise à jour : en effet, ceci accroît le nombre d'itérations sans pour autant garantir d'atteindre l'optimum local. Le deuxième inconvénient est que le fait de se limiter aux 4 voisins n'apporte pas assez d'informations pour que la mise à jour soit efficace. Pour nous affranchir de ces limites, nous proposons un nouvel algorithme de propagation des croyances qui diffère des algorithmes existants par :

1. L'introduction des mesures de confiances lors du processus de passage des messages dans le graphe : chaque mesure de confiance influence la décision prise par les pixels voisins. Un pixel ayant une mesure de confiance élevée a ainsi plus d'influence sur le choix des disparités des pixels voisins.

2. La prise en compte d'un voisinage étendu et sélectif : un sous-ensemble des pixels voisins sont considérés lors de la mise à jour des disparités. Il s'agit de ne sélectionner que les *k-meilleurs proches voisins (k-mpv)*. La façon de choisir les *k-mpv* est détaillée par la suite.

La propagation de croyance sélective que nous proposons est résumée dans l'algorithme 2 :

2.8.2.1 Ré-estimation des disparités et propagation de croyance

Nous allons décrire dans ce paragraphe la deuxième étape de l'algorithme 2. Cette étape consiste en la mise à jour des disparités des pixels classés comme mal appariés. Les pixels mal appariés sont identifiés par leurs mesures de confiance faibles. Contrairement à la configuration des 4-voisins connexes, nous considérons un voisinage plus étendu, noté $\mathcal{N}$. Le voisinage $\mathcal{N}_p$ d'un pixel p mal apparié est choisi comme l'ensemble des pixels appartenant à une fenêtre 3D de côté variable β et centrée sur le pixel p. Cependant, ces pixels voisins ne sont pas tous considérés lors de la mise à jour de la disparité du pixel p. Seul un sous- ensemble de pixels, noté $\dot{\mathcal{N}}_p$, parmi les pixels de l'ensemble des voisins $\mathcal{N}_p$, sont pris en compte pour initialiser la mise à jour. Ces pixels sont choisis en fonction de leurs mesures de confiance $\dot{\mathcal{N}}_p = \{q \ ; \psi_q \geq \varrho_{mc} \ \ et \ \ q \in \mathcal{N}_p\}$. Le paramètre ϱ_{mc} est un seuil de confiance. Il se peut que certains pixels de l'ensemble $\dot{\mathcal{N}}_p$ aient une mesure de confiance élevée alors qu'il s'agit d'un mauvais appariement : ceci influence le processus de mise à

Algorithme 2 Propagation de Croyance Sélective PCS.

1 - Initialisation du graphe :

a - Attribuer un label à chaque pixel du graphe de l'image de référence. Un label correspond à la disparité du meilleur candidat ayant le score minimal obtenu par la fonction de vraisemblance DCMP (voir section 2.5.2). Ceci correspond au terme d'étiquetage, $E_{etiquetage}(f)$, défini dans l'équation 2.5 de la section 2.3.2.

b - Calculer une mesure de confiance ψ à chaque appariement. La fonction d'évaluation de la qualité d'appariement est détaillée dans la section 2.6.2.4. La mesure de confiance est vue comme un poids attribué à chaque nœud du graphe, permettant de définir le degré d'importance de chaque message.

2 - Pour chaque pixel p ayant une mesure de confiance inférieure à un seuil ϱ_{mc} :

a - Identifier l'ensemble $\mathcal{N}_p^*$ des meilleurs proches voisins *k-mpv* selon l'équation 2.24.

i - Si $\mathcal{N}_p^* = \emptyset$ alors retourner au début de l'étape 2. Sinon passer à l'étape suivante.

b - Mettre à jour la disparité du pixel p ainsi que la mesure de confiance correspondante. La nouvelle disparité est donnée par l'équation 2.26, et la mesure de confiance associée, ψ_p, est mise à jour par ϱ_{mc}.

c - Une fois que tous les pixels du graphe sont parcourus, calculer une énergie globale $E(f)$ pour la solution f, et passer à mettre à jour le seuil de confiance ϱ_{mc} = ϱ_{mc} - ϱ_{step}.

d - Arrêter si ϱ_{mc} = 0 ou si $E(f)$ est inférieur à un seuil d'énergie $E_{min}(f)$.

jour, puisque l'initialisation de l'algorithme de propagation des croyances dépend principalement des pixels considérés comme bien appariés. Ce genre d'erreur n'est pas facile à détecter puisqu'en réalité nous ne disposons pas des vraies disparités. Afin de réduire l'impact des pixels ayant des disparités aberrantes, nous proposons une fonction permettant l'identification des pixels aberrants. Nous définissons un nouvel ensemble, noté $\mathcal{N}_p^* \subset \dot{\mathcal{N}}_p$, contenant les pixels voisins au pixel p ayant une mesure de confiance élevée, et après avoir éliminé les pixels aberrants. Ce sous-ensemble de pixels choisis correspond au *k-meilleurs proches voisins (k-mpv)*. L'ensemble $\mathcal{N}_p^*$ est obtenu à partir de l'équation 2.24 :

$$\mathcal{N}_p^* : \Big\{ q_i / \underbrace{\frac{1}{\sharp\dot{\mathcal{N}}_p - 1} \sum_{i=1}^{\#\dot{\mathcal{N}}_p} \Delta_{q_i,q_j}}_{\widetilde{d}_i} < \mathcal{D} \Big\}; \quad \mathcal{D} = \beta\sqrt{3}, \;\; q_i \in \dot{\mathcal{N}}_p, \;\; et \;\; q_j \in \dot{\mathcal{N}}_p \backslash q_i \tag{2.24}$$

La quantité $\mathcal{D} = \beta\sqrt{3}$ représente la diagonale de la fenêtre 3D cubique de coté β. La distance euclidienne cumulée, noté Δ_{q_i,q_j}, est calculée dans l'espace $\mathbb{R}^3$. Il s'agit de la moyenne des sommes des distances euclidiennes entre un pixel voisin et le reste des voisins d'un pixel p de l'ensemble $\mathcal{N}_p^*$. Elle est donnée pour un pixel q_i par l'équation 2.25 :

$$\Delta_{q_i,q_j} = \left(\sum_{\substack{q_i \in \dot{\mathcal{N}}_p \\ q_j \in \dot{\mathcal{N}}_p \backslash q_i}} \Big(q_i(u) - q_j(u)\Big)^2 + \Big(q_i(v) - q_j(v)\Big)^2 + \Big(\hat{d}_{q_i} - \hat{d}_{q_j}\Big)^2 \right)^{1/2} \tag{2.25}$$

où $q_i(u)$ et $q_i(v)$ correspondent respectivement aux positions verticale et horizontale (ligne et colonne) du pixel q_i dans le graphe. La quantité $\hat{d}_{q_i}$ correspond à la disparité estimée au pixel q_i. La notation $\sharp\dot{\mathcal{N}}_p$ correspond au nombre des pixels voisins définis par l'ensemble $\dot{\mathcal{N}}_p$. La variable $\widetilde{d}_i$ correspond à la moyenne des distances euclidiennes cumulées pour une fenêtre 3D centrée sur le pixel i. L'équation 2.24 permet d'ignorer les pixels voisins ayant des disparités aberrantes mais des mesures de confiance élevées. Une disparité est erronée si elle n'est pas homogène avec les disparités voisines. Cette étape permet de minimiser les erreurs transmises dans les messages entre les pixels. La solution optimale est en effet rapidement atteinte puisque le nombre d'itérations se réduit considérablement. Cependant, un pixel p ne reçoit des messages que de ses *k-meilleurs proches voisins* $k - mpv \in \mathcal{N}_p^*$ ayant des mesures de confiance acceptables et des disparités homogènes. La disparité d'un pixel p est mise à jour en fonction des disparités transmises par les meilleurs voisins selon l'équation 2.26 :

$$\hat{d}_p^{\,t} = \frac{1}{\sum_{q_i \in \mathcal{N}_p^*} w_i} \sum_{q_i \in \mathcal{N}_p^*} \underbrace{(\mathcal{D} - \widetilde{d}_i)}_{w_i} . \hat{d}_{q_i}^{(t-1)} \tag{2.26}$$

Le processus de mise à jour se fait de manière séquentielle à l'intérieur des régions de couleur uniforme, obtenues lors de la segmentation par la méthode MeanShift [CM02]. La disparité mise à jour d'un pixel $p(u, v)$ est prise en compte pour la mise à jour de la disparité du pixel $p(u, v + 1)$ dans la même région. Après la mise à jour de la disparité du pixel p, la mesure de confiance associée ψ_p prend la valeur du seuil de confiance ϱ_{mc}. Le pixel p est alors utilisé lors de la mise à jour des disparités des pixels voisins à l'itération suivante. L'algorithme de mise à jour est itératif, et à chaque itération un ensemble de pixel $\mathcal{P}' \subset \mathcal{P}$ est mis à jour. À la fin de chaque itération, la solution f est quantifié par une énergie globale que l'on cherche à minimiser. Nous rappelons qu'une solution f est une configuration de labels, ou disparités, sur l'ensemble des pixels du graphe. Une énergie correspond au coût permettant l'obtention de la solution f. À chaque itération, le seuil de confiance diminue de ϱ_{step} et devient $\varrho_{mc} = (\varrho_{mc} - \varrho_{step})$. L'algorithme s'arrête quand l'énergie est suffisamment petite, ou, que le seuil de confiance a atteint une valeur minimale.

Nous rappelons que la fonction d'énergie à minimiser $E(f)$ se compose de deux termes : un terme d'étiquetage $E_{etiquetage}(f) = \sum_{p \in \mathcal{P}} D_p(l_p)$, et un terme de lissage $E_{lissage}(f) = \sum_{\substack{p \in \mathcal{P} \\ q \in \mathcal{N}_p}} V_{p,q}(l_p, l_q)$. Le terme d'étiquetage, est donné par l'équation 2.27 :

$$D_p(l_p) = \alpha . \sum_{\substack{p_g \in \mathcal{P} \\ p_d \in S}} \phi_{DCMP}(p_g, p_d) \tag{2.27}$$

La quantité $\phi_{DCMP}(p_g, p_d)$ correspond au coût de mise en correspondance obtenu lors de l'appariement des pixels p_g et p_d par la fonction de vraisemblance ϕ_{DCMP} décrite au §2.5.2. Le pixel p_g est un pixel de l'image gauche tel que $p_g \in \mathcal{P}$, et p_d est un pixel appartenant au support $\mathcal{S}$ de l'image droite. Le paramètre α est donné par l'équation 2.28 :

$$\alpha = \begin{cases} \psi_{p_g} & \text{si } \psi_{p_g} \geqslant \varrho_{mc} \\ 0 & \text{sinon} \end{cases} \tag{2.28}$$

ψ_{p_g} est la mesure de confiance obtenue pour le pixel p de l'image gauche (image de référence). Le terme de lissage est défini, pour un pixel p, par la somme des différences de la disparité l_p d'un pixel p et les disparités des *k-meilleurs proches voisins*.

La figure 2.22 illustre un exemple de régions contigues. Chaque région regroupe un ensemble de pixels dont les propriétés colorimétriques sont similaires. La région centrale correspond aux

pixels appariés avec des disparités différentes. Les pixels blancs correspondent aux pixels mal appariés. Le pixel noir correspond à un pixel mal apparié dont le processus de mise à jour de disparité est en cours. Seuls les pixels voisins appartenant à une fenêtre donnée et ayant une mesure de confiance élevée sont activés lors de la mise à jour. Les pixels verts correspondent aux voisins réellement bien appariés ayant une mesure de monfiance élevée. Le pixel rouge représente un pixel dont la confiance est importante alors qu'en réalité il s'agit d'un mauvais appariement. Aucun message n'est transmis par ce pixel lors de la mise à jour.

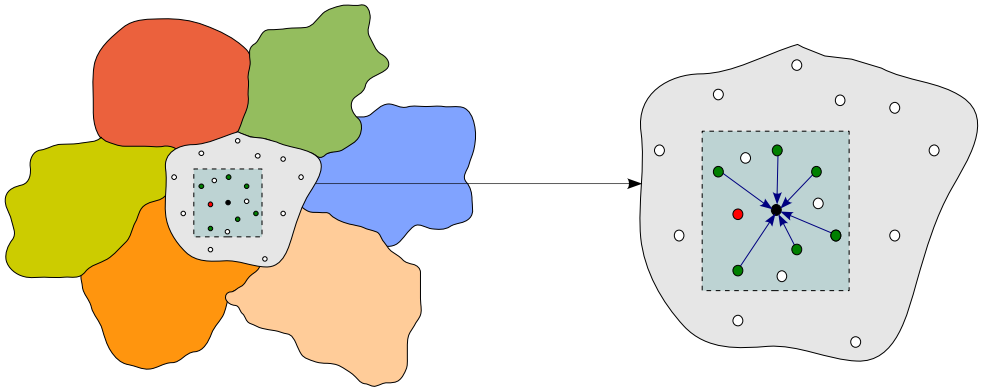


FIGURE 2.22 – Principe de propagation de croyance dans une région de couleur homogène.

La figure 2.23 présente le principe de passage de message dans l'algorithme de propagation de croyances sélective. Le pixel central, en cours mise à jour, reçoit des messages des *k-meilleurs proches voisins* (*k-mpv*). Les *k-mpv* sont des pixels ayant une mesure de confiance supérieure à un seuil donné. À titre d'illustration, les flèches montrent l'implication des meilleurs voisins dans le calcul de la distance cumulée $\Delta_{q_i,q_{j.}}$ entre un pixel q_i et le reste des meilleurs voisins $q_{j.}$. Cette distance sert à identifier les pixels voisins aberrants. C'est le cas du pixel q_{j_3}, qui est éliminé puisque sa distance cumulée est plus grande que $\mathcal{D}$.

2.9 Résultats expérimentaux

Nous présentons dans cette section une évaluation qualitative de notre algorithme d'appariement. L'évaluation porte sur les trois étapes de l'algorithme d'appariement proposé : l'estimation d'une première carte de disparités (approche locale), l'estimation de la qualité d'appariement par calcul de mesures de confiance (approche semi-globale), et la ré-estimation des disparités par propagation de croyance sélective (approche globale).

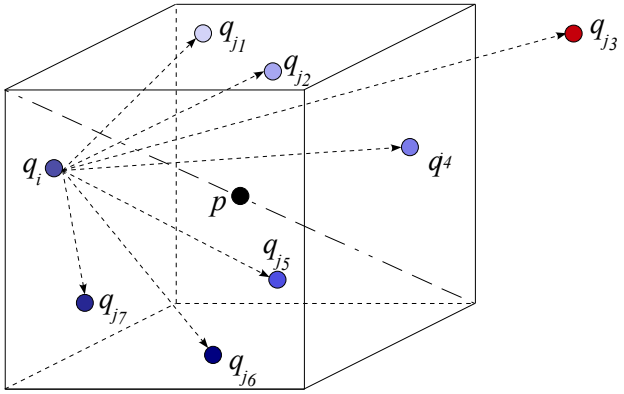


FIGURE 2.23 – Principe de la propagation de croyance sélective dans une fenêtre 3D.

2.9.1 Base d'images stéréoscopiques considérée

L'évaluation des performances d'un algorithme d'appariement consiste à mesurer sa capacité à produire une carte de disparités correcte à partir d'une paire d'images. En stéréovision, l'évaluation des performances d'un algorithme d'appariement se fait par l'analyse et l'évaluation du résultat d'appariement. Cependant, la connaissance préalable du résultat d'appariement n'est pas toujours possible, surtout dans le cas d'appariement d'images réelles sur lesquelles une "vérité terrain" n'est pas facile à obtenir. Une vérité terrain correspond à une carte contenant les disparités réelles obtenues généralement à partir d'autres techniques telles que le scanner laser ou la lumière structurée. Ces techniques sont souvent appliquées sur des scènes intérieures dont les conditions d'illumination sont bien contrôlées. Il existe d'autres moyens permettant l'obtention des vérités terrain, tels que les images de synthèse. L'avantage de ce genre de méthodes est que la vérité terrain obtenue pour une paire d'images stéréoscopiques est d'une précision absolue. Un premier inconvénient réside dans le fait que l'obtention d'images de synthèse stéréoscopiques est un processus long à réaliser. Un deuxième inconvénient est que les images traitées ne sont pas tout à fait réalistes, puisqu'elles n'intègrent pas les difficultés souvent rencontrées dans des scènes réelles et complexes. Parfois, on ne dispose que de paire d'images sans vérité terrain. Le seul moyen permettant l'évaluation des performances d'appariement est d'effectuer le test de symétrie, dit aussi test de vérification croisée, proposé dans [BC01].

Dans le cadre de notre évaluation, nous avons utilisé les paires d'images avec des vérités terrain, largement utilisées par la communauté en stéréovision. La première base est celle proposée par D. Scharstein et R. Szeliski, qui est disponible en ligne à l'adresse `http://vision.middlebury.edu/stereo/data/` [mid]. Des vérités terrain, obtenues par la méthode des lumières structurées, sont disponibles pour chaque paire d'images. Nous avons choisi six paires d'images intitulées "Cones", "Teddy", "Venus", "Tsukuba", "Sawtooth", et "Moebius". Les figures

2.24 et 2.25 sont les paires d'images utilisées pour l'appariement, ainsi que les vérités terrain correspondantes.

Nous détaillons ci-dessous le protocole d'évaluation sur lequel nous nous sommes basés, ainsi que les méthodes utilisées pour la comparaison. La section 2.9.2 traite des critères retenus pour l'évaluation de notre algorithme.

2.9.2 Méthodes comparées

Nous rappelons que notre algorithme est conçu pour apparier des images couleur. L'apport de la couleur dans le processus d'appariement stéréoscopique a été discuté dans plusieurs travaux récents tels que [Cha05], qui a démontré que la prise en compte de la couleur diminue le taux de faux appariements. Les vérités terrain des figures 2.24 et 2.25 correspondent aux disparités réelles obtenues pour chaque pixel de l'image de référence. Une faible disparité, qui correspond à un point éloigné des caméras, est illustrée par la couleur bleue. Une forte disparité, qui correspond à un point proche des caméras, est illustrée par la couleur rouge. Pour chaque paire d'images, les disparités varient dans l'intervalle $[d_{min}, d_{max}]$. Les zones en noir désignent les pixels dont la disparité est inconnue.

La première étape de notre algorithme consiste à estimer une carte initiale de disparité. Les fonctions de vraisemblance donnent des cartes de disparité denses, de sorte qu'une disparité est systématiquement attribuée à chaque pixel de l'image de référence. La fonction de vraisemblance proposée est comparée avec les trois méthodes *SAD*, *SSD*, et *NCC* :

– SAD - Somme des valeurs absolues des différences. La fonction de vraisemblance est donnée par l'équation 2.29 dans l'espace couleur *RVB*. Le score obtenu pour un couple de pixels varie, dans l'intervalle $[0, +\infty]$. Plus le score est petit, plus l'appariement est vraisemblable :

$$\phi_{SAD}(I_g(u,v), I_d(u,v')) = \sum_{i,j \in ZA} \sum_{c \in \{RVB\}} \left| I_g^c(u+i, v+j) - I_d^c(u+i, v'+j) \right| \quad (2.29)$$

– SSD - Somme des carrés des différences. La fonction de vraisemblance est donnée par l'équation 2.30 dans l'espace couleur *RVB*. Le score obtenu pour un couple de pixels, varie dans l'intervalle $[0, +\infty]$. Plus le score est petit, plus l'appariement est vraisemblable.


Image gauche *Cones*


Image droite *Cones*

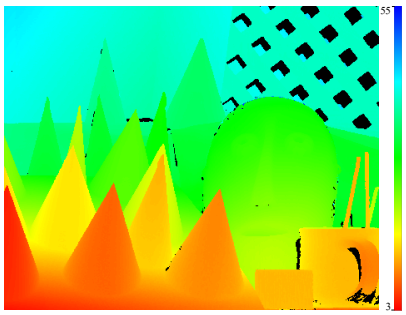

Vérité terrain *Cones*


Image gauche *Teddy*


Image droite *Teddy*

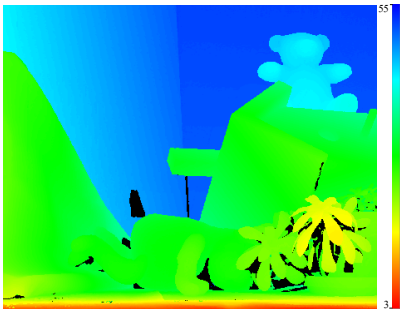

Vérité terrain *Teddy*


Image gauche *Venus*


Image droite *Venus*

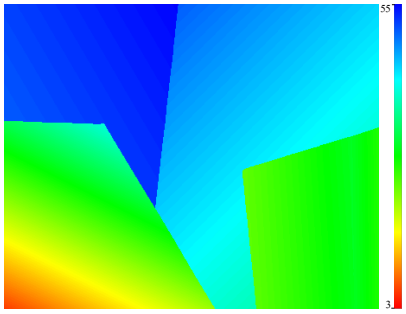

Vérité terrain *Venus*


Image gauche *Tsukuba*

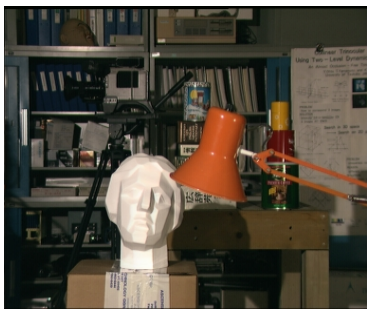


Image droite *Tsukuba*

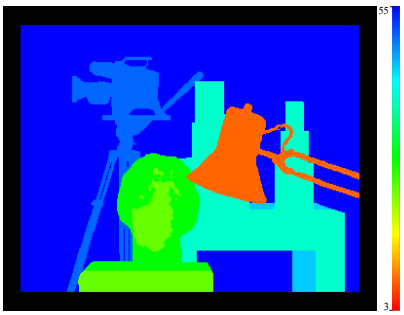
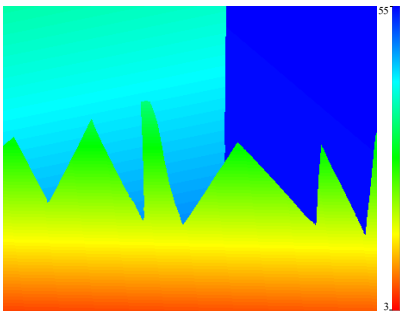

Vérité terrain *Tsukuba*

FIGURE 2.24 – Base d'images stéréoscopiques considérée pour l'évaluation.


Image gauche *Sawtooth*

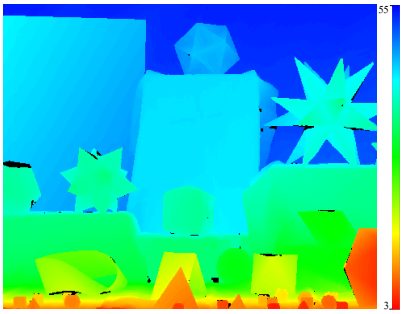

Image droite *Sawtooth*



Vérité terrain *Sawtooth*


Image gauche *Moebius*


Image droite *Moebius*



Vérité terrain *Moebius*

FIGURE 2.25 – Base d'images stéréoscopiques considérée pour l'évaluation (suite).

$$\phi_{SSD}(I_g(u,v), I_d(u,v')) = \left(\sum_{i,j \in ZA} \sum_{c \in \{RVB\}} \left(I_g^c(u+i, v+j) - I_d^c(u+i, v'+j) \right)^2 \right)^{1/2} \tag{2.30}$$

– NCC -Mesure de corrélation croisée normalisée. La fonction de vraisemblance est donnée par l'équation 2.31 dans l'espace couleur *RVB*. Le score obtenu pour un couple de pixels, varie dans l'intervalle $[0, 1]$. Plus le score est proche de 1, plus l'appariement est vraisemblable.

$$\phi_{NCC}(I_g(u,v), I_d(u,v')) = \frac{\sum_{i,j \in ZA} \sum_{c \in \{RVB\}} \left(I_g^c(u+i, v+j) . I_d^c(u+i, v'+j) \right)}{\left\| \mathcal{N}_{I_g} \right\| . \left\| \mathcal{N}_{I_d} \right\|} \tag{2.31}$$

$\mathcal{N}_{I_g \wedge I_d}$ correspond au voisinage des pixels I_g et I_d. Dans le cadre des fonctions de vraisemblance étudiées, le voisinage appartient à une fenêtre carrée de côté e, centrée sur les pixels à apparier. Les indices i et j appartiennent à l'intervalle $\left[-\left\lfloor \frac{e}{2} \right\rfloor, +\left\lfloor \frac{e}{2} \right\rfloor\right]$, où $\lfloor x \rfloor$ correspond à la partie entière de x. Dans l'équation 2.31, $\left\|\mathcal{N}_{I_{g \vee d}}\right\|$ est un coefficient de normalisation calculé sur l'ensemble du voisinage $\mathcal{N}$ de chaque pixel I_g et I_d :

$$\left\|\mathcal{N}_{I_{g \vee d}}\right\| = \left(\sum_{i,j \in \mathcal{N}_{I_{g \vee d}}} \sum_{c \in \{RVB\}} \left(I^c_{g \vee d}(u+i, v+j) \right)^2 \right)^{1/2} \tag{2.32}$$

La deuxième partie de notre algorithme consiste à identifier les paires de pixels bien mis en correspondance à partir de leurs mesures de confiance. A ce niveau, nous n'évaluons que l'apport des mesures de confiance dans l'identification des appariements corrects. De ce fait, nous nous basons sur les disparités des vérités terrain pour estimer le *Taux d'Appariements Corrects TAC* en fonction de la densité de la carte de disparité. La dernière partie consiste en la réestimation itérative des disparités permettant l'obtention de cartes de disparité denses. Les cartes obtenues sont comparées avec les trois méthodes globales suivantes : H-Cut [MMI09], MaxProduct [FH06], et PhaseBased [EEAHM07].

2.9.3 Protocole d'évaluation

L'évaluation des performances des différentes fonctions de vraisemblance est effectuée sur l'ensemble des pixels de l'image de référence, noté $\mathcal{P}_{all}$. Le nombre d'éléments de l'ensemble $\mathcal{P}_{all}$ est donné par la notation $Card(\mathcal{P}_{all})$. L'ensemble des appariements corrects est noté par AC^s_{all}. Le nombre d'appariements corrects, noté $\#AC^s_{all}$, correspond au nombre de pixels tels que la différence de disparités entre les disparités réelles et les disparités estimées ne dépasse pas un seuil s. Ce nombre, calculé sur l'ensemble des pixels de l'image de référence, est donné par l'équation 2.33. Notons par TAC^s_{all} le taux d'appariement correct calculé pour un seuil s. Le TAC^s_{all} est donné par l'équation 2.34 :

$$\#AC^s_{all} = \left\{ p \in \mathcal{P}_{all} / \left| d_p - \hat{d}_p \right| \leq s \right\} \tag{2.33}$$

$$TAC^s_{all} = \frac{\#AC^s_{all}}{Card(\mathcal{P}_{all})} \tag{2.34}$$

Dans l'équation 2.33, d_p et $\hat{d}_p$ correspondent respectivement aux disparités réelle et estimée pour le pixel p. Nous proposons ci-après une évaluation qualitative de la fonction de vraisemblance proposée, notée ϕ_{DCMP}, comparée avec les trois fonctions ϕ_{SAD}, ϕ_{SSD}, et ϕ_{NCC} données par les équations 2.29, 2.30, et 2.31.

L'évaluation des appariements en fonction des mesures de confiance, ainsi que la ré-estimation des disparités erronées par propagation de croyance sélective (PCS), sont effectués sur trois types de régions : l'ensemble des pixels de l'image $\mathcal{P}_{all}$, les pixels non occultés $\mathcal{P}_{nonocc}$, et l'ensemble des pixels appartenant à une zone de discontinuité de profondeur $\mathcal{P}_{disc}$. Le TAC est calculé de la même façon que pour l'ensemble des pixels. L'ensemble $\mathcal{P}_{nonocc}$ correspond aux pixels de l'image de référence non occultés. L'ensemble $\mathcal{P}_{disc}$ correspond aux pixels de l'image de référence qui correspondent aux discontinuité en profondeur. Les régions de discontinuité en profondeur et les régions occultées n'existent que dans les images "Cones", "Teddy", "Venus", et "Tsukuba". La figure 2.26 illustre les régions sur lesquelles l'évaluation s'est basée. Les images (a), (b), (c), et (d) correspondent respectivement aux vérités terrain des paires d'images "Cones", "Teddy", "Venus", et "Tsukuba". Les régions occultées sont illustrées par des pixels noirs, tandis que les régions de discontinuité en profondeur sont illustrées par des pixels blancs.

2.9.4 Evaluations

La figure 2.27 montre les courbes correspondant aux taux d'appariements corrects obtenus pour les quatre fonctions de vraisemblance pour les paires d'images "Cones", "Teddy", "Venus", "Tsukuba", "Sawtooth", et "Moebius" de la base stéréoscopique Middlebury [mid]. Le taux d'appariements corrects augmente avec la taille de la fenêtre d'agrégation, puisque le voisinage pris en compte lors de l'appariement comporte un grand nombre de pixels. L'application d'une petite fenêtre de taille 3×3 permet de conserver les contours des objets, préserve les discontinuités en profondeur, et permet d'obtenir une carte dense de disparité en temps réel, mais ne permet pas d'avoir un TAC élevé, en raison des erreurs au niveau des régions de couleur et de texture uniformes. À l'inverse, l'utilisation d'une grande fenêtre a l'avantage de couvrir une large zone lors de l'appariement, permettant de résoudre en partie les inconvénients des fenêtres de petite taille. L'inconvénient majeur réside dans le temps de calcul important, et dans la non conservation des contours dans les régions de discontinuités de profondeur. Expérimentalement, il a été montré dans [CS09] que le temps de calcul des disparités varie exponentiellement en fonction de la taille de la fenêtre d'agrégation. La figure 2.27 montre que la fonction de vraisemblance que nous avons développée donne un TAC supérieur aux autres méthodes évaluées sur des fenêtre de petite taille. L'écart diminue à mesure que la taille de la fenêtre d'agrégation augmente.

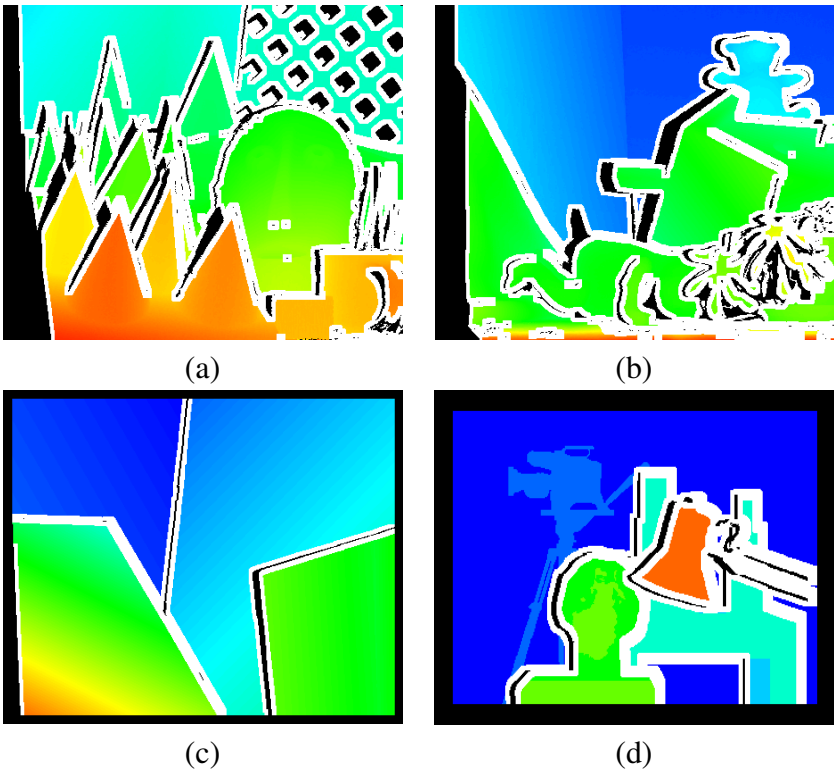


FIGURE 2.26 – Les régions considérées lors de l'évaluation. Les pixels noirs correspondent aux régions occultés . Les pixels blancs constituent les régions de discontinuité de profondeur.

Les figures 2.28 et 2.29 montrent les cartes de disparités denses des paires d'images "Cones" et "Tsukuba" respectivement, obtenues avec la fonction de vraisemblance ϕ_{DCMP} ainsi que les trois autres fonctions de vraisemblance évaluées. Nous avons choisi des fenêtres de taille 3×3, 11×11 et 19×19 afin d'étudier l'effet de la taille de la fenêtre d'agrégation sur la qualité d'appariement. Les images (a.1), (a.2), (a.3), et (a.4) montrent les cartes de disparités obtenues pour une fenêtre de taille 3×3 pour les fonctions de vraisemblance ϕ_{DCMP}, ϕ_{SAD}, ϕ_{SSD} et ϕ_{NCC} respectivement. Les images (b) et (c) correspondent aux cartes de disparités obtenues pour des fenêtres de tailles 11×11 et 19×19 respectivement. La dernière ligne montre les erreurs d'appariement (les pixels noirs), en se référant à la vérité terrain. Les erreurs d'appariement concernent la carte de disparités obtenue avec une fenêtre 3×3. Afin de réduire le temps de traitement, nous choisissons pour la suite de l'évaluation la fonction de vraisemblance ϕ_{DCMP} avec une fenêtre d'agrégation de taille 3×3.

Nous rappelons qu'une fonction de vraisemblance donne des scores pour chaque paire de pixels. La stratégie "*Winner-take-all*" consiste à choisir le couple de pixels candidats ayant le meilleur score. Les fonctions de vraisemblance ϕ_{SAD} et ϕ_{SSD} donnent des scores dans l'inter-

valle $[0, +\infty[$ alors que la fonction ϕ_{NCC} donne des scores dans l'intervalle $[0, 1]$. Les scores de la fonction ϕ_{DCMP} sont dans l'intervalle $[0, +\infty]$ de telle sorte qu'un score faible correspond à un bon appariement. Une paire de pixels ayant un score minimal ne correspond donc pas systématiquement à un bon appariement. La figure 2.30 présente des cartes des scores, données pour chaque fonction de vraisemblance. Les scores sont dans l'intervalle $[0, 255]$ de telle sorte qu'un score faible est représenté par un pixel sombre.

L'introduction d'une mesure de confiance permet de classer les pixels appariés en deux classes : pixels bien appariés et mal appariés. Les appariements classés comme corrects correspondent aux pixels ayant une mesure de confiance supérieure à un seuil ϱ_{mc}. La figure 2.31 montre des courbes qui correspondent aux TAC^1, où l'indice 1 correspond à un seuil (c'est à dire la différence entre une disparité réelle et une disparité estimé ne doit pas dépassé 1 pixel), calculés sur les ensembles $\mathcal{P}_{all}$, $\mathcal{P}_{nonocc}$, et $\mathcal{P}_{disc}$ des images "Cones", "Teddy", "Venus", et "Tsukuba". Plus le seuil de confiance augmente, plus le TAC^1 augmente. Le rapport entre le TAC^1 et la densité de la carte de disparités est présenté dans le tableau 2.1. L'exemple concerne trois seuils de mesure de confiance pour l'image "Teddy". Plus le seuil augmente, plus le nombre $\#AC^1_{all}$ d'appariements corrects diminue.

ϱ_{mc}	10%	40%	80%
TAC^1_{all}	90.29%	91.73%	93.03%
Densité	35.60%	21.56%	5.60%

TABLE 2.1 – Le rapport entre le taux d'appariements corrects TAC^1_{all} et la densité de la carte de disparité pour l'image "Teddy". La densité correspond au rapport entre le nombre de pixels, noté $\#AC^1_{all}$, ayant une mesure de confiance supérieure à ϱ_{mc} et le nombre de pixels total de l'image $\#\mathcal{P}_{all}$.

Dans ce qui suit, l'algorithme de Propagation de Croyance Sélective (PCS) est évalué pour une configuration donnée de paramètres. Certains paramètres sont fixés pour la segmentation par *MeanShift* (α_{ms}, β_{ms}, et γ_{ms}), la largeur de la Fenêtre d'Agrégation FA (L_{FA}), et les seuils retenus pour la propagation elle même (L_{FC}, $k - mpv$, et ϱ_{mc}). Le paramètre L_{FC} désigne la largeur de la fenêtre cubique prise en compte lors de la propagation. La figure 2.32 illustre les différentes étapes de l'algorithme proposé sur l'ensemble des images "Cones", "Teddy", "Venus", et "Tsukuba".

Le tableau 2.3 résume les Taux d'Appariement Incorrect $TAI^1 = 1 - TAC^1$ calculés sur les ensembles $\mathcal{P}_{all}$, $\mathcal{P}_{nonocc}$, et $\mathcal{P}_{disc}$ des images "Cones", "Teddy", "Venus", et "Tsukuba". Les TAI^1 obtenus avec notre méthode sont comparés avec les trois méthodes globales *H-Cut* [MMI09], *Max-Product* [FH06], et *PhaseBased* [EEAHM07].

MeanShift			$DCMP$	*PCS*		
α_{ms}	β_{ms}	γ_{ms}	L_{FA}	L_{FC}	$k-mpv$	ϱ_{mc}
7	7	15	3	4	6	0.6

TABLE 2.2 – Les paramètres pris en compte pour l'évaluation de l'algorithme de propagation de croyance sélective.

Algorithme	*Tsukuba*			*Venus*			*Teddy*			*Cones*		
	nonocc	*all*	*disc*	*nonocc*	*all*	*disc*	*nonocc*	*all*	*disc*	*nonocc*	*all*	*disc*
H-Cut	2.85	4.86	14.4	1.73	3.14	20.2	10.7	19.5	25.8	5.46	15.6	15.7
PCS	**4.87**	**5.04**	**8.47**	**3.42**	**3.99**	**10.5**	**17.5**	**20.8**	**28.0**	**7.46**	**12.5**	**13.3**
Max-Product	1.88	3.78	10.1	1.31	2.34	15.7	24.6	32.4	34.7	21.2	28.5	30.1
PhaseBased	4.26	6.53	15.4	6.71	8.16	26.4	14.5	23.1	25.5	10.8	20.5	21.2

TABLE 2.3 – Comparaison des TAI^1 de l'algorithme proposé et des méthodes de références sur la base [mid].

Nous montrons dans la figure 2.33 les différentes étapes de notre algorithme d'appariement sur deux régions extraites manuellement des images de référence "Cones" et "Teddy". Le premier exemple montre que les disparités d'une région non occultée sont homogènes et lisses. La région extraite du deuxième exemple correspond à une peluche ainsi que les zones de discontinuité en profondeur. La propagation des disparités respecte les limites de régions, les disparités ne se propageant qu'à l'intérieur des régions de couleur uniforme. En se référant au tableau 2.3, la précision de l'appariement est amélioré par la contrainte de cohérence colorimétrique. Nous montrons par la figure 2.32 une comparaison visuelle des cartes de disparités denses des méthodes évaluées et celle obtenue avec notre algorithme.

2.10 Conclusion

Nous avons présenté dans ce chapitre un nouvel algorithme d'appariement stéréoscopique combinant les avantages des méthodes locales et globales. La qualité des appariements et le temps de traitement sont les deux critères pris en compte lors de la conception de l'algorithme. La première étape consiste à attribuer une disparité à chaque pixel de l'image de référence, en utilisant une méthode locale avec une fenêtre d'agrégation de taille réduite. Une première carte de disparités est ainsi obtenue. Cette première carte comprend habituellement des erreurs d'appariement dues aux occultations et aux régions de couleur uniforme et de texture répétitive. Ensuite, nous avons proposé un nouvel algorithme de réestimation et de rectification de disparité, basé sur la propagation de croyance. L'algorithme consiste à propager itérativement des messages dans un graphe formé par les pixels de l'image de référence. À chaque itération, les disparités sont rectifiées en fonc-

tion des messages reçus des *k-mpv*. Afin d'accélérer le traitement, nous avons introduit les deux contraintes de cohérence spatiale et colorimétrique, et la notion de confiance lors du processus de propagation de croyance. Nous avons appelé cette dernière "*Propagation de Croyance Sélective*" puisque nous avons contraint le nombre de messages qui circulent durant une itération dans une région spatialement et colorimétriquement cohérente, aux pixels ayant une confiance élevée. Cette procédure a permis de réduire la complexité algorithmique et une accélération du traitement, comparé aux algorithmes standards, tout en améliorant la qualité d'appariement.

Nous nous sommes basés sur la constatation qu'une région de couleur uniforme appartient généralement à un même objet physique. L'intégration de la contrainte de cohérence spatiale et colorimétrique est justifiée par le fait que les disparités dans une même région varient d'une façon homogène et lisse. Cette contrainte permet de diminuer à chaque itération le nombre candidats, qui dépendent des disparités des pixels d'une même région ayant une forte mesure de confiance. À ce stade, une question se pose sur la possibilité d'intégrer d'autres contraintes afin d'accélérer et d'améliorer des performances d'appariement, et l'applicabilité de notre algorithme sur des images réelles, souvent bruitées. Puisque nous nous sommes intéressés aux applications de détection d'obstacles dans des environnements réels, une nouvelle contrainte peut s'ajouter dans le processus d'appariement. Il s'agit de *la contrainte de mouvement*. Comme pour les contraintes de cohérence spatiale et colorimétrique, la contrainte de mouvement stipule que les disparités d'une même région en mouvement varient aussi de manière homogène. Le fait de limiter l'appariement sur les régions affectées par le mouvement permet d'une part d'accélérer le processus d'appariement, et d'autre part d'améliorer la qualité d'appariement. Cependant, le mouvement d'un objet ne peut être estimé que sur une séquence d'images. La segmentation des images selon le critère de mouvement est un problème encore posé en vision artificielle. Dans le cas des environnements extérieurs, il existe différents problèmes à résoudre, tels que les variations continues d'illumination et le mouvement répétitif et continu de certains objets du fond. Nous proposons dans le chapitre suivant un nouvel algorithme permettant la détection des régions affectées par un mouvement dans une séquence d'images.

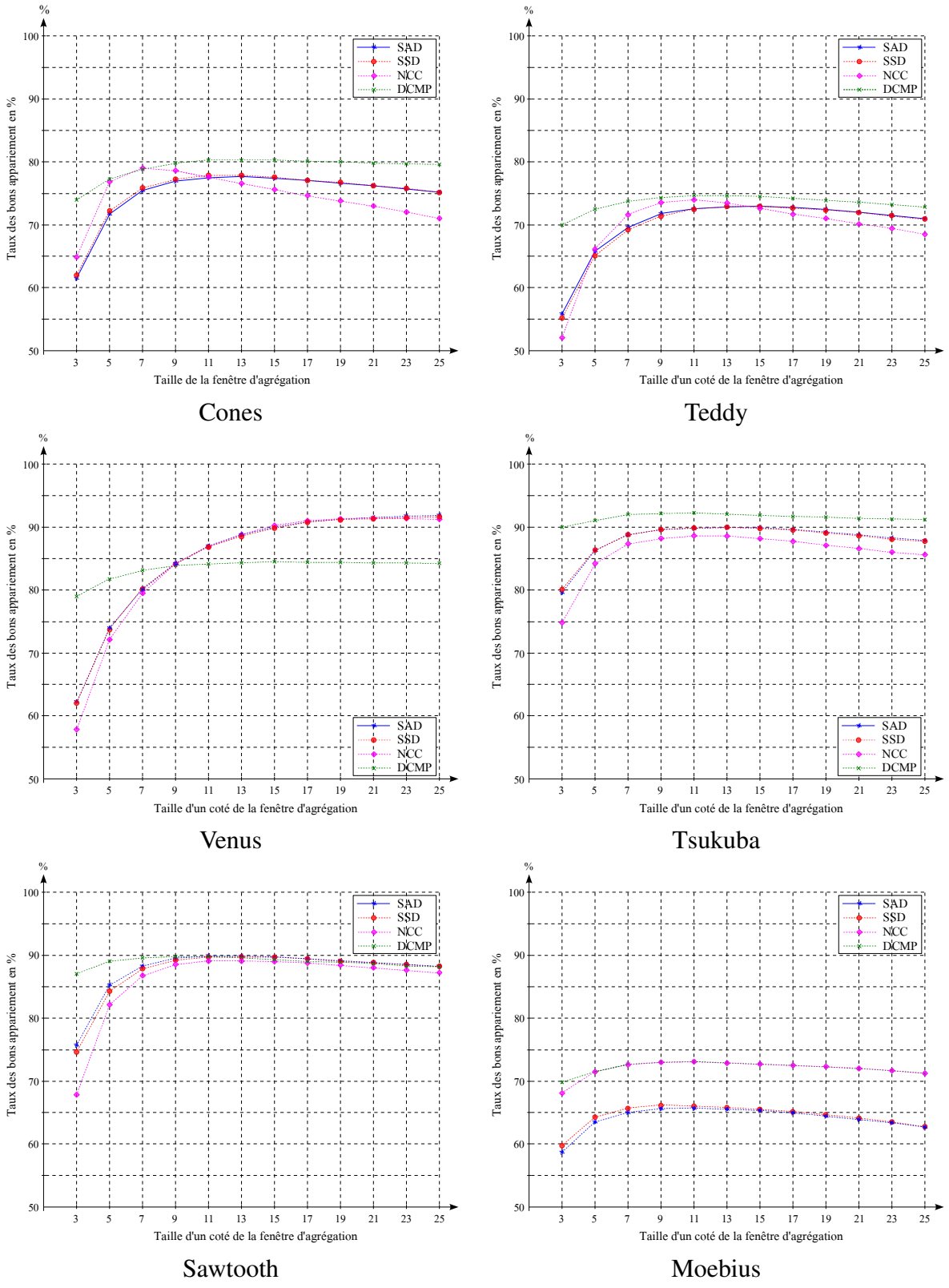
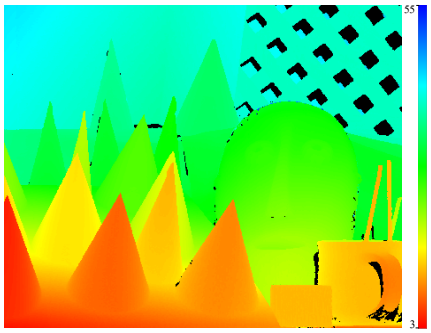


FIGURE 2.27 – Courbes des taux d'appariements corrects obtenues avec les fonctions de vraisemblance ϕ_{DCMP}, ϕ_{SAD}, ϕ_{SSD} et ϕ_{NCC} en fonction de la taille des fenêtres d'agrégation, qui varient de 3×3 jusqu'à 25×25. Les TAC^1_{all} sont calculé pour un seuil $s = 1$.

Image gauche



Vérité terrain

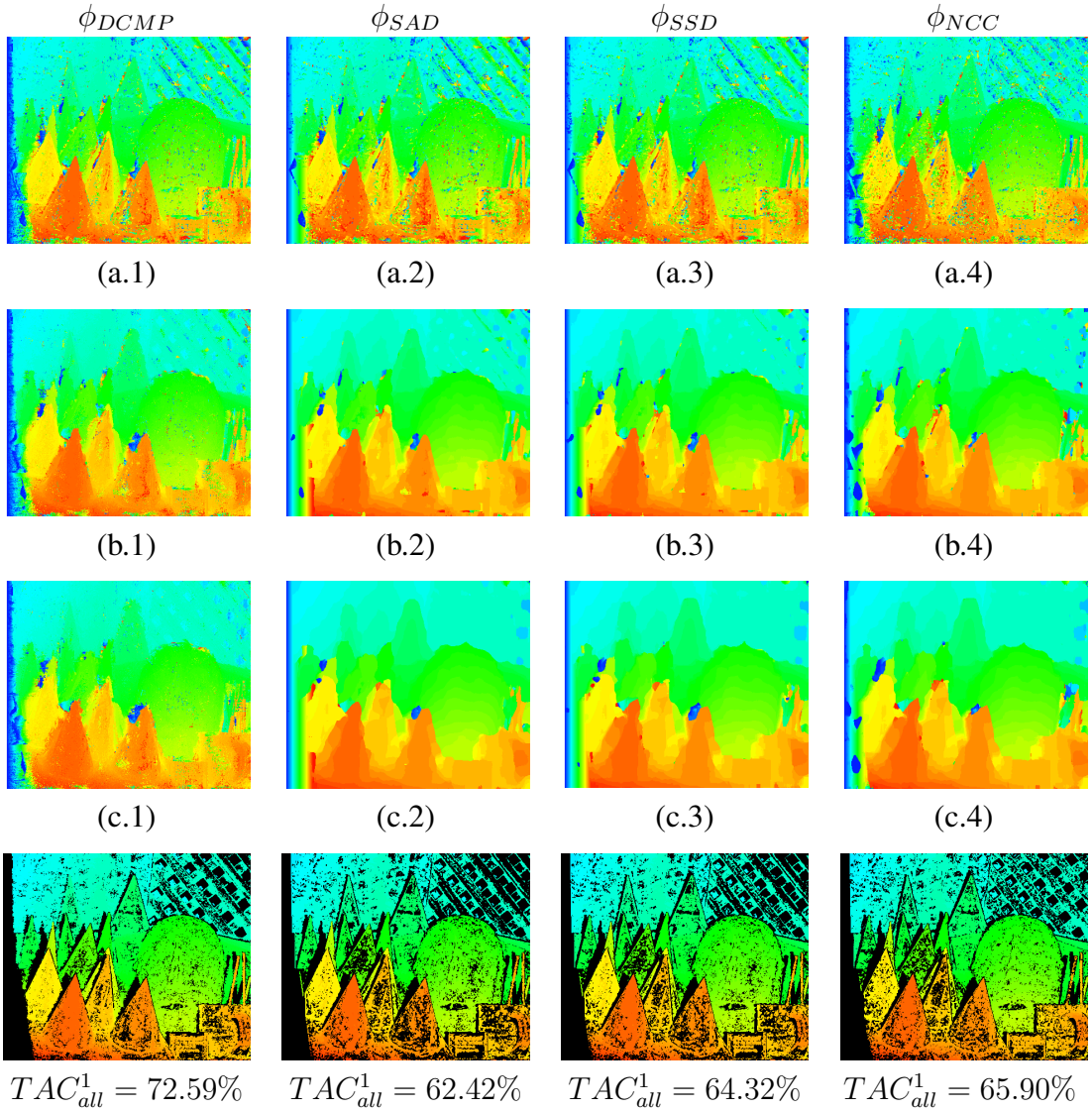


FIGURE 2.28 – Cartes de disparité denses obtenues avec les fonctions de vraisemblance ϕ_{DCMP}, ϕ_{SAD}, ϕ_{SSD} et ϕ_{NCC} pour des fenêtres de taille (a.) 3×3, (b.) 11×11 et (c.) 19×19. Les images de la dernière ligne montrent les erreurs d'appariement (couleur noire). Les taux d'appariements corrects TAC^1_{all} sont calculés pour une fenêtre de taille 3×3 et un seuil de 1.

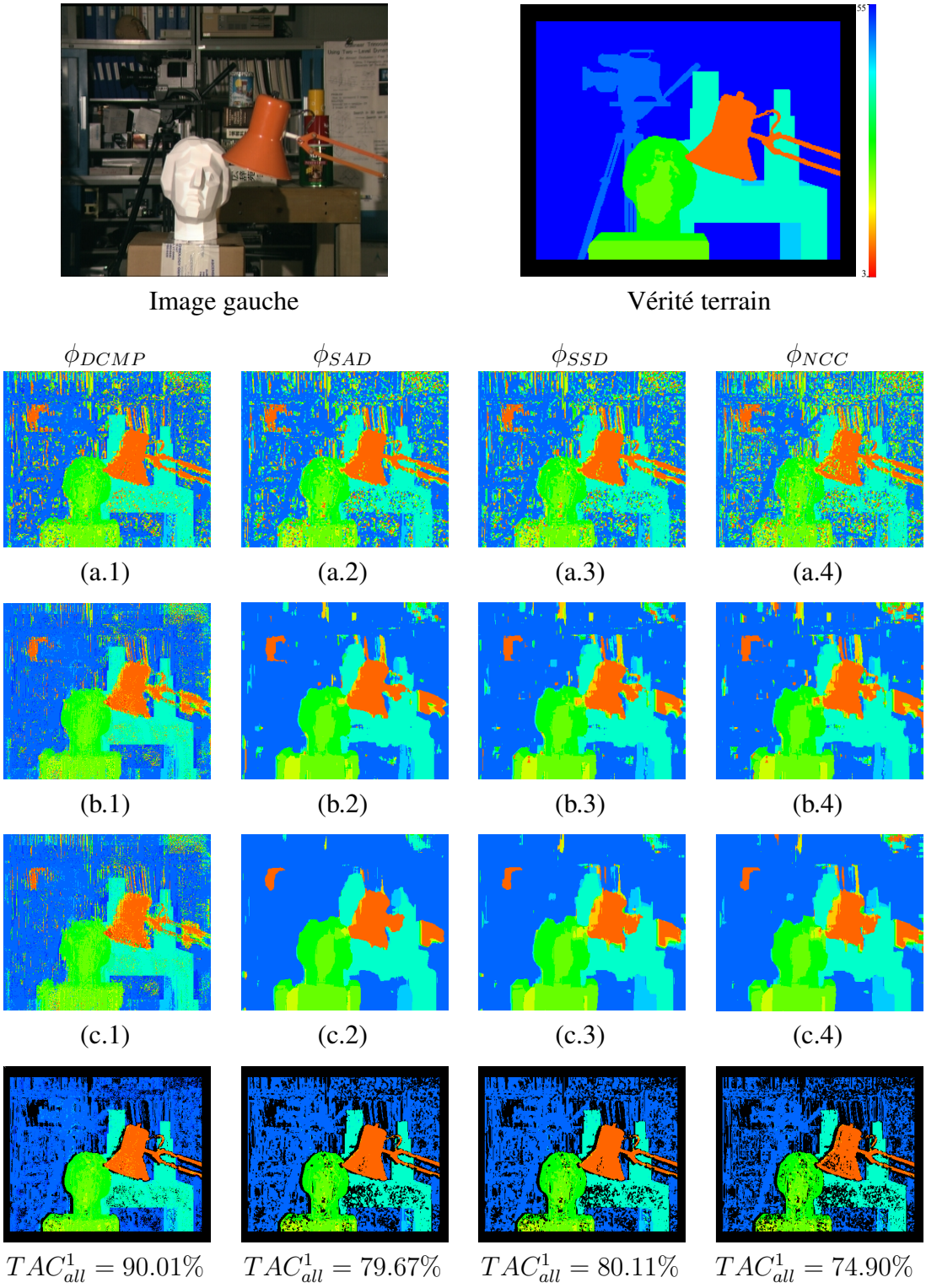


FIGURE 2.29 – Cartes de disparités denses obtenues avec les fonctions de vraisemblance ϕ_{DCMP}, ϕ_{SAD}, ϕ_{SSD} et ϕ_{NCC} pour des fenêtres de taille (a.) 3×3, (b.) 11×11 et (c.) 19×19. Les images de la dernière ligne montrent les erreurs d'appariement (couleur noire). Les taux d'appariements corrects TAC^1_{all} sont calculés pour une fenêtre de taille 3×3 et un seuil de 1.

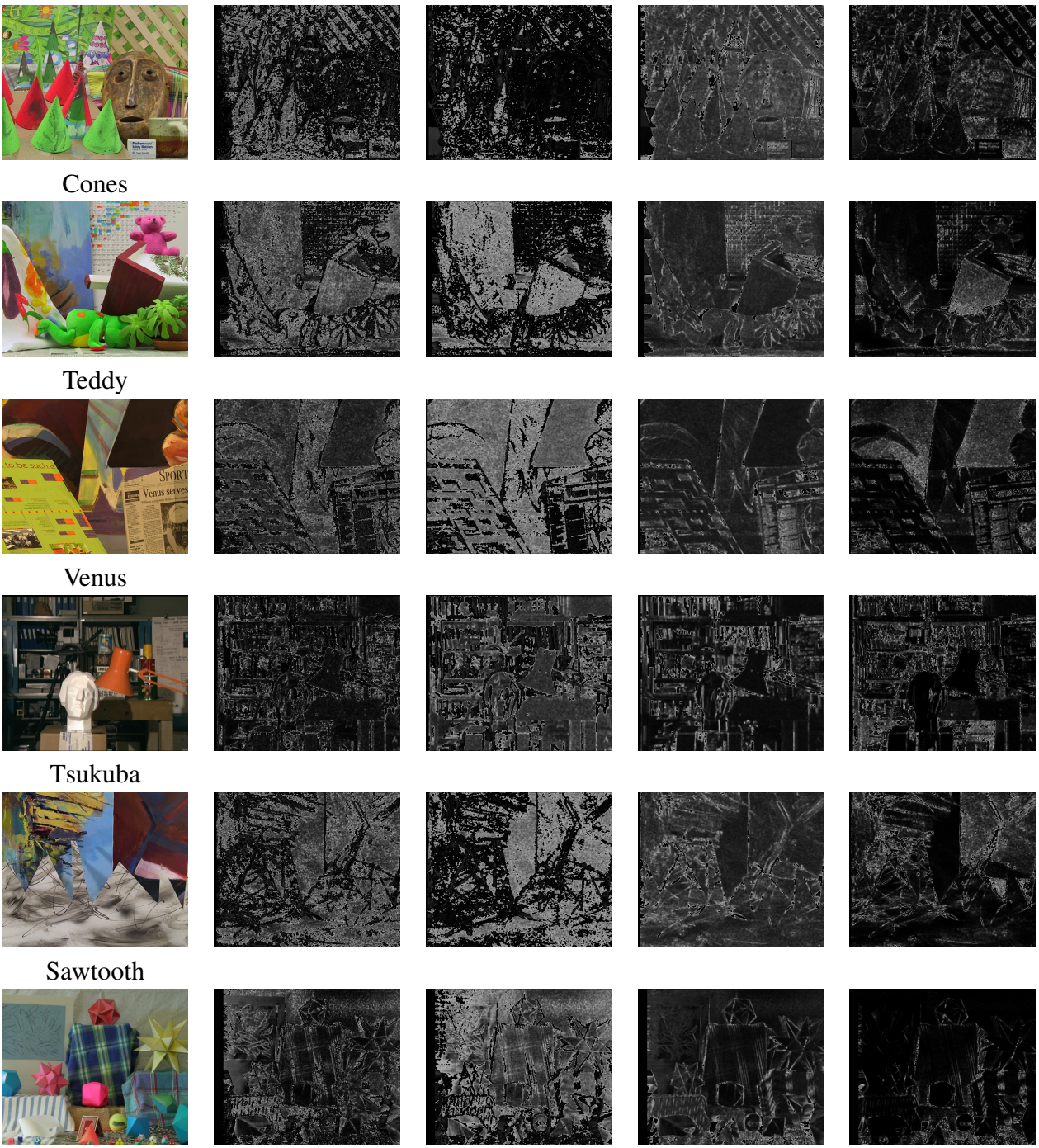


FIGURE 2.30 – Cartes des coûts obtenues avec les fonctions de vraisemblance ϕ_{DCMP}, ϕ_{SAD}, ϕ_{SSD} et ϕ_{NCC} pour une fenêtre de taille 3×3. La couleur noire désigne le score minimal. Plus la couleur d'un pixel est sombre, plus son score est faible, et vice versa.

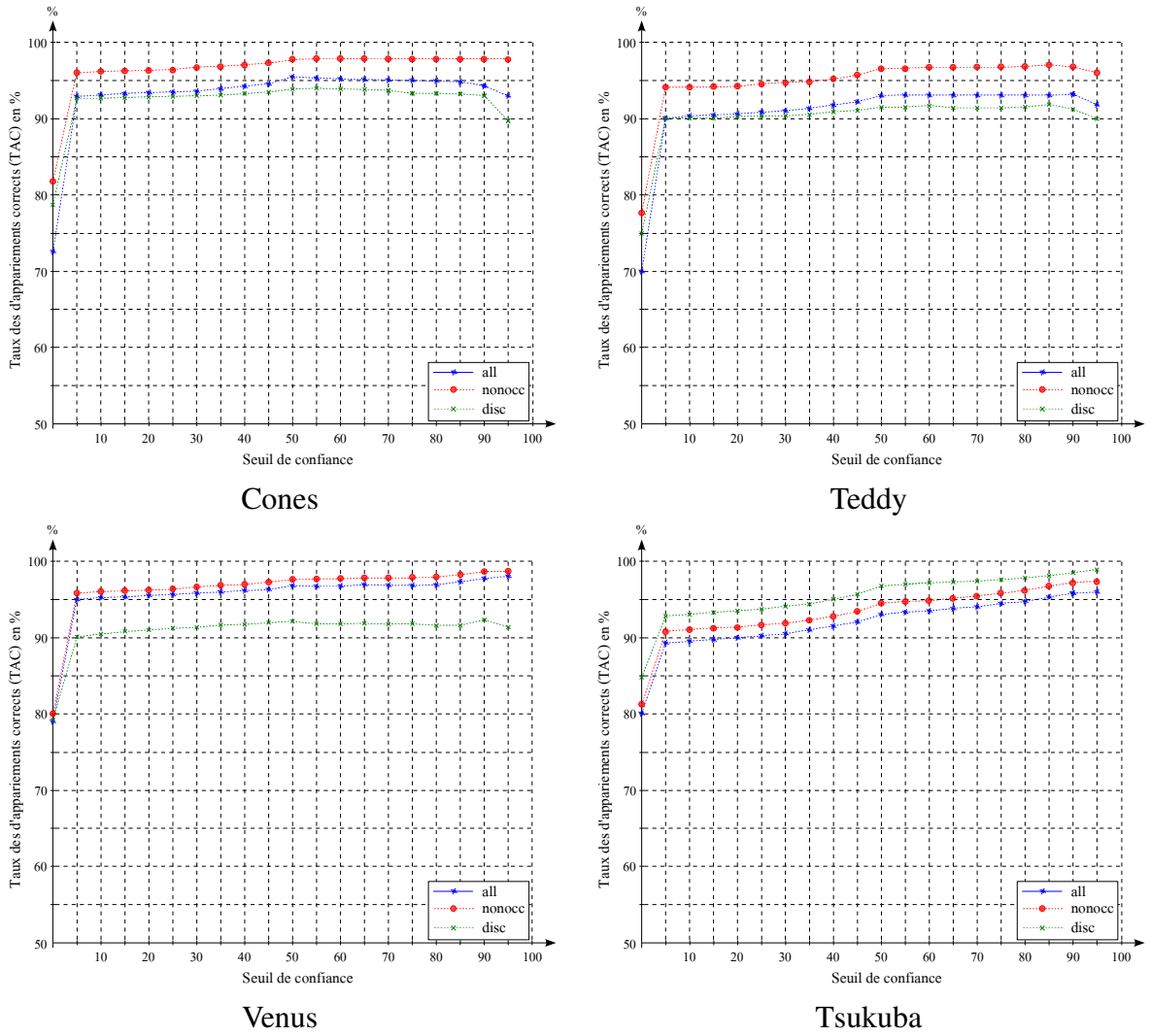
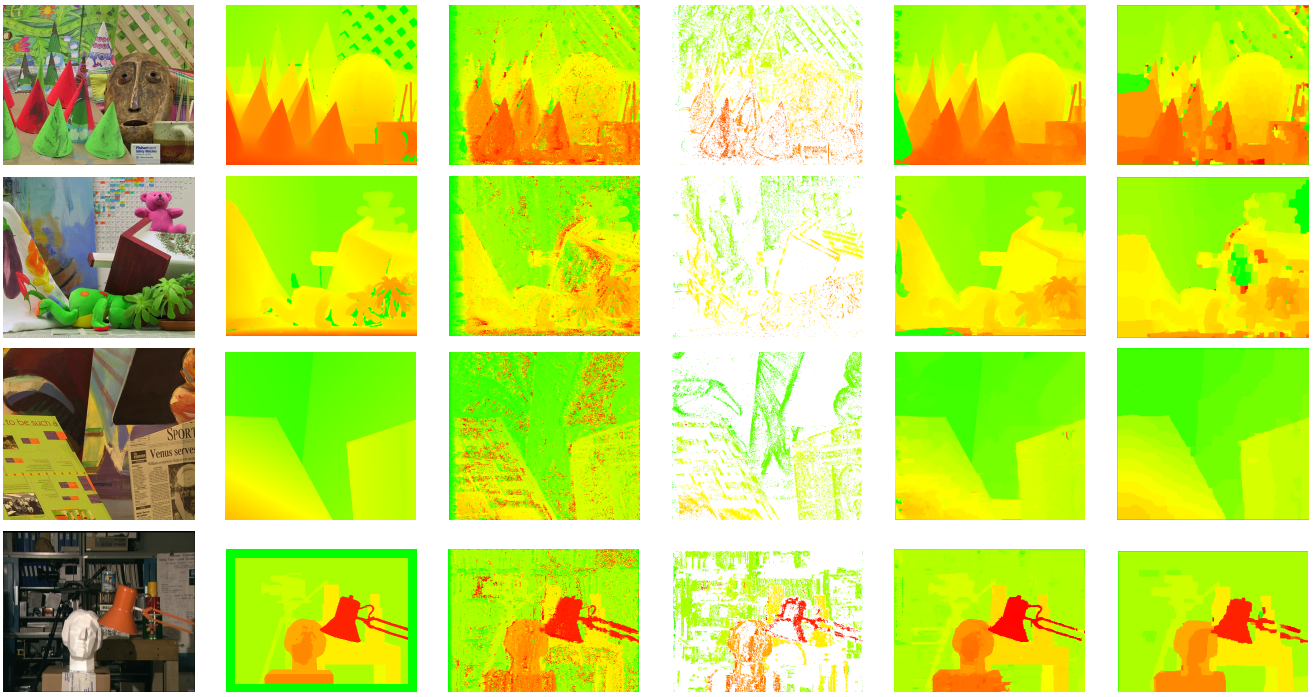
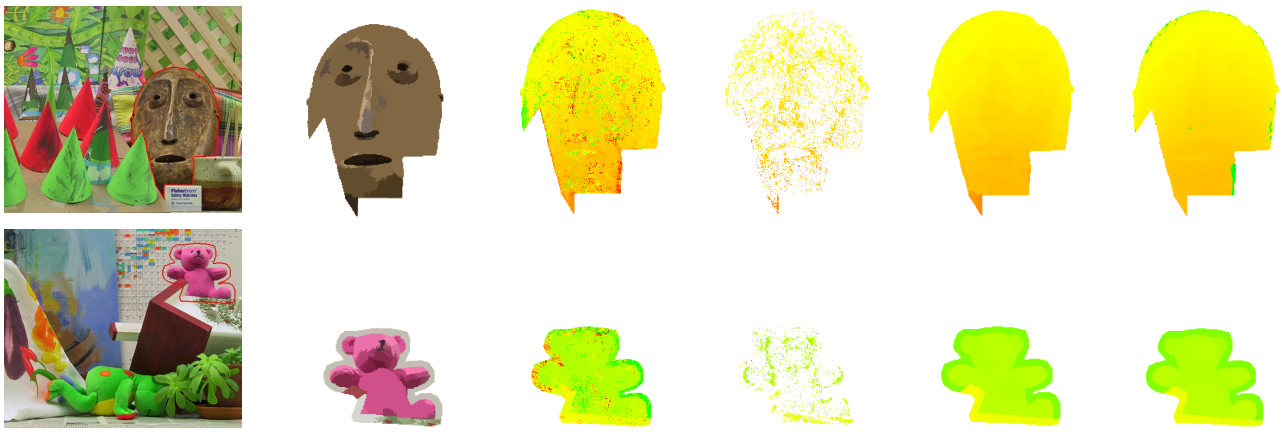


FIGURE 2.31 – Courbes des TAC^1 obtenues sur les ensembles $\mathcal{P}_{all}$, $\mathcal{P}_{nonocc}$, et $\mathcal{P}_{disc}$ pour les images "Cones", "Teddy", "Venus", et "Tsukuba".



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

FIGURE 2.32 – Cartes de disparités denses obtenues avec notre algorithme d'appariement et l'algorithme max-product [FH06] (a) les images de références. De haut en bas : "Cones", "Teddy", "Venus", et "Tsukuba" (b) vérités terrain. (c) cartes denses de disparité obtenues avec la fonction de vraisemblance ϕ_{DCMP} (d) cartes éparses de disparité obtenues avec un seuil de confiance de 60% : seuls les appariements ayant mesure de confiance supérieure à ce seuil sont retenus (e) cartes denses de disparité obtenues aprés réestimation et rectification des disparités par notre algorithme de propagation de croyance sélective (f) cartes denses de disparité obtenues avec [FH06].



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

FIGURE 2.33 – Les différentes étapes de l'algoritme d'appariement proposé, appliquées sur deux régions manuellement segmentées des images "Cones" et "Teddy". (a) les images gauches "Cones" (haut) et "Teddy" (bas) (b) deux régions extraites manuellement des images segmentées par *MeanShift* (c) carte de disparités dense obtenues pour les deux régions par la fonction de vraisemblance ϕ_{DCMP} (d) carte de disparités éparse après avoir appliqué un seuil de confiance de 60% (e) réestimation et rectification des disparités par notre algorithme Propagation de Croyance Sélective (*PCS*) (f) vérité terrain.

Chapitre 3

Estimation du Mouvement par Analyse en Composantes Indépendantes

3.1 Introduction

L'estimation du relief par vision stéréoscopique fait l'objet de nombreux travaux de recherches. Leur principal objectif est d'améliorer la qualité d'appariement tout en réduisant le temps de traitement, qui est essentiel dans certaines applications. L'amélioration de la qualité d'appariement consiste à réduire les erreurs de calcul de disparités des régions partiellement occultées, des régions de couleur uniforme ou de textures répétitives. Ces aspects ont fait l'objet du chapitre précédent dans lequel nous avons proposé un nouvel algorithme d'appariement permettant à la fois une amélioration de la qualité de la mise en correspondance, et une réduction du temps de traitement. Nous avons vu que l'introduction de certains critères (unicité, symétrie, ordre) permettaient la réduction de l'espace de recherche et l'élimination de faux appariements.

Une nouvelle possibilité d'améliorer la mise en correspondance peut être utilisée si l'on dispose de séquences d'images et non d'images fixes uniqument. Il s'agit d'une contrainte de mouvement. Dans le cadre des applications de vidéosurveillance, une ou plusieurs caméras, fixes ou embarquées, supervisent un environnement composé de régions affectées par du mouvement, dites premier-plan, ou non affectées par du mouvement, dites arrière-plan. Dans le cas d'une caméra fixe, les disparités des pixels appartenant au fond ne varient pas significativement au fil du temps. Par contre, les disparités des pixels affectés par le mouvement varient continuellement. Seules les disparités des pixels affectés par du mouvement peuvent alors être estimées. Le principe est que le correspondant d'un pixel classé en mouvement de l'image gauche, est aussi en mouvement dans l'image droite. Ceci permet de réduire l'ensemble des candidats possibles pour chaque pixel en mouvement à apparier.

Dans ce qui suit, nous nous intéressons aux méthodes permettant l'estimation du mouvement et la classification des pixels en deux catégories : premier-plan (appelé aussi région, forme ou objet), et arrière-plan (appelé aussi fond de scène). La première partie présente un état de l'art sur les méthodes existantes d'extraction de fond à partir d'une séquence d'images issue d'une caméra fixe. Nous détaillons ensuite une nouvelle méthode d'estimation du mouvement dans des environnements extérieurs difficiles.

3.2 Méthodes basées sur une image de référence

L'extraction d'objets en mouvement est une étape critique pour grand nombre d'applications en vision par ordinateur. Même si ce problème a été et est toujours largement étudié, l'extraction d'objets se déplaçant dans des environnements complexes et non contraints est encore loin d'être complètement résolue. La diversité des recherches menées dans ce domaine est liée à la difficulté de la tâche : les méthodes proposées doivent être robustes aux fluctuations des intensités observées et au contenu dynamique de la scène étudiée, tout en étant suffisamment sensibles pour permettre la détection d'objets sur des images peu contrastées. Ce paragraphe a pour but de fournir un bref état de l'art sur les méthodes de détection d'objets en mouvement basées sur une image de référence. Une vaste littérature existe dans ce domaine [EESA08]. Les méthodes existantes peuvent être divisées en deux grandes catégories :

– Les méthodes basées sur une modélisation non-paramétrique du fond.

– Les méthodes basées sur une modélisation paramétrique du fond.

3.2.1 Les méthodes non-paramétriques

Cette famille de méthodes d'extraction des régions affectées par du mouvement dans une séquence d'images, est connue par sa simplicité d'implémentation. La procédure de classification se divise généralement en deux parties : une période d'apprentissage et une période de détection. Pour que les méthodes non-paramétriques soient efficaces, la période d'apprentissage doit être suffisamment longue. Durant cette période, l'établissement d'un modèle du fond consiste à sauvegarder les états possibles d'un pixel (intensités, couleurs). El-Gamal et al. [EHD00] proposent de modéliser chaque pixel du fond par un vecteur contenant les intensités observées durant la période d'apprentissage. La classification fond/objet d'un pixel donné, durant la période de détection, est basée sur la probabilité qu'un pixel issu du modèle du fond ait une intensité égale à l'intensité du pixel à classer. La probabilité correspond à une estimation non-paramétrique basée sur un noyau Gaussien. Un pixel est classé comme appartenant au premier-plan si la probabilité correspondante estimée ne dépasse pas un certain seuil. La mise à jour du modèle du fond se base tout simple-

ment sur le principe "*first-in first-out*" qui consiste à remplacer l'observation la plus ancienne par l'observation courante. Cette technique, comparée avec la méthode du mélange de Gaussiennes, montre qu'elle est moins sensible aux changements brusques d'illumination.

Un modèle adaptatif, appelé modèle à valeurs médianes, a été proposé par S. Greenhill [GVW04] pour l'extraction d'objets en mouvement dans des conditions d'illumination dégradées. En se référant aux différents états de chaque pixel durant une période d'apprentissage, un modèle du fond est ainsi élaboré. Chaque pixel du modèle correspond à la valeur (intensité, couleur, ...) minimisant la distance avec chacune des valeurs possibles. Le modèle du fond est constamment mis à jour pour chaque nouvelle image de telle sorte qu'un vecteur V de valeurs médianes est construit sur les $N/2$ dernières images, sachant que N correspond au nombre d'images utilisées durant l'étape d'apprentissage. Une distance notée D correspond à la différence entre la valeur minimale et la valeur maximale du vecteur V. La classification fond/objet se fait simplement par seuillage sur la distance entre la valeur du pixel à classer et son correspondant dans le modèle du fond. Afin de tenir compte des changements d'illumination, le seuil tient compte de la distance D et un facteur de corrélation. Cette technique a prouvé sa capacité dans la mesure où elle permet d'ignorer les "faux" objets, qui correspondent souvent à de simples changements dûs aux variations rapides d'illumination.

Une des méthodes non-paramétriques les plus répandues est celle intitulée "*Codebook*", proposée par K. Kim [KCHD05]. L'auteur propose de modéliser le fond en se basant sur une séquence d'observations de chaque pixel durant une longue période (plusieurs minutes). Cependant, les occurrences similaires d'un pixel donné sont réprésentées sous forme d'un vecteur appelé "*Codeword*". Deux "*Codewords*" sont différents si la distance, dans l'espace vectoriel, dépasse un certain seuil. Un "*Codebook*", qui est un ensemble de "*Codewords*", est construit pour chaque pixel. Notons que l'idée de sauvegarder les états possibles d'un même pixel est très similaire aux méthodes paramétriques basées sur un mélange de Gaussiennes. La classification fond/objet d'un pixel donné se base sur une simple différence entre sa valeur courante (intensité, couleur, etc.) et chacun des "*Codewords*" correspondants.

Dans [SHT+06], l'auteur propose une méthode relativement simple pour l'extraction de silhouettes humaines dans un contexte de vidéo-surveillance. Chaque pixel dans le modèle du fond est représenté par un vecteur de dimension 3, dont les éléments sont les valeurs minimales et maximales des intensités, et la différence maximale d'intensité entre deux images successives durant l'étape d'apprentissage qui, selon l'auteur, est constituée d'une cinquantaine d'images en moyenne. Cette méthode est évaluée sur une base d'images assez simple, dans laquelle les conditions dégradées d'illumination sont ignorées. Contrairement à ce qui était présenté, les auteurs dans [LH07] remplacent la modélisation pixellaire du fond par une méthode basée sur un principe de segmen-

tation. Les différentes régions de l'image segmentée sont prises comme références dans l'étape de détection. Dans ce cas, aucune période d'apprentissage n'est nécessaire, ce qui conduit à un gain de mémoire non négligeable.

3.2.2 Les méthodes paramétriques

3.2.2.1 Modélisation du fond au niveau des pixels

La plupart des approches d'extraction d'objets en mouvement se fondent sur l'évolution temporelle de chaque pixel de l'image. Une séquence d'observations est utilisée pour construire un modèle de fond en chaque pixel. Elle peut inclure l'intensité lumineuse, la couleur, ou d'autres caractéristiques de texture. Le processus de détection consiste alors à classifier chaque pixel de manière indépendante dans les classes fond et objet, en fonction des observations courantes. De nombreuses approches ont été proposées pour construire le modèle du fond de scène, nous en détaillons ici les principales.

Estimation du modèle de fond par filtrage : le modèle du fond peut être estimé par des filtres de type moyens ou médians [MS95], [LV01], [CGPP03], un filtre max-min [HHD00]. Il est généralement défini lors d'une phase d'initialisation puis constamment mis à jour afin de s'adapter aux changements éventuels de la scène observée. Le principal inconvénient de cette approche est que le modèle s'adapte très lentement aux changements soudains du fond. Cependant, l'utilisation d'un même seuil de détection en chaque pixel rend ainsi la méthode très peu adaptative.

Modélisation du fond par une distribution Gaussienne : une manière d'adapter le seuil en chaque pixel consiste à modéliser la distribution des intensités lumineuses de chaque pixel par une loi Gaussienne [WADP97]. Ce modèle peut s'adapter à des changements lents de la scène, tels que des changements d'illumination progressifs, en se mettant à jour de manière récursive grâce à un filtre adaptatif. Différentes extensions de ce modèle ont été développées en changeant les caractéristiques utilisées au niveau pixel. G. Gordon et al. [GDHW99] représentent chaque pixel par un vecteur composé de quatre composantes indépendantes qui sont les trois composantes couleur et la profondeur. Les auteurs dans [JDWR00] représentent le fond par deux distributions Gaussiennes distinctes : une utilisant la couleur et la seconde exploitant les contours.

Modélisation du fond par un Mélange de Gaussiennes : la modélisation par une simple distribution Gaussienne présente de bonnes performances pour des scènes intérieures en milieu contrôlé. Dans le cas où les scènes sont bruitées et évolutives, la densité de probabilité des niveaux de gris en chaque pixel devient multi-modale. Une amélioration consiste alors à modéliser l'évolution temporelle des pixels par un mélange de Gaussiennes [FR97]. Ainsi, C. Stauffer et W.E.L. Grimson [SG99] [SG00] modélisent la couleur de chaque pixel par un mélange de i Gaus-

siennes. Le nombre de Gaussiennes i doit être ajusté en fonction de la complexité de la scène observée. Afin de simplifier les calculs, la matrice de covariance est supposée diagonale, ce qui revient à considérer les trois canaux couleur de manière indépendante. Le modèle de mélange de Gaussiennes est mis à jour à chaque itération en utilisant l'algorithme des "*k-means*" [MNV09]. En effet, l'utilisation initiale de l'algorithme "*Expectation Maximisation EM*" [DLR77] en chaque pixel s'est montré trop coûteuse. M. Harville et al. [HGW01] proposent d'utiliser le modèle de mélange de Gaussiennes dans un espace combinant la profondeur et l'espace couleur YUV. Ils améliorent la méthode en modulant le taux d'apprentissage du modèle en fonction de l'activité de la scène. Bien que cette approche soit largement utilisée dans la littérature, elle présente un certain nombre d'inconvénients : elle est très sensible aux variations brusques du fond, tel que le changement global d'illumination. Ainsi, un compromis doit être trouvé entre le taux d'apprentissage et l'adaptation aux changements du fond. Un faible taux d'apprentissage produira énormément de fausses détections lors d'un changement d'éclairage, tandis qu'un fort taux d'apprentissage aura tendance à inclure les objets en mouvement dans le modèle du fond.

Modélisation du fond grâce à des techniques de quantification/agglomération : D. Butler et al. [BSJ03][BS05] ont proposé un algorithme de segmentation d'objets en mouvement basé sur la technique d'agglomération, appelée "*clustering*" par la communauté anglophone. Ils modélisent chaque pixel par un groupe de "*clusters*", où chaque *cluster* se compose d'une valeur moyenne de couleur et d'un poids. Les clusters sont ordonnés en fonction de leur poids et s'adaptent pour gérer les variations de fond et d'éclairage. Le principe de cet algorithme est semblable à celui de Stauffer et Grimson [SG99], mais il a la capacité de traiter des vidéos de taille 320×240 en temps réel sur des hardwares modestes.

Modélisation du fond par des chaînes de Markov : tous les modèles précédemment cités ne considèrent pas réellement l'évolution temporelle des pixels lorsque l'on modélise les k dernières valeurs d'intensité lumineuse d'un pixel par une Gaussienne, l'ordre d'arrivée des niveaux de gris en ce pixel n'est pas pris en compte, alors que cette information peut-être utile. Une solution consiste alors à modéliser l'évolution des niveaux de gris en chaque pixel par une chaîne de Markov. J. Rittscher et al. [RKJB00] emploient une chaîne de Markov à 3 états : objet, fond et ombre. Tous les paramètres de la chaîne, qui sont la probabilité initiale, la probabilité de transition et la probabilité d'observation, sont estimés hors ligne, sur une séquence d'apprentissage. B. Stenger et al. [SRPB01] ont proposé une amélioration, puisqu'après un court apprentissage préalable, le modèle de la chaîne et ses paramètres continuent à être mis à jour. Cette mise à jour, réalisée durant la période de détection, pemet de mieux gérer les états non-stationnaires dûs par exemple à de brusques changements d'illumination.

3.2.2.2 Modélisation du fond par des régions

A l'inverse de l'approche précédente, la modélisation du fond par des régions qui a l'avantage de prendre en compte les relations spatiales entre pixels, a gagné beaucoup d'intérêt ces dernières années. K. Toyama et al. [TKBM99] proposent un système à trois échelles : pixel, région et image. Au niveau pixel, un modèle de fond est maintenu pour chaque pixel, ce qui permet de prendre une première décision de classification. Le niveau région considère les relations spatiales entre pixels, ce qui améliore la segmentation. Enfin, le niveau image permet de gérer les changements globaux d'illumination.

Une autre méthode globale proposée par N.M. Oliver et al. [ORP00] consiste à construire un espace propre qui modélise le fond et ses changements d'apparence. Cet espace propre est formé à partir de la moyenne et de la matrice de covariance des images d'une séquence. Le fond est modélisé par les vecteurs propres correspondant aux plus grandes valeurs propres de la matrice de covariance. La détection est alors réalisée simplement de la manière suivante : les zones statiques de l'image sont décrites avec précision par une somme pondérée des vecteurs propres, ce qui n'est pas le cas des zones affectées par du mouvement. J. Zhong et al. [ZS03] présentent un algorithme qui modélise le fond texturé et non-stationnaire par un modèle de Box-Jenkins (modèle autorégressif et moyenne mobile). Un filtre de *Kalman* est ensuite employé pour estimer itérativement les paramètres dynamiques de la texture. Les régions en mouvement sont alors obtenues par seuillage de la fonction de pondération utilisée dans le filtre de *Kalman*.

Dans [SS05], Sheikh et Shah modélisent le fond dans un espace conjoint couleur-position grâce à une estimation non-paramétrique de la densité de probabilité utilisant des fonctions noyau. L'auteur propose également de modéliser explicitement l'objet afin d'exploiter la persistance temporelle et d'améliorer la détection. Enfin, et afin d'introduire le contexte spatial, un algorithme de segmentation par champs de Markov est proposé. Il permet de trouver une solution globale optimale de la classification des pixels en fond/objet. M. Heikkila et al. [HP06] [HPS09] proposent de diviser l'image en blocs de taille égale, se chevauchant partiellement, en utilisant une structure de grille et de modéliser des caractéristiques statistiques de chaque bloc par les histogrammes de motif binaire local. S.-D. Chen et al. [CR03] étendent cette idée en combinant l'approche à base de blocs et une approche à base de pixels. Les auteurs utilisent un descripteur discriminant appelé histogramme contraste pour caractériser chaque bloc, puis construisent un cadre de détection en combinant ce descripteur avec un mélange de Gaussiennes au niveau des pixels. Ce cadre de détection permet de combiner efficacement les informations détectées par la modélisation du fond par des blocs et celle au niveau des pixels.

3.3 Analyse en Composante Indépendante (ACI)

Plusieurs techniques existent pour l'extraction des régions affectées par du mouvement dans une séquence d'images. Bien que ces méthodes présentent de bons résultats, plusieurs inconvénients subsistent. Ces différentes méthodes prouvent leur efficacité sur des scènes intérieures. Les scènes réelles prises à l'extérieur sont plus difficiles à traiter à cause des contraintes comme le changement continu de luminosité et la présence d'ombres irrégulières. Une nouvelle méthode basée sur le principe d'Analyse en Composante Indépendante a été récemment proposée pour la soustraction du fond dans une séquence d'images [HJ86], [ZC06], [TL09]. Cette technique a été initialement introduite dans le domaine du traitement du signal dans le cadre de la séparation de sources. Compte tenu de la complexité algorithmique, de la qualité des résultats, et du temps de traitement, nous allons développer dans ce qui suit une nouvelle méthode permettant une amélioration en termes de temps de traitement et de qualité de détection, par rapport aux méthodes existantes.

3.3.1 Historique

L'*Analyse en Composantes Indépendantes (ACI)* est une méthode introduite dans les années 1980 par J. Hérault [HA84] pour résoudre des problèmes neuro-physiologiques. L'ACI était largement répandue dans la communauté française de chercheurs avec une ouverture limitée vers l'international, qui dans la même période, traitaient des problèmes de rétro-propagation, des réseaux de Hopfield, et des cartes d'auto-organisation de Kohonen. L'ACI se base sur les méthodes traitant des observations vectorielles afin d'en extraire des composantes linéaires qui soient aussi indépendantes que possible. En 1989, des méthodes traitant de l'analyse spectrale d'ordre élevé ont été proposées. J.-F. Cardoso a proposé dans [Car89] une méthode algébrique traitant les tenseurs cumulants d'ordre élevé, qui avait conduit au fameux algorithme *JADE*. L'ACI a pris davantage d'intérêt après la publication de A.J. Bell et T.J. Sejnowski [BS95] qui proposaient une approche basée sur le principe d'infomax [BS95]. Cette idée a été raffinée par S.-I Amari [ACY96], en introduisant la méthode du Gradient en connexion fondamentale avec l'estimation du maximum de vraisemblance. Quelques années plus tard, un nouvel algorithme intitulé *fixed-point* ou *FastICA* a été proposé dans [HO97] et [Hyv99]. Cet algorithme a permis l'utilisation de l'ACI dans plusieurs applications grâce à sa rapidité.

Nous citons ci-après quelques applications de l'ACI comme en télécommunication pour la restitution de signaux et élimination du bruit. L'ACI est appliquée aussi sur des données Électro-encéphalographiques (EEG) qui consiste à séparer des signaux issus d'un ensemble d'électrodes mesurant les activités cérébrales, sur des groupes de signaux temporairement indépendants [MJaB$^+$97]. L'ACI est aussi largement utilisée en vision artificielle. Dans ce cas, une image est interprétée comme étant un signal discret ayant certaines caractéristiques. Une des premières applications

était la restitution et la compréhension des activitées cérébrales par analyse d'imagerie par résonance magnétique fonctionnelle (*fMRI*) [Sto99] [CAPP01]. Un signal *fMRI* est donné pour chaque voxel. Le signal correspond aux variations hémodynamiques associées avec un bruit de mesure. L'ACI spatio-temporelle est appliquée sur chaque signal obtenu au fil du temps pour l'extraction des signaux sources indépendants. L'ACI est aussi utilisée pour la caractérisation d'images, pour la compression d'images, la réduction de bruits dans des images réelles, la reconnaissance faciale [HHC+02]. L'ACI spatiale a été utilisée par [THM98] sur des images couleur représentant la peau humaine pour la caractérisation de la mélanine et de l'hémoglobine.

3.3.2 Principe

L'ACI est une méthode statistique permettant la séparation d'un signal source complexe, qui est un mélange de signaux inconnus a priori, en une combinaison de plusieurs signaux, dits signaux estimés. L'ACI est définie par un modèle génératif permettant l'estimation d'un ensemble de signaux à partir d'un signal donné. Le signal observé est supposé être une combinaison linéaire des signaux à estimer. Nous avons choisi la linéarisation afin de réduire la complexité de réprésentation et garantir une meilleure compréhension du modèle de l'ACI. Nous définissons un signal d'entrée par un ensemble de variables observées simultanément. Le nombre de variables est noté par $\mathcal{V}$ et le nombre d'observations est noté par $\mathcal{O}$. Le signal d'entrée peut être représenté sous forme d'une matrice X dont chaque élément est donné par $x_{i,j}$ tels que $i = 1, \ldots, \mathcal{V}$ et $j = 1, \ldots, \mathcal{O}$. L'ACI consiste à estimer les signaux sources notés $\tilde{S}$. Pour un signal X donné, le modèle mathématique de l'ACI s'écrit de la manière suivante :

$$X = \tilde{A}.\tilde{S} \tag{3.1}$$

Dans le modèle de l'équation 3.1, nous constatons qu'il y a deux inconnues à estimer. Il s'agit de la combinaison linéaire des signaux sources séparés $\tilde{S}$ et une matrice de coéfficients $\tilde{A}$, formant le signal d'entrée. La séparation des sources revient à estimer les paramètres d'une matrice W, qui est l'inverse de la matrice $\tilde{A}$. Le problème revient alors à estimer les paramètres de la matrice de passage, dite "*de-mixing matrix*", notée W permettant d'obtenir les sources séparées donnés par la matrice $\tilde{S}$. Le modèle d'ACI est reformulé ainsi :

$$\begin{pmatrix} s_{1,1} & \ldots & s_{1,j} \\ \vdots & \ddots & \vdots \\ s_{i,1} & \ldots & s_{i,j} \end{pmatrix} = \begin{pmatrix} w_{1,1} & \ldots & w_{1,i} \\ \vdots & \ddots & \vdots \\ w_{i,1} & \ldots & w_{i,i} \end{pmatrix} \times \begin{pmatrix} x_{1,1} & \ldots & x_{1,j} \\ \vdots & \ddots & \vdots \\ x_{i,1} & \ldots & x_{i,j} \end{pmatrix}$$

$$\tilde{S} = \tilde{A}^{-1}.X = \tilde{W}.X \tag{3.2}$$

D'une manière classique, le modèle d'ACI peut être résolu en définissant une fonction d'énergie et un algorithme optimisant cette fonction comme par exemple le gradient stochastique ou la méthode de Newton. Ces méthodes classiques d'optimisation échouent parfois à cause de la complexité du signal à séparer. Le choix d'une telle formulation dépend des propriétés à satisfaire. Celles-ci dépendent fortement des données à traiter. Des critères dans le choix de la fonction à optimiser existent tels que la cohérence, la variance asymptotique, et la robustesse. Les critères pouvant être pris en compte lors de la conception d'un algorithme d'optimisation peuvent être la vitesse de convergence, les exigences en mémoire, et la stabilité numérique.

3.3.2.1 Le principe d'indépendance et de non-corrélation

Le principe d'indépendance statistique entre les composantes à estimer est une des réflexions pour la résolution d'un système par ACI. Les statistiques d'ordre élevé sont utilisées pour trouver les composantes indépendantes d'un signal en maximisant l'indépendance statistique entre les composantes estimées. Cependant, deux variables aléatoires s_1 et s_2 sont mutuellement indépendantes si la réalisation de l'une ne dépend pas de la réalisation de l'autre. Autrement dit, la variable s_1 n'apporte pas de l'information sur l'autre, et vice versa. Il existe plusieurs techniques pour la séparation des sources. L'indépendance peut être définie par les densités de probabilité. Notons $p(s_1, s_2)$ la densité de probabilité jointe des deux variable aléatoires s_1 et s_2. Notons ainsi $p_1(s_1)$ et $p_2(s_2)$ les probabilités marginales des deux variables s_1 et s_2 respectivement. La probabilité marginale est représentée en fonction de la densité de probabilité jointe de la façon suivante :

$$p_1(s_1) = \int p(s_1, s_2) ds_2 \tag{3.3}$$

Deux variables aléatoires s_1 et s_2 sont définies comme indépendantes si et seulement si la densité de probabilité jointe est le produit des densités de probabilités marginales de ces deux variables, comme donné par l'équation 3.4. La non-corrélation signifie que la covariance entre les variables aléatoires est égale à zéro. Pour établir la relation entre indépendance et corrélation, deux variables indépendantes sont ainsi non corrélées.

$$p(s_1, s_2) = p_1(s_1).p_2(s_2) \tag{3.4}$$

3.3.2.2 La non-Gaussienneté

Une restriction fondamentale de l'ACI est que la distribution des composantes indépendantes ne doit pas être Gaussienne. En effet, la matrice de mélange W est non identifiable pour des composantes indépendantes Gaussiennes. Pour expliquer cette contrainte, on suppose que la matrice de mélange est orthogonale et que les sources sont gaussiennes ; par conséquent, les composantes de la matrice des observations sont gaussiennes, non-corrélées et de covariance égale à la matrice Identité. La densité de probabilité jointe de deux variables gaussiennes est définie par :

$$p(s_1, s_2) = \frac{1}{2\pi}\left(-\frac{s_1^2 + s_2^2}{2}\right) \tag{3.5}$$

La non-Gaussienneté d'une variable aléatoire peut être mesurée par le coefficient d'aplatissement "*Kurtosis*" ou par néguentropie "*negentropy*".

– ***Le coefficient d'aplatissement ou "Kurtosis"***

Il s'agit d'une version normalisée du moment d'ordre quatre $E\{s_1^4\}$. Ce coéfficient permet de mesurer la non-Gaussianité d'une variable s_1, comme donné par l'équation 3.6.

$$kurt(s_1) = E\{s_1^4\} - 3\left(E\{s_1^2\}\right)^2 \tag{3.6}$$

L'égalité de l'équation 3.6 peut être simplifiée à $E\{s_1^4\} - 3$ si la variance de la variable s_1 est supposée égale à une unité. Dans le cas où la distribution de la variable s_1 est Gaussienne, le moment d'ordre 4 sera égal à $3\left(E\{s_1^2\}\right)^2$. Donc le coefficient d'aplatissement est égal à zéro pour une variable Gaussienne. La non-Gaussienneté est mesurée par la valeur absolue ou le carré du coéfficient d'aplatissement. Cette mesure est égale à zéro pour une variable aléatoire dont la distribution est Gaussienne, et supérieure à zéro autrement. En se basant sur le coefficient d'aplatissement, il existe plusieurs algorithmes d'optimisation tels que ceux basés sur le gradient ou les méthode de point fixe "*fixed-point*". L'optimisation consiste à trouver les bons coefficients de la matrice W. Le *Kurtosis* est minimisé dans les directions des composantes indépendantes qui correspondent aux directions dans lesquelles la valeur

absolue du coefficient d'aplatissement est maximale. Une optimisation adaptative dite aussi en-ligne peut être donnée par $\Delta_\gamma \propto \left(\left(W^T x\right)^4 - 3\right) - \gamma$, où x est une observation, et γ est une estimation du *Kurtosis* obtenue par une moyenne temporelle sur l'observation. Le principal inconvénient de cette méthode est qu'elle est très sensible aux valeurs aberrantes.

– ***La néguentropie ou "negentropy"***

Il s'agit d'une quantité issue de la théorie de l'information, et basée sur les entropies différentielles. Elle est introduite comme alternative pour mesurer la non-Gaussienneté d'une variable aléatoire. Une idée fondamentale en théorie de l'information stipule que l'entropie $\mathcal{H}$ d'une variable aléatoire ayant une distribution gaussienne est plus grande que les entropies des autres variables aléatoires ayant une autre distribution non gaussienne. Ceci justifie que l'utilisation de l'entropie peut mesurer la non-gaussienneté d'une variable aléatoire. La non-gaussienneté d'une variable aléatoire était souvent mesurée par l'utilisation d'une version normalisée de l'entropie différentielle. La néguentropie est donnée par une fonction, notée $\mathcal{J}$, pour une variable aléatoire x. Elle est définie ainsi :

$$\mathcal{J} = \mathcal{H}(x_{gauss}) - \mathcal{H}(x) \tag{3.7}$$

où x_{gauss} est une variable aléatoire Gaussienne ayant la même matrice de covariance que la variable x. Il est à noter que la néguentropie a une propriété intéressante dans le fait qu'elle est invariante face à des transformations linéaires inversibles. La néguentropie est considérée comme le meilleur estimateur de la non-Gaussienneté. Cette mesure est robuste mais reste compliquée à estimer. Il existe une méthode classique pour l'approximation de la néguentropie telle que donnée par l'équation 3.8 pour une variable aléatoire x dont la distribution suit la loi Normale.

$$\mathcal{J} \approx \frac{1}{12} E\{x^3\}^2 + \frac{1}{48} kurt(x)^2 \tag{3.8}$$

Cette dernière approximation est considérée comme non robuste. Cependant, il existe une approche robuste qui consiste à estimer une fonction non quadratique, notée G. Des estimations robustes de la fonction G ont été proposées dans la littérature. Elles sont données par les équations 3.9 et 3.10 :

$$G_1(x) = \frac{1}{a_1} log(cosh\ \ a_1 x) \tag{3.9}$$

$$G_2(x) = -exp(-x^2/2) \tag{3.10}$$

où a_1 est une constante comprise entre 1 et 2, souvent fixée à 1. Ces dernières mesures ont l'avantage d'être robustes et faciles à calculer.

3.3.2.3 Une méthode tensorielle : algorithme JADE

Les composantes indépendantes d'une variable aléatoire peuvent être estimées en se basant sur des tenseurs cumulants d'ordre élevé. Un tenseur cumulant est une généralisation de la matrice de covariance, qui est un tenseur cumulant d'ordre deux. Nous allons nous intéresser dans le reste de la présente section aux tenseurs cumulants d'ordre quatre. Un tenseur cumulant d'ordre 4 est un opérateur linéaire défini par les cumulants d'ordre 4. Il s'agit d'un tableau de dimension 4 dont les entrées sont les cumulants croisés d'ordre 4 obtenus à partir des données initiales, soit le vecteur x. Un tenseur cumulant est noté par $cum(x_i, x_j, x_k, x_l)$ tels que les indices i, j, k, et l sont dans l'intervalle $[1, n]$. Les lecteurs intéressés par les tenseurs cumulants peuvent se rapporter à la référence [HKO01]. Une variable aléatoire possède des composantes indépendantes si tous les cumulants d'un tenseur, avec au moins deux indices différents, sont égaux à zéro.

Un tenseur cumulant possède une décomposition en valeurs propres. La matrice des valeurs propres d'un tenseur est donnée par M. La diagonalisation de cette application linéaire permet d'effectuer la séparation voulue sous certaines conditions. Dans le cas où les valeurs propres sont distinctes, l'indépendance entre les composantes d'une variable aléatoire est facilement estimée. Ce problème peut être résolu par l'algorithme intitulé "*Joint approximate Diagonalization of Eigenmatrices (JADE)*" proposé dans [CS93]. La matrice W diagonalise $F(M)$ pour chaque matrice M, tel que $F(M)$ est une fonction de transformation linéaire. En d'autres termes, la matrice $WF(M)W^T$ est diagonale. L'idée de l'algorithme $JADE$ est de rendre la matrice $Q = WF(M_i)W^T$ aussi diagonale que possible, ceci pour chaque matrice M_i tels que $i = 1, ..., k$ et k est le nombre de matrices pour un tenseur donné. En pratique, nous ne pouvons pas obtenir une diagonalisation exacte à cause des erreurs d'échantillonnage. La diagonalité de la matrice Q peut être mesurée par la somme des carrés des éléments hors diagonale. La minimisation de la somme des carrés des éléments hors diagonale est équivalent à la maximisation de la somme des éléments de la diagonale. Cette idée peut être formulée ainsi :

$$\mathcal{J}_{JADE} = \sum_i \left\| diag(WF(M_i)W^T) \right\|^2 \tag{3.11}$$

où $\|diag(.)\|^2$ désigne la somme des carrés de la diagonale. L'algorithe $JADE$ nécessite le calcul de tous les cumulants d'ordre 4 et a donc une complexité en $\mathcal{O}(n^4)$.

3.3.2.4 Estimation par maximum de vraisemblance : algorithme FastICA

La maximisation de la non-Gaussienneté peut être estimée par une méthode basée sur la descente de Gradient [HKO01]. La convergence de ce genre de méthode est lente et dépend du choix des taux d'apprentissage. Cependant, un mauvais choix des paramètres d'apprentissage rend parfois impossible la convergence vers une solution. Les algorithmes basés sur le principe de "point fixe" est une alternative pour accélérer la convergence tout en améliorant l'exactitude de la solution estimée. L'algorithme le plus connu est celui proposé par A. Hyvarinen [Hyv99], qui s'intilule "*FastICA*". L'algorithme FastICA est itératif, et consiste à estimer les paramètres de la matrice de séparation W, dite aussi de démixage, tels que $W \times X$ maximise la non-Gaussienneté. La non-Gaussienneté est mesurée par approximation de la néguentropie $\mathcal{J}(S)$. Une approximation de la néguentropie a été proposée dans [Hyv99] par $\mathcal{J}(S) \propto [E\{G(S)\} - E\{G(v)\}]^2$ où S est une variable aléatoire, v une variable Gaussienne centrée réduite, et G est une fonction non-quadratique. Nous pouvons détailler l'algorithme *FastICA* de la manière suivante.

Soit g la dérivée de la fonction non-quadratique G. À titre d'exemple, les dérivées des fonctions données par les équations 3.9 et 3.10 sont $g_1(x) = tanh(a_1 x)$ et $g_2(x) = x.exp(-x^2/2)$ respectivement. Le paramètre a_1 est une constante comprise entre 1 et 2, souvent fixée expérimentalement à 1. Les différentes étapes de l'algorithme *FastICA* sont résumées dans l'algorithme 3.

Algorithme 3 FastICA.

1 - Initialiser aléatoirement les paramètres de la matrice de démixage W.

2 - Mettre à jour la matrice de démixage par $W^+ = E\{Xg(W^T X)\} - E\{g'(W^T X)\}W$.

3 - Normaliser la matrice de démixage : $W = W^+ / \|W^+\|$.

4 - S'il n'y a pas de convergence, aller à l'étape 2.

L'algorithme 3 est appliqué sur des données prétraitées. Le prétraitement consiste à *blanchir*

les données initiales "*whitening data*" [1]. Nous rappelons que la variance de $W^T X$ est égale à 1. Dans le cas des données blanchies, ceci correspond à contraindre la norme de W à l'unité. Le but de l'algorithme est de trouver la direction, qui correspond à une configuration de paramètres de la martice W, permettant de maximiser la non-Gaussienneté de la projection $W^T X$. La convergence signifie que les valeurs de la matrice de démixage W obtenues pour deux itérations successives tendent vers la même direction. Autrement dit, la convergence est atteinte dans le cas où la norme du produit, $|W_n.W_{n-1}|$ est égale à 1. La norme du produit est justifiée par le fait que le signe de $W_n.W_{n-1}$ n'a pas d'importance puisque W et $-W$ définissent la même direction.

3.3.3 L'ACI pour l'extraction des régions en mouvement

L'ACI a été récemment exploitée pour l'extraction des régions en mouvement à partir d'une séquence d'images. En se référant à la littérature, nous n'avons recensé que peu de publications récentes dans ce domaine [CZ03], [ZC06] et [TL09]. Dans [CZ03] et [ZC06], les auteurs proposent une méthode d'extraction d'objets en mouvement à partir d'une séquence d'images. La méthode proposée est divisée en deux parties. La première consiste à estimer les composantes indépendantes en se basant sur le principe d'infomax [BS95] et une décomposition en valeurs singulières. La deuxième partie consiste en un post-traitement permettant la localisation des régions d'intérêt par analyse multi-échelle et par une décomposition en ondelettes sur les signaux indépendants obtenus par l'ACI.

La recherche la plus récente est celle menée par D.-M. Tsai [TL09]. L'auteur propose un nouvel algorithme de séparation de sources, basé sur le principe d'indépendance statistique entre les composantes à estimer. L'indépendance est directement mesurée par la différence des fonctions de densité de probabilités jointes "*PDFs*", ainsi que le produit des fonctions de densité de probabilités marginales. Les *PDFs* sont estimées à partir d'un histogramme mesurant la fréquence des distributions des données observées. Un algorithme basé sur la méthode d'optimisation intitulée "optimisation par essaims particulaires" permettant l'estimation de la matrice de séparation optimale, est proposé. L'auteur a evalué la méthode proposée sur quelques séquences d'images acquises dans des environnement intérieurs avec des conditions d'illumination assez controlées.

3.4 Méthode proposée

Nous présentons dans ce paragraphe la méthode permettant la séparation entre les objets en mouvement et les fonds de scène à partir d'une séquence d'images. L'algorithme proposé se divise

1. Le blanchiment de données consiste à transformer linéairement un vecteur d'observations en un autre de sorte que la matrice de covariance du vecteur estimé est égale à la matrice d'identité. Les composantes d'un vecteur d'observations blanchi sont non-corrélées, et leur variances sont égales à l'unité.

en deux étapes :

– Une étape d'apprentissage.

– Une étape de détection.

La première étape consiste à estimer le modèle de bruit ainsi que la matrice de séparation, notée W^* obtenue par analyse en composantes indépendantes. La deuxième étape se divise aussi en deux parties : la première partie consiste à approximer les régions affectées par du mouvement en utilisant la matrice de séparation estimée pendant l'étape d'apprentissage. La deuxième partie permet de raffiner les régions extraites en introduisant une étape de lissage, effectuée en minimisant une énergie dans un cadre de champs de markov aléatoires. Les différentes étapes de l'algorithme proposé sont décrites sur la figure 3.1.

3.4.1 Description détaillée du synoptique de traitement

La première étape de l'algorithme consiste à modéliser le bruit à partir de quelques images successives non affectées par du mouvement, c'est à dire avec un fond de scène vide. Ces images de fond sont en général instables en termes d'illumintaion et de mouvement. Ces instabilitées sont généralement dûes aux variations d'illumination, au processus d'échantionnage au niveau du capteur *CCD*, ou aux mouvements continus d'objets appartenant au fond, comme le mouvement de branches d'arbres par exemple. Nous considérons ce genre de mouvement comme du bruit dans notre algorithme. Le modèle de bruit entre deux images successives est obtenu en appliquant une analyse en composantes indépendantes dont la matrice de données est formée de deux images successives correspondant au fond de scène. Les sources estimées correspondent alors à deux signaux. Le premier signal est une estimation du fond alors que le deuxième correspond aux changements survenus entre les deux images successives.

A la fin de la période d'apprentissage, les deux images les plus récentes sont retenues pour l'estimation de la matrice de séparation, notée W^*, par analyse en composantes indépendantes. La matrice de données est formée de deux images. La première correspond à une image de la séquence ne contenant que du fond, alors que la deuxième correspond à une image contenant du fond, sur laquelle un objet est ajouté. La deuxième étape consiste en l'étape de détection. La matrice de séparation, W^*, précédemment estimée est utilisée dans l'étape de détection pour l'estimation de l'objet. Les régions corespondant à des objets sont estimées par la multiplication de deux matrices : une matrice de données, et une matrice de séparation. La matrice de données est formée de deux images. La première correspond à l'image de fond la plus récente (IFPR) alors que la deuxième est l'image courante de la séquence. La matrice résultat correspond aux sources séparées estimées.

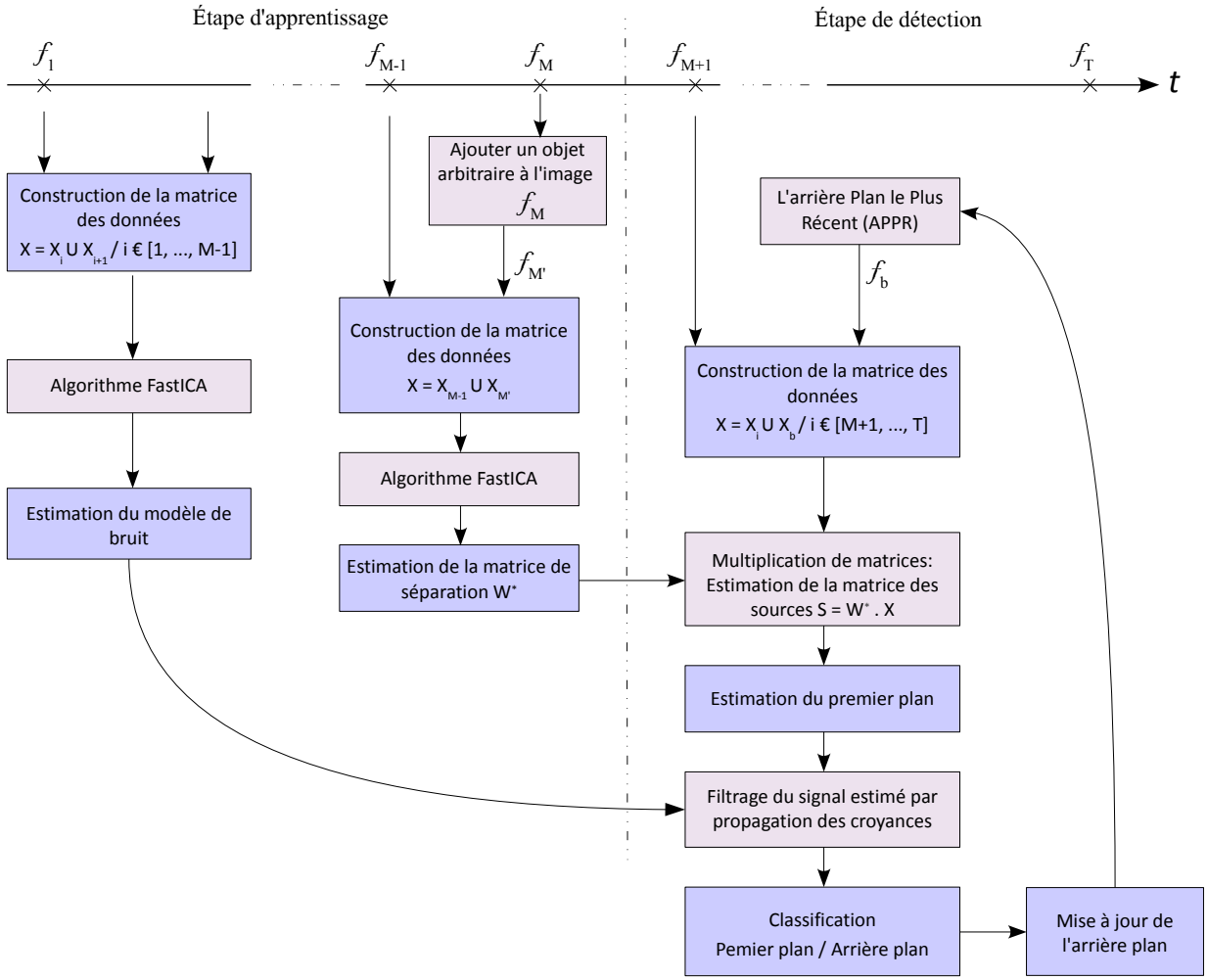


FIGURE 3.1 – Vu générale de l'algorithme d'extraction des régions en mouvement.

Il s'agit de deux sources estimées : le fond et les objets. C'est cette dernière source, les objets en mouvement, qui nous intéresse pour le reste de l'algorithme.

La composante indépendante, qui correspond aux objets affectés par du mouvement, est une image contenant à la fois des régions affectées par du mouvement, sous un fond uniforme. Les pixels appartenant au fond sont de couleur très similaires, de sorte qu'aucun détail n'est visible du vrai fond de la scène observée. Les pixels affectés par du mouvement sont alors bien mis en valeur.

La dernière étape de l'algorithme consiste à "post-traiter" le signal correspondant aux objets détectés. Le "post-traitement" est un lissage spatio-temporel permettant une meilleure classification des pixels selon qu'ils appartiennent au fond ou à un objet. Le problème du lissage est largement abordé dans la littérature. Cependant, les différentes méthodes existantes se basent généralement sur l'aspect spatial. Le sur-lissage génère dans la plupart des cas la non conservation des frontières

des objets. Nous introduisons l'aspect temporel dans la conception d'une nouvelle méthode de lissage spatio-temporelle. Le lissage est effectué par une propagation de croyance spatio-temporelle dont le but est de rendre homogène le fond tout en conservant la frontière des objets. Les différentes étapes sont résumées dans l'algorithme 4.

Algorithme 4 Extraction des régions affectées par du mouvement dans une séquence d'images.

Etape d'apprentissage

1. Estimation du modèle de bruit cumulé par analyse en composantes indépendantes. La matrice de données est constituée de deux images consécutives non affectées par du mouvement. Le modèle de bruit obtenu correspond à un vecteur de dimension 6 contenant la moyenne et l'écart-type de chaque composante couleur.

2. Estimer la matrice de séparation, notée W, en effectuant une ACI sur une matrice de données X. La matrice de données est composée de deux images. La première correspond à une image du fond et la deuxième correspond à une image du fond sur laquelle un objet arbitraire est ajouté.

Etape de détection

3. Première phase : détection d'objets en mouvement. La composante qui correspond aux objets est obtenue en multipliant la matrice de séparation précédemment estimée par une matrice de donnée formée par l'image du fond la plus récente et l'image courante de la séquence.

4. Deuxième phase : filtrage de l'image source obtenue à l'étape 3 par propagation de croyance spatio-temporelle. Cette étape permet de rendre le fond aussi homogène que possible pour une meilleure extraction des objets en mouvement.

3.4.2 Formulation du problème d'extraction d'objets en mouvement par ACI

Comme nous l'avons mentionné au début du présent chapitre, les méthodes existantes ont permis de résoudre en partie les problèmes liés à l'extraction comme les changements d'illumination et le mouvement continu des faux objets tels que les branches d'arbres. La conception d'un tel algorithme dépend des facteurs suivants : la complexité algorithmique et la mémoire nécessaire pour accomplir cette tâche. Un compromis doit être trouvé entre : la minimisation du nombre de paramètres en les rendant le plus possible adaptatifs, le temps mis pour la mise à jour, et la qualité

d'extraction des régions affectées par du mouvement.

Le point de départ consiste à définir les données à partir desquelles les sources sont estimées. La formulation proposée par la suite prend en compte l'information couleur contrairement à ce qui a été proposé dans [ZC06] et [TL09]. Pour que l'ACI ait un sens, la contrainte de l'indépendance entre les composantes à estimer doit être vérifiée. Ce principe est facilement justifiable puisque l'état d'un pixel affecté par du mouvement ne dépend pas de son état antérieur (avant qu'il soit affecté par du mouvement). Autrement dit, il n'existe aucune corrélation entre les objets et le fond dans le sens où le mouvement est indépendant de l'état du fond : il s'agit de deux évènements indépendants. Notons que le cas de la transparence entre un objet et du fond ne correspond pas à une indépendance en terme de mouvement. Un exemple d'une transparence est le cas d'une partie du fond, vu à travers d'une vitre d'un véhicule en mouvement. Il s'agit dans ce cas d'une forte corrélation entre le fond et l'objet.

Une vidéo est définie par une séquence d'images, notée par $\mathcal{S}$. Une image acquise à l'instant t est notée par I_t, tels que $t = 1, \ldots, T$, T correspond à la longueur de la séquence d'images, et $\mathcal{S} = \bigcup_{t \in T} I_t$. Chaque image I_t est linéarisée, et est représentée sous forme d'un vecteur de taille $K = M \times N$, où M et N sont la longueur et la largeur de l'image. Un pixel $p(i, j)$ d'une image est représenté par $I_{p,t}$, où i et j correspondent aux positions verticale et horizontale respectivement. Nous avons mentionné précédemment que nous introduisons l'information couleur dans la formulation du problème. Dans le cas de l'espace couleur RVB, la notation c^i correspond à l'une des composantes de telle sorte que c^1 correspond à la composante rouge, c^2 à la composante verte, et c^3 à la composante bleu. La matrice de données, X, est de deux lignes et $(K \times 3)$ colonnes. Chaque ligne de la matrice X correspond à un vecteur représentant les composantes couleur contigües d'une image, organisées d'une manière adjacente dans le vecteur. La première ligne correspond à l'image vecteur représentant le fond, tandis que la deuxième correspond au vecteur image représentant le fond sur lequel un objet est ajouté. La matrice X est par conséquent la suivante :

$$X = \begin{pmatrix} I^{c^1}_{IFPR}, & I^{c^2}_{IFPR}, & I^{c^3}_{IFPR} \\ I^{c^1}_{IFPR+Ob}, & I^{c^2}_{IFPR+Ob}, & I^{c^3}_{IFPR+Ob} \end{pmatrix} \tag{3.12}$$

Les notations $I^{c^i}_{APPR}$ et $I^{c^i}_{APPR+PP}$ dans l'équation 3.12 désignent la composante c^i de l'image I qui correspond à *l'image de fond la plus récente (IFPR)* et à *l'image de fond la plus récente avec un objet ajouté (IFPR+Ob)*, respectivement. Le modèle basé sur l'ACI est donné alors par :

$$\tilde{S} = \tilde{W}.X + \tilde{n} \tag{3.13}$$

Rappelons que la matrice $\tilde{S}$ est la matrice des composantes indépendantes séparées de même dimension que la matrice X. La matrice $\tilde{W}$ est la matrice de séparation estimé par l'algorithme *FastICA*. Nous avons introduit dans l'équation 3.13 le vecteur $\tilde{n}$ qui représente le bruit présent dans les données, soit la matrice X.

3.5 Estimation des régions affectées par le mouvement

3.5.1 Modélisation du bruit

La modélisation de bruit est une étape importante permettant d'identifier les pixels du fond ayant un mouvement continu et répétitif. L'algorithne *FastICA* est exécuté sur chaque image du fond exploitée lors de la période d'apprentissage. Pour chaque étape de l'apprentissage, une matrice de données formée par deux images consécutives du fond est traitée par ACI pour estimer la composante qui correspond au fond non bruité et la composante du fond contenant du bruit seulement. Cette dernière correspond en général à une image dont les pixels affectés par le mouvement sont mis en valeur, et possède des couleurs différentes. À chaque fois qu'un modèle de bruit est estimé, un vecteur $V = \langle(\mu_R, \sigma_R), (\mu_V, \sigma_V), (\mu_B, \sigma_B)\rangle$ de dimension 6 est mis à jour pour chaque pixel. Chaque couple (μ_i, σ_i) correspond à la moyenne mobile et l'écart type de la composante couleur $i \in RVB$ dans le cas de l'espace couleur RVB. À la fin de l'étape d'apprentissage, un modèle final du bruit est obtenu en fixant un seuil sur les écarts-types. Les pixels ayant des écarts-type supérieurs au seuil sont classés comme des pixels affectés par du bruit. Il seront alors pris en compte lors de l'étape de classification fond/objet. La figure 3.2 illustre le principe d'estimation des composantes indépendantes par ACI.

3.5.2 Détection d'objets stationnaires ou en mouvement

L'élément clé de la résolution d'un modèle linéaire basé sur l'ACI consiste en l'estimation des paramètres de la matrice $\tilde{W}$ permettant une meilleure séparation des sources cachées. La variation entre les données dans le temps n'est pas considérable puisque le fond varie d'une manière homogène et progressive. La matrice de séparation n'est estimée qu'une seule fois durant la phase d'apprentissage. La matrice de données est formée de deux images. La première est une image du fond obtenue à l'instant $t = (m - 1)$, selon la figure 3.1. La deuxième est l'image du fond obtenue à l'instant $t = m$ sur laquelle un objet est ajouté. Nous avons utilisé l'algorithme FastICA

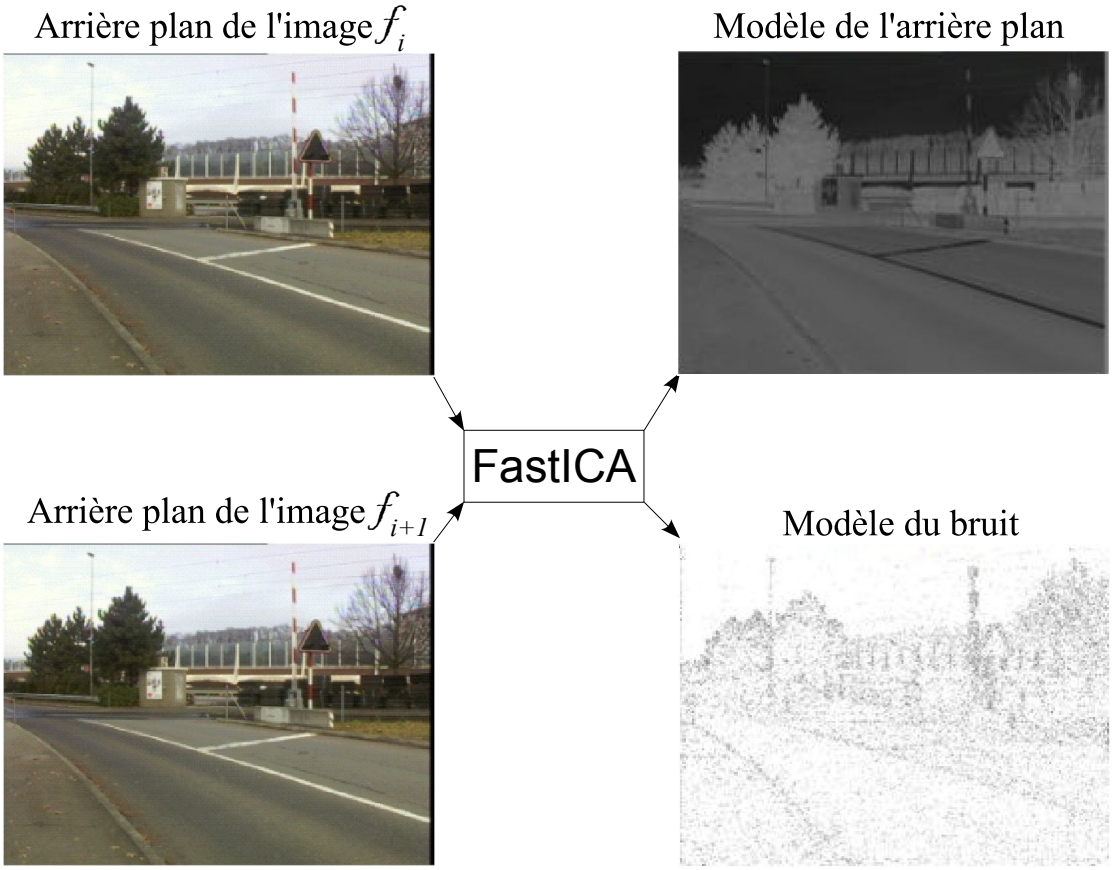


FIGURE 3.2 – Estimation des composantes indépendantes par ACI. La martice de données est formée par deux images successives correspondant à l'arrière-plan. Les composantes estimées correspondent à l'arrière-plan non bruité, et au bruit.

pour l'estimation de la matrice de séparation $\tilde{W}$. Cette matrice est ensuite utilisée dans l'étape de détection.

À chaque image de la séquence, la matrice de données X est mise à jour. La première ligne correspond à l'image de fond la plus récente (IFPR). La deuxième ligne correspond à une image de la séquence. La matrice $\tilde{S}$ contenant les composantes indépendantes, est estimée par une multiplication de la matrice de données et la matrice de séparation. Le modèle de l'ACI est alors le suivant :

$$\begin{pmatrix} I_{AP}^{c^1} & I_{AP}^{c^2} & I_{AP}^{c^3} \\ I_{PP}^{c^1} & I_{PP}^{c^2} & I_{PP}^{c^3} \end{pmatrix} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \times \begin{pmatrix} I_{IFPR}^{c^1}, & I_{IFPR}^{c^2}, & I_{IFPR}^{c^3} \\ I_{IFPR+Ob}^{c^1}, & I_{IFPR+Ob}^{c^2}, & I_{IFPR+Ob}^{c^3} \end{pmatrix} \tag{3.14}$$

Il est à noter que la matrice de dé-mixage $\tilde{W}$, dite aussi matrice de séparation, est de taille 2×2. Les paramètres de la matrice $\tilde{W}$ représentent un filtre permettant la séparation des composantes indépendantes en attribuant un poids à chaque ligne de la matrice de données X. La première ligne de la matrice $\tilde{S}$ correspond au fond estimé sans le/les objets. La deuxième ligne correspond uniquement aux objets affectés par du mouvement, sans aucun détail sur le fond. C'est cette deuxième ligne qui nous intéresse. L'image peut être reconstruite par une simple conversion d'un espace 1D vers un espace 2D. Nous rappelons qu'une image est linéarisée et représentée sous forme d'un vecteur unidimensionnel de telle sorte que les composantes couleur sont adjacentes. Le passage inverse peut être obtenu par une transformation inverse. Tout pixel 2D, qui correspond à l'échelle image, possède trois composantes couleur. À titre d'exemple, les composantes couleur $p^R(i,j)$ pour le rouge , $p^V(i,j)$ pour le vert, et $p^B(i,j)$ pour le bleu, d'un pixel p de coordonnées (i,j) de l'image de l'arrière-plan, sont obtenues par la transformation donnée par l'équation 3.15. Les composantes $p^R(i,j) = \tilde{S}(u,v_1)$, $p^V(i,j) = \tilde{S}(u,v_2)$, et $p^B(i,j) = \tilde{S}(u,v_3)$ sont des éléments de la ligne $u = 0$ de la matrice $\tilde{S}$ tels que v_1, v_2, et v_3 sont les positions des composantes couleur dans la matrice.

$$p^c(i,j) \rightleftharpoons \tilde{S}(0,(l*K)+(i*M)+j) \tag{3.15}$$

où c correspond à une composante couleur de l'espace RVB, le paramètre $K = M \times N$ correspond à la taille de l'image, et l dépend de la composante couleur de telle sorte que $l = 0$ pour la composante rouge, $l = 1$ pour la composante verte, et $l = 2$ pour la composante bleu. La figure 3.3 illustre le principe de séparation des sources. Les deux images en haut à gauche et en bas à gauche correspondent, respectivement, au fond, et une image de la séquence sur laquelle une voiture est ajoutée. Ces deux images forment la matrice des données sur laquelle l'algorithme ***FastICA*** est appliqué pour l'estimation de la matrice de dé-mixage, et pour la séparation des signaux sources. Les composantes indépendantes estimées correspondent au fond, en haut à droite, et les objets seulement sans aucun détail sur le fond, en bas à droite.

La figure 3.4 illustre un exemple de scène où l'on cherche à détecter les objets en mouvement. Les images en haut à gauche et en haut à droite de la figure 3.4 correspondent à l'image du fond, et une image de la séquence contenant une voiture et deux piétons. L'image en bas à gauche correspond aux objets détectés par analyse en composantes indépendantes. Pour avoir plus de détails, nous avons zoomé sur une petite région (carré rouge) afin de bien distinguer les pixels appartenant aux objets de ceux constituant le fond. Ceci représente une première étape de la séparation entre le fond et les objets.

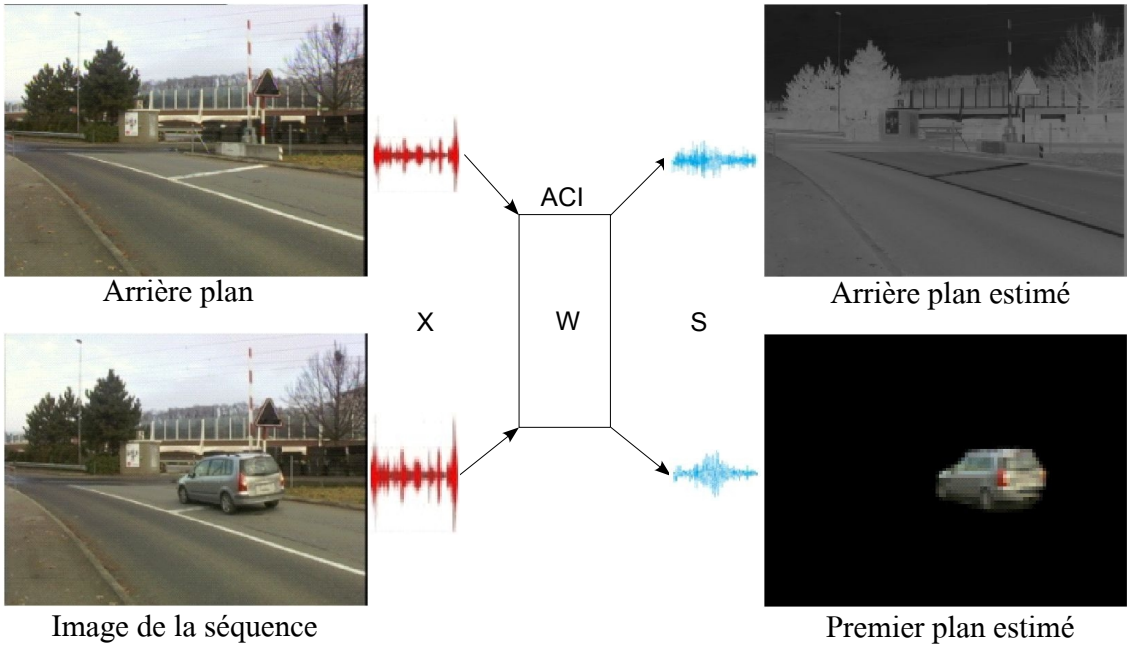


FIGURE 3.3 – Principe de séparation des deux composantes, fond et objet, par analyse en composantes indépendantes. La matrice de données est formée par deux images : la première ne contenant que le fond et la deuxième correspond à une image de la séquence. Les deux composantes estimées correspondent au fond, et à l'objet séparés.

3.6 Filtrage spatio-temporel pour la classification fond/objets

3.6.1 Principe

Seule l'image contenant les objets en mouvement détectés est prise en compte pour la classification fond/objets. Le fond de scène est relativement homogène sur un plan colorimétrique tandis que les pixels affectés par du mouvement sont mis en valeur. Bien que les pixels qui correspondent aux objets en mouvement sont bien distincts, une classification automatique est nécessaire pour une séparation fond/objets, efficace et fiable. Nous proposons dans ce paragraphe une nouvelle approche permettant un filtrage spatio-temporel de l'image correspondant aux objets détectés, notée $\tilde{S}_{Ob}$. Ce filtrage est un post-traitement permettant d'éliminer des pixels aberrants résiduels lors de l'extraction des objets. Le filtrage consiste à mettre à jour, itérativement, la couleur de chaque pixel tout en tenant compte des couleurs des pixels dans un voisinage spatio-temporel. L'idée s'inspire de la propagation de croyance sélective proposée dans le chapitre précédent. L'algorithme proposé dans ce paragraphe se base sur un cadre général similaire à celui détaillé dans le chapitre précédent.

En se basant sur les modèles graphiques de représentation d'image, tel que détaillé au §3.3.1

FIGURE 3.4 – Exemple d'un objet approximé par analyse en composantes indépendantes. Les images en haut à gauche et à droite correspondent à l'image du fond et une image de la séquence, respectivement. L'image en bas à gauche est la composante de l'objet estimée à partir de l'ACI appliquée sur les deux images en haut à gauche et à droite. En voyant de près une petite région (rectangle rouge), nous remarquons que l'objet est mis en valeur par rapport à l'arrière-plan. La classification fond / objet n'est pas tout à fait évidente puisque le fond est bruité.

du chapitre précédent, l'image $\tilde{S}_{Ob}$ est modélisée par un graphe non orienté $\mathcal{G} = \langle P, E \rangle$ tel que P correspond à l'ensemble des pixels, dit aussi noeuds, et E l'ensemble des arêtes qui correspondent aux dépendances spatiales entre les pixels. L'état d'un pixel p est modélisé par un vecteur de paramètres noté $\dot{s}_p \langle l_{t-1}, \ldots, l_{t-k} \rangle$. Les paramètres l_{t-i} correspondent aux labels, $i = 0, \ldots, k$, correspondant au pixel p pour les k dernières images.

En se basant sur le principe des 4-voisins connexes, la probabilité jointe dans le champ de Markov aléatoire MRFs est donnée par :

$$P(\mathcal{G}) = \prod_{p \in P} \Phi(\dot{s}_p) + \prod_{\substack{p \in P \\ q \in \mathcal{N}_{s,p}}} \Psi(\dot{s}_p, \dot{s}_q) + \prod_{p(t-i) \in \mathcal{N}_{t,p}} \Theta\left(\dot{s}_{p(t)}, \dot{s}_{p(t-i)}\right) \tag{3.16}$$

Dans l'équation 3.16, la fonction Φ correspond au terme permettant l'estimation d'un message

à transmettre dans le graphe. La fonction Ψ mesure le degré de corrélation spatiale entre deux pixels voisins, et Θ mesure le degré de corrélation temporelle d'un même pixel à des instants différents. $\mathcal{N}_{s,p}$ et $\mathcal{N}_{t,p}$ correspondent au voisinage spatial et temporel du pixel p. La notation $\mathcal{N}_{s,p}$ désigne le voisinage spatial d'un pixel p. $p(t-i)$ désigne un pixel à l'instant $t-i$ tel que $i = 1, \ldots, k$. Comme dans le chapitre précedent, le problème d'inférence dans le graphe est NP-hard, d'où l'idée d'approximer la solution de labellisation optimale par propagation de croyance. Le problème revient alors à optimiser une fonction d'énergie par Maximum a Posteriori (MAP) ou par l'estimateur du minimum du carré de l'erreur moyenne "*(MMSE)*". La figure 3.5 illustre le principe de propagation de croyance spatio-temporelle utilisé.

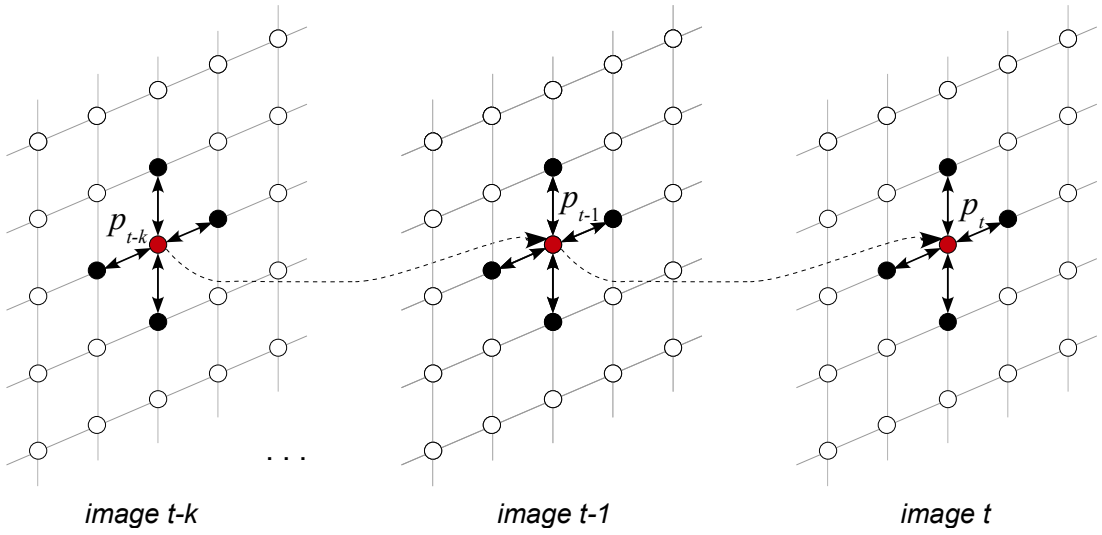


FIGURE 3.5 – Principe de propagation de croyance spatio-temporelle. La figure illustre, pour un pixel donné, le voisinage spatial qui correspond aux 4 voisins connexes, et le voisinage temporel qui correspond au même pixel à des instants antérieurs.

3.6.2 Propagation de croyance spatio-temporelle

Le problème d'inférence peut être reformulé en un problème de minimisation d'une fonction d'énergie par la recherche d'une fonction de labellisation, notée $f^* \subset \mathcal{F}$, minimisant une certaine énergie. $\mathcal{F}$ correspond à l'ensemble des solutions possibles. Une labellisation consiste à attribuer un label l à chaque pixel du graphe. L'énergie à minimiser est une somme linéaire de trois termes : terme de données, terme de lissage spatial, et terme de lissage temporel.

$$E(\mathcal{G}) = E_{data}(\hat{f}) + E_{lissage\_spatial}(\hat{f}) + E_{filtrage\_temporel}(\hat{f}) \tag{3.17}$$

Maximiser la probabilité donnée par l'équation 3.16 est équivalent à minimiser la négative du logarithme de chaque terme de la fonction objectif. En se basant sur les égalités $\phi = -\log \Phi$, $\psi = -\log \Psi$, et $\theta = -\log \Theta$, la fonction peut être reformulée de la façon suivante :

$$E(\mathcal{G}) = \sum_{p \in P} \phi(\dot{s}_p) + \sum_{\substack{p \in P \\ q \in \mathcal{N}_{s,p}}} \psi(\dot{s}_p, \dot{s}_q) + \sum_{p(t-i) \in \mathcal{N}_{t,p}} \theta\left(\dot{s}_{p(t)}, \dot{s}_{p(t-i)}\right) \tag{3.18}$$

Les fonctions ϕ, ψ, et θ dans l'équation 3.18 correspondent aux termes de données, de lissage spatial, et de filtrage temporel, respectivement. La configuration de labels permettant de minimiser la fonction d'énergie $E(\mathcal{G})$ est retenue comme la solution de labellisation optimale f^*, sachant qu'une solution intermédiaire est notée par $\hat{f}$.

3.6.2.1 Terme de données

En vision stéréoscopique, le terme de données correspond au coût permettant la mise en correspondance de deux pixels appartenant aux images gauche et droite. D'une manière similaire, le terme de données correspond à la différence d'intensité ou de couleur d'un même pixel. La différence représente le coût d'attribuer un label $l = \langle c^1, c^2, c^3 \rangle$ à un pixel p, $\dot{s}_{p,l}$, étant donné l'état courant du même pixel $\dot{s}_p$. La différence entre deux états d'un même pixel, notés $\dot{s}_p$ et $\dot{s}_{p,l}$, est exprimée par la distance Euclidienne dans l'espace $\mathbb{R}^3$. Le terme de données est le suivant :

$$\phi(\dot{s}_p) = \begin{cases} 0 & \text{si } |\dot{s}_p - \dot{s}_{p,l}| \leq \varepsilon \\ |\dot{s}_p - \dot{s}_{p,l}| & \text{sinon} \end{cases} \tag{3.19}$$

Il faut noter que ε est une valeur constante et que $\dot{s}_{p,l}$ est l'état d'un pixel p ayant un label l.

3.6.2.2 Terme de lissage spatial

Le choix du terme de lissage est compliqué à déterminer. Plusieurs formulations ont été proposées dont chacune dépend fortement du problème à résoudre. Le terme de lissage proposé dans [PTK85] rend le graphe uniformément lisse, ce qui réduit les performances du résultat en particulier à la frontière des objets. D'autres formulations du terme de lissage ont été récemment proposées dans le cadre de l'appariement stéréoscopique tel que dans [AKT08]. Le terme de lissage proposé dans [IM06] est donné par $\psi(\dot{s}_p, \dot{s}_q) = min(a, b|\dot{s}_p - \dot{s}_q|)$, où b est un paramètre choisi

empiriquement, et a un paramètre calculé d'une manière dynamique par $a = a_0 \exp(-\|\nabla I\| / \varepsilon)$, où a_0 est une constante, $\|\nabla I\|$ le gradient de l'image de référence, et ε la moyenne de l'image de gradient. En se basant sur le modèle de Potts [Wu82], nous définissons le terme de lissage spatial par la différence d'intensité ou de couleur entre un pixel et ses 4 voisins connexes. Il est donné par l'équation 3.20 :

$$\psi(\dot{s}_p, \dot{s}_q) = \begin{cases} 0 & \text{si } \Delta_{p,q} \leq \xi \\ T.\Delta_{p,q} & \text{sinon} \end{cases} \tag{3.20}$$

Dans cette l'équation, $\Delta_{p,q}$ correspond à la différence colorimétrique entre les pixels p er q, ξ une constante, et T une constante dite *température du modèle de Potts*.

3.6.2.3 Terme de lissage temporel

Le troisième terme de la fonction d'énergie permet une amélioration des performances du processus de labellisation. Le terme de lissage temporel, dit aussi filtrage temporel, assure la cohérence temporelle entre les labels. L'état $\dot{s}_{p,t}$ d'un pixel p à l'instant t est inconnu et nous cherchons à estimer. L'historique et la dépendance temporelle entre les labels d'un même pixel permet de renforcer la contrainte de continuité des labels. Cependant, la classe d'appartenance d'un pixel dépend des classes d'appartenance du même pixel aux instants précédents. Pour un pixel donné et pour un voisinage donné, plus les labels sont temporellement proches du pixel donné et plus leur influence en terme de décision de classification fond/objet est grande. Nous illustrons ci-dessous ce phénomène avec un voisinage temporel de $i = 4$ images :

$$F_{p,t-4} \rightarrow F_{p,t-3} \rightarrow F_{p,t-2} \rightarrow F_{p,t-1} \rightarrow F_{p,t} \vee Ob_{p,t} \tag{3.21}$$

$$Ob_{p,t-4} \rightarrow Ob_{p,t-3} \rightarrow Ob_{p,t-2} \rightarrow Ob_{p,t-1} \rightarrow F_{p,t} \vee Ob_{p,t} \tag{3.22}$$

$$Ob_{p,t-4} \rightarrow Ob_{p,t-3} \rightarrow F_{p,t-2} \rightarrow F_{p,t-1} \rightarrow F_{p,t} \vee Ob_{p,t} \tag{3.23}$$

$$F_{p,t-4} \rightarrow F_{p,t-3} \rightarrow Ob_{p,t-2} \rightarrow Ob_{p,t-1} \rightarrow F_{p,t} \vee Ob_{p,t} \tag{3.24}$$

Les notations $F_{p,t-i}$ et $Ob_{p,t-i}$ désignent qu'à l'intant $(t - i)$, le pixel p est classé en tant que

fond et objet respectivement. La représentation $\vee$ désigne l'opérateur logique "OU". Nous partons du principe que l'influence de l'état du pixel p à l'instant $(t-1)$ est plus importante que celle à l'intant $(t-2)$, et ainsi de suite. Si nous prenons l'exemple de l'équation 3.21, nous constatons que le pixel p à l'instant t correspond probablement à un pixel du fond puisqu'il était classé comme fond aux instants précédents les plus proches. Le même principe est observé pour les exemples 3.22, 3.23, et 3.24. L'expression du lissage temporel est fournie par l'équation suivante :

$$\theta\left(\dot{s}_{p(t)}, \dot{s}_{p(t-i)}\right) = \sum \kappa . \left\| \dot{s}_{p(t)} - \dot{s}_{p(t-i)} \right\| \tag{3.25}$$

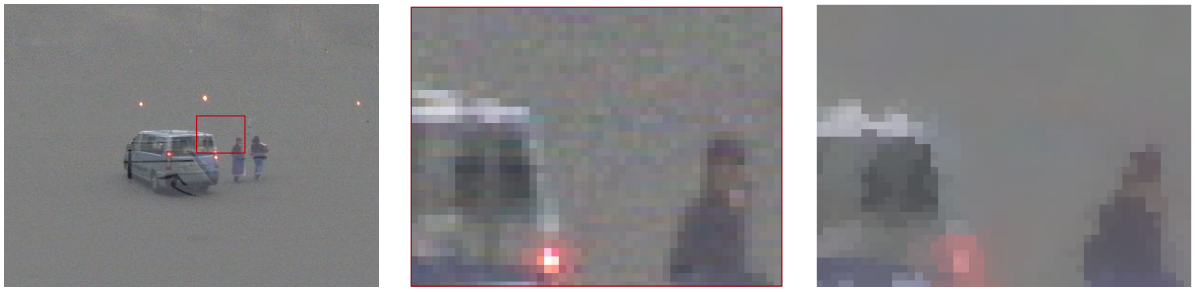
Le paramètre κ est une valeur binaire permettant d'activer ou de ne pas activer les états précédents dans le calcul du terme de lissage temporel. Il dépend de la classe d'appartenance du pixel le plus temporellement proche. Dans le cas où le pixel p à l'instant $(t-1)$ est classé comme du fond, le paramètre κ vaut 1 pour tous les pixels classés comme fond aux instants précédents, et 0 autrement. Dans le cas où le pixel p à l'instant $(t-1)$ est classé comme appartenant à un objet, le paramètre κ vaut 1 pour tous pixels classés comme appartenant à un objet aux instants précédents, et 0 autrement.

3.6.3 Processus de passage de messages entre pixels

Au départ, chaque pixel du graphe possède son propre label (caractéristique colorimétrique). Un label est un vecteur de dimension 3 qui correspond à la moyenne de chaque composante couleur calculée sur les pixels 4 voisins connexes d'un pixel donné. L'étape suivante consiste à faire passer des messages (échange des caractéristiques colorimétriques) entre les pixels. Pour chaque pixel, le label est mis à jour à chaque itération en fonction des messages reçus par ses voisins. Le sens de passage des messages est le suivant : de gauche à droite $(G \rightarrow D)$, de haut en bas $(H \rightarrow B)$, de droite à gauche $(D \rightarrow G)$, puis de bas en haut $(B \rightarrow H)$. Autrement dit, si le sens de propagation des message est $(G \rightarrow D)$ à l'itération i, alors le sens de propagation des messages à l'itération $(i+1)$ est $(H \rightarrow B)$. Il s'agit d'un processus itératif tel qu'à chaque itération une solution $\hat{f}$ est estimée. A chaque itération une énergie est calculée pour une solution donnée selon l'équation 3.18. La solution optimale est atteinte dans le cas où l'énergie est inférieure à un seuil. Le message m transmis d'un pixel p vers un pixel q à l'itération i est exprimé par l'équation suivante :

$$m^{i}_{p \rightarrow q} = \arg\min_{l} \;\; \phi(l_p) + \psi(l_p, l_q) + \theta(l_{p(t)}, l_{p(t-i)}) + \sum_{x \in \mathcal{N}_s, p \backslash q} m^{i-1}_{x \rightarrow p} \tag{3.26}$$

La notation $\mathcal{N}_s, p \backslash q$ désigne le voisinage spatial du pixel p sauf le pixel q. La propagation de croyance est une technique permettant une estimation rapide et efficace d'une solution relative à un problème complexe. Une solution consiste à attribuer un label, qui est une couleur, à chaque pixel de l'image de sorte que les pixels du fond ont des labels similaires et homogènes, souvent différents des labels des objets. Nous cherchons à travers cette méthode à rendre le fond aussi lisse que possible afin de pouvoir facilement séparer les objets du fond. La classification des pixels de l'image selon deux classes, fond et objets, ne nécessite pas un nombre élevé d'itérations puisque les pixels du fond sont bruités et leur variation colorimétrique est homogène. La solution optimale préserve les frontières des objets grâce à la contraite de discontinuité introduite dans la fonction d'énergie. La figure 3.6 fournit le résultat de filtrage de l'image estimée par analyse en composantes indépendantes de la figure 3.4. Nous remarquons que l'image filtrée 3.6 (c) préserve les frontières des objets.



(a) (b) (c)

FIGURE 3.6 – Exemple illustrant le résultat de filtrage appliqué sur la composante du premier plan approximée par ACI. Reprenons l'exemple de la figure 3.4, image (a). L'image (b) montre l'objet approximé par ACI sur une petite région (rectangle en rouge). L'image (c) illustre la même région après avoir appliqué le filtrage spatio-temporel.

La figure 3.7 illustre l'évolution de l'énergie à chaque itération dans les cas suivants : sans mise à jour de l'image de fond (courbe en rouge), et avec l'image de fond le plus récent IFPR (courbe en vert). La mise à jour régulière du modèle du fond permet de minimiser l'énergie et d'atteindre la solution optimale plus rapidement. Ceci s'explique par le fait que la mise à jour de l'image du fond permet de tenir compte des changements continus des intensités des pixels au fil du temps. En effet, la composante estimée par ACI, qui correspond aux objets en mouvement, est beaucoup moins bruitée. Les courbes de la figure 3.7 correspondent à l'exemple de la figure 3.4. Sans la mise à jour du fond, la solution optimale est atteinte après 17 itérations tandis que l'optimatité est atteinte après 7 itérations dans le cas de la mise à jour du fond.

FIGURE 3.7 – Évolution de l'énergie en fonction du nombre d'itérations dans les deux cas suivants : sans mise à jour du fond et avec l'image du fond la plus récente (IFPR). Nous remarquons que l'utilisation de l'APPR permet d'atteindre plus rapidement l'optimalité. La convergence est assurée dans ce cas avec peu d'itérations.

3.6.4 Extraction des objets

Cette étape consiste à séparer les objets en mouvement du fond. Il s'agit de classer les pixels de l'image filtrée en deux classes : fond et objets. Les pixels qui correspondent au fond de l'image filtrée sont de couleur uniforme et homogène, alors que les pixels formant les objets en mouvement sont bien distingués. Cependant, il s'avère intéressant à ce stade de détecter les contours dans l'image filtrée, notée I_f. Nous avons choisi d'appliquer le détecteur de contour $Sobel$ (*Sob*) sur chaque composante couleur de l'image I_f. Le gradient spatial est appliqué dans les deux directions horizontale et verticale de chaque composante couleur. Pour une composante donnée, l'application de l'opérateur $Sobel$ sur les deux directions donne une image notée $E(I_f^{c^i})$ telle que c^i est la composante i d'un espace couleur donné. L'image des contours est donnée par l'équation 3.27 :

$$E(I_f^{c^i}) = Sob(I_f^{c^i})_{dx} \cup Sob(I_f^{c^i})_{dy} \tag{3.27}$$

La carte finale des contours, notée $E(I_f)$, est obtenue en superposant les cartes des contours obtenues avec chaque composante couleur de l'image I_f. Un pixel est considéré comme pixel de contour s'il est détecté comme pixel de contour dans chaque composante couleur. La carte $E(I_f)$ est donnée par l'équation 3.28 :

$$E(I_f) = E(I_f^{c_1}) \cap E(I_f^{c_2}) \cap E(I_f^{c_3}) \tag{3.28}$$

Une segmentation couleur de l'image filtrée par *Meanshift* est ensuite effectuée. En fixant un seuil, les pixels sont classés en fonction de leur couleur. Le seuil est considéré comme la moyenne de la couleur de la plus grande région de couleur homogène. Nous obtenons ainsi une image binaire, noté $S(I_f)$, tels que les objets en mouvement sont extraits. À la fin, les deux cartes, $E(I_f)$ et $S(I_f)$, sont fusionnées pour obtenir la carte finale des objets en mouvement.

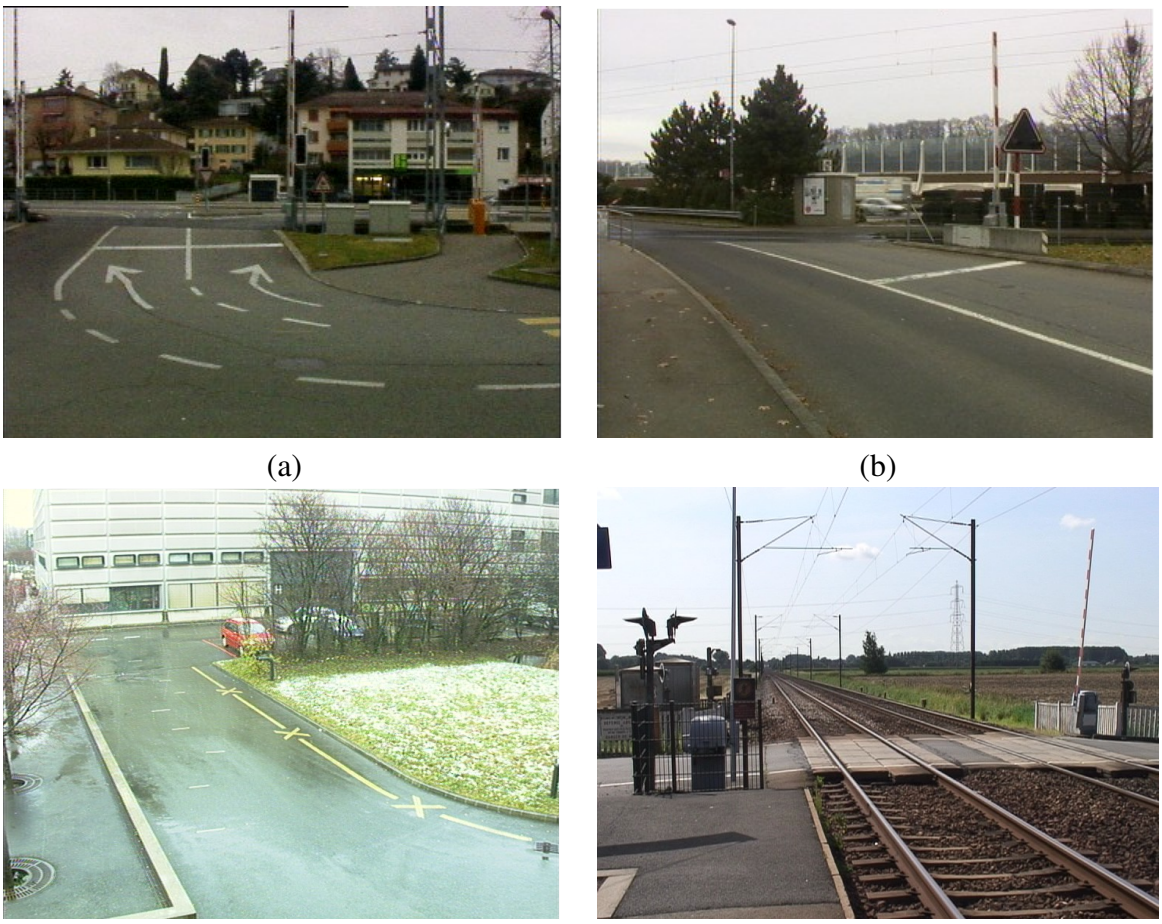
3.7 Résultats expérimentaux

3.7.1 Bases de données évaluées

La méthode proposée dans la section précédente est évaluée sur quatre bases d'images réelles couleur obtenues dans des environnements extérieurs. Les séquences d'images sont issues de caméras que nous avons installées à quatre endroits différents. Les bases intitulées "Pontet", "Chamberonne", et "EPFL" ont été acquises à Lausanne en Suisse et la quatrième base intitulée "PAN" a été acquise en France. Les bases "Pontet", "Chamberonne", et "PAN" ont bénéficié de système de vidéosurveillance dédiés à la sécurité aux passages à niveau. Ces trois dernières bases sont acquises durant des temps nuageux avec de la pluie intermittente. La base "EPFL" est acquise sur un parking surveillé par une caméra disposée à 7 mètres de hauteur par temps de neige. Les images des bases "Pontet", "Chamberonne", et "EPFL" ont une résolution de 384×288, alors que les images de la base "PAN" ont une résolution de 720×576. Les différentes bases d'images sont illustrées sur la figure 3.8 à l'aide d'une image.

3.7.2 Protocoles d'évaluations

Pour évaluer la qualité de l'extraction des objets du fond, nous avons extrait manuellement des objets de 1000 images des bases "Pontet" et "Chamberonne". Cette "vérité terrain" va permettre d'évaluer quantitativement les précisions de l'extraction à l'aide de deux indicateurs : *Rappel* et *Précision*. Nous avons aussi comparé notre méthode avec deux autres méthodes de la littérature :



(a) (b)

(c) (d)

FIGURE 3.8 – Les différentes bases d'images considérées dans l'évaluation (a) la base "Pontet" : un premier passage à niveau à Lausanne en Suisse (b) la base "Chamberonne" : un deuxième passage à niveau à Lausanne en Suisse (c) la base "EPFL" : un parking surveillé par une caméra de surveillance (d) la base "PAN" : un troisième passage à niveau en France.

mélange de gaussiennes "*MOG*" [MNV09] et "*Codebook*" [KCHD05]. Par ailleurs, nous avons aussi appliqué des invariants d'illuminant sur les images initiales : les invariants *maxRVB*, *Grey World (G.W.)*, *Normalisation Affine (A.N.)*, et *Equalisation d'Histogramme (H.E.)* ont été utilisés [LP06b]. Ce sont les invariants *G.W.* et *A.N.* qui ont fourni les meilleurs résultats. C'est pourquoi les résultats donnés par la suite ne tiennent compte que de ces invariants. Les invariants sont appliqués sur les images originales et l'évaluation consiste à analyser l'impact des différents invariants sur l'estimation des composantes indépendantes par ACI. Nous avons implémenté notre méthode ainsi que les deux autres méthodes considérées pour la comparaison sur Visual Studio 2008 avec les deux bibliothèques OpenCV version 2.0 pour la gestion des images (importation, traitement et enregistrement des résultats) et IT++ version 4.0.7 pour le calcul matriciel et vectoriel. Nous terminons l'évaluation par une comparaison des temps d'exécution des différentes étapes de notre

méthode, et une comparaison du temps d'exécution global entre les méthodes *MOG*, *Codebook* et la nôtre.

3.7.3 Discussion sur les invariants

Dans le cas où les données sont réelles, les composantes indépendantes obtenues par ACI ne sont généralement pas connues à l'avance. L'introduction d'un a priori dans le modèle lors de l'ACI permet une extraction efficace. Dans notre cas, le nombre de composantes à extraire des données sont deux : le fond et les objets. Comme expliqué dans la section 3.3, l'extraction des composantes indépendantes n'est pas toujours possible, ceci est le cas quand les composantes sont totalement ou partiellement corrélées. Même si la séparation des sources est possible, la qualité de la séparation dépend de la nature des données. Un des éléments clés pour la réussite du processus d'extraction des régions affectées par du mouvement est de rendre le modèle d'ACI moins sensible au bruit. Ceci revient à réduire le bruit au niveau des données et les rendre aussi invariantes que possible face aux changements continus du fond au fil du temps. Cependant, l'idée revient à passer d'un espace couleur vers un autre ou d'appliquer une transformation linéaire ou non linéaire sur un espace couleur donné.

Afin d'analyser l'impact des invariants sur le résultat de l'ACI, nous proposons d'appliquer une transformation linéaire sur les données. Les invariants permettent de rendre les données moins sensible aux changements minimes d'intensité entre deux images successives. Afin d'évaluer les performances de chaque invariant, nous proposons de mesurer l'effet de chaque invariant sur une matrice de données composée de deux images successives correspondant au fond. L'ACI estime deux composantes qui sont : le fond non bruité et du bruit. Dans le cas où les données sont non bruitées, les pixels du modèle de bruit estimé sont uniformes, et chaque pixel est de couleur donnée par le vecteur $\langle 127, 127, 127 \rangle$. En présence de bruit, la couleur de chaque pixel varie autour de cette valeur moyenne. Cette couleur sera prise comme une moyenne à partir de laquelle la valeur efficace, dite aussi valeur quadratique ou *"Root Mean Square RMS"*, est calculée. La valeur efficace permet de mesurer la variation de la couleur de chaque pixel autour de la couleur moyenne $\langle 127, 127, 127 \rangle$. Le tableau 3.1 montre les $RMS = \langle RMS_{c^1}, RMS_{c^2}, RMS_{c^3} \rangle$ calculés pour chaque invariant sur les quatres bases d'images.

L'espace couleur RVB est souvent remplacé par des invariants, utilisés pour réduire l'impact du bruit dû au capteur CCD. Nous décrivons ci-après les différents invariants utilusés dans notre expérimentation.

- $maxRVB = \langle r(maxRVB), v(maxRVB), b(maxRVB) \rangle$ tels que $r(maxRVB) = R/max_R$, $v(maxRVB) = V/max_V$, et $b(maxRVB) = B/max_B$. Les notations max_R, max_V, et max_B correspondent aux valeurs maximales des composantes rouge, vert, et du bleu, res-

	Pontet	**Chamberonne**	**EPFL-Parking**	**PAN**
RVB	$\langle 3.83, 3.01, 4.4 \rangle$	$\langle 4.30, 3.21, 6.04 \rangle$	$\langle 8.71, 5.81, 14.36 \rangle$	$\langle 2.45, 2.24, 2.36 \rangle$
$maxRVB$	$\langle 4.58, 3.76, 9.66 \rangle$	$\langle 4.30, 3.21, 6.04 \rangle$	$\langle 8.71, 5.81, 14.36 \rangle$	$\langle 2.45, 2.24, 2.36 \rangle$
$G.W.$	$\langle$**3.20, 2.57, 4.39**$\rangle$	$\langle$**3.44, 2.64, 5.47**$\rangle$	$\langle$**8.40, 5.07, 14.04**$\rangle$	$\langle$**2.06, 1.91, 2.01**$\rangle$
$A.N.$	$\langle$**3.62, 3.00, 4.07**$\rangle$	$\langle$**3.73, 2.80, 4.89**$\rangle$	$\langle$**8.24, 5.36, 12.13**$\rangle$	$\langle$**2.18, 1.83, 1.78**$\rangle$
$H.E.$	$\langle 16.23, 13.39, 18.43 \rangle$	$\langle 9.58, 7.98, 12.15 \rangle$	$\langle 13.82, 9.16, 19.22 \rangle$	$\langle 3.43, 3.36, 3.74 \rangle$

TABLE 3.1 – Evaluation du taux de bruits obtenu avec les invariants $maxRVB$, $Grey\ World$, $Affine\ Normalization$, et $Histogram\ \ Egualization$. Il s'agit des valeurs efficaces, RMS, obtenues sur chaque composante couleur de chaque invariant. Les RMS sont calculées sur les bases "Pontet", "Chamberonne", "EPFL-Parking", et "PAN".

pectivement, obtenues sur toute l'image.

- $Grey\ World\ \ G.W. = \langle r(G.W.), v(G.W.), b(G.W.) \rangle$ tels que $r(G.W.) = R/\mu_R, v(G.W.) = V/\mu_V$, et $b_(G.W.) = B/\mu_B$. Les notations μ_R, μ_V, et μ_B correspondent aux moyennes des composantes rouge, vert, et du bleu, respectivement, calculées sur toute l'image.

- $Normalisation\ \ Affine\ \ A.N. = \langle r(A.N.), v(A.N.), b(A.N.) \rangle$ tels que $r(A.N.) = |R - \mu_R| / \sigma_R$, $v(A.N.) = |V - \mu_V| / \sigma_V$, et $b(A.N.) = |B - \mu_B| / \sigma_B$. Les notations σ_R, σ_V, et σ_B correspondent aux écarts type des composantes rouge, vert, et du bleu, respectivement, calculées sur toute l'image.

- $Equalisation\ \ d'Histogramme\ \ H.E. = \langle r(H.E.), v(H.E.), b(H.E.) \rangle$ tels que $r(H.E.) = \sum_{c=0,R}^{255} H_c/\#pix$, $v(H.E.) = \sum_{c=0,V}^{255} H_c/\#pix$, et $b(H.E.) = \sum_{c=0,B}^{255} H_c/\#pix$. La notation H_c désigne le nombre de pixels de l'image ayant une valeur inférieure à c. $\#pix$ correspond au nombre total de pixels de l'image.

Le tableau 3.1 montre que les invariants $G.W.$ et $A.N.$ fournissent de meilleurs résultats par rapport aux autres invariants et l'espace couleur RVB seul. Nous avons retenu l'invariant $G.W.$ dans le reste de notre expérimentation.

3.7.4 Evaluation en termes de Rappel et Précision

Afin de mesurer les performances en termes de détection des différentes méthodes, nous avons procédé à une segmentation manuelle, des régions affectées par du mouvement, de mille images issues des bases "Pontet" et "Chamberonne". Le résultat de l'extraction manuelle des régions affectées par du mouvement est qualifié de vérité terrain VT. Cette segmentation manuelle va servir

de référence pour établir les taux de vraie détection et les taux de fausse détection. Nous proposons les deux mesures suivantes : *Rappel* et *Précision*. La mesure intitulée *Rappel* est un indicatif du taux des vraies détections. Le paramètre *Précision* mesure la précision de la détection. Ces derniers paramètres peuvent être exprimés en termes de vrai positif T_p, faux positif F_p, et faux négatif F_n. Les mesures *Rappel* et *Precision* sont données par les équations 3.29 et 3.30 respectivement :

$$Rappel = \frac{T_p}{T_p + F_n} \tag{3.29}$$

$$Precision = \frac{T_p}{T_p + F_p} \tag{3.30}$$

En se référant à la vérité terrain, le vrai positif T_p correspond au nombre de pixels correctement détectés par l'algorithme. Il s'agit des pixels affectés réellement par du mouvement, et correctement détectés comme des pixels affectés par du mouvement par l'algorithme utilisé. Le faux positif F_p ou fausse alarme correspond au nombre de pixels classés comme appartenant à un objet alors qu'ils appartiennent en réalité au fond. Le faux négatif F_n correspond aux pixels classés comme du fond alors qu'ils appartiennent réellement à un objet. La quantité $(T_p + F_n)$ correspond aux objets obtenus avec la vérité terrain, et la quantité $(T_p + F_p)$ correspond aux objets estimé par un algorithme donné. Le résultat de l'évaluation quantitative figure dans le tableau 3.2 :

	MOG	**Codebook**	**ACI+Filtrage**
Rappel	94.76%	93.49%	96.14%
Précision	95.87%	91.72%	97.34%

TABLE 3.2 – Évaluation quantitative selon les mesures *Rappel* et *Précision*.

Les figures 3.9, 3.10, 3.11, et 3.12 fournissent une comparaison visuelle des différentes méthodes de détection d'objets en mouvement issus de différentes bases d'images.

La figure 3.9 montre une image d'une scène extérieure acquise par une caméra sur un passage à niveau. L'image (b) illustre la présence d'une voiture se rapprochant de la zone de croisement, deux piétons traversant la route ainsi que deux autres voitures lointaines roulant en parallèle avec la voie ferrée. L'aspect colorimétrique d'une grande partie des objets est différent de celui du fond, sauf dans certains endroits où la couleur est très proche du fond. Les méthodes *MOG* (e) et *Codebook* (f) permettent d'extraire les objets du fond tout en manquant quelques zones, classées comme appartenant au fond. La méthode MOG dépend de plusieurs paramètres tels que le nombre

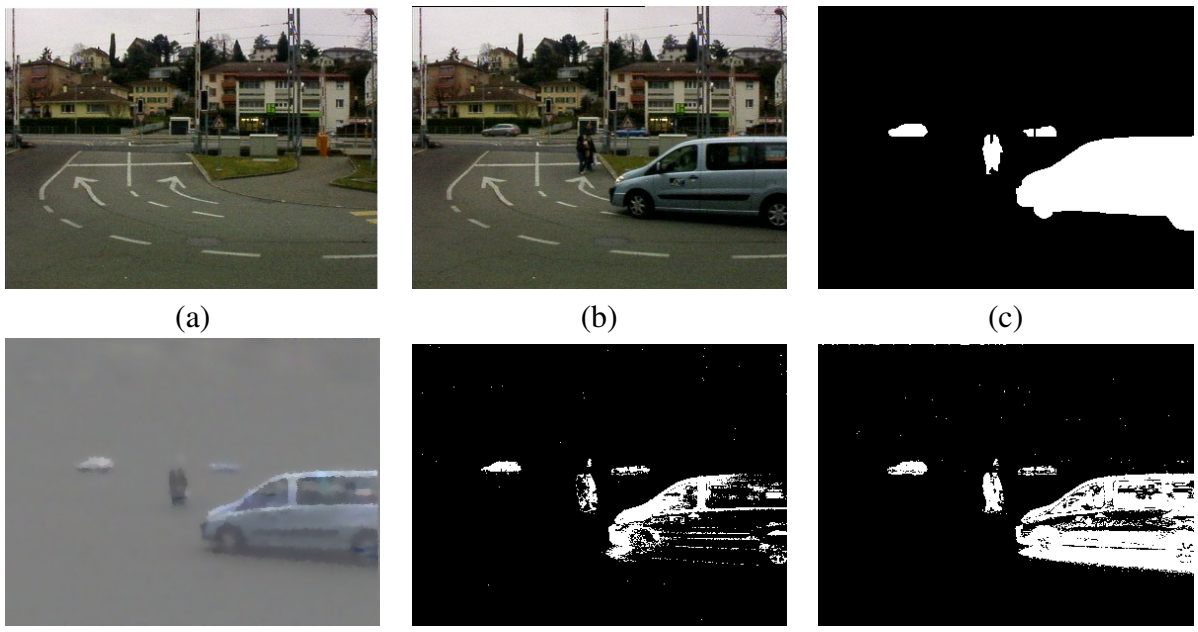
de gaussiennes. Un choix inapproprié de ce paramètre influe sur la décision de classification. La méthode Codebook échoue dans la classification de certains pixels puisque à cause du nombre d'images d'apprentissage important qu'exige cette méthode. Le résultat de classification obtenu avec notre méthode (d) montre que les limites des autres méthodes sont surmontées à cause des raisons suivantes : la sensibilité vis à vis du fond est traitée par le processus de filtrage qui rend lisse l'ensemble des pixels du fond, et le fait que l'ACI n'est pas très sensible aux données traitées.

La figure 3.11 montre une image d'une scène extérieure acquise pendant un temps de neige. Le fond contient des arbres dont les branches bougent continuellement. La chaussée est mouillée et un reflet apparait lors du passage d'un objet (voiture, piéton). L'image (b) contient deux voitures blanches, et une troisième partiellement occultée par les branches d'arbres. La méthode *MOG* (e) ne permet pas de bien détecter les objets mais une partie des branches est détectée comme des objets en mouvement. La méthde *Codebook* (f) donne un résultat de classification meilleur que celui obtenu avec la méthode *MOG*. L'image (d) illustre le résultat de classification fond/objets obtenu avec notre méthode. Nous remarquons que les pixels correspondant aux mouvement des branches d'arbres sont classés comme du fond. Les pixels formant les objets en mouvement sont bien identifiés et sont différents du fond.

3.7.5 Temps de traitement

Nous proposons dans cette section d'évaluer le temps d'exécution des différentes étapes de l'algorithme proposé ainsi qu'une comparaison du temps d'exécution global de différentes méthodes. Nous avons implémenté l'algorithme proposé ainsi que les méthodes évaluées sur la plateforme Visual Studio C++ 2008 en utilisant les bibliothèques OpenCV 2.0 permettant l'importation, le traitement, et l'enregistrement des images, et la bibliothèque IT++ 4.0.7 optimisée pour le calcul vectoriel et matriciel. Avec un ordinateur de bureau de processeur Intel 32-bit 3.1-GHz, les temps d'exécutions partiels de notre algorithme fugurent sur le tableau 3.3 pour la base "Pontet" dont les images sont de 384×288 pixels. En ne tenant pas compte du temps mis pour l'estimation de la matrice de séparation et pour l'estimation du modèle du bruit, le temps de traitement s'élève à 13 images par seconde.

Le temps de traitement de notre algorithme est comparé avec le temps de traitement des méthodes Mélange de Gaussiennes "*MOG*" et "*Codebook*". Dans le tableau 3.4, nous constatons que notre algorithme est plus rapide que les méthodes évaluées.



(a) (b) (c)

(d) (e) (f)

FIGURE 3.9 – Exemple de la base "Pontet" illustrant le résultat de détection de notre méthode face à d'autres méthodes de la littérature, *MOG* et *Codebook*. Les images (a) et (b) correspondent à l'image de l'arrière-plan et une image de la séquence contenant trois voitures et deux piétons, respectivement. L'image (c) correspond à la vérité terrain. Les objets en blanc sont segmentés manuellement. Les images (d), (e), et (f) montrent le résultat de la détection automatique obtenus avec notre méthode, la méthode *MOG*, et la méthode *Codebook* respectivement.

Les différentes étapes de l'algorithme	**Temps d'exécution**
Estimation de la matrice de séparation	466.324 *ms*
Etape d'apprentissage	7.424 *ms*
Approximation de l'objet	39.735 *ms*
Propagation de croyance	41.528 *ms*

TABLE 3.3 – Le temps d'exécution obtenu sur la base "Pontet".

Algorithmes	algorithme proposé	Codebook	MOG
Temps de traitement	81.263 *ms*	118.402 *ms*	286.588 *ms*

TABLE 3.4 – Le temps de traitement obtenu sur la base "Pontet" pour les différentes méthodes testées.

3.8 Conclusion

Nous avons proposé dans ce chapitre une nouvelle méthode permettant une extraction automatique des régions affectées par du mouvement dans une séquence d'images. Après une revue des techniques existantes, nous nous sommes basés sur le principe de séparation des signaux indépendants non connus a priori, à partir d'un mélange de ces signaux. Cette idée servait principalement

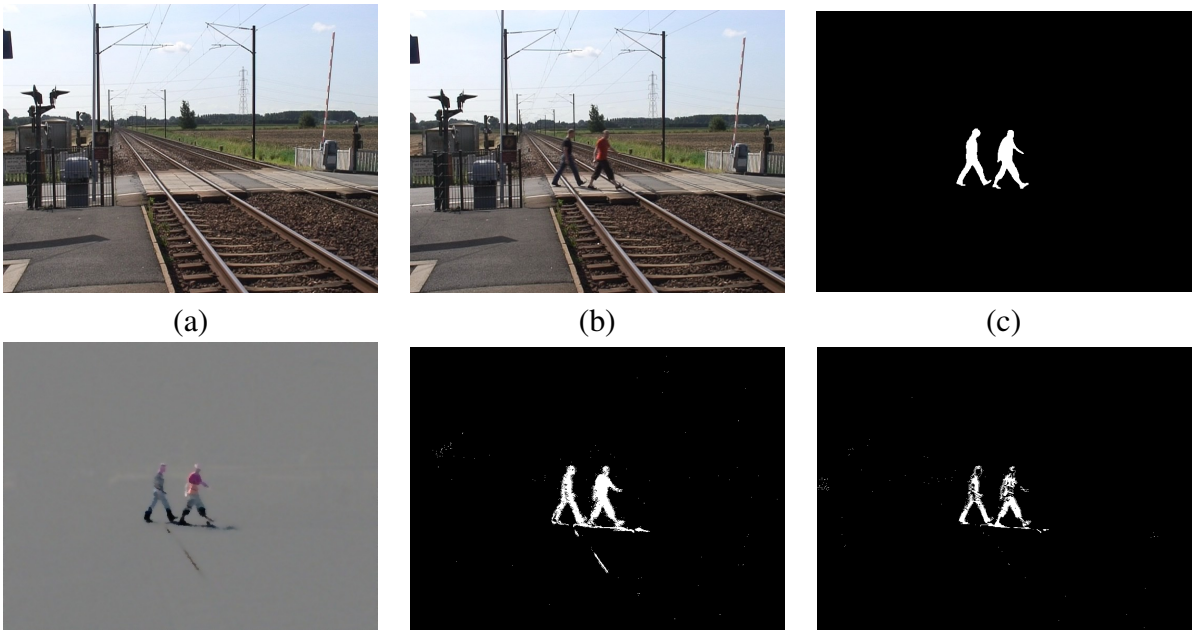


FIGURE 3.10 – Exemple de la base "PAN" illustrant le résultat de détection de notre méthode face à d'autres méthodes de la littérature, *MOG* et *Codebook*. Les images (a) et (b) correspondent à l'image de l'arrière-plan et une image de la séquence contenant deux piétons traversant un passage à niveau, respectivement. L'image (c) correspond à la vérité terrain. Les objets en blanc sont segmentés manuellement. Les images (d), (e), et (f) montrent le résultat de la détection automatique obtenue avec notre méthode, la méthode *MOG*, et la méthode *Codebook* respectivement.

dans le domaine du traitement du son. Adaptées au problème d'estimation du mouvement, nous n'avons recensé que deux travaux relativement récents se basant sur l'analyse en composantes indépendantes pour l'extraction des régions affectées par du mouvement à partir d'une séquence d'images. Ces deux travaux ne traitaient que le cas d'images en niveau de gris. Notre contribution consiste en l'introduction de l'information colorimétrique dans le modèle d'ACI et le développement d'une méthode de filtrage spatio-temporel basée sur la propagation de croyance.

L'algorithme proposé est conçu pour être beaucoup moins sensible aux variations continues d'illuminations par rapport aux méthodes existantes. Nous avons choisi d'évaluer notre algorithme sur des séquences d'images prises dans des environnements extérieurs difficiles. Dans ce genre d'environnement, nous nous sommes confrontés à plusieurs problèmes tels que les conditions d'éclairage qui sont difficiles à contrôler, la présence d'ombres des objets sur le fond, le mouvement continu des branches d'arbres, et des conditions météorologiques différentes. Ces contraintes sont bien gérées par l'algorithme proposé : d'un côté, l'ACI est moins sensible au bruit qui correspond aux changements continus d'une partie ou de l'ensemble de l'image à cause des variations d'illumination et aux faux mouvements tels que le mouvement des branches d'arbres. Dans le cas d'un temps neigeux, la neige ne doit pas être détectée comme appartenant à des régions affectées

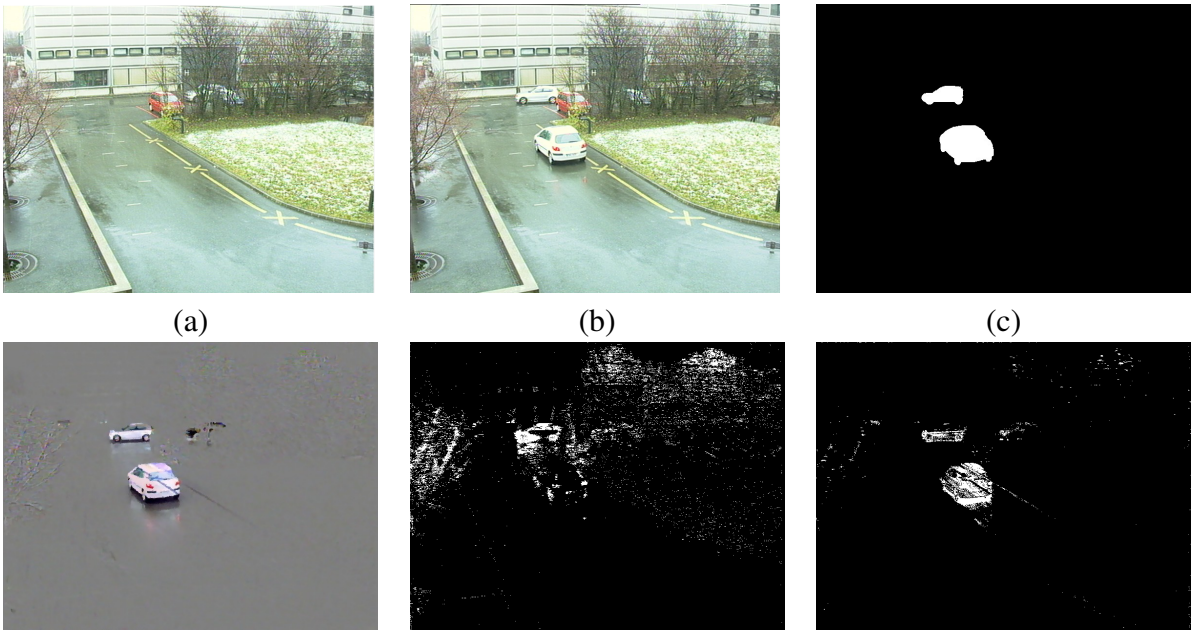


FIGURE 3.11 – Exemple de la base "EPFL-Parking" illustrant le résultat de détection de notre méthode face à d'autres méthodes de la littérature, *MOG* et *Codebook*. Les images (a) et (b) correspondent à l'image de l'arrière-plan et une image de la séquence contenant trois voitures, respectivement. L'image (c) correspond à la vérité terrain. Les objets en blanc sont segmentés manuellement. Les images (d), (e), et (f) montrent le résultat de la détection automatique obtenue avec notre méthode, la méthode *MOG*, et la méthode *Codebook* respectivement.

par du mouvement. Les méthodes classiques basées sur une modélisation de l'arrière-plan ne permettent pas de gérer ce genre de situation. Avec notre méthode, la neige n'a pas d'effet important sur le processus de détection puisqu'elle est facilement éliminée lors de l'étape de filtrage par propagation de croyance. Par ailleurs, l'ombre des objets correspond à des régions transparentes dont la couleur et la teinte sont différentes de ceux du fond, mais de texture semblable à celle du fond.

L'introduction du mouvement en tant que contrainte permet de réduire les ambiguïtés d'appariement. Supposons deux séquences d'images stéréoscopiques. Le correspondant, dans l'image droite, d'un pixel à apparier de l'image gauche et affecté par du mouvement, est probablement lui aussi affecté par du mouvement. Dans le cas des applications de vidéosurveillance, les objets d'intérêt, en l'occurrence les objets mobiles, ont plus d'importance que le fond qui correspond à l'environnement où ces objets interagissent. Dans le chapitre suivant, nous proposons d'appliquer l'algorithme de localisation tridimensionnelle sur des objets d'intérêt extraits automatiquement à partir de séquences d'images réelles. Il s'agit d'une application de vidéosurveillance dont le but est d'accroître la sécurité aux passages à niveau. Le chapitre suivant fait l'objet d'une étude détaillée sur l'apport de la vision artificielle, en l'occurrence de la vision stéréoscopique, sur l'accroissement de la sécurité aux passages à niveau. Nous détaillons dans le chapitre suivant jusqu'à quel point les

algorithmes que nous proposons peuvent répondre, en termes de détection et de localisation, aux exigences sécuritaires de l'application concernée.

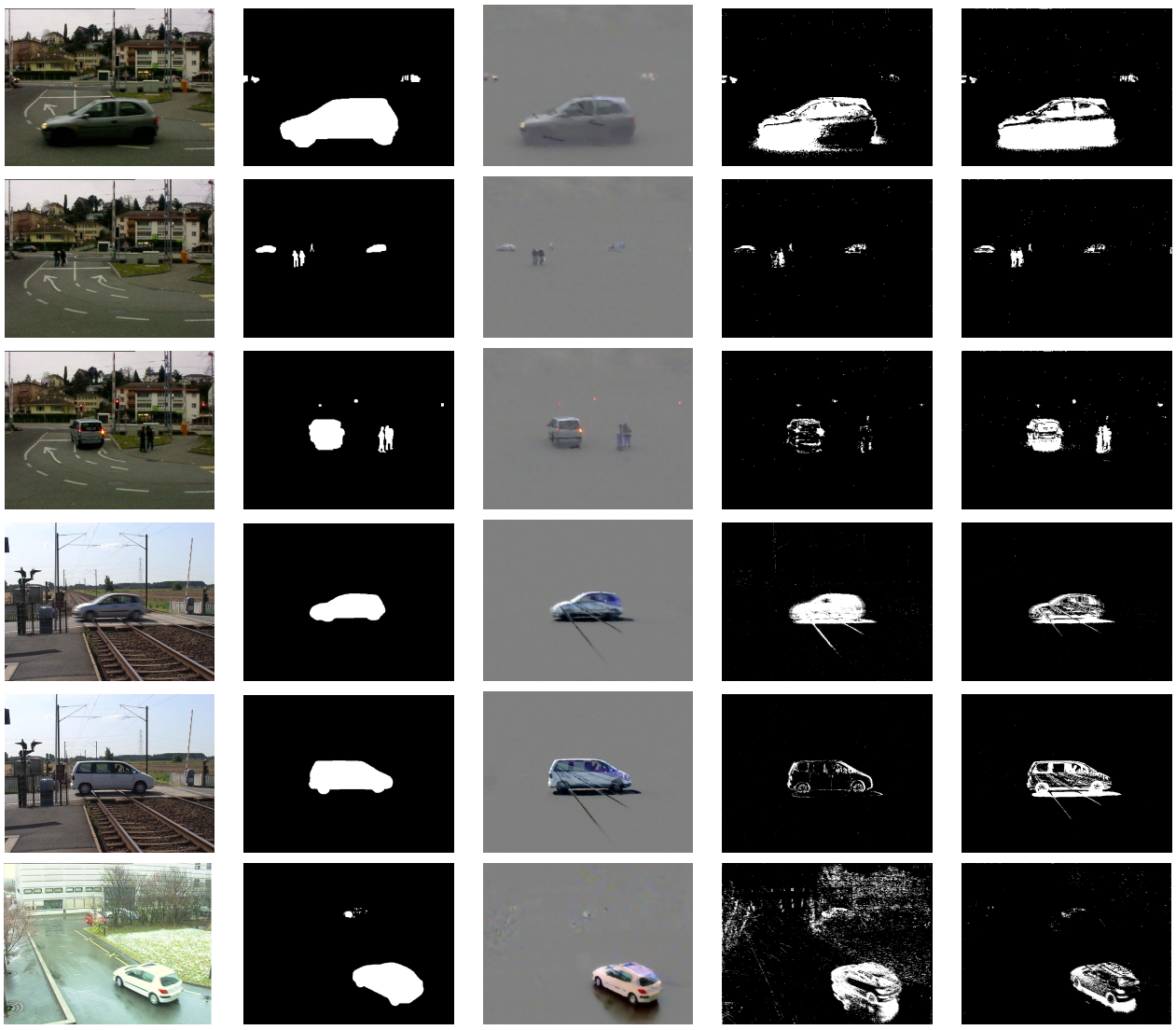


FIGURE 3.12 – Exemples issus des différentes bases d'images illustrant le résultat de détection de notre méthode face à d'autres méthodes de la littérature, *MOG* et *Codebook*. La première colonne correspond aux images contenant des objets stationnaires ou en mouvement. La deuxième colonne correspond aux vérités terrain. Les colonnes (c), (d), et (e) montrent le résultat de détection obtenu avec notre méthode, la méthode *MOG*, et la méthode *Codebook* respectivement.

Chapitre 4

Localisation tridimensionnelle d'obstacles aux passages à niveau

4.1 Introduction

La sécurité des personnes et des équipements est un élément capital dans le domaine des transports routiers et ferroviaires. De nos jours, le train représente le moyen de transport terrestre le plus sûr. Les différents travaux d'automatisation du fonctionnement des réseaux de transport ont montré une autonomie significative des moyens de transport ferroviaires dans le sens de la gestion du trafic et de l'anticipation des problèmes de collision train/train et train/obstacle. La portion commune entre l'infrastructure routière et ferroviaire est le lieu de croisement des trafics routier et ferroviaire : il s'agit des Passages à Niveau (PN). Afin d'accroître la sécurité des usagers des PN, une étude approfondie doit être menée sur le sujet. Nous devons comprendre les différents types et les modes de fonctionnement des PN. Cette étape permet d'évaluer le niveau de risque que peut représenter un PN particulier, ainsi que le niveau de sécurité qui lui est associé. Afin de réduire les risques d'accidents et d'accroître le niveau de sécurité, une analyse des comportements des différents acteurs qui interviennent dans le fonctionnement des PN doit être menée en se basant sur l'analyse des causes des différents scénarios d'accidents répertoriés, et les résultats des enquêtes qui ont été effectuées dans le cadre de plusieurs travaux à l'échelle internationale, tels que au Canada, aux Etat Unis, au Japon, et aux états membres de l'Union Européenne. C'est dans ce cadre que le projet européen intitulé SELCAT [Sel] a vu le jour, dont le principal objectif est d'étudier les différentes façons d'harmoniser le fonctionnement des passages à niveau à l'échelle européenne et de minimiser le nombre d'accidents de 50% à court terme. Les recommandations en termes de solutions technologiques et non technologiques proposées par le projet SELCAT sont développées dans le cadre du projet national intitulé PANsafer [Pan] qui a comme objectif d'étudier l'apport des nouvelles technologies dans l'anticipation des risques et l'accroîssement de la sécurité aux PN. Le présent chapitre est organisé en deux parties :

– La première partie permet de recenser les différentes architectures possibles des PN et une Analyse Préliminaire des risques. Cette étape est primordiale puisqu'elle permet de mieux comprendre le mode de fonctionnement des PN ainsi que les intéractions possibles entre acteur/acteur et acteur/infrastructure.

– La deuxième partie détaille l'apport de la vision artificielle dans l'anticipation des situations dangereuses. La compréhension des exigeances en termes de sécurité et de fiabilité influence le choix de l'architecture ainsi que les techniques utilisées. Nous présentons dans cette partie le dispositif expérimental et les résultats obtenus en termes de détection et localisation tridimentionnelle en situation réelle.

4.2 Analyse fonctionnelle

4.2.1 Différents types de passages à niveau

En Europe, plusieurs classifications de PN existent à ce jour. La classification dépend des équipements installés sur chacun d'entre eux. La sécurité et les risques associés à un PN varient selon le type de PN. Globalement, les différentes catégories sont les suivantes :

– ***Les passages à niveau actifs*** sont equipés d'un système de signalisation automatique ou manuelle, dont l'activation des dispositifs de sécurité est faite par un membre de l'équipe du train ou par une autre personne chargée de controler la sécurité d'un tel passage a niveau lors du rapprochement du train. Ces systèmes de signalisation indiquent le rapprochement des trains (abaissement des barrières, allumage des feux rouges, déclenchement des alertes sonores).

– ***Les passages à niveau passifs*** sont caractérités par l'absence d'un système de signalisation automatique. C'est aux piétons et aux conducteurs de véhicules d'évaluer le niveau de risque associé lors d'une traversée du PN. Les passages piétons, dont l'accès est restreint aux piétons, est un exemple de passages à niveau passif.

Les passages à niveau actifs peuvent aussi être répartis selon le mode de fonctionnement des équipements. [MMG06] [SK04] proposent de classer les PN comme suit :

– ***Les passages à niveau automatiques*** : nous citons AHB (Automatic Half Barrier), ABCL (Automatic Barrier Crossing Locally monitored), AOCL (Automatic Open Crossing Locally monitored), AOCR (Automatic Open Crossing Remotely monitored), et UWC+MWL (User-

Worked Crossing with Miniature Warning Light).

– ***Les passages à niveau manuels*** : MG (Manual Gate), MCB (Manually Controlled Barrier), et MCB+CCTV (Manually Controlled Barrier protected by Closed Circuit TeleVision).

4.2.2 Évaluation de la sécurité aux passages à niveau

4.2.2.1 Pannes liées au système

Les pannes liées au système et aux équipements sont rares, et n'interviennent pas souvent dans les accidents aux PN car à la suite d'une panne, le système de sécurité s'active et se met en mode sécurisé : abaissement automatique des barrières et allumage continu des feux rouges. Dans le cas où le système ne se met pas en mode sécurisé, le premier train qui passe par le PN signale le défaut de fonctionnement du système afin de le réparer rapidement.

4.2.2.2 Erreurs humaines

Des statistiques d'accidents [Nel02] montrent que 44% des usagers des PN passifs ont une mauvaise perception de l'environnement et une mauvaise évaluation des risques associés aux passages à niveau, 35% des accidents sont liés à la vitesse des trains, et 2% sont liés à un niveau de trafic ferroviaire élevé. Selon la même étude, il y avait en moyenne 14 morts par an dans des accidents sur des passages a niveau passifs avec un taux moyen annuel croissant de 20% pour la periode entre 1983 et 1994 en Grande Bretagne. La majorité des victimes sont des piétons, des conducteurs et des passagers de véhicules. Cette dernière étude a recensé les facteurs de risques associés aux passages à niveau passifs. La sécurité liée à ce type de passage dépend principalement du comportement des usagers de la route. L'absence d'un système automatique de signalisation nécessite plus de vigilance de la part des acteurs des passages à niveau. Les facteurs qui influencent la sécurité des PN sont les suivants :

– Le comportement des usagers de la route vis-a-vis du passage a niveau (piétons, conducteur et occupants d'un véhicule,...).

– Le niveau de perception du risque avant et au moment de la traversée du passage à niveau.

– Le temps mis par les usagers de la route pour traverser la zone de danger.

– L'état de la signalisation sonore des locomotives des trains en approche.

– La fréquence de circulation des trains et la fréquence d'utilisation des passages par les usagers de la route.

– La vitesse de circulation des trains.

D'après [Gri04], les erreurs humaines interviennent dans près de 99% des cas d'accidents dont 93% sont causés par les usagers de la route. Les causes d'accidents aux passages à niveau peuvent être regroupées en trois catégories selon l'étude de "*Rail Safety and Standards Board*" sur le comportement des usagers de la route [SK04] :

– ***Bonne Utilisation*** : cette classe représente les acteurs routiers qui traversent entièrement la voie ferrée en respectant les indications et les panneaux de signalisation (automatique ou non automatique) avec un événement imprévu qui conduit à un accident. Il s'agit d'un dysfonctionnement du système de signalisation.

– ***Erreur d'utilisation*** : elle regroupe les usagers de la route qui traversent la voie ferrée alors qu'un train est imminent, avec une mauvaise estimation du temps d'approche du train et une mauvaise évaluation des signaux d'alertes. Nous citons à titre d'exemple la traversée d'un passage à niveau passif à plus d'une voie, ou la traversée non prudente d'un passage à niveau non gardé.

– ***Violation d'utilisation*** : elle représente les usagers de la route qui traversent la voie ferrée avec un train imminent et un non respect des panneaux et dispositifs d'alertes de rapprochement du train. Nous citons le cas d'un franchissement des barrières baissées, ou le non respect des feux rouges d'interdiction du passage, ou encore le non respect d'un panneau de stop.

Le degré de danger lors d'une violation des dispositifs de sécurité est proportionnel à la durée qui sépare le déclenchement des alertes d'interdiction de passage, et au temps mis par l'usager de la route pour franchir le passage. Nous détaillons ci-après trois types de comportements à risque.

4.3 Analyse Préliminaire des Risques (APR)

Toute solution technologique ou non pour améliorer le niveau de sécurité aux PN doit prendre en compte les normes de sécurité imposées par les experts du domaine ferroviaire. Pour cela, avant d'étudier la faisabilité et les performances d'une solution technologique, nous sommes amenés à définir préalablement les normes et les niveaux de sécurité exigés, relatifs à notre système d'infor-

mation.

Selon [DLQV05], un incident peut être défini comme "*un évènement imprévu pendant le fonctionnement d'un système ou le déroulement d'une activité dont les conséquences sont un dysfonctionnement du système, une perturbation de l'activité ou l'occurrence de dégâts matériels légers*". Selon le même auteur, un risque représente un ensemble de "*caractéristiques d'un événement, définies conjointement par sa vraisemblance d'occurrence (définie en termes de fréquence d'apparition ou de probabilité d'occurrence pendant une période de temps ou un nombre d'opérations) et la gravité de ses conséquences*". L'identification et l'analyse des risques sont en général indissociables. L'identification des risques peut correspondre, suivant le contexte à tout ou partie de :

- L'identification des dangers,
- L'identification des situations dangereuses ou à risque,
- L'identification des situations accidentelles,
- L'identification des scénarios d'accident.

La recherche doit porter a priori sur l'ensemble des évènements, qui peuvent mettre en contact les dangers et les éléments potentiellement vulnérables inclus dans le périmètre de l'analyse des risques. Les résultats de l'identification et de l'analyse des risques sont :

- La liste des dangers,
- La liste des évènements indésirables et des facteurs de risques (évènement amorce, situations dangereuses),
- La liste des scénarios d'évènements indésirables et leurs conséquences.

Un scénario d'accident est défini selon [DLQV05] par "*une Suite et/ou combinaison de circonstances favorisant l'apparition d'évènements aboutissant à un accident*". Un scénario d'accident est une suite d'événements dont l'origine première est la présence d'un danger qui, suite à la survenance d'un événement contact, met le système en situation dangereuse qui peut elle-même, suite à la survenance d'un événement amorce, conduire à une situation accidentelle ou un accident. Un scénario d'accident est caractérisé par la gravité des conséquences de l'accident auquel il aboutit. Nous présentons ci-après les spécificités de chaque étape et l'état du système conduisant à une situation accidentelle ou à un accident.

- ***Danger*** : il peut être considéré comme "*un potentiel de dommages ou de préjudices portant atteinte aux personnes, aux biens ou à l'environnement*" [DLQV05]. Les usagers des PN sont constamment exposés aux dangers au moment de leur traversée de la zone de danger, à cause de la circulation continue des trains. La violation, volontaire ou par erreur, des dispositifs de sécurité par les usagers du passage à niveau affecte la sécurité et la fiabilité du système. L'état "en danger" d'un système se déclenche lors de la présence d'un flux routier à proximité ou sur la zone de danger. Le train étant l'entité prioritaire, nous considérons que le fonctionnement des trains est un fonctionnement habituel, et la présence d'autres entités (piéton, voiture, camion, objet) dans sa zone de circulation représente un évènement indésirable. Une situation de danger se définit par l'approche d'un train de la zone de croisement.

- ***Évènement Contact*** : il s'agit d'un "*évènement dont la survenance, en présence de danger, met le système en situation dangereuse*" [DLQV05]. Cet évènement peut être l'aboutissement d'un scénario issu d'une défaillance matérielle du système, ou d'erreurs humaines. La proximité d'un usager routier au lieu de croisement voie-ferrée/route, couplé avec un dysfonctionnement d'un des dispositifs de sécurité, représente un évènement contact qui met le système en un état critique. Cet évènement indésirable met le système en une situation dangereuse.

- ***Situation dangereuse*** : elle est définie par [DLQV05] comme "*la situation dans laquelle les éléments du système sont exposés à un danger imminent. La Situation Dangereuse (SD) est un état du système en présence d'un danger imminent. Autrement dit, elle résulte de la conjonction d'un danger (D) et d'un évènement contact (EC)*". Cet évènement peut être le résultat d'un ensemble de circonstances. Les entités exposées au danger sont les personnes, les biens, et l'environnement dans lequel le système évolue. Certains paramètres, comme la durée de l'exposition de l'évènement contact, peuvent influer sur la dangerosité d'une telle situation. Une situation dangereuse peut être définie par le fait qu'un train est imminent, un ou plusieurs usagers sont à proximité de la zone de danger. Cette situation peut être plus grave en présence d'un évènement amorce, permettant de faire évoluer le système vers une situation accidentelle.

- ***Évènements amorce*** : il s'agit d'un "*évènement dont l'occurrence peut entrainer une situation accidentelle ou un accident quand le système est en situation dangereuse. L'évènement amorce, appelé aussi évènement déclencheur est l'évènement qui peut initier un accident, lorsque les autres conditions requises sont également réunies*" [DLQV05]. Cet évènement peut être l'aboutissement d'un scénario issu de la combinaison de défaillances du matériel, et d'erreurs humaines. Nous pouvons definir un évènement amorce, dont son occurrence en-

traine une situation accidentelle, par la présence d'un obstacle ou d'un usager routier sur la zone de danger sachant qu'un train est imminent.

– ***Situation accidentelle (SA) ou Accident (A)*** : la situation accidentelle peut etre considérée comme "*une situation dangereuse dans laquelle des éléments vulnérables du système ont été mis en présence d'un danger les affectant directement*" [DLQV05]. La situation accidentelle est généralement assimilée à l'accident lui-même. Le passage en situation accidentelle d'un système initialement en situation dangereuse résulte de la survenance d'un évènement amorce qui accroît la dangerosité des éléments vulnérables du système. Un accident est défini, selon le même auteur par "*un évènement redouté, soudain, involontaire et imprévu dont les conséquences sont la mort, l'invalidité ou les blessures graves aux personnes, l'atteinte grave à l'environnement ou la destruction partielle ou totale du système. L'accident est l'évènement résultant d'un enchainement d'évènements élémentaires ou de scénarios de situations dangereuses*". Une situation accidentelle est définie par la présence simultanée d'un train et, un ou plusieurs obstacles ou/et des usagers.

Les conséquences d'un tel accident touchent généralement l'être humain et les équipements. Un accident dans un passage à niveau peut entrainer :

– La mort et/ou la blessure grave des utilisateurs du passage à niveau (surtout les usagers routiers).

– Des dégats sur les installations du PN.

– Des dégats sur les moyens de transports routiers (voitures, camions ou autres, impliqués dans l'accident).

– Le déraillement du train en cas d'accident grave.

4.4 Détection et localisation d'obstacles aux passages à niveau

Nous avons détaillé dans la première partie du présent chapitre les sources éventuelles d'accidents liés aux passages à niveau. L'introduction de la vision artificielle en tant que source complémentaire d'informations s'avère un choix réfléchi. Un tel système de vision doit être capable de superviser la zone de danger d'un passage à niveau, de détecter, de localiser, et d'analyser le comportement des usagers. Les éléments ainsi fournis par le système de vision seront couplés avec d'autres informations relatives à l'infrastructure tels que l'état des barrières et des feux de signali-

sation, et à la vitesse et la distance du train le plus proche du passage à niveau. L'objectif principal est de fournir des informations aussi sûres que possible sur l'environnement surveillé. Les informations fournies concernent la position 3D précise d'éventuels obstacles autour d'un passage à niveau.

4.4.1 Système de vision proposé

Contrairement à d'autres techniques telles que le laser ou le radar, la vision artificielle fournie beaucoup plus d'informations sur l'environnement surveillé. Compte tenu des exigeances en termes de sécurité, nous proposons un système de vision doté des deux fonctionnalitées suivantes : la détection d'obstacles, et leur localisation tridimentionnelle. Les usagers routiers et les différents objets pouvant affecter la fiabilité et le fonctionnement optimal du PN, sont considérés comme des obstacles. Tout objet fixe, en mouvement ou stationnaire, est considéré comme un obstacle potentiel puisqu'il interagit avec le passage à niveau. Nous partons du principe qu'une situation dangereuse causée par un piéton ou par un grand camion a le même niveau de criticité. La figure 4.1 illustre sur trois images le dispositif expérimental mis en place lors de l'évaluation :

FIGURE 4.1 – Vue générale du système de vision proposé.

Le système de vision proposé est composé de deux caméras stéréoscopiques couleur dont le champ de vision est orienté vers la zone de danger. Cette zone représente la région de croisement définie par l'intersection entre une ou plusieurs voies ferrées, et une ou plusieurs voies routières. Le système proposé doit être capable de détecter n'importe quels types d'objets, et de pouvoir les localiser dans un espace 3D. Les caméras sont installées sur un support dédié permettant de les aligner et de les orienter dans la même direction. Compte tenu des distances minimale et maximale des obstacles à surveiller, et à la position des caméras par rapport au centre de la zone de danger, la distance qui sépare les caméras est fixée à 50 cm. Le support portant les caméras est tenu par un trépied permettant d'atteindre une hauteur de 150 cm environ. Nous avons utilisé deux caméras *Sony DXC-390/390P 3-CCD* et un objectif *Cinegon 3 CCD Lens 5.3mm FL* (figure 4.2).

FIGURE 4.2 – Modèle de caméra et objectif utilisé.

4.4.2 Démarches proposées

Le système de vision proposé assure les deux principales fonctionnalitées qui sont la détection d'obstacles et leur localisation tridimensionnelle. Le premier module permet l'extraction des régions d'intérêt à partir d'une séquence d'images issue des caméras gauche et droite. Les régions d'intérêt correspondent aux obstacles qui peuvent être en mouvement ou stationnaires. Les obstacles en mouvement correspondent aux objets qui bougent et interagissent avec leur environnement, en l'occurrence des piétons, des véhicules. Les obstacles fixes correspondent aux objets laissés volontairement ou involontairement sur le PN. La localisation 3D permet d'estimer la position relative des différents objets détectés. Cette information est cruciale puisque le niveau de dangerosité dépend du comportement spatio-temporel des obstacles. À titre d'exemple, une voiture temporairement arrêtée derrière les barrière en dehors de la zone de croisement ne représente pas le même niveau de risque qu'une voiture arrêtée dans la zone de danger. La figure 4.3 présente le symphonique général des algorithmes proposés.

Selon le diagramme de la figure 4.3, une étape de pré-traitement est tout d'abord appliquée sur les images issues de chaque caméra. Un premier pré-traitement consiste à corriger les aberrations causées par le système optique des caméras. Ce traitement consiste à rectifier les courbures des lignes, bien marquées sur les bords des images. Les images ainsi rectifiées subissent un pré-traitement supplémentaire afin de faciliter l'algorithme d'appariement stéréoscopique. Il s'agit du calibrage du système stéréoscopique afin de rendre les lignes épipolaires (utilisation du modèle du sténopé) confondues. Cette modélisation permet un gain significatif en terme de temps de traitement. Nous avons utilisé la méthode de calibrage de [Tsa86] dont le code source est disponible dans la bibliothèque OpenCV 2.0. Nous détaillons ci-après les deux principales fonctionnalitées de l'algorithme de détection et de localisation 3D proposées.

4.4.2.1 Détection 2D des régions d'intérêts

La détection des régions d'intérêt consiste à identifier les zones affectées par du mouvement à partir d'une séquence d'images. Étant donné que les caméras sont fixes, nous nous sommes

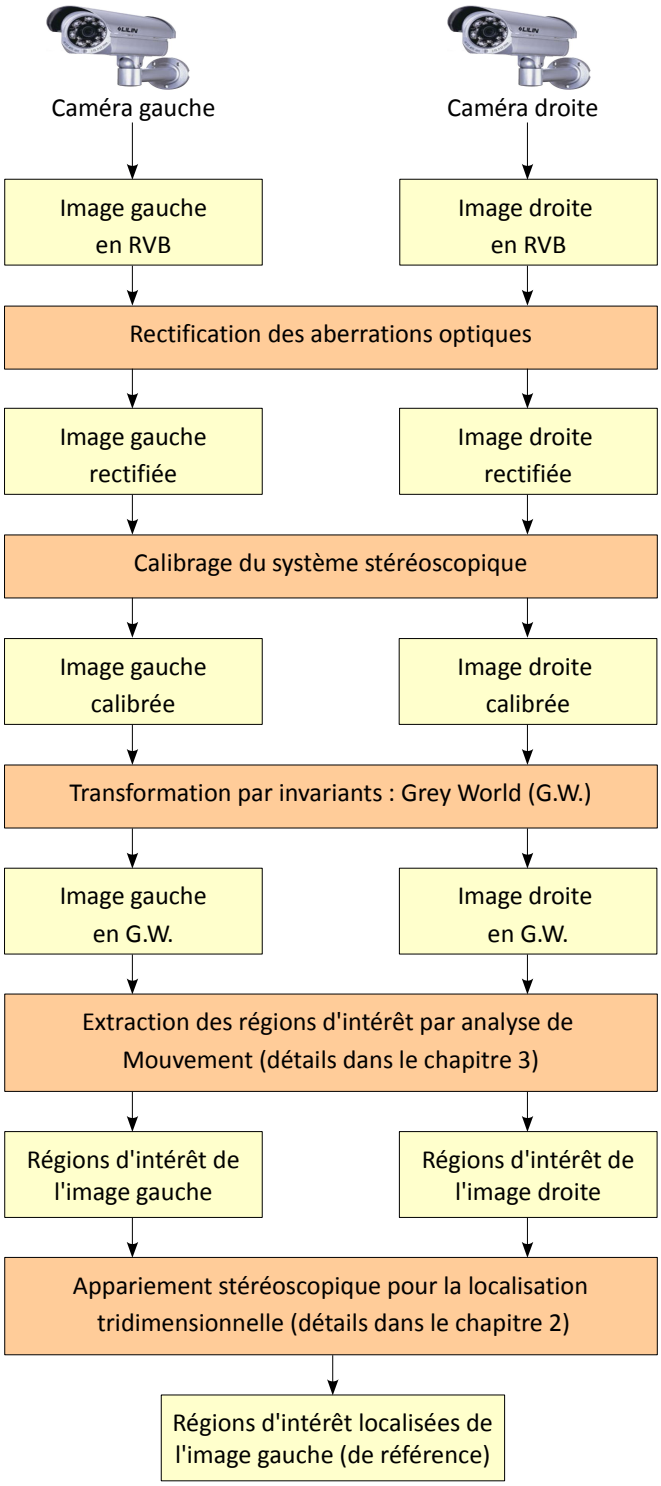


FIGURE 4.3 – Principaux modules de traitements permettant la détection et la localisation 3D d'obstacles.

orientés vers les techniques de soustraction de fond basées sur le principe d'image de référence. Les régions affectées par du mouvement sont extraites à partir de l'Analyse en Composante Indépendante adaptée au problème d'estimation du mouvement à partir d'une séquence d'images. Un tel modèle identifie les pixels dont l'intensité ou la couleur diffère entre les deux images. Il existe plusieurs méthodes assurant cette tâche. Ces méthodes diffèrent à la fois dans la façon où l'image de référence est mise à jour, et dans la prise en compte ou non des conditions dégradées de l'environnement observé. Nous avons proposé dans le chapitre 3 une nouvelle méthode d'extractions d'objets en mouvement à partir d'une séquence d'images d'un environnement extérieur. Cette méthode est basée sur l'Analyse en Composantes Indépendantes qui permet d'estimer les composantes indépendantes, qui sont le fond et les régions d'intérêt, à partir d'un mélange de données, qui sont l'image de fond et une image de la séquence. Cette méthode est couplée avec un module de filtrage basé sur le principe de propagation de croyance spatio-temporelle que nous avons développé. Nous avons montré dans le chapitre précédent que cette méthode donne de meilleurs résultats sur des images lorsque l'espace couleur est transformé selon l'invariant *Grey World*.

L'apport de la méthode d'extraction des régions d'intérêt a été justifié en la comparant avec deux autres techniques connues dans la littérature : le mélange de gaussiennes, et la méthode Codebook. Les régions d'intérêt peuvent correspondre aux piétons, aux objets posés par des personnes, à des véhicules de différentes tailles. L'extraction est appliquée en parallèle sur les deux images issues des deux caméras. Nous obtenons ainsi deux images représentant les régions affectées par du mouvement de la scène observée selon deux points de vue différents. Ces deux images seront prises comme des entrées dans le module de localisation tridimensionnelle.

4.4.2.2 Localisation 3D des régions d'intérêt

Un des problèmes de la vision monoculaire est que les vraies distances 3D de deux objets, en occultation, ne peuvent pas être correctement estimées. L'introduction de la stéréoscopie permet d'estimer la distance de chaque objet même en cas d'occultation. La deuxième étape consiste alors à localiser les objets précédemment détectés dans un espace 3D. L'importance de cette étape se reflète dans la mesure où des objets partiellement occultés sont facilement distingués. En effet, un algorithme d'appariement stéréoscopique est appliqué sur chaque pixel affecté par du mouvement. Un tel algorithme doit pouvoir gérer le bruit dû à la qualité dégradée des images acquises, et les problèmes classiques de la vision stéréoscopique tels que les occultations, les régions de textures répététives, et les régions de couleur homogène. L'algorithme utilisé est celui présenté et détaillé dans le chapitre 2. Nous avons introduit ainsi la contrainte de mouvement en se basant sur l'hypothèse suivante : l'homologue dans l'image droite, d'un pixel à apparier affecté par du mouvement dans l'image gauche, est aussi affecté par du mouvement. Ceci réduit davantage le nombre de candidats possibles.

La première étape consiste à estimer une première carte de disparités par l'appliquation d'une méthode locale d'appariement. Cette première carte comporte des erreurs dûes à l'ambigüté de certains appariements. Les erreurs concernent généralement les pixels appartenant à des régions de couleur homogène. La deuxième étape résoud ce problème par l'introduction du principe de confiance : à chaque appariement une mesure de confiance est attribuée, les appariements ayant une faible mesure de confiance sont ignorés. La troisième étape permet de ré-estimer les disparités ignorées, ceci par la méthode de propagation de croyance sélective. La propagation se fait dans les régions de couleur homogène. La carte finale de disparités peut se traduire en une carte de profondeur contenant les distance 3D que sépare chaque point des caméras. Cette information est facilement obtenue puisque nous disposons des paramètres intrinsèques et extrinsèques du système de vision stéréoscopique.

4.5 Bases de données recueillies

Nous avons validé les algorithmes développés sur deux jeux de données recueillis à deux passages à niveau en exploitation à Lausanne en Suisse.

4.5.1 La base "Pontet"

La première base est acquise par deux caméras stéréoscopiques supervisant la zone de danger d'un passage à niveau près de la station "Pontet". Il s'agit d'un passage à niveau à deux voix ferrées, croisées par un passage piéton et une route à trois voies : deux voies dans un sens et une voie dans un autre sens (figure 4.4). Le passage piéton est gardé par deux demi barrières de chaque coté du passage. Le reste du passage à niveau est protégé par deux demi-barrières d'un coté, et d'une demi barrière de l'autre. En terme de signalisation, le passage à niveau est équipé de deux systèmes de feux tricolores installés sur deux poteaux. Au cas où les feux sont mis au rouge, le premier poteau (à droite dans l'image) ne présente pas un danger direct sur la sécurité du PN puisque les objets, tels que des voitures, camions, ou piétons, sont temporairement immobilisés en dehors de la zone de danger. Le deuxième poteau (au milieu) est positionné de sorte qu'il présente un danger potentiel. Il est installé à quelques mètres après la zone de croisement de sorte qu'une voiture (à titre d'exemple), après la traversée de la zone de danger, se trouve souvent obligée de s'arrêter. Dans le cas où cette voiture est suivie par d'autres voitures ou par un long camion, la situation devient dangereuse puisque un ou plusieurs obstacles sont temporairement immobilisés sur la zone de croisement. La figure 4.4 illustre le passage à niveau "Pontet" :

FIGURE 4.4 – Le passage à niveau "Pontet".

4.5.2 La base "Chamberonne"

Le même système de vision stéréoscopique que celui de la base "Pontet", est utilisé pour l'acquisition d'une deuxième base intitulée "Chamberonne". Il s'agit aussi d'un passage à niveau situé dans une zone urbaine. La zone de danger correspond à la zone de croisement de deux voies ferrées et d'une route à deux voies. Le passage à niveau est gardé par un système de signalisations lumineuses composé de deux feux bicolores installés sur deux poteaux de part et d'autre de la zone de croisement. Le passage à niveau est équipé par une demi barrière de chaque coté du passage ; ceci ne bloque pas l'accès à la zone de croisement. Le passage à niveau est situé dans une portion de route faisant l'objet d'un virage. La distance qui sépare les deux demi barrières est grande, ce qui fait que la zone de danger ne peut pas être totalement surveillé. Le système de vision est installé d'un seul coté du passage supervisant l'accès le plus court vers la zone de croisement. Une autre caractéristique du passage à niveau est celle de la voie ferrée qui présente un virage proche de la zone de croisement. La visibilité d'un train en approche est alors très limitée dans le sens "de gauche vers la droite" selon la direction du champ de vision. Un exemple d'une violation d'utilisation consiste à franchir le passage au moment ou après l'abaissement des barrières par un mouvement de "zig-zag" effectué par un usager (voiture, camion, etc.). Ce comportement rend le passage à niveau dans une situation dangereuse affectant la sécurité des personnes et l'état des installations. La figure 4.5 illustre le passage à niveau "Chamberonne" :

FIGURE 4.5 – Le passage à niveau "Chamberonne".

4.6 Scénarios traités

Le système de vision installé est orienté de sorte que la zone de croisement est totalement surveillé. Le champ de vision du passage à niveau "Pontet" comprend les deux croisements, passage piéton et croisement routes/rails, ainsi que les équipements de sécurité tels que les barrières et les feux de signalisations lumineuses. Le champs de vision du passage à niveau "Chamberonne" correspond à la zone de croisement ainsi qu'une demi barrière et un feu de signalisation installés du même coté que le système de vision. Dans chaque passage, l'acquisition de séquences d'images a duré environ une heure en non continue. Chaque séquence dure 15 minutes en moyenne. Le fond de la scène observée contient des éléments fixes, des arbres, et du ciel. Les objets surveillés correspondent à des piétons seuls ou en groupes, des voitures, des camions, et des cyclistes. Ces objets interagissent entre eux et génèrent des scénarios différents. La complexité des scénarios dépend du nombre et de la position des différents objets. Une scène contenant une voiture traversant la zone de croisement est un exemple d'un scénario non complexe. En plus des scénarios obtenus avec des objets en mode de fonctionnement normal, nous avons fait participer quelques piétons afin de générer des scénarios correspondant à des situations dangereuses. Les scénarios joués consistent essentiellement à traverser la zone de croisement après l'abaissement des barrières.

La figure 4.6 illustre un scénario acquis sur le passage "Chamberonne". Le scénario comprend quatre piétons et quelques voitures. La première image (a) illustre la présence de quatre piétons sur et à proximité de la zone de danger. L'image (b) illustre le cas d'une occultation partielle entre deux piétons dans la zone de croisement, les deux autres piétons s'éloignant de la zone de danger. Les piétons partiellement occultés ont des caractéristiques colorimétriques différentes. L'image (c) illustre toujours les quatre piétons : deux dans la zone de danger et deux hors de la zone de

danger. Les images (d) et (e) illustrent chacune quatre piétons dont deux sont partiellement occultés, mais cette fois les objets occultés ont des caractéristiques colorimétriques similaires ; l'un est dans la zone de danger et l'autre hors de la zone de danger. Il n'existe pas d'occultation entre les deux autres piétons. L'image (f) illustre les quatre piétons non occultés à des distances différentes. L'image (g) illustre les 4 piétons dont un est totalement occulté par un autre. L'image (h) illustre les quatre piétons dont deux s'éloignent de la zone de croisement alors que les deux autres sont dans la zone de danger. L'image (i) illustre 3 piétons dont un est hors de la zone de danger et les deux autres se trouvent dans la zone de croisement sachant que la barrière est abaissée. L'image (j) contient deux voitures dont une qui traverse la zone de croisement et l'autre hors de la zone de croisement, quatre piétons dont un est totalement occulté par une voiture et un autre portant un objet rectangulaire occultant une partie de ce dernier. L'image (k) illustre une voiture hors de la zone de croisement occultée par un piéton, un piéton partiellement occulté par une deuxième voiture traversant le passage, un piéton portant un objet rectangulaire et un quatrième piéton. La dernière image (l) contient quatre piétons à proximité de la zone de croisement dont un porte un objet rectangulaire.

La séquence d'images illustrée sur la figure 4.7 correspond à un scénario d'une situation dangereuse non prévue, acquise sur le passage à niveau "Pontet". L'image (a) du scénario contient deux piétons traversant la rue à coté de la zone de croisement, et deux voitures circulant sur une route parallèle aux voies ferrées. L'image (b) illustre deux voitures et un piéton situés au delà de la zone de croisement, les deux piétons de l'image (a) continuant leur traversés de la rue, ainsi qu'une voiture se rapprochant de la zone de croisement. L'image (c) illustre quelques voitures circulant sur la route parallèle à la voie ferrée, deux piétons continuant leur traversée de la rue et une voiture blanche en approche de la zone de croisement. Les feux de signalisation du rapprochement du train se sont déclenchés. L'image (d) illustre la voiture blanche se rapprochant de plus en plus de la zone de croisement. Les piétons sont totalement occultés et les feux de signalisation sont toujours déclenchés. L'image (e) illustre deux piétons, et la voiture blanche. Les feux de signalisation sont au rouge. L'image (f) montre la voiture blanche en train de franchir la zone de croisement sans respecter la signalisation lumineuse d'interdiction de passage. La voiture se trouve en partie dans la zone de danger. L'image (g) montre l'abaissement de la demi barrière la plus proche de la voiture, encore temporairement stationnée. L'image (h) montre l'abaissement des barrières et l'approche de la demi barrière de la voiture stationnée. L'image (i) montre que la barrière la plus proche de la voiture blanche arrêtée heurte cette dernière. L'image (j) illustre la voiture blanche après avoir forcé le passage en endomageant en partie la barrière. La voiture se trouve en pleine zone de croisement étant donné que le train est imminent. L'image (k) illustre la voiture blanche stationnée juste après la zone de croisement. La voiture a réussi à quitter la zone de danger puisque cette partie du passage n'est pas équipée par une demi barrière. Dans l'image (l), la voiture blanche s'éloigne progressivement du passage à niveau.

FIGURE 4.6 – Scénario d'une situation dangereuse sur le passage à niveau "Chamberonne".

FIGURE 4.7 – Scénario d'une situation accidentelle sur le passage à niveau "Pontet".

4.7 Illustrations des résultats de localisation 3D

Nous présentons dans cette section les résultats de localisation tridimensionnelle sur les scénarios choisis sur les passages à niveau "Pontet" et "Chamberonne". Les disparités minimale et maximale du scénario "Chamberonne" sont mesurées à 5 et 55 respectivement, tandis que les

disparités minimale et maximale du scénario de la base "Pontet" sont mesurées à 3 et 22 respectivement. Les figures allant de 4.8 à 4.19 illustrent les différentes étapes de l'algorithme de mise en correspondance détaillé dans le chapitre 2. L'image (a) de chaque figure correspond à l'image originale rectifiée prise par la caméra gauche. Sur l'image (b) de chaque figure, les objets en mouvement sont extraits et sont représentés sur un fond vert pour des raisons de visibilité. La première étape de l'algorithme d'appariement consiste à initialiser la carte de disparités en utilisant la méthode de vraisemblance *DCMP* (image (c) de chaque figure). Nous précisons que la disparité minimale est représentée par la couleur bleu tandis que la disparité maximale avec la couleur rouge. Cette première carte de disparités présente des erreurs d'appariement. La localisation 3D des objets n'est donc pas précise à ce stade. La deuxième étape de l'algorithme consiste à identifier automatiquement les pixels bien appariés à l'aide des mesures de confiance calculées pour chaque appariement. Avec un seuil de confiance de 40%, l'image (d) de chaque figure contient les pixels identifiés comme bien appariés, et ayant donc une mesure de confiance supérieure à 40%. Nous remarquons qu'une grande partie des pixels mal appariés sont alors éliminés. La troisième étape de l'algorithme permet de ré-estimer les disparités des pixels ignorés à l'étape précédente. Les pixels bien appariés représentent le point de départ de la propagation de croyance sélective *PCS*. L'image (e) de chaque figure illustre le résultat de la propagation de croyance de sorte qu'une disparité est attribuée à chaque pixel appartenant aux objets en mouvement.

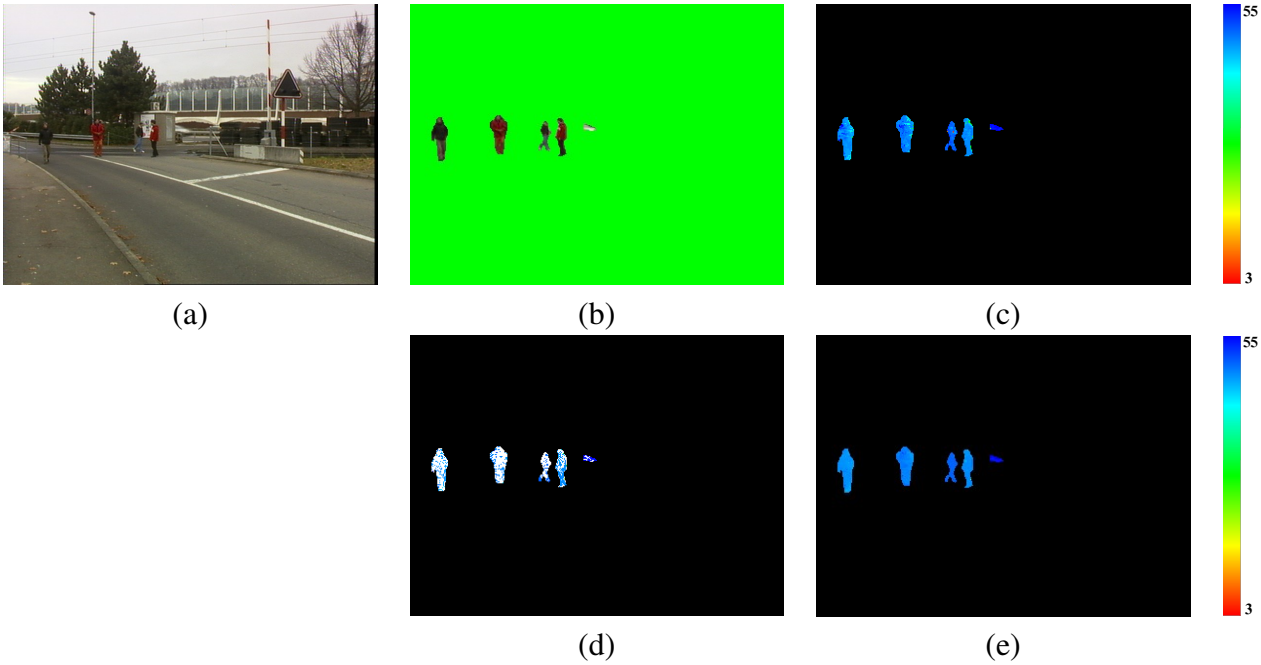


FIGURE 4.8 – Résultat de la localisation 3D de l'image (a) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

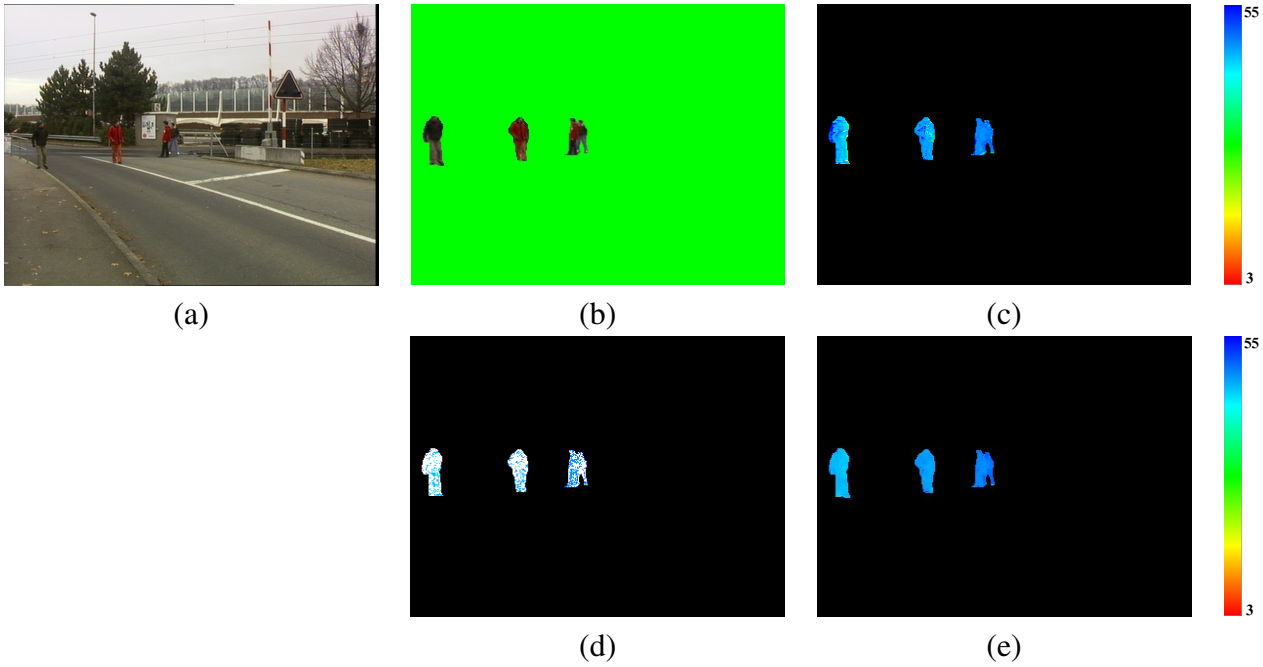


FIGURE 4.9 – Résultat de la localisation 3D de l'image (b) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

4.8 Interprétation des résultats

Nous avons testé les algorithmes de détection et de localisation tridimensionnelle d'objets en mouvement sur des scénarios divers issus de passages à niveau réels. La complexité des scénarios traités se reflète dans le nombre, le type d'objets en mouvement ou en stationnement temporaire, et dans la complexité de leurs intéractions. La plupart des scénarios sur les PN sont préparés à l'avance et sont joués par des acteurs afin de les rendre aussi complexes que possible. Dans le PN "Chamberonne", les caméras sont placées à 20 mètres de la barrière la plus proche. La zone de danger, représentée par le croisement routes/rails se trouve à une distance entre 20 et 30 mètres du capteur stéréoscopique. À ces distances, un pixel caractérise environ 10 centimètres d'espace dans la scène réelle. En effet, un objet dans ou au delà de la zone de danger est illustré par une petite région dans l'image, ceci correspond à une relative basse résolution. Une autre source de difficulté de la base traitée est le bruit du capteur *CCD* ainsi que les conditions d'illumination non stables. En faisant un zoom sur une région donnée de la scène de couleur homogène, nous constatons une variation considérable de la couleur entre deux pixels voisins. Malgré ces difficultés, l'algorithme d'extraction d'objets stationnaires ou en mouvement a prouvé son efficacité dans le cas des objets de petite taille. En terme de localisation 3D, les résultats obtenus sur les scénarios traités montrent la précision et la robustesse de la localisation, en particulier sur des objets lointains. Le scénario de la figure 4.6 illustre des piétons séparés, en groupe, des voitures et un objet déposé par un piéton.

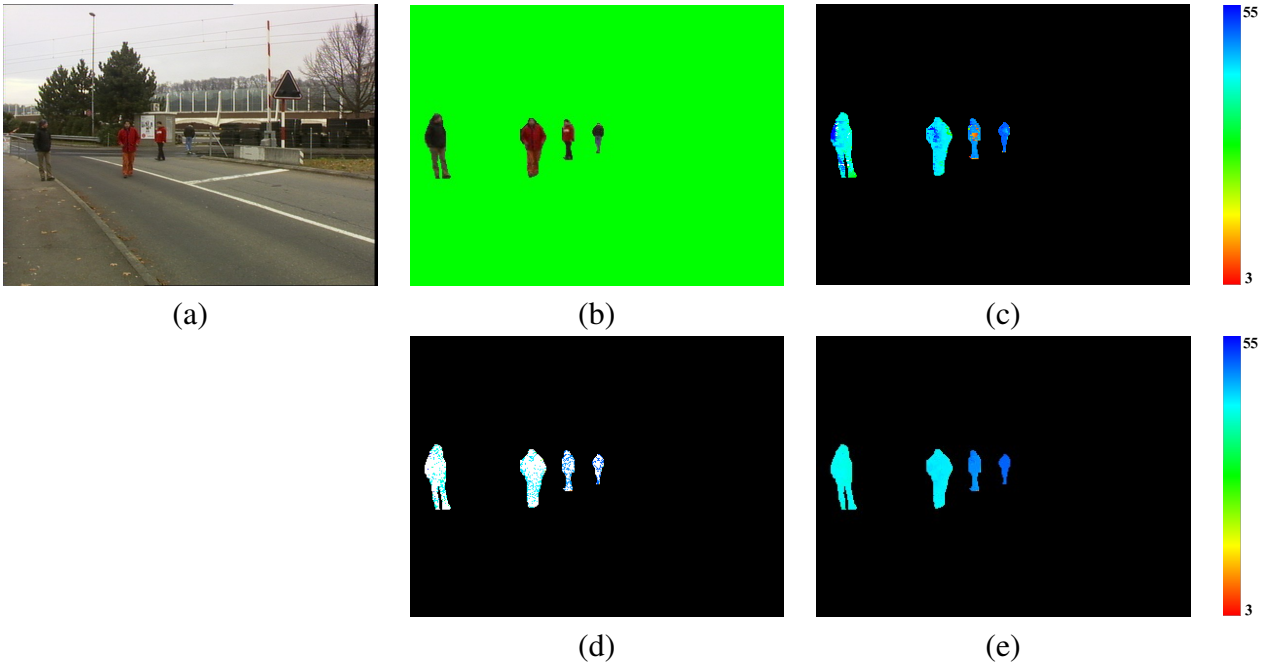


FIGURE 4.10 – Résultat de la localisation 3D de l'image (c) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

Les voitures se déplacent en s'éloignant des caméras, alors qu'une partie des piétons est sur la zone de croisement et une autre partie se rapproche des caméras. La précision et la robustesse de la localisation 3D est dûe principalement à la propagation de croyance basée sur le principe de mesures de confiance. Un tel algorithme d'appariement stéréoscopique s'est montré efficace dans le cas où les objets, dont la disparité est à estimer, ne sont pas ou peu occultés. Par contre, la chaine globale de mise en correspondance fonctionne de manière satisfaisante pour des objets très éloignés des caméras. Ceci montre la forte sensibilité de l'extraction des objets en mouvement et de la mise en correspondance des pixels représentant ces objets. Les figures 4.8, 4.10, 4.13, 4.14, 4.17, 4.18 et 4.19 illustrent des piétons non ou peu occultés à des distances différentes par rapport aux caméras. Les cartes finales de disparités montrent l'efficacité de notre approche malgré la mauvaise qualité des images traitées.

La plupart des algorithmes d'appariement stéréoscopique échouent à cause de la présence de forte occultations, ou sur des surfaces de couleurs homogènes. Les figures 4.9, 4.11, 4.12, 4.15 et 4.16 illustrent des cas difficiles dont l'appariement stéréoscopique peut échouer. La figure 4.9 illustre quatre piétons dont deux sont partiellement occultés tels que l'un est très proche de l'autre. La segmentation couleur des objets en mouvement permet de séparer les deux piétons partiellement occultés puisque ces derniers ont des caractéristiques colorimétriques différentes. La propagation de croyance dans les régions de couleur homogène respecte en effet les bords des objets.

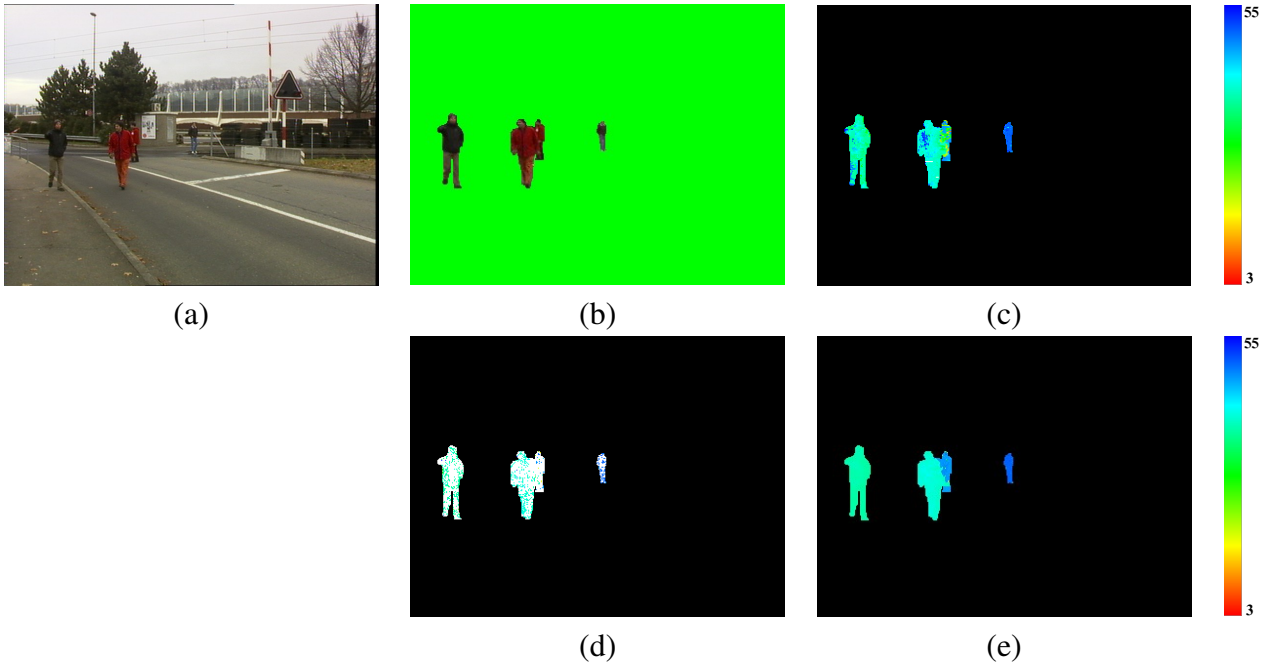


FIGURE 4.11 – Résultat de la localisation 3D de l'image (d) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

La localisation 3D est précise dans ce dernier cas. Quatre piétons sont identifiés dans l'image 4.11 dont deux sont partiellement occultés, à des distances différentes par rapport aux caméras et ayant des caractéristiques colorimétriques similaires (deux piétons portant des vêtements rouges). Nous constatons dans ce cas que la propagation de croyance a permis de bien séparer les piétons, pour les raisons suivantes :

– La partie occultée entre les deux piétons n'est pas importante et la surface de chacun d'eux est suffisamment grande.

– Une grande surface augmente la probabilité d'avoir des pixels bien appariés permettant par la suite la ré-estimation des disparités erronées.

La propagation de croyance se fait dans les deux sens dans la région commune des deux piétons : de la zone rouge du piéton 1 vers la zone rouge du piéton 2, et vice-versa (figure 4.20). Ceci explique le manque de précision dans l'estimation des disparités à la frontière des deux piétons. L'algorithme peut échouer dans le cas d'une grande occultation : il ne reste pratiquement plus de pixels bien appariés dans la petite partie visible de l'objet partiellement occulté. Dans ce cas, les disparités de la partie visible seront ceux de l'objet occultant. La figure 4.12 illustre un cas où l'algorithme d'appariement échoue. Les deux piétons partiellement occultés portent les mêmes vê-

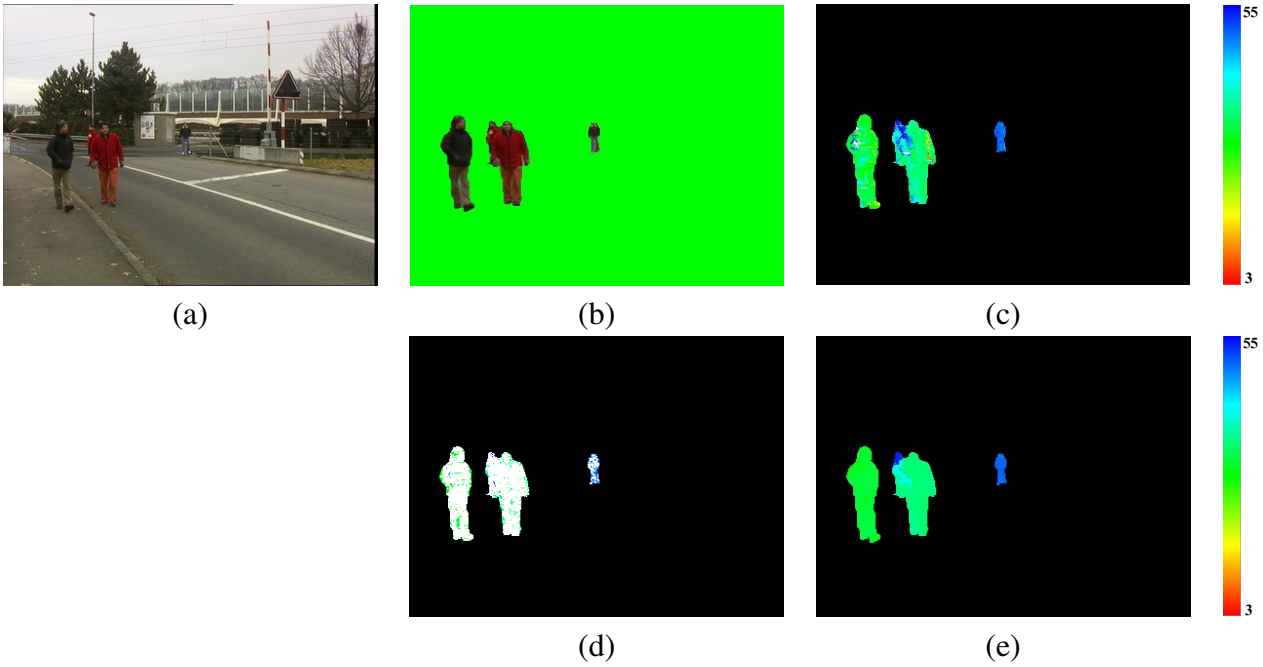


FIGURE 4.12 – Résultat de la localisation 3D de l'image (e) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

tements (rouge). Une erreur de localisation figure dans la partie commune des deux piétons. Seule une petite partie de l'objet occulté est bien appariée. La figure 4.20 montre une vue proche des ces deux piétons partiellement occultés.

Comme illustré dans la figure 4.15, seules les disparités d'une partie du piéton à gauche de l'image sont estimées. Ce défaut de localisation est lié au nombre très réduit de pixels bien appariés, ayant une mesure de confiance élevée, appartenant à cette région. Dans ce cas, les disparités de cette région ne peuvent être estimées qu'à partir des disparités des régions voisines. Sur la figure 4.16, nous remarquons que les disparités d'une partie de la région à gauche de l'image (e) ne sont pas estimées, et le reste de la région est mal estimé. Ceci est justifié par le fait que cette région est bien visible dans l'image gauche alors qu'elle ne l'est pas dans l'image droite (figure 4.21).

Dans notre application, la présence d'occultations n'est pas critique et ne représente pas un vrai problème puisque la configuration du système de vision utilisée est loin d'être optimale. Le système de vision est actuellement placé à 1.5 mètres du sol ce qui nous confronte à des cas d'occultations difficiles à gérer. Ce problème peut être aisément résolu en positionnant les caméras à plusieurs mètres au dessus du sol (5 à 6 mètres) de façon à ce que le champ de vision des caméras soit plus concentré sur la zone de croisement. Une telle configuration réduit considérablement les problèmes d'occultations et permet a priori de réduire les erreurs de localisation. Notons ainsi qu'en situation

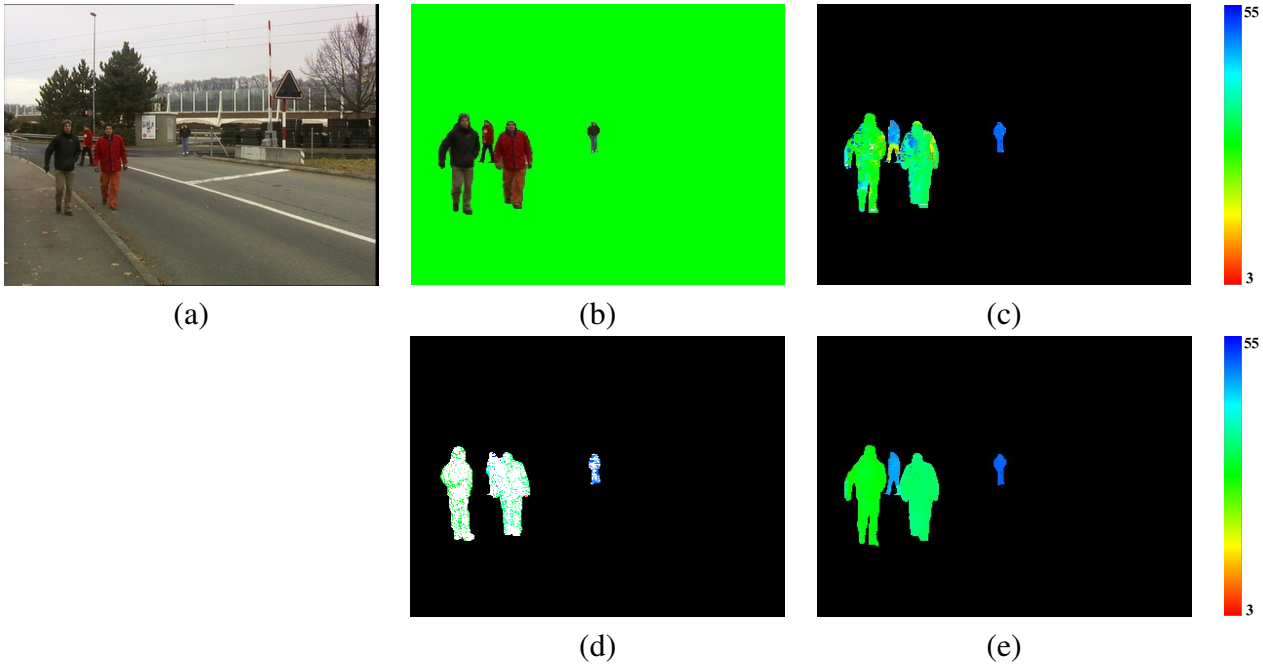


FIGURE 4.13 – Résultat de la localisation 3D de l'image (f) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

réelle, les scénarios sont généralement beaucoup moins complexes que les scénarios pris pour l'évaluation, ceci en termes du nombre d'objets présents simultanément sur un passage à niveau, et de leurs intéractions. Un des principaux objectifs de la vision artificielle dans les PN est de pouvoir détecter et localiser un ou plusieurs obstacles sur la zone de danger, une tâche que notre système de vision est maintenant capable de réaliser.

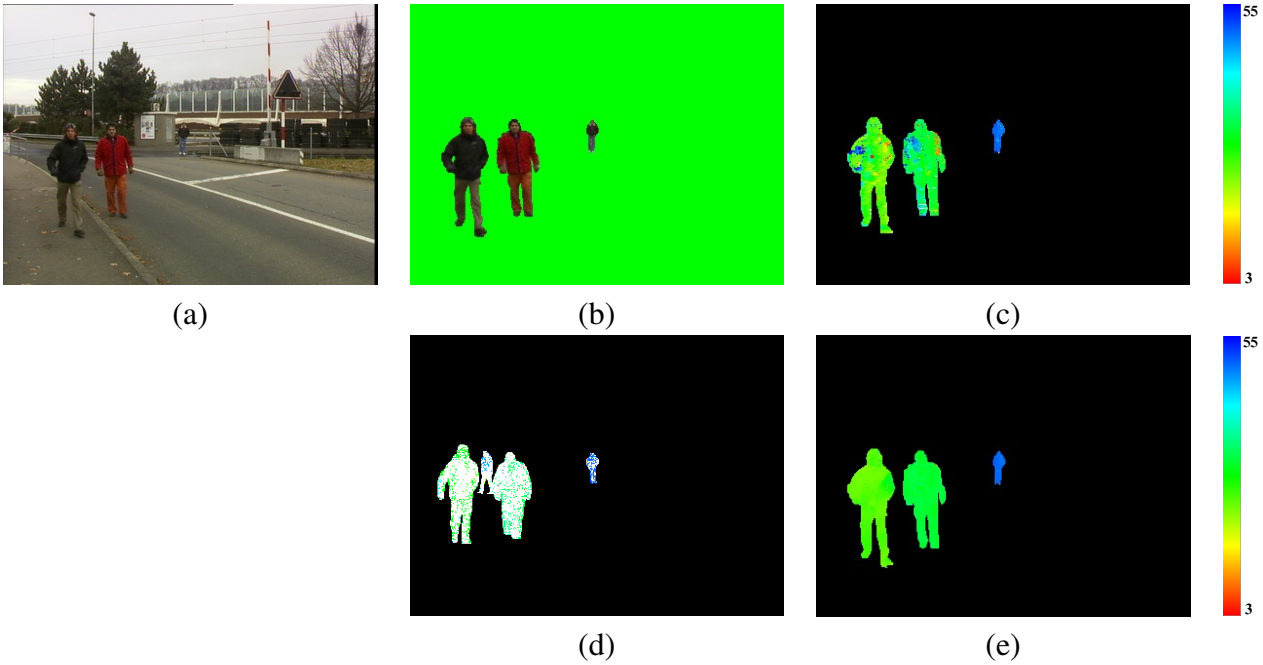


FIGURE 4.14 – Résultat de la localisation 3D de l'image (g) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

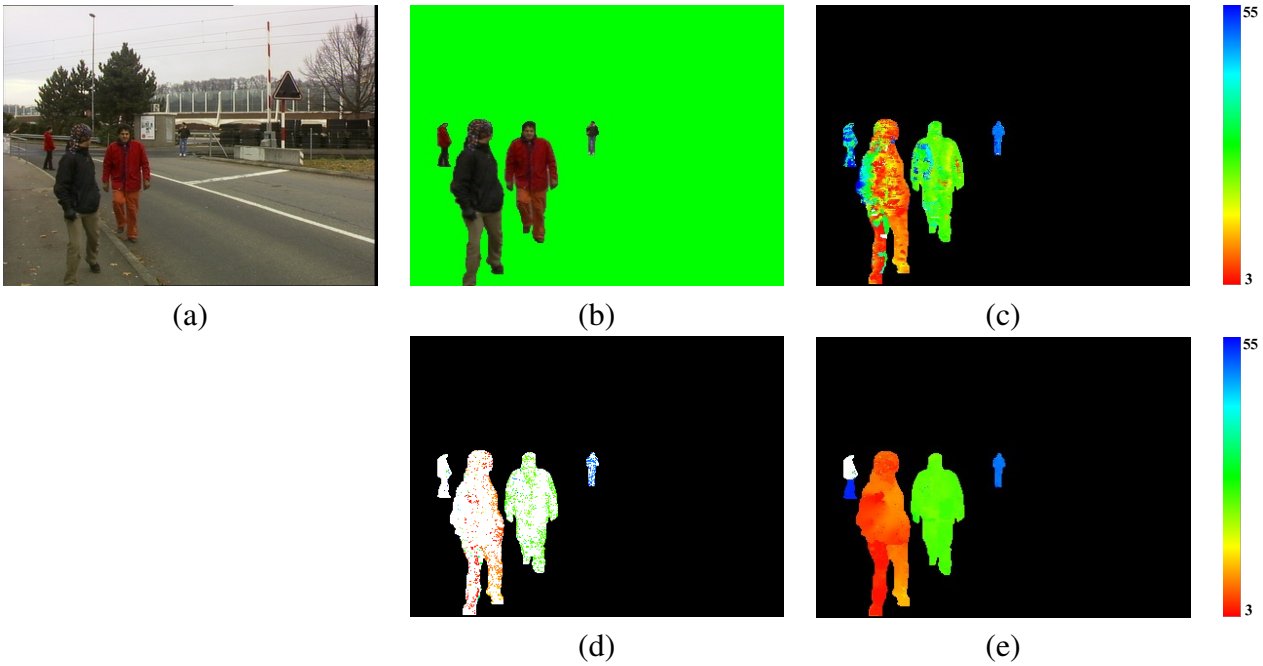


FIGURE 4.15 – Résultat de la localisation 3D de l'image (h) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

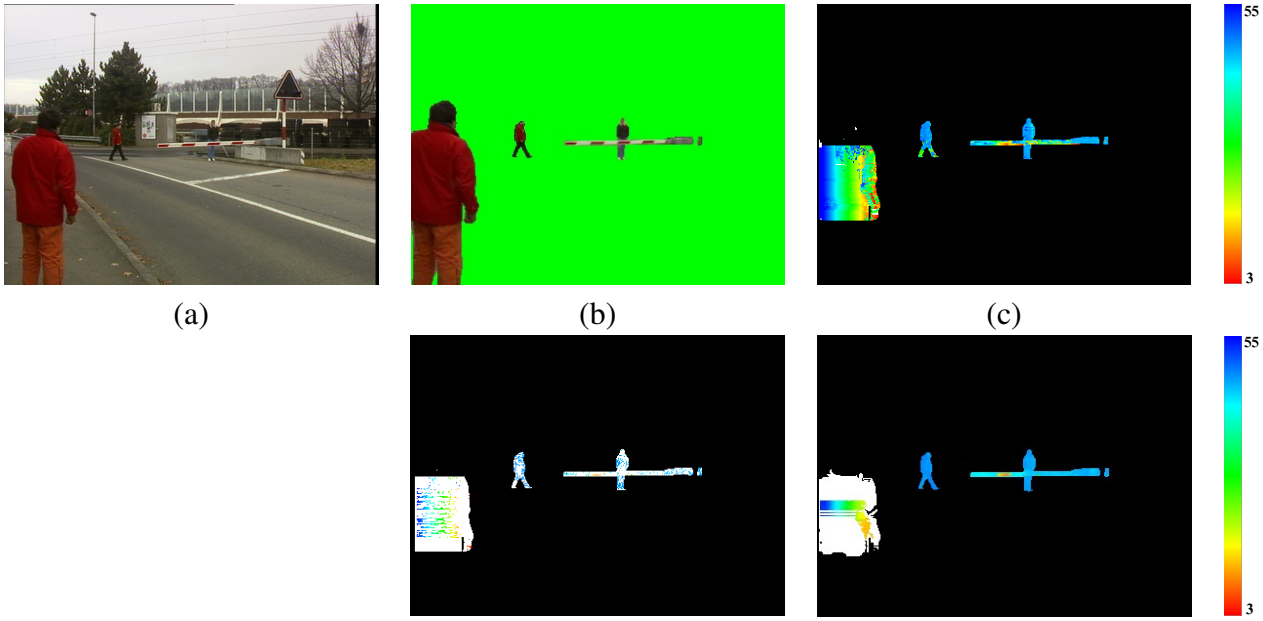


FIGURE 4.16 – Résultat de la localisation 3D de l'image (i) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

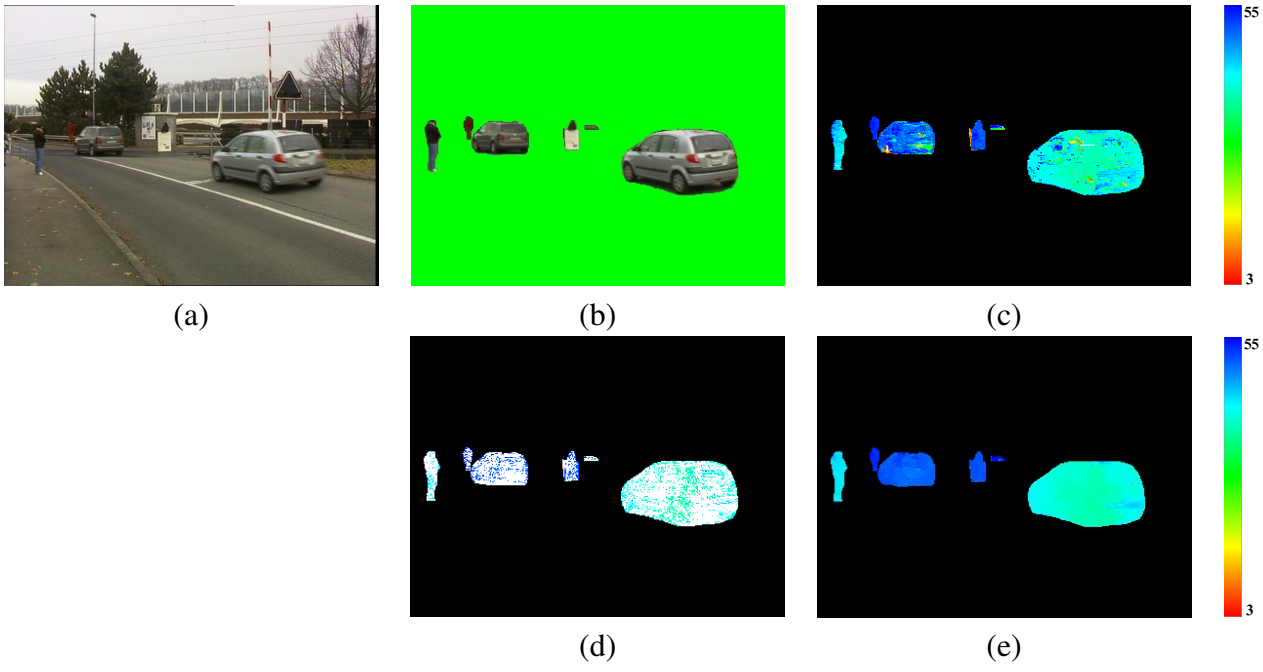


FIGURE 4.17 – Résultat de la localisation 3D de l'image (j) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

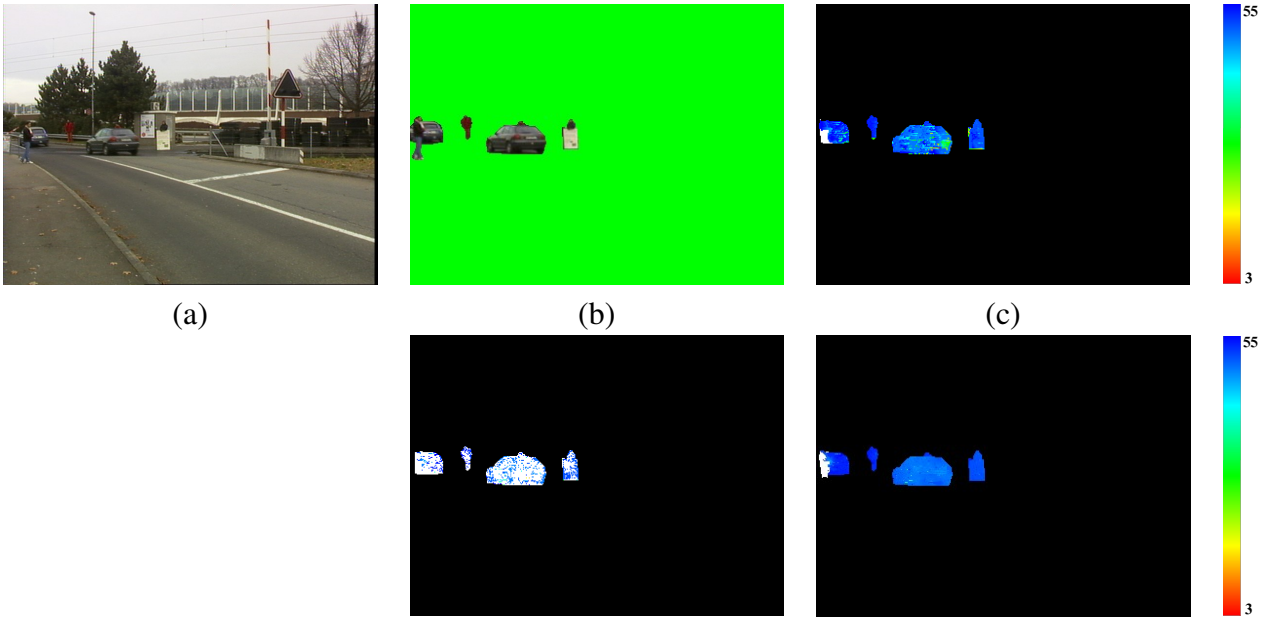


FIGURE 4.18 – Résultat de la localisation 3D de l'image (k) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.

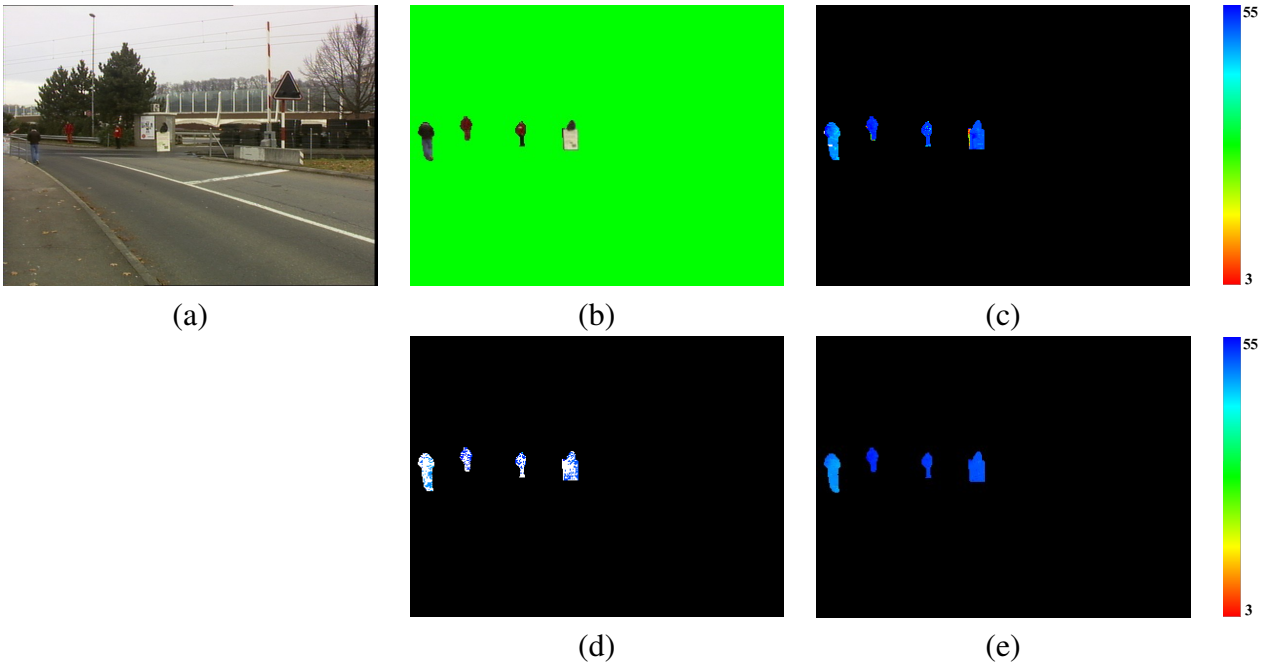
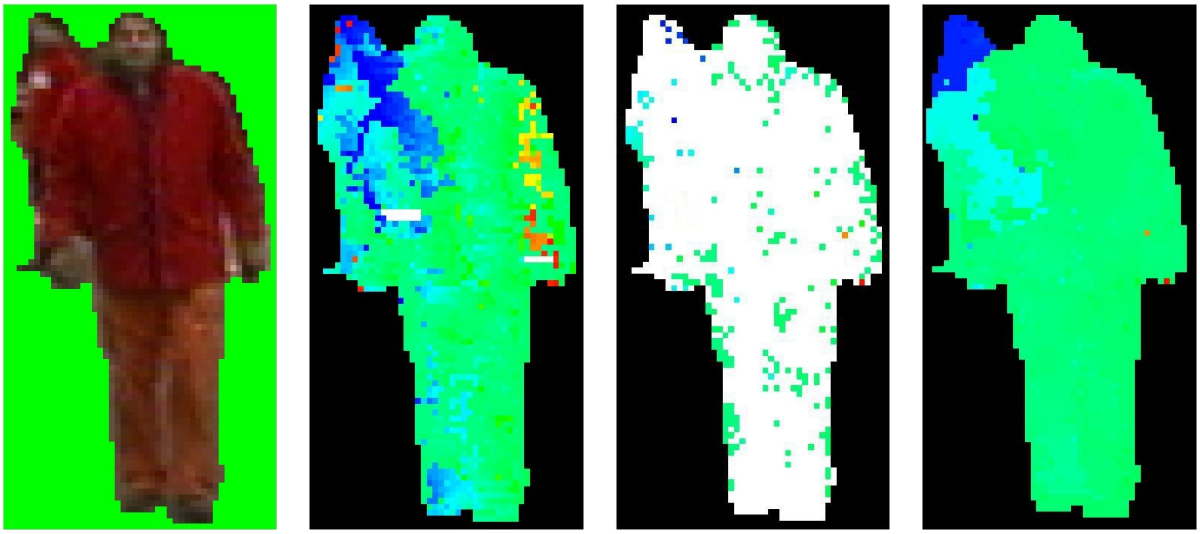


FIGURE 4.19 – Résultat de la localisation 3D de l'image (l) du scénario de la figure 4.6. (a) image originale, (b) objets en mouvement, (c) carte de disparités obtenue avec la méthode de vraisemblance *DCMP* sur les pixels affectés par du mouvement, (d) pixels identifiés comme bien appariés, (e) carte de disparités améliorée par la *PCS*.



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

FIGURE 4.20 – Résultat de localisation 3D illustré sur deux piétons partiellement occultés et ayant des caractéristiques colorimétriques similaires. (a) deux piétons extrait par ACI (b) localisation 3D par l'algorithme DCMP (c) après l'application d'un seuil de confiance de 60% (d) après propagation de croyance sélective.

(a) (b)

FIGURE 4.21 – Cas d'un objet visible par une caméra et non visible par l'autre.

Conclusion générale

La perception des reliefs d'une scène à partir de la vision artificielle est devenue un besoin croissant dans diverses applications. La troisième dimension est utile dans des applications où les distances 3D des primitives sont indispensables. En sécurité routière, la distance 3D d'obstacles routiers est une information cruciale pour éviter d'éventuels accidents et pour l'aide à la conduite. Dans ce cas, une reconstruction partielle de primitives est souvent suffisante. La perception 3D peut être effectuée à partir d'un système de vision monoculaire, binoculaire, ou n-oculaires. La vision monoculaire est souvent dédiée aux traitements bas niveau par analyse locale d'images telle que la détection des contours, la segmentation basée sur l'intensité, la couleur, ou le mouvement, l'analyse des textures, etc. La perception tridimensionnelle d'un environnement est aussi possible par l'utilisation d'une seule caméra, ceci à l'aide de la connaissance préalable du modèle géométrique du système de vision, tels que les paramètres intrinsèques de la caméra, et des a priori sur la scène observée. Cette configuration présente des limites dans la localisation 3D et elle n'est pas capable de gérer les problèmes d'occultations partielles. La vision binoculaire, dite aussi vision stéréoscopique, est un moyen performant souvent adopté pour combler les limites de la localisation 3D d'un système de vision monoculaire. Il s'agit dans ce cas de deux caméras supervisant simultanément la même scène. Les caméras sont souvent placées l'une à coté de l'autre puisque c'est le décalage d'un même pixel ou primitive entre les deux vues qui permet d'estimer les distances 3D, ceci à l'aide d'autres informations telles que les paramètres extrinsèques et intrinsèques relatifs à chaque caméra. Cette technique ne permet pas d'avoir un rendu tridimensionnel parfait puisque nous ne disposons pas suffisamment de vues pour une même scène. Cette technique peut être utilisée pour des applications dont le système de vision est embarqué tel que la navigation de robots, la détection d'obstacles dans un environnement routier, ou des applications dont le système de vision est fixe tel que le suivi de cibles, et la reconnaissance de scénarios de situations dangereuses pour la vidéosurveillance.

Une étape fondamentale de la vision stéréoscopique est la mise en correspondance, dit aussi appariement stéréoscopique. Etant données deux images, gauche et droite, représentant la même scène, l'appariement consiste à faire apparier chaque pixel, ou primitive, de l'image gauche, considérée comme image de référence, en cherchant le pixel ou primitive le plus homologue dans l'image droite. Deux approches sont possibles pour réaliser cette tâche : la famille des approches

locales, et la familles des approches globales. En se basant sur les approches locales, seul le voisinage du pixel à apparier est considéré lors de l'appariement. Pour chaque pixel de l'image gauche, une fonction d'appariement est appliquée afin de choisir le pixel le plus homologue dans l'image droite. Le couple de pixels avec lequel la fonction de vraisemblance donne une valeur optimale, sont considérés comme homologues. Le décalage entre deux pixels homologues s'appelle disparité. Une carte de disparités est ainsi obtenue pour l'ensemble des pixels de l'image de référence. L'avantage de cette approche est que l'appariement peut se faire en temps réel. Les méthodes locales présentent la limite de ne pas pouvoir apparier les pixels partiellement occultés, appartenant à des régions d'intensité ou de couleur homogènes, ou appartenant à des régions de textures répétitives. L'amélioration de la qualité des appariements fait l'objet de plusieurs travaux de recherche, mais il reste des limites qui ne peuvent être gérées que par des méthodes globales. La deuxième catégorie de méthodes d'appariement introduit des contraintes d'ordre globale lors du processus de mise en correspondance, telles que la contrainte de symétrie, d'ordre, de continuité, etc. Les méthodes globales établissent les relations spatiales et colorimétriques entre les pixels ou primitives à apparier afin de lever certaines ambiguités de mise en correspondance. Malgré une amélioration considérable de la qualité d'appariement, les méthodes globales restent inefficaces dans certains cas et le temps de traitement élevé reste une limite à surmonter.

C'est dans ce cadre que nous nous positionnons afin de proposer un nouvel algorithme tirant profil des méthodes locales et globales tout en considérant la contrainte temps de traitement. Nous avons introduit dans le premier chapitre le contexte applicatif, ainsi que le cadre scientifique et académique de notre recherche. Les contraintes imposées par l'application sont pris en compte lors du choix, de la conception, et du développement de notre solution. Nous avons présenté dans le deuxième chapitre la géométrie d'un tel système de vision stéréoscopique et les paramètres nécessaires permettant l'estimation des distances 3D. La deuxième partie du deuxième chapitre est consacrée à un état de l'art récent sur les différentes techniques permettant la mise en correspondance stéréoscopique. Nous avons présenté dans l'état de l'art les méthodes locales et globales.

Compte tenu des avantages et des limites des méthodes locales et globales, nous avons présenté dans le troisième chapitre une nouvelle méthode d'appariement stéréoscopique. Le principal objectif était d'améliorer la qualité d'appariement d'images acquises dans un environnement réel, extérieur, et bruité, et dont l'illumination diffère entre les images gauche et droite. La première étape de l'algorithme proposé consiste à estimer une première carte de disparités en appliquant une méthode locale intitulée "*Différence de Couleur Moyenne Pondérée*" que nous avons proposés. Cette méthode a prouvé son efficacité en terme d'amélioration de la qualité d'appariement comparée avec d'autres méthodes existantes dans la littérature telles que la méthode *SAD* ou *SSD*. En analysant la carte dense de disparités obtenue, nous avons constaté que les erreurs d'appariement concernent généralement les pixels partiellement occultés, ou les pixels appartenant à des

régions de couleur homogène, ou des régions de textures répétitives. Nous avons alors cherché à repérer automatiquement ces pixels en se basant sur des paramètres d'ordre local. L'établissement d'une mesure de confiance à chaque couple de pixels homologues a permis de repérer les pixels correctement appariés. En fixant un seuil de confiance, les couples ayant une mesure de confiance supérieure à ce seuil sont gardés pour la prochaine étape, alors que le reste des pixels est ignoré momentanément. Une carte éparse est alors obtenue. La dernière étape consiste à retrouver les disparités des pixels ignorés. De ce fait, nous avons opté pour l'utilisation d'une méthode globale par le choix de la méthode de propagation de croyance. Le choix d'une telle méthode est justifié par l'étendue de cette méthode dans diverses applications dans lesquelles elle a prouvé son efficacité, par sa simplicité d'implémentation, et par les possibilités d'optimisation qu'offre cette méthode. L'algorithme de propagation de croyances développé consiste à faire passer d'une manière hiérarchique, itérative, et sélective des messages entre les pixels. La hiérarchie réside dans le fait que les messages sont transmis des pixels ayant des mesures de confiances élevées, vers les pixels ayant des mesures de confiances faibles. Une autre contrainte est introduite lors du processus de propagation des messages qui consiste à ne faire passer des messages qu'à l'intérieur des régions de couleur homogène. Après quelques itérations, l'algorithme d'appariement ré-estime les disparités manquantes et fournit alors une carte de disparités dense. Notre méthode d'appariement est comparée avec d'autres méthodes globales telles que la méthode de coupure de graphe, et a conduit à des résultats satisfaisants en termes de temps de traitement et de qualité d'appariement.

Sachant que le système de vision stéréoscopique est formé de deux caméras fixes, la distance des points non affectés par du mouvement reste la même, seule la distance des pixels affectés par du mouvement varie au fil du temps. Nous avons eu l'idée d'exploiter le mouvement afin d'accélérer le processus d'appariement, par l'introduction d'une nouvelle contrainte qui dit que l'homologue d'un pixel affecté par du mouvement est lui aussi affecté par du mouvement. L'extraction des régions affectées par du mouvement à partir d'une séquence d'images fait alors l'objet du quatrième chapitre. Nous avons débuté ce chapitre par un état de l'art sur les méthodes de soustraction de fond, basée sur le principe de comparaison avec une image de référence. Nous avons proposé une méthode innovante basée sur l'Analyse en Composantes Indépendantes (ACI) et un nouveau filtre basé sur la propagation de croyance spatio-temporelle. L'ACI est adaptée dans notre cadre pour l'extraction d'objets en mouvement. Le choix de l'ACI est justifié par sa rapidité en termes de temps de traitement, et par son insensibilité des variations continues d'illumination. Le filtrage spatio-temporel est introduit afin de réduire le bruit et afin d'extraire aisément les régions affectées par du mouvement. Nous avons testé et comparé notre méthode avec d'autres méthodes de la littérature telles que la méthode "*Codebook*" ou la méthode de mélange de gaussiennes. L'évaluation est effectuée sur quelques séquences d'images issues d'environnements extérieurs, dont 1000 images sont segmentées manuellement pour établir une vérité terrain. En termes de Rappel et Précision, notre méthode a présenté des résultats très encourageants.

Par rapport à l'application, nous avons proposé dans la dernière partie de ce mémoire une étude sur la sécurité des passages à niveau et les possibilités que peut offrir un tel système de vision. Nous avons commencé le cinquième chapitre par une analyse préliminaire des risques liée à la sécurité des passages à niveau. Cette analyse nous a permis de comprendre les principales causes des accidents. Nous avons identifié le facteur humain comme étant la source de la majorité des incidents et des accidents. En effet, l'analyse du comportement des usagers des passages à niveau est un des moyens pour l'anticipation des risques et pour la réduction du nombre d'accidents. La vision artificielle est identifiée comme une technique émergeante permettant la surveillance des zones critiques. Le nombre et la qualité des informations que nous pouvons acquérir à partir d'un tel système de vision sont d'une importance majeure. Nous proposons ainsi dans la deuxième partie de ce chapitre une évaluation des algorithmes développés sur des bases de données réelles. Une illustration de quelques scénarios ainsi que le résultat de localisation 3D d'obstacles est ainsi proposée.

Perspectives scientifiques

– *Fusion spatio-temporelle pour la localisation 3D* : L'algorithme de localisation tridimensionnelle détaillé dans le deuxième chapitre passe par les étapes suivantes. La première étape consiste à estimer une carte dense de disparités. La deuxième étape permet de calculer une mesure de confiance à chaque appariement afin d'identifier les couples bien mis en correspondance de ceux qui ne le sont pas. Nous obtenons ainsi une carte éparse de disparités. La dernière étape consiste à ré-estimer les disparités manquantes par propagation des croyances sélective. Le deuxième algorithme de soustraction de fond à partir d'une séquence d'images, détaillé dans le quatrième chapitre, est totalement indépendant de l'algorithme d'appariement stéréoscopique. Le mouvement est intégré dans notre cadre comme une contrainte temporelle introduite dans le but de réduire le temps de traitement et l'ambiguité d'appariement. Une fusion des deux algorithmes est possible, telle que l'algorithme développé par J. Zhu [ZWGY10] qui propose une méthode spatio-temporelle pour l'estimation d'une carte de profondeur. Le système proposé par l'auteur consiste en un sous système de vision stéréoscopique couplé avec un capteur actif dit Time-Of-Flight (TOF). Nous proposons ainsi d'étendre nos travaux en exploitant le mouvement comme suit : La fusion des deux algorithmes d'extraction des régions en mouvement et de localisation 3D sous un même cadre général. Il s'agit dans ce cas d'ajouter un terme de mouvement dans la fonction objectif à optimiser. La contrainte temporelle permet d'estimer la disparité d'un pixel à l'instant (t+1) étant donné les disparités du même pixel à des instants antérieurs.

– *Optimisation par extraction de points d'intérêts* : Il est possible d'optimiser l'algorithme d'appariement stéréoscopique en partant du principe qu'un pixel appartenant à une région de couleur homogène a probablement une mesure de confiance faible. De ce fait, l'estimation d'une disparité à ces pixels semble être inutile. Nous proposons alors de remplacer la première étape d'estimation d'une carte de disparités dense par une autre permettant d'estimer une carte éparse. Il suffit d'identifier les pixels ou les primitives ayant des caractéristiques spatiales ou colorimétriques particulières. Nous pouvons appliquer un détecteur de points caractéristiques, tel que la méthode intitulée Scale-invariant feature transform SIFT [Low99], ou la méthode intitulée Speeded Up Robust Features SURF" [BTG06]. Une méthode locale d'appariement est alors appliquée sur cet ensemble de pixels. La fonction mesurant le degré de confiance des couples appariés est appliquée ainsi sur les pixels caractéristiques.

– *Amélioration de la fonction locale d'appariement* : Nous avons proposé dans le deuxième chapitre une nouvelle méthode locale d'appariement stéréoscopique. L'estimation de la corrélation entre deux pixels se base sur les lignes horizontales, verticales, et les diagonales, passant par le pixel à apparier dans une fenêtre carrée. Les segments ont le même degré d'importance dans le calcul d'une mesure de vraisemblance. Nous proposons d'étendre la méthode proposée en introduisant les idées suivantes :

  1. Choisir une forme appropriée de la fenêtre d'agrégation : Il a été justifié dans la littérature qu'un choix approprié de la forme de la fenêtre d'agrégation peut améliorer la qualité d'appariement de certains pixels.

  2. Attribuer des poids adaptatifs à chaque segment : Le fait d'attribuer un poids à chaque segment permet a priori d'améliorer la qualité d'appariement des pixels partiellement occultés. Le voisinage d'un pixel situé dans une région à côté d'une discontinuité en profondeur de l'image gauche, n'est pas tout à fait le même que le voisinage du pixel le plus homologue dans l'image droite.

– *Proposition de nouvelles méthodes d'appariement stéréoscopique* : Durant la recherche et la rédaction de l'état de l'art, nous avons identifié une méthode pouvant être une piste intéressante pour résoudre en partie les problèmes de la vision stéréoscopique. Cette méthode s'intitule "*optimisation par essaims particulaires*" qui est une méta-heuristique d'optimisation inventée par R. Eberhart [ESD96] et J. Kennedy [KE95]. Cet algorithme s'inspire à l'origine du monde du vivant, et s'appuie notamment sur un modèle développé par C. Reynolds à la fin des années 1980, permettant de simuler le déplacement d'un groupe d'oiseaux. Cette méthode d'optimisation se base sur la collaboration des individus entre eux. Elle a

d'ailleurs des similarités avec les algorithmes de colonies de fourmis, qui s'appuient eux aussi sur le concept d'auto-organisation. Cette idée veut qu'un groupe d'individus peu intelligents puisse posséder une organisation globale complexe. Ce principe a été récemment introduit dans la vision artificielle tel que dans le suivi 3D [KKD08], [KHD10], et [JT10], le calibrage de caméras [KRW09]. Cette méthode a aussi été récemment utilisée pour la reconstruction 3D et l'estimation des cartes denses de disparités à partir d'un système de vision multi-vues [WN10] et [SZ09]. La méthode d'optimisation par essaims particulaires pourra être un cadre intéressant pour la combinaison des modules d'appariement stéréoscopique et de suivi 3D.

Perspectives en termes d'application

Le système de vision proposé permet de fournir des informations complémentaires sur l'état du passage à niveau. Les usagers qui sont à coté ou dans la zone de croisement sont détectés et localisés. L'état des feux de signalisation ainsi que l'état des barrières ne sont pas détectés par le système de vision pour deux raisons : la première est que les feux de signalisation ainsi que les barrières ne sont pas toujours visibles. La deuxième raison est que ces informations peuvent être fournies par d'autres capteurs. À ce stade, plusieurs questions se posent alors sur la manière d'exploiter les informations de détection issues du système de vision. En particulier, la question qui se pose est de savoir avec qui, et comment, partager ces informations générées par le système de vision.

- Sous quelles formes les informations échangées doivent-elles être traduites : quelques images-clés, position de l'obstacle, mode de représentation ?

- À qui faut-il envoyer le résultat de la détection et de localisation 3D d'obstacles (centre de contrôle/commande, conducteur du train, usagers en approche du PN) ?

- Quelles sont les technologies permettant d'échanger ces informations (GPS, panneaux d'affichage variable, antennes de transmission, etc.) ?

L'analyse du degré de dangerosité d'une telle situation est possible par vision artificielle, nous proposons d'étendre notre travail selon deux directions :

- **Traitement en ligne** : Cette étape consiste à développer un module de suivi d'obstacles. Le suivi permettra d'anticiper les scénarios de situations accidentelles en analysant individuellement le comportement de chaque objet.

– **Traitement hors ligne** : Il est possible d'alimenter une base de données par des statistiques sur les types d'usagers, le nombre d'intéractions avec le PN par type d'usagers, la fréquence d'utilisation du PN par type d'usagers, etc. Ces statistiques pourront être utiles dans le sens où l'opérateur pourra prendre des mesures préalables permettant de maximiser le niveau de sécurité aux passages à niveau. Pour cela, un module de reconnaissance d'objets peut être ajouté à la chaine de traitement.

Liste des publications

Journaux internationaux

– N. Fakhfakh, L. Khoudour, E.M. El-Koursi, J.-L. Bruyelle, A. Dufaux, et J. Jacot : 3D objects localization using fuzzy approach and hierarchical belief propagation : application at level crossings. *In EURASIP Journal on Image and Video Processing*, No. 4, pp. 1–15, January 2011.

– N. Fakhfakh, L. Khoudour, E.M. El-Koursi, J. Jacot, et A. Dufaux : A Video-Based Object Detection System for Improving Safety at Level Crossings. *In Open Transportation Journal*, Vol. 5, pp. 1–15, 2011.

– N. Fakhfakh, L. Khoudour, E.M. El-Koursi, J.-L. Bruyelle, A. Dufaux, et J. Jacot : Robust Background Subtraction in Complex Environment by Independent Component Analysis and Spatio-Temporal Belief Propagation. Submitted to *IEEE Transactions on Image Processing*, 2011.

Communications internationales avec actes et comité de lecture

– N. Fakhfakh, L. Khoudour, E.M. El-Koursi, J.-L. Bruyelle, A. Dufaux, et J. Jacot : Robust 3D Objects Localization using Hierarchical Belief Propagation in Real World Environment. *In International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV'10)*, pp. 30–35, Las Vegas, USA, July 2010.

– N. Fakhfakh, L. Khoudour, E.M. El-Koursi, J.-L. Bruyelle, A. Dufaux, et J. Jacot : Dense Stereo Matching by Hierarchical Belief Propagation based on Fuzzy Confidence Approach. *In 14th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES'10)*, Cardiff, Wales, UK, 2010.

– N. Fakhfakh, L. Khoudour, E.M. El-Koursi, J.-L. Bruyelle, A. Dufaux, et J. Jacot : Background Subtraction and 3D Localization of Moving and Stationary Obstacles at Level Cros-

sings. *In International Conference on Image Processing Theory, Tools and Applications (IPTA'10)*, pp. 72–78, Paris, France, 2010.

– N. Fakhfakh, L. Khoudour, E.M. El-Koursi, J. Jacot, et A. Dufaux : A New Selective Confidence Measure-Based Approach for Stereo Matching. *In 13th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, Springer-erlag Berlin Heidelberg*, Vol. 5711, pp. 184–191, Santiago, Chile, 2009.

– N. Fakhfakh, L. Khoudour, E.M. El-Koursi, J. Jacot, et A. Dufaux : Mise en Correspondance Stereoscopique d'Images Couleur pour la Detection d'Objets Obstruant la Voie aux Passages a Niveau. *In Colloque International TELECOM'09 & 6th JFMMA*, pp. 206–209, Agadir, Maroc, 2009.

– L. Khoudour, N. Fakhfakh, et E.M. El-Koursi : Localisation tri-dimensionnelle robuste d(objets par propagation de croyance hiérarchique. *In Colloque International Telecom'2011 & 7èmes JFMMA*, 13-16 mars Tanger, Maroc 2011.

Séminaires internes avec actes et comité de lecture

– N. Fakhfakh : Détection de régions en mouvement et suivi 3D d'obstacles aux passages à niveau. *In Journée des doctorants de l'INRETS*, Villeneuve d'Ascq, France. Avril 2009.

– N. Fakhfakh : Étude d'un système multicapteurs pour la reconnaissance de situations potentiellement dangereuses aux passages à niveau. *In Journée des doctorants de l'INRETS*, Villeneuve d'Ascq, France. Avril 2008.

Rapports techniques internes

– N. Fakhfakh : Vers un système multi-capteurs pour la détection et la reconnaissance de situations potentiellement dangereuses. Application aux passages à niveau. *ISRN : INRETS/RA-08-736-FR*, Janvier 2008.

Lauréats

– N. Fakhfakh : A Video-Based Obstacles Localization System for Safety at Level Crossings. *Finalist in the competition of the Young European Arena of Research, Bruxelles*, 2010.

– N. Fakhfakh : président d'une session intitulée "Motion and Orientation Estimation + Tar-

get and Object Tracking + Extraction + Video Editing" de la conférence : *International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV'10). The World Congres in Computer Science, Computer Engineering and Applied Computing*. In Las Vegas, Nevada, USA July 12-15, 2010.

Bibliographie

[ACRB06] N. ANANTRASIRICHAI, C.N. CANAGARAJAH, D.W. REDMILL et D.R. BULL : Dynamic programming for multi-view. *In Proc. on IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[ACY96] S.-I. AMARI, A. CICHOCKI et H.H. YANG : A new learning algorithm for blind source separation. *In In Advances in Neural Information Processing Systems, MIT Press*, pages 757–763, 1996.

[AGV10] R. S. ALLISON, B.J. GILLAM et E. VECELLIO : Binocular depth discrimination and estimation beyond interaction space. *Journal of Vision*, 9(1):1–14, 2010.

[AKN07] S.A. ADHYAPAK, N. KEHTARNAVAZ et M. NADIN : Stereo matching via selective multiple windows. *Journal of Electronic Imaging*, 16(1):1–14, 2007.

[AKT08] K. ALAHARI, P. KOHLI et P.H.S. TORR : Reduce, reuse & recycle : Efficiently solving multi-label mrfs. *In CVPR*, 2008.

[Arn78] D. ARNOLD : Local context in matching edges for stereo vision. *In Proc. on Image Understanding Workshop, Science Apphcations*, pages 65–72, May 1978.

[Arn83] R. D. ARNOLD : Automated stereo perception. Rapport technique, Technical Report AIM-351, AI Lab, Stanford University, 1983.

[AZK08] N. ALLDRIN, T. ZICKLER et D. KRIEGMAN : Photometric stereowith non-parametric and spatially-varying reflectance. *In CVPR*, 2008.

[Bak80] H. BAKER : Edge-based stereo correlation. *In Proc. Image Understanding Workshop, Science Applications*, pages 168–175, Apr 1980.

[BB81] H.H. BAKER et T.O. BINFORD : Depth from edge and intensity based stereo. *In In : Proc. 7th Inter. Joint Conf. on Artificial Intelligence*, pages 631–636, 1981.

[BB01] J. BANKS et M. BENNAMOUN : Reliability analysis of the rank transform for stereo matching. *IEEE Trans. on Systems, Man and Cybernetics (SMC)*, 31(6):870–880, Dec 2001.

[BBH03] M.Z. BROWN, D. BURSCHKA et G.D. HAGER : Advances in computational stereo. *IEEE Trans. on PAMI*, 25:993–1008, Aug 2003.

[BC01] J. BANKS et P. CORKE : Quantitative evaluation of matching methods and validity measures for stereo vision. *International Journal of Robotics Research*, 20(7):512–532, 2001.

[BCFG05] A. BROGGI, C. CARAFFI, R.I. FEDRIGA et P. GRISLERI : Obstacle detection with stereo vision for off-road vehicle navigation. *In Obstacle Detection with Stereo Vision for off-road vehicle navigation*, page 65, 2005.

[Ben84] M. BENARD : Automatic stereophotogrammetry a method bared on feature detecton and dynamic programming. *Pattern Recognition in Photogrammetry*, 39:169–181, 1984.

[BES05] G.J. BROUWER, R.V. EE et J. SCHWARZBACH : Activation in visual cortex correlates with the awareness of stereoscopic depth. *Journal of Neuroscience*, 25(45):10403–10413, 2005.

[BF82] S.T. BARNARD et M.A. FISCHLER : Computational stereo. *ACM Computing Surveys (CSUR)*, 14:553–572, 1982.

[BG04] M. BLEYER et M. GELAUTZ : A layered stereo algorithm using image segmentation and global visibility constraints. *In ICIP*, 2004.

[BG05] M. BLEYER et M. GELAUTZ : A layered stereo matching algorithm using image segmentation and global visibility constraints. *ISPRS Journal of Photogrammetry & Remote Sensing*, 59:128–150, 2005.

[BG07] M. BLEYER et M. GELAUTZ : Graph-cut-based stereo matching using image segmentation with symmetrical treatment of occlusions. *Signal Processing : Image Communication*, 22:127–143, 2007.

[BHM05] R. BROCKERS, M. HUND et B. MERTSCHING : Stereo vision using cost-relaxation with 3d support regions. *In Image and Vision Computing New Zealand*, 2005.

[BI99] A.F. BOBICK et S.S. INTILLE : Large occlusion stereo. *IJCV*, 33(3):181–200, 1999.

[BJK07] R. BASRI, D.W. JACOBS et I. KEMELMACHER : Photometric stereo with general, unknown lighting. *IJCV*, 72(3):239–257, Mai 2007.

[BKP09] M. BUJNAK, Z. KUKELOVA et T. PAJDLA : 3d reconstruction from image collections with a single known focal length. *In ICCV*, 2009.

[BKR+08] G. BLOHM, A.Z. KHAN, L. REN, K.M. SCHREIBER et J. D. CRAWFORD : Depth estimation from retinal disparity requires eye and head orientation signals. *Journal of Vision*, 8(16):1–23, 2008.

[BMB08] J. BELTOWSKA, K. MUSETH et D. BREEN : Investigations of tensor voting modeling. *In WSCG'08*, 2008.

[BMD96] A. BENSRHAIR, P. MICHE et R. DEBRIE : Fast and automatic stereo vision matching algorithm based on dynamic programming method. *in Pattern Recognition Letters*, 17:457–466, 1996.

[BN97] D.N. BHAT et S.K. NAYAR : Ordinal measures for image correspondence. Rapport technique, Technical Report CUCS-009-96, Department of Computer Science, Columbia University, 1997.

[BN98] D.N. BHAT et S. K. NAYAR : Ordinal measures for image correspondence. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 20(4):415 423, April 1998.

[Bro09] R. BROCKERS : Cooperative stereo matching with color-based adaptive local support. *In Proc. of the 13th Inter. Conf. on Computer Analysis of Images and Patterns (CAIP)*, volume LNCS 5702, pages 1019–1027, 2009.

[BS95] A.J. BELL et T.J. SEJNOWSKI : An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[BS05] D.E. BUTLER et S. SRIDHARAN : Real-time adaptive foreground/background segmentation. *EURASIP journal on applied signal processing*, 14:2292–2304, 2005.

[BSA98] S. BAKER, R. SZELISKI et P. ANANDAN : A layered approach to stereo reconstruction. *In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 434–441, Juin 1998.

[BSJ03] D. BUTLER, S. SRIDHARAN et V.M. Bove JR. : Real-time adaptive background segmentation. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 349–352, 2003.

[BT80] S.T. BARNARD et W.B. THOMPSON : Disparity analysts of images. *IEEE Trans. on PAMI*, 4:333–340, July 1980.

[BT98] S. BIRCHFIELD et C. TOMASI : Depth discontinuities by pixel-to-pixel stereo. *In Proc. of the IEEE Inter. Conf. on Computer Vision*, 1998.

[BT99] S. BIRCHFIELD et C. TOMASI : Multiway cut fro stereo and motion with slanted surfaces. *In Inter. Joint Conf. on Artificial Intelligence*, volume 1, pages 489–495, 1999.

[BTG06] H. BAY, T. TUYTELAARS et L.V. GOOL : Surf : Speeded up robust features. *In 9th European Conference on Computer Vision*, pages 7–13, Graz, Autriche, Mai 2006.

[BVI98] Y. BOYKOV, O. VEKSLER et R. Zabih. . IN : Markov random fields with efficient approximations. *In IEEE Conference on CVPR*, 1998.

[BVZ98] Y. BOYKOV, O. VEKSLER et R. ZABIH : A variable window approach to early vision. *IEEE Trans. on PAMI*, 20:1283–1294, 1998.

[CAPP01] V.D. CALHOUN, T. ADALI, G.D. PEARLSON et J.J. PEKAR : Spatial and temporal independent component analysis of functional mri data containing a pair of task-related waveforms. *Hum Brain Mapp.*, 13(1):43–53, 2001.

[Car89] J.-F. CARDOSO : Blind identification of independent signals. *In In Proc.Workshop on Higher-Order Specral Analysis*, 1989.

[CC06] C.P. CHEN et C.S. CHEN : The 4-source photometric stereo under general unknown lighting. *In ECCV III*, volume LNCS Vol. 3953, pages 72–83, Mai 2006.

[CCD+07] C. CHANDRASEKARAN, V. CANON, J.C. DAHMEN, Z. KOURTZI et A.E. WELCHMAN : Neural correlates of disparity-defined shape discrimination in the human brain. *Journal of Neurophysiology*, 97:1553–1565, 2007.

[CF09] Y.-C. CHAI et B. FARELL : From disparity to depth : How to make a grating and a plaid appear in the same depth plane. *Journal of Vision*, 9(10):1–19, 2009.

[CGPP03] R. CUCCHIARA, C. GRANA, M. PICCARDI et A. PRATI : Detecting moving objects, ghosts, and shadows in video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337–1342, 2003.

[Cha05] S. CHAMBON : *Mise en correspondance stéréoscopique d'images couleur en présence d'occultations*. Thèse de doctorat, These de doctorat, Université de Toulouse III, 2005.

[CJ08] H.-S. CHUNG et J. JIA : Efficient photometric stereo on glossy surfaces with wide specular lobes. *In CVPR*, 2008.

[CLR90] T.H. CORMEN, C.E. LEISERSON et R.L. RIVEST : *Introduction to Algorithms*. The MIT Press ; first edition, 1990.

[CM02] D. COMANICIU et P. MEER : Mean shift : A robust approach toward feature space analysis. *ITTT Trans. on PAMI*, 24(5):603–619, 2002.

[CR03] S.-D. CHEN et A.-R. RAMLI : Minimum mean brightness error bi-histogram equalization in contrast enhancement. *IEEE Transactions on Consumer Electronics*, 49(4):1310–1319, 2003.

[CRF+08] B. CLIPP, R. RAGURAM, J.M. FRAHM, G. WELCH et M. POLLEFEYS : A mobile 3d city reconstruction system. *In Proc. IEEE Virtual Reality workshop on Cityscapes*, March 2008.

[Cro97] A. CROUZIL : *Perception du relief et du mouvement par analyse d'une séquence stéréoscopique d'images*. Thèse de doctorat, Thèse de doctorat, Université Paul Sabatier, UPS, Toulouse, Septembre 1997.

[CS93] J.-F CARDOSO et A. SOULOUMIAC : Blind beamforming for non-gaussian signals. *In IEE Proceedings F on Radar and Signal Processing,*, volume 140, pages 362–370, 1993.

[CS09] B. CYGANEK et J.P. SIEBERT : *An Introduction to 3D computer vision techniques and algorithms*. John Wiley & Sons, 2009.

[CVHC08] N. CAMPBELL, G. VOGIATZIS, C. HERNANDEZ et R. CIPOLLA : Using multiple hypotheses to improve depth-maps for multiview stereo. *In ECCV*, pages 766–779, 2008.

[CW81] M. COURTOIS et G. WEILL : The spot satellite remote sensing mission. *In Photogrametric Engineering and Remote Sensing*, pages 1163–1171, 1981.

[CWD03] S.O.-Y. CHAN, Y.-P. WONG et J.K. DANIEL : Dense stereo correspondence based on recursive adaptive size multi-windowing. *In Image and Vision Computing NZ*, pages 256–259, Nov 2003.

[Cyg04] B. CYGANEK : Comparison of nonparametric transformations and bit vector matching for stereo correlation. *In Combinatorial Image Analysis*, volume LNCS 3322, pages 534–547, 2004.

[Cyg05] B. CYGANEK : Adaptive window growing technique for efficient image matching. *In Pattern Recognition and Image Analysis*, volume 3522, pages 308–315, 2005.

[CZ03] Z. CHEN et X.-P. ZHANG : Video sequences processing based on spatiotemporal independent component analysis. *In Proceedings of 2003 Canadian Conference on Electrical and Computer Engineering*, 2003.

[Dar59] C.R. DARWIN : *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London John Murray, 1859.

[Dar98] T. DARREL : A radial cumulative similarity transform for robust image correspondence. *In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 656–662, 1998.

[DC07] J.D. DUROU et F. COURTEILLE : Integration of a normal field without boundary condition. *In Proc. of the 1st Int. Workshop on Photometric Analysis For Computer Vision (PACV)*, 2007.

[DC08] J.D. DUROU et F. COURTEILLE : Integration of a normal field without boundary condition. *In In Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle (RFIA)*, pages 331–340, 2008.

[Dem86] T. DEMETRI : Regularization of inverse visual problems involving discontinuities. *IEEE Trans. on PAMI*, 8:413–424, 1986.

[DGL08] M. DAROUICH, S. GUYETANT et D. LAVENIER : Architecture flexible pour la stéréovision embarquée. *In Court article pour le colloque national du GDR*, 2008.

[DJMMR01] O. DE-JOINVILLE, G. MAILLET, H. MAÎTRE et M. ROUX : Evaluation a priori de la qualité d'un mns. *In Actes du congrès francophone de Vision par Ordinateur, ORASIS*, pages 67–76, Juin 2001.

[DLQV05] A. DESROCHES, A. LEROY, J.F. QUARANTA et F. VALLEE : *Dictionnaire d'analyse et de gestion des risques*. Lavoisier, December 2005.

[DLR77] A.P. DEMPSTER, N.M. LAIRD et D.B. RUBIN : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[EEAHM07] S. EL-ETRIBY, A. AL-HAMADI et B. MICHAELIS : Desnse stereo correspondance with slanted surface using phase-based algorithm. *In IEEE International Symposium on Indistrual Electronics*, pages 1807–1813, 2007.

[EESA08] S.Y. ELHABIAN, K.M. EL-SAYED et S.H. AHMED : Moving object detection in spatial domain using background removal techniques - state-of-art. *Recent Patents on Computer Science*, 1(1):32–54, 2008.

[Egn00] G. EGNAL : Mutual information as a stereo correspondence measure. Rapport technique, Technical Report MS-CIS-00-20, Comp. and Inf. Science, Univ. of Pennsylvania, 2000.

[EHD00] A. ELGAMMAL, D. HARWOOD et L. DAVIS : Non-parametric model for background subtraction. *In Proceedings of the 6th European Conference on Computer Vision (ECCV) -Part II*, June 2000.

[EMW04] G. EGNAL, M. MINTZ et R.P. WILDES : A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image and Vision Computing, ELSEVIER*, 22:943–957, 2004.

[ESD96] R. EBERHART, P. SIMPSON et R. DOBBINS : *Computational intelligence PC tools*. Academic Press Professional, San Diego, CA, USA, 1996.

[FCF10] B. FARELL, Y.-C. CHAI et J.M. FERNANDEZ : The horizontal disparity direction vs. the stimulus disparity direction in the perception of the depth of two-dimensional patterns. *Journal of Vision*, 10(4):1–15, 2010.

[FCSS09] Y. FURUKAWA, B. CURLESS, S.M. SEITZ et R. SZELISKI : Manhattan-world stereo. *In CVPR*, pages 1422–1429, 2009.

[FH06] P.F. FELZENSZWALB et D.P. HUTTENLOCHER : Effcient belief propagation for early vision. *International Journal of Computer Vision (IJCV)*, 70(1):41–54, 2006.

[FJLV07] P. FOGGIA, J.-M. JOLION, A. LIMONGIELLO et M. VENTO : Stereo vision for obstacle detection : A graph-based approach. *In Graph-Based Representations in Pattern Recognition*, volume 4538, pages 37–48, 2007.

[FK10] R. FABBRI et B. KIMIA : 3d curve sketch - flexible curve-based stereo reconstruction and calibration. *In CVPR*, 2010.

[FL95] P. FUA et Y. LECLERC : Object-centered surface reconstruction. *IJCV*, 16:35–56, 1995.

[FLM04] B. FARELL, S. LI et S.P. MCKEE : Coarse scales, fine scales, and their interactions in stereo vision. *Journal of Vision*, 4:488–499, 2004.

[For98] G.L. FORESTI : A real-time system for video surveillance of unattended outdoor environments. *IEEE Trans. on Circuits Syst. Video Technol.*, 8(6):697–704, 1998.

[FP07] Y. FURUKAWA et J. PONCE : Accurate, dense, and robust multi-view stereopsis. *In CVPR*, 2007.

[FP08] Y. FURUKAWA et J. PONCE : Accurate, dense, and robust multi-view stereopsis. *IEEE Trans. on PAMI*, 1(1):2008, August 2008.

[FPC+09] J.-M. FRAHM, M. POLLEFEYS, B. CLIPP, D. GALLUP, R. RAGURAM, C.C. WU et C. ZACH : 3d reconstruction of architectural scenes from uncalibrated video sequences. *In 3D Virtual Reconstruction and Visualization of Complex Architectures (3D-ARCH)*, February 2009.

[FR97] N. FRIEDMAN et S. RUSSELL : Image segmentation in video sequences : A probabilistic approach. *In Thirteenth Conference on Uncertainty in Artificial Intelligence*, volume 94720, pages 175–181, 1997.

[FR00] A. FUSIELLO et V. ROBERTO : Symmetric stereo with multiple windowing. *Inter. Journal of Pattern Recognition and Artificial Intelligence*, 14(8):1053–1066, 2000.

[FRT97] A. FUSIELLO, V. ROBERTO et E. TRUCCO : Efficient stereo with multiple windowing. *In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 858–863, 1997.

[FTVV00] A. FUSIELLO, E. TRUCCO, A. VERRI et R. VERRI : A compact algorithm for rectification of stereo pairs. *In Machine Vision and Applications*, volume 12, pages 16–22, 2000.

[FYO+04] S. FORSTMANN, Y.KANOU, J. OHYA, S. THUERING et A. SCHMITT : Real-time stereo by using dynamic programming. *In CVPR*, pages 29–35, 2004.

[FZ10] P.F. FELZENSZWALB et R. ZABIH : Dynamic programming and graph algorithms in computer vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, PP:1–51, 2010.

[GC10] R.K. GUPTA et S.-Y. CHO : Real-time stereo matching using adaptive binary window. *In Inter. Symposium 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010.

[GDHW99] G. GORDON, T. DARRELL, M. HARVILLE et J. WOODFILL : Background estimation and removal based on range and color. *In IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 459–464, 1999.

[GG84] S. GEMAN et D. GEMAN : Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 6:721–741, 1984.

[GL03] J. Y. GOULERMAS et P. LIATSIS : A collective-based adaptive symbiotic model for surface reconstruction in area-based stereo. *IEEE Trans. on Evolutionary Computation*, 7(5):482–502, 2003.

[GLY92] D. GEIGER, B. LADENDORF et A. YUILLE : Occlusions and binocular stereo. *In Proc. on European Conf. Computer Vision (ECCV)*, pages 425–433, 1992.

[GM79] W.E.L. GRIMSON et D. MARR : A computer implementation of a theory of human stereo vision. *In Proc. Image Understanding Workshop, Science Applications*, pages 41–47, Apr 1979.

[Gri04] E. GRIFFIOEN : Improving level crossings using findings from human behaviour studies. *In Proc. of 8th Inter. Level Crossing Symposium*, Sheffield, Storbritannien, 2004.

[Gro73] S. GROSSBERG : Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 52:213–257, 1973.

[GVW04] S. GREENHILL, S. VENKATESH et G. WEST : Adaptive model for foreground extraction in adverse lighting conditions. *In Pacific Rim international conference on artificial intelligence*, volume 3157, pages 805–811, 2004.

[GY02] M. GONG et Y.-H. YANG : Genetic-based stereo algorithm and disparity map evaluation. *International Journal of Computer Vision (IJCV)*, 47(1-3):63–77, Avril 2002.

[GY03] M. GONG et Y.H. YANG : Fast stereo matching using reliability-based dynamic programming and consistency constraints. *In IEEE 9th ICCV*, volume 1, pages 610–617, 2003.

[HA84] J. HERAULT et B. ANS : Circuits neuronaux à synapses modifiables : décodage de messages composites par apprentissage non supervisé. *In C.-R. de l'Académie des Sciences*, volume 299(III-13), pages 525–528, 1984.

[HA89] W. HOFF et N. AHUJA : Surfaces from stereo : Integrating feature matching, disparity estimation, and contour detection. *IEEE Trans. on PAMI*, 11:121–136, 1989.

[Han80] M.J. HANNAH : Bootstrap stereo. *In in Proc. on Image Understanding Workshop (College Park, Md.,)*, pages 201–208, 1980.

[HB10] D.M. HOFFMAN et M.S. BANKS : Focus information is used to interpret binocular images. *Journal of Vision*, 10(5):1–17, 2010.

[HBG10] A. HOSNI, M. BLEYER et M. GELAUTZ : Near real-time stereo with adaptive support weight approaches. *In Inter. Symposium 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010.

[HBGR09] A. HOSNI, M. BLEYER, M. GELAUTZ et C. RHEMANN : Local stereo matching using geodesic support weights. *In 16th IEEE ICIP*, pages 2093–2096, 2009.

[HC05] D.A. HINKLE et C.E. CONNOR : Quantitative characterization of disparity tuning in ventral pathway area v4. *Journal of Neurophysiology*, 94:2726–2737, 2005.

[HGW01] M. HARVILLE, G. GORDON et J. WOODFILL : Foreground segmentation using adaptive mixture models in color and depth. *In IEEE Workshop on Detection and Recognition of Events in Video*, pages 1–13, 2001.

[HHC+02] C. HAVRAN, L. HUPET, J. CZYZ, J. LEE, L. VANDENDORPE et M. VEREYSEN : Independent component analysis for face authentification. *In Knowledge-Based Intelligent Information and Engineering Systems*, pages 1207–1211, 2002.

[HHD00] I. HARITAOGLU, D. HARWOOD et LS DAVIS : W4 :real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 2000.

[HIG02] H. HIRSCHMULLER, P.R. INNOCENT et J. GARIBALDI : Real-time correlation-based stereo vision with reduced border errors. *IJCV*, 47(1-3):229–246, 2002.

[Hir05] H. HIRSCHMULLER : Accurate and efficient stereo processing by semi-global matching and mutual information. *In In Proc. CVRP*, volume 2, pages 807–814, June 2005.

[HJ86] J. HERAULT et C. JUTTEN : Space or time adaptive signal processing by neural networks model. *In in Proc. Of International Conference on Neural Networks for Computing*, pages 206–211, April 1986.

[HKLP09] V.H. HIEP, R. KERIVEN, P. LABATUT et J.-P. PONS : Towards high-resolution large-scale multi-view stereo. *In CVPR*, pages 1430–1437, 2009.

[HKO01] A. HYVARINEN, J. KARHUNEN et E. OJA : *Independent Component Analysis*. JOHN WILEY & SONS, INC, 2001.

[HL06] P.D. L. HOWE et M.S. LIVINGSTONE : V1 partially solves the stereo aperture problem. *Journal of Cerebral Cortex*, 16:332–337, 2006.

[HMJI09] T. HIGO, Y. MATSUSHITA, N. JOSHI et K. IKEUCHI : A hand-held photometric stereo camera for 3-d modeling. *In ICCV*, 2009.

[HO97] A. HYVARINEN et E. OJA : A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[Hol75] J. HOLLAND : *Adaptation in natural and artificial systems : An introductory analysis with applications to biology, control and, artificial intelligence*. The MIT Press, 1975.

[HP06] M. HEIKKILA et M. PIETIKAINEN : A texture-based method for modeling the background and detecting moving objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(4):657–662, 2006.

[HPHH08] K. HUHA, J. PARK, J HWANG et D. HONG : A stereo vision-based obstacle detection system in vehicles. *Optics and Lasers in Engineering*, 46:168–178, Feb 2008.

[HPS09] M. HEIKKILA, M. PIETIKAINEN et C. SCHMID : Description of interest regions with local binary patterns. *Pattern Recognition*, 42(3):425–436, 2009.

[HRK03] M. HARITI, Y. RUICHEK et A. KOUKAM : A fast stereo matching method for obstacle detection in front of a vehicle. *In Colloque TILT, Lille*, 2003.

[HS07] H. HIRSCHMULLER et D. SCHARSTEIN : Evaluation of cost functions for stereo matching. *In CVPR*, pages 1–8, 2007.

[HSC+01] K.-P. HAN, K.-W. SONG, E.-Y. CHUNG, S.-J. CHO et Y.-H. HA : Stereo matching using genetic algorithm with adaptive chromosomes. *The Journal of the Pattern Recognition Society, PR*, 34:1729–1940, Sept 2001.

[Hua07] X. HUANG : Cooperative optimization for energy minimization in computer vision : A case study of stereo matching. *In CVPR*, 2007.

[HVB+07] C. HERNANDEZ, G. VOGIATZIS, G.J. BROSTOW, B. STENGER et R. CIPOLLA : Non-rigid photometric stereo with colored lights. *In ICCV*, 2007.

[HVC07] C. HERNANDEZ, G. VOGIATZIS et R. CIPOLLA : Probabilistic visibility for multi-view stereo. *In CVPR*, pages 1–8, 2007.

[HVC08a] C. HERNANDEZ, G. VOGIATZIS et R. CIPOLLA : Multi-view photometric stereo. *IEEE Trans. on PAMI*, 30:548–554, March 2008.

[HVC08b] C. HERNANDEZ, G. VOGIATZIS et R. CIPOLLA : Shadows in three-source photometric stereo. *In ECCV*, 2008.

[Hyv99] A. HYVARINEN : Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

[IM06] M. ISARD et J. MCCORMICK : Dense motion and disparity estimation via loopy belief propagation. *In Asian Conference on Computer Vision (ACCV)*, pages 23–41, 2006.

[IRP05] H. ISSA, Y. RUICHEK et J.G. POSTAIRE : A specific encoding scheme for genetic stereo correspondence searching : Application to obstacle detection. *Journal of Systemics, Cybernetics and Informatics*, 1(1):9–17, 2005.

[JDWR00] S. JABRI, Z. DURIC, H. WECHSLER et A. ROSENFELD : Detection and location of people in video images using adaptive fusion of color and edge information. *In International Conference on Pattern Recognition*, volume 4, pages 4627–4630, 2000.

[JOK09] L. JANSEN, S. ONAT et P. KONIG : Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1):1–19, 2009.

[JSLKS05] Y. J. SUN, LI, S.B. KANG et H.-Y. SHUM : Symmetric stereo matching for occlusion handling. *In CVPR*, 2005.

[JT10] V. JOHN et E. TRUCCO : Multiple view human articulated tracking using charting and particle swarm optimisation. *In Proc. of the 1st inter. workshop on 3D video processing 3DVP*, 2010.

[KC06] J.H. KIM et M.J. CHUNG : Absolute motion and structure from stereo image sequences without stereo correspondence and analysis of degenerate cases. *IEEE Trans. on PAMI*, 39:1649–1661, 2006.

[KCHD05] K. KIM, T. H. CHALIDABHONGSE, D. HARWOOD et L. DAVIS : Real-time foreground-background segmentation using codebook model. *Real-Time Imaging, Special Issue on Video Object Processing*, 11:172–185, June 2005.

[KE95] J. KENNEDY et R. EBERHART : Particle swarm optimization. *In Proc. of the 1995 IEEE Inter. Conf. on Neural Networks*, pages 1942–1948, 1995.

[KGB+09] L. KHOUDOUR, M. GHAZEL, F. BOUKOUR, M. HEDDEBAUT et E.-M. EL-KOURSI : Towards safer level crossings : existing recommendations, new applicable technologies and a proposed simulation model. *European Transport Research Review*, 1(1):35–45, 2009.

[KHD10] U. KIRCHMAIER, S. HAWE et K. DIEPOLD : Dynamical information fusion of heterogeneous sensors for 3d tracking using particle swarm optimization. *Journal of Information Fusion*, In Press:1–9, 2010.

[KKB07a] K. KOLEV, M. KLODT et T. BROX : Continuous global optimization in multiview 3d reconstruction. *IJCV*, 84(1):80–96, 2007.

[KKB+07b] K. KOLEV, M. KLODT, T. BROX, S. ESEDOGLU et D. CREMERS : Continuous global optimization in multiview 3d reconstruction. *In CVPR*, 2007.

[KKD08] F. KEYROUZ, U. KIRCHMAIER et K. DIEPOLD : Three dimensional object tracking based on audiovisual fusion using particle swarm optimization. *In 11th International Conference on Information Fusion*, pages 1–5, Cologne, 2008.

[KKZ03] J. KIM, V. KOLMOGOROV et R. ZABIH : Visual correspondence using energy minimization and mutual information. *In In Proc. ICCV*, 2003.

[KLCL05] J.C. KIM, K.M. LEE, B.T. CHOI et S.U. LEE : A dense stereo matching using two-pass dynamic programming with generalized ground control points. *In CVPR*, volume 2, pages 1075–1082, 2005.

[KN09] L. KRATZ et K. NISHINO : Factorizing scene albedo and depth from a single foggy image. *In ICCV*, 2009.

[KO94] T. KANADE et M. OKUTOMI : A stereo matching algorithm with adaptive window - theory and experiment. *IEEE Trans. on PAMI*, 16(9):920–932, Sept 1994.

[KOY00] M. KUMANO, A. OHYA et S. YUTA : Obstacle avoidance of autonomus mobile robot using stereo sensor. *In Second International Symposium on Robotic and Automation ISRA*, pages 497–502, 2000.

[KPC10] K. KOLEV, T. POCK et D. CREMERS : Anisotropic minimal surfaces integrating photoconsistency and normal information for multiview stereo. *In ECCV*, 2010.

[KRW09] S. KUMAR, B. RAMAN et J. WU : Neuro-calibration of a camera using particle swarm optimization. *In Proc. of the 2009 Second Inter. Conf. on Emerging Trends in Engineering & Technology ICETET*, 2009.

[KS80] R. KINDERMANN et J.L. SNELL : *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.

[KS06] M. KOLOMENKIN et I. SHIMSHONI : Image matching using photometric information. *In CVPR*, pages 2506–2514, Octobre 2006.

[KSC01] S. B. KANG, R. SZELISKI et J. CHAI : Handling occlusions in dense multi-view stereo. *In CVPR*, volume 1, pages 103–110, 2001.

[LAC04] C. LEUNG, B. APPLETON et C.S.. CHANGMING : Fast stereo matching by iterated dynamic programming and quadtree subregioning. *In British Machine Vision Conference*, 2004.

[LBC08] Y. LIU, A.C. BOVIK et L.K. CORMACK : Disparity statistics in natural scenes. *Journal of Vision*, 8(11):1–14, 2008.

[LCCB01] L.J. LUO, D.R. CLEWER, C.N. CANAGARAJAH et D.R. BULL : Genetic stereo matching using complex conjugate wavelet pyramids. *In ICIP*, volume 2, pages 153–156, 2001.

[Lef08] S. LEFEBVRE. : *Approche monodimensionnelle de la mise en correspondance stéréoscopique par corrélation - Application à la détection d'obstacles routiers*. Thèse de doctorat, Thèse de doctorat de l'Ecole Centrale de Lille 1, 2008.

[LG94] J.-L. LOTTI et G. GIRAUDON : Adaptive window algorithm for aerial image stereo. *In Proc. of the 12th ICPR*, pages 701–703, 1994.

[LH07] L. LU et G. HAGER : *Dynamic foreground/background extraction from images and videos using random patches*. The Massachusetts Institute of Technology (MIT) press, 2007.

[LHYJ09] Z. LIU, Z. HAN, Q. YE et J. JIAO : A new segment-based algorithm for stereo matching. *In Int. Conf. on Mechatronics and Automation*, pages 999–1003, 2009.

[LLL+02] Y. LI, S. LIN, H. LU, S.-B. KANG et H.-Y. SHUM : Multibaseline stereo in the presence of specular reflections. *In Proc. of the 16th ICPR'02*, volume 3, pages 573–576, 2002.

[LMC10] Y. LIN, G. MEDIONI et J. CHOI : Accurate 3d face reconstruction fromweakly calibrated wide baseline images with profile contours. *In CVPR*, 2010.

[Low99] D.G. LOWE : Object recognition from local scale-invariant features. *In Proc. of the Inter. Conf. on Computer Vision*, volume 2, pages 1150–1157, 1999.

[LP06a] H.S. LIM et H. PARK : A dense disparity estimation method using color segmentation and energy minimization. *In ICIP*, pages 1033–1036, 2006.

[LP06b] R. LUKAS et K.N. PLATANIOTIS : *Color Image Processing : Methods and Applications*. Boca Raton, FL, CRC Press / Taylor & Francis, 2006.

[LR94] J.H. LEE et N.L. RICKER : Extended kalman filter based nonlinear model predictive control. *In Industrial and Engineering Chemistry Research*, volume 33, pages 1530–1541, 1994.

[LSY06] C. LEI, J. SELZER et Y.H. YANG : Region-tree based stereo using dynamic programming optimization. *In CVPR*, pages 2378 – 2385, 2006.

[LV01] B.P.L. LO et S.A. VELASTIN : Automatic congestion detection system for underground platforms. *In International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 158–161, 2001.

[LY10] J. LEVEILLE et A. YAZDANBAKHSH : Speed, more than depth, determines the strength of induced motion. *Journal of Vision*, 10(6):1–9, 2010.

[MAW+07] P. MERRELL, A. AKBARZADEH, L. WANG, P. MORDOHAI, J.-M. FRAHM, R. YANG, D. NISTER et M. POLLEFEYS : Real-time visibility-based fusion of depth maps. *In ICCV*, 2007.

[May03] H. MAYER : Analysis of means to improve cooperative disparity estimation. *In In International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume 34, pages 25–31, Sept 2003.

[MG04] D. MARKOVIC et M. GELAUTZ : Experimental combination of intensity and stereo edges for improved snake segmentation. *In n 7th International Conference on Pattern Recognition and Image Analysis*, pages 18–23, 2004.

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

[mis] http ://mars.jpl.nasa.gov/.

[MJaB+97] S. MAKEIG, T.P. JUNG, a.J. BELL, D. GHAHREMANI et T.J. SEJNOWSKI : Blind separation of auditory event-related brain responses into independent components. *In Proc Natl Acad Sci*, volume 94, pages 10979–10984, 1997.

[MM06] P. MORDOHAI et G. MEDIONI : Stereo using monocular cues within the tensor voting framework. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 28(6):968–982, 2006.

[MMG06] V. MARIE, B. MARCUS et S. GEOFF : Red light violations. *In International Level Crossing Symposium*, 2006.

[MMI09] D. MIYAZAKI, Y. MATSUSHITA et K. IKEUCHI : Interactive shadow removal from a single image using hierarchical graph cut. in : Asian conference on computer vision. *In ACCV*, 2009.

[MMN89] R. MOHAN, G. MEDIONI et R. NEVATIA : Stereo error detection, correction, and evaluation. *IEEE Trans. on PAMI*, 11(2):113–120, 1989.

[MNV09] M. MAHAJAN, P. NIMBHORKAR et K. VARADARAJAN : The planar k-means problem is np-hard. *In WALCOM '09 Proceedings of the 3rd International Workshop on Algorithms and Computation*, pages 274–285, 2009.

[Mor79] H. MORAVEC : Visual mapping by a robot rover. *In in Proc. 6th Inter. Joint Conf. on Artificial Intell. (IJCAI'79)*, volume 1, pages 598–600, Aug 1979.

[MP76] D. MARR et T. PAGGIO : Cooperative computation of stereo disparity. *Science, New Series*, 194(4262):283–287, 1976.

[MP79] D. MARR et T. POGGIO : A computational theory of human stereo vision. *In Proceedings of the Royal Society of London. Series B, Biological Sciences*, volume 204, pages 301–328. Proceedings of the Royal Society of London. Series B, Biological Sciences, 1979.

[MPP09] W. MILED, J.C. PESQUET et M. PARENT : A convex optimization approach for depth estimation under illumination variation. *IEEE Trans. on Image Processing*, 18(4):813–830, Avril 2009.

[MS95] N.J.B. MCFARLANE et CP SCHOFIELD : Segmentation and tracking of piglets in images. *Machine Vision and Applications*, 8(3):187–193, 1995.

[MS09] Y. MOSES et I. SHIMSHONI : 3d shape recovery of smooth surfaces dropping the fixed viewpoint assumption. *IEEE Trans. on PAMI*, 31:1310 – 1324, July 2009.

[MT10] S.P. MCKEE et D.G. TAYLOR : The precision of binocular and monocular depth judgments in natural settings. *Journal of Vision*, 10(10):1–13, 2010.

[MVF05] S.P. MCKEE, P. VERGHESE et B. FARELL : Stereo sensitivity depends on stereo matching. *Journal of Vision*, 5:783–792, 2005.

[MVPG02] G.V. MEERBERGEN, M. VERGAUWEN, M. POLLEFEYS et L.V. GOOL : A hierarchical symmetric stereo algorithm using dynamic programming. *IJCV*, 47:275–285, 2002.

[MYS06] D. MIN, S. YOON et K. SOHN : Segment-based stereo matching using energy-based regularization. *In Multimedia Content Representation, Classification and Security*, volume 4105, pages 761–768, 2006.

[NB09] G. NIERADKA et B. BUTKIEWICZ : Features stereo matching based on fuzzy logic. *In Int. Fuzzy Systems Association - European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT)*, 2009.

[Nel02] A. NELSON : The uk approach to managing risk at passive level crossings. *In Inter. Symposium on RailRoad-Highway Grade Crossing Research and Safety, 7th*, Melbourne, Victoria, Australia, 2002.

[Ner05] P. NERI : A stereoscopic look at visual cortex. *Journal of Neurophysiology*, 93: 1823–1826, 2005.

[Oht05] M. OHTA : Level crossing obstacle detection system using stereo camera. *Quarterly Report of Railway Technical Research Institute (RTRI)*, 46(2):110–117, 2005.

[OK85a] Y. OHTA et T. KANADE : Stereo by intra- and inter- scanline search using dynamic programming. *IEEE Trans. on PAMI*, 7(2):139–154, March 1985.

[OK85b] Y. OHTA et T. KANADE : Stereo by two-level dynamic programming. *In Proc. of the 9th inter. joint conference on Artificial intelligence*, volume 2, pages 1120–1126, 1985.

[OK92] M. OKUTOMI et T. KANADE : A locally adaptive window for signal matching. *IJCV*, 7(2):143–162, 1992.

[OK93] M. OKUTOMI et T. KANADE : A multiple baseline stereo. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 15(4):353–363, April 1993.

[OKO02] M. OKUTOMI, Y. KATAYAMA et S. OKA : A simple stereo algorithm to recover precise object boundaries and smooth surfaces. *Inter. Journal on Computer Vision (IJCV)*, 47(1-3):261–273, 2002.

[OLC93] M. OTT, J. LEWIS et I.J. COX : Teleconferencing eye contact using a virtual camera. *In Interchi'93 Adjunct Proceedings*, 1993.

[ORP00] N.M. OLIVER, B. ROSARIO et A.P. PENTLAND : A bayesian computer vision system for modeling human interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.

[Pan] http ://pansafer.inrets.fr/.

[PCBC10] T. POCK, D. CREMERS, H. BISCHOF et A. CHAMBOLLE : Global solutions of variational models with convex regularization. *SIM Journal on Imaging Sciences*, 3(4):1122–1145, 2010.

[PCF06] N. PARAGIOS, Y. CHEN et O. FAUGERAS : *Handbook of Mathematical Models in Computer Vision*. Springer, 2006.

[PD96] D.V. PAPADIMITRIOU et T.J. DENNIS : Epipolar line estimation and rectification for stereo image pairs. *IEEE Trans. on Image Processing*, 5:672–676, 1996.

[PGG+10] S. PALMISANO, B. GILLAM, D.G. GOVAN, R.S.ALLISON et J.M. HARRIS : Stereoscopic perception of real depths at large distances. *Journal of Vision*, 10(6):1–16, 2010.

[PNF+08] M. POLLEFEYS, D. NISTER, J.-M. FRAHM, A. AKBARZADEH, P. MORDOHAI, B. CLIPP, C.ENGELS, D. GALLUP, S.-J. KIM, P. MERRELL, C. SALMI, S. SINHA, B. TALTON, L. WANG, Q. YANG, H. STEWENIUS, R. YANG, G. WELCH et H. TOWLES : Detailed real-time urban 3d reconstruction from video. *IJCV*, 78(2-3):134–167, 2008.

[Pra85] K. PRAZDNY : Detection of binocular disparities. *In Biomedical and Life Sciences*, volume 52, pages 93–99, 1985.

[PTK85] T. POGGIO, V. TORRE et C. KOCH : Computational vision and regularization therory. *Journal Nature*, 317:314–319, 1985.

[PU07] M. PAULINAS et A. USINSKAS : A survey of genetic algorithms applications for image enhancement and segmentation. *Information Technology and Control*, 36(3): 278–284, 2007.

[RGS03] C. RAO, A. GRITAI et M. SHAH : View-invariant alignment and matching of video sequences. *In IEEE Inter. Conf. on Computer Vision (ICCV)*, pages 939–945, Oct 2003.

[RKJB00] J. RITTSCHER, J. KATO, S. JOGA et A. BLAKE : A probabilistic background model for tracking. *In European Conference on Computer Vision*, pages 336–350, 2000.

[RM02] M. RIVERA et J.L. MARROQUIN : Adaptive rest condition potentials - second order edge-preserving regularization. *In In Proceedings of the ECCV*, volume 1, pages 113–127, 2002.

[RN10] et A.J. Davison R.A. NEWCOMBE : Live dense reconstruction with a single moving camera. *In CVPR*, 2010.

[RPBD07] A.W. ROE, A.J. PARKER, R.T. BORN et G.C. DEANGELIS : Disparity channels in early vision. *Journal of Neuroscience*, 27(44):11820–11831, 2007.

[RSC04] C. RABAUD, O. STRAUSS et F. COMBY : Fuzzy dense stereoscopic matching. *In Rencontres Francophones sur la Logique Floue et ses Application (LFA)*, 2004.

[RSG06] A.M. RAUSCHECKER, S.G. SOLOMON et A. GLENNERSTER : Stereo and motion parallax cues in human 3d vision : Can they vanish without a trace ? *Journal of Vision*, 6:1471–1485, 2006.

[SAHK09] L.T. SACH, K. ATSUTA, K. HAMAMOTO et S. KONDO : A robust road profile estimation method for low texture stereo images. *In ICIP*, pages 4273–4276, 2009.

[Sch99] D. SCHARSTEIN : *View synthesis using stereo vision*, volume 1583. Springer, 1999.

[SD10] G. SCHINDLER et F. DELLAERT : Probabilistic temporal inference on reconstructed 3d scenes. *In CVPR*, 2010.

[Sel] http ://www.iva.ing.tu-bs.de/levelcrossing/selcat/.

[SG98] R. SZELISKI et P. GOLLAND : Stereo matching with transparency and matting. *In Proceedings of the Sixth (ICCV'98)*, pages 517–524, Jan 1998.

[SG99] C. STAUFFER et W. GRIMSON : Adaptive background mixture models for real-time tracking. *In IEEE International Conference on Computer Vision and Pattern Recognition*, pages 246–52, 1999.

[SG00] C. STAUFFER et WEL GRIMSON : Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 747–757, 2000.

[SHT+06] S. SULAIMAN, A. HUSSAIN, N.M. TAHIR, A.M. MUAD et M.M. MUSTAFA : Scene analysis for human silhouette extraction. *In 4th Student Conference on Research and Development, SCOReD*, pages 124–126, Selangor, Malaysia, 2006.

[SK04] A. SMALL et R. KENNEDY : Road vehicle level crossings. Special topic report, Departmental Head of Safety Strategy and Risk Rail Safety and Standards Board (DHSS and RSSB), London, January 2004.

[SK07] X. SU et T.M. KHOSHGOFTAAR : A progressive edge-based stereo correspondence method. *In ISVC'07 Proceedings of the 3rd international conference on Advances in visual computing*, volume LNCS 4841, pages 248–257, 2007.

[SM95] H. SAITO et M. MORI : Application of genetic algorithms to stereo matching of images. *Pattern Recognition Letters (PRL)*, 16:815–821, Aug 1995.

[SMM08] H. SADEGHI, P. MOALLEM et S.A. MONADJEMI : Feature based dense stereo matching using dynamic programming and color. *International Journal of Engineering and Mathematical Sciences*, 3(4):179–186, 2008.

[SMP07] S.N. SINHA, P. MORDOHAI et M. POLLEFEYS : Multi-view stereo via graph cuts on the dual of an adaptive tetrahedral mesh. *In ICCV*, 2007.

[SO06] T.M. SANADA et I. OHZAWA : Encoding of three-dimensional surface slant in cat visual areas 17 and 18. *Journal of Neurophysiology*, 95:2768–2786, 2006.

[SR98] et I. Cox S. ROY : A maximum-flow formulation of the n-camera stereo correspondance problem. *In ICCV*, pages 492–499, 1998.

[SRPB01] B. STENGER, V. RAMESH, N. PARAGIOS et F. Coetzee & BUHMANN : Topology free hidden markov models : application to background modeling. *In Proceedings International Conference On Computer Vision*, 2001.

[SS02] D. SCHARSTEIN et R. SZELISKI : A taxonomy and evaluation of dense two-frame stereo correspondance agorithms. *IJCV*, 47(1-2-3):7–42, 2002.

[SS05] Y. SHEIKH et M. SHAH : Bayesian modeling of dynamic scenes for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(11): 1778–1792, 2005.

[SSE+09] J. SALMEN, M. SCHLIPSING, J. EDELBRUNNER, S. HEGEMANN et S. LÜKE : Real-time stereo vision making more out of dynamic programming. *In Computer Analysis of Images and Patterns*, volume LNCS 5702, pages 1096–1103, 2009.

[SSHN09] Y. SWIRSKI, Y.Y. SCHECHNER, B. HERZBERG et S. NEGAHDARIPOUR : Stereo from flickering caustics. *In ICCV*, 2009.

[SSS09] S.N. SINHA, D. STEEDLY et R. SZELISKI : Piecewise planar stereo for image-based rendering. *In ICCV*, 2009.

[STG03] C. STRECHA, T. TUYTELAARS et L.Van GOOL : Dense matching of multiple wide-baseline views. *In ICCV*, pages 1194–1201, 2003.

[Sto99] J. STONE : Spatial, temporal, and spatiotemporal independent component analysis of fmri data. *In Proceedings of the 18th Leeds Statistical Research Workshop on Spatial-Temporal Modeling and Its Applications*, pages 23–28, 1999.

[SZ98] C. SCHMID et A. ZISSERMAN : The geometry and matching of curves in multiple views. *In in Proc. Eur. Conf. Comput. Vis. (ECCV)*, pages 394–409, Jun 1998.

[SZ09] Z.-M. SHAO et J.-Y. ZHU : Dense disparity map estimation using pso algorithm with adaptive hierarchical images. *Journal of Image and Graphics*, 4:Unknown, 2009.

[SZJ09] B.M. SMITH, L. ZHANG et H. JIN : Stereo matching with nonparametric smoothness priors in feature space. *In IEEE Conf. on CVPR*, pages 1–8, 2009.

[SZS+06] R. SZELISKI, R. ZABIH, D. SCHARSTEIN, O. VEKSLER, V. KOLMOGOROV, A. AGARWALA, M. TAPPEN et C. ROTHER : A comparative study of energy minimization methods for markov random fields. *In Proc. ECCV*, 2006.

[TF03] M.F. TAPPEN et W.T. FREEMAN : Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. *In Proc. ICCV*, 2003.

[THM98] N. TSUMURA, H. HANEISHI et Y. MIYAKE : Independent component analysis of skin color image. *In Sixth Color Imaging Conference : Color Science, Systems and Applications*, pages 177–180, 1998.

[TKBM99] K. TOYAMA, J. KRUMM, B. BRUMITT et B. MEYERS : Wallflower : Principles and practice of background maintenance. *In International Conference on Computer Vision*, pages 29–35, 1999.

[TKWB08] T. TEICHERT, S. KLINGENHOEFER, T. WACHTLER et F. BREMMER : Depth perception during saccades. *Journal of Vision*, 8(14):1–13, 2008.

[TL09] D.M. TSAI et S.C. LAI : Independent component analysis-based background subtraction for indoor surveillance. *IEEE Trans. on Image Processing*, 18(1):158–167, 2009.

[TM08] H. TRINH et D. MCALLESTER : Particle-based belief propagation for structure from motion and dense stereo vision with unknown camera constraints. *In Proceedings of the 2nd international conference on Robot vision*, pages 16–28, 2008.

[TS00] H. TAO et H.S. SAWHNEY : Global matching criterion and color segmentation based stereo. *In 5th IEEE Workshop on Applications of Computer Vision*, pages 246 – 253, 2000.

[Tsa86] R.Y. TSAI : An efficient and accurate camera calibration technique for 3d machine vision. *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 364–374, Miami Beach, FL, 1986.

[TTJ+00] M. TAIRA, K.-I. TSUTSUI, M. JIANG, K. YARA et H. SAKATA : Parietal neurons represent surface orientation from the gradient of binocular disparity. *Journal of Neurophysiology*, 83(5):3140–3146, 2000.

[TWA10] I. TSIRLIN, L.M. WILCOX et R.S. ALLISON : Monocular occlusions determine the perceived shape and depth of occluding surfaces. *Journal of Vision*, 10(6):1–12, 2010.

[TWZ08] Y. TAGUCHI, B. WILBURN et C.L. ZITNICK : Stereo reconstruction with mixed pixels using adaptive over-segmentation. *In CVPR*, pages 1–8, 2008.

[Vek02] O. VEKSLER : Dense features for semi-dense stereo correspondenc. *IJCV*, 47(1-3):247–260, 2002.

[Vek03] O. VEKSLER : Fast variable window for stereo correspondence using integral images. *In Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 556–561, 2003.

[VG08] F. VERBIEST et L.V. GOOL : Photometric stereo with coherent outlier handling and confidence estimation. *In CVPR*, 2008.

[VHTC07] G. VOGIATZIS, C. HERNÁNDEZ, P. H. S. TORR et R. CIPOLLA : Multi-view stereo via volumetric graph-cuts and occlusion robust photo-consistency. *IEEE Trans. on PAMI*, 29:2241–2246, 2007.

[WADP97] CR WREN, A. AZARBAYEJANI, T. DARRELL et AP PENTLAND : Pfinder : Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.

[WB09] J. WATTAM-BELL : Stereo and motion dmax in infants. *Journal of Vision*, 9(6):1–9, 2009.

[WEEW10] D.A. WISMEIJER, C.J. ERKELENS, R.v. EE et M. WEXLER : Depth cue combination in spontaneous eye movements. *Journal of Vision*, 10(6):1–15, 2010.

[Wes09] G. WESTHEIMER : The third dimension in the primary visual cortex. *Journal of Physiology*, 12:2807–2816, 2009.

[WHVW10] R.F.v.d. WILLIGEN, W.M. HARMENING, S. VOSSEN et H. WAGNER : Disparity sensitivity in man and owl : Psychophysical evidence for equivalent perception of shape-from-stereo. *Journal of Vision*, 10(1):1–11, 2010.

[WKI09] Y. WATANABE, T. KOMURO et M. ISHIKAWA : High-resolution shape reconstruction from multiple range images based on simultaneous estimation of surface and motion. *In ICCV*, 2009.

[WKS04] L. WANG, S.B. KANG et H.-Y. SHUM : Cooperative segmentation and stereo using perspective space search. *In Proc. Asian Conf. Computer Vision*, volume 1, pages 366–371, 2004.

[WLGRY06] L. WANG, M. LIAO, M. GONG et D. Nister R. YANG : High-quality real-time stereo using adaptive cost aggregation and dynamic programming. *In Proc. of the Third Int. Symposium on 3D Data Processing, Visualization, and Transmission*, 2006.

[WMK+08] J. WANG, T. MIYAZAKI, H. KOIZUMI, M. IWATA, J. CHONG, H. YAGYU, H. SHIMAZU, T. IKENAGA et S. GOTO : Rectangle region based stereo matching for building reconstruction. *Journal of Ubiquitous Convergence Technology*, 1(1):9–17, 2008.

[WN10] Y.-P. WONG et B.-Y. NG : 3d reconstruction from multiple views using particle swarm optimization. *In IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8, 2010.

[Wu82] F.Y. WU : The potts model. *Reviews of Modern Physics*, 54(1):235–268, Janvier 1982.

[WWLHG08] W. WANG, Y. WANG, andQ. Huang L. HUO et W. GAO : Symmetric segment-based stereo matching of motion blurred images with illumination variations. *In ICPR*, pages 1–4, 2008.

[WYJT10] T.-P. WU, S.-K. YEUNG, J. JIA et C.-K. TANG : Quasi-dense 3d reconstruction using tensor-based multiview stereo. *In CVPR*, pages 1482–1489, 2010.

[WZ08] Z.F. WANG et Z.G. ZHENG : A region based stereo matching algorithm using cooperative optimization. *In CVPR*, pages 1–8, 2008.

[XCJ09] W. XIONG, H. S. CHUNG et J. JIA : Fractional stereo matching using expectation-maximization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31:428–443, 2009.

[XJ07] W. XIONG et J. JIA : Stereo matching on objects with fractional boundary. *In CVPR*, pages 1–8, 2007.

[XJ08] L. XU et J.Y. JIA : Stereo matching - an outlier confidence approach. *In ECCV*, volume IV, pages 775–787, 2008.

[XWFS02] Y. XU, D. WANG, T. FENG et H.-Y. SHUM : Stereo computation using radial adaptive windows. *In Proc. Inter. Conf. Pattern Recognition*, volume 3, pages 595–598, 2002.

[Yah07] T. YAHIAOUI : *Une approche de stéréovision dense intégrant des contraintes de similarité*. Thèse de doctorat, These de doscorat de l'Université de Lille 1, 2007.

[YIC10] F. YI, D.R. ISKANDER et M.J. COLLINS : Estimation of the depth of focus from wavefront measurements. *Journal of Vision*, 10(4):1–9, 2010.

[YK06] K.J. YOON et I.S. KWEON : Adaptive support-weight approach for correspondence search. *IEEE Trans. on PAMI*, 28:650–656, 2006.

[YP05] R. YANG et M. POLLEFEYS : A versatile stereo implementation on commodity graphics hardware. *Journal of Real-Time Imaging*, 11:7–18, 2005.

[YWA10] Q. YANG, L. WANG et N. AHUJA : A constant-space belief propagation algorithm for stereo matching. *In CVPR*, 2010.

[YWY+09] Q. YANG, L. WANG, R. YANG, H. STEWENIUS et D. NISTE : Stereo matching with color-weighted correlation, hierachical belief propagation. *IEEE Trans. on PAMI*, 31(3):492–504, 2009.

[ZC06] X.-P. ZHANG et Z. CHEN : An automated video object extraction system based on spatiotemporal independent component analysis and multiscale segmentation. *EURASIP Journal on Applied Signal Processing*, 2006:1–22, 2006.

[ZGY08] Y. ZHANG, M. GONG et Y.-H. YANG : Local stereo matching with 3d adaptive cost aggregation for slanted surface modeling and sub-pixel accuracy. *In ICPR*, pages 1–4, 2008.

[Zhe10] M.J. ZHENG : Depth map estimation from uncalibrated stereo video sequences. *In The 23rd IPPR Conf on Computer Vision, Graphics, and Image Processing*, volume 16, 2010.

[ZHK+03] T.E. ZICKLER, J. HO, D.J. KRIEGMAN, J. PONCE et P.N. BELHUMEUR : Binocular helmholtz stereopsis. *In ICCV*, volume 2, pages 1411–1417, Oct 2003.

[ZJWB08] G. ZHANG, J. JIA, T.T. WONG et H. BAO : Recovering consistent video depth maps via bundle optimization. *In CVPR*, 2008.

[ZK00] C.L. ZITNICK et T. KANADE : A cooperative algorithm for stereo matching and occlusion detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 22(7):675–684, 2000.

[ZK02] Y. ZHANG et C. KAMBHAMETTU : Stereo matching with segmentation-based cooperation. *In In Proc. of the European Conference on Computer Vision (ECCV)*, volume 2, pages 556–571, 2002.

[ZK07] C.L. ZITNICK et S.B. KANG : Stereo for image-based rendering using image oversegmentation. *IJCV*, 75(1):49–65, 2007.

[ZKU+04] C.L. ZITNICK, S.B. KANG, M. UYTTENDAELE, S. WINDER et R. SZELISKI : High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics (TOG)*, 23(3):600–608, 2004.

[ZS03] J. ZHONG et S. SCLAROFF : Segmenting foreground objects from a dynamic textured background via a robust kalman filter. *In IEEE International Conference on Computer Vision*, pages 44–50, 2003.

[ZW94] R. ZABIH et J. WOODFILL : Non-parametric local transforms for computing visual correspondence. *In Proc. of the Third European Conf. on Computer Vision (ECCV)*, volume 2, pages 151–158, 1994.

[ZWGY10] J. ZHU, L. WANG, J. GAO et R. YANG : Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs. *IEEE Trans. on PAMI*, 32(5):899–909, May 2010.

[ZZW09] F. ZHANG, L. ZHANG et D. WANG : Progressive correspondence approach based on edge. *In Inter. Conf. on Communications and Mobile Computing (CMC'09)*, pages 195–197, 2009.

Titre : Détection et localisation tridimensionnelle par stéréovision d'objets en mouvement dans des environnements complexes.

Résumé : La sécurité des personnes et des équipements est un élément capital dans le domaine des transports routiers et ferroviaires. Depuis quelques années, les Passages à Niveau (PN) ont fait l'objet de davantage d'attention afin d'accroître la sécurité des usagers sur cette portion route/rail considérée comme dangereuse. Nous proposons dans cette thèse un système de vision stéréoscopique pour la détection automatique des situations dangereuses. Un tel système permet la détection et la localisation d'obstacles sur ou autour du PN. Le système de vision proposé est composé de deux caméras supervisant la zone de croisement. Nous avons développé des algorithmes permettant à la fois la détection d'objets, tels que des piétons ou des véhicules, et la localisation 3D de ces derniers. L'algorithme de détection d'obstacles se base sur l'Analyse en Composantes Indépendantes et la propagation de croyance spatio-temporelle. L'algorithme de localisation tridimensionnelle exploite les avantages des méthodes locales et globales, et est composé de trois étapes : la première consiste à estimer une carte de disparité à partir d'une fonction de vraisemblance basée sur les méthodes locales. La deuxième étape permet d'identifier les pixels bien mis en correspondance ayant des mesures de confiances élevées. Ce sous-ensemble de pixels est le point de départ de la troisième étape qui consiste à ré-estimer les disparités du reste des pixels par propagation de croyance sélective. Le mouvement est introduit comme une contrainte dans l'algorithme de localisation 3D permettant l'amélioration de la précision de localisation et l'accélération du temps de traitement.

Title : Detection and 3D localization of moving and stationary obstacles by stereo vision in complex environments. Application at level crossings.

Abstract : Within the past years, railways undertakings became interested in the assessment of Level Crossings (LC) safety. We propose in this thesis an Automatic Video-Surveillance system (AVS) at LC for an automatic detection of specific events. The system allows automatically detecting and 3D localizing the presence of one or more obstacles which are motionless at the level crossing. Our research aims at developing an AVS using the passive stereo vision principles. The proposed imaging system uses two cameras to detect and localize any kind of object lying on a railway level crossing. The cameras are placed so that the dangerous zones are well (fully) monitored. The system supervises and estimates automatically the critical situations by detecting objects in the hazardous zone defined as the crossing zone of a railway line by a road or path. The AVS system is used to monitor dynamic scenes where interactions take place among objects of interest (people or vehicles). After a classical image grabbing and digitizing step, the processing is composed of the two following modules : moving and stationary objects detection and 3-D localization. The developed stereo matching algorithm stems from an inference principle based on belief propagation and energy minimization. It takes into account the advantages of local methods for reducing the complexity of the inference step achieved by the belief propagation technique which leads to an improvement in the quality of results. The motion detection module is considered as a constraint which allows improving and speeding up the 3D localisation algorithm.

Discipline : Automatique, Génie Informatique, Traitement du signal et Images

Mots clés : Mise en correspondance, Propagation de Croyance Sélective, Mesure de confiance, Analyse en composantes indépendantes, Propagation de croyance spatio-temporelle, Passages à niveau.

Laboratoires : IFSTTAR/ESTAS-LEOST, 20 rue Élisée Reclus, F-59650, Villeneuve d'Ascq, France