Cooperation stereo mouvement pour la detection des objets dynamiques

Adrien Bak

UNIVERSITÉ PARIS SUD XI

Ecole Doctorale
Sciences et Technologies de l'Information, des Télécommunications
et des Systèmes

Thèse Présentée par :
Adrien BAK

Formation Doctorale

en vue de l'obtention du grade de

DOCTEUR

spécialité: Traitement des Images

Coopération Stéréo-Mouvement pour la Détection des Objets Dynamiques

Soutenue publiquement le : 28 août 2011 devant le jury :

Madame Sylvie LE HÉGARAT	Présidente du Jury
Monsieur Jack-Gérard POSTAIRE	Rapporteur
Monsieur Fawzi NASHASHIBI	Rapporteur
Monsieur Vincent FRÉMONT	Examinateur
Monsieur Frédéric LARGE	Membre Invité
Monsieur Didier AUBERT	Directeur de Thèse
Madame Samia BOUCHAFA	Encadrante

Remerciements

En premier lieu, je tiens à remercier Mme Sylvie LE HÉGARAT pour m'avoir fait l'honneur de présider mon jury de thèse, mais également pour son rôle en tant que responsable du département ACCIS, où j'ai eu beaucoup de plaisir à travailler pendant ces trois années.

Je tiens également à remercier Messieurs POSTAIRE et NASHASHIBI pour avoir accepté de rapporter ce travail. Leurs remarques et conclusions m'ont été d'une grande utilité pour tirer les enseignements de mon travail de thèse.

Mes remerciements vont également à Messieurs FRÉMONT et LARGE pour avoir accepter de prendre part au jury de ce travail.

J'adresse également mes plus sincères remerciements à Didier Aubert et Samia Bouchafa pour avoir su m'accorder leur confiance et m'avoir guider et conseiller tout au long de ces trois années passées ensemble.

Ces trois années n'auraient pas été les mêmes sans tout ceux avec qui j'ai eu le plaisir de les partager

Je remercie finalement mes amis et ma famille pour m'avoir toujours soutenu et sans qui je n'en serait pas là.

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

Depuis quelques années, les différents secteurs de la robotique connaissent un développement important. Si la robotique est emmenée par un marché grand public florissant, celui-ci reste majoritairement centré sur un public d'enthousiastes ou sur quelques applications simples (on peut citer par exemple le robot-aspirateur Roomba®). De nombreux secteurs économiques bénéficient, plus ou moins explicitement, des avancées dans les domaines connexes de la robotique. Celui de la défense, bien sûr, avec le recours de plus en plus systématique aux drones, les services à la personne, en particulier au Japon, où le recours à une présence robotique pour la prise en charge des personnes âgées dépendantes est de plus en plus envisagé comme une solution au vieillissement de la population de l'archipel.

D'un point de vue académique, le secteur le plus dynamique est probablement celui des véhicules autonomes, ou intelligents. En effet, un grand nombre de challenges, aussi bien technologiques que scientifiques restent à relever. En particulier, les véhicules ayant vocation à évoluer en extérieur, et plus spécifiquement les automobiles concentrent un grande partie de l'effort de recherche. Le poids économique des acteurs industriels du domaine, ainsi que la prévalence d'un des grands problèmes de santé publique, celui des décès liés à la route, font des véhicules automatisés et des aides avancées à la conduite (*Advanced Driver Assistance System, ADAS*) un des centres d'intérêt majeurs de la recherche académique en robotique.

Dès l'introduction de l'automobile, ses effets en termes d'efficacité logistique ont rapidement été mis en regard de sa dangerosité. Par exemple, dès 1865, le parlement britannique n'hésita pas à légiférer dans ce sens en introduisant le *Locomotive Act* qui imposait notamment à chaque véhicule à moteur d'être précédé par un piéton muni d'un drapeau rouge. Néanmoins, pendant la première moitié du XX$^{\text{è}}$ siècle, les pouvoirs publics ont assisté à une importante croissance de la mortalité sur les routes.

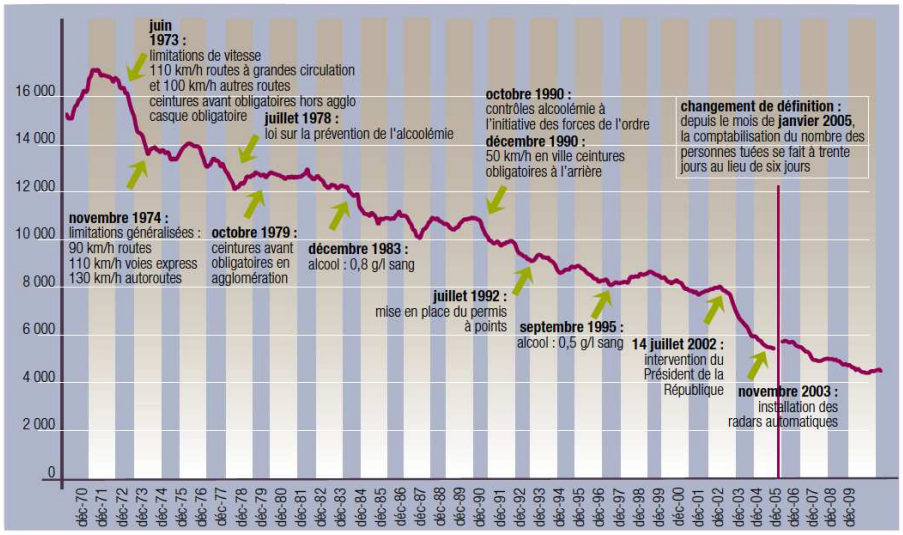


FIGURE 1 – Evolution du nombre de tués sur les routes de France entre 1970 et 2009. *Source ONISR*

A partir des années 70, d'importantes évolutions ont permis d'inverser cette tendance. Dans un premier temps, ces évolutions furent purement législatives (mise en place des limitations de vitesse, obligation du port de la ceinture de sécurité, ...). Dans un second temps, un effort technologique de développement des dispositifs de sécurité (airbag, ABS, *etc.*), couplé à un effort de prévention et d'éducation du public en matière de sécurité et de prudence routière ont eu un très net impact sur la mortalité routière. Toutefois, il apparaît que, depuis les deux dernières années années, la mortalité stagne. Cela peut s'expliquer par plusieurs facteurs. Tout d'abord, l'*indice de circulation* (nombre de kilomètres parcourus chaque année) augmente encore légèrement, bien que sa croissance diminue d'année en année, ce qui peut contribuer à masquer une baisse faible du nombre de tués sur la route par kilomètre parcouru.

Aujourd'hui, nous pouvons globalement considérer que 90% des accidents mortels sont dûs en premier lieu au comportement du conducteur[1]. Aussi, les développements récents et à venir ne doivent plus se contenter de pouvoir assurer une sécurité passive des usagers de la route, mais devenir pro-actifs. En particulier, deux pistes nous semblent pertinentes : l'évitement et la mitigation de collision.

[1] Ces chiffres proviennent du rapport 2008 de la Sécurité Routière, que le lecteur peut consulter sur http://www.securite-routiere.equipement.gouv.fr

Il est important de noter, cependant, qu'un facteur important est le comportement des conducteurs. On peut en effet considérer que les usagers vont agir en termes de *risque perçu constant*. Paradoxalement, un système éliminant les collisions aurait ainsi de fortes chances d'encourager les prises de risques des conducteurs, là où un système qui se contente de limiter les dégâts liés à une collision (mitigation) pourrait être plus judicieux.

Contexte

De nombreuses équipes, aussi bien académiques qu'industrielles se sont emparées de ces thématiques. Les défis qui se présentent à nous relèvent en effet de nombreux domaines scientifiques. Les communications intra- et extra- véhiculaire, le développement de systèmes de commandes (embarqués ou déportés) sont et seront bien évidemment indispensables, mais le défi le plus important est sans doute celui des capteurs proprio et surtout extéroceptifs. En effet, la perception de l'environnement va jouer un rôle prépondérant dans la prise de décision et dans le représentation de cet environnement. Plusieurs pistes technologiques sont utilisées afin d'aboutir à une représentation efficace de l'environnement d'un véhicule autonome (LIDAR, radars, caméras, ...). Comme nous le verrons par la suite, les capteurs de vision présentent plusieurs avantages décisifs. Tout d'abord, leur coût très réduit permet d'envisager des applications industrielles à court terme. Ensuite, la richesse de l'information qu'ils permettent de recueillir en font un outil particulièrement polyvalent. Ainsi, nous nous attacherons à démontrer qu'un système stéréoscopique soigneusement conçu permet d'obtenir des résultats comparables à ceux d'une centrale inertielle beaucoup plus onéreuse, tout en étant également capable de produire une représentation 3D de l'environnement dans lequel évolue le véhicule, ainsi que la détection et une localisation des menaces potentielles.

Dans ce contexte, le cluster Digiteo[2] a choisi de financer le présent travail de recherche, en collaboration avec l'Université Paris-Sud et l'IFSTTAR (institut issu de la fusion entre le LCPC et l'INRETS).

Objectifs et Contributions

L'objectif initial de ce travail de thèse était de proposer des moyens innovants, permettant d'établir une collaboration efficace entre les méthodes de traitement des images propres au mouvement et propres à la stéréo-vision. La détection

[2] http://www.digiteo.fr

d'objets mobiles à partir d'un capteur, lui même mobile, est la finalité que nous souhaitons atteindre. A cette fin, une méthode d'odométrie visuelle sera mise en place et évaluée, avant d'être exploitée à des fins de détection. Au court de ce travail, nous nous attacherons à formaliser un certains de nombre de recommandations, issues de notre connaissance approfondie du système développé et de sa généralisation. Ces recommandations et observations ont vocation à être utilisées par de futurs concepteurs de systèmes de vision. Parallèlement à ce travail, centré sur l'utilisation explicite de l'information structurelle fournie par la stéréo-vision, une étude centrée sur l'utilisation d'un capteur monoculaire sera également menée. Cette seconde phase, centrée autour d'une utilisation nouvelle du concept de C-Vélocité, permet d'entrevoir les futurs développements d'un système purement monoculaire, qui permettrait une estimation jointe du mouvement propre du capteur et de la structure de la scène observée.

Organisation du présent mémoire

Le chapitre 1 reviendra sur les bases de la vision par ordinateur, qu'elle soit monoculaire ou binoculaire, ainsi que sur l'état de l'art des techniques actuelles d'estimation du mouvement et de calcul de cartes de disparité. En effet, la communauté scientifique de traitement des images est extrêmement active et diversifiée, et les outils mis au point sont très nombreux. Nous exposerons les bases théoriques communes à plusieurs types de systèmes, avant de revenir sur les principales approches algorithmiques.

Le chapitre 2 développera le point particulier de l'Odométrie Visuelle. Après être revenu plus précisément sur les différentes méthodes actuelles. Le système développé sera présenté en détails et évalué. Cette évaluation portera sur les performances intrinsèques du système, mais également sur sa comparaison avec différents appareils commerciaux, et surtout sur sa complémentarité avec d'autres capteurs proprioceptifs.

Le chapitre 3 présentera les moyens envisagés afin de remplir l'objectif premier de ce travail : détecter les objets mobiles, à partir d'un capteur lui même en mouvement. Plusieurs propositions seront faites et comparées. Les performances de ces systèmes seront comparées. Nous nous attarderons également sur une étude des facteurs limitant d'un tel système, ainsi que sur certaines pistes d'améliorations. En particulier, ce travail peut servir à émettre un certain nombre de recommandations à l'attention de futurs concepteurs de systèmes de vision embarqués.

Le chapitre 4 décrira une technique monoculaire d'estimation du mouvement d'un capteur de vision, la C-Vélocité Inverse (CVI). La CVI trouve ses racines dans les travaux sur la C-Vélocité, concept permettant d'exploiter un champ de flot optique d'une manière novatrice afin de récupérer des informations sur la structure de la scène observée. La CVI utilise une information structurelle afin de déterminer le mouvement d'un capteur. Nous reviendrons donc sur les bases de la C-Vélocité, avant d'introduire la CVI. Ce procédé sera évalué et les nombreuses possibilités qu'il ouvre seront discutées.

Finalement, nous conclurons sur les contributions proposées au cours de ce travail de thèse, ainsi que sur les différentes pistes d'amélioration et de développement pour l'avenir. Nous reviendrons également sur un certains nombre d'applications qui pourraient, à court terme, bénéficier des travaux présentés.

Chapitre 1

Prérequis - Modèles et Notations

«*Where there is no vision, there is no hope*»
Georges Washington Carver

Sommaire

L'objet de ce chapitre est, dans un premier temps d'introduire les bases mathématiques utilisées au cours de ce travail. Par la suite, nous nous pencherons sur les capteurs de vision, qu'ils soient monoculaire ou binoculaire. Nous présenterons les différents modèles de représentation. Nous nous pencherons également sur les différentes techniques, abordées dans la littérature, et permettant d'extraire l'information temporelle (mouvement) et structurelle (stéréo) de ces capteurs. Le lecteur familier avec la géométrie projective et avec l'état de l'art en estimation monoculaire du mouvement et en estimation binoculaire de profondeur peut poursuivre la lecture avec le chapitre 2.

1.1 Objets Mathématiques

Un certain nombre d'objets mathématiques seront manipulés. A des fins de clarté, les notations utilisées sont présentées ci-après.

Objet / Concept	Représentation Utilisée
Variable Monodimensionelle	x
Variable Vectorielle	$\boldsymbol{X}$
Coordonnées d'un vecteur $\boldsymbol{M}$	$\boldsymbol{M} = \left\vert \begin{matrix} x_M \\ y_M \\ z_M \end{matrix} \right.$
Ensemble	$\mathcal{X}$
Fonction d'une variable y	$x(y)$
Fonction d'une variable y, paramétrisée par $\boldsymbol{P}$	$x_{(\boldsymbol{P})}(y)$
Dérivée partielle	$\frac{\partial x}{\partial y}$
Moyenne d'une variable aléatoire	$\hat{x}$
Estimée, Mesure	$\tilde{x}$
Proportionalité entre deux grandeurs	$\propto$

TABLE 1.1 – Tables des Notations

Il est également à noter que nous utiliseront les termes *objet* et *image*, empruntés à l'Optique, afin de différencier le monde physique de sa représentation dans le plan focale d'un capteur.

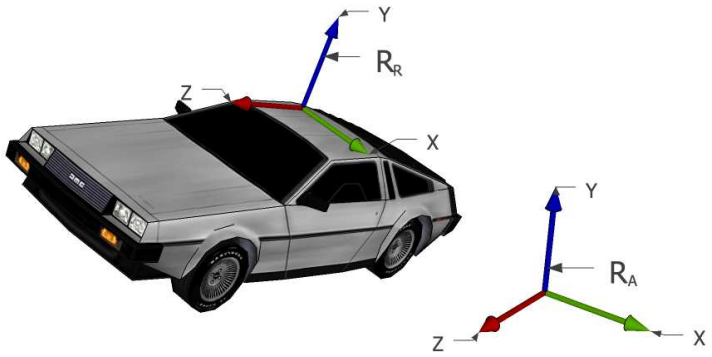


FIGURE 1.1 – Représentation des systèmes d'axes considérés

1.2 Modélisation d'un Monde 4D

Le monde dans lequel nous évoluons, et dans lequel un véhicule autonome est amené à évoluer est complexe. Il n'est pas plan, ni statique, ni composé d'objets rigides. Nous considérons que le monde est représenté par une base de temps et deux repères cartésiens. Le premier, noté $\mathcal{R}_a$, est absolu, alors que le second, noté $\mathcal{R}_r$, est relatif et lié au mobile considéré. Ces systèmes d'axes sont représentés en figure 1.1.

Deux instants t_0 et t_1 sont considérés. En $t = t_0$, les deux repères sont alignés et leurs origines sont confondues, le cas échéant, elles sont positionnées au point principal d'un des deux capteurs. Entre les deux instants, un mouvement non-contraint peut exister entre les deux repères. Ce mouvement est composé d'une composante translationelle $\boldsymbol{T}$ et d'une composante rotationelle $\boldsymbol{\Omega}$:

$$\boldsymbol{T} = \begin{vmatrix} T_X \\ T_Y \\ T_Z \end{vmatrix} \quad ; \quad \boldsymbol{\Omega} = \begin{vmatrix} \omega_X \\ \omega_Y \\ \omega_Z \end{vmatrix} \tag{1.1}$$

On suppose que les instants t_0 et t_1 sont suffisamment proches pour pouvoir linéariser les lignes trigonométriques. Dès lors, pour un point physique M, ses coordonnées peuvent s'exprimer dans $\mathcal{R}_a$, ou dans $\mathcal{R}_r$:

$$M(t_0) = \begin{vmatrix} X_M \\ Y_M \\ Z_M \end{vmatrix}_{\mathcal{R}_a} = \begin{vmatrix} X_M \\ Y_M \\ Z_M \end{vmatrix}_{\mathcal{R}_r} \tag{1.2}$$

Après le déplacement sus-mentionné, ces coordonnées deviennent :

$$M(t_1) = \left|\begin{matrix} X_M \\ Y_M \\ Z_M \end{matrix}\right|_{\mathcal{R}_a} = \left|\begin{matrix} X_M \\ Y_M \\ Z_M \end{matrix}\right|_{\mathcal{R}_r} + \left|\begin{matrix} X_M \\ Y_M \\ Z_M \end{matrix}\right| \wedge \left|\begin{matrix} \omega_X \\ \omega_Y \\ \omega_Z \end{matrix}\right| - \left|\begin{matrix} T_X \\ T_Y \\ T_Z \end{matrix}\right|_{\mathcal{R}_r} = \left|\begin{matrix} X_M - \omega_Y Z_M + \omega_Z Y_M - T_X \\ Y_M + \omega_X Z_M - \omega_Z X_M - T_Y \\ Z_M + \omega_Y X_M - \omega_X Y_M - T_Z \end{matrix}\right|_{\mathcal{R}_r} \tag{1.3}$$

Maintenant que nous avons vu comment nous allons modéliser les différents mouvements intervenants, nous allons nous intéresser de plus près aux capteurs de vision que nous utiliserons pour l'imager. Dans un premier temps, nous allons nous intéresser à la vision monoculaire, en choisissant un modèle adéquat pour les capteurs, en revenant sur la notion de géométrie projective, ainsi que sur le mouvement, son image et les différentes façon de l'imager. Dans un second temps, nous nous pencherons sur la vision binoculaire et plus particulièrement sur la stéréo-vision. Nous enrichirons le modèle choisi, avant de présenter les différentes méthodes envisagées dans la littérature permettant de résoudre les problèmes d'appariements liés à la stéréo-vision. L'expression de l'image du mouvement 3D sera également présentée, ainsi qu'un aperçu des technologies de mesure de la profondeur, analogue à la stéréo-vision.

1.3 Vision Monoculaire

1.3.1 Modélisation d'un capteur monoculaire

Nous utiliserons au cours de ce travail un ou plusieurs capteurs de vision. Fondamentalement, un capteur peut être vu comme étant composé d'une optique et d'une matrice photo-sensible. Il existe plusieurs façons de modéliser un tel système, prenant en compte un plus ou moins grand nombre de paramètres et d'aberrations.

Il existe plusieurs modèles couramment admis dans la littérature[HZ03, HM95].

Une modélisation fine - Tout d'abord, une modélisation précise, illustrée en figure 1.2, des différents phénomènes physiques à l'origine de la formation des images peut apparaître intéressante. Ainsi, les auteurs de [KMH95] cherchent à obtenir la modélisation la plus fidèle d'une optique, à des fins d'imagerie numérique. La plupart des aberrations géométriques sont prises en compte par ce modèle, hautement non linéaire. Si une telle précision peut présenter un intérêt dans la reproduction photo-réaliste d'images, il n'est pas certain que nos applications puissent en bénéficier.

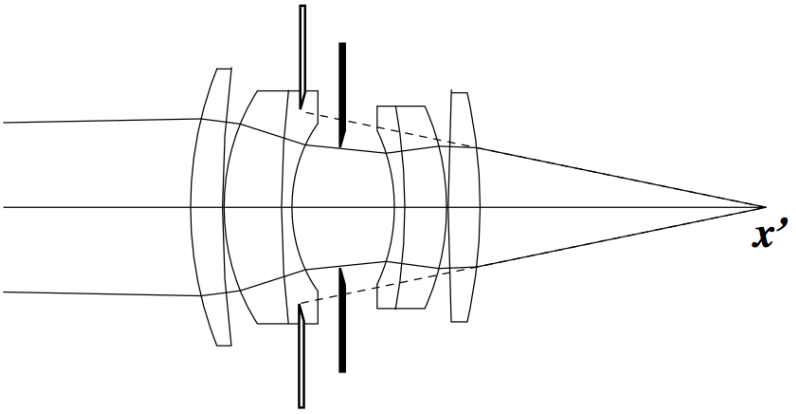


FIGURE 1.2 – Modèle d'optique complet, prenant en compte les différentes aberrations optiques[1].

Modèle Sténopé - À l'autre extrémité du spectre, le modèle sténopé (*pinhole* dans la littérature anglophone) est extrêmement simple [HZ03] [HM95]. Il permet d'assimiler le capteur de vision à un opérateur linéaire. Bien que très simple, il est suffisant dans un grand nombre d'applications, il est important de revenir brièvement sur les limitations de ce modèle :

- L'approximation de Gauss est valide sur tout le champ, les optiques "grand angle" ne sont donc pas prises en compte.
- De la même façon, les systèmes sujets aux aberrations ne peuvent pas être modélisés par un sténopé. En particulier, les systèmes très intégrés (type téléphone portable) pour lesquels la taille de la tâche de diffraction est de l'ordre du pixel.
- Toute considération photométrique est écartée. Il suppose ainsi que le vignettage est inexistant, ce qui n'est pas nécessairement le cas dans un système réel.

Une illustration du modèle sténopé peut être trouvé en figure 1.3. Un capteur de vision modélisé par un sténopé peut être entièrement défini par les grandeurs suivantes :

- La longueur focale de son optique : f' (millimètres).

[1] Image tirée de [KMH95], reproduite avec l'autorisation des ayant-droits.

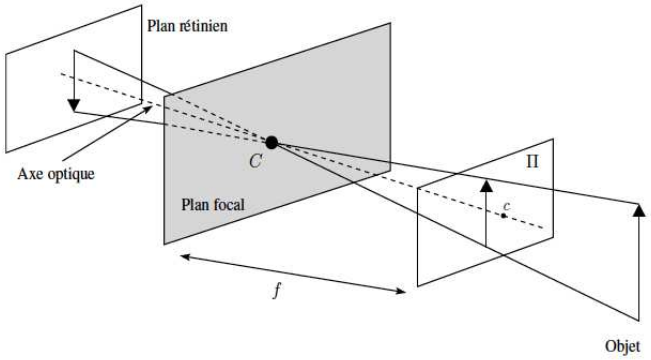


FIGURE 1.3 – Illustration du modèle Sténopé

- Les dimensions verticale et horizontale de ses pixels : α_v et α_u (pixels par millimètre).

- Les coordonnées de l'origine du capteur dans $\mathcal{R}_r$: u_0 et v_0 (pixels).

- Le coefficient de non-orthogonalité du capteur s_{uv} (pixels par millimètre). Ce dernier peut être négligé dans l'immense majorité des cas.

Modèle Sténopé étendu - Finalement, l'intérêt d'un modèle intermédiaire a plusieurs fois été avancé [Bey92] [Bro02]. Dans ce type de modèle, l'approche linéaire du modèle sténopé est enrichie par des termes d'ordre supérieurs, correspondant aux distorsions, à l'aberration sphérique, etc. Ce type de modélisation présente un intérêt notable dans les cas où il convient de compenser les aberrations afin d'obtenir une précision métrologique.

Malgré les limitations avancées, nous allons travailler avec le modèle sténopé. En effet, les non-linéarités introduites par des modèles plus complexes, en complexifiant les équations de projection, pourraient être rédhibitoires. De plus, les limitations du modèle sténopé ne constituent pas nécessairement des éléments bloquants : nous ne travaillerons pas avec des optiques grand angle et, d'une manière générale, les optiques utilisées seront de qualité suffisante pour ne pas avoir besoin de prendre en compte les aberrations[2].

1.3.2 Équations du Modèle Sténopé

Le capteur de vision peut être vu comme deux systèmes distincts. Tout d'abord, la conjugaison optique proprement dite peut être modélisée au moyen d'une transformation projective. Ensuite, la matrice photosensible, qu'elle soit naturelle (cônes et bâtonnets) ou synthétique (pixels CMOS ou CCD) va intégrer et discrétiser le signal et opérer un changement de repère[3]. Ces deux transformations peuvent être envisagées séparément.

Conjugaison Optique - Un point objet $M = \left| \begin{array}{c} X_M \\ Y_M \\ Z_M \end{array} \right|_{\mathcal{R}_c}$ est imagé sur le plan focal du dispositif en $m = \left| \begin{array}{l} x_m = f\frac{X_M}{Z_M} \\ y_m = f\frac{Y_M}{Z_M} \end{array} \right|_{\mathcal{R}_c}$. Cette transformation n'est pas linéaire. Toutefois, l'utilisation des coordonnées homogènes (ou projectives) permet de définir la transformation projective comme un opérateur matriciel [HZ03]. Un point M d'un espace de dimension n admet comme coordonnées projective, le lieu de la droite vectorielle d'un espace de dimension $(n+1)$ engendrée par $\begin{pmatrix} M \\ 1 \end{pmatrix}$. Ainsi, les coordonnées projectives d'un point ne sont définies qu'à une constante multiplicative près. En utilisant ce formalisme, la projection opérée par le sténopé devient un opérateur linéaire :

$$\left| \begin{array}{c} x_m \\ y_m \\ 1 \end{array} \right| = \begin{pmatrix} f' & 0 & 0 & 0 \\ 0 & f' & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \cdot \left| \begin{array}{c} X_M \\ Y_M \\ Z_M \\ 1 \end{array} \right| \tag{1.4}$$

Toutefois, et il s'agit du problème fondamental de la vision monoculaire, cette transformation est non-injective. Mathématiquement, cela est traduit par, justement, le fait que les coordonnées projectives ne sont définies qu'à un facteur d'échelle près. Physiquement, cela traduit le fait que tous les points objets situés sur une même droite passant par le centre optique du capteur vont admettre la même image. Cette non-injectivité est à l'origine des principales limitations des systèmes monoculaires.

[2]Exception faite de la distorsion qui est correctible par étalonnage

[3]Rigoureusement, ce changement de repère n'est pas dû à la rétine, mais il est d'usage de définir que le repère image lié à un capteur a son origine dans l'un des coin de la matrice, plutôt que sur le point principal.

Influence de la rétine - Comme nous l'avons vu, le rôle de la rétine est d'intégrer spatialement et temporellement l'information lumineuse. Dans cette étude, nous ne nous pencherons pas sur l'aspect temporel de cette intégration, nous faisons l'hypothèse que les temps d'intégration permettent d'avoir des images contrastées et nettes. En revanche, l'intégration spatiale, s'accompagne d'un changement de référence, plus adapté au cadre informatique des traitements à venir. Les coordonnées des points vont être exprimées en pixels et l'origine va être rapportée au coin de l'image.

Cette seconde transformation peut également s'exprimer sous forme matricielle :

$$\begin{vmatrix} u_m \\ v_m \\ 1 \end{vmatrix} = \begin{pmatrix} \alpha_u & s_{uv} & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{vmatrix} x_m \\ y_m \\ 1 \end{vmatrix} \tag{1.5}$$

Système Complet - En combinant les équations 1.4 et 1.5 nous pouvons exprimer l'opérateur équivalent à notre capteur de vision complet :

$$\begin{vmatrix} u_m \\ v_m \\ 1 \end{vmatrix} = K_C \cdot \begin{vmatrix} X_M \\ Y_M \\ Z_M \\ 1 \end{vmatrix} \tag{1.6}$$

Avec :

$$K_C = \begin{pmatrix} f_u & s_{uv} & u_0 & 0 \\ 0 & f_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{1.7}$$

Dans cette dernière équation, $f_u = \frac{f'}{\alpha_u}$ et $f_v = \frac{f'}{\alpha_v}$ sont les longueurs focales équivalentes horizontale et verticale, exprimées en pixels. Par la suite, nous considèrerons que $f_u = f_v = f$, ce qui est usuellement le cas. De plus, nous considèrerons que $s_{uv} = 0$.

Les paramètres apparaissant dans la matrice K_C sont appelés paramètres intrinsèques de la caméra, ils suffisent à décrire complètement le capteur de vision et peuvent être mesurés par calibrage préalable [HM95].

L'équation 1.6 peut finalement être réécrite :

$$\begin{vmatrix} u_m = u_0 + f\frac{X_M}{Z_M} \\ v_m = v_0 + f\frac{Y_M}{Z_M} \end{vmatrix} \tag{1.8}$$

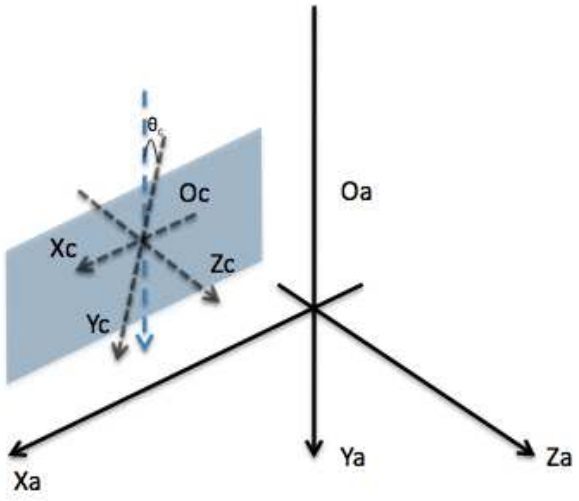


FIGURE 1.4 – Illustration d'un repère dédié à la caméra

Positionnement dans l'Espace - Contrairement aux conventions que nous avons prises, il est possible de considérer un troisième repère, tel qu'illustré en figure 1.4, lié à la caméra et dont les axes sont alignés avec ceux du capteur, alors que les repères $\mathcal{R}_a$ et $\mathcal{R}_r$ ont des axes prenant en compte la géométrie du terrain, ainsi que la mobilité du véhicule.

Classiquement, cela permet de rendre compte du tangage du capteur par rapport à la direction de déplacement privilégié du véhicule (noté θ_c sur la figure 1.4). Il peut alors être souhaitable de localiser les objets dans le repère absolu, et de tenir compte de ce tangage dans les équations de projections, qui deviennent alors :

$$\left| \begin{array}{l} u_m \\ v_m \end{array} = \right| \begin{array}{l} u_0 + f_u \frac{X_M - X_O}{\sin\theta_c(Y_M - Y_O) + \cos\theta_c(Z_M - Z_O)} \\ v_0 + f_v \frac{\cos\theta_c(Y_M - Y_O) - \sin\theta_c(Z_M - Z_O)}{\sin\theta_c(Y_M - Y_O) + \cos\theta_c(Z_M - Z_O)} \end{array} \tag{1.9}$$

Toutefois, pour plusieurs raisons, nous pensons que cette modélisation est inadaptée. Tout d'abord, elle alourdit considérablement les notations. Ensuite, notre approche est centrée sur le ou les capteurs de vision. Nous nous intéressons en premier lieu au mouvement du capteur par rapport à son environnement (ou des objets par rapport à ce capteur). Finalement, en termes d'implémentation, une prise en compte partielle[4] de ce positionnement limite nécessairement la généralisation de

[4]Ne considérer que le tangage, au détriment des autres angles par exemple.

notre travail, alors que la prise en compte totale de ces termes a un impact négatif et non nécessaire sur les performances.

1.3.3 Image 2D du mouvement 3D

Au cours de ce travail, nous aurons maintes fois l'occasion de nous pencher sur le mouvement et son image. A cette fin, il est souhaitable de présenter sa possible expression. Les développements mathématiques permettant d'aboutir aux équations présentées ci-après sont disponibles en annexe A.

En particulier, l'expression de l'image du mouvement nous est donnée par A.15 :

$$\begin{vmatrix} \mu \\ \nu \end{vmatrix} = \begin{vmatrix} \frac{x_m y_m}{f}\omega_X - (f + \frac{{x_m}^2}{f})\omega_Y + y_m\omega_Z - \frac{fT_X}{Z_M} + \frac{x_m T_Z}{Z_M} \\ (f + \frac{{y_m}^2}{f})\omega_X - \frac{x_m y_m}{f}\omega_Y - x_m\omega_Z - \frac{fT_Y}{Z_M} + \frac{y_m T_Z}{Z_M} \end{vmatrix} \tag{1.10}$$

Il est intéressant de noter que ces équations peuvent être décomposées en une partie purement rotationelle :

$$\begin{vmatrix} \mu_{rot} \\ \nu_{rot} \end{vmatrix} = \begin{vmatrix} \frac{x_m y_m}{f}\omega_X - (f + \frac{{x_m}^2}{f})\omega_Y + y_m\omega_Z \\ (f + \frac{{y_m}^2}{f})\omega_X - \frac{x_m y_m}{f}\omega_Y - x_m\omega_Z \end{vmatrix} \tag{1.11}$$

et une partie purement translationnelle :

$$\begin{vmatrix} \mu_{trans} \\ \nu_{trans} \end{vmatrix} = \begin{vmatrix} \frac{x_m T_Z - fT_X}{Z_M} \\ \frac{y_m T_Z - fT_Y}{Z_M} \end{vmatrix} \tag{1.12}$$

Il est important de noter que la partie rotationnelle du mouvement ne dépend pas de la profondeur des points objets considérés, et ne porte donc aucune information structurelle sur la scène observée. Inversement, la partie translationnelle du mouvement dépend de cette profondeur, c'est ce qu'on appelle la parallaxe du mouvement. De plus, et comme nous le verrons plus loin, cette dépendance entre structure de la scène et mouvement translationnel est à l'origine du concept de stéréovision.

1.3.4 Estimation et Mesure de l'image 2D du mouvement 3D

La question de mesurer ou, à défaut, d'estimer ce mouvement est consubstantielle de la vision par ordinateur. C'est en effet une des questions qui ont été le plus fréquemment abordées. Plusieurs types d'approches ont ainsi été envisagées au cours des dernières décennies. On peut en distinguer deux. Tout d'abord, les

méthodes denses (ou semi-denses) pour lesquelles on va chercher à estimer numériquement le champ de déplacement sur toute l'image ou au moins un ensemble de parties connexes de cette image. A l'inverse, les approches éparses où l'on va chercher à extraire d'une image un ensemble de points particuliers avant de chercher leur correspondant dans l'autre image.

1.3.4.1 Approches Denses d'Estimation du Mouvement

Toutes les méthodes denses partent de la même hypothèse de départ : l'hypothèse dite de constance de la luminosité (*constant brightness* dans la littérature anglophone)[BB95]. Il s'agit de considérer que deux points images successifs imageant le même point objet ont la même luminosité[5]. Mathématiquement, cette hypothèse s'énonce :

$$\nabla I \cdot \begin{vmatrix} \mu \\ \nu \end{vmatrix} + \frac{\partial I}{\partial t} = \mathbf{0} \tag{1.13}$$

Où I représente l'intensité lumineuse en un point, ∇ représente l'opérateur vectoriel de différenciation spatiale. Cette contrainte de luminosité constante est à l'origine des approches d'estimation de l'image du mouvement, communément appelées méthodes de Flot Optique[6].

Cette équation définit un problème sous-contraint (ou mal posé pour Hadamard), cette sous-définition du problème est connue sous le nom de "problème d'ouverture" (*aperture problem* dans la littérature anglophone) : il n'est possible de connaître absolument que la projection de l'image du mouvement sur la direction orthogonale au gradient local de l'image [Ull79]. Le problème d'ouverture est illustré en figure 1.5.

Approches Locales - Dès lors, si l'on cherche à évaluer le flot optique, il va être nécessaire d'ajouter une contrainte supplémentaire. Les tenants d'une approche dite locale considèrent ainsi que le flot optique est localement constant, ce qui permet de réduire localement le problème à un système linéaire, résolu aux moindres carrés par une descente de gradient [LK81]. Les améliorations proposées dans la littérature portent sur deux aspects de cette méthode : la méthode de résolution numérique d'une part et l'estimation de grands mouvements d'autre part.

[5] Dans le cas d'images discrétisées, le même niveau de gris.

[6] La similitude entre l'équation 1.13 et l'expression de la dérivée Lagrangienne en mécanique des fluides pourrait être à l'origine de cette dénomination.

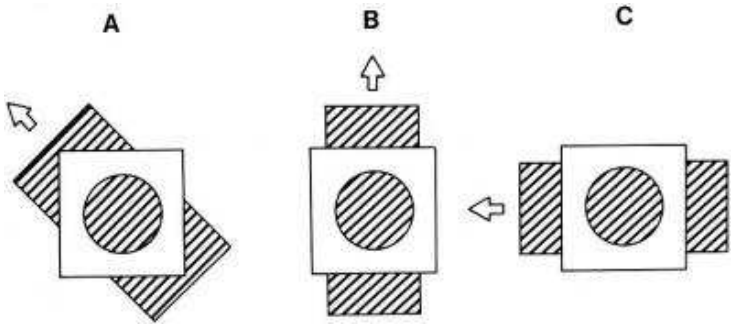


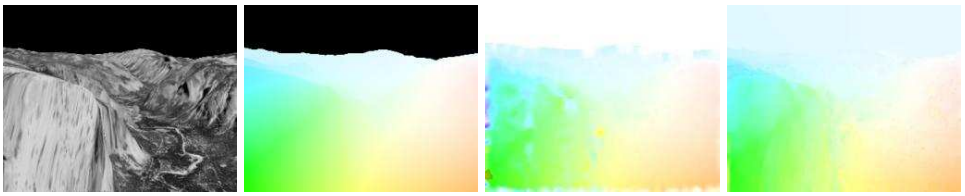
FIGURE 1.5 – Problème d'ouverture : dans les scénarios A, B et C, le mouvement perçu à travers l'ouverture sera le même, même si les mouvements physiques à l'origine de l'observation sont différents.

Si d'autres approches, telle la décomposition des différences [Gle97] ou la régression linéaire [CET01] ont été envisagées, la descente de gradient reste largement utilisée. Alors que l'algorithme originel de LUCAS et KANADE repose sur une approche incrémentale du mouvement, il a été avancé qu'une approche compositionelle [SS02b] pouvait améliorer significativement les résultats. Ces différentes méthodes ont été longuement décrites et comparées par BAKER *et al.* [BM04].

L'amélioration de la méthode initiale de LUCAS et KANADE passe également par la supression de la limitation concernant les grands mouvements. Ainsi, une implémentation multi-résolution peut participer à cette amélioration [Bou99]. Une approche itérative, où l'une des deux images est recalée en utilisant l'estimation courante du flot optique est également fréquemment utilisée afin d'affiner une estimation [BAHH92]. Ce type d'approche itérative peut également être couplée avec une méthode multi-résolutions [SJHG99]. Finalement, des améliorations concernant la convergence générale de l'algorithme, ou sa capacité à être implémenté en temps-réel peuvent être apportées [BC05].

Approches Globales - Inversement, les tenants de l'approche globale vont, au contraire, chercher à définir un terme de régularisation, permettant de rendre le problème soluble. Un tel terme de régularisation est généralement un terme de lissage, pénalisant les gradients trop élevés [HS94]. Cette introduction d'un terme de régularisation permet de réduire le problème de détermination du flot optique à un problème de minimisation d'une fonction quadratique de la forme :

$$\mathcal{E}^2 = \iint \alpha^2 \mathcal{E}_c^2 + \mathcal{E}_b^2 \tag{1.14}$$



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

FIGURE 1.6 – Comparaison de Différentes méthodes de flot optique, réalisée à partir de la séquence Yosemite : (a) Une des images source; (b) Image 2D du mouvement - Vérité Terrain; (c) Flot optique obtenu par une méthode locale [BC05]; (d) Flot optique obtenu par une méthode globale [WTPW09]

Où $\mathcal{E}_b$ est le terme modélisant l'équation fondamentale 1.13, ou terme de données, $\mathcal{E}_c$ est le terme de régularisation et α est un facteur de raideur, ou de relaxation. Ce type de problème a fait l'objet d'études extensives, permettant d'avoir aujourd'hui de nombreuses méthodes d'optimisation numériques à notre disposition, comme par exemple, la méthode de Levenberg-Marquardt [Lev44, Mar63]. Le lecteur peut se tourner vers [Bjö96] pour une étude plus approfondie des méthodes de résolution des problèmes de moindres carrés linéaires ou non-linéaires.

De nombreuses améliorations à ce travail ont été apportées. Si le cadre théorique de résolution des problèmes de moindres carrés est relativement bien maîtrisé, l'essentiel des efforts s'est porté sur la formulation de ce problème. Ainsi, l'introduction de la variation totale, dans le terme de régularisation a permis d'améliorer sensiblement l'estimation du mouvement, en particulier dans les zones faiblement texturées [BW04]. Si certains travaux portent encore sur l'amélioration de la prise en compte du terme de données [PBB$^+$06], l'essentiel de l'effort de recherche est porté sur l'amélioration du terme de régularisation. SHULMAN *et al.* sont ainsi les premiers à proposer l'utilisation de la fonction de HUBER dans le terme de régularisation qui permet une pénalisation plus fine des gradients dépendamment de leur norme [SJ89, Hub73]. Les travaux les plus récents s'attachent à rendre ce terme de régularisation plus fidèle au contenu des images [WTPW09] ou encore, à proposer des modifications algorithmiques permettant des implémentations efficaces [ZPB07].

L'objet de ce travail n'est pas de redéfinir, ou d'étudier différentes méthodes de flot optique, dans l'absolu. Pour une étude plus poussée des (nombreuses) méthodes existantes, BEAUCHEMIN et BARRON [BB95], ou plus récemment WEICK-

ERT *et al.* [WBBN06] ont longuement analysé de façon critique les différentes approches, tandis que BAKER*et al.* établissent une méthodologie permettant de mener une étude comparative des différentes approches [BSL+11].

1.3.4.2 Approches Éparses d'Estimation du Mouvement

Les approches présentées dans la partie précédente visent à établir le flot optique sur tout le champ image. Or il est évident que, surtout dans des images issues de capteurs embarqués de faible résolution et/ou dynamique, il peut exister de nombreux points images porteur de peu, voire pas, d'information. C'est notamment le cas des zones peu texturées (route, murs lisses, etc.). Une approche envisageable peut alors être de sélectionner un ensemble de points, porteurs d'information, afin de ne travailler que sur ceux-ci. C'est l'objectif des méthodes éparses.

Points privilégiés pour l'estimation du mouvement - Une première approche consiste à mettre en œuvre une méthode d'estimation locale du flot optique, mais uniquement sur des points potentiellement plus fiables, de part le contenu fréquentiel de leur voisinage. Ainsi, le tracker KLT [TK92] constitue une modification de l'algorithme original de LUCAS et KANADE sur un ensemble de points discret. Cet ensemble de points peut être l'ensemble des points pour lesquels les valeurs propres de la matrice des gradients sont au dessus d'un seuil spécifique : les *Good Features to Track* [ST94], ou encore des points de Harris [HS88]. Ces méthodes sont relativement peu coûteuses en temps de calcul et connaissent une certaine popularité. Toutefois, les limitations du tracker KLT restent les mêmes que celles du flot optique. Les mouvements sont estimés correctement tant que l'on reste dans l'approximation des petits mouvements. De plus, le mouvement reste estimé à partir d'une information différentielle, alors qu'il serait possible de réaliser des appariements rigoureux.

Recherche de Correspondants - Afin de réaliser une stricte mise en correspondance entre deux ensembles de points, nous devons disposer de deux outils :

- Tout d'abord, nous devons pouvoir identifier les candidats à la mise en correspondance.
- Ensuite, nous devons disposer d'une mesure de similarité afin de rechercher effectivement les correspondances.

Les méthodes de recherches de points d'intérêts vues précédemment remplissent le premier de ces deux objectifs. Dans ce cas, plusieurs mesures de similarité sont possibles.

Le premier exemple qui vient à l'esprit est le calcul d'une corrélation entre les voisinages de deux candidats à la mise en correspondance. Toutefois, ce calcul, mené sur un grand nombre de points peut s'avérer couteux. C'est pourquoi plusieurs approximations ont été proposées. Les mesures SAD[7], SSD[8], ainsi que leur variantes centrées autour de 0 : ZSAD et ZSSD, peuvent jouer le rôle d'approximation du score de corrélation. Ces approximations (en particulier SAD) connaissent actuellement un développement très important du point de vue des implémentations, notamment à cause de leurs nombreuses applications en compression vidéo [Ric03]. Une comparaison expérimentale de ces différentes techniques d'approximation de la corrélation se trouve dans [MC95].

Ces mesures présentent plusieurs avantages, notamment en termes de rapidité, mais restent sensibles aux changements d'illumination relative et aux occultations. Afin de surmonter ce problème, la transformée du census peut être utilisée [ZW94]. Cette méthode repose sur la comparaison de motifs : à chaque pixel on attribue un vecteur de bits qui va représenter les intensités relatives des points du voisinage du pixel considéré. La mesure d'information mutuelle [VW95] a également été introduite comme une mesure de similitude insensible aux défauts de capteurs, comme les changements d'illumination. Cette dernière est également très utilisée pour déterminer des appariements stéréo, notamment en raison de sa robustesse à de grands changements de point de vue.

Finalement, un certain nombre de méthodes proposent conjointement une extraction de points d'intérêts et un ensemble de descripteurs particuliers permettant de mettre deux points en correspondance. Par exemple, les points *SIFT*[9] sont définis comme étant les maxima de gradients dans l'espace des échelles (pyramide gaussienne construite à partir d'une image originale) auxquels sont adjoints des descripteurs construits à partir de plusieurs histogrammes des gradients orientés [Low04]. Bien que très robuste, cette méthode est relativement coûteuse, de part les nombreux traitements mis en place (calcul de la pyramide, extraction, puis affinement de la position des points d'intérêts, calculs des descripteurs). Cette lourdeur est l'une des motivations à l'introduction de la méthode *SURF*[10] [BTV06]. Le descripteur proposé repose cette fois sur une décomposition en ondelette de Haar et sur l'utilisation d'images intégrales afin de rendre les calculs beaucoup plus rapides. AGRAWAL a récemment proposé un nouveau type de détecteur [AKB08], dont l'extraction est plus rapide que SURF, sans toutefois bénéficier d'un descripteur facilitant la mise en correspondance.

[7] *Sum of Absolute Differencies* - somme des différences absolues
[8] *Sum of Squared Differencies* - somme des différences au carré
[9] *Scale Invariant Feature Transform*
[10] *Speeded-Up Robust Features*

FIGURE 1.7 – Paire stéréo installée dans le prototype CARLLA du LIVIC

Cette mise en correspondance fait également l'objet de recherches algorithmiques. Ainsi si des stratégies naïves (type *bruteforce*) peuvent être employées, d'autres méthodes, comme les mariages stables, peuvent améliorer les temps de calcul ou diminuer le nombre de fausses associations [Suv06].

1.4 Vision Binoculaire

La vision binoculaire peut être vue comme un cas particulier de mouvement vu d'un capteur monoculaire. Plus spécifiquement, si nous considérons que deux (ou plus) capteurs sont synchronisés, et qu'entre ces capteurs il existe un déplacement purement translationnel et connu, l'équation 1.12 peut nous permettre de retrouver les profondeurs absolues des points observés. L'objet de cette section est de présenter une modélisation possible d'un capteur de stéréo-vision, les méthodes récentes de calcul d'une carte de disparité et l'image du mouvement que nous pouvons obtenir avec un tel capteur. Finalement, nous reviendrons sur d'autres technologies embarquables permettant d'obtenir des mesures de distances. Notre exposé se limitera aux capteurs horizontaux, qui sont plus fréquents, et plus adaptés aux contraintes automobiles. Toutefois, la généralisation à des capteurs verticaux est immédiate.

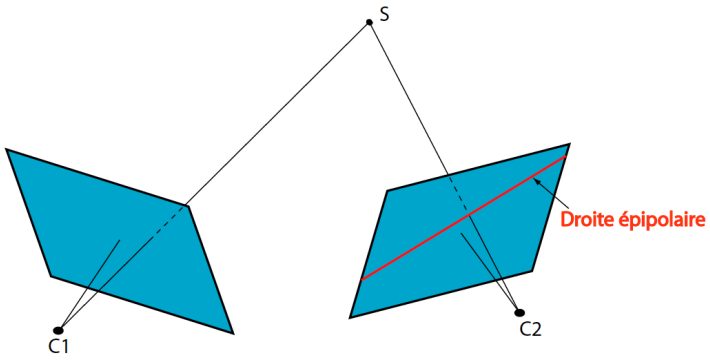


FIGURE 1.8 – Fondement de la Stéréovision - Géométrie Epipolaire

1.4.1 Modélisation d'un capteur binoculaire - Étalonnage et Rectification

1.4.1.1 Modélisation - Rectification

Le principe de la stéréovision repose sur la triangulation. Étant donné deux capteurs, leurs positions relatives (*via* étalonnage par exemple) et les deux images d'un même point par ces deux capteurs, il est aisé de retrouver la profondeur du point objet [HZ03].

Toutefois, la mise en pratique de ce concept simple n'est pas évidente. En effet, on se retrouve confronté à un problème de mise en correspondance, analogue au problème de flot optique, mais où les fenêtres de recherches sont potentiellement beaucoup plus grandes. Afin de trouver les moyens de contrebalancer ce problème, nous devons nous pencher sur la question de la géométrie épipolaire et le problème de la rectification d'images stéréoscopiques.

Dans le cas général, illustré en figure 1.8 où l'on dispose de deux capteurs de vision, sans aucune connaissance sur leur positionnement relatif, on sait que l'antécédent d'un point image du capteur C_1 est la droite qui passe par ce point et par le centre optique du capteur (1.4). L'image de cette droite par le capteur C_2 est la droite épipolaire, il s'agit également et surtout du domaine dans lequel nous devons chercher le correspondant du point S imagé par C_1.

Ce cas général ne prend pas en compte la géométrie particulière de la matrice d'un capteur de vision (où les pixels sont carrés dans une immense majorité des cas), et encore moins le formalisme informatique qui va être utilisé. En effet, il est beaucoup plus efficace et simple de devoir effectuer une recherche de correspondance sur une ligne de l'image. Pour cela, il est nécessaire de rectifier la

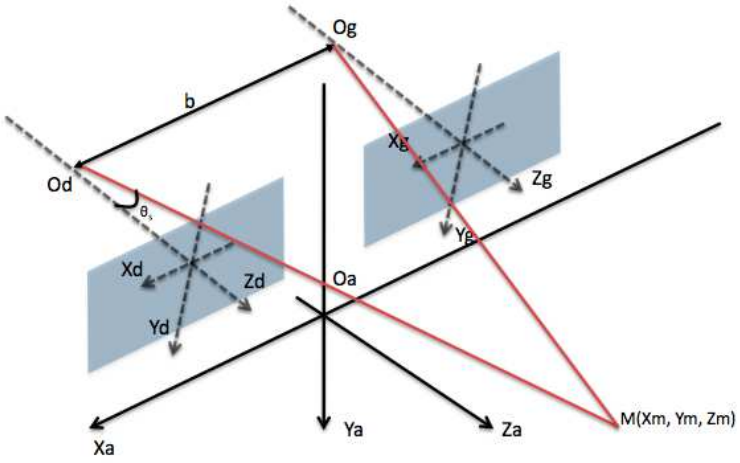


FIGURE 1.9 – Modélisation d'un capteur de stéréo-vision

géométrie épipolaire, de façon à ce que toutes les lignes épipolaires correspondent aux lignes du capteur. Cette rectification est possible en connaissant les positions relatives des deux capteurs. Ces positions relatives peuvent être calculées en utilisant des images d'un motif connu [Zha98]. La matrice fondamentale obtenue peut ensuite être utilisée afin de calculer les transformations à effectuer sur chacune des deux images [LZ99, FTV00]. En pratique, nous pouvons considérer que notre capteur de stéréovision est systématiquement rectifié, tel qu'illustré en figure 1.9.

Dans ce modèle, on considère que les deux matrices sont parfaitement parallèles et alignées. les coordonnées des deux centres optiques dans le repère $\mathcal{R}_a$ sont respectivement $O_G \begin{vmatrix} X_G \\ Y_G \\ Z_G \end{vmatrix}$ pour le capteur gauche et $O_D \begin{vmatrix} X_D \\ Y_D \\ Z_D \end{vmatrix}$ pour le capteur droit.

1.4.1.2 Equations du capteur de Stéréovision

D'après l'équation 1.8, nous pouvons exprimer la position des images d'un point $M = \begin{vmatrix} X_M \\ Y_M \\ Z_M \end{vmatrix}$ à travers les capteurs gauche et droit :

$$\begin{vmatrix} u_g \\ v_g \end{vmatrix} = \begin{vmatrix} u_{0g} + f_g \frac{X_M - X_G}{Z_M - Z_G} \\ v_{0g} + f_g \frac{Y_M - Y_G}{Z_M - Z_G} \end{vmatrix} \quad \begin{vmatrix} u_d \\ v_d \end{vmatrix} = \begin{vmatrix} u_{0d} + f_d \frac{X_M - X_D}{Z_M - Z_G} \\ v_{0d} + f_d \frac{Y_M - Y_D}{Z_M - Z_D} \end{vmatrix} \tag{1.15}$$

Or, si l'on considère d'une part que les capteurs sont identiques : $u_{0g} = u_{0d} = u_0$, $v_{0g} = v_{0d} = v_0$ et $f_g = f_d = f$. Par ailleurs, par rectification : $Y_G = Y_D =$

Y, $Z_G = Z_D = Z$ et $X_D = X_G + b_s$, où la distance b_s est la *base* du capteur. Les équations 1.15 deviennent donc :

$$\left| \begin{array}{c} u_g \\ v_g \end{array} = \right| \left| \begin{array}{c} u_0 + f\frac{X_M - X_G}{Z_M - Z} \\ v_0 + f\frac{Y_M - Y}{Z_M - Z} \end{array} \right| \quad \left| \begin{array}{c} u_d \\ v_d \end{array} = \right| \left| \begin{array}{c} u_0 + f\frac{X_M - X_G - b_s}{Z_M - Z} \\ v_0 + f\frac{Y_M - Y}{Z_M - Z} \end{array} \right| \tag{1.16}$$

On retrouve ici le fait que deux points images correspondant à un même point objet ont bien la même ordonnée. De plus, il vient que l'information relative à la profondeur du point objet peut être résumée dans la grandeur $\delta = u_g - u_d = f\frac{b_s}{Z_M - Z}$. Cette grandeur est classiquement appellée *disparité*[11]. L'image faisant correspondre à chaque point la disparité qui lui est associée est communément appelée une carte de disparité.

Dès lors, nous pouvons poser, en centrant notre repère sur le centre optique du capteur gauche, les équations régissant le capteur de stéréo-vision :

$$\left| \begin{array}{c} u = u_0 + f\frac{X_M}{Z_M} \\ v = v_0 + f\frac{Y_M}{Z_M} \\ \delta = f\frac{b_s}{Z_M} \end{array} \right. \tag{1.17}$$

Ces équations nous permettent de passer d'un repère image, dans lequel nous connaissons parfaitement les correspondances gauche-droite à l'espace objet. Toutefois, une grande partie du problème de stéréo-vision consiste en l'acquisition de ces correspondances. C'est ce point que nous allons évoquer dans la suite.

1.4.2 Construction d'une Carte de Disparité

La construction de cartes de disparité répond à la même problématique générale que le calcul du flot optique : il s'agit de construire un champ, dense ou éparse, de correspondance entre deux images. Ainsi, des travaux ont porté sur les deux problématiques [LK81], ou sur le transfert d'un problème vers l'autre [SBW05].

La seule différence fondamentale entre ces deux procédés vient de la géométrie qui, en limitant l'espace de recherche de correspondants, ajoute une contrainte intéressante au problème de stéréo-vision. Dès lors, les différentes approches envisagées dans la littérature respectent la même typologie : d'une part les approches locales, et d'autre part, les approches globales.

Nous ne reviendrons pas sur les méthodes de mises en correspondance de points d'intérêts robustes, telles qu'abordées au 1.3.4.2. Les méthodes d'extrac-

[11] Par convention, nous avons pris l'hypothèse que le capteur de référence est le capteur gauche, il s'agit, bien évidemment, d'une convention purement arbitraire

tion et d'appariement de points d'intérêt sont directement transposables au cas de la stéréo, à ceci près que les points d'intérêts devront nécessairement être invariants en fonction d'un changement de point de vue relativement grand.

Afin de rester succincts, nous nous contenterons de présenter ici les approches majeures. Pour une revue plus exhaustive, ainsi qu'une évaluation et une méthodologie de comparaison des différents algorithmes proposés, le lecteur peut se tourner vers [SS02a].

1.4.2.1 Méthodes Locales

Les méthodes locales de calcul d'une carte de disparité, aussi appelées "approches par corrélation" ou "*Block-Matching*" reposent sur des comparaisons ponctuelles. Pour un point donné de l'image de référence, on va chercher à calculer un score reflétant sa similarité avec tous les candidats. Ces candidats sont les points situés sur la droite épipolaire. Cette mesure de similarité est généralement une approximation de corrélation, telles qu'on les a présentées au 1.3.4.2.

Cette opération repose sur plusieurs hypothèses. Tout d'abord, il est implicite que les objets imagés sont lambertiens, ce qui garantit que le motif d'illumination ne change pas suivant la direction d'observation.

Ensuite, il est implicite que, dans une fenêtre de corrélation, tous les points sont à la même profondeur. Plusieurs scénarios peuvent aller à l'encontre de cette hypothèse, dite de plan fronto-parallèle.

Le plan de la route, en premier lieu, ne peut pas être correctement étudié. Ce problème peut être contourné en utilisant une fenêtre de corrélation unidimensionnelle (hauteur nulle) les points de cette fenêtre se retrouvant alors *de facto* à la même profondeur [LAC06]. Une autre méthode permet de conserver le caractère discriminant d'un grand voisinage, sans souffrir des problèmes liés à la géométrie de la route. Williamson propose ainsi d'appliquer une homographie à l'une des deux images de façon à la rectifier [Wil99]. Ce procédé a, par la suite, été amélioré en appliquant cette homographie à la fenêtre de corrélation elle même [Per08].

Ensuite, il est fréquent que la fenêtre de corrélation contienne plusieurs objets, à différentes profondeurs, par exemple des zones de contours. Dans ces cas, l'utilisation de plusieurs fenêtres différentes a été avancée [HIG02], le recours à une approche multi-échelles peut également permettre d'affiner les résultats dans ce type de situation [BT99].

1.4.2.2 Méthodes Globales et Semi-Globales

Inspirées par l'algorithme d'HORN et SCHUNK, de nombreuses équipes ont cherché à représenter le problème de mise en correspondance stéréo sous la forme d'une fonctionnelle quadratique de la forme :

$$E^2 = \iint \Psi\left((I_{gauche}(u+d\,,v) - I_{droite}(u\,,v))^2\right) + \alpha(\nabla I_{gauche}(u+d\,,v) - \nabla I_{droite}(u\,,v))^2 \tag{1.18}$$

Où le premier terme impose le respect de la géométrie épipolaire et le second terme est un terme de lissage, venant pénaliser les forts gradients dans la carte de disparité.

Plusieurs méthodes d'optimisation numériques ont été proposées pour résoudre ce problème : des méthodes variationnelles [SM07], des algorithmes génétiques [GY02] ou encore des graphs-cuts [BVZ01].

Ces méthodes globales ont généralement été considérées comme très fiables, mais trop couteuses pour être implémentées en temps réel [SS02a]. En raison de ces problèmes d'implémentation, des méthodes dites semi-globales ont été introduites. Ces méthodes sont à mi-chemin entre les méthode locales, pour lesquelles la disparité d'un point ne dépend que de son voisinage, et les méthodes globales, pour lesquelles une optimisation sur tout le champ image va être menée.

Par exemple, une optimisation sur chaque ligne de l'image a été envisagée [SS02a]. Le formalisme de la programmation dynamique se prête en effet très bien à ce type de problème. Toutefois, les erreurs liées à cette approche sont fréquemment fortement corrélées. Les travaux d'HIRSCHMÜLLER ont apporté un éclairage nouveau à cette problématique. S'il s'agit toujours d'utiliser la programmation dynamique, plusieurs directions d'optimisation sont utilisées, typiquement 16. Cela a pour effet d'approximer une contrainte globale, sans pour autant avoir le même impact sur les ressources calculatoires[12][Hir05]. Ce type d'approche permet d'obtenir de bons résultats, tout en n'interdisant pas des implémentations temps-réel [EH08].

1.4.2.3 Filtrage - Améliorations

Quelle que soit le type de méthode envisagée, une performance temps-réel se paie généralement aux prix d'erreurs d'appariements. Un filtrage *a posteriori* des cartes de disparité est généralement nécessaire.

[12]Cependant, les ressources mémoires restent importantes et cela peut demeurer un problème pour certaines architectures.

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

FIGURE 1.10 – Comparaison de différentes méthodes de calculs de cartes de disparité, à partir de la séquences Tsukuba : (a) Une des images source; (b) Carte de disparité - Vérité Terrain; (c) Carte de disparité obtenue par une méthode locale [SS02a]; (d) Carte de Disparité obtenue par une méthode Semi-Globale [Hir05]

Tout d'abord, les zones d'occlusion peuvent poser problème[13]. Elles peuvent néanmoins être identifiées et retirées en réalisant une vérification croisée [ER02]. Il s'agit de calculer deux cartes de disparité, en prenant chacune des deux images comme référence, les points pour lesquels les résultats ne sont pas identiques sont supprimés. Étant donné que les coûts d'appariements n'ont besoin d'être calculés qu'une seule fois, le sur-coût occasionné est très faible.

Ensuite, des contraintes spécifiques au problème étudié peuvent être envisagées. On peut par exemple choisir de supprimer tous les points ne faisant pas partie d'un alignement vertical [BFT06]. Il est également possible d'utiliser une détection préalable de la route au moyen de la V-disparité afin d'optimiser les résultats [HLPA06].

La principale amélioration que l'on peut apporter aux résultats obtenus concerne leur précision. En effet, de nombreux développements ont été réalisés, visant à fournir des résultats sub-pixelliques. Une technique d'interpolation portant sur une image source [BT99] ou les deux [SS02a] permet ainsi d'obtenir une précision d'appariement sous la barre du pixel. Il est également possible de réaliser cette interpolation directement sur la carte de disparité [Cha05, HPON10].

1.4.3 Image 3D du mouvement 3D

S'il peut être immédiat de chercher à considérer directement le mouvement 3D dans l'espace objet, ce n'est pas nécessairement la meilleure solution. En effet, on peut généralement considérer que l'erreur commise lors du calcul d'une carte de disparité est isotrope (il est équivalent de surestimer une disparité donnée ou de

[13] On appelle zones d'occlusion les parties de la scène qui ne sont visibles que dans une des deux images

la sous-estimer). En revanche, ce caractère isotrope disparaît lors de la rétroprojection dans l'espace objet [DD02]. Dès lors, il peut être plus judicieux d'utiliser l'espace image pour exprimer les déplacements. Ainsi, l'image m' d'un point fixe M après un déplacement arbitraire est explicité en A.2 et est rappelé ici :

$$\left|\begin{array}{l} x'_m = \dfrac{x_m + y_m\omega_Z - f\omega_Y - \frac{T_X\delta_m}{b_s}}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z\delta_m}{fb_s} + 1} \\ y'_m = \dfrac{y_m - x_m\omega_Z + f\omega_X - \frac{T_Y\delta_m}{b_s}}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z\delta_m}{fb_s} + 1} \\ \delta'_m = \dfrac{\delta_m}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z\delta_m}{fb_s} + 1} \end{array}\right. \tag{1.19}$$

On choisit de noter :

$$\left|\begin{array}{l} x'_m \\ y'_m \\ \delta'_m \end{array}\right. = P_{(\boldsymbol{\Omega},\boldsymbol{T})}(m_i) \tag{1.20}$$

où $P_{(\boldsymbol{\Omega},\boldsymbol{T})}$ est alors la projection du point m_i dûe au mouvement $(\boldsymbol{\Omega},\boldsymbol{T})$.

Contrairement au résultat présenté en 1.3.3, l'expression ci-dessus est exacte et ne repose pas sur une approximation au premier ordre, généralement valide lorsque l'on s'intéresse au calcul du flot optique, lui même généralement limité par l'hypothèse des petits déplacements. Plus spécifiquement, la simplification apportée en 1.3.3 consiste à supposer que la profondeur des objets observés ne change pas entre deux images, ce qu'on ne peut plus considérer en s'intéressant au mouvement 3D.

1.4.4 Technologies de mesure de distance

D'autres technologies ont été développées afin de pouvoir obtenir des cartes de profondeur. Si elles peuvent parfois permettre d'obtenir des résultats plus précis, certaines présentent toutefois des inconvénients qui les rendent incompatibles, du moins à court terme, avec une application industrialisable. La plupart de ces solutions sont également plus limitées que la vision en termes de richesse de l'information fournie.

FIGURE 1.11 – RADAR utilisé pour un *Adaptative Cruise Control*, développé par Continental Automotive Systems

Radar - Le RADAR est un dispositif permettant de mesurer la distance d'un objet grâce à une mesure de temps de vol ou de la dispersion d'une impulsion électromagnétique, dans le domaine des radios-fréquence, et qui peut permettre d'estimer la vitesse relative de ces objets grâce à l'effet Doppler.

En particulier, les dispositifs dits *Adaptative Cruise Control* qui permettent de réguler la distance entre le véhicule équipé et le(s) véhicule(s) le précédant peuvent être implémentés grâce à des radars [GJB+00]. Le recours à la bande radio rend le RADAR robuste au brouillard, mais très sensible à la pluie.

Le RADAR présente cependant plusieurs inconvénients majeurs. Tout d'abord, certains objets sont peu réfléchissants dans la bande radios. En particulier, des piétons peuvent facilement ne pas être détectés. Ensuite, pour des RADARS à longue portée ($150m$), l'angle de vision est très étroit (environ 10^o) ce qui limite, dans les faits l'utilité, de l'AAC au domaine autoroutier. Finalement, le prix de ce type de dispositif (plusieurs centaines d'euros) le limite, au moins pour l'instant, aux véhicules haut de gammes.

Radar de Recul - Télémètres Acoustiques - L'appellation *radar de recul* est impropre, les dispositifs auxquels elle fait référence sont en effet des télémètres à impulsion acoustique, des sonars. Ce type de capteur est, en revanche, très bon marché et a déjà fait son apparition dans de nombreux modèles commerciaux. En revanche, la portée limitée de ce type de capteur (environ $3m$) le restreint en

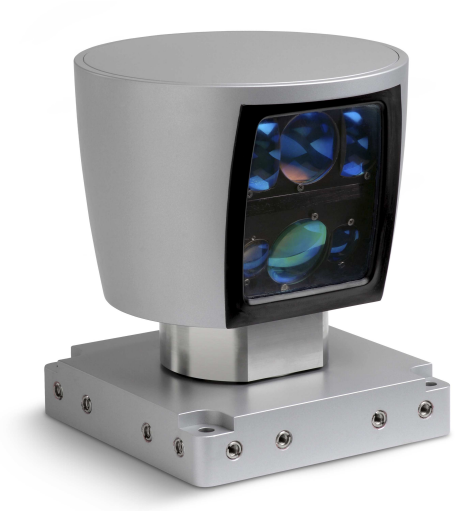


FIGURE 1.12 – LIDAR 64 nappes, commercialisé par *Velodyne*

pratique à des applications d'assistance au stationnement.

LIDAR - Les LIDAR sont des dispositifs purement optiques, reposant sur le principe de la télémétrie par temps de vol d'une impulsion laser. Le LIDAR génère un ou plusieurs plans de coupe, en balayant l'espace avec autant de nappes lasers. Ce type de dispositif allie une grande précision (qui peut être de l'ordre du centimètre) à une grande portée (80 à 100 mètres). Une nappe est généralement générée en utilisant un miroir rotatif, bien que des méthodes électro-optiques puissent être envisagées. Chaque nappe va générer un ensemble de points de mesure, qui pourrons ensuite être fusionnés [LRGA05].

Bien qu'étant un outil puissant, le LIDAR présente toutefois plusieurs inconvénients. Tout d'abord, son coût et sa fragilité (lié à la présence de pièces mécaniques) le rendent difficilement intégrable. Ensuite, et ce dernier point n'est pas exclusif au LIDAR, l'information fournie, bien que très précise est également très pauvre, faisant *in fine* du LIDAR un capteur qui ne peut être utilisé seul, mais qui va devoir être utilisé en conjonction avec d'autres sources d'information (centrale inertielle, caméras, etc.).

Éclairage structuré - Finalement, une dernière possibilité pour obtenir des relevés de profondeur à partir d'un capteur embarquable est le recours à l'illumination structurée. Il s'agit de projeter un motif géométrique connu sur la scène à

FIGURE 1.13 – Capteur Kinect®

imager et d'utiliser l'image rétro-diffusée de ce motif afin d'obtenir une lecture de la profondeur des différents points de la scène [FSMA10].

En particulier, le capteur Kinect®, commercialisé par Microsoft utilise ce principe. Même si ses capacités le cantonnent pour l'instant à des applications en intérieur (une portée d'environ $5m$ notamment), son prix et ses capacités d'imagerie en font un capteur intéressant pour la robotique.

1.5 Conclusion

Au cours de ce chapitre, nous sommes revenus sur plusieurs fondements de l'imagerie numérique. Nous avons cherché une modélisation pertinente de l'environnement de notre capteur, ainsi que de son mouvement. Nous sommes revenus sur deux des domaines de recherche les plus actifs en vision par ordinateur : l'estimation du flot d'une part et la stéréo-vision d'autre part. S'il est vrai que ces problèmes présentent certaines similitudes, il n'en demeure pas moins indispensable d'exploiter leur différences pour parvenir à des résultats précis.

Nous avons ainsi pu présenter l'état de l'art dans ces deux domaines et avons pu extraire plusieurs méthodes permettant d'estimer un champ de flot optique ou une carte de disparité. La connaissance de ces différentes techniques est indispensable avant de mener une réflexion sur une possible coopération entre les approches structurelle et dynamique.

Dans l'état actuel de notre réflexion, nous ne pouvons nous permettre d'arrêter des choix technologiques. En revanche, la connaissance que nous avons des différentes méthodes et approches envisagées dans la littérature nous permettra de

prendre les choix qui nous semblent les plus pertinents.

Chapitre 2

Odométrie Visuelle

«The greatest thing in this world
is not so much where we are,
but in which direction we are moving.»
Oliver Wendell Holmes Jr.

Sommaire

Pour un conducteur, qu'il soit humain ou robotique, il est fondamental de disposer d'informations sur le mouvement du véhicule contrôlé par rapport à son environnement. C'est ainsi que l'un des premiers capteurs effectivement embarqué dans les automobiles modernes est le compteur de vitesse. Cette information, bien qu'utile est extrêmement incomplète.

Une première étape vers un affinement de cette mesure est de considérer certaines rotations. Le modèle le plus simple est de prendre en compte uniquement la translation longitudinale et l'angle de lacet.

Toutefois, dès lors que l'hypothèse d'un monde plan n'est plus valide, ce modèle de mouvement est inadéquat. Il peut alors être envisagé d'évaluer la translation longitudinale, ainsi que les trois rotations, lacet, tangage et roulis. Si ce modèle permet de rendre plus fidèlement compte d'un monde complexe, il est en revanche mis à mal dans plusieurs situations. Tout d'abord, en cas de glissement des roues, l'absence de prise en compte d'une translation latérale peut rendre le système inutilisable. Ensuite, un écart angulaire entre le dispositif de mesure et l'horizontale engendre également des erreurs de mesure liées à l'absence de prise en compte d'une composante translationnelle verticale. En pratique, il est fréquent que les dispositifs de prise de vue stéréoscopique ne soient pas rigoureusement horizontaux, mais légèrement dirigés vers le bas, ceci pour optimiser le champ utile.

La connaissance précise des 6 composantes du mouvement 3D est incontournable pour la résolution d'un ensemble de problèmes. Tout d'abord, l'égo-localisation dans un monde tri-dimensionnel, ainsi que son corolaire, la planification de trajectoire, ne peuvent être envisagés de manière robuste qu'en ayant une connaissance complète de l'égo-mouvement. De la même manière, afin de pouvoir distinguer efficacement les objets dynamiques du fond statique, un système doit connaître son propre mouvement afin de pouvoir le compenser.

Chez l'Homme, deux modalités sensorielles remplissent cette fonction de localisation et d'estimation du mouvement. Le système vestibulaire (ou oreille interne) et la vision. Dans un contexte robotique, nous disposons d'un grand nombre de systèmes dédiés à l'estimation du mouvement. En particulier, les centrales inertielles, comme celle présentée en Fig. 2.1, permettent une estimation assez fine des taux de rotations et des accélérations translationnelles d'un véhicule. Cependant, la vision reste, chez l'Homme une composante indispensable de l'estimation du mouvement[1].

Le mimétisme anthropomorphique n'est cependant pas le seul argument en faveur d'un développement d'un système d'odométrie visuelle. En premier lieu,

[1] En particulier, un conflit entre les systèmes vestibulaire et visuelle est à l'origine du mal de mer.

si la localisation d'une automobile est généralement possible au moyen de GPS, il n'est pas toujours possible de disposer d'un réseau fiable de satellites [CMM06]. De plus, l'argument financier, mais également l'argument de la polyvalence jouent un rôle non-négligeable. La où la centrale inertielle ne peut répondre qu'à une problématique, la vision, en revanche, apporte une information extrêmement riche qui peut résoudre un grand nombre de problèmes [BBC+02, Dic02]. Dès lors, la vision apparaissant comme une solution meilleure marché et plus polyvalente, le développement de méthodes d'odométrie visuelle fiables et performantes est une évidence.

Dans un premier temps, nous reviendrons sur les différents systèmes décrits dans la littérature, permettant de remplir cette tâche, leurs avantages, ainsi que leurs inconvénients respectifs. Ensuite, en nous fondant sur cet état de l'art, nous décrirons brièvement les bases fondamentales d'un système d'odométrie visuelle. Nous décrirons ensuite le système, tel qu'il a été implémenté, nous présenterons et justifierons les différentes choix techniques qui ont été faits. Finalement, nous présenterons plusieurs méthodes d'évaluation des performances intrinsèques du systèmes, mais aussi en tant que maillon d'un réseau de capteurs dont les informations sont fusionnées.

FIGURE 2.1 – Centrale Inertielle Crossbow VG400®

2.1 L'Odométrie Visuelle - Retour sur 20 ans d'évolutions

L'odométrie visuelle est un problème complexe. En effet, nous avons vu au chapitre 1 que l'image d'un déplacement translationnel et la structure 3D de la scène sont intimement liées. Il est donc évident que pour pouvoir estimer les 6 composantes du mouvement $(\boldsymbol{T}, \boldsymbol{\Omega})$, il est nécessaire de disposer d'appariements temporels et d'une information 3D sur la scène observée. Cette nécessité d'une collaboration entre approches temporelles et approches structurelles a longtemps freiné les développements de méthodes performantes.

Malgré cet état de fait, certaines méthodes purement monoculaires ont été proposées. Toutefois, ces méthodes reposent sur des hypothèses fortes et limitantes. Par exemple, SCARAMUZZA *et al.* ont proposé récemment une méthode purement monoculaire reposant sur l'utilisation d'une caméra omnidirectionnelle et d'une paramétrisation du mouvement à deux dimensions uniquement [SFS09]. De même, STEIN *et al.* proposent de limiter les degrés de liberté du véhicule à 3, avant de procéder à plusieurs estimations locales du mouvement, et de fusionner ces estimations de façon à rendre le procédé robuste [SMS00]. Irani *et al.* proposent de compenser les rotations en recalant l'image d'un plan sur elle même [IRP97], puis d'estimer les translations par localisation du Foyer d'Expansion[2]. Cette localisation du FoE est une approche fréquente, en effet, le FoE regroupe toute l'information translationnelle accessible à un système monoculaire. Plusieurs approches ont donc été proposées, depuis le début des années 80.

Ainsi, les premières méthodes de localisation du FoE reposaient sur l'utilisation de primitives robustes de l'image [Pra79, RA80]. En effet, au début des années 80, les contraintes technologiques limitaient alors les approches envisageables. Ces méthodes n'étaient pas très robustes, du fait de leur utilisation très partielle de l'information. Avec l'émergence de méthodes de calculs du flot optique, des méthodes plus globales furent introduites. La plupart utilise le fait que le FoE est le point de convergence des lignes de champ du flot optique. Ainsi, une méthode commune est de recourir à un vote cumulatif, chaque vecteur élémentaire va voter pour les points appartenant à la demi-droite qu'il porte [SJBK08]. L'utilisation d'un filtre, reposant sur le changement de signe des deux composantes du flot a également été envisagée [SRR04]. Des méthodes plus classiques de résolution ont également été investiguées, comme l'utilisation d'une régression aux moindres carrés [XD92]. L'exploitation de certaines propriétés du flot optique du point de vue de l'analyse vectorielle a également été envisagée [HS93]. Toutefois, ce genre de méthode repose sur des approches différentielles (*via* le calcul de la divergence et du rotationnel) ce qui les rend très sensibles au bruit. Finalement, des méthodes reposant sur des considérations géométriques ont été décrites [WWH07, TSO+05]. Ces dernières présentent un inconvénient majeur en termes de temps de calcul qui les rend, pour l'instant, inadaptées aux besoins d'applications temps-réel.

Parmi les approches monoculaires, une place particulière doit être réservée aux

[2]Le Foyer d'Expansion (*Focus of Expansion*, *FoE*) est le point d'intersection des lignes de champ du flot optique, ce point résume également l'intégralité de l'information translationnelle disponible pour un système monoculaire. Nous reviendrons plus avant sur sa description lors du chapitre 4.

approches de type *SLAM*[3]. Le *problème SLAM* consiste, pour un robot autonome, à estimer en même temps la topographie de son environnement et sa position dans celui-ci. Ce problème a été très discuté au cours de la dernière décennie, avant d'atteindre un stade de relative maturité. Les méthodes de SLAM reposent donc sur une estimation jointe de la structure de la scène et du mouvement du capteur. Cette estimation, probabiliste, est fréquemment fondée sur une approche particulaire [MTKW02, MTKW03] ou encore sur un filtrage de KALMAN étendu [DNC+01]. Pour faire face à la grande dimensionnalité du problème, la plupart des auteurs emploient un modèle limité du mouvement [DWB06, BDW06]

Cependant, parmi les approches de SLAM, l'approche MonoSLAM [DRMS07] présente la particularité de ne reposer que sur l'utilisation de la vision. Toutefois, des hypothèses assez fortes doivent être prises, de façon à lever certaines indéterminations, en l'occurrence, il est nécessaire, pour cette méthode, d'être initialisée sur un motif prédéfini.

Dans l'absolu, les seules méthodes fondées uniquement sur la vision, et permettant une extraction complète de l'égo-mouvement d'un véhicule sont celles tirant partie de l'information stéréoscopique. Ces méthodes reposent sur une collaboration entre une information temporelle et l'information issue de la stéréovision. Le déroulement de ces méthodes reste sensiblement constant :

1. Acquisition d'une paire stéréo.
2. Extraction/Appariement de points d'intérêt.
3. Rétro-projection de ces points dans l'espace objet.
4. Minimisation de l'erreur de rétro-projection.

Une revue de différentes approches respectant ce schéma peut être trouvée dans [SP07]. Diverses contributions ont été apportées à ce modèle de base. Ainsi une contrainte additionnelle de lissage peut être ajoutée [Bad07], de manière à rendre le procédé plus robuste. De la même façon, un filtrage de type KALMAN est fréquemment envisagé, sous la forme d'une filtre de KALMAN étendu [CMR07] ou non-parfumé [JU97, KGL10].

Une contrainte qui peut également être utilisée, notamment pour la résolution du problème de minimisation, est celle de rigidité du faisceau lumineux imagé. Cette contrainte est à l'origine des approches de *Bundle Adjustement*, et en particulier de *Sparse Bundle Adjustement*(SBA) quand on s'intéresse aux approches

[3] *Simlutaneous Localization and Mapping*, Localisation et Cartographie Simultanées

éparses [ESN06]. Ce type d'approche a notamment été utilisé dans [KAS11] et [LDD+06].

Toutefois, cette description reste schématique. En effet, par exemple DEMIRDJIAN *et al.* ne rétro-projettent pas les points extraits dans l'espace objet et préfèrent rester dans l'espace image où le bruit d'appariement reste isotrope [DD02].

Une approche très populaire est l'utilisation du tenseur tri- ou quadri-focal [CMR07, CMR10, KGL10, RFB09]. Il s'agit fondamentalement d'une utilisation des concepts issus de la stéréo-vision multi-caméras [HZ03]. Implicitement ou explicitement, les auteurs de ces études considèrent que les paramètres extrinsèques du capteur stéréo sont variables dans le temps et les ré-estiment à chaque itération. Pour notre part, nous considérons que ce n'est pas une piste à privilégier. En effet, même s'il est indéniable que tout dispositif mécanique est susceptible de se déformer, nous pensons que ces déformations peuvent être considérées comme négligeables devant le mouvement global du véhicule. Si ces déformations peuvent être prises en compte périodiquement, chercher à les estimer systématiquement complexifie inutilement le problème d'odométrie visuelle. Considérer que les paramètres extrinsèques du capteurs stéréo sont variables se fait nécessairement au détriment de la précision ou des temps de calculs.

Certaines méthodes atypiques s'éloignent de cette approche. Ainsi, OBDRZALEK *et al.* proposent une méthode reposant sur un vote cumulatif permettant d'estimer les rotations dans un premier temps puis d'effectuer un recalage plus traditionnel pour estimer les translations [OM10]. Bien qu'innovante, cette méthode n'est pas, pour l'instant, très performante, ne serait-ce que parce qu'elle repose sur l'extraction de points situés à l'infini.

L'approche que nous proposons et présentons dans la suite est sensiblement différente. En effet, nous montrerons que le problème d'estimation du mouvement peut être posé de façon purement linéaire, sans pertes de précision, mais avec des performances supérieures en termes de temps de calculs. Un tel formalisme a été esquissé dans [TM04]. Toutefois, notre approche est moins restrictive en cela qu'elle ne repose pas sur des approximations de l'expression du mouvement. De plus, notre approche permet l'exploitation complète de l'information et mène à des résultats supérieurs à ceux présentés[4].

[4] TALUDKER *et al.* ne présentent pas de résultats quantifiés, mais il apparaît que leur méthode mène à une extraction assez mauvaise de l'égo-mouvement.

2.2 Approche Proposée

L'implémentation proposée repose sur l'expression de la projection 3D du mouvement que nous avons exposé dans le chapitre précédent, en équation 1.19 :

$$\left|\begin{array}{l} x'_m = \dfrac{x_m + y_m\omega_Z - f\omega_Y - \frac{T_X\delta_m}{b_s}}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z\delta_m}{fb_s} + 1} \\ \\ y'_m = \dfrac{y_m - x_m\omega_Z + f\omega_X - \frac{T_Y\delta_m}{b_s}}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z\delta_m}{fb_s} + 1} \\ \\ \delta'_m = \dfrac{\delta_m}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z\delta_m}{fb_s} + 1} \end{array}\right. \tag{2.1}$$

Comme nous l'avons vu, nous avons besoin de deux informations complémentaires : l'information structurelle 3D sur la scène et l'information temporelle de mouvement. A partir de ces informations, nous pouvons construire un ensemble, dense ou éparse, de correspondances entre les deux images :

$$\mathcal{S} = \{m_i, i \in [1;N]\} \rightarrow \left\{m'_i, i \in [1;N]\right\} \tag{2.2}$$

où les m_i sont des points images de coordonnées :

$$m_i = \left|\begin{array}{l} x_{m_i} \\ y_{m_i} \\ \delta_{m_i} \end{array}\right. \tag{2.3}$$

A partir de cet ensemble de correspondances, et des équations vues en annexe A.17, il est possible de construire le système suivant, linéaire en $(\boldsymbol{\Omega}, \boldsymbol{T})$:

$$\begin{pmatrix} -\frac{x_{m'_1}.y_{m_1}}{f} & \frac{x_{m_1}.x_{m'_1}}{f} + f & -y_{m_1} & \frac{\delta_{m_1}}{b_s} & 0 & \frac{\delta_{m_1}.x_{m'_1}}{b_s} \\ f - \frac{y_{m_1}.y_{m'_1}}{f} & \frac{x_{m_1}.y_{m'_1}}{f} & -x_{m_1} & 0 & \frac{\delta_{m_1}}{b_s} & \frac{\delta_{m_1}.y_{m'_1}}{b_s} \\ \frac{\delta_{m'_1}.y_{m_1}}{f} & -\frac{\delta_{m'_1}.x_{m_1}}{f} & 0 & 0 & 0 & \frac{\delta_{m_1}.\delta_{m'_1}}{b_s} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -\frac{x_{m'_N}.y_{m_N}}{f} & \frac{x_{m_N}.x_{m'_N}}{f} + f & -y_{m_N} & \frac{\delta_{m_N}}{b_s} & 0 & \frac{\delta_{m_N}.x_{m'_N}}{b_s} \\ f - \frac{y_{m_N}.y_{m'_N}}{f} & \frac{x_{m_N}.y_{m'_N}}{f} & -x_{m_N} & 0 & \frac{\delta_{m_N}}{b_s} & \frac{\delta_{m_N}.y_{m'_N}}{b_s} \\ \frac{\delta_{m'_N}.y_{m_N}}{f} & -\frac{\delta_{m'_N}.x_{m_N}}{f} & 0 & 0 & 0 & \frac{\delta_{m_N}.\delta_{m'_N}}{b_s} \end{pmatrix} \cdot \left|\begin{array}{l} \boldsymbol{\Omega} \\ \boldsymbol{T} \end{array}\right. = \begin{pmatrix} x_{m_1} - x_{m'_1} \\ y_{m_1} - y_{m'_1} \\ \delta_{m_1} - \delta_{m'_1} \\ \ldots \\ x_{m_N} - x_{m'_N} \\ y_{m_N} - y_{m'_N} \\ \delta_{m_N} - \delta_{m'_N} \end{pmatrix} \tag{2.4}$$

Que l'on peut réécrire :

$$\boldsymbol{\Sigma}.\left|\begin{array}{c}\boldsymbol{\Omega}\\ \boldsymbol{T}\end{array}\right. = \boldsymbol{M} \tag{2.5}$$

Il est intéressant de noter que, sous cette forme, le problème d'estimation de l'égo-mouvement d'un capteur stéréo est un problème linéaire, là où la plupart des auteurs travaillent avec des formulations fortement non-linéaires. Nous ne pouvons pas expliquer, *a priori*, le fait que cette formulation, pourtant très naturelle, ne soit pas plus utilisée. Il est possible que le fait de raisonner de bout en bout dans l'espace image mène plus facilement à ce résultat.

Dans le cas général, ce système est sur-contraint et n'admet donc pas de solution régulière, il est alors équivalent de trouver le couple $\left(\tilde{\boldsymbol{\Omega}}, \tilde{\boldsymbol{T}}\right)$ qui minimise l'erreur suivante :

$$\varepsilon^2 = \|\boldsymbol{M'} - \boldsymbol{\Sigma}.\left|\begin{array}{c}\boldsymbol{\Omega}\\ \boldsymbol{T}\end{array}\right. \|_2 \tag{2.6}$$

Cette dernière erreur peut encore être réécrite de la façon suivante :

$$\varepsilon^2 = \sum_{\mathcal{S}} \|m'_i - P_{(\boldsymbol{\Omega},\boldsymbol{T})}(m_i)\|_2 \tag{2.7}$$

Avec cette dernière expression, il est alors facile d'interpréter la quantité minimisée comme étant l'erreur de recalage commise en estimant incorrectement le mouvement.

Un certain nombre de points méritent d'être étudiés. Tout d'abord, l'ensemble $\mathcal{S}$ peut contenir des points aberrants, il est nécessaire de recourir à un procédé robuste à ce niveau. Ensuite, la résolution aux moindres carrés de l'équation 2.6 peut être obtenue de plusieurs façons différentes que nous allons comparer. Finalement, nous nous pencherons sur l'impact des techniques choisies pour réaliser les appariements stéréo et temporels sur les résultats de notre système d'odométrie visuelle.

2.2.1 Élimination des Points Aberrants

un point aberrant (ou *outlier*) peut être la conséquence de trois phénomènes :

- Il peut tout d'abord être l'image d'un point objet en mouvement, et pour lequel un modèle de mouvement global n'est pas satisfaisant.
- Une erreur lors des appariements temporels peut survenir. C'est notamment le cas pour les motifs répétitifs produisant des ambiguïtés d'appariement (marquage au sol, mobilier urbain, ...).

Entrées : $\mathcal{S}$, Th, $N_{itérations}$, $N_{échantillons\_mini}$, modèle_précédent
Sorties : ε, modèle, consensus

début
  $\varepsilon = \infty$
  consensus $= \{\emptyset\}$
  modèle = modèle_précédent
  si `Taille(`$\mathcal{S}$`)` <10 **alors**
    retourner ε, modèle, consensus
  fin
  itrations $= 0$
  tant que itérations $< N_{itérations}$ **faire**
    échantillon = `ChoisirÉchantillon(`$\mathcal{S}$`,3)`
    modèle$_{\text{hypothèse}}$ = `Resoudre(`échantillon`)`
    pour tous les $(m_i \rightarrow m'_i) \in \mathcal{S}$ **faire**
      si `Dist(`$m'_i, P_{\text{modèle}_{\text{hypothèse}}}$`)` $<$ Th **alors**
        consensus$_{\text{hypothèse}}$ =`Ajoute(`i`)`
      fin
    fin
    si `Taille(`consensus$_{\text{hypothèse}}$`)` $> N_{échantillons\_mini}$ **alors**
      modèle$_{\text{hypothèse}}$=`Resoudre(`consensus`)`
      erreur$_{\text{hypothèse}}$ =`Erreur(`modèle$_{\text{hypothèse}}$, consensus$_{\text{hypothèse}}$`)`
      si erreur$_{\text{hypothèse}}$ $<$ ε **alors**
        $\varepsilon =$ erreur$_{\text{hypothèse}}$
        modèle = modèle$_{\text{hypothèse}}$
        consensus = consensus$_{\text{hypothèse}}$
      fin
    fin
  fin
fin

Algorithme 1: RANSAC - cas particulier de l'estimation de l'égo-mouvement

- De la même façon, des erreurs d'appariements peuvent survenir lors du calcul des appariements stéréo.

Dans ces circonstances, il est nécessaire d'utiliser un processus robuste vis-à-vis de ces points aberrants, potentiellement nombreux. Nous avons choisi d'utiliser un processus de RANSAC (*RANdom SAmple Consensus*) [FB81]. Cet algorithme est illustré dans l'algorithme 1.

Le grand intérêt de cette méthode est qu'elle suppose uniquement qu'une majorité relative de points vont être statiques. Dans notre cas, nous considérons qu'il n'y aura jamais moins de 30% d'*inliers*[5]. C'est pourquoi si $\mathcal{S}$ contient moins de 10 appariements, nous considèrerons le système insoluble. En effet, 3 points sont nécessaires pour résoudre le système, 2 points pouvant aboutir à des indéterminations, et plus particulièrement à la construction de système de rang inférieur à 6. En pratique toutefois, cette limite n'a jamais été atteinte, quelles que soient les résolutions d'images et méthodes d'appariement utilisées.

Les autres paramètres de cet algorithme sont également fixés en prenant comme bases des considérations statistiques.

Nombres d'Itérations - Le nombre d'itérations est fixé de façon à ce que la probabilité de sélectionner au moins une fois uniquement des points pertinents (*inliers*) soit supérieure à un p fixé.

Pour cela, nous notons q la probabilité de piocher un *inlier* dans $\mathcal{S}$, dans notre cas, nous avons évalué le pire cas possible à $q = 0,3$. La probabilité que l'échantillon initial contienne au moins un *outlier* est donc :

$$1 - q^3 \tag{2.8}$$

Il s'agit ici d'une borne supérieure. En effet, les tirages effectués sont sans remise, la probabilité de piocher un *outlier sachant que* nous venons d'en piocher un est donc légèrement inférieure à $(1-q)$. Toutefois, nous cherchons ici à évaluer le pire cas possible, ce n'est donc pas gênant.

La probabilité que toutes les itérations donnent lieu à des tirages comprenant au moins un *outlier* est donc :

$$(1 - q^3)^{N_{\text{itérations}}} \tag{2.9}$$

Notre condition s'énonce donc :

$$1 - p = (1 - q^3)^{N_{\text{itérations}}} \tag{2.10}$$

D'où :

$$N_{\text{itérations}} = \frac{\ln(1-p)}{\ln(1-q^3)} \tag{2.11}$$

La relation numérique entre le nombre d'itérations, et la probabilité de réaliser au moins un *bon* tirage est illustré en table **??**[6].

[5] Points respectant le modèles dominant, par opposition à *outliers*

[6] Un *bon* tirage est un tirage ne comportant que des *inliers*

Probabilité de réaliser au moins un bon tirage	Nombre d'itérations
0,9	85
0,99	169
0,999	253
0,9999	337

TABLE 2.1 – Influence du nombre d'Itérations de RANSAC sur la probabilité de fausse mesure

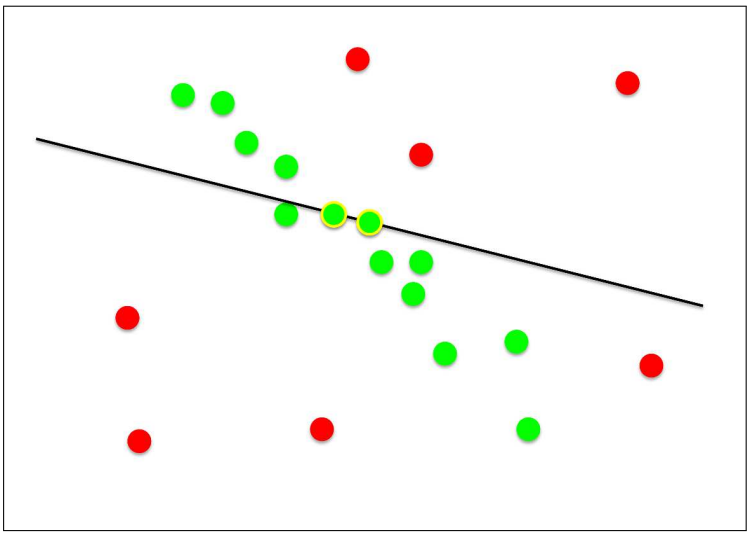


FIGURE 2.2 – Extraction d'*inliers* menant à une mauvaise estimation du modèle dominant

Toutefois, il est important de se souvenir que p n'est que la probabilité de ne sélectionner que des *inliers*. Ce n'est pas la probabilité d'estimer correctement le modèle à partir d'un échantillon donné. Ainsi, la figure 2.2 illustre un cas d'utilisation de RANSAC pour estimer un modèle de droites à partir de données bruitées contenant des *outliers*, pour lequel la sélection d'un échantillon composé uniquement d'*inliers* conduit à une mauvaise estimation.

Dans ces conditions, fixer un nombre maximum d'itérations nous semble vain. C'est pourquoi notre implémentation de l'algorithme RANSAC ne repose pas sur un nombre maximum d'itérations, mais sur un temps maximum fixe. Ce temps est fixé à 10*ms*. Nous avons constaté que cela correspond (sur la machine sur laquelle les développements de ce travail ont été menés) à une moyenne d'environ 300 itérations, soit une probabilité de sélectionner au moins un échantillon composé

exclusivement d'*inliers* de $0,9997$.

Comme nous le verrons en 2.2.2, le temps d'exécution par itération n'est pas constant, et est influencé par l'échantillon initial et le nombre de points dans le consensus. En d'autres termes, un échantillon initial qui permet d'estimer un "bon modèle" pour lequel le consensus va être important conduira à une itération plus longue, car le système à résoudre (en ligne 18 de l'algorithme 1) sera plus grand. En contrepartie, cette itération plus longue aura plus de chances de mener à une estimation fiable du mouvement qu'une itération pour laquelle l'échantillon initial est composé d'*outliers*, et dont le consensus restera vide. Nous pensons que cette approche permet d'une part de faciliter le respect des contraintes temps réel, et d'autre part d'accorder mécaniquement un poids plus important aux itérations "intéressantes" du processus.

2.2.2 Résolution du système

Le formalisme que nous employons, nous permet de considérer le problème à résoudre, présenté en équation 2.6, comme étant un problème des moindres carrés linéaires. Ce type de problème a longuement été disserté, notamment dans [Bjö96]. Nous allons ici comparer plusieurs méthodes de résolution, en termes de précision et de performances. Nous allons nous intéresser plus particulièrement aux méthodes suivantes [SB02, LH95] :

- la Décomposition en Valeurs Singulières (*SVD, Singular Values Decomposition*).
- la Factorisation *QR*.
- la Décomposition Orthogonale Complète (*Complete Orthogonal Decomposition, COD*).

Ces méthodes sont décrites en Annexe B.

Afin de pouvoir arrêter un choix, nous allons comparer ces différentes méthodes en termes de précision et de temps de calcul. Les implémentations que nous allons utiliser sont librement disponibles au sein de la librairie LAPACK[7], considérée comme un mètre étalon en termes de précision numérique et de performances.

Les critères de comparaison sont définis comme suit :

- Le temps de calcul nécessaire à la résolution d'un système, exprimé en nombre de cycles processeur.

[7] http://www.netlib.org/lapack

Méthode	Précision	Temps de Calcul
SVD	$1.0 \cdot 10^{-3}$	50 730
QR	$1.4 \cdot 10^{-3}$	24 200
COD	$1.9 \cdot 10^{-3}$	25 400

TABLE 2.2 – Benchmark des 3 méthodes de résolution numériques envisagées sur un système de petite taille

- La précision des résultats obtenus, calculée comme étant l'erreur relative entre une vraie solution et la solution obtenue.

Dans un souci d'exhaustivité, nous considèrerons deux scénarios différents, représentatifs du déroulement de l'algorithme :

- Tout d'abord, nous nous pencherons sur un système généré à partir de 3 correspondances, ce qui correspond à la résolution intervenant à la ligne 11 de l'algorithme 1.

- Ensuite, nous considèrerons un cas plus ouvert, en faisant évoluer le nombre d'éléments à l'origine du système, ce cas représente la résolution intervenant en ligne 18 de l'algorithme 1.

Dans les deux cas, nous fixons un mouvement arbitraire et aléatoire, et générons un ensemble de correspondances, dénuées de bruit.

Les résultats de ces tests sont disponibles en table **??** et en figure 2.3. Il apparaît en particulier que la COD ne présente pas d'intérêt dans notre cas particulier. En effet, comme décrit en annexe B, la COD peut être vue comme une variante de la décomposition QR, garantissant une grande robustesse aux systèmes mal conditionnés. Cette robustesse ne présente pas d'intérêt pour nos applications et la perte conséquente de précision est rédhibitoire. Ces tests nous permettent également et surtout de sélectionner, pour chaque type de résolution la meilleure méthode :

- Les *petits* systèmes (ligne 11 de l'algorithme 1) seront résolus par factorisation *QR*. En effet, ces résolutions seront les plus nombreuses, une méthode rapide est donc intéressante. Si l'erreur numérique commise est légèrement plus élevée qu'avec d'autres méthodes, cette différence n'est pas pertinente car elle n'impacte pas les résultats finaux.

- Les systèmes complets (ligne 18) seront résolus par Décomposition en Valeurs Singulières. En effet, si pour des petits systèmes, les temps de calculs

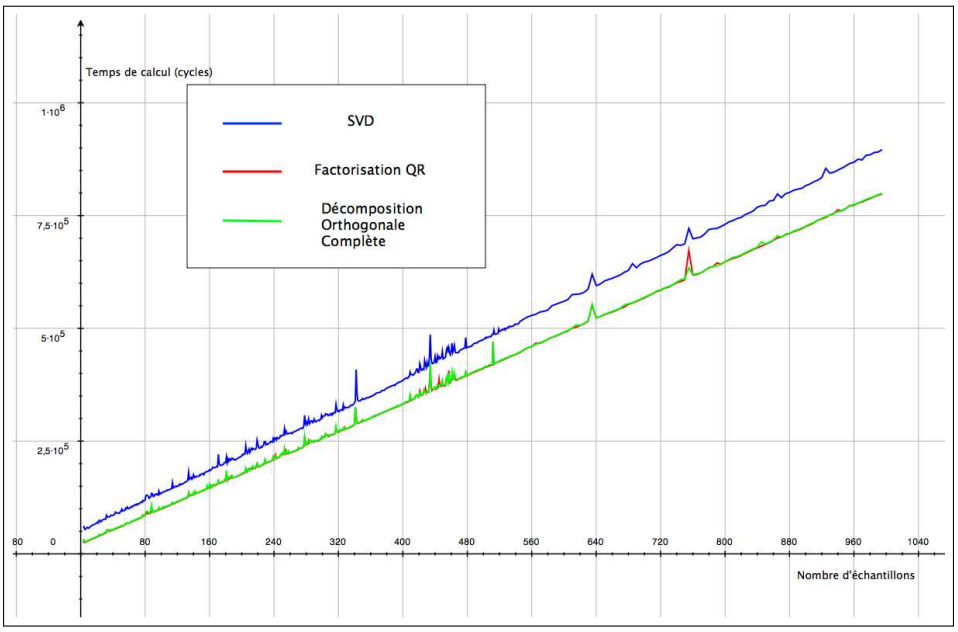


FIGURE 2.3 – Benchmark des 3 méthodes de résolution numériques envisagées sur un système de taille variable

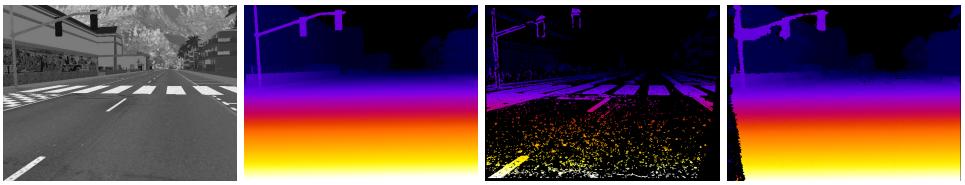
pour chaque méthode varient du simple au quadruple, pour des systèmes plus lourds, cette différence s'estompe. D'autre part, pour ces systèmes, il peut être intéressant de disposer du maximum de précision.

2.2.3 Extraction de l'information 3D

Comme nous l'avons vu chapitre 2, de nombreuses méthodes de calculs de carte de disparité sont disponibles. Si l'on exclut toute considération d'ordre calculatoire, pour ne se focaliser que sur les résultats obtenus, deux aspects nous semblent importants :

- La densité des cartes de disparité obtenues.
- Le caractère subpixellique de l'estimation réalisée.

Nous allons donc mener une étude de l'impact de ces deux facteurs sur les résultats obtenus. Pour cela, deux algorithmes de calcul de cartes de disparité seront utilisés. Tout d'abord, l'algorithme de Block-Matching (BM) initialement développé au Livic [HLPA06], qui nous permet d'obtenir des cartes semi-denses et



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

FIGURE 2.4 – Comparaison des deux méthodes de calcul de cartes de disparité : (a) : Image Source (b) : Carte Calculée à partir du Z-Buffer de SiVIC (Vérité Terrain); (c) : Carte calculée à partir d'une méthode locale semi-dense [HLPA06]; (d) Carte Calculée à partir d'une méthode Semi-Globale [EH08]

pixelliques et d'autre part une version modifiée[8] de l'implémentation de l'algorithme d'HIRSCHMULLER(SG) que l'on peut trouver au sein de la bibliothèque libre OpenCV[9]. Nous utiliserons ici le simulateur SiVIC (décrit en annexe D.1) qui nous permet d'obtenir une vérité terrain absolue des cartes de profondeur attendues (VT). Cette vérité terrain, ainsi que les résultats des deux algorithmes mentionnés sont en figure 2.4.

De façon à comparer ces différentes approches, nous avons généré une séquence de 600 paires stéréo. Cette séquence prend place en milieu urbain, dans des conditions de trafic légères. Ce scénario comprend notamment plusieurs éléments traditionnellement problématiques pour ce genre de méthode (ronds-points, motifs répétitifs, ...). Notre méthode requérant des points d'intérêt, nous avons choisi d'utiliser des appariements temporels obtenus par la mise en correspondance de points SURF, étant donné qu'il s'agit *a priori* de la méthode la plus précise. Les résultats obtenus vont être évalués en utilisant comme critère la moyenne des écarts-types (obtenus par répétabilité) sur toute la séquence. Ces résultats seront normalisés de façon à ce que la carte parfaite présente une erreur unitaire.

Afin de pouvoir déterminer l'impact de la densité de carte de disparité, et l'impact de la précision subpixellique de celle-ci, les différents cas de figures suivants seront évalués :

- Carte parfaite (VT).
- Carte semi-dense pixellique (BM).

[8]Ces modifications ont pour effet une amélioration sensible des performances, ainsi qu'une meilleure prise en charge des bords de l'image. Elles n'ont pas été intégrées dans la version principale de la librairie.

[9]http://opencv.willowgarage.com

Méthode	Erreur Normalisée
VT	1,0
BM	1,5
SG	1,1
SG pixellique	1,2
SG semi-dense	1,4

TABLE 2.3 – Comparaison de l'influence des différentes stratégies de calculs de cartes de disparité sur l'odométrie visuelle

- Carte dense sub-pixellique (SG).
- Carte dense pixellique (SG sans affinement sub-pixellique).
- Carte semi-dense sub-pixellique (SG filtrée de façon à présenter la même densité que la méthode BM).

Les résultats de ces tests figurent en tableau 2.3. Il apparaît que l'impact de la densité de la carte de disparité est prépondérant, par rapport à la précision sub-pixellique de celle-ci. En effet, avec une carte de disparité plus dense, le nombre d'appariements utilisés pour la résolution du problème d'odométrie est plus important. Il est également important de considérer que cet algorithme d'odométrie visuelle constitue la première étape d'un système de détection d'objets dynamiques. Comme nous le verrons au chapitre 3, la densité de la carte de disparité joue un rôle primordial lors de cette détection. En effet, une détection ne sera possible que dans les zones de l'image pour lesquelles deux valeurs de disparité consécutives sont disponibles. C'est également pour cette raison que nous n'avons pas testé de méthode complètement éparse, comme par exemple l'extraction/appariement de points SURF dans les images gauche et droite.

2.2.4 Extraction de l'information temporelle

Tout comme pour l'extraction de l'information 3D, il existe plusieurs moyens, abordés au chapitre 1, qui peuvent nous permettre d'obtenir des correspondances temporelles. Afin de pouvoir arrêter un choix technique pertinent, nous allons comparer les résultats obtenus en utilisant différentes méthodes, en conjonction avec une carte de disparité obtenue par une méthode semi-globale, telle qu'exhibée plus haut.

Les méthodes que nous avons choisi de comparer sont les suivantes :

- L'extraction et l'appariement de points SURF [BTV06].
- Un suiveur KLT appliqué sur des points de HARRIS [TK92, HS88].
- L'information dense issue d'un flot optique de type LUCAS et KANADE [BC05].

En effet, le KLT s'est imposé comme le standard *de facto* pour les applications de robotique mobiles, les points SURF peuvent *a priori* nous apporter plus de précision et de robustesse, notamment vis-à-vis de grands mouvements.

Les temps de calculs pour ces différentes méthodes étant sensiblement identiques, le seul critère de choix pertinent va donc être la précision des résultats obtenus. C'est donc en ces termes que nous allons comparer les trois approches. Le protocole de comparaison que nous allons suivre reste sensiblement le même qu'au 2.2.3, à un détail près, nous allons cette fois faire varier la cadence des images. En effet, si, pour de nombreux auteurs, la cadence est considéré comme un indicateur de performances, nous pensons qu'il s'agit, au même titre que la longueur focale ou la base stéréoscopique d'une variable d'ajustement. En particulier, nous pensons que le choix de la cadence image peut avoir un impact fondamental sur la sensibilité d'un système de détection d'objets dynamiques. Nous reviendrons plus en détails sur ce point au chapitre 3.

Il apparaît que l'information issue de correspondances de points SURF est très peu, voire pas du tout, influencée par une variation de cadence image. Au contraire, les correspondances obtenues à partir d'une méthode différentielle tendent à se dégrader, voire à devenir occasionnellement instables (voir fig. 2.5).

En conclusion, si pour des vitesses "urbaines" et des cadences d'acquisition de l'ordre d'une vingtaine de Hertz, les différentes méthodes sont équivalentes, cela n'est plus vrai pour des cadences plus basses, ou des vitesses plus élevées. La plupart des méthodes denses d'estimation du mouvement, nous l'avons vu précédemment, peuvent recourir à un schéma pyramidal ou itératif afin d'être robustes vis-à-vis de grands déplacements. C'est le cas des méthodes que nous avons utilisé ici. Toutefois, l'appariement de points SURF présente un avantage considérable : la polyvalence. En effet, et contrairement aux méthodes fondées sur le flot optique, à temps de calcul constant, et surtout sans un paramétrage supplémentaire lié à la configuration matérielle, l'appariement de points SURF peut être utilisé, quelques soient la cadence image et la vitesse de déplacement attendue du véhicule. Or, à ce stade de notre étude, il est impossible de statuer quant à l'importance de cette polyvalence, et devant la différence très marginale en termes de temps de calcul, le choix d'utiliser la mise en correspondance de points SURF pour l'extraction de l'information temporelle nous semble évident.

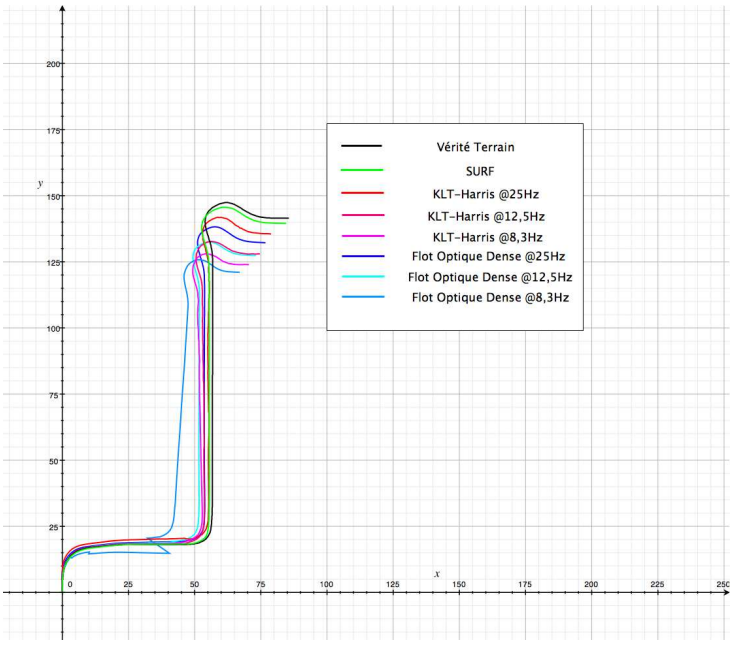


FIGURE 2.5 – Trajectoires obtenues par Odométrie Visuelle, en utilisant différentes méthodes d'extraction de l'information temporelle

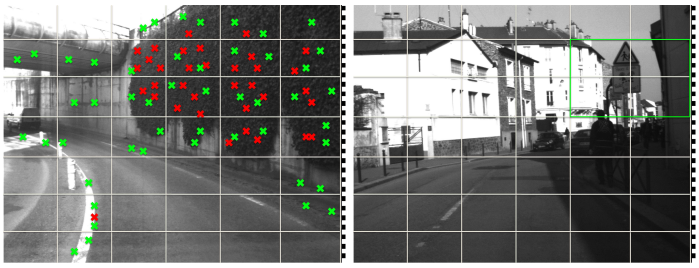
2.2.5 Homogénéité Spatiale

Sous certaines conditions, il est possible que les points d'intérêt extraits de l'image source ne soient pas répartis de façon homogène. Cela est particulièrement vrai lorsque, par exemple, un bâtiment très texturé occupe l'arrière plan de l'image. Dans ces conditions, un consensus non représentatif peut être extrait, et conduire à la construction d'un système mal conditionné, et donc à une estimation erronée de l'égo-mouvement.

Deux solutions à ce problème ont été envisagées :

- Le "*bucketing*" [ZDFL95], qui va consister à supprimer *a posteriori* des points dont l'information est trop redondante.
- L'extraction multi-passes, qui va consister à rechercher des points supplémentaires dans les zones sous-représentées.

Bucketing - Le "bucketing" consiste à découper l'image en un nombre prédéfini de zones (des "*buckets*") et à ne conserver qu'un nombre maximum de point d'intérêt dans chaque "*bucket*". Cette technique a, en particulier, été utilisée dans [KGL10].



(a) (b)

FIGURE 2.6 – Principe et limite du bucketing; (a) Illustration du principe de "*bucketing*" Vert : points conservés, rouge : points éliminés; (b) Limite du "*bucketing*" : les quatre "*buckets*" en vert sont composés d'objets à plusieurs profondeur

La figure 2.6(a) illustre ce procédé. En l'absence de "*bucketing*", un consensus aurait été trouvé dans la zone supérieure droite de l'image, sur-représentée lors de l'extraction de points d'intérêt. Le "*bucketing*" permet alors d'améliorer la diversité des points utilisés pour la résolution numérique du problème.

Toutefois, la validité du "*bucketing*" repose sur une hypothèse implicite : au sein d'un même "*bucket*", les points sont tous porteurs d'une information comparable. Ce n'est pas vrai, notamment, lorsque plusieurs objets à différentes profondeurs sont présents dans un même "*bucket*. Le mobilier urbain est riche de tels exemples 2.6(b).

De façon à pouvoir prendre ces cas particuliers en compte, nous proposons de définir un "*bucketing* en 3 dimensions, tel qu'illustré en figure 2.7. Un "*bucket*" ne prend plus en compte que les deux coordonnées traditionnelles du point mais également sa disparité, ce qui nous permet de lever certaines ambiguïtés et de ne pas éliminer des points dont l'information pourrait nous être utile. Le sur-coût calculatoire lié à cette prise en considération de la disparité est négligeable, seuls les besoins en mémoire sont sensiblement plus importants. D'un autre côté, cette méthode nous donne l'assurance de ne pas éliminer de points qui sont porteurs d'une information utile.

Extraction Multi-Passes - Alors que le "*bucketing*" cherche à éliminer l'information redondante, il est également souhaitable de chercher à l'enrichir dans les zones sous-représentées. Ces zones peuvent correspondre à des parties peu texturées de l'image (comme la route), où, à l'inverse à des motifs très répétitifs. Dans ce dernier cas toutefois, le problème n'est pas tant le manque de points d'intérêt que le manque de points d'intérêt *fiables*. Les faux-appariements sont en effet

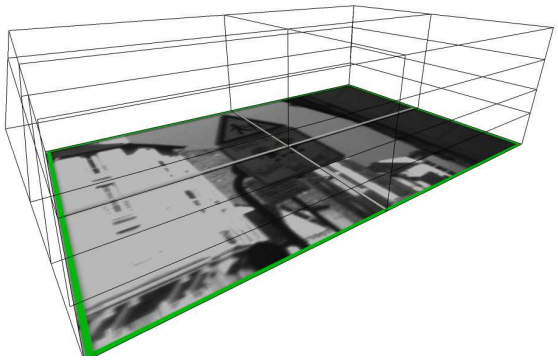


FIGURE 2.7 – "*Bucketing*" 3D - L'information supplémentaire, portée par le volume, est la disparité des points extraits.

beaucoup plus fréquents sur ces motifs.

Dans ces conditions, il peut être profitable d'identifier les zones faiblement représentées afin d'y opérer une seconde passe de détecteur de points d'intérêt. Toutefois, nous pensons qu'il n'est pas nécessairement judicieux de chercher à réutiliser le même détecteur, ou un détecteur moins discriminant, dans la mesure où de faux appariements apparaîtraient plus facilement.

Nous proposons une approche en deux étapes :

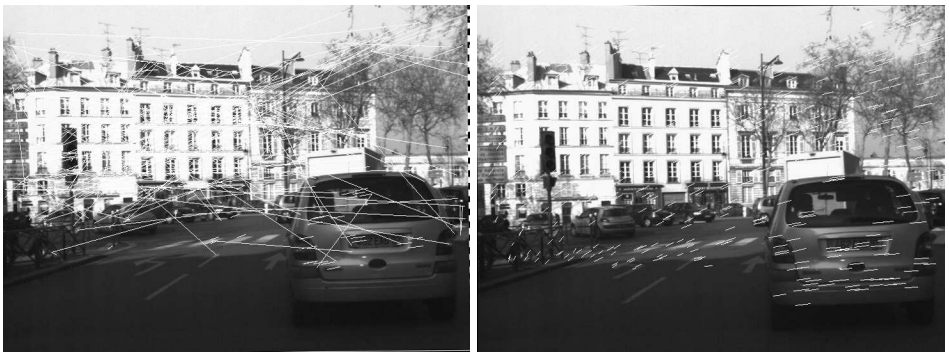
1. Tout d'abord, nous procédons à une extraction "normale", et à une première estimation de l'égo-mouvement.

2. Ensuite, nous utilisons cette première estimation comme indication pour faciliter la résolution du problème d'appariement en réduisant considérablement le domaine de recherche pour chaque point d'intérêt

Cette méthode nous permet ainsi de supprimer une immense majorité des faux appariements liés aux motifs trop répétitifs (objets manufacturés), avant de procéder à un affinement de l'estimation de l'égo-mouvement. La figure 2.8 permet de visualiser cette suppression des faux appariements.

2.2.6 Filtrage

Dans un contexte sensible, comme celui de la sécurité routière, il est important de pouvoir détecter d'éventuelles défaillances et comportements anormaux, tout



(a) (b)

FIGURE 2.8 – Comparaison des stratégies d'extraction de points d'intérêts : (a) : extraction simple ; (b) extraction après une première estimation frustre du mouvement, la majorité des faux appariements ont disparu.

comme il est important de pouvoir fournir une indication quant à ces défaillances. De plus, il est important de pouvoir, même en cas de défaillance, fournir une estimation du mouvement courant. Une telle estimation, pour être pertinente, doit être fondée sur l'évolution du signal mesuré.

Détection de Défaillance - Il existe plusieurs raisons qui peuvent entraîner une défaillance de notre système d'odométrie visuelle. Tout d'abord, des problèmes matériels peuvent survenir :

- Perte de synchronisme des caméras.
- Perte d'Image.
- Éblouissement des caméras.

Pour ces différentes raisons, il est intéressant de pouvoir disposer d'un indicateur de performances en sortie de l'évaluation odométrique. Nous proposons d'utiliser à cette fin l'erreur correspondant au meilleur consensus trouvé lors du processus de RANSAC. Les figures 2.15 et 2.16 illustrent ainsi cet indicateur de performances en regard des données odométriques calculées et extraites de la centrale inertielle. La corrélation entre les hautes valeurs de l'indicateur de performances[10] et les écart élevés entre l'odométrie visuelle et la centrale inertielle nous permet de confirmer le choix de cet indicateur de performances.

[10]Un point absent de la courbe indique une valeur infinie

FIGURE 2.9 – Le camion blanc au premier plan occupe la majorité du champ. En cas d'extraction du mouvement, le mouvement dominant apparent sera le mouvement relatif entre le camion et l'égo-véhicule.

Ensuite, un certain nombre de scénarios peuvent se produire et gêner la bonne extraction de l'égo-mouvement. Ces scénarios correspondent à des violations de nos hypothèses initiales. Par exemple, le processus de RANSAC utilisé repose implicitement sur le fait qu'une majorité relative de points utilisés sont les images de points objets statiques. Il est possible que cette hypothèse soit fausse. Par exemple, dans la figure 2.9, le mouvement extrait sera le mouvement relatif entre le camion et l'égo-véhicule.

Afin d'éviter ou, à tout le moins, de détecter, ce type de scénario, nous proposons d'utiliser la détection d'objets dynamiques qui sera présentée dans le chapitre 3. En détectant les points qui imagent des objets dynamiques, il est possible de les éliminer du processus d'extraction de points d'intérêt, et donc d'éviter certains cas "pathologiques", comme celui illustré en figure 2.9.

Intégration Temporelle - Filtrage de KALMAN - Pour rappel, le filtre de KALMAN [Kal60] repose d'une part sur la modélisation d'un processus d'évolution et d'autre part d'un processus de mesure. Le processus d'évolution peut s'exprimer :

$$X_k = f(X_{k-1}, u_k, w_k) \tag{2.12}$$

Où f est la fonction d'évolution du système, X_k est son état à l'instant k, u_k est une commande bruitée et w_k modélise le bruit inhérent au système.

Et le processus de mesure :

$$Z_k = h(X_k, v_k) \tag{2.13}$$

Où Z_k est la mesure et v_k est l'erreur de mesure. Dans le cas qui nous intéresse, le système est composé des 6 composantes du mouvement. On considère que la mesure est l'état et que donc la fonction h est l'identité.

Les équations du filtres de KALMAN nous permettent, d'une part, de prédire une mesure, en sachant l'état précédent, et, d'autre part, de mettre le modèle à jour en fonction de la dernière mesure disponible. Finalement, le filtrage de KALMAN repose sur une connaissance *a priori* des matrices de covariance, caractérisant le bruit du système. Dans sa forme originale, le filtrage de KALMAN ne permet d'estimer que des systèmes dont l'évolution est linéaire. Plusieurs contributions ont été apportées, de façon à étendre cette techniques à des systèmes non-linéaires, notamment le filtre de KALMAN Étendu qui repose sur une linéarisation locale de la fonction d'évolution autour de l'état courant, ou encore le filtre de KALMAN non-parfumé [JU97] qui repose sur l'utilisation d'échantillons pris autour de l'état courant et caractérisant l'ellipse d'incertitude.

Dans notre cas, nous ne cherchons pas à prédire systématiquement l'état à venir de notre système. Nous cherchons à obtenir une estimation du mouvement en cas de défaillance temporaire de l'odométrie visuelle. Dans ces conditions, un modèle simple est suffisant, en l'occurrence, nous modélisons notre système comme étant linéaire.

2.2.7 Résumé

La figure 2.10 propose un résumé graphique du fonctionnement global de l'algorithme d'odométrie visuelle mis en place dans le cadre de ce travail.

2.3 Résultats

Cet algorithme aura vocation à être utilisé comme étape initiale du système de détection d'objets dynamiques que nous décrirons dans le chapitre 3. Il est donc fondamental de pouvoir le caractériser finement. Les différents systèmes, réels et simulés, utilisés pour obtenir ces résultats sont décrits dans l'annexe D. Dans un premier temps, nous ne nous intéresserons qu'à ses performances intrinsèques. Dans un second temps, nous chercherons à évaluer les performances de l'odométrie visuelle, au sein d'un dispositif de fusion de données, en lieu et place d'une centrale inertielle classique. Nous pourrons ainsi comparer les performances de deux dispositifs ne différant que par l'origine de leur données odométriques. Cette dernière comparaison nous permettra de statuer quant à la viabilité d'une solution d'égo-localisation reposant sur la vision.

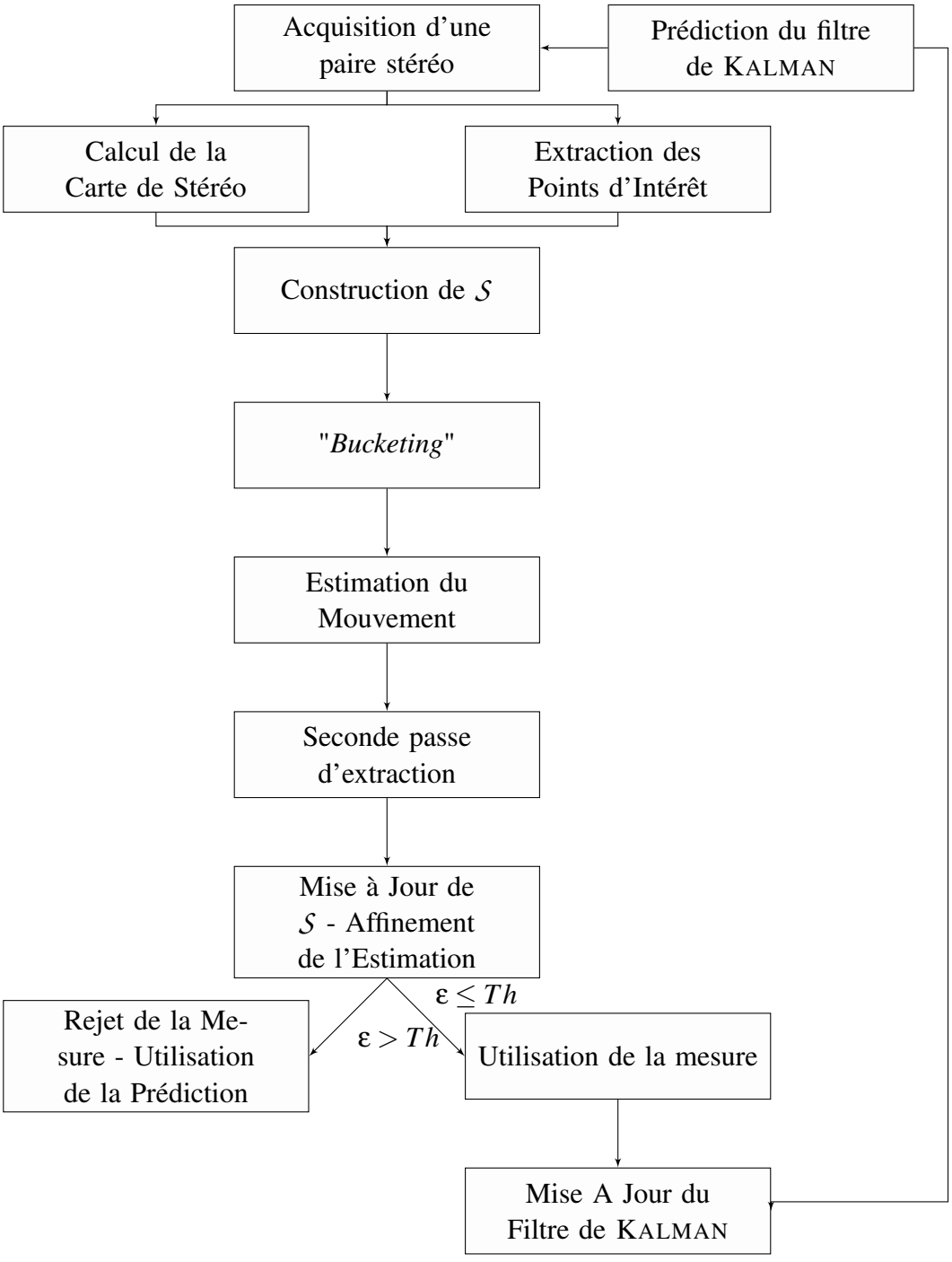


FIGURE 2.10 – Résumé du Fonctionnement de l'Odométrie Visuelle

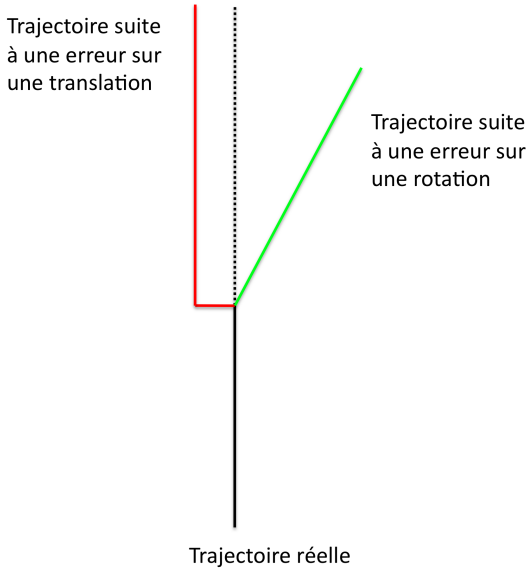


FIGURE 2.11 – Illustration de l'erreur d'Abbe

2.3.1 Performances Intrinsèques

2.3.1.1 Simulation

Une reproduction numérique de la piste de Satory a été utilisée afin de mener les premiers test de notre méthode d'odométrie. Il s'agit d'une séquence de 2700 paires stéréo représentant un parcours de $1090m$. L'erreur de positionnement finale sur ce trajet est de $14,8$ mètres soit $1,3\%$ de la distance parcourue totale. Toutefois, nous pensons que l'erreur de positionnement finale n'est pas un bon indicateur des performances de l'odométrie visuelle. En effet, tout comme une centrale inertielle, un algorithme d'odométrie visuelle est sensible à l'erreur d'Abbe. C'est à dire au fait que les erreurs de dérive vont s'accumuler aussi longtemps que le système va fonctionner :

$$\varepsilon = h \sin \theta \tag{2.14}$$

où ε est l'erreur finale en distance, h la distance totale parcourue et θ l'erreur angulaire initiale. En d'autres termes, si les erreurs commises sur les translations vont se compenser à l'infini, ce n'est pas le cas des erreurs en rotations. Ceci est illustré en figure 2.11.

Il nous paraît donc plus judicieux d'examiner l'erreur instantanée commise sur l'égo-mouvement.

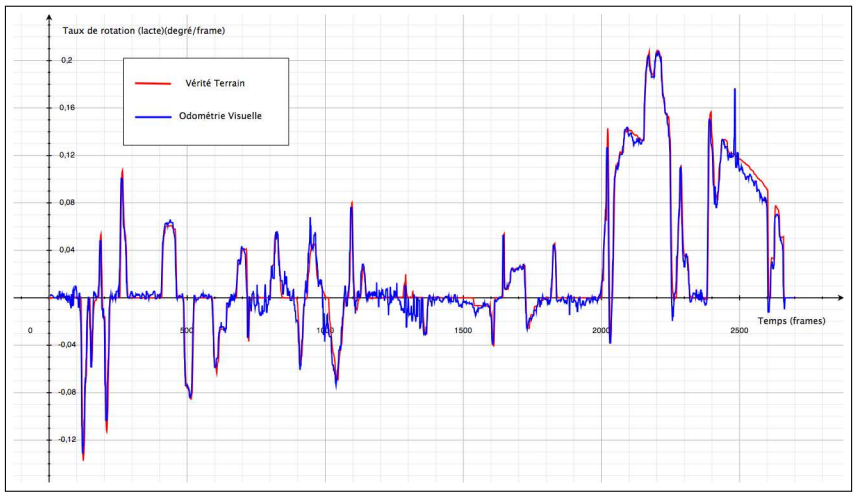


FIGURE 2.12 – Evolution du taux de rotation suivant l'axe de lacet

A cette fin, les figures 2.12 et 2.13 représentent l'évolution du taux de rotation suivant l'angle de lacet et de la translation axiale au cours de cette séquence de 2700 paires stéréo.

Sur cette seconde figure, la dernière partie est en particulier révélatrice d'un problème qui peut survenir sur des environnements simulés : le manque de textures sur les surfaces composant l'environnement virtuel. En effet, SiVIC étant un logiciel en cours de développement, tous ses composants ne sont pas finalisés. En particulier, la seconde partie de cette séquence consiste en un long passage très peu texturé, où tous les objets présentent des motifs extrêmement répétitifs, peu représentatifs de la réalité et susceptibles d'introduire de nombreuses erreurs d'estimation. C'est ce que l'on peut voir sur la figure 2.14. Ce manque de textures est très marqué entre les images 1200 et 1400.

En raison de ces erreurs, il est nécessaire, dans un premier temps, de multiplier les scénarios simulés et dans un second temps, d'utiliser des données issues de sources réelles. Ainsi, le scénario utilisé en section 2.2.4 et illustré en figure 2.5 est également utilisé afin d'obtenir des résultats globaux sur données synthétiques. Ces résultats sont disponibles en table 2.4.

2.3.1.2 Images Réelles

Comme nous l'avons vu, si les données issues d'un simulateur peuvent s'avérer pratiques afin d'attester du bon fonctionnement de notre méthode, des résultats issus d'expérimentations sur images réelles sont irremplaçables. A ces fins,

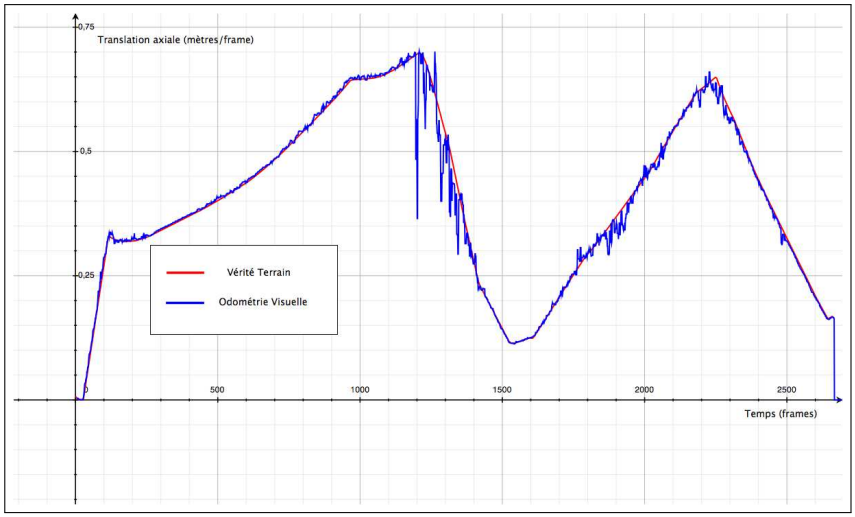


FIGURE 2.13 – Evolution de la translation axiale

FIGURE 2.14 – Extrait d'une séquence générée par SiVIC, les seules zones suffisamment texturées de l'image pour permettre une extraction efficace de points d'intérêts sont également les zones présentant des motifs très répétitifs (route, grillage, ...), ce qui empêche une estimation efficace du mouvement.

Composante	Erreur Moyenne
Rotationelle	$0,02\text{deg}.s^{-1}$
Translationelle	$0,04\text{m}.s^{-1}$

TABLE 2.4 – Résumé des résultats d'Odométrie Visuelle - SiVIC

Composante	Erreur Moyenne
Rotationelle	$0,04\text{deg}.s^{-1}$
Translationnelle	$0,03\text{m}.s^{-1}$

TABLE 2.5 – Résumé des résultats d'Odométrie Visuelle - CARLLA

nous présentons ci-après deux jeux de résultats obtenus respectivement à partir des moyens expérimentaux du LIVIC (véhicule CARLLA) et de l'Université de Karlsruhe.

CARLLA - Le véhicule de test CARLLA du Livic, présenté en D.2.2 a été utilisé afin de réaliser plusieurs séquences de tests, prises dans la ville de Versailles et ses environs. Ces séquences présentent donc des environnements urbains, péri-urbains et autoroutiers.

Ces séquences représentent un circuit d'une heure en environnement ouvert, soit environ 30 000 paires stéréo, dont les résultats sont résumés en table 2.5.

A titre indicatif, les figures 2.15 et 2.16 proposent une visualisation du taux de rotation selon l'axe vertical, ainsi que de la translation longitudinale au cours d'une de ces séquences. Ces courbes font également figurer l'évolution de l'indicateur de défaillance, décrit en 2.2.6. En particulier, deux passages proches de la fin de la séquence mettent en évidence une forte corrélation entre les hautes valeurs de l'indicateur de défaillance et des valeurs mesurées éloignées de la vérité terrain. Il apparaît que ces deux passages font apparaître deux cas de figures dans lesquels notre méthode est impuissante. Ces cas "pathologiques" sont illustrés en figure 2.17. Dans le premier cas, les points extraits appartiennent tous à un plan fronto-parallèle, le problème à résoudre devient alors insuffisamment conditionné (de plus, le véhicule au premier plan étant mobile, l'hypothèse de mouvement dominant est invalide), alors que dans le second cas, les conditions d'imagerie ne permettent pas d'avoir des résultats convenables.

Karlsruhe - Les bases de données de l'Université de Karlsruhe représentent environ 9000 paires stéréo, réparties en 9 séquences, prenant places en environnement urbain. Sur l'ensemble de ces séquences, les résultats intrinsèques de notre méthode s'avèrent encore du niveau d'une bonne centrale inertielle.

A titre indicatif, la figure 2.19 présente la trajectoire suivie lors d'une séquence, re calculée et projetée sur une image aérienne de la zone parcourue. La figure 2.20 présente l'extraction du taux de rotation selon l'axe vertical par la

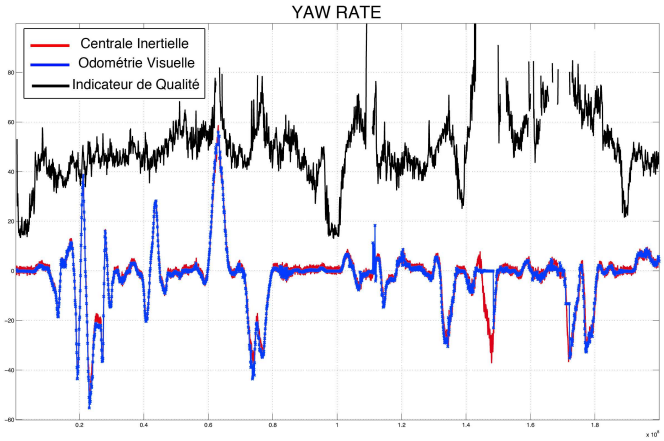


FIGURE 2.15 – Taux de rotation (lacet) extrait par la centrale inertielle (rouge) et l'odométrie visuelle (bleu)

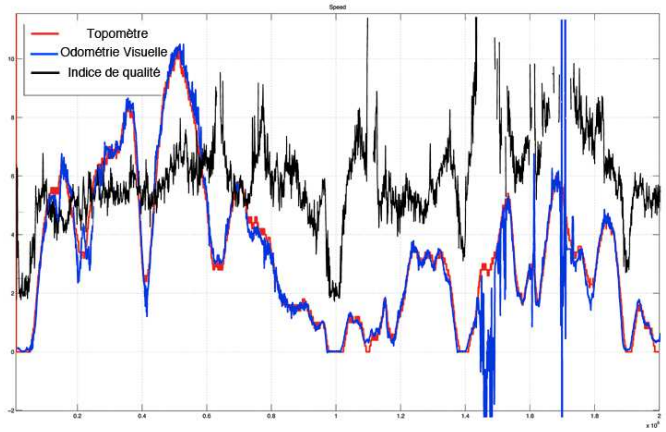


FIGURE 2.16 – Translation Longitudinale extraite par les topomètres embarqués (rouge) et l'odométrie visuelle (bleu)

Composante	Erreur Moyenne
Rotationelle	$0,06 \text{deg}.s^{-1}$
Translationelle	$0,04 \text{m}.s^{-1}$

TABLE 2.6 – Résumé des résultats d'Odométrie Visuelle - Karlsruhe

(a) (b)

FIGURE 2.17 – Illustration des défaillances de l'odométrie visuelle : (a) : le champ est intégralement occupé par un plan fronto-parallèle ; (b) la qualité de l'imagerie est dégradée par des conditions météorologiques dégradées

FIGURE 2.18 – Trajectoire recalculée à partir du mouvement extrait - recalée sur une vue aérienne - CARLLA

FIGURE 2.19 – Trajectoire Recalculée à partir du mouvement extrait - recalée sur une vue aérienne - Karlsruhe

centrale inertielle embarquée et par notre méthode d'odométrie visuelle. Il est intéressant de noter que les données issues de la centrale inertielle (en rouge sur la figure) présentent de nombreuses aberrations et instabilités. Les résultats numériques obtenus sur ces séquences sont disponibles en table 2.6.

2.3.2 Localisation par Fusion Multi-Capteurs

L'algorithme d'Odométrie Visuelle décrit dans le présent document a été mis à contribution dans plusieurs projets de localisation par fusion multi-capteurs. En effet, comme nous l'avons vu, la méthode présentée est susceptible de défaillir dans un certain nombre de cas, tout comme d'autres types de capteurs sont susceptibles de défaillance. Par exemple, un système GPS est mis à mal par une constellation de satellite insuffisante, de la même façon, des odomètres embarqués sont inutiles en cas de glissement des roues, où lorsque la géométrie du terrain n'est plus plane. Bien que cela ne représente pas le cœur du travail présenté, il nous semble important de présenter les différentes conclusion auxquelles ces projets ont pu aboutir, de façon à pouvoir mettre en évidence la contribution qu'à pu avoir le résultat de notre travail sur les avancées d'autres chercheurs.

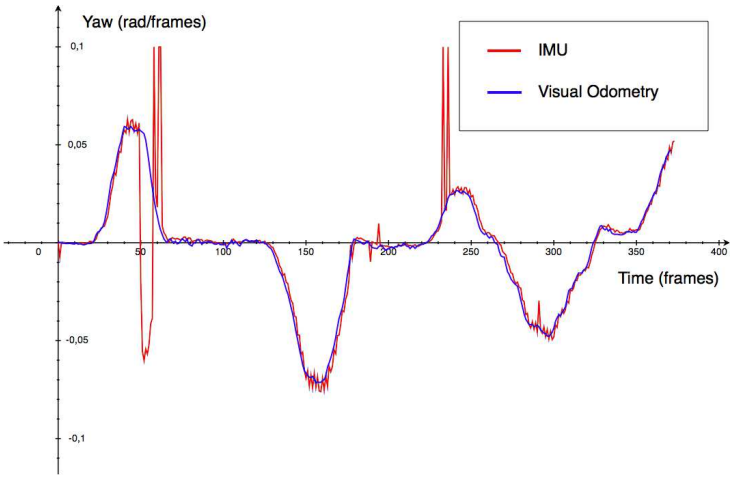


FIGURE 2.20 – Evolution du taux de lacet, mesuré par Odométrie Visuelle (bleu) et par Centrale Inertielle (rouge)

2.3.2.1 Fusion par Filtre de KALMAN Etendu

La méthode d'Odométrie Visuelle, décrite ci-dessus a tout d'abord fait l'objet d'une intégration dans le système de fusion multi-capteurs par filtrage de KALMAN étendu, développé par Dominique GRUYER[11], dans le cadre du projet européen CVIS [GPGD10] et illustré en figure 2.21. L'objectif de ce travail était de démontrer que l'Odométrie Visuelle, considérée comme un capteur proprioceptif, pouvait remplacer une centrale inertielle. Ce travail a été réalisé en utilisant les données acquises à l'aide du véhicule CARLLA, équipé d'une paire stéréo, mais également d'un GPS, ainsi que d'odomètres classiques et d'une centrale inertielle, qui a servi de base de comparaison.

La figure 2.22 présente des résultats partiels de ce travail. Sur ces résultats, le comportement du GPS a été idéal, ce qui nous permet de l'utiliser comme référence terrain. Toutefois, il est important de noter que les performances des systèmes GPS sont sujettes à de nombreuses perturbations, ainsi, si la constellation de satellites visibles est insuffisante (bâtiments hauts, tunnels, obstructions, ...), les performances sont nettement dégradées.

Entre autres conclusions, il est apparu que l'utilisation de l'odométrie visuelle au lieu de la centrale inertielle menait à des résultats aussi bons, voire occasionnellement meilleurs. Ce travail nous a permis de positionner l'odométrie visuelle comme alternative viable aux technologies traditionnelles. En effet, nous avons pu

[11] Chargé de Recherche INRETS au LIVIC

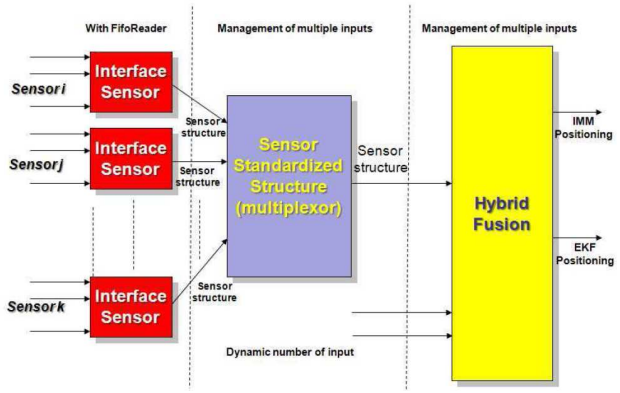
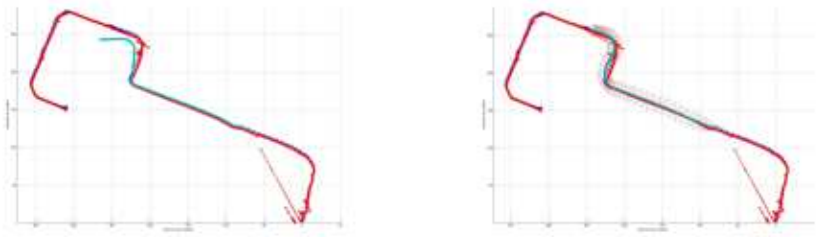


FIGURE 2.21 – Vue globale du système de fusion d'information par filtre de KALMAN étendu



(a) (b)

FIGURE 2.22 – Résultats de Fusion Multi-Capteurs utilisant l'Odométrie Visuelle en lieu et place d'une centrale inertielle ; La coube rouge correspond aux donnée GPS, la courbe bleue correspond aux données issues de la fusion : (a) Fusion Odométrie Visuelle + topomètres ; (b) Fusion Odométrie Visuelle + Topomètres + GPS.

prouver que les systèmes de vision sont non seulement beaucoup moins coûteux mais également au moins aussi performants que les centrales inertielles.

Ce travail fait toujours l'objet de développements, au LIVIC, notamment pour complexifier et diversifier le modèle cinématique considéré, mais également pour utiliser l'interaction entre différents modèles (approche Multi-Modèle).

2.3.2.2 Fusion par Théorie des Croyances

Ce second projet repose sur le formalisme de la théorie des croyances [Sha76, Sme90]. Le cadre de cette fusion a été développé par Arnaud ROQUEL dans le cadre de son travail de thèse[12]. Pour ce travail, nous avons utilisé la plateforme Mini-Truck, décrite en D.2.3.

L'objectif premier de ce travail a été de réaliser une localisation du mobile, en fusionnant les informations provenant de plusieurs sources, en l'occurrence, une méthode de FAST-Slam mise au point par Bastien VINCKE dans le cadre de son travail de thèse[13].

Par la suite, ce cadre a été utilisé afin de réaliser de la détection et de la résolution de conflit entre capteurs, afin de pouvoir mettre en évidence l'apport général de la fusion de données. Dans le cas de notre méthode d'odométrie visuelle, une défaillance a été induite en manœuvrant le mobile très près (environ 20 cm) d'une surface plane, de façon à être sous la distance minimale efficace du capteur Kinect®.

Le cadre de cette fusion par fonctions de croyances est décrit par la figure 2.23. A l'heure actuelle, ces travaux ne sont pas terminés, et il nous est donc impossible de présenter des résultats ou conclusions.

2.4 Conclusion

Au cours de ce chapitre, nous sommes revenus sur deux décennies d'évolution des techniques d'odométrie visuelle avant de présenter notre approche. Cette approche présente l'avantage de ne reposer que sur un formalisme linéaire et évite donc les inconvénients classiques des méthodes de résolution des problèmes de moindres carrés non linéaires : divergence et lenteur. Différents choix techniques ont du être pris lors de la conception de ce système. Ces différents choix ont été

[12] Thèse menée à l'Institut d'Electronique Fondamentale, sous la direction de Sylvie LE HÉGARAT et Isabelle BLOCH

[13] Thèse menée à l'Institut d'Electronique Fondamentale, sous la direction d'Alain MÉRIGOT, co-encadrée par Alain LAMBERT et Abdelhafid EL OUARDI.

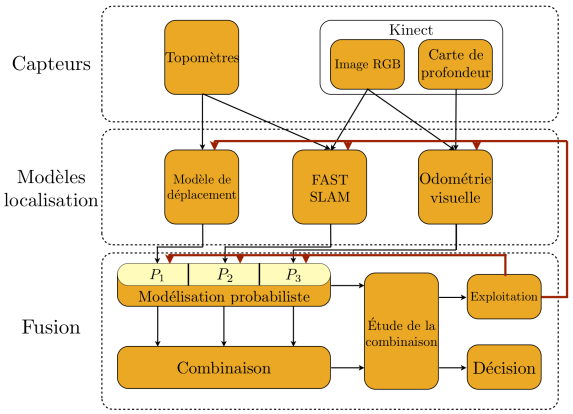


FIGURE 2.23 – Vue globale du système de fusion par masses de croyances

défendus et motivés. Nous avons vu que si la plupart de ces choix peuvent se résumer à un compromis entre temps de calcul et performances, certains, notamment le choix de la technique utilisée pour réaliser des appariements spatiaux, ont un impact notable sur la polyvalence du système complet.

Le système final que nous proposons présente de bons résultats, avec une erreur moyenne comparable à celle obtenue par une centrale inertielle traditionnelle de très bonne qualité. Notre système n'a pas été évalué qu'à l'aune de ses performances intrinsèques, nous avons également cherché à évaluer son efficacité en tant que maillon d'un réseau de capteurs plus complexes. Là encore, notre approche a pu prouver sa compétitivité vis-à-vis de solutions plus onéreuses.

A notre sens, l'odométrie visuelle mérite donc d'être considérée comme un choix viable d'extraction de l'égo-mouvement lors de la conception de véhicules intelligents. Outre un avantage de coût non-négligeable, la vision fournit une information beaucoup plus riche qu'un appareil dédié à la mesure des accélérations linéaires et rotationelles, et cette information peut être utilisée pour de nombreuses tâches, autres que l'odométrie.

Chapitre 3

Détection du Mouvement Indépendant

«We have no idea if there really are aliens. The next question would be, If there are aliens, can we detect it? The answer is yes. With this telescope we now have a much better chance.»
Bruce Betts

Sommaire

Motivations

La prévention ou la mitigation de collision est une thématique importante de la robotique générale et en particulier des véhicules intelligents. En effet, l'immense majorité des accidents de la route résulte d'une collision entre l'égo-véhicule et un tiers. Aussi bien pour l'intégrité de l'égo-véhicule que pour la sécurité des autres usagers de la route, cette tâche de détection et de prévention est donc fondamentale. Dans ce contexte, de nombreuses approches ont été envisagées afin de permettre à un conducteur artificiel de pouvoir évoluer sans risque dans un environnement aussi complexe que les villes du XXI$^{\text{è}}$ siècle.

3.1 Art Antérieur

3.1.1 Classification - Reconnaissance

Une première solution à ce problème est de chercher à identifier les objets dont on sait *a priori* qu'ils peuvent présenter un danger. Ainsi, une reconnaissance des véhicules peut être menée, en se fondant sur des attributs particuliers. La symétrie par exemple [BBB$^+$01, BBFN00], la texture [KTS98], ou encore la couleur [BD98] sont autant d'indices qui permettent d'identifier des véhicules.

Toutefois, si ces approches constituent une première base de travail, elles présentent de nombreuses limites. Tout d'abord, les véhicules peuvent présenter une grande variabilité, ne serait-ce que suivant l'angle de visualisation, mais surtout, les véhicules ne représentent qu'une partie des acteurs de la route, qui plus est, la moins vulnérable.

Les piétons présentent, en effet, une très grande vulnérabilité, tout en exhibant une variabilité (déformation, différences d'aspects, ...) rendant toute approche basée sur une reconnaissance *simple* plus ardue.

L'émergence de solutions de classification et de reconnaissance robustes et performantes permet, dans ce contexte, d'envisager d'autres solutions. Les techniques de classifications binaires en cascade [GB00, Alf02], qui permettent de construire un arbre de décision complexe à partir de décision élémentaires binaires ainsi que les techniques de *boosting* telles qu'AdaBoost [FS95, FHT98, HTFF05] qui permettent de construire un classifieur *fort* à partir de classifieurs élémentaires *faibles*

Parmi les classifieurs utilisés dans la littérature, les solutions les plus fréquentes sont les Machines à Vecteurs de Support[1] [Bur98, Vap99] ou les réseaux

[1] *Support Vector Machine, SVM*, on trouve également Séparateur à Vaste Marge.

FIGURE 3.1 – Exemples d'images extraites de la banque INRIA

de neurones [Hay99, SP98, MG06]. Ces méthodes jouissent en effet d'une bonne capacité à travailler sur des espaces de haute dimensionnalité.

Cet espace sur lequel le classifieur va être entraîné, puis utilisé, repose sur la définition d'une base de représentation dans laquelle les régions d'intérêts de l'image vont être projetées. La solution la plus populaire est actuellement l'utilisation d'histogrammes de gradients orientés[2] [DT05, ZYCA06, SRBB06]. D'autres méthodes ont été proposées comme les *Joint Ranking of Granules* [HN10], la décomposition en ondelettes de HAAR [VJS05, LGLL10], ou encore une analyse en composantes principales (ACP, *Principal Components Analysis, PCA* dans la littérature anglophone) [TP91].

Ces méthodes de classification reposent également sur une phase d'apprentissage *hors-ligne*. Cet apprentissage se déroule en présentant au classifieur des populations de négatifs et de positifs, tels que ceux illustrés en figure 3.1. Cette phase va jouer un rôle critique, en effet, la représentativité des bases d'apprentissage va fortement conditionner l'aptitude du classifieur à discerner différents objets.

Il nous semble que les méthodes de reconnaissance d'une façon générale souffrent d'un manque de généralité. Même si certaines idées permettant de rendre le processus de reconnaissance plus robuste aux variations intra-classes ont été avancées, comme par exemple, la segmentation d'un piéton en sous objets (jambes, torse, etc.) [SGH04], il n'en reste pas moins que les approches par apprentissage vont être pertinentes dans la limite de la représentativité de leur base d'apprentissage. De la même façon, les méthodes qui partent d'un constat concernant l'aspect visuel des véhicules vont être mises à mal dès que certaines hypothèses ne sont plus valides.

Dès lors, il est nécessaire d'envisager des méthodes qui ne reposent pas sur l'aspect des objets potentiellement dangereux, mais sur des caractéristiques plus

[2] *Histograms of Oriented Gradients, HOG*

globales, comme le mouvement ou la structure de la scène.

3.1.2 Détection Générique

Une telle détection plus générale des obstacles ou des autres usagers de la route peut reposer sur deux modalités principales :

- l'image du mouvement ;
- la structure de la scène.

Méthodes Temporelles - Le mouvement seul est rarement utilisé pour lui même. Il est en effet fréquent de le voir utilisé après une phase de détection par reconnaissance, afin de procéder à un suivi [BBFT04, GM07], ou avant cette phase de détection, afin de réduire la zone de recherche [EKG08].

Toutefois l'exploitation du flot optique seul reste possible. Ici encore des approches par reconnaissance sont possibles et trouvent leur fondements, par exemple, dans une modélisation de l'image du mouvement humain [FSWG06].

L'exploitation intrinsèque d'une information purement monoculaire, en l'occurrence le flot optique a été mise en avant afin de réaliser une détection d'obstacle. Ces approches vont généralement reposer sur une segmentation du flot optique [MMP08, Dum09], segmentation qui peut faire intervenir une connaissance *a priori* des objets [BPCL06], des considérations géométriques [IA96] ou encore une évaluation du temps à collision [Dum09].

Les méthodes monoculaires fondées sur le mouvement trouvent cependant des limites théoriques, en particulier l'impossibilité de pouvoir estimer les translations de manière absolue et non pas à un facteur d'échelle près. Dès lors, il n'est pas surprenant de constater que la plupart des approches proposées reposent sur une estimation structurelle ou sur une collaboration entre les deux modalités.

Méthodes Structurelles - Compte tenu du fait que la majorité des objets attendus ne sont pas transparents[3], un obstacle potentiel va être imagé comme son plan fronto-parallèle équivalent. Ce constat est à l'origine, notamment, de la V-Disparité [LAT02], mais également de la C-Vélocité [BPZ09], qui sera décrite au chapitre suivant. Ces méthodes ont pour objectif de construire un espace dans lequel les plans fronto-parallèles vont être transformés en droites, faciles à extraire, par exemple par transformée de HOUGH.

[3]Nous parlons ici de la transparence au sens large, que ce soit dans le visible, l'IR ou encore les radio-fréquences

Au delà de la simple détection de plans fronto-parallèles, la construction de cartes, ou de grilles d'occupation, connaît un certain succès [VBA08, KZP+08, NBT09, YPP+10]. L'un des principaux intérêts de cette approche est que la collaboration entre plusieurs capteurs est alors immédiate. En effet, une même carte d'occupation peut être peuplée en utilisant indifféremment des points issus d'un LIDAR, d'un RADAR ou de la stéréovision.

D'une manière plus générale, la collaboration entre LIDAR et stéréovision est une piste fréquemment envisagée. Ainsi, le LIDAR peut être utilisé afin de fournir des hypothèses de détection que la vision viendra ensuite confirmer [RFBC10, Per08].

Ce problème de localisation des obstacles peut également être abordé par son dual : l'identification de l'espace libre devant le véhicule. La problématique n'est plus alors de chercher à éviter les menaces potentielles, mais de chercher à définir l'espace dans lequel il est possible pour l'égo-véhicule de manœuvrer [BMVF08, SPLA07, LS09].

Toutefois, si ce type d'approche peut se montrer relativement performant, certaines configurations peuvent se révéler difficiles, voire impossibles à résoudre. Ce type de cas est illustré par la figure 3.2, ainsi une des deux cibles mise en évidence est partiellement occultée par un véhicule, ce qui rend son extraction difficile, alors que l'autre est assez éloignée du capteur, ce qui peut, là encore, rendre sa détection difficile. Ce sont précisément ces challenges que proposent de relever les méthodes collaboratives.

Méthodes Collaboratives - L'idée de faire collaborer estimation du mouvement et estimation structurelle n'est pas neuve. Les travaux faisant intervenir activement ces deux approches se succèdent depuis le début des années 2000, c'est à dire depuis que la puissance de calcul disponible permet de mener ces deux processus de front.

Ainsi, par ses travaux, centrés sur l'exhibition d'un invariant de l'image, en l'occurence le rapport de la norme du flot optique sur la distance au capteur, HEINRICH fait figure de précurseur [Hei02]. Si l'approche développée est séduisante par bien des aspects, elle demeure trop contraignante et il faut encore attendre quelques années avant de voir émerger des méthode réellement efficaces.

Le principe de 6D-Vision, exhibé par FRANKE dans [FRBG05], puis exploité dans les différents travaux de son équipe [WRV+08, PF10, RFG07] constitue une approche intéressante. Elle repose sur le suivi de points d'intérêts en utilisant des filtres de KALMAN, accordés sur les mouvements susceptibles d'animer les objets de la scène. Les travaux présentés dans ce cadre ont constitué des avancées importantes pour la conception de systèmes intelligents, exploitant plusieurs capteurs.

Toutefois, il nous semble important de continuer le développement de système fondés sur la vision seule.

Comme nous l'avons vu plus haut, le formalisme des grilles d'occupations permet une intégration aisée de différents capteurs. Il est donc naturel de le retrouver exploité ici afin de faire coopérer les différentes modalités de la vision artificielle. Ce formalisme peut être exploité afin de construire une représentation de la scène observée [DC00], ou simplement enrichi de l'information temporelle [BPU+08, LCCG07]

Les approches recevant le plus grand intérêt de la communauté sont cependant les approches centrées sur l'évaluation du *scene-flow*, soit l'extension du flot optique à un espace tridimensionnel. Pour cela, il est possible de suivre des points d'intérêt [LZGR11] ou d'intégrer la prise en compte de la stéréo dans une méthode de calcul du flot optique type HORN & SCHUNK [PKF07, WRV+08]. A partir de ce champ de correspondances, des techniques de segmentations classiques peuvent être utilisées afin d'obtenir une représentation de la scène en fonction du mouvement apparent des objets. Nous pensons que cette approche, de part sa simplicité théorique, est amenée à se démocratiser. Toutefois, le fait de raisonner dans l'espace objet n'est pas très judicieux, de part l'anisotropie du bruit que nous avons déjà présenté. Il faut cependant considérer que ce choix peut être dicté par l'utilisation de capteurs additionnels ce qui, là encore, ne nous semble pas être une voie à privilégier pour l'instant.

Dans ce contexte, l'étude que nous proposons se singularise par plusieurs aspects. Tout d'abord, il ne repose que sur la vision, là où de nombreux auteurs utilisent un ou plusieurs capteurs additionnels (LIDAR, centrale inertielle, etc.), nous pensons en effet qu'il est nécessaire d'explorer les limites de la vision seule avant de proposer des systèmes multi-capteurs.

Ensuite et surtout, la détection proposée est dense, là où la majorité des méthodes existantes proposent une détection éparse. Quelle que soit la méthode choisie pour extraire les points d'intérêts, une méthode éparse de détection ne nous paraît pas acceptable. En effet, il nous semble que, surtout dans un cadre sensible comme celui de la sécurité routière, la détection d'un objet dangereux ou vulnérable ne doit pas reposer sur son aspect visuel plus ou moins texturé.

Finalement, les méthodes qui reposent sur une segmentation du flot optique, ou du *scene flow*, sans compenser préalablement l'égo-mouvement, sont sujettes à des faux négatifs correspondant aux objets lentement mobiles [LZGR11], mais également à des faux positifs que pourraient représenter les objets saillants, en raison de la parallaxe du mouvement.

Dans ces conditions, nous pensons qu'un système efficace de détection des

FIGURE 3.2 – Exemples de cibles présentant potentiellement un danger, difficilement identifiable par des méthodes structurelles : 1) Le piéton, à une grande distance est partiellement dissimulé par les véhicules garés ; 2) le piéton, à une très grande distance est trop loin pour être distingué des bâtiments à l'arrière plan.

objets dynamiques doit proposer une détection dense, *via* une compensation de l'égo-mouvement. De plus, et c'est probablement l'un des aspects les plus importants de notre travail, nous pensons qu'une réflexion de fond doit être menée sur cette compensation de l'égo-mouvement et surtout sur son impact sur les résultats en termes de détection du système. Comme nous le verrons, différents paramètres du système de vision peuvent être ajustés, de façon à optimiser les performances du système.

Dans la section suivante, nous allons nous attacher à présenter le système développé, en justifiant les différents choix techniques et technologiques qui ont été pris. Avant de présenter des résultats quantitatifs et qualitatifs, nous allons nous attarder sur une étude des limitations de notre système, ainsi que sur différentes pistes d'amélioration possibles. Cette étude devrait pouvoir être, *in fine*, considérée comme un ensemble de recommandations à l'attention de futurs développeurs de système de vision intelligents.

3.2 Système Proposé

Le système de détection que nous proposons peut être vu comme la succession de trois étapes distinctes.

- Tout d'abord, le mouvement de l'égo-véhicule, connu, doit être compensé.
- Ensuite, le mouvement résiduel doit être estimé et, éventuellement, filtré.
- Finalement, les résultats de cette détection peuvent être intégrés dans le temps, de façon à améliorer la détection d'objets faiblement mobiles, et à éliminer les faux-positifs.

Facultativement, une étape de segmentation peut être utilisée de façon à obtenir une information de plus haut niveau.

Un résumé de notre approche est disponible en fin de section, en figure 3.11.

3.2.1 Compensation de l'Égo-Mouvement

Principe - Implémentation - Nous supposons dans cette partie qu'une estimation de l'égo-mouvement du capteur entre les instants t_0 et t_1 est disponible. Cette estimation est notée $\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}}\right)$. En pratique, cette estimation nous est fournie par le système décrit au chapitre 2, bien qu'il soit possible d'utiliser n'importe quel système autorisant une mesure du mouvement 6D, dès lors que ce système soit suffisamment bien caractérisé en termes de bruit et de précision.

Nous supposons également qu'une carte de disparité est disponible, nous permettant d'attribuer aux points de l'image une valeur de disparité δ. Idéalement cette carte est dense, mais n'importe quelle méthode de calcul est utilisable. Outre les zones d'occultation, il est possible que certaines zones faiblement texturées de l'image (une route lisse par exemple) soient filtrées et éliminées de la carte de disparité. Nous utiliserons également les deux images successives délivrées par le capteur choisi comme référence de la paire stéréo, ces images sont notées I_0 et I_1.

L'ensemble des points de $I_{k\in\{0,1\}}$ pour lesquels une information de disparité est disponible est noté $\mathcal{E}_{k\in\{0,1\}}$

Dès lors, pour chaque point m de $\mathcal{E}_0$, nous pouvons évaluer le déplacement image dû à $\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}}\right)$, en utilisant l'équation A.17. Ce déplacement est noté :

$$\boldsymbol{\Pi}_{\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}}\right)}\left(\begin{vmatrix} x_m \\ y_m \\ \delta_m \end{vmatrix}\right) = \begin{vmatrix} \mu \\ \nu \\ \xi \end{vmatrix} \tag{3.1}$$

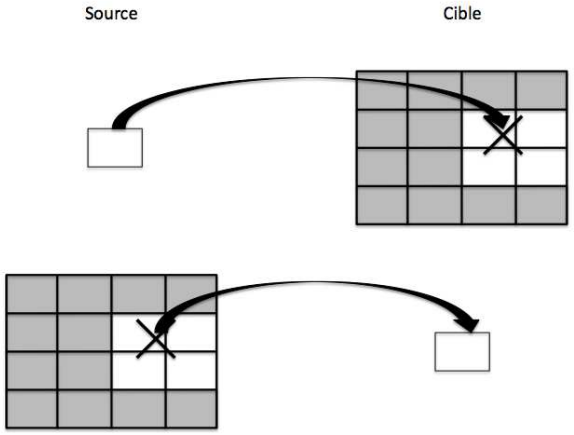


FIGURE 3.3 – Comparaison de deux stratégies d'interpolation : ligne supérieure : les coordonnées non-entières sont celles de l'image d'arrivée ; ligne inférieure : les coordonnées non-entières sont celles de l'image de départ

Ce déplacement nous permet de définir, sur $\mathcal{E}_0$, la déformation subie par l'image I_0, en d'autres termes, de connaître, pour un point $m = \begin{vmatrix} x_m \\ y_m \\ \delta_m \end{vmatrix}$ les coordonnées où il se trouvera dans l'image I_1. Toutefois, ces coordonnées sont très fréquemment non-entières, et il est donc indispensable de recourir à des techniques d'interpolation.

Considérations Numériques - Cette formulation du problème n'est pas très adaptée. Afin de réaliser une interpolation, il est souhaitable que les coordonnées non-entières soient les coordonnées de départ et non d'arrivée (du point de vue de la déformation), cela est illustré en figure 3.3. Le problème indirect (ligne inférieure) est beaucoup plus immédiat que le problème direct (ligne supérieure). En effet, dans le cas indirect, il suffit d'attribuer au pixel cible une valeur résultat de l'interpolation des valeurs des pixels encadrant les coordonnées non-entières.

Dans ces conditions, nous choisissons de formuler le problème inverse. Au lieu de calculer le déplacement de I_0 vers I_1, nous calculons, à partir des équations A.17 le déplacement de I_1 vers I_0, défini pour chaque point m de $\mathcal{E}_1$ comme :

$$\mathbf{\Pi}^{-1}_{(-\tilde{\boldsymbol{T}},-\tilde{\boldsymbol{\Omega}})} \begin{pmatrix} x_m \\ y_m \\ \delta_m \end{pmatrix} = \begin{vmatrix} -\mu \\ -\nu \\ -\xi \end{vmatrix} \tag{3.2}$$

Cette déformation permet de définir, pour chaque point de $\mathcal{E}_1$ les coordonnées *d'où il vient* dans l'image I_0, ce qui est plus adapté à l'utilisation des méthodes d'interpolation.

Cette transformation peut alors être appliquée à tous les points de $\mathcal{E}_1$, de façon à déformer, par parties[4], l'image I_0 en fonction du mouvement estimé $\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}}\right)$. Ce processus est illustré en figure 3.4.

Pour des raisons numériques (la détection d'objets dynamiques étant réalisée à partir de calculs sur des voisinages) et de performances (les calculs menés sur des zones de mémoire contigües étant plus efficaces) il peut être intéressant de chercher à "combler" les zones vides de l'image, qui correspondent à des zones creuses de la carte de disparité. Nous choisissons de combler ces zones en utilisant l'information de l'image I_1. L'image résultante complète est notée I'_1 et est illustré en figure 3.5.

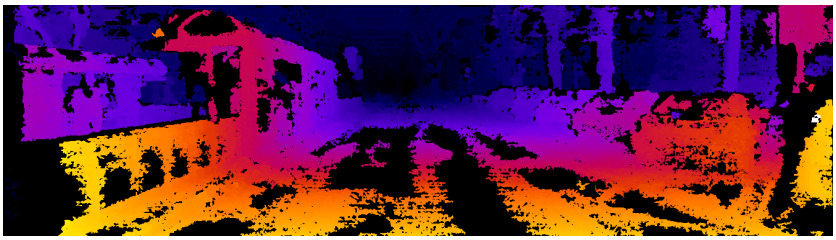
Cette procédure de compensation de l'égo-mouvement est résumée dans l'algorithme 2.

Précision - Ce recalage d'une image sur l'autre n'est pas parfait. Il est notamment entaché des erreurs liées à l'extraction du mouvement. Par propagation des incertitudes, ces erreurs peuvent être exprimées :

$$\left\{\begin{array}{l} \partial\mu = \frac{\frac{y_m}{f}(x+y\omega_z-f\omega_y-\frac{T_X\delta_m}{b_s})}{a^2}\partial\omega_x + \frac{y_m\omega_x+\frac{T_Z\delta_m}{b_s}-f-\frac{x_m^2}{f}-\frac{x_m y_m\omega_z}{f}+\frac{x_m T_X\delta_m}{fb_s}}{a^2}\partial\omega_y \\ \qquad\qquad +\frac{y_m}{a}\partial\omega_z+\frac{\delta_m/b_s}{a}\partial T_X+\frac{\frac{\delta_m}{fb_s}(x+y\omega_z-f\omega_y-\frac{T_X\delta_m}{b_s})}{a^2}\partial T_Z \\ \partial\nu = \frac{x\omega_y-\frac{T_Z\delta_m}{b_s}+f+\frac{y_m^2}{f}-\frac{x_m y_m}{f}\omega_z-\frac{y_m T_Y\delta_m}{fb_s}}{a^2}\partial\omega_x+\frac{\frac{x_m}{f}(y_m-x_m\omega_z+f\omega_x-\frac{T_Y\delta_m}{b_s})}{a^2}\partial\omega_y+ \\ \qquad\qquad \frac{x_m}{a}\partial\omega_z+\frac{T_Y/b_s}{a}\partial T_Y+\frac{\frac{\delta_m}{fb_s}(y_m-x_m\omega_z+f\omega_x-\frac{T_Y\delta_m}{b_s})}{a^2}\partial T_Z \\ \partial\xi = \frac{y_m\delta_m}{fa^2}\partial\omega_x+\frac{x_m\delta_m}{fa^2}\partial\omega_y+\frac{\delta_m^2}{fb_s a^2}\partial T_Z \end{array}\right. \tag{3.3}$$

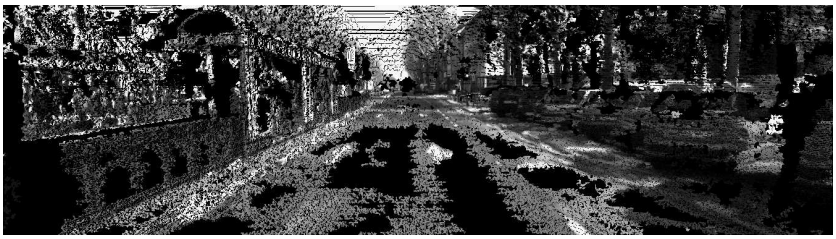
avec :

[4]Cette déformation n'est possible que pour les points pour lesquels une information de disparité est disponible, dès lors, il est rarement possible de pouvoir traiter toute l'image.



(a)

(b)



(c)

(d)

FIGURE 3.4 – Déformation induite par le mouvement : (a) : Carte de Disparité à t_1 ; (b) : Image I_1 ; (c) : Image I_0 déformée ; (d) : Image I_0. Pour des raisons d'illustration, le mouvement a été amplifié ($t_1 - t_0 = 2s$)

(a)

(b)

FIGURE 3.5 – Image I'_1, (a) : $t_1 - t_0 = 2s$ les artefacts visibles (déchirement, traînée, changement d'illumination) sont liés à l'amplification du mouvement utilisée pour des raisons de visibilité ; (b) : $t_1 - t_0 = 0.1s$, les artefacts précédemment observés sont absents.

Entrées : $\mathcal{E}_1$,I_1, I_0, $\tilde{\boldsymbol{T}}$, $\tilde{\boldsymbol{\Omega}}$
Sorties : I'_1

début
  $\mu = 0$
  $\nu = 0$
  $\xi = 0$
  pour tous les $m \in \mathcal{E}_1$ **faire**
    $\pi =$ `CalculeDeplacementLocal`$(m, -\tilde{\boldsymbol{T}}, -\tilde{\boldsymbol{\Omega}})$
    $\mu(m) = \pi.\mu$
    $\nu(m) = \pi.\nu$
    $\xi(m) = \pi.\xi$
  fin
  $I'_1 =$ `AppliqueTransformationInterpolée`(I_0, μ, ν, ξ)
  pour tous les $m \notin \mathcal{E}_1$ **faire**
    $I'_1(m) = I_1(m)$
  fin
fin

Algorithme 2: Algorithme de Compensation de l'Égo-Mouvement

$$a = \frac{x_m}{f}\omega_y - \frac{y_m}{f}\omega_x - \frac{T_Z \delta_m}{f b_s} + 1$$

On peut d'ordinaire considérer que $a \approx 1$[5], et en ne conservant que les termes dominants :

$$\begin{cases} \partial\mu = \frac{x_m y_m}{f}\partial\omega_x + \frac{x_m^2}{f}\partial\omega_y + y\partial\omega_z + \frac{\delta_m}{b_s}\partial T_X + \frac{x_m \delta_m}{f b_s}\partial T_Z \\ \partial\nu = \frac{y_m^2}{f}\partial\omega_x + \frac{x_m y_m}{f}\partial\omega_y + x\partial\omega_z + \frac{\delta_m}{b_s}\partial T_Y + \frac{y_m \delta_m}{f b_s}\partial T_Z \\ \partial\xi = \frac{y_m \delta_m}{f}\partial\omega_x + \frac{x_m \delta_m}{f}\partial\omega_y + \frac{\delta_m^2}{f b_s}\partial T_Z \end{cases} \tag{3.4}$$

Les équations 3.4 nous permettent de définir, pour chaque point m de I_1, un intervalle de confiance, dans lequel son correspondant m' dans I'_1 se trouve :

$$\begin{aligned} \mathcal{W}^m_{(\tilde{\mathbf{\Omega}},\tilde{\boldsymbol{T}})} &= [x_m - \Delta\mu; x_m + \Delta\mu] \\ &\times [y_m - \Delta\nu; y_m + \Delta\nu] \\ &\times [\delta_m - \Delta\xi; \delta_m + \Delta\xi] \end{aligned} \tag{3.5}$$

Plusieurs stratégies sont alors possibles pour estimer $\Delta\mu$, $\Delta\nu$ et $\Delta\xi$. Tout d'abord, il est possible de majorer les dérivées partielles par les estimations de précisions, évaluées en 2.3.1. Cette stratégie a notamment été mise en œuvre dans [BBA10] et à mener à de bons résultats. Elle présente l'avantage de pouvoir être utilisée avec n'importe quel système d'estimation de l'égo-mouvement, dès lors qu'il est suffisamment finement caractérisé, ce qui ne pose pas de problème majeur.

En revanche, cette approche ne repose que sur une base statistique, et ne permet pas de prendre en compte la précision instantanée atteinte par l'odométrie visuelle. Afin d'obtenir une estimation plus fine, nous préférons utiliser l'erreur renvoyée par l'algorithme de RANSAC, présentée en 2.2.6. Cette erreur nous fournit un point fixe de l'erreur de recalage commise, point fixe à partir duquel nous pouvons construire, par extrapolation, l'intervalle de confiance pour tout point de l'image.

Si dans une grande majorité de cas, la différence entre les deux approches est négligeable, cela permet en revanche une meilleure intégration du système. Il n'est ainsi pas nécessaire de prévoir explicitement un cas particulier correspondant aux défaillances de l'odométrie visuelle, une erreur renvoyée par RANSAC très élevée ou infinie entraînant des intervalles de confiance très grands, donc des résultats filtrés car peu crédibles.

[5] En effet, les rotations sont exprimées en radians, et il est généralement possible de considérer que la distance des objets est largement supérieure à la translation axiale du capteur.

Cette estimation de l'imprécision du recalage a plusieurs conséquences, en termes théoriques et pratiques.

- En termes pratiques tout d'abord, cela va pouvoir nous permettre de filtrer les valeurs de déplacement résiduel qui peuvent correspondre à une erreur liée au recalage. Cela peut être fait de plusieurs façons. Si un simple seuillage est envisageable, cette connaissance de l'imprécision peut également être mise à profit au cours du filtrage temporel, qui sera décrit en 3.2.3.

- D'autre part, d'un point de vue plus théorique, la précision du recalage opéré limite ultimement les performances d'un système de détection, ces valeurs seront donc utilisées dans notre étude des facteurs limitants et des pistes d'améliorations envisagées, en 3.3.

3.2.2 Estimation du Mouvement Indépendant

Une fois que l'égo mouvement a été estimé et compensé, le mouvement résiduel observable correspond au mouvement des objets mobiles indépendants de l'égo-véhicule. Ce mouvement résiduel est noté :

$$\boldsymbol{A} = \begin{vmatrix} \mu_A \\ \nu_A \\ \xi_A \end{vmatrix} \tag{3.6}$$

Méthode - Il est important de considérer un déplacement résiduel à 3 dimensions, la troisième étant la variation de disparité. En effet, dans certains cas, le déplacement résiduel dans le plan de l'image peut être nul, alors que le mouvement des objets dynamiques est important, comme illustré en figure 3.6. Ceci est d'autant plus important que ce genre de cas de figure se produit majoritairement pour des objets qui se dirigent le long de l'axe optique, c'est-à-dire sur une trajectoire de collision avec l'égo-véhicule. Le choix de travailler avec la variation de disparité plutôt qu'avec son dual, la variation de profondeur, est dicté par deux constatations :

- Tout d'abord, il nous semble plus cohérent de raisonner de bout en bout dans le domaine image.

- Ensuite, il apparaît que l'espace image présente certaines particularités, en particulier un bruit isotrope, qui facilite les démarches mathématiques [DD02].

FIGURE 3.6 – Le motard mis en évidence se déplace rapidement, toutefois son mouvement latéral est quasiment nul.

Dès lors, il est nécessaire de procéder en deux étapes :

- Tout d'abord, on estime le mouvement latéral par une recherche de correspondants ;
- Ensuite, on calcule le mouvement longitudinal en comparant la disparité que le point considéré *devrait* avoir s'il était immobile, et celle qu'il a *effectivement*.

Cette démarche est illustrée dans l'algorithme 3.

La recherche de correspondants, effectuée en ligne 2 de l'algorithme 3 est une étape critique pour laquelle plusieurs méthodes ont été envisagées. Tout d'abord, une recherche de correspondant au sens de la corrélation croisée[6] a été envisagée [BBA10].

Si les résultats étaient satisfaisants, deux points nous ont conduit à envisager des méthodes de flot optique qui proposent également de résoudre ce problème d'appariement entre deux images :

- En premier lieu, et malgré l'utilisation d'optimisations algorithmiques et matérielles, les temps de calculs étaient prohibitifs pour de grandes images.

[6] Ou de ses approximations

Entrées : I_1, I'_1
Sorties : μ_A, ν_A, ξ_A
début
  $\{\mu_A, \nu_A\}$ =`ChercheCorrespondants(`I'_1, I_1`)`
  pour tous les $m \in I'_1$ **faire**
    $m' = m + \begin{vmatrix} \mu_A \\ \nu_A \end{vmatrix}$
    $\xi_A(m) = m.\delta - m'.\delta$
  fin
fin

Algorithme 3: Détection du mouvement indépendant

Si l'exploitation de cartes de disparité semi-denses était envisageable, en revanche, le coût lié à l'exploitation de cartes denses est rapidement devenu prohibitif.

- Ensuite, une méthode fondée sur des calculs de corrélation entre pixel est très sensible à la définition du domaine dans lequel le correspondant d'un point m va être recherché. Si la taille de ce domaine de recherche est surestimée, les temps de calculs vont inutilement croître polynomialement. A l'inverse, si le domaine de recherche est sous-estimé, des objets au mouvement apparent trop grand seront paradoxalement impossible à percevoir, ou, du moins, impossible à distinguer des points occultés.

Ces problèmes ne se posent pas en utilisant une méthode différentielle, dès lors que l'on a recours à une approche multi-résolution ou itérative.

Si nous avons conclu en 2.2.4 que les méthodes différentielles peuvent poser problème pour l'estimation de l'égo-mouvement, le problème posé résidait majoritairement dans le manque de généralité que l'on pouvait espérer de ces méthodes. Toutefois, nous ne nous trouvons plus dans le même cas de figure et l'amplitude des mouvements image observés est plus faible. En particulier, l'utilisation du flot optique pouvait poser de gros problème lorsque le déplacement apparent dans le domaine image été très grand, notamment lorsque les taux de rotations étaient très élevés (virages, demi-tours, etc.). Ce n'est plus le cas ici, car les rotations de l'égo-véhicule sont compensées et que les rotations des objets d'intérêts vont être soit faibles et sans incidence, soit suffisamment importante pour entraîner un changement de point de vue, insolvable quelque soit la méthode utilisée.

De plus, le recours à des méthodes de flot optique présente l'intérêt majeur d'autoriser une détection dense, ce qui est un de nos principaux objectifs.

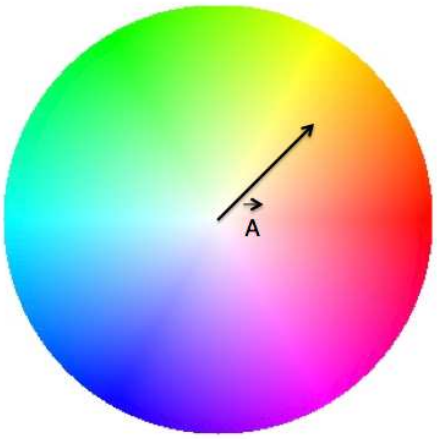


FIGURE 3.7 – Représentation du mouvement résiduel

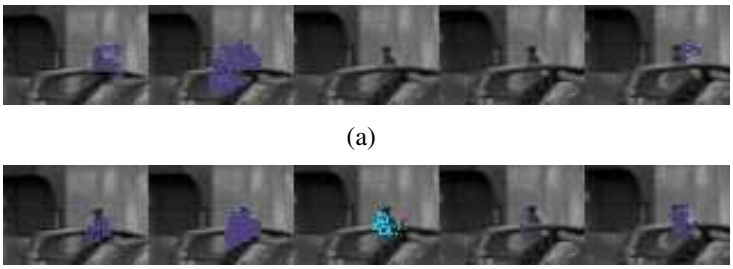
Représentation - Afin de ne pas surcharger les représentations à venir, nous choisissons de procéder de deux manières distinctes. Tout d'abord, les composantes latérale et verticale du mouvement : μ_A et ν_A seront représentées suivant un code couleur inspiré de la norme CIE La*b* et illustré en figure 3.7. La teinte correspond à la direction du mouvement, et la saturation correspond à son amplitude. Cette représentation nous permet de superposer le mouvement extrait et les images d'origine pour plus de clarté. Dans une majorité de cas, cette représentation est suffisante pour attester visuellement du mouvement indépendant des cibles. Toutefois, certains cas particulier, tel que celui exposé en figure 3.6 nécessite une visualisation de la composante axiale du mouvement résiduel. Dans ces conditions, nous ferons figurer cette composante sous la forme d'une carte en niveaux de gris de l'amplitude du déplacement axial résiduel.

3.2.3 Intégration Temporelle

Il nous semble primordial d'ajouter une prise en compte temporelle à notre méthode de détection. En effet, de nombreux faux négatifs ne présentant pas de cohérence temporelle peuvent être éliminés de la sorte (par exemple, des faibles sautes brusques d'intensité), comme illustré en figure 3.8. De la même façon, certains objets faiblement mobiles, comme illustré en figure 3.9, peuvent être plus facilement détectés si l'information temporelle est prise en compte.

Nous voyons deux grandes stratégies permettant de mettre en oeuvre une intégration temporelle des résultats de détection. Tout d'abord, un suivi, à l'échelle

FIGURE 3.8 – Exemple de faux positifs, dûs à un brusque changement d'illumination, correctibles par une intégration temporelle des résultats de détection.



(a)

(b)

FIGURE 3.9 – (a)Exemple de faux négatifs, si l'objet est mobile pendant les 5 images, en revanche, il n'est détecté que dans 3 de ces images ; (b) L'ajout de l'intégration temporelle nous permet d'obtenir une détection continue de l'objet.

d'une cible peut être opéré, par exemple en utilisant la méthode Mean-Shift [CM02] ou encore un filtrage de Kalman. Cette approche permet de réduire considérablement la quantité d'objets à suivre, tout en accomplissant une détection et une représentation de haut-niveau. Toutefois, ce type de méthodes nécessite également un mécanismes d'extraction de cibles, qui pourrait être sujet à des erreurs supplémentaires. De plus, un certains nombre de problèmes, comme le suivi de cibles partiellement ou totalement occultées pourrait survenir.

A l'opposé, une méthode pixellique nous autorise une plus grande souplesse à ces différents niveaux, tout en n'interdisant pas l'extraction d'information de haut-niveau par la suite. Pour cela nous utilisons la fonction logistique symétrique suivante :

$$f_{(Th,\lambda)}(x) = \frac{1}{1+e^{-\lambda(|x|-Th)}} \tag{3.7}$$

Le taux de croissance λ sera utilisé pour ajuster le système, de façon à avoir une détection rapide pour des objets rapides, et une détection robuste pour les objets lents. Le seuil Th est défini de manière indépendante pour les trois composantes du mouvement, nous utilisons ici les bornes des intervalles de confiance, définis en 3.2.1.

Nous pouvons alors définir une image de confiance instantanée, ICI, comme :

$$ICI(m) = max\left\{f_{(Th_i,\lambda_i)}(i_A(m)) \; \forall \; i \in \{\mu, \nu, \xi\}\right\} \tag{3.8}$$

Cette image indique, pour le point m, la crédibilité de l'hypothèse "*m est mobile*". Puis nous pouvons définir une version intégrée dans le temps de cette image de confiance, CI, comme :

$$CI(m) = \alpha \, ICI(m) + (1-\alpha) CI(m) \tag{3.9}$$

Où le facteur α permet d'ajuster le taux de rafraichissement de cette image, en pratique, nous fixons $\alpha = 0,3$, ce qui permet de fonder la détection sur les trois dernières images. Cela nous semble être un bon compromis entre la prise en compte de l'information temporelle et le retard de la détection. De plus, cette image de "crédibilité du mouvement" nous sert à exclure certaines zones de l'image de l'extraction de points d'intérêt, préalable à l'estimation du mouvement par odométie visuelle.

Il faut toutefois noter que ces images de "crédibilité de mouvement" doivent également être compensées vis-à-vis du mouvement de l'égo-véhicule et des objets mobiles de la scène. Pour cette raison, on applique successivement la compen-

sation de l'égo-mouvement présentée en 3.2.1, puis une compensation du mouvement résiduel $\boldsymbol{A}$.

3.2.4 Segmentation

Dans une optique d'intégration à des systèmes plus complets, il nous semble important de pouvoir fournir une détection de plus haut niveau qu'une cartographie du mouvement indépendant. En effet, il nous semble que certains sous-systèmes d'un véhicule intelligent pourraient bénéficier d'une information plus synthétique.

A ces fins, nous avons également mis en place une procédure d'agglomération (ou *clustering* en anglais) des cibles. Pour cela, nous avons utilisé une méthode de clustering en K-Moyennes (*K-Means*), en particulier la méthode "K-Means++", présentée par ARTHUR et VASSILVITSKII [AV07]. La méthode des K-Moyennes permet d'extraire un nombre arbitraire et défini de clusters. Pour cette raison, nous testons toutes les valeurs de K possibles, entre 1 et 30. Cette borne supérieure nous semble suffisamment élevée pour ne pas être pris au dépourvu par un nombre particulièrement élevé de cibles présentes à l'image.

Afin d'identifier la meilleure partition, et donc le meilleur nombre de cibles, nous comparons les 30 partitions obtenues en utilisant leur compacité comme métrique. La compacité d'une partition peut s'exprimer :

$$c_K = \sum_i \sum_{x \in C_i} \|x - \bar{x}_{C_i}\| \tag{3.10}$$

où C_i est le $i^{\text{ème}}$ cluster de la partition, et $\bar{x}_{C_i}$ est son centroïde.

Cette méthode, très simple, ne présente pas d'innovation particulière. Toutefois, à notre sens, il est important que le système développé dans le cadre de ce travail de thèse puisse être considéré comme une *preuve de concept*, et cela ne serait pas possible sans cette dernière étape d'intégration spatiale des cibles détectées.

3.3 Performances, limitations et pistes d'améliorations

Au cours de cette section, nous allons étudier plusieurs caractéristiques du système de détection, afin de pouvoir évaluer l'importance des différents paramètres sur ces performances.

(a) (b)

FIGURE 3.10 – Illustration de la méthode de Clustering : (a) Image de détection initiale ; (b) : clusters extraits

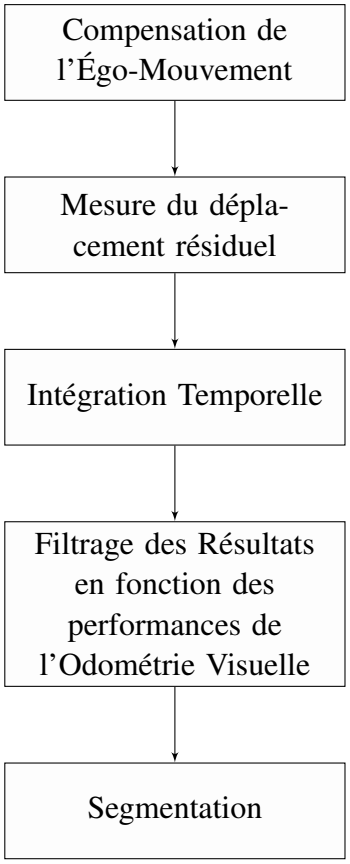


FIGURE 3.11 – Résumé du fonctionnement de la détection d'objets mobiles

Tout d'abord, nous allons nous pencher sur la sensibilité de détection, c'est-à-dire sur les plus petits mouvements discernables. Ensuite, nous évaluerons le pouvoir de séparation, c'est-à-dire, la capacité du système à différencier deux objets proches. Finalement, nous étudierons l'impact de la résolution temporelle, qui est un paramètre généralement fixé par défaut, voire considéré comme un indicateur de performances.

3.3.1 Sensibilité de Détection

Nous avons précédemment mis en lumière une limitation, dûe à l'imprécision de l'odométrie visuelle[7], qui mène à la définition de seuils, dans l'espace image, en dessous desquels un déplacement ne peut pas être considéré comme étant l'image d'un mouvement indépendant. L'objet de cette section est d'évaluer l'impact de cette limitation sur la détection d'objets mobiles. Cela sera fait en évaluant sous quelles conditions (position, mouvement objet), un point objet dynamique peut effectivement être détecté. Pour cela, considérons un objet ponctuel mobile M_0:

$$M_0 = \begin{vmatrix} X_{M_0} \\ Y_{M_0} \\ Z_{M_0} \end{vmatrix} \tag{3.11}$$

Son image par le système de vision est m_0 :

$$m_0 = \begin{vmatrix} x_{m_0} \\ y_{m_0} \\ \delta_{m_0} \end{vmatrix} \tag{3.12}$$

Son mouvement propre est noté :

$$\boldsymbol{T}^{M_0} = \begin{vmatrix} T_X^M \\ T_Y^M \\ T_Z^M \end{vmatrix} \tag{3.13}$$

Étant donné que cet objet est considéré ponctuel, il n'est pas pertinent de considérer une composante rotationelle de son mouvement propre. Son mouvement par rapport au capteur est alors :

$$M_1 = \begin{vmatrix} X_{M_1} \\ Y_{M_1} \\ Z_{M_1} \end{vmatrix} = \begin{vmatrix} X_{M_0} \\ Y_{M_0} \\ Z_{M_0} \end{vmatrix} + \boldsymbol{\Omega} \wedge \begin{vmatrix} X_{M_0} \\ Y_{M_0} \\ Z_{M_0} \end{vmatrix} - \begin{vmatrix} T_X \\ T_Y \\ T_Z \end{vmatrix} - \begin{vmatrix} T_X^M \\ T_Y^M \\ T_Z^M \end{vmatrix} \tag{3.14}$$

[7] mais partagée *a priori* par toutes les méthodes existantes d'estimation de l'égo-mouvement

Dès lors, son déplacement image peut s'écrire :

$$\left|\begin{array}{l} \mu_m = \dfrac{x_{m_0} + y_{m_0}\omega_Z - f\omega_Y - \frac{\delta_{m_0}(T_X+T_X^M)}{b_s}}{\frac{x_{m_0}}{f}\omega_Y - \frac{y_{m_0}}{f}\omega_X - \frac{\delta_{m_0}(T_Z+T_Z^M)}{fb_s} + 1} - x_{m_0} \\ \\ \nu_m = \dfrac{y_{m_0} - x_{m_0}\omega_Z + f\omega_X - \frac{\delta_{m_0}(T_Y+T_Y^M)}{b_s}}{\frac{x_{m_0}}{f}\omega_Y - \frac{y_{m_0}}{f}\omega_X - \frac{\delta_{m_0}(T_Z+T_Z^M)}{fb_s} + 1} - y_{m_0} \\ \\ \xi_m = \delta_{m_0}\dfrac{y_{m_0}\omega_X - x_{m_0}\omega_Y + \frac{T_Z+T_Z^M}{b_s}}{x_{m_0}\omega_Y - y_{m_0}\omega_X - \frac{T_Z+T_Z^M}{b_s} + 1} \end{array}\right. \tag{3.15}$$

Le point M sera considéré, à tort, comme étant statique s'il satisfait les inéquations suivantes :

$$\left\{\begin{array}{l} \left|\mu_m - \mu_{(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}})}(m)\right| \leq \Delta\mu \\ \left|\nu_m - \nu_{(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}})}(m)\right| \leq \Delta\nu \\ \left|\xi_m - \xi_{(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}})}(m)\right| \leq \Delta\xi \end{array}\right. \tag{3.16}$$

L'équation 3.16 introduit un système de 3 inéquations, pour 6 inconnues : x_{m_0}, y_{m_0}, δ_{m_0}, T_X^M, T_Y^M et T_Z^M. Même en considérant que, la plupart du temps, T_Y^M est très contraint, ce système reste fortement indéterminé.

Néanmoins, nous pouvons identifier plusieurs solutions évidentes :

- Tout d'abord, un objet dont le mouvement propre est inférieur au bruit d'extraction ne pourra pas être identifié.

- Ensuite, un objet situé à l'infini ($\delta_{m_0} = 0$) ne pourra pas être détecté non plus, quelque soit son mouvement. Les termes dépendants du mouvement indépendant dans l'équation 3.15 étant annulés.

Toutefois, d'autres solutions sont possibles. Ainsi, par souci de simplification, nous considérons que l'égo-véhicule connaît un mouvement purement translationnel, les inéquations 3.16 deviennent :

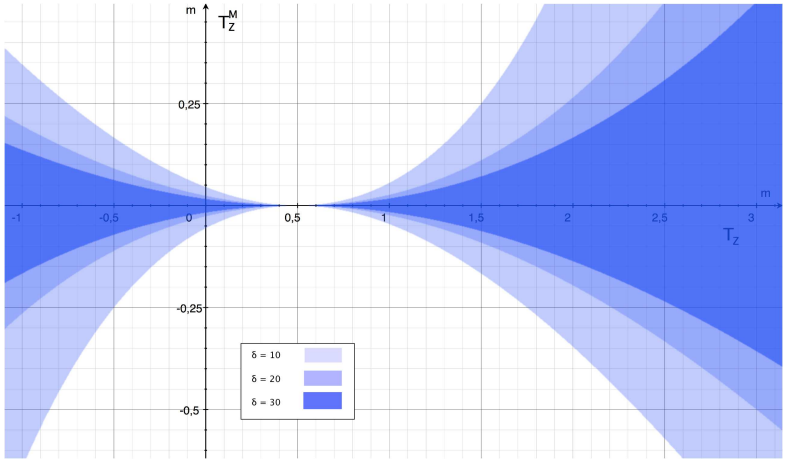


FIGURE 3.12 – Répartition des valeurs de T_Z^M satisfaisant l'équation 3.17 , en fonction de T_Z pour trois objets présentant des disparités de 10, 20 et 30 pixels, c'est à dire des distances d'environ 50, 25 et 17m. Une valeur réaliste de $\Delta\xi = 1$ a été considérée, la zone en bleu représente l'étendu des mouvements indétectables. La base considérée pour ce calcul est de 50cm.

$$\begin{cases} \left|\mu_m - \mu_{(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}})}(m)\right| \leq \Delta\mu \\ \left|\nu_m - \nu_{(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}})}(m)\right| \leq \Delta\nu \\ \left|\xi_m - \xi_{(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{\Omega}})}(m)\right| \leq \Delta\xi \end{cases} \Rightarrow \begin{array}{l} \left|\frac{\frac{x_{m_0}}{f}\frac{\delta_{m_0}T_Z^M}{b_s} - \frac{\delta_{m_0}T_X^M}{b_s}}{(1-\frac{T_Z}{b_s})(1-\frac{T_Z}{b_s}-\frac{T_Z^M}{b_s})}\right| \leq \Delta\mu \\ \left|\frac{\frac{y_{m_0}}{f}\frac{\delta_{m_0}T_Z^M}{b_s} - \frac{\delta_{m_0}T_Y^M}{b_s}}{(1-\frac{T_Z}{b_s})(1-\frac{T_Z}{b_s}-\frac{T_Z^M}{b_s})}\right| \leq \Delta\nu \\ \left|\frac{\delta_{m_0}T_Z^M}{b_s(1-\frac{T_Z}{b_s})(1-\frac{T_Z}{b_s}-\frac{T_Z^M}{b_s})}\right| \leq \Delta\xi \end{array} \tag{3.17}$$

Il apparaît que les solutions en T_X^M et T_Y^M sont conditionnées par T_Z^M. Il est donc nécessaire de commencer par trouver les T_Z^M satisfaisant la troisième inéquation. Des solutions à ce problème sont illustrées en figures 3.12 et 3.13.

En particulier, la figure 3.12 illustre l'importance du choix de la base du capteur stéréoscopique. En effet, pour $T_Z \approx b_s$ les possibilité de mouvement non-détectable sont quasiment nulles. Dès lors, la base du capteur stéréo doit être dimensionnée en ayant connaissance des données suivantes :

- La plage de vitesses opérationnelle du véhicule

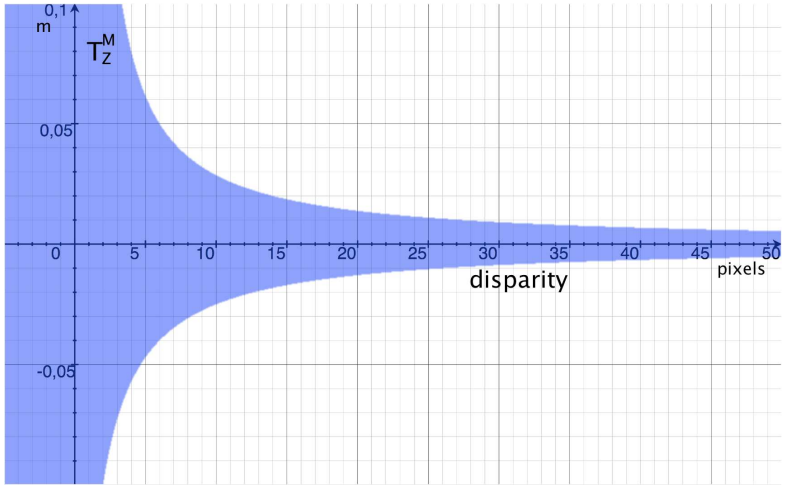


FIGURE 3.13 – Répartition des valeurs de T_Z^M satisfaisant l'équation 3.17, en fonction de d, en considérant une translation axiale de $1m$, soit une vitesse de $50km.h^{-1}$ à $15Hz$. Une valeur réaliste de $\Delta\xi = 1$ a été considérée, la zone en bleu représente l'étendu des mouvements indétectables.

- La cadence d'acquisition des images

Dans le cas du véhicule de test CARLLA (cf D.2.2), nous avons choisi une base stéréoscopique de 50 cm qui est optimale à une vitesse de $22km.h^{-1}$, à une cadence d'acquisition de $12,5Hz$. Si nous avions choisi une base optimale pour des vitesses de l'ordre de $50km.h^{-1}$, l'erreur à basse vitesse aurait été beaucoup plus importante. En choisissant d'optimiser notre système pour une vitesse plus basse[8], nous nous assurons que l'erreur sera minimale sur une plage de 0 à $50km.h^{-1}$. En particulier, le déplacement longitudinal indétectable le plus important sera de $0,03m$ soit $1,5km/h^{-1}$.

Une fois que l'espace des solutions en T_Z^M a été identifié, et que T_Z^M a été fixé, il reste à trouver les solutions en T_X^M et T_Y^M :

[8]Idéalement, une base de $60cm$ serait optimale pour ce type de vitesse, mais les contraintes mécaniques, inhérentes à une intégration réelle nous ont empêché de choisir avec précision la base la plus efficace pour notre système.

$$\left| \frac{\frac{x_{m_0}}{f}\frac{\delta_{m_0} T_Z^M}{b_s} - \frac{\delta_{m_0} T_X^M}{b_s}}{(1-\frac{T_Z}{b_s})(1-\frac{T_Z}{b_s}-\frac{T_Z^M}{b_s})} \right| \leq \Delta\mu$$
$$\left| \frac{\frac{y_{m_0}}{f}\frac{\delta_{m_0} T_Z^M}{b_s} - \frac{\delta_{m_0} T_Y^M}{b_s}}{(1-\frac{T_Z}{b_s})(1-\frac{T_Z}{b_s}-\frac{T_Z^M}{b_s})} \right| \leq \Delta\nu \tag{3.18}$$

L'existence de solutions dépend de plusieurs facteurs :

- La distance entre le capteur et l'objet : $Z \propto \frac{b_s}{\delta_{m_0}}$
- Les rapports entre longueur focale et coordonnées images : $\frac{x_{m_0}}{f}$ et $\frac{y_{m_0}}{f}$

C'est principalement ce second terme qui peut nous intéresser. Si nous n'avons aucune prise sur la distance à laquelle se trouvent les objets à localiser, en revanche les caractéristiques des capteurs sont définis par nos soins. En d'autres termes, plus la focale du système utilisé va être courte, plus l'espace des solutions possibles, c'est-à-dire des mouvements indétectables, grandit.

Dans le cas du système mis en place sur le véhicule CARLLA, nous avons cherché à contenir cette source d'erreur en fixant un rapport $\frac{x_{m_0}}{f}$ relativement bas. Des problèmes peuvent en revanche se poser pour des systèmes utilisant une optique grand angle, comme par exemple le système mis en place par l'Université de Karlsruhe. Ainsi, plus les objets seront éloignés du centre de l'image, plus l'étendue de leur mouvements indétectables va croître. Cela peut se montrer particulièrement critique pour des objets ayant un mouvement relativement faible, comme des piétons. S'il peut paraître intéressant d'augmenter le champ des caméras pour améliorer les capacités de détection, il est nécessaire de considérer que la détection sera d'autant moins efficace que les points considérés sont loin du centre de l'image.

Bien que nous ayons considéré un mouvement de l'égo-véhicule purement translationnel, cela ne change en rien nos conclusions. En effet, les termes dépendant des rotations sont également en $\frac{x_{m_0}}{f}$ ou en $\frac{y_{m_0}}{f}$.

Si certaines de conclusions étaient relativement connues, à notre connaissance, cette réflexion sur la sensibilité du système n'a jamais été menée. Il apparaît que les mouvements indétectables sont très fortement impactés par les paramètres des capteurs de vision. Il nous semble important de prendre ces considérations en compte lors de la conception d'un système de vision ayant pour objectif de détecter les objets dynamiques.

FIGURE 3.14 – Pouvoir de Résolution : Les deux cibles mises en évidence sont coïncidentes, il est toutefois nécessaire de pouvoir les différencier car l'une peut présenter un danger, alors que l'autre n'est pas sur une trajectoire de collision avec l'égo-véhicule.

3.3.2 Pouvoir de Séparation - Résolution Temporelle

Considérons maintenant l'aptitude de notre système à distinguer deux objets présentant deux mouvements différents. Ce type de scénario peut en effet se révéler particulièrement important. On peut par exemple envisager le cas d'un piéton traversant devant une voiture en mouvement, comme illustré sur la figure 3.14.

Afin d'étudier le pouvoir de résolution de notre système, et surtout l'influence que vont avoir les différents paramètres du système sur ce dernier, nous considérons deux points objets M et N, dont les mouvements propres sont $\boldsymbol{T}^M$ et $\boldsymbol{T}^N$. Nous nous plaçons dans le pire cas possible, c'est à dire que nous considérons que les deux points M et N sont coïncidents, et que leurs coordonnées sont :

$$M = N = \begin{vmatrix} X_{MN} \\ Y_{MN} \\ Z_{MN} \end{vmatrix} \tag{3.19}$$

Les coordonnées de leur image sont :

$$m = n = \left| \begin{array}{l} x_{mn} \\ y_{mn} \\ \delta_{mn} \end{array} \right. \tag{3.20}$$

Considérer que les deux points sont coïncidents est, bien sûr, totalement hypothétique, mais il s'agit, ici de déterminer le pouvoir de résolution uniquement dû à l'extraction du mouvement indépendant.

La capacité à différencier ces deux points ne peut reposer que sur la différence de leur mouvement résiduel. Étant donné que les deux points sont coïncidents, tout se passe comme si ils n'étaient pas affectés par le bruit inhérent au recalage des images[9]. Dès lors, la différence de mouvement résiduel peut être écrite :

$$\boldsymbol{A}_M - \boldsymbol{A}_N = \left| \begin{array}{l} \dfrac{\frac{b\delta_{mn}}{b_s}(T_X^M - T_X^N)}{b^2} \\ \dfrac{\frac{b\delta_{mn}}{b_s}(T_Y^M - T_Y^N)}{b^2} \\ \dfrac{\delta_{mn}^2}{fb_s} \dfrac{T_Z^N - T_Z^M}{(b - \frac{\delta_{mn} T_Z^M}{fb_s})(b - \frac{\delta_{mn} T_Z^N}{fb_s})} \end{array} \right. \tag{3.21}$$

avec $b = \frac{x_{mn}}{f}\omega_y - \frac{y_{mn}}{f}\omega_x - \frac{\delta_{mn} T_Z}{fb_s} + 1$

Il apparaît que le pouvoir de résolution ne dépend *a priori* que de la distance des objets au capteur et de la différence absolue entre leur mouvement. Il n'est pas possible ici de jouer sur les paramètres du système de vision pour optimiser la faculté à discerner deux objets. En revanche, la cadence d'acquisition peut jouer un rôle primordial, rôle que nous allons examiner dans la suite.

L'importance de la résolution temporelle est rarement discuté en détails. Toutefois, il est trivial que le mouvement entre deux instants peut être réécrit comme étant :

$$\begin{array}{rl} \overrightarrow{T} = \left| \begin{array}{l} T_X \\ T_Y \\ T_Z \end{array} \right. & = \frac{1}{f_{acq}} \overrightarrow{V} = \frac{1}{f_{acq}} \left| \begin{array}{l} V_X \\ V_Y \\ V_Z \end{array} \right. \\ \overrightarrow{\Omega} = \left| \begin{array}{l} \omega_X \\ \omega_Y \\ \omega_Z \end{array} \right. & = \frac{1}{f_{acq}} \overrightarrow{R} = \frac{1}{f_{acq}} \left| \begin{array}{l} R_X \\ R_Y \\ R_Z \end{array} \right. \end{array} \tag{3.22}$$

[9] L'imprécision ne dépend en effet que de la position du point considéré.

où f_{acq} est la fréquence d'acquisition des images, $V_{i,i\in\{X,Y,Z\}}$ sont les vitesses instantanées et $R_{i,i\in\{X,Y,Z\}}$ sont les taux de rotation instantanés. En utilisant cette notation, il est immédiat que plus la cadence est basse, plus le pouvoir de résolution sera élevé. Toutefois, une baisse de cette cadence d'acquisition présente plusieurs inconvénients.

1. Une diminution de la cadence d'acquisition des images se traduit également par une augmentation du temps nécessaire à la détection des cibles.

2. L'estimation de l'égo-mouvement peut être perturbé par une cadence image trop basse. En effet, dans l'expression du mouvement utilisée en 2.2, nous faisons explicitement l'hypothèse que les lignes trigonométriques peuvent être linéarisée. Si cela ne pose généralement pas de problème, il est possible qu'en diminuant la cadence d'acquisition, les rotations deviennent trop importantes pour pouvoir prendre une telle hypothèse.

Étant donné que l'effet du premier inconvénient est immédiat, nous allons nous attacher à quantifier le second.

Dans des conditions de circulation normales, la rotation la plus importante est le lacet (rotation autour de l'axe vertical de l'égo-véhicule). Dès lors, même si les autres rotations ne sont pas nulles, la première à violer l'hypothèse de linéarité des lignes trigonométriques sera le lacet. En conséquence, nous allons nous concentrer sur cette dernière. Nous ne considérons évidemment pas les composantes translationnelles du mouvement. Finalement, en nous contentant d'une approximation au second ordre, les équations de l'image du mouvement deviennent :

$$\left|\begin{array}{l} \mu = \dfrac{x_m + \frac{y_m}{f_{acq}} R_Z \cos\frac{R_Y}{f_{acq}} - f\frac{R_Y}{f_{acq}}}{x_m R_Y - y_m R_X \cos\frac{R_Y}{f_{acq}} + 1} - x_m \\ \\ \nu = \dfrac{y_m - \frac{x_m R_Z}{f_{acq}} \cos\frac{R_Y}{f_{acq}} + f\frac{R_X}{f_{acq}} \cos R_Y f_{acq}}{x_m R_Y - y_m R_X \cos\frac{R_Y}{f_{acq}} + 1} - y_m \\ \\ \xi = \delta_m \dfrac{y_m R_X \cos\frac{R_Y}{f_{acq}} - x_m R_Y}{x_m R_Y - y_m R_X \cos\frac{R_Y}{f_{acq}} + 1} \end{array}\right. \tag{3.23}$$

L'erreur commise sur les trois composantes du mouvement en linéarisant à tort les lignes trigonométriques est ainsi représentée en figure 3.15, dans le cas suivant :

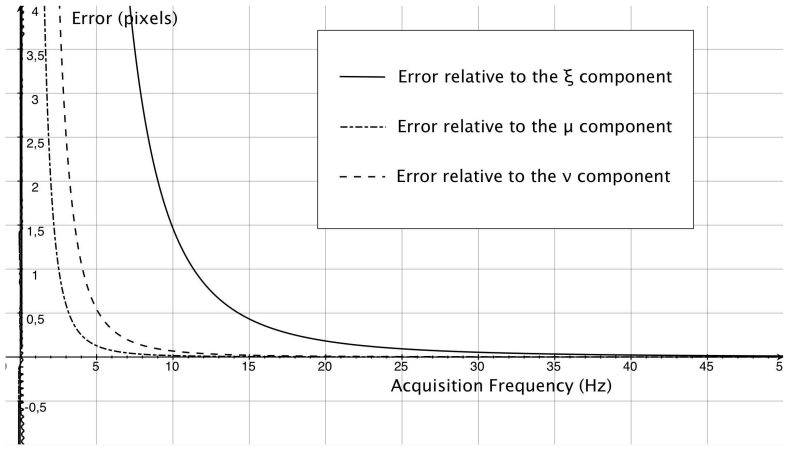


FIGURE 3.15 – Erreur commise en linéarisant à tort les lignes trigonométriques pour un scénario correspondant au pire cas possible.

- les taux de rotations considérés correspondent aux valeurs les plus importantes observées lors des expérimentations urbaines menées avec CARLLA.

- les paramètres intrinsèques du système sont identiques à ceux du système embarqué sur CARLLA.

- le point considéré est en bord de champ, ce qui constitue le pire cas possible.

Il apparaît que l'erreur sur les trois composantes augmente dramatiquement dès que l'on passe sous la barre des $10Hz$. Il apparaît également que des cadences d'acquisition entre $10Hz$ et $15Hz$ permettent d'augmenter le pouvoir de résolution, tout en assurant une détection rapide (moins d'un dixième de seconde) et en n'introduisant qu'une erreur négligeable, sur une grande partie du champ image, due à la linéarisation des lignes trigonométriques.

En conclusion, il nous apparaît que les standards $25Hz$ (Europe) et $30Hz$ (Amérique du Nord), s'ils sont bien adapté à une prise d'images destinée à l'œil humain ne conviennent pas à nos traitements. Cela est rarement pris en compte, or, l'intérêt, illustré en figure 3.16 est immédiat. Une prise en compte fine de ce paramètre nous apporte plusieurs avantages. Le pouvoir de résolution de notre système augmente, ainsi que sa sensibilité, mais il est également notable que cela nous permet de relaxer légèrement les contraintes en termes de temps de calculs de nos algorithmes.

FIGURE 3.16 – Améioration du pouvoir de résolution obtenue en diminuant la cadence d'acquisition : à gauche la cadence est de $15Hz$, les deux cibles sont correctement prises en compte ; à droite, la cadence est de $30Hz$ et le piéton, plus lent, n'est pas détecté.

3.4 Résultats

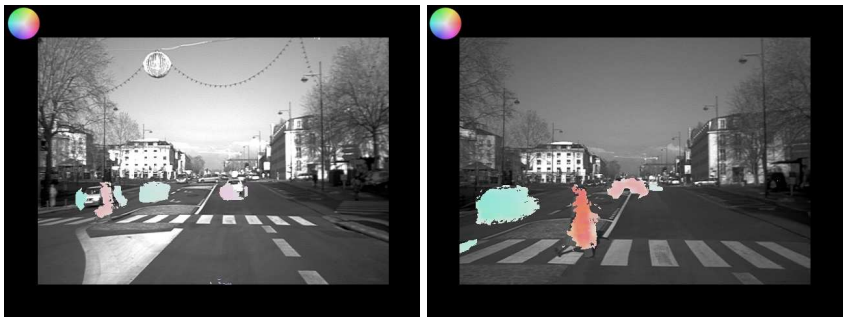
Dans cette section, nous allons tout d'abord présenter des résultats généraux, illustrés et quantifiés, issus d'expérimentations sur des images réelles. Dans un second temps, nous allons nous pencher sur plusieurs cas problématiques qui illustrent un certain nombre de limitations de notre système. Ces limitations sont autant de pistes d'améliorations futures.

3.4.1 Résultats Généraux

Dans un premier temps, nous allons présenter et commenter un certain nombre de résultats ponctuels, avant de présenter les résultats quantitatifs finaux de notre étude. Toutefois, nous sommes conscients que des images fixes rendent mal compte de la détection réalisée, aussi, plusieurs séquences vidéos sont disponibles en tant que pièces jointes numériques à ce manuscrit.

Deux des systèmes à notre disposition répondent aux exigences que nous avons pu définir précédemment : le système de l'UTC (cf. annexe D.2.5) et le véhicule CARLLA (cf. annexe D.2.2). Les séquences issues du projet LoVe remplissent notre cahier des charges en termes d'imagerie, mais leur manque de cohérence temporelle rend leur utilisation relativement ardue, la figure 3.17 illustre l'application de notre méthode sur ces séquences, en particulier dans un cas relativement complexe.

Dans un premier temps, nous allons présenter des résultats généraux, aussi bien qualitatifs que quantitatifs, avant de revenir sur un certain nombre de cas problématiques que nous avons rencontrés.



(a) (b)

FIGURE 3.17 – Résultats de détection obtenus à partir d'une séquence LoVE. Les images (a) et (b) sont séparées de quelques seconde. Dans les deux cas, les cibles son toutes bien extraites

Vrais Positifs	Faux Positifs
89%	11%

TABLE 3.1 – Répartition des détections - Séquence UTC

Données UTC - La séquence mise à disposition par l'UTC consiste en 2100 paires stéréos rectifiées, prises à $30Hz$, soit 1050 paires à $15Hz$, fréquence que nous préfèrerons utiliser. Au cours de cette séquence, 354 détections ont été réalisées, réparties entre faux et vrais positifs, comme indiqué en table 3.1. Les faux positifs sont majoritairement dûs à l'apparition d'un motif très répétitif (grillage), illustré en figure 3.18.

Malgré ces inconvénients, notre système a donné de bons résultats, comme, par exemple, en figure 3.19. Aucun faux négatif n'est survenu durant cette séquence. Pour information, nous considérons qu'une non-détection est un faux négatif si les deux conditions suivantes sont remplies :

- L'objet mobile non-détecté est présent à l'image depuis 3 images.
- L'information de disparité est présente sur tout ou partie de l'objet mobile non-détecté depuis au moins 3 images.

En effet, notre système est tributaire du calcul de carte de disparité, et celui-ci n'est pas infaillible.

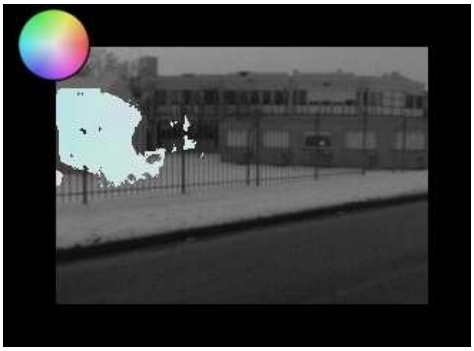


FIGURE 3.18 – Exemple de faux positifs survenus dans la séquence UTC

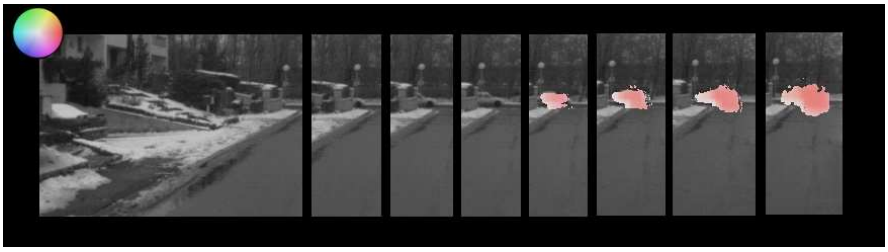


FIGURE 3.19 – Exemples de résultats obtenus. 8 frames successives sont représentées. Le véhicule apparaît en frame 2, son mouvement est intégré en frame 2, 3 et 4. La détection commence en frame 5. L'égo-véhicule avance lentement, en tournant vers sa droite.

Vrais Positifs	Faux Positifs
89%	11%

TABLE 3.2 – Répartition des détections - Séquences CARLLA

FIGURE 3.20 – Exemple de résultats obtenus. 12 frames successives sont représentées. Le véhicule est présent pendant ces douze frames, mais, jusqu'à la frame 5, sa distance au capteur est trop importante pour qu'il soit détecté, ceci illustre ce qui a été vu en 3.3.1

Données CARLLA - Les données récoltées par le système CARLLA constituent également une très bonne source pour l'évaluation de notre système. Nous avons toutefois concentré nos efforts sur les séquences pertinentes de cette base de données, c'est-à-dire celles prenant places dans un environnement urbain, dans lequel évoluent des objets mobiles. Cela représente environ 5 000 paires stéréo exploitables pour cette étude.

On peut retrouver en table 3.2 les résultats en termes de faux positifs et de bonnes détections obtenus sur l'ensemble des séquences évaluées. Au total, 2515 détections ont été dénombrées, dont 283 faux positifs. Les résultats s'avèrent très comparables à ceux obtenus sur les données de l'UTC. Très peu de faux négatifs ont été dénombrés : 46 sur l'ensemble des séquences utilisées. Ces faux négatifs correspondent en grande majorité à des objets insuffisamment texturé pour permettre une bonne évaluation du flot optique, ou a des zones très faiblement contrastées, ce dernier point pouvant être amélioré de façon matériel, en jouant sur les capteurs utilisés.

La figure 3.21 présente la carte d'intégration *CI* obtenue pour une image particulière. Cette carte est utilisée afin de filtrer les résultats et d'aliminer les faux positifs, mais aussi afin de supprimer les points d'intérêt présents sur les objets mobiles lors de l'odométrie visuelle. Il apparaît ici que les deux véhicules en mouvement sont bien détecté, tandis que le véhicule arrêté au milieu du carrefour apparaît statique. Quelques faux positifs, notamment sur la façade du bâtiment en arrière plan sont efficacement filtrés par notre intégration temporelle.

FIGURE 3.21 – A gauche : image de détection ; A droite : carte d'intégration temporelle des résultats

La figure 3.22 revient sur l'importance de la prise en compte de la composante axiale du mouvement résiduel des objets. La détection du véhicule au premier plan est ainsi facilitée, sinon due, à cette prise en compte.

3.4.2 Limitations - Cas Problématiques

Au cours de ces expérimentations, nous avons également été en mesure d'isoler et d'analyser les différentes circonstances qui vont mener à de mauvais résultats (faux-négatifs ou faux-positifs). Il s'agit dans l'immense majorité de cas problématiques pour les deux procédés utilisés : la mise en correspondance stéréoscopique et la mise en correspondance temporelle.

Reflets Spéculaires - En premier lieu, les réflexions sont une importante source de problème. Les reflets vont ainsi être à l'origine de nombreux faux-positifs. En effet, les objets réfléchissants vont se comporter comme les objets qu'il reflètent : leur profondeur et leur mouvement deviennent donc indéterminés.

Dans un environnement urbain, donc hautement manufacturé, les sources de réflexions spéculaires sont très nombreuses :

- les carrosseries des autres véhicules
- Les vitrines et parois vitrées
- Les flaques d'eau

Le problème de l'élimination de ces réflexions n'est pas neuf en traitement des images, et a déjà été abordé à de nombreuses reprises. Toutefois, un traitement logiciel de ces réflexions repose nécessairement sur l'utilisation de l'information

FIGURE 3.22 – Illustration d'une détection reposant sur le mouvement axiale. Le véhicule détecté présente un mouvement purement longitudinal. Les composantes latérales de son mouvement résiduel sont faibles, contrairement à la composante axiale de son mouvement résiduel, qui permet d'assurer une bonne détection. Il est toutefois notable que cette détection est limité par la densité de la carte de disparité, qui peut présenter des "trous".

colorimétrique [Sha85, YK04]. Les méthodes décrites dans ce document ont été conçues de façon à exploiter des images en niveaux de gris, et il ne nous semble pas judicieux de tripler la masse de données à traiter, et donc la charge computationelle afin de supprimer les réflexions spéculaires. Nous pensons que la solution idéale se trouve en amont des traitements, en exploitant une propriété physique des images réfléchies : leur polarisation [Wol89]. En effet, l'ajout de polariseurs en amont des caméras ne pose pas de problème d'intégration, ces composants peuvent être très peu coûteux.

Motifs Répétitifs - De même, il est naturellement apparu que les objets qui peuvent se montrer problématiques pour la stéréovision ou l'imagerie du mouvement demeurent une source d'erreurs pour notre méthode qui est le fruit d'une collaboration stéréo/mouvement. Ces objets sont principalement ceux qui vont présenter un motif répétitif, c'est notamment le cas des marquages au sol, barrières, routes pavées, façades d'immeubles, etc. Si nous avons réussi à limiter partiellement l'impact de tels points en ce qui concerne l'odométrie visuelle (cf. 2.2.5), le problème demeure entier pour la détection des objets dynamiques. Nous n'avons en effet pas été en mesure d'exhiber une méthode robuste permettant de s'en prémunir.

Variations d'Intensité - Le dernier étage de notre approche reposant sur une méthode de flot optique, il est normal que des variations d'intensité, locales ou globales, viennent perturber les résultats obtenus. Ces variations d'intensité peuvent

FIGURE 3.23 – Variations d'intensité dues à des surfaces non-lamberienes. Les variations ont été artificiellement augmentées à des fins d'illustration.

avoir deux causes principales :

- Variations Globales : l'éclairement global de la scène peut évoluer brusquement (c.f. Fig. 3.8) que ce soit parce que les sources d'éclairage évoluent (soleil, éclairage public, ...) ou parce que les paramètres des caméras (temps d'exposition et gain) ne sont pas fixes.

- Variations Locales : les surfaces constituant la scène observée ne sont pas lambertiennes. En d'autres termes, leur luminosité est susceptible d'évoluer, en fonction de l'angle d'observation. Si cela ne pose généralement pas de problème, la procédure de compensation de l'égo-mouvement que nous utilisons repose sur l'exploitation de deux images différentes, avec des points de vue différents. Dès lors, l'image recalée I'_1 peut présenter des disparités d'intensité, comme illustré en figure 3.23. Ces disparités vont être à l'origine de contours fantômes qui vont perturber les résultats des algorithmes de flot optique.

Dans une grande majorité des cas, ce genre de problème peut être éliminé par l'intégration temporelle. Toutefois, il serait souhaitable de pouvoir corriger ce problème *a priori*. Les différentes approches que nous avons mises en œuvre (normalisation photométrique) se sont montrées infructueuses.

3.5 Conclusion

Au cours de chapitre, nous nous sommes intéressés à la détection d'objets dynamiques, depuis un capteur lui même en mouvement. Après être revenu sur l'état de l'art, nous avons constaté qu'une approche fondée sur la vision doit nécessairement faire appel à la stéréovision afin d'être pertinente.

Forts de ce constat, nous avons décrit notre système qui repose sur une compensation de l'égo-mouvement, puis sur une mesure du mouvement résiduel 3D des objets de l'image. Une segmentation des résultats permet de fournir une information haut niveau, tandis qu'une intégration temporelle assure leur cohérence temporelle. Cette intégration temporelle permet de supprimer de nombreux faux-positifs, tout en améliorant la détection d'objets lents ou lointains. Notre procédé permet une détection dense et rapide dans un très grand nombre de cas.

Nous nous sommes également attachés à une étude approfondie de ce système, à travers ses sources d'erreurs possibles, et ses possibilités d'améliorations. Nous avons ainsi pu constater que les performances finales du système vont dépendre fortement des différents paramètres du capteur de vision initialement choisi. Nous avons la ferme conviction que ces conclusions ne dépendent pas de notre approche ou de notre implémentation, mais sont les conséquences des mécanismes de formation des images. A ce titre, nous pensons que ces conclusions sont généralisables à tout système de vision ayant pour objet la détection d'objets dynamiques. En effet, il est apparu que les différents paramètres intrinsèques du système stéréo vont avoir un impact non-négligeable sur l'aptitude à détecter ou a différencier des objets.

Chapitre 4

Estimation Monoculaire du Mouvement : C-Vélocité Inverse

«When you start with a portrait and search for a pure form, a clear volume, through successive eliminations, you arrive inevitably at the egg. Likewise, starting with the egg and following the same process in reverse, one finishes with the portrait.»
Pablo Picasso

Sommaire

Alors que la première partie de ce travail se focalisait sur une approche faisant intervenir activement la stéréovision et aboutissant à une détection des objets dynamiques, il nous semble également important de nous pencher sur les approches monoculaires d'estimation du déplacement, et d'apporter notre contribution à ce segment de la vision. Ainsi, nous avons introduit le concept de C-Vélocité Inverse, immédiatement dérivé de la C-Vélocité Directe [BPZ09], une méthode permettant d'inférer la structure d'une scène à partir d'un capteur de vision monoculaire dont le mouvement, translationnel, est connu.

Après avoir rappelé les fondements de la C-Vélocité, nous proposerons et présenterons la C-vélocité Inverse, qui propose de résoudre le problème de l'estimation de l'égo-mouvement translationnel à partir d'images monoculaires. Après avoir présenté les résultats de cette approche, nous aborderons les perspectives que la C-Vélocité Inverse ouvre au développement de nouvelles applications de vision monoculaire.

4.1 C-Vélocité : Une Approche Cumulative Pour l'Estimation Structurelle

Si la reconnaissance de plans dans des images vidéos peut être une tâche relativement complexe, la détection de formes paramétrées dans une image binarisée est en revanche bien maitrisée, en particulier depuis la formalisation de la Transformée de Hough [DH72]. Dès lors, la reconnaissance de structures peut être grandement simplifiée par la définition d'un espace dans lequel ces structures sont représentées par des formes simples. Le concept de V-Disparité [LAT02] propose une solution à ce problème dans le cadre d'images stéréoscopiques, en l'espèce, deux espaces cumulatifs dans lesquels les plans sont imagés par des segments de droites. La C-Velocité [BZ09], quant à elle, propose une solution à ce même problème, mais dans le cas d'imagerie purement monoculaire. Plus spécifiquement, la C-Vélocité repose sur la définition d'un processus de vote cumulatif et d'espaces de votes associés, permettant de mettre en évidence certaines structures particulières de l'espace objet.

Au cours de cette section, nous allons revenir sur la C-Vélocité Directe[1]

[1]Par souci de clarté, nous ferons maintenant référence à la *C-Vélocité Directe* pour décrire le processus d'estimation de la structure de la scène et de *C-Vélocité Inverse* pour évoquer le processus d'estimation de l'égo-mouvement.

4.1.1 Objectifs - Hypothèses

Le premier objectif de la C-Vélocité est d'utiliser le flot optique afin d'identifier certains plans d'intérêt dans l'image. Ces plans sont choisis de façon à pouvoir représenter au mieux l'environnement urbain dans lequel le système est destiné à opérer. Ce sont respectivement :

- Le plan horizontal de la route
- Les plans verticaux, fronto-parallèles, statiques ou dynamiques, correspondants aux obstacles potentiels.
- Les plans verticaux latéraux, correspondants aux bâtiments.

La détection de ces différents plans peut, en effet, permettre de résoudre un certain nombre de problèmes. Tout d'abord, la localisation du plan de la route peut permettre, *via* une détermination de l'espace libre, de planifier et définir des trajectoires possibles. Ensuite, La détection des obstacles a bien évidemment des conséquences immédiates en termes de sécurité. Finalement, la localisation des bâtiments vient affiner la connaissance de cet environnement, et facilite la reconnaissance de l'espace navigable.

Plusieurs hypothèses sont prises de façon à mener ce travail, et la principale concerne le mouvement du capteur et la connaissance que nous en avons.

Ce mouvement est considéré comme étant purement translationnel, mais non-nécessairement axial. De plus, ce mouvement est connu, à un facteur d'échelle près. Cette connaissance du mouvement peut être résumée dans la position du Foyer d'Expansion (*Focus of Expansion, FoE*).

En effet, dans le cas d'un mouvement translationnel[2], le FoE est défini comme le point d'intersection entre le plan image et le vecteur déplacement du capteur. Ses coordonnées dans l'image sont :

$$\text{FoE} = \left| \begin{array}{l} x_{FoE} = f\frac{T_X}{T_Z} \\ y_{FoE} = f\frac{T_Y}{T_Z} \end{array} \right. \tag{4.1}$$

Ce point présente deux propriétés intéressantes :

- Tout d'abord, ses coordonnées regroupent toute l'information translationnelle que l'on peut obtenir avec un système purement monoculaire.

[2]Si la définition initiale du FoE est valable dans le cas d'un mouvement translationnel, il est possible d'étendre cette définition à des cas plus généraux, voire d'introduire une notion analogue à celle de FoE pour les rotations [FA97]

- Ensuite, si le mouvement est translationnel pur, il est le point de convergence des vecteurs déplacement de l'image.

Pour vérifier ce second point, il suffit de rappeler les équations de l'image de la composante translationnelle du mouvement (Équation 1.12), pour un point $M = \begin{vmatrix} X_M \\ Y_M \\ Z_M \end{vmatrix}$, imagé en $m = \begin{vmatrix} x_m \\ y_m \end{vmatrix}$:

$$\begin{vmatrix} \mu_{trans} \\ \nu_{trans} \end{vmatrix} = \begin{vmatrix} \dfrac{x_m T_Z - f T_X}{Z_M} \\ \dfrac{y_m T_Z - f T_Y}{Z_M} \end{vmatrix} \tag{4.2}$$

En introduisant les coordonnées du FoE, ces équations deviennent :

$$\begin{vmatrix} \mu_{trans} \\ \nu_{trans} \end{vmatrix} = \begin{vmatrix} \dfrac{T_Z}{Z_M}(x_m - x_{FoE}) \\ \dfrac{T_Z}{Z_M}(y_m - y_{FoE}) \end{vmatrix} \tag{4.3}$$

Il apparaît donc que le FoE est le point de convergence de toutes les droites engendrées par les vecteurs déplacement de l'image. Cette dernière propriété est illustrée par la figure 4.1. Le FoE extrait est bien situé à l'intersection des lignes de champ du flot optique.

Elle propose, en utilisant une information monoculaire de mouvement, et la connaissance de la position du FoE, d'identifier, dans l'image, les différents plans correspondant aux catégories route, bâtiments et obstacles.

4.1.2 Définition des $C - value$

Afin de pouvoir construire un espace de représentation nous permettant d'identifier aisément les différents plans que nous cherchons, nous allons tout d'abord nous pencher sur l'expression de l'image du mouvement d'un plan rigide.

Rappelons tout d'abord l'expression générale de l'image du mouvement, telle que développée dans l'annexe A :

$$\begin{cases} \mu = -\frac{xy}{f}\omega_X - \left(\frac{x^2}{f} + f\right)\omega_Y + y\omega_Z + \frac{xT_Z - fT_X}{Z} \\ \nu = \left(\frac{y^2}{f} + f\right)\omega_X + \frac{xy}{f}\omega_Y - x\omega_Z + \frac{yT_Z - fT_Z}{Z} \end{cases} \tag{4.4}$$

Considérons maintenant un plan quelconque, Π, d'équation :

FIGURE 4.1 – Image issue d'une séquence simulée, flot optique calculé par la méthode FOLKI [BC05] et Foyer d'Expansion extrait par C-Vélocité Inverse

$$\boldsymbol{n}^T \boldsymbol{P} = d \tag{4.5}$$

Où $\boldsymbol{P}$ est le vecteur position d'un point de Π, $\boldsymbol{n} = \begin{vmatrix} n_x \\ n_y \\ n_z \end{vmatrix}$ son vecteur normal unitaire et d sa distance à l'origine du repère.

En considérant ce plan, 4.4 devient :

$$\begin{cases} \mu = \frac{1}{fd}\left(a_1x^2 + a_2xy + a_3fx + a_4fy + a_5f^2\right) \\ \nu = \frac{1}{fd}\left(a_1xy + a_2y^2 + a_6fy + a_7fx + a_8f^2\right) \end{cases} \tag{4.6}$$

Où les coefficients a_n sont :

$$\begin{array}{ll} a_1 = -d\omega_y + T_Z n_x & a_2 = d\omega_x + T_Z n_y \\ a_3 = T_Z n_z - T_X n_x & a_4 = d\omega_z - T_X n_y \\ a_5 = -d\omega_y - T_X n_z & a_6 = T_Z n_z - T_Y n_y \\ a_7 = -d\omega_z - T_Y n_x & a_8 = d\omega_x - T_Y n_z \end{array} \tag{4.7}$$

Nous choisissons de ne considérer que les mouvements translationnels, dès lors : $\omega_X = \omega_Y = \omega_Z = 0$. De plus, nous ne recherchons que 3 types de plans par-

Type	$\boldsymbol{n}$	Distance à l'Origine
Bâtiment	$[1,0,0]$	d_b
Route	$[0,1,0]^T$	d_r
Obstacle	$[0,0,1]$	d_o

TABLE 4.1 – Hypothèses de plans considérées

ticuliers, représentatifs des scènes urbaines attendues. Ces plans et leurs équations peuvent être retrouvés en table 4.1.

Dans ces conditions, en simplifiant les coefficients a_n présentés en 4.7, il est possible d'exprimer l'image du mouvement pour chacune des hypothèses de plan. Pour des raisons de clarté, dans la suite de cet exposé, nous ne présenterons que le cas d'un plan correspondant à l'hypothèse "Bâtiment" :

$$\left\{ \begin{array}{l} \mu = \frac{1}{fd_b}\left(x^2 T_Z - fxT_X\right) \\ \nu = \frac{1}{fd_b}\left(xyT_Z - fxT_Y\right) \end{array} \right\} \tag{4.8}$$

En reconnaissant dans l'équation 4.8 les coordonnées du FoE, $\left| \begin{array}{l} x_{FoE} = f\frac{T_X}{T_Z} \\ y_{FoE} = f\frac{T_Y}{T_Z} \end{array} \right.$, nous pouvons écrire :

$$\left\{ \begin{array}{l} \mu = \frac{T_Z x}{fd}\left(x - x_{FoE}\right) \\ \nu = \frac{T_Z x}{fd}\left(y - y_{FoE}\right) \end{array} \right. \tag{4.9}$$

Afin d'une part de simplifier le problème en réduisant sa dimensionalité, et d'autre part de rendre le procédé insensible aux erreurs commises sur l'orientation des vecteurs du flot optique, nous ne considérons alors plus que la norme w du flot optique :

$$w = \sqrt{\mu^2 + \nu^2} = \frac{T_Z}{fd}\,|x|\,\|m - FoE\| \tag{4.10}$$

Il apparaît que la norme du flot optique, pour un point appartenant à un plan "bâtiment" est le produit de deux facteurs bien distincts. Le premier terme, qui est constant sur tout le plan considéré : $\frac{T_Z}{fd}$ n'est fonction que de la valeur absolue de la translation axiale (T_Z) et de la géométrie de la scène et du système ($\frac{1}{fd}$), alors que le second terme, variable, ne dépend que du positionnement relatif du FoE et du point considéré : $|x|\,\|m - FoE\|$. Ce second terme est défini comme étant la $C-value$ associée à l'hypothèse de plan "bâtiment". Il est immédiat de définir les autres $C-Values$, présentées en table 4.2.

Type de Plan	$C-value$ associée
Bâtiment	$c_b = \|x\| \, \|\|m - FoE\|\|$
Route	$c_r = \|y\| \, \|\|m - FoE\|\|$
Obstacle	$c_o = \|\|m - FoE\|\|$

TABLE 4.2 – Définition des $c-value$ spécifiques à chaque hypothèse de plan

L'équation 4.10 implique donc, avec la définition formelle de c_b :

$$w \propto c_b \tag{4.11}$$

Malgré l'existence d'une relation linéaire directe, il sera plus aisé, pour des raisons numériques et de représentation, de travailler avec $\sqrt{c_b}$, la relation 4.11 devient alors :

$$w \propto \sqrt{c_b}^2 \tag{4.12}$$

En effet, il n'est pas rare que les $C-values$ soient bien supérieures à la norme du flot optique (plusieurs ordres de grandeurs). Par abus de langage, et dans la mesure où il existe une bijection parfaite entre les deux représentations, nous parlerons indifféremment de la représentation linéaire ou parabolique des plans dans l'espace C-Vélocité.

4.1.3 Transformation C-Vélocité - Reconnaissance des Plans

Finalement, il nous faut définir un formalisme permettant d'exploiter cette relation entre $c-value$ et norme du flot optique. Pour cela, nous allons définir une transformation de l'espace image vers des espaces de représentation *ad-hoc*. Cette transformation sera par la suite appelée Transformation C-Vélocité.

Le formalisme choisi pour la transformation C-Vélocité est celui du vote cumulatif. En effet, ce type de procédure nous garantit une assez bonne robustesse vis-à-vis du bruit d'estimation du flot optique. En particulier la définition d'espaces robustes de représentations des différents plans de l'espace objet est immédiate. L'équation 4.10 nous apprend que, pour un plan "bâtiment" donné, la norme du déplacement image et la $c-value$ adéquate seront proportionnelles. Dès lors, un plan "bâtiment" sera représenté dans l'espace (w, c_b) par une droite. Les autres espaces peuvent être définis par analogie et sont présentés en table 4.3.

La transformation proprement dite est illustrée par l'algorithme 14. Pour chaque point de l'image, en connaissant la position du FoE, on calcule les trois $c-values$

Type de Plan	Espace de représentation
Bâtiment	(w, c_b)
Route	(w, c_r)
Obstacle	(w, c_o)

TABLE 4.3 – Définition des espaces de vote spécifiques à chaque hypothèse de plan

Entrées : $Flot\_Optique$, FoE, $Image$
Sorties : $Espace\_Batiments$, $Espace\_Route$, $Espace\_Obstacles$

début
  `Vider`($Espace\_Obstacles$)
  `Vider`($Espace\_Route$)
  `Vider`($Espace\_Batiments$)
  pour tous les points $m \in Image$ **faire**
    $w = \|Flot\_Optique[m]\|$
    $c_b = |x| \, \|m - FoE\|$
    $c_r = |y| \, \|m - FoE\|$
    $c_o = \|m - FoE\|$
    `Incrémente`($Espace\_Batiments[w, c_b]$)
    `Incrémente`($Espace\_Route[w, c_r]$)
    `Incrémente`($Espace\_Obstacles[w, c_o]$)
  fin
fin

Algorithme 4: Transformation C-Vélocité

correspondant aux trois hypothèse de plans, puis on va incrémenter les trois espaces de votes conformément à ces trois valeurs.

Vues les définitions des différentes $c-value$, deux points appartenant au même plan vont voir leurs contributions respectives alignées, et confirmées par les autres points de ce plan. D'autre part, vu l'orthogonalité des différentes $c-value$, un plan validant une hypothèse donnée ne pourra pas donner lieu à une réponse cohérente dans un des autres espaces de vote.

A partir de ces espaces de vote, des techniques classiques de traitement d'images peuvent être mises en œuvre afin d'identifier les meilleures réponses aux différentes transformations, et donc les ensembles de points les plus significativement coplanaires. De telles techniques peuvent être une transformée de Hough parabolique ou un clustering en K-moyennes [Llo82]. L'ensemble de cette méthode d'estimation de la structure de la scène est résumé en figure 4.2. Dans l'exemple présenté, un seul plan est présent, vérifiant l'hypothèse "route". Ainsi, sa représentation dans l'espace "Bâtiments" est diffuse, alors que sa représentation dans l'espace "route" fait apparaître une parabole bien définie, qui va facilement être extractible par une transformée de HOUGH parabolique. Des exemples de résultats de segmentation obtenus sur des séquences réelles sont disponibles en figure 4.3.

4.2 C-Vélocité Inverse : Une Approche Monoculaire d'Estimation de l'Égo-mouvement

La C-Vélocité Directe, telle que nous l'avons présentée, repose donc sur une connaissance fine de l'égo-mouvement du capteur (pas de rotation, translations connues à un facteur d'échelle près) et permet d'extraire la structure de la scène, à un facteur d'échelle près. Or, il nous semble que la connaissance de l'égo-mouvement est au moins aussi importante que la connaissance de la géométrie de la scène. Dans ce contexte, nous avons développé le concept de C-Vélocité Inverse. La C-Vélocité Inverse constitue notre contribution aux bases d'un système unifié d'estimation de la structure et du mouvement, ce système unifié étant un des axes directeurs d'un travail de thèse engagé à l'Institut d'Électronique Fondamentale[3].

L'idée à la base de la C-Vélocité Inverse est la suivante. Nous savons que, connaissant le FoE, les plans de l'espace objet seront imagés par la transforma-

[3]Cette thèse est menée par Qiong NIE, sous la direction d'Alain MÉRIGOT et la tutelle de Samia BOUCHAFA

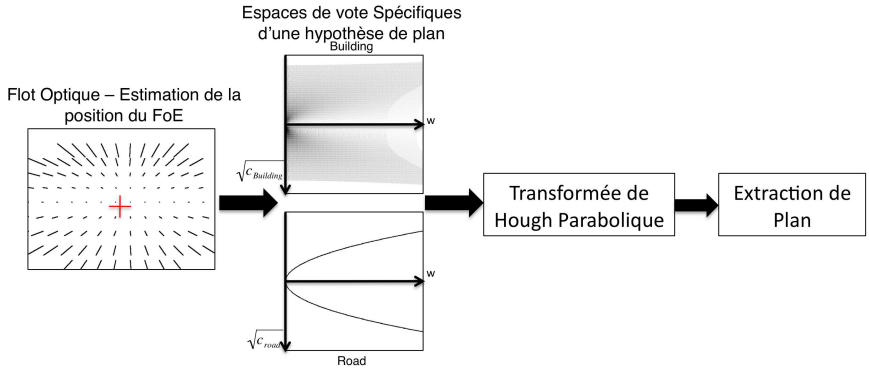


FIGURE 4.2 – Résumé du fonctionnement du processus de C-Vélocité Directe

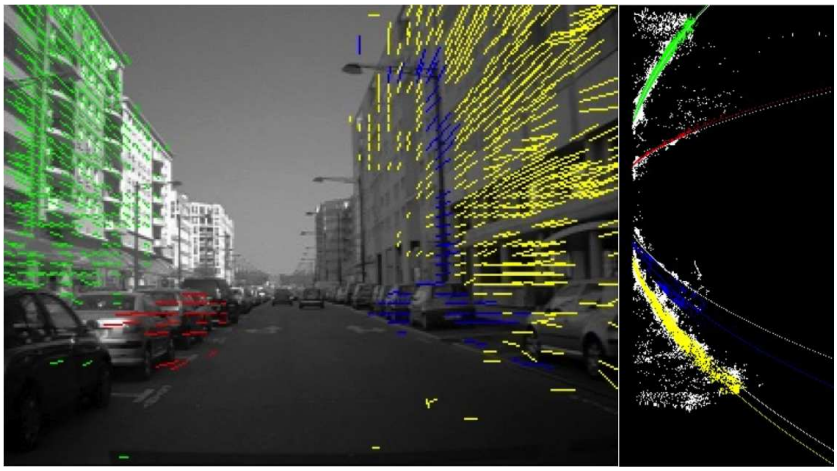


FIGURE 4.3 – Exemples de résultats de segmentation obtenus par C-Vélocité Directe, ainsi que les espaces de vote correspondants. Les paraboles parfaites sont calculées à partir des résultats de l'algorithme des K-moyennes et du pic de l'espace de HOUH.

tion C-Vélocité par des paraboles. Dès lors, supposons que nous puissions isoler différents plans objets. Les images de ces plans dans les espaces de votes adéquats seront d'autant plus paraboliques que le FoE utilisé sera proche du vrai FoE.

Deux problèmes sont soulevés par cette formulation approximative. Tout d'abord, il est nécessaire de pouvoir isoler les images de différents plans. Ensuite, et surtout, il est nécessaire de pouvoir quantifier l'aspect parabolique des représentations de ces plans, en exhibant une métrique reflétant, à travers la qualité des représentations C-Vélocité des plans, la distance séparant un FoE *supposé* du FoE réel.

Dans un premier temps, nous allons nous attacher à décrire une métrique possible, avant de proposer la solution que nous avons mise en œuvre afin de réaliser la nécessaire extraction de plans initiale.

4.2.1 Localisation du FoE

Nous supposons, dans cette partie, que l'image contient au moins un plan correspondant à une des trois hypothèses présentées en 4.1. Chacun de ces plans sera traité dans un espace de vote indépendant, afin d'éliminer le bruit lié au vote "croisé"[4]. Ce plan est noté Π, dans les équations suivantes, nous allons considérer qu'il s'agit d'un plan "bâtiment", la généralisation est toutefois immédiate.

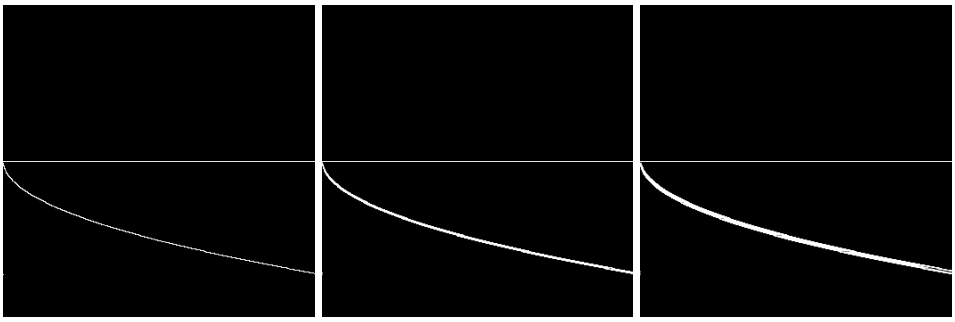
La détection de plans issue de la C-Vélocité Directe est dépendante de la précision de la localisation initiale du Foyer d'Expansion. En effet, c'est par rapport à ce FoE supposé que vont être calculées les différentes *c-values*. Il est assez intuitif que les espaces de votes de la C-Vélocité soient fortement impactés par la justesse de cette estimation initiale. Cela a pu être vérifé par BOUCHAFA *et al.* [BPZ09]. Il apparaît notamment que la dispersion au sein des espaces de votes augmente avec la distance séparant le FoE *supposé* du FoE *réel*. Cette dispersion est en particulier illustrée par la figure 4.4. Il apparaît clairement que si le modèle parabolique attendu est bien valide dans le cas où le FoE est correctement estimé, ce modèle n'est plus valide dans le cas contraire.

La dispersion de la représentation d'un plan dans l'espace de vote peut être formalisée comme étant :

$$\text{dispersion}_{(FoE)}(\Pi) = \sum_{m \in \Pi} \left(c_{observée} - c_{moyenne}(w(m))\right)^2 \tag{4.13}$$

Dans cette équation, $c_{observée}$ est la $c-value$ du point m, $c_{moyenne}(w(m))$ est la moyenne de $c-values$ des points pour lesquels la norme du flot optique est iden-

[4] Par exemple, un point appartenant à la route vote dans les espaces "bâtiments" et "obstacles", ce qui peut générer un bruit de fond.



(a) (b) (c)

FIGURE 4.4 – Visualisation de l'effet d'un décalage du FoE sur la C-Vélocité : (a) : Image d'un plan dans l'espace de vote adéquat avec une estimation correcte de la position du FoE ; (b) : Image d'un plan dans l'espace de vote adéquat avec une mauvaise estimation de la position du FoE (décalage de 10 pixels) ; Image d'un plan dans l'espace de vote adéquat avec une mauvaise estimation de la position du FoE (décalage 20 pixels)

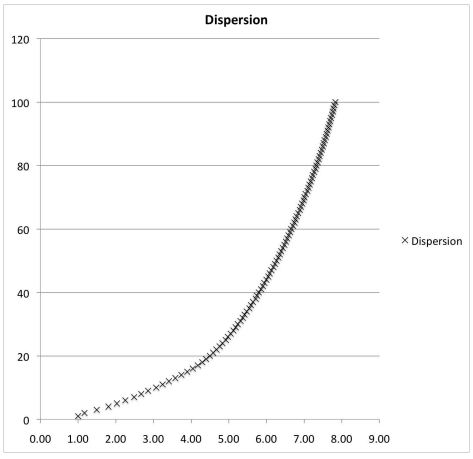


FIGURE 4.5 – Dispersion normalisée de la représentation C-Vélocité d'un plan dans l'espace de vote adéquat, en fonction du décalage du FoE

tique à celle de m. En d'autre termes, cette dispersion est la somme des largeurs élémentaires de la courbe représentant Π, ces largeurs étant considérées comme les carrés des écarts à la moyenne locale. Cette grandeur est représentée en figure 4.5. Cette courbe a été obtenue en calculant la dispersion pour différentes valeurs du décalage entre le FoE *réel* et le FoE *supposé*.

L'intuition à l'origine de la C-Vélocité Inverse est qu'il est possible d'utiliser cette dispersion de la représentation C-Vélocité d'un plan comme une métrique, monotone avec la distance séparant le FoE *supposé* du FoE *réel*. En particulier, la figure 4.5, renforce cette intuition première. Il apparaît en effet que la dispersion est convexe et peut permettre l'utilisation de techniques d'optimisation classiques afin de retrouver la position du FoE. Dans la suite de cette partie, nous allons nous attacher à renforcer mathématiquement cette intuition.

Afin de pouvoir exprimer la dispersion dans l'espace de vote, nous choisissons de considérer le lieu des points à w constant dans l'image. Ce lieu, noté $\mathcal{C}$, est en effet fixe, quelque soit l'hypothèse de FoE considérée. Nous exprimons ensuite les $c-values$ associées aux points de ce lieu. Les développements de ce calcul sont disponibles en Annexe C. Il vient que, en rappelant l'équation C.6 :

$$\begin{aligned} c^2_{batiment}\left(m \middle| \begin{matrix} x \\ y \end{matrix} \in \mathcal{C}\right) = K^2 - \Delta x x^2(2x - 2x_{FoE} - \Delta x) + \\ \Delta y^2 x^2 \mp \Delta y \sqrt{x^2K^2 - x^4(x - x_{FoE})} \end{aligned} \tag{4.14}$$

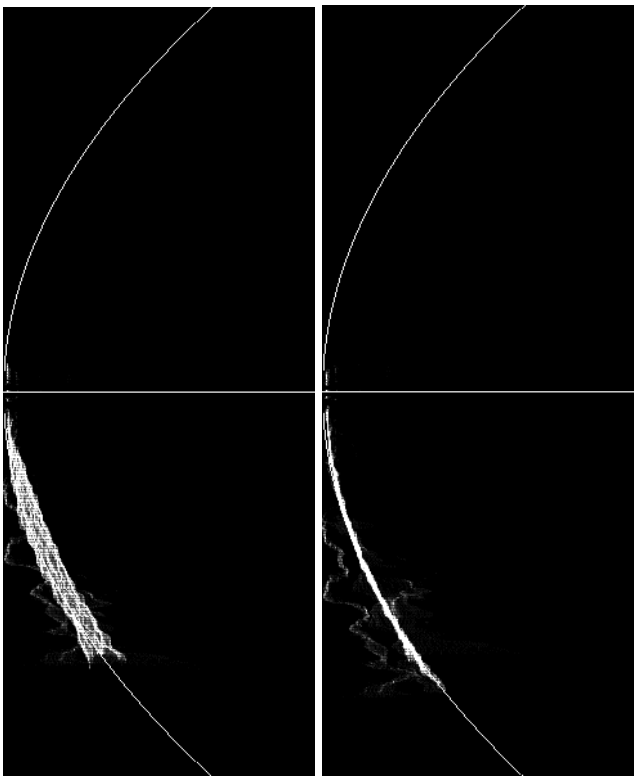
Dans cette équation, K^2 est une constante, qui correspond à la $c-value\ théorique$ à laquelle on aboutit en menant les calculs avec un FoE *supposé* coïncidant avec le FoE *réel*. Le décalage du FoE *supposé*, par rapport au FoE *réel* est exprimé par Δx et Δy, x_{FoE} correspond à la coordonnée du *vrai* FoE.

Dans un cas comme dans l'autre, il est apparent que les variations de $c_{batiment}$, donc la dispersion, vont directement être croissantes en fonctions des deux composantes du décalage du FoE, Δx et Δy. Dès lors, la mesure de la dispersion, exhibée en 4.13 peut être utilisée comme métrique reflétant la qualité d'un FoE supposé. Il peut être numériquement vérifié[5] que cette métrique est convexe, comme illustré en figure 4.5.

A partir de cette métrique, il est possible de formaliser le problème de localisation du FoE comme étant un problème de résolution aux moindres carrés :

$$\varepsilon^2 = \sum_{\Pi \in \mathcal{P}} \text{dispersion}_{(FoE)}(\Pi) \tag{4.15}$$

[5]L'expression de $c_{batiment}$ le long d'une $iso-w$ est non-intégrable.



(a) (b)

FIGURE 4.6 – Visualisation de l'espace de vote "bâtiment" avant (a) et après optimisation (b). La déformation de la représentation C-Vélocité, induite par la position du FoE avant optimisation (a) est visiblement compensée (b).

où $\mathcal{P}$ est l'ensemble des plans présents dans l'image.

Un tel problème peut être résolu par un schéma d'optimisation classique. Dans la suite de ce travail, nous avons utilisé une descente de gradient.

4.2.2 Estimation de la Structure de l'Espace Objet

Avant de présenter le système utilisé, nous allons nous attarder sur les exigences que nous avons vis-à-vis de ce système.

4.2.2.1 Cahier des Charges

Un système efficace d'estimation de la structure de la scène doit être en mesure de fournir un ensemble de labels, le type de plan auquel chaque label correspond,

ainsi qu'une carte représentant l'appartenance de chaque pixel de l'image à l'un ou l'autre label[6].

Il n'est pas nécessaire, et c'est un point important, de connaître les équations complètes des plans concernés. En particulier, la distance du plan à l'origine n'est pas nécessaire. Il n'est donc pas nécessaire de recourir à des méthodes fournissant une mesure absolue de la distance entre deux points (type stéréo-vision ou LIDAR), des méthodes fournissant une estimation relative de cette distance suffisant (méthodes monoculaire).

Toutefois, et pour des raisons pratiques, nous avons choisi d'employer une méthode utilisant la stéréo-vision. Cette méthode est décrite dans la suite.

4.2.2.2 Système Utilisé

L'extraction de plans utilisée dans ce travail repose donc sur l'utilisation d'une carte de disparité complète, fournie par le même algorithme que celui utilisé au chapitre 2.

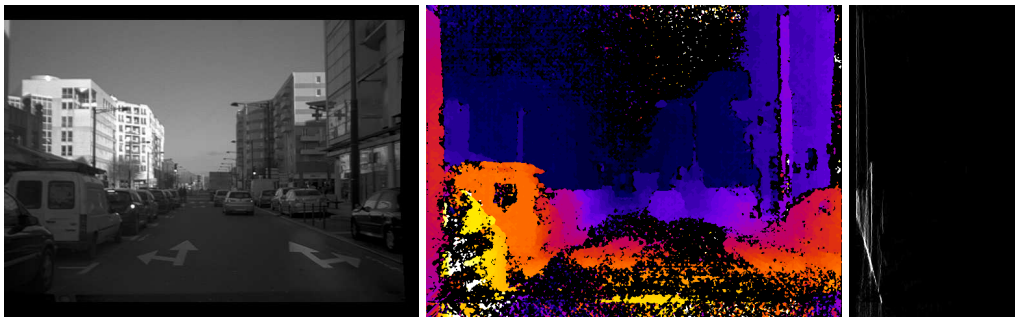
Extraction de la route - Pour l'extraction de la route, nous avons choisi d'utiliser une technique éprouvée : la V-Disparité [LAT02]. Cette technique repose sur l'accumulation des valeurs de disparité le long des lignes horizontales de la carte de disparité. Dans l'image résultante, la route est représentée par une droite, qui est alors facilement extractible par une transformée de HOUGH classique. Cette transformée est illustrée en figure 4.7. Les lignes verticales dans l'imagette 4.7(c) sont la représentation V-Disparité des plans fronto-parallèle, et la droite oblique est l'image de la route dans le plan V-Disparité.

Extraction des plans verticaux - En ce qui concerne l'extraction de plans verticaux, nous avons développé une transformée de HOUGH généralisée, très proche de celle présentée par IOCCHI *et al.* [IKB01]. Dans cette transformation, nous utilisons un espace de vote à deux dimensions :

$$\mathcal{V} = \{\rho, \theta\} \tag{4.16}$$

Dans cet espace ρ représente la distance du plan à l'origine, et θ représente l'angle entre l'axe optique et le plan. A partir de cet espace de vote, il nous est possible d'extraire de l'image les points appartenant à des plans verticaux, satisfaisant les hypothèses "bâtiments" et "obstacles".

[6] Il est également possible d'appartenir simultanément à plusieurs catégories, c'est notamment le cas des points situés à l'intersection de la route et d'un bâtiment.



(a) (b) (c)

FIGURE 4.7 – Illustration de la V-Disparité : (a) : Image Source (LoVE) ; (b) : carte de disparité ; (c) : Espace V-Disparité

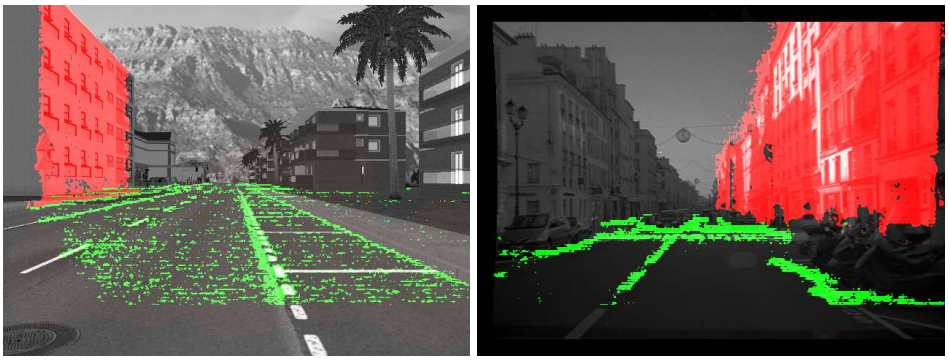
La combinaison de ces deux processus nous permet donc de pouvoir identifier dans l'image les différents plans d'intérêt, comme sur la figure 4.8. Un dernier point mérite que l'on s'y attarde, il s'agit de la qualité des résultats attendus. Une segmentation complète et dense de l'image n'est pas nécessaire. En revanche, il est important que tous les points qui partagent le même label appartiennent *effectivement* au même plan objet.

4.2.3 Validation - Images de Synthèse

Avant d'être utilisée sur des images réelles et pseudo-réalistes, cette méthode a été testée sur des flots optiques purement synthétiques, tels qu'en figure 4.9. Cela nous a permis, tout d'abord, d'obtenir une validation de notre approche, mais aussi de pouvoir étudier la sensibilité de cette approche à plusieurs facteurs :

- le bruit sur le flot optique ;
- l'influence des rotations, non prises en compte dans notre modèle ;
- le nombre de plans utilisés pour l'extraction.

Validation - Sur un flot optique synthétique, et en l'absence de bruit, une extraction exacte de la position du FoE peut être correctement réalisée et ce, quelque soit sa position. De plus, nous avons pu vérifier que, conformément à nos attentes, le système proposé est complètement insensible à un bruit sur l'orientation du flot optique, qui est l'erreur la plus commune parmi les techniques d'estimation actuelles [BSL+11].



(a) (b)

FIGURE 4.8 – Résultats du Système d'Estimation Structurelle : (a) : Images Simulées ; (b) : Images Réelles

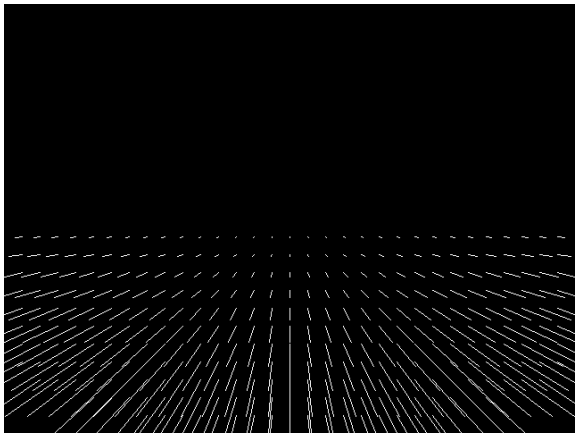


FIGURE 4.9 – Flot Synthétique

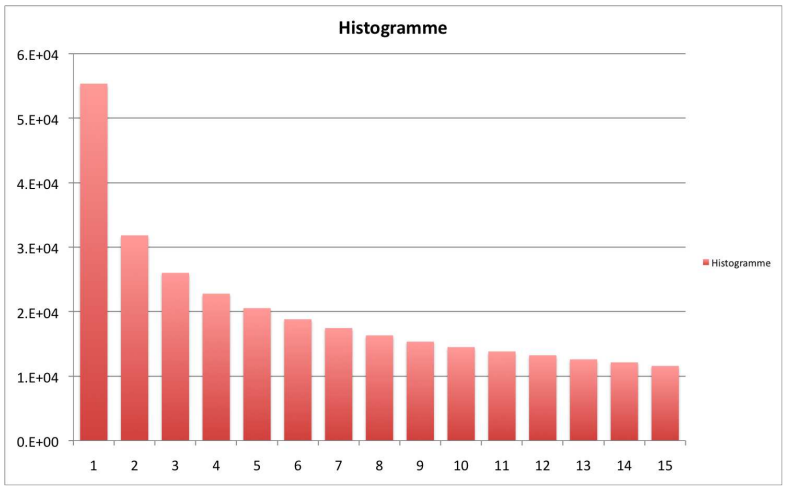


FIGURE 4.10 – Histogramme des normes du flot synthétique

Influence du bruit - Afin de pouvoir étudier l'impact du bruit sur l'extraction du FoE par notre méthode, nous avons ajouté au flot synthétique un bruit additif, normal, centré sur 0. La figure 4.10 représente l'histogramme des normes du flot synthétique, afin de pouvoir constater que le bruit introduit est plus que significatif.

A l'issue de ce premier test, il apparaît que le processus de localisation du FoE est très robuste au bruit, en effet, vu la figure 4.10, la majorité des vecteurs générés ont une norme inférieure à 5 pixels. Toutefois, le bruit additif considéré n'a pas d'impact sur la précision de la localisation jusqu'à un écart-type de 3 pixels. En effet, le schéma cumulatif utilisé est reconnu pour sa robustesse vis-à-vis du bruit.

Influence des Rotations - Dans cette étude, et, plus généralement, dans le formalisme de la C-Vélocité, nous ne considérons pas les rotations, et cette hypothèse peut paraître limitante. Toutefois, la littérature est riche de méthodes permettant de s'affranchir de ces rotations. On peut notamment citer des approches reposant sur le suivi de points et l'estimation d'un tenseur trifocal [RASP96], ou encore fondées sur les propriétés du flot optique, considéré comme un champ vectoriel [HS93].

Malgré cette possibilité que nous avons de supprimer les rotations des images sources, il nous semble pertinent d'observer l'impact de faibles taux de rotations sur l'extraction du Foyer d'Expansion.

La figure 4.12 illustre cet impact sur les performances de notre méthode. Il apparaît, que pour des taux de rotation important (au dessus de 3deg.$frame^{-1}$), une

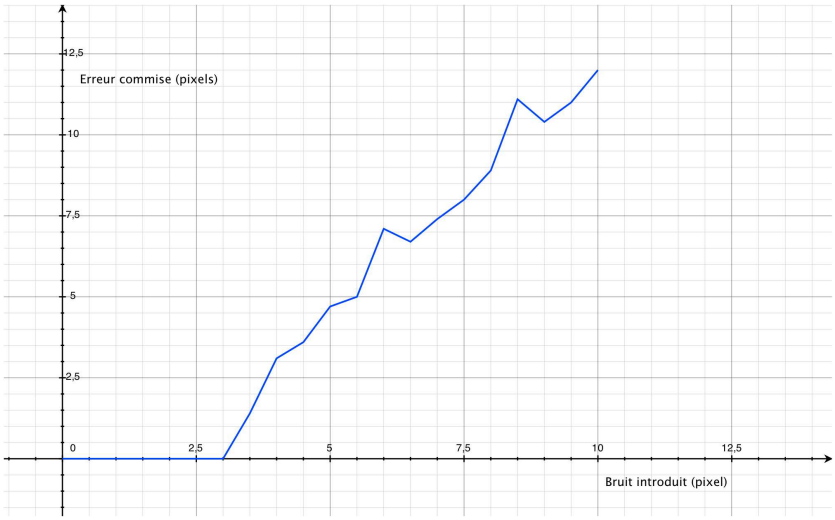


FIGURE 4.11 – Impact du bruit sur la localisation du FoE

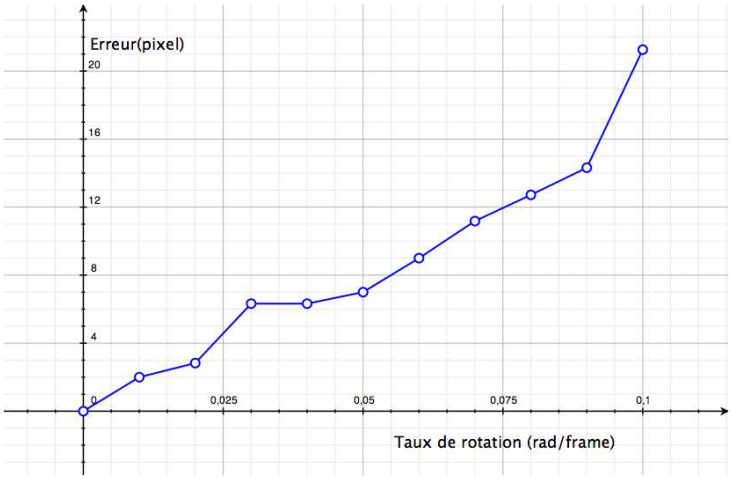


FIGURE 4.12 – Influence des rotations (ici du lacet) sur la précision de l'extraction du FoE

erreur assez importante est introduite. Pour des rotations plus faibles, de l'ordre de grandeur de celles qui peuvent apparaître lors d'une trajectoire en "ligne droite", cette erreur est beaucoup plus contenue.

Influence du nombre de plans utilisés - Au cours de ce travail, nous n'avons pas noté d'influence notable, en termes de résultats, du nombre de plans utilisés. Il est à noter que n'avons jamais utilisé de plan de type "obstacles" pour la localisation du FoE. En effet, ceux-ci ne sont pas nécessairement présents dans l'image. De plus, ils sont généralement plus petits que leurs homologues "route" et "bâtiment".

D'autre part, il est important de noter que chaque plan supplémentaire pris en compte augmente d'autant le temps de calcul nécessaire. En conséquence, nous choisissons de ne travailler qu'avec un plan répondant à chaque hypothèse. Si cela tombe sous le sens en ce qui concerne l'hypothèse "route"[7], concernant l'hypothèse "bâtiment", nous n'utiliserons donc que le plan ayant la meilleure représentation dans l'espace de HOUGH généralisé utilisé pour la détection.

4.3 Résultats

Cette méthode a été testée, dans un premier temps sur les images pseudo-réalistes générées par le simulateur SiVIC (cf. D.1), avant d'être mise en œuvre sur les images issues des bases de données réalisées pour le projet LoVE (cf. D.2.1). Le flot optique utilisé a été calculé en utilisant la méthode FOLKI [BC05].

4.3.1 Images Pseudo-Réalistes

Le simulateur a été utilisé afin de générer une séquence de 250 paires stéréo, prenant place dans un environnement urbain au trafic modéré. Au cours de cette séquence, le mouvement du véhicule est majoritairement translationnel. Un exemple d'extraction du FoE peut être vu en figure 4.1. Il est intéressant de noter que le FoE extrait repose sur la ligne d'horizon, ce qui est cohérent avec le mouvement connu du mobile. Plus spécifiquement, pour cette image, l'erreur commise entre le FoE extrait et le vrai FoE, recalculé à partir des composantes du mouvement, était de $2,2$ pixels.

Sur la totalité de cette séquence, l'erreur moyenne commise était de $5,2$ pixels, avec une erreur maximale de $15,6$ pixels. Cette erreur peut paraître importante,

[7] Il est toutefois possible, dans certains cas d'obtenir plusieurs plans répondant à cette hypothèse, l'un correspondant à la route proprement dite, et l'autre aux trottoirs.

toutefois, il nous semble pertinent de relativiser. En effet, s'il est naturel d'exprimer l'erreur commise sur la localisation du FoE en pixel, il est important d'insister sur le fait que le FoE n'est pas localisé au même titre qu'un point d'intérêt, mais qu'il s'agit avant tout d'une mesure des différentes translations d'un véhicule.

En particulier, si nous considérons le système optique simulé (qui présente une longueur focale de $f = 10mm$ et une taille de pixel de $10\mu m$), une erreur de 1 pixel sur la position du FoE se traduit par une erreur de 10^{-3} sur le rapport :

$$\frac{\|\boldsymbol{T}\|}{T_Z} \tag{4.17}$$

où $\boldsymbol{T} = \begin{vmatrix} T_X \\ T_Y \end{vmatrix}$. Cela représente, dans le cas d'un véhicule se déplaçant à $50km.h^{-1}$, une translation latérale de $0,6mm$. Dans notre cas, l'erreur commise sur la localisation du FoE correspond donc à une erreur d'estimation du rapport $\frac{\|\boldsymbol{T}\|}{T_Z}$ de 0.8%. A titre de comparaison, nous avons également utilisé une méthode de localisation par vote cumulatif, telle qu'illustrée dans [SJBK08]. Cette dernière méthode conduit à une erreur moyenne de localisation du FoE de $10,6$ pixels, soit une erreur de $1,6\%$ sur l'estimation du rapport $\frac{\|\boldsymbol{T}\|}{T_Z}$.

4.3.2 Images Réelles

Malgré les moyens expérimentaux utilisés, aussi bien par nous-même, que par les différentes équipes qui nous ont autorisé l'accès à leurs bases de données, il n'a pas été possible d'utiliser de capteurs suffisamment précis pour pouvoir être utilisé comme vérité terrain, et donc comme base de comparaison. Au mieux, en utilisant les données de l'Université de Karlsruhe nous pouvons espérer une estimation de la position du FoE précise à environ 13 pixels près, ce qui n'est pas suffisant pour pouvoir constituer une base solide de comparaison.

Nous sommes donc contraints de nous rabattre sur des considérations qualitatives afin d'évaluer la qualité de notre extraction du FoE. Ainsi, les figures 4.13 et 4.14 présentent le Foyer d'Expansion extrait par notre méthode pour plusieurs images réelles. Dans tous les cas, ce FoE correspond à la connaissance que nous avons du mouvement approximatif de l'égo-véhicule. De plus, il peut être intéressant de considérer les lignes de champ du flot optique comme des indicateurs visuels de la qualité du FoE extrait.

Un autre indicateur qualitatif de la qualité du FoE extrait peut être la comparaison entre l'espace de vote initial (calculé avec un FoE *supposé* au centre de l'image) et l'espace de vote final (calculé avec le FoE *extrait*). Une telle comparaison peut être trouvée en figure 4.6 et en figure 4.15. Dans les deux cas, il

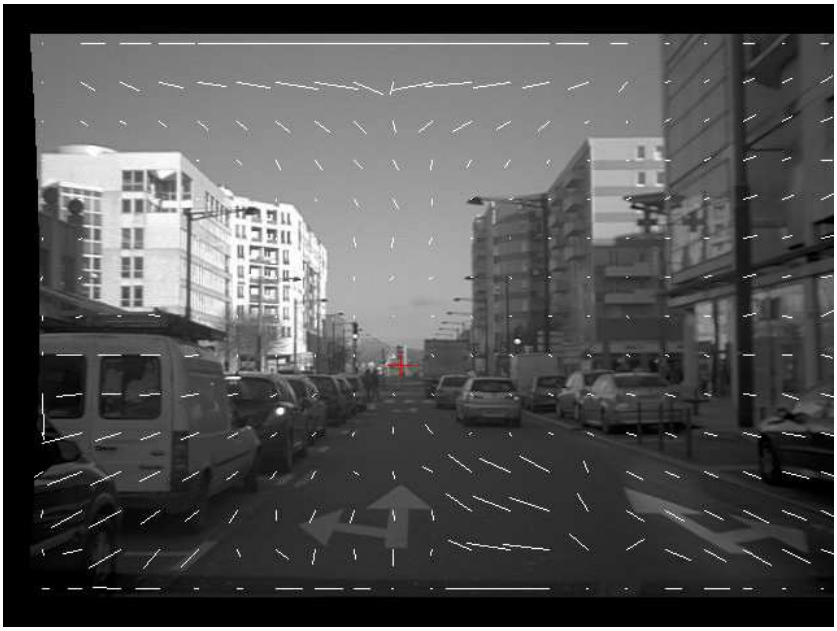
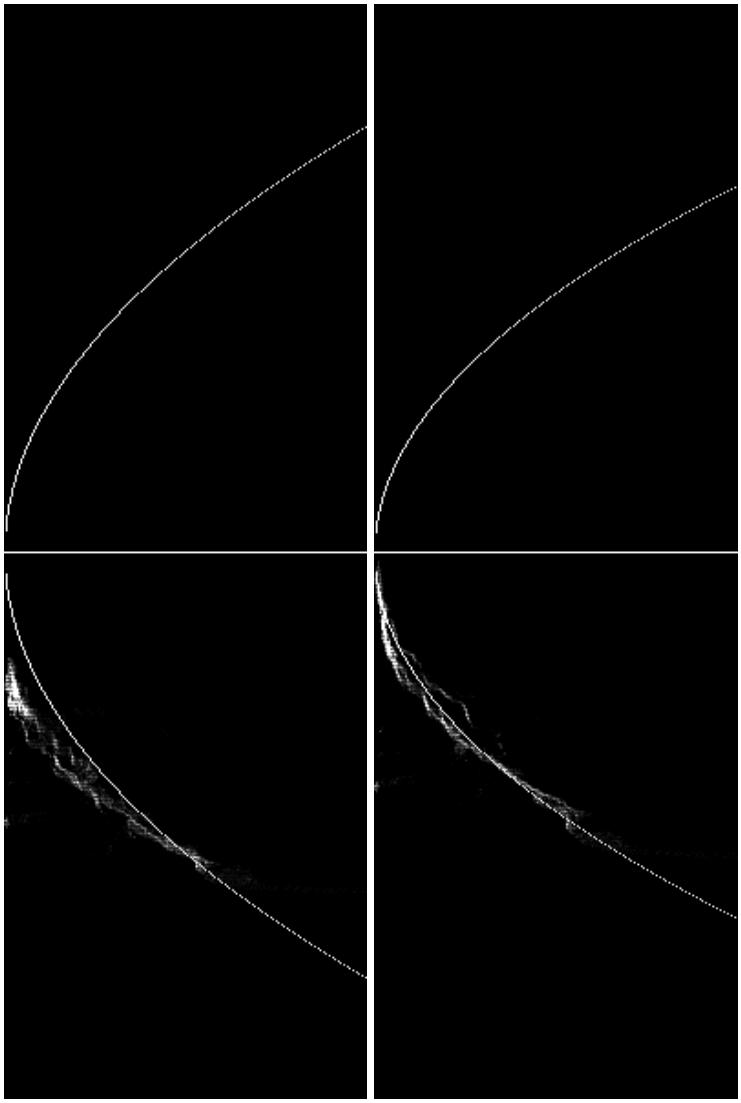


FIGURE 4.13 – Exemple de résultats obtenus sur image réelle

FIGURE 4.14 – Exemple de résultats obtenus sur image réelle

apparaît que, à l'issue de l'optimisation réalisée, la représentation C-Vélocité des plans observés est bien plus conforme au modèle parabolique attendu, signe que la position estimée du FoE est plus proche du FoE réel après optimisation qu'avant.



(a) (b)

FIGURE 4.15 – Comparaison des Espaces de vote avant (a) et après (b) recherche du FoE. Les paraboles parfaites correspondent à une extraction par transformée de HOUGH, par construction, les paraboles sont contraintes et doivent passer par l'origine.

4.4 Conclusion

Au cours de cette section, nous avons présenté une nouvelle méthode permettant de localiser le Foyer d'Expansion dans des séquences d'images monoculaires. Contrairement à l'immense majorité des méthodes existantes, la C-Vélocité Inverse n'utilise qu'une partie de l'information contenue dans le flot optique, en l'occurrence, la norme relative. Ce faisant, notre méthode est insensible au bruit sur l'orientation des vecteurs et au biais sur l'estimation de ces vecteurs. Toutefois, et contrairement à d'autres méthodes, la C-Vélocité Inverse repose sur une estimation préalable de la structure de la scène, à un facteur d'échelle près. Les résultats obtenus sont encourageants. Malgré le recours à un matériel de mesure de pointe, nous n'avons pu obtenir de mesures suffisamment précises en situation réelle, limitant, *de facto*, l'évaluation quantitative de notre méthode à des séquences simulées. Cependant, nous pensons que le niveau de réalisme obtenu par le simulateur SiVIC est suffisant pour pouvoir étendre nos conclusions aux images réelles.

De plus, nous pensons, et cela fait actuellement l'objet d'un travail de thèse, que C-Vélocité Directe et Inverse peuvent être utilisées conjointement, afin d'obtenir une estimation simultanée de la structure de la scène et de l'égo-mouvement translationnel. Ce nouveau cadre pourrait permettre, à terme, d'ouvrir la voie vers de nouveaux systèmes monoculaires, efficaces et compacts.

Conclusion et Perspectives

«The most exciting phrase to hear in science,
the one that heralds the most discoveries,
is not "Eureka!" but "That's funny..."»
Isaac Asimov

Conclusion

Au cours de ce travail de thèse, nous avons apporté plusieurs contributions à la vision artificielle en générale, et plus particulièrement au domaine des véhicules "intelligents". Nous avons ainsi décrit 3 systèmes distincts. Si deux d'entre eux ont été explicitement conçus afin d'être complémentaires, il n'en reste pas moins que ces trois systèmes sont trois entités distinctes, qui nous ont permis de tirer différentes conclusions.

Tout d'abord, nous avons tenu à apporter notre contribution à l'odométrie visuelle. Contrairement à la totalité des approches existantes, nous avons montré qu'il est possible d'utiliser un formalisme entièrement linéaire afin d'estimer les 6 composantes du mouvement. Cela nous permet d'éviter les inconvénients généralement associés aux problèmes non-linéaires : instabilité et lenteur. Malgré l'apparente simplicité de notre formalisme, nous ne sacrifions pas la précision et les performances. Nous avons ainsi pu exhiber une précision du niveau d'une bonne, voire très bonne, centrale inertielle. Si l'odométrie visuelle peut se montrer distancée dans quelques cas, lorsque les caméras sont éblouies par exemple, il n'en reste pas moins vrai qu'une solution à base de vision présente d'excellentes performances, à un coût très modique. Non contents de pouvoir tester notre

système extensivement sur différents systèmes, nous avons également cherché, à travers des collaborations avec le LIVIC et avec d'autres chercheurs de l'IEF, à monter différents projets de localisation mettant en œuvre plusieurs capteurs. Il est ressorti de ces différents projets que, une fois encore, notre implémentation de l'odométrie visuelle pouvait tirer son épingle du jeu, et se montrer compétitive vis-à-vis, notamment, de centrales inertielles haut de gamme.

Par la suite, nous nous sommes concentrés sur ce qui était l'objectif principal de cette thèse, à savoir la détection d'objets dynamiques, à partir d'un capteur de stéréovision. Le système que nous avons mis en place repose sur une compensation de l'égo-mouvement qui peut être mesuré à partir d'un capteur tierce (pourvu qu'il soit suffisamment bien caractérisé) ou de notre méthode d'odométrie visuelle. Nous avons tenu à mettre l'accent sur l'évaluation des facteurs d'amélioration des performances. Ainsi, nous avons pu établir différentes recommandations à l'attention de futurs concepteurs de système de détection. Parmi ces recommandations, nous pouvons citer le choix de la base du système stéréo, mais aussi de la taille relative de la focale et des matrices des capteurs, ou encore l'influence de la cadence de prise des images sur les performances.

Finalement, nous avons apporté notre contribution à un segment légèrement différent de la vision artificielle : celui des systèmes monoculaires. Le formalisme de la C-Vélocité Directe a fait ses preuves en ce qui concerne la cartographie de l'espace et la détection de plans particuliers dans l'image. Toutefois, il nous semble que la connaissance, même partielle, de l'égo-mouvement est au moins aussi importante que la localisation des plans de l'image. C'est dans cet état d'esprit que nous avons défini et conçu la C-Vélocité Inverse. Si notre procédure de localisation du FoE a pu être testée avec succès sur des images issues d'un simulateur de haut-niveau, en revanche, nous n'avons pas pu émettre de conclusion quantitative quant à ses performances sur des images réelles. En effet, il nous est apparu que des centrales inertielles de haut niveau ne pouvaient pas fournir des mesures d'une précision suffisante pour pouvoir être utilisées comme vérité terrain dans ce cas précis. Nous avons donc dû nous contenter d'appréciations qualitatives du positionnement du FoE sur des images réelles.

Perspectives

Malgré des résultats encourageants, notre travail a mis au jour un certain nombre de pistes qu'il serait intéressant de poursuivre.

Tout d'abord, concernant la C-Vélocité Inverse, un des points faibles de la méthode que nous avons décrite est l'utilisation de la stéréovision. En effet, même si la méthode que nous avons décrite n'utilise pas l'information de profondeur, nous avons exploité, par commodité, une carte de disparité afin d'extraire la structure de la scène. Il serait intéressant, voire indispensable de se dispenser de cette information supplémentaire afin de mettre en place une collaboration entre C-Vélocité Directe et Inverse, par exemple, au sein d'un processus itératif. Ainsi, nous disposerions d'une méthode monoculaire intégrée pour l'estimation jointe de la structure de la scène et du FoE, donc de l'égo-mouvement.

De la même manière, si la C-Vélocité s'est montrée robuste à des taux de rotations relativement faibles cela n'est, en revanche, pas le cas pour des orientations de plans ne respectant pas les hypothèses initiales. Il nous semble que le cadre de la C-Vélocité pourrait être étendu à des orientations de plans quelconque, mais également à des mouvements non-contraints. A ce titre, nous pensons que la thèse de Cornelia FERMÜLLER pourrait représenter une base intéressant à une réflexion de fond sur la structure du champ de flot optique [FA97]. En particulier, la réflexion qu'elle a initiée, sur la définition d'un "Foyer des Rotations", notion analogue au FoE, nous semble prometteuse et pourrait ouvrir la voie à une meilleure prise en compte des rotations.

En ce qui concerne le système intégré d'estimation de l'égo-mouvement et de localisation des objets dynamiques, nous pensons qu'il y a, ici encore, plusieurs pistes de réflexion intéressantes.

En premier lieu, la compensation de l'égo-mouvement est apparue, ce qui a confirmé l'intuition que nous pouvions avoir, comme une étape fondamentale du travail. Toutefois, un point en particulier de notre approche n'est pas complètement satisfaisant. Il s'agit du *remplissage* que l'on opère afin de combler les zones de l'image impossibles à recaler à cause d'une information de disparité non-disponible. Si cette procédure nous a permis d'obtenir des résultats satisfaisants, elle est également à l'origine de plusieurs défaillances, notamment lorsque les objets observés ne sont pas lambertiens (ou que les conditions d'illuminations changent). Nous pensons que l'amélioration de ce point particulier, par exemple en proposant une méthode de *remplissage* qui prendrait en compte l'illumination locale, pourrait être intéressante.

Ensuite, si la méthode imposant la cohérence temporelle des résultats que nous avons implémenté s'est montré efficace la plupart du temps, certains cas demeurent problématiques. En particulier, notre méthode est efficace tant que la

direction des objets mobiles est conservée, c'est à dire tant que ces objets peuvent être considérés comme étant rigides. C'est le cas des véhicules bien sûr, mais aussi des piétons partiellement dissimulés. Dès que les objets deviennent déformables, les résultats sont de moins bonne qualité, en particulier, les membres des piétons, de part la nature indépendante de leur mouvement, sont moins bien imagés. Il nous semble donc intéressant de poursuivre le travail par une amélioration de cette prise en charge, de façon à améliorer la détection des objets déformables.

Nous sommes convainus que le recours à la vision seule nous a permis d'aboutir à une meilleure compréhension des mécanismes mis en jue. Toutefois, une intégration efficace, notamment en termes de fiabilité n'est possible qu'à condition d'utiliser également d'autres capteurs, afin de pouvoir, notamment, compenser les éventuelles défaillances de l'un ou l'autre. Nous pensons que ce travail sur la fusion multi-capteurs est de première importance. Si nous avons tenté, à travers différentes collaborations, de poser des fondations dans cette direction, nous sommes conscients qu'un travail plus ambitieux est nécessaire et cette direction devrait être privilégiée dans les années à venir. De plus, il nous semble que la fusion entre différentes sources d'information ne devrait pas se limiter à la mesure de l'égo-mouvement, mais également être étendue à la détection d'objets dynamiques, par exemple en exploitant l'information issue de LIDAR ou d'autres procédés utilisant la vision, tel que la reconnaissance.

Annexe A

Expression du mouvement rigide

A.1 Image 2D du mouvement 3D

La présente annexe développe les calculs qui mènent à l'expression de l'image du mouvement rigide sur la rétine d'un capteur de vision traditionnel. Nous nous plaçons dans le cadre des hypothèses et modélisations présentées en 1.3

Soit un point $M = \left| \begin{array}{c} X_M \\ Y_M \\ Z_M \end{array} \right|_{\mathcal{R}_a} = \left| \begin{array}{c} X_M \\ Y_M \\ Z_M \end{array} \right|_{\mathcal{R}_r}$ de l'espace objet. Nous considérons que ce point est correctement imagé par le capteur, et que son image est le point :

$$m = \left| \begin{array}{c} x_m \\ y_m \end{array} \right. = \left| \begin{array}{c} f\frac{X_M}{Z_M} \\ f\frac{Y_M}{Z_M} \end{array} \right. \tag{A.1}$$

Entre les instants t_0 et t_1, le capteur est animé d'un mouvement arbitraire :

$$\boldsymbol{T} = \left| \begin{array}{c} T_X \\ T_Y \\ T_Z \end{array} \right. \quad ; \quad \boldsymbol{\Omega} = \left| \begin{array}{c} \omega_X \\ \omega_Y \\ \omega_Z \end{array} \right. \tag{A.2}$$

Dès lors, en t_1, les coordonnées du point M peuvent s'exprimer :

$$M = \left| \begin{array}{l} X'_M = X_M - \omega_Y Z_M + \omega_Z Y_M - T_X \\ Y'_M = Y_M + \omega_X Z_M - \omega_Z X_M - T_Y \\ Z'_M = Z_M + \omega_Y X_M - \omega_X Y_M - T_Z \end{array} \right|_{\mathcal{R}_r} \tag{A.3}$$

Et son mouvement relatif peut s'écrire :

$$\left|\begin{array}{l} \Delta X_M = X'_M - X_M = -\omega_Y Z_M + \omega_Z Y_M - T_X \\ \Delta Y_M = Y'_M - Y_M = \omega_X Z_M - \omega_Z X_M - T_Y \\ \Delta Z_M = Z'_M - Z_M = \omega_Y X_M - \omega_X Y_M - T_Z \end{array}\right. \tag{A.4}$$

A.1.1 Calcul Exact

En reprenant et en inversant les équations de projection 1.4 :

$$X_M = \frac{x_m Z_M}{f} \quad Y_M = \frac{y_m Z_M}{f} \tag{A.5}$$

Dès lors :

$$\left|\begin{array}{l} \Delta X_M = \frac{x'_m Z'_M}{f} - \frac{x_m Z_M}{f} = -\omega_Y Z_M + \omega_Z Y_M - T_X \\ \Delta Y_M = \frac{y'_m Z'_m}{f} - \frac{y_m Z_M}{f} = \omega_X Z_M - \omega_Z X_M - T_Y \end{array}\right. \tag{A.6}$$

Soit :

$$\left|\begin{array}{l} x'_m Z'_M - x_m Z_M = f(-\omega_Y Z_M + \omega_Z \frac{y_m Z_M}{f} - T_X) = -f\omega_Y Z_M + \omega_Z y_m Z_M - f T_X \\ y'_m Z'_M - y_m Z_M = f(\omega_X Z_M - \omega_Z \frac{x_m Z_M}{f} - T_Y) = f\omega_X Z_M - \omega_Z x_m Z_M - f T_Y \end{array}\right. \tag{A.7}$$

D'où

$$\left|\begin{array}{l} x'_m \frac{Z'_M}{Z_M} = x_m - f\omega_Y + y_m \omega_Z - \frac{f T_X}{Z_M} \\ y'_m \frac{Z'_M}{Z_M} = y_m + f\omega_X - x_m \omega_Z - \frac{f T_Y}{Z_M} \end{array}\right. \tag{A.8}$$

Or

$$\Delta Z = X_M \omega_Y - Y_M \omega_X - T_Z \tag{A.9}$$

Soit, en réutilisant A.5

$$Z'_M = \omega_Y \frac{x_m Z_M}{f} - \omega_X \frac{y_m Z_M}{f} - T_Z + Z_M \tag{A.10}$$

D'où :

$$\left|\begin{array}{l} x'_m \left(\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z}{Z_M} + 1 \right) = x_m - f\omega_Y + y_m \omega_Z - \frac{f T_X}{Z_M} \\ y'_m \left(\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z}{Z_M} + 1 \right) = y_m + f\omega_X - x_m \omega_Z - \frac{f T_Y}{Z_M} \end{array}\right. \tag{A.11}$$

Finalement :

$$\left| \begin{array}{l} x'_m = \frac{x_m + y_m\omega_Z - f\omega_Y - f\frac{T_X}{Z_M}}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z}{Z_M} + 1} \\ y'_m = \frac{y_m - x_m\omega_Z + f\omega_X - f\frac{T_Y}{Z_M}}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z}{Z_M} + 1} \end{array} \right. \tag{A.12}$$

Ce résultat, bien qu'exact, est cependant relativement difficile à manipuler.

A.1.2 Approximation au premier ordre

C'est pourquoi il est fréquent de considérer que l'image du mouvement est la dérivée de la position de l'image :

$$\left\{ \begin{array}{l} \mu = \dot{x}_m = \frac{\partial}{\partial t}\left(f\frac{X_M}{Z_M}\right) = f\left(\frac{\dot{X}_M}{Z_M} - X_M\frac{\dot{Z}_M}{Z_M^2}\right) \\ \nu = \dot{y}_m = \frac{\partial}{\partial t}\left(f\frac{Y_M}{Z_M}\right) = f\left(\frac{\dot{Y}_M}{Z_M} - Y_M\frac{\dot{Z}_M}{Z_M^2}\right) \end{array} \right. \tag{A.13}$$

Dès lors, en considérant que l'intervalle ∂t est suffisamment court pour approximer les dérivées au premier ordre, en combinant A.4 et A.13, il vient que :

$$\left\{ \begin{array}{l} \mu = f(\frac{X_M Y_M}{{Z_M}^2}\omega_X - (1 + \frac{{X_M}^2}{{Z_M}^2})\omega_Y + \frac{Y_M}{Z_M}\omega_Z - \frac{T_X}{Z_M} + \frac{X_M T_Z}{{Z_M}^2} \\ \nu = f((1 + \frac{{Y_M}^2}{{Z_M}^2})\omega_X - \frac{X_M Y_M}{{Z_M}^2}\omega_Y - \frac{X_M}{Z_M}\omega_Z - \frac{T_Y}{Z_M} + \frac{Y_M T_Z}{{Z_M}^2}) \end{array} \right. \tag{A.14}$$

En reprenant la formulation de l'image d'un points objet A.1, nous arrivons finalement aux expressions suivantes :

$$\left\{ \begin{array}{l} \mu = \frac{x_m y_m}{f}\omega_X - (f + \frac{{x_m}^2}{f})\omega_Y + y_m\omega_Z - \frac{f T_X}{Z_M} + \frac{x_m T_Z}{Z_M} \\ \nu = (f + \frac{{y_m}^2}{f})\omega_X - \frac{x_m y_m}{f}\omega_Y - x_m\omega_Z - \frac{f T_Y}{Z_M} + \frac{y_m T_Z}{Z_M} \end{array} \right. \tag{A.15}$$

Cette formulation est en particulier adaptée aux approches fondées sur le flot optique, en effet l'approximation des petits déplacements, et l'approximation au premier ordre réalisée ici sont généralement compatibles. Toutefois, il est important de noter que cette approximation revient à considérer que la profondeur des objets observés varie peu entre deux instants, ce qui peut s'avérer faux dans de nombreux cas.

A.2 Image 3D du mouvement 3D

Dans cette section, nous revenons sur l'expression de l'image du mouvement 3D par un capteur de stéréovision idéal, tel que présenté en 1.4.

Il peut apparaître intuitif de chercher à exprimer toutes les grandeurs dans le domaine objet, de façon à reconstruire immédiatement le mouvement des objets physiques observés. Il a cependant été démontré, dans [DD02], que ce formalisme, en faisant apparaître des termes d'incertitudes anisotropes, pouvait conduire à des solutions sous optimales pour un grand nombre d'applications. Il est donc préférable d'exprimer tous les développements dans le domaine image.

A cette fin, nous considérons un point objet $M \left| \begin{matrix} X_M \\ Y_M \\ Z_M \end{matrix} \right.$ correctement imagé par le système de vision en $m \left| \begin{matrix} x_m \\ y_m \\ \delta_m \end{matrix} \right.$.

Le capteur subit le mouvement arbitraire A.2. Nous conservons les notations introduites au A.1, pour écrire le déplacement relatif que subit le point rigide :

$$\left| \begin{array}{l} \Delta X_M = X'_M - X_M = -\omega_Y Z_M + \omega_Z Y_M - T_X \\ \Delta Y_M = Y'_M - Y_M = \omega_X Z_M - \omega_Z X_M - T_Y \\ \Delta Z_M = Z'_M - Z_M = \omega_Y X_M - \omega_X Y_M - T_Z \end{array} \right. \tag{A.16}$$

Les résultats A.12 concernant la partie 2D du mouvement restent conservés. Par un raisonnement analogue en tous points, et en remplaçant $\frac{1}{Z_M} = \frac{\delta}{f b_s}$ on abouti à l'expression complète des coordonnées du points m':

$$\left| \begin{array}{l} x'_m = \dfrac{x_m + y_m \omega_Z - f\omega_Y - \frac{T_X \delta_m}{b_s}}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z \delta_m}{f b_s} + 1} \\ \\ y'_m = \dfrac{y_m - x_m \omega_Z + f\omega_X - \frac{T_Y \delta_m}{b_s}}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z \delta_m}{f b_s} + 1} \\ \\ \delta'_m = \dfrac{\delta_m}{\frac{x_m}{f}\omega_Y - \frac{y_m}{f}\omega_X - \frac{T_Z \delta_m}{f b_s} + 1} \end{array} \right. \tag{A.17}$$

Finalement, nous choisissons de noter le déplacement relatif subi par le point

$$\left| \begin{matrix} \mu \\ \nu \\ \xi \end{matrix} \right. = \left| \begin{matrix} x'_m - x_m \\ y'_m - y_m \\ \delta'_m - \delta_m \end{matrix} \right. \tag{A.18}$$

Annexe B

Méthodes Numériques de Résolution de Systèmes Linéaires

Au cours de cette section, nous allons décrire au mieux les différentes méthodes de résolution de systèmes linéaires aux moindres carrés utilisées en 2.2.2. Ces méthodes peuvent être séparées en deux catégories :

- Les méthodes qui visent à établir une Décomposition en Valeurs Singulière de la matrice étudiée.
- Les méthodes visant à établir une factorisation *QR* de la matrice étudiée.

Dans la suite, nous noterons M la matrice étudiée, et le système à résoudre est donc :

$$Mx = b \tag{B.1}$$

B.1 Décomposition en Valeurs Singulières

La Décomposition en Valeurs Singulières (*Singular Values Decomposition, SVD*) vise à écrire la matrice M sous la forme :

$$M = U\Sigma V^* \tag{B.2}$$

où Σ est une matrice diagonale, dont les coefficients sont dans $\mathbb{R}^+$, dont l'inverse est donc immédiat, U et V sont des matrices orthogonales[1]. Il est intéressant

[1] Dans le cas générale, ces matrices sont unitaires, ce qui est équivalent à des matrices orthogonales à valeurs dans $\mathbb{R}$.

de noter, en considérant que les coefficients de Σ sont ordonnés, que cette décomposition est unique.

Le chercheur en traitement des images peut considérer la décomposition en valeurs singulières comme se rapprochant de l'analyse en composantes principales.

Le système B.1 à résoudre devient donc trivial. En effet, les différentes dimensions sont entièrement découplées par la SVD, et les matrices orthogonales sont facilement inversibles. Toutefois, il arrive que certaines des valeurs singulières soient numériquement beaucoup plus faibles que les autres et peuvent donc entraîner des instabilité numériques. Dès lors, annuler ces valeurs permet d'obtenir une plus grande stabilité et *in fine* une plus grande précision.

La méthode que nous avons testée repose sur l'utilisation des transformations de HOUSEHOLDER [Hou58], qui sont une généralisation des réflexions à un espace à n dimensions. L'utilisation de ces transformations, depuis son introduction par GOLUB [GL96] rend la SVD algorithmiquement efficace.

B.2 Factorisation *QR*

Contrairement à la SVD, la factorisation QR consiste à écrire la matrice M sous la forme :

$$\Sigma = QR \tag{B.3}$$

Où R est une matrice triangulaire, et Q est une matrice orthogonale. Nous avons étudié deux variantes de cette transformation, la première est la factorisation simple décrite plus haut, tandis que la seconde est la Décomposition Orthogonale Complète, qui consiste à écrire M sous la forme :

$$\Sigma = QRZ^* \tag{B.4}$$

Où Q et Z snt deux matrices orthogonales.

Ici, la matrice Z est utilisée de façon à annuler les dernières colonnes de R, et donc d'apporter plus de robustesse, dans le cas où R serait trapézoïdale supérieure, et non triangulaire supérieure. Cela est notamment utile pour la prise en compte de certains problèmes pathologiques, comme le mauvais conditionnement des matrices à pseudo-inverser.

Contrairement à la SVD, cette décomposition n'est pas unique. Plusieurs solutions sont possibles pour parvenir à une telle décomposition. Nous avons choisi de ne travailler qu'avec celles qui reposent sur les transformations de HOUSEHOLDER.

De plus, et c'est là une différence significative avec la SVD, il n'est plus ici possible d'annuler les termes non-significatifs de la décomposition, pour la simple raison qu'il n'est pas possible de les identifier, dès lors, les résultats sont nécessairement moins précis. Toutefois, cette perte de précision s'accompagne d'un gain de rapidité, les transformations à calculer étant moins nombreuses, et moins complexes.

Annexe C

Effet d'un décalage du FoE sur la représentation C-Vélocité d'un plan

L'objet de cette annexe est de développer certains calculs exploités dans le chapitre 4, et en particulier dans la section 4.2.1.

Afin de quantifier l'effet d'un décalage du FoE sur la représentation C-Vélocité d'un plan, nous commençons par définir l'ensemble $\mathcal{S}$ des points imageant le même plan. Au sein de cet ensemble, nous définissons la courbe $iso-w$, $\mathcal{C}$. Dans un premier temps, nous allons commencer par déterminer l'équation de cette courbe $iso-w$ avant d'évaluer, dans un second temps, les c-values des points appartenant à cette courbe. En effet, vu le processus cumulatif de la C-Vélocité, les seuls points pouvant interagir dans l'espace de vote sont les points partageant le même w.

C.1 Détermination des Iso-W

Par définition, et en utilisant l'équation 4.10 :

$$\mathcal{C} = \left\{ m \middle| \begin{matrix} x \\ y \end{matrix} | \frac{T_Z\,|x|}{fd} \sqrt{(x-x_{FoE})^2+(y-y_{FoE})^2} = w_{constant} \right\} \tag{C.1}$$

D'où il vient que :

$$\mathcal{C} = \left\{ m \middle| \begin{matrix} x \\ y \end{matrix} | \,|x| \sqrt{(x-x_{FoE})^2+(y-y_{FoE})^2} = K \right\} \tag{C.2}$$

où $K = \frac{fd}{w_{constant}T_Z}$ est une constante. En isolant y :

$$\mathcal{C} = \left\{ m \middle| \begin{array}{c} x \\ y = \pm\sqrt{\frac{K^2}{x^2} - (x - x_{FoE})^2} + y_{FoE} \end{array} \right\} \tag{C.3}$$

C.2 Calculs des C-Values, le long d'une iso-w

Dans cette seconde partie, nous allons nous attacher à évaluer $c_{batiment}$ le long d'une courbe iso-w, pour un candidat FoE dont les coordonnées sont : $\left| \begin{array}{l} x_{FoE} + \Delta x \\ y_{FoE} + \Delta y \end{array} \right.$. De la définition de $c_{batiment}$, il vient que :

$$c_{building}\left(m \middle| \begin{array}{c} x \\ y \end{array} \in \mathcal{C} \right) = |x| \sqrt{(x - x_{FoE} - \Delta x)^2 + (y - y_{FoE} - \Delta y)^2} \tag{C.4}$$

En combinant l'équation précédente à l'équation C.3 :

$$c_{building}\left(m \middle| \begin{array}{c} x \\ y \end{array} \in \mathcal{C} \right) = |x| \sqrt{(x - x_{FoE} - \Delta x)^2 + \left(\pm\sqrt{\frac{K^2}{x^2} - (x - x_{FoE})^2} - \Delta y \right)^2} \tag{C.5}$$

Si $\left| \begin{array}{l} \Delta x = 0 \\ \Delta y = 0 \end{array} \right.$, l'équation C.5 nous mène à $c_{building}(m \in \mathcal{C}) = K$. En d'autres mots, tous les points pour lesquelles la norme du flot optique est identique partage un même $c_{batiment}$, au signe près. Nous retrouvons bien ici le principe fondateur de la C-Vélocité Directe.

En élevant au carré et en développant Eq. C.5, il vient que $c^2_{building}$ peut exprimer comme suit :

$$\begin{aligned} c^2_{building}\left(m \middle| \begin{array}{c} x \\ y \end{array} \in \mathcal{C} \right) = K^2 - \Delta x x^2 (2x - 2x_{FoE} - \Delta x) + \\ \Delta y^2 x^2 \mp \Delta y \sqrt{x^2 K^2 - x^4 (x - x_{FoE})} \end{aligned} \tag{C.6}$$

Donc, si l'on calcule $c_{batiment}$ le long d'une $iso-w$, pour un candidat FoE, distant du FoE réel, les c-values résultantes présentent une dispersion qui est dûe à deux termes distincts.

Le premier

$$\Delta x x^2 (2x - 2x_{FoE} - \Delta x) \tag{C.7}$$

dépend de Δx et d'une fonction du troisième ordre en x, alors que le second :

$$\Delta y^2 x^2 \mp \Delta y \sqrt{x^2 K^2 - x^4 (x - x_{FoE})} \tag{C.8}$$

dépend uniquement de Δy et d'une courbe d'ordre $2,5$ en x.

Par unicité de la décomposition en série entière, il ne peut pas exister de couples $(\Delta x, \Delta y)$ qui annule globalement toutes variations. De plus, il est immédiat que l'amplitude de ces variations est directement croissante en fonction des deux paramètres Δx et Δy.

Annexe D

Moyens Expérimentaux

Au cours de ce travail de thèse, nous avons tenu à mettre un accent sur le travail expérimental. Bien que les moyens expérimentaux du LIVIC et de l'IEF nous permettent de mener à bien l'intégralité des tests requis par notre travail, nous avons tenu à exploiter également des bases de données aussi diversifiées que possible. En effet, il nous semble que l'exploitation de différentes configurations matérielles présente un intérêt non négligeable.

Nous tenons donc à remercier le laboratoire Heudiasyc de l'Université de Technologie de Compiègne[1] [RFBC10], ainsi que l'Institute of Measurement and Control System de l'Institut de Technologie de Karlsruhe[2] [KGL10] d'avoir rendus leurs données disponibles.

Dans cette annexe, nous présenterons les différents systèmes, aussi bien réels que simulés sur lesquels nous avons pu mener nos expérimentations.

D.1 Système Simulé - Le Logiciel SiVIC

Le logiciel a été développé au sein du Livic [GRL+06] avant de conduire à la création de la start-up Civitec[3]. L'objectif premier de ce logiciel est de simuler des systèmes multi-capteurs complexes. SiVIC permet de générer des bases de données multi-capteurs réalistes et intégrant un grand nombre de paramètres (déformations optiques, conditions atmosphériques adverses, etc.).

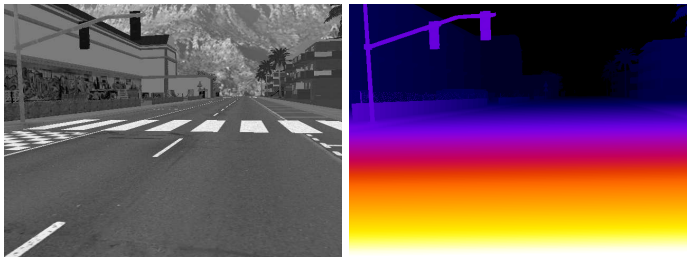
Ce système nous permet, notamment, de générer des séquences stéréo, parfaitement rectifiées et photoréalistes. De plus, il nous est également possible d'obte-

[1] http://www.hds.utc.fr/

[2] http://www.mrt.kit.edu/

[3] http://www.civitec.net/

FIGURE D.1 – Image générée par SiVIC



(a) Image générée par SiVIC (b) Carte de profondeur générée par SiVIC

nir des cartes de profondeur, permettant ainsi de mesurer l'influence de la qualité des appariements stéréos. Finalement, l'ensemble des paramètres dynamiques du véhicule simulé sont également disponibles. Ces paramètres incluent la position, les vitesses, accélérations et taux de rotation.

D.2 Systèmes Réels

Les données réelles que nous avons utilisées au cours de ce travail proviennent de cinq sources :

- Le Projet LoVE (Logiciel d'Observation des Vulnérables), mené par le LAS-

FIGURE D.2 – Image issue des bases de données LoVE

MEA[4], projet auquel le groupe ACCIS de l'IEF et le LIVIC (IFSTTAR) ont pris part.

- Les moyens propres du Livic, en l'occurrence, le véhicule expérimental CARLLA.
- Le prototype Mini-Truck de l'IEF, équipé d'un capteur Kinect.
- l'Université de Karlsruhe.
- L'Université de Technologie de Compiègne.

D.2.1 Système LoVE

Le système LoVE a été mis au point lors du projet ANR du même nom. Il consiste en une paire stéréo de caméra. Aucun capteur additionnel n'est disponible pour ces images. De plus, si la synchronisation des caméras est garantie, la continuité temporelle des données ne l'est pas.

D.2.2 CARLLA

Le véhicule CARLLA[5] fait partie de la flotte de prototypes du LiVIC. Outre une paire stéréo, plusieurs capteurs sont embarqués :

- Une centrale inertielle Crossbow VG400®.

[4] http://http//www.lasmea.univ-bpclermont.fr/

[5] Contrôleur d'Assistance Routière Longitudinale et LAtérale

FIGURE D.3 – Image prise par l'un des capteurs du véhicule CARLLA

- Un GPS centimétrique RTK.
- Un GPS "commercial" Garmin.
- des odomètres permettant de mesurer la distance parcourue par les roues.

D.2.3 Mini-Truck

Mini-Truck est une plateforme d'expérimentation robotique mise en place à l'Institut d'Electronique Fondamentale, par Abdelhafid EL OUARDI. Ce véhicule est équipé d'odomètres fixés sur ses roues et d'une plateforme permettant la mise en place d'une caméra ou de télémètres à ultra-sons. Lors de l'utilisation que nous en faisons, nous avons installé un capteur Kinect®à l'avant du véhicule. Ce capteur (décrit en 1.4.4) nous permet d'acquérir simultanément des images couleur et une carte de distance dense, limitée à une profondeur de 5 mètres.

D.2.4 Karlsruhe

Les bases de données mises à disposition par l'Université de Karlsruhe consistent en des séquences stéréos, des données odométriques, ainsi que des données GPS. Les images sont prises par des caméras PointGrey Flea2.

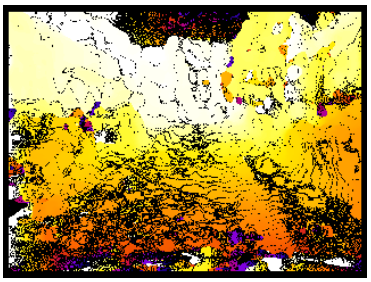
(a) (b)

FIGURE D.4 – Images obtenues via le capteur Kinect®: (a) image couleur ; (b) : carte de profondeur

FIGURE D.5 – Prototype Mini-Truck

FIGURE D.6 – Image Issue de la base de données Karlsruhe

(a) Image Source - Système Videre (UTC)



(b) Carte de Disparité générée par le système Videre (UTC)

Les données odométriques et GPS sont issues d'un capteur OXTS RT3003 de grande précision. En particulier, les taux de rotation sont mesurés avec une précision annoncée de 0.01deg.s^{-1} et les accélérations avec une précision annoncée de 0.01m.s^{-2}. Ce système est également pourvu d'un GPS centimétrique. Les séquences disponibles présentent une grande variété de scénarios différents.

D.2.5 UTC

Le système utilisé par l'Université de Compiègne est constitué d'un système de stéréo-vision *on-chip*, commercialisé par la société Videre[6]. Ce système délivre à la fois les paires stéréos et les cartes de disparité, calculées par une méthode non dévoilée.

Le véhicule utilisé était également équipé d'un GPS Septentrio PolaRx. Les autres capteurs utilisés étaient ceux normalement utilisés par le système ESP : des capteurs de vitesses de rotation des roues et un gyromètre mesurant la vitesse de lacet.

D.2.6 Récapitulatif

[6]http://www.videredesign.com/

Base de Données	Résolution	Focale (pixels)	Base (mm)	Cadence
SiVIC	variable	variable	variable	variable
LoVE	640x480	645	650	variable
CARLLA	768x578	1016	495	25Hz
Mini-Truck	800x600		N.D.	30Hz
Karlsruhe	1344x372	894	570	10Hz
UTC	320x240	381	438	30Hz

TABLE D.1 – Récapitulatif des caractéristiques des systèmes de vision utilisés

Bibliographie

[AKB08] M Agrawal, K Konolige, and M R Blas. Censure: Center surround extremas for realtime feature detection and matching. *European Conference on Computer Vision, Marseille, France*, 5305:102–115, Oct 2008.

[Alf02] M Alfonso. On cascading small decision trees. *PhD Dissertation, Universitat Autonoma de Barcelona, Espagne*, 2002.

[AV07] D Arthur and S Vassilvitskii. k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM Symposium On Discrete Algorithms, New Orleans, USA*, pages 1027–1035, Jan 2007.

[Bad07] H Badino. A robust approach for ego-motion estimation using a mobile stereo platform. *Lecture Notes In Cmputer Science*, 3417:198–208, 2007.

[BAHH92] J Bergen, P Anandan, K Hanna, and R Hingorani. Hierarchical model-based motion estimation. *European Conference on Computer Vision, Santa Margherita Ligure, Italie*, 588:237–252, Mai 1992.

[BB95] S Beauchemin and J Barron. The computation of optical flow. *ACM Computing Surveys (CSUR)*, 1995.

[BBA10] A Bak, S Bouchafa, and D Aubert. Detection of independently moving objects through stereo vision and ego-motion extraction. In *Intelligent Vehicles Symposium*, Jun 2010.

[BBB+01] A Bensrhair, M Bertozzi, A Broggi, P Miche, S Mousser, and G Toulminet. A cooperative approach to vision-based vehicle detection.

IEEE Intelligent Transportation Systems Conference, Oakland, USA, pages 207–212, Aou 2001.

[BBC+02] M Bertozzi, A Broggi, M Cellario, A Fascioli, P Lombardi, and M Porta. Artificial vision in road vehicles. *Proceedings of the IEEE*, 90(7):1258–1271, Juil 2002.

[BBFN00] M Bertozzi, A Broggi, A Fascioli, and S. Nichele. Stereo vision-based vehicle detection. *IEEE Intelligent Vehicle Symposium, Dearborn, USA*, pages 39–44, Oct 2000.

[BBFT04] M Bertozzi, A Broggi, A Fascioli, and A Tibaldi. Pedestrian localization and tracking system with kalman filtering. *IEEE Intelligent Vehicles Symposium, Parma, Italie*, pages 584–589, 2004.

[BC05] G Le Besnerais and F Champagnat. Dense optical flow by iterative local window registration. *IEEE International Conference on Image Processing, Genoa, Italie*, Sep 2005.

[BD98] S D Buluswar and B A Draper. Color machine vision for autonomous vehicles. *Engineering Applications of Artificial Intelligence*, 11(2):245–256, Avr 1998.

[BDW06] T Bailey and H Durrant-Whyte. Simultaneous localization and mapping (slam): Part ii. *Robotics & Automation Magazine*, 13(3):108–117, Sep 2006.

[Bey92] H Beyer. Accurate calibration of ccd-cameras. *Computer Vision and Pattern Recognition, Champaign, USA*, pages 96–101, Juin 1992.

[BFT06] A Broggi, RL Fedriga, and A Tagliati. Pedestrian detection on a moving vehicle: an investigation about near infra-red images. *IEEE Intelligent Vehicle Symposium, Tokyo, Japon*, pages 431–436, Juin 2006.

[Bjö96] Åke Björck. Numerical methods for least-squares problems, siam. 1996.

[BM04] S Baker and I Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[BMVF08] H Badino, R Mester, T Vaudrey, and U Franke. Stereo-based free space computation in complex traffic scenarios. *Image Analysis and Interpretation*, pages 189–192, 2008.

[Bou99] J Bouguet. Pyramidal implementation of the lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 1999.

[BPCL06] C Braillon, C Pradalier, JL Crowley, and C Laugier. Real-time moving obstacle detection using optical flow models. *IEEE Intelligent Vehicles Symposium, Tokyo, Japon*, pages 466–471, Juin 2006.

[BPU+08] C Braillon, C Pradalier, K Usher, J Crowley, and C Laugier. Occupancy grids from stereo and optical flow data. *Experimental Robotics*, 39:367–376, 2008.

[BPZ09] S Bouchafa, A Patri, and B Zavidovique. Efficient plane detection from a single moving camera. *International Conference on Image Processing, Le Caire, Egypte*, pages 3493–3496, Nov 2009.

[Bro02] D C Brown. Close-range camera calibration. *Photogrammetric Engineering*, 37(8):855–866, Nov 2002.

[BSL+11] S Baker, D Scharstein, J Lewis, S Roth, and M Black. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

[BT99] S Birchfield and C Tomasi. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293, 1999.

[BTV06] H Bay, T Tuytelaars, and L VanGool. Surf: Speeded up robust features. *European Conference on Computer Vision, Graz, Autriche*, 3951:404–417, Mai 2006.

[Bur98] C Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[BVZ01] Y Boykov, O Veksler, and R Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[BW04] T Brox and J Weickert. A tv flow based local scale measure for texture discrimination. *European Conference on Computer Vision, Prague, République Tchèque*, 3022:578–590, Mai 2004.

[BZ09] S Bouchafa and B Zavidovique. C-velocity: A cumulative frame to segment objects from egomotion. *Pattern Recognition and Image Analysis*, 19(4):583–590, 2009.

[CET01] T Cootes, G J Edwards, and C J TAYLOR. Active appearance models. *Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[Cha05] S Chambon. Mise en correspondance stéréoscopique d'images couleur en présence d'occultations, université paul sabatier, toulouse, france. *PhD thesis*, 2005.

[CM02] D Comaniciu and P Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[CMM06] Y Cheng, M W Maimone, and L Matthies. Visual odometry on the mars exploration rovers. *Robotics & Automation Magazine*, 13(2):54–62, Jan 2006.

[CMR07] A Comport, E Malis, and P Rives. Accurate quadrifocal tracking for robust 3d visual odometry. *IEEE International Conference on Robotics and Automation, Rome, Italie*, pages 40–45, Avr 2007.

[CMR10] A Comport, E Malis, and P Rives. Real-time quadrifocal visual odometry. *The International Journal of Robotics Research*, 29(2-3):245–266, 2010.

[DC00] F Dornaika and R Chung. Cooperative stereo-motion: Matching and reconstruction. *Computer Vision and Image Understanding*, 79(2):408–427, Sep 2000.

[DD02] D Demirdjian and T Darrell. Motion estimation from disparity images. *International Conference on Computer Vision, Vancouver, Canada*, 1:213–218, Jul 2002.

[DH72] R O Duda and P E Hart. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972.

[Dic02] E Dickmanns. The development of machine vision for road vehicles in the last decade. *IEEE Intelligent Vehicle Symposium, Londres, Royaume-Uni*, 1:268–281, Juin 2002.

[DNC+01] M Dissanayake, P Newman, S Clark, H F Durrant-Whyte, and M Csorba. A solution to the simultaneous localization and map building (slam) problem. *IEEE Transactions on Robotics and Automation*, 17(3):229–241, 2001.

[DRMS07] A Davison, I Reid, N D Molton, and O Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007.

[DT05] N Dalal and B Triggs. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition*, 1:886–893, 2005.

[Dum09] Y Dumortier. Perception monoculaire de l'environnement pour les systèmes de transport intelligents. *PhD thesis, Mines Paristech*, 2009.

[DWB06] H Durrant-Whyte and T Bailey. Simultaneous localisation and mapping (slam): Part 1 the essential algorithms. *Robotics and Automation Magazine*, 13(2):99–110, 2006.

[EH08] I Ernst and H Hirschmüller. Mutual information based semi-global stereo matching on the gpu. *Advances in Visual Computing*, 5358:228–239, 2008.

[EKG08] M Enzweiler, P Kanter, and D Gavrila. Monocular pedestrian recognition using motion parallax. *IEEE Intelligent Vehicles Symposium, Eindhoven, Pays-Bas*, pages 792–797, 2008.

[ER02] G Egnal and R.P.Wildes. Detecting binocular half-occlusions: Empirical comparisons of five approaches. *Pattern Analysis and Machine Intelligence*, 24(8):1127–1133, 2002.

[ESG09] E Einhorn, C Schroeter, and H M Gross. Monocular obstacle detection for real-world environments. *Autonome Mobile Systeme 2009*, pages 1–33, 2009.

[ESN06] C Engels, H Stewénius, and D Nister. Bundle adjustment rules. *In Photogrammetric Computer Vision*, Jan 2006.

[FA97] C Fermüller and Y Aloimonos. On the geometry of visual correspondence. *International Journal of Computer Vision*, 21(3):223–247, 1997.

[FB81] M A Fischler and R C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24, 1981.

[FHT98] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. Aug 1998.

[FRBG05] U Franke, C Rabe, H Badino, and S Gehrig. 6d-vision: Fusion of stereo and motion for robust environment perception. *Lecture Notes in Computer Science*, 3663:216–223, 2005.

[FS95] Y Freund and R Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. *Lecture Notes in Computer Science*, 904:23–37, 1995.

[FSMA10] B Freedman, A Shpunt, M Machline, and Y Arieli. Depth mapping using projected patterns. *US2010118123*, 2010.

[FSWG06] B Fardi, I Seifert, G Wanielik, and J Gayko. Motion-based pedestrian recognition from a moving vehicle. *IEEE Intelligent Vehicle, Tokyo, Japon*, pages 219–224, 2006.

[FTV00] A Fusiello, E Trucco, and A Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22, Sep 2000.

[GB00] J Gama and P Brazdil. Cascade generalization. *Machine Learning*, 41:315–343, 2000.

[GJB+00] I Gresham, N Jain, T Budka, A Alexanian, N Kinayman, B Ziegner, S Brown, and P Staecker. A 76-77 ghz pulsed-doppler radar module for autonomous cruise control applications. *IEEE Microwave Symposium Digest*, 3:1551–1554, Jan 2000.

[GL96] Gene Howard Golub and Charles F. Van Loan. Matrix computations. *JHU Press*, page 694, Jan 1996.

[Gle97] M Gleicher. Projective registration with difference decomposition. *IEEE Conference on Computer Vision and Pattern Recognition, San Juan, Porto Rico*, page 331, Juin 1997.

[GM07] D Gavrila and S Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *International Journal of Computer Vision*, 73:41–59, 2007. 10.1007/s11263-006-9038-7.

[GPGD10] D Gruyer, S Pechberti, D Gingras, and F Dupin. Robust positioning in safety applications for the cvis project. *IEEE Intelligent Vehicles Symposium, San Diego, USA*, pages 262–268, Juin 2010.

[GRL+06] D Gruyer, C Royere, N Du Lac, G Michel, and J Blosseville. Sivic and rtmaps, interconnected platforms for the conception and the evaluation of driving assistance systems. *IEEE Internationa Conference on Intelligent Transportation Systems, Toronto, Canada*, Sep 2006.

[GY02] M Gong and Y Yang. Genetic-based stereo algorithm and disparity map evaluation. *International Journal of Computer Vision*, 47(1-3):63–77, 2002.

[Hay99] Simon S. Haykin. Neural networks: a comprehensive foundation. page 842, Jan 1999.

[Hei02] S Heinrich. Fast obstacle detection using flow/depth constraint. *Intelligent Vehicle Symposium, Londres, Royaume-Uni*, 2:658–665, Juin 2002.

[HIG02] H Hirschmüller, PR Innocent, and J Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47(1-3):229–246, 2002.

[Hir05] H Hirschmüller. Accurate and efficient stereo processing by semi-global matching and mutual information. *IEEE Conference en Computer Vision and Pattern Recognition, San Diego, USA*, 2:807–814, Juin 2005.

[HLPA06] N Hautière, R Labayrade, M Perrollaz, and D Aubert. Road scene analysis by stereovision: a robust and quasi-dense approach. *IEEE International Conference on Control, Automation, Robotics and Vision, Singapore*, pages 1–6, Dec 2006.

[HM95] R Horaud and O Monga. Vision par ordinateur: outils fondamentaux. *Hermes*, 1995.

[HN10] C Huang and R Nevatia. High performance object detection by collaborative learning of joint ranking of granules features. *IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, USA*, pages 41–48, Juin 2010.

[Hou58] A Householder. Unitary triangularization of a nonsymmetric matrix. *Journal of the ACM (JACM)*, 1958.

[HPON10] I Haller, C Pantilie, F Oniga, and S Nedevschi. Real-time semi-global dense stereo solution with improved sub-pixel accuracy. *IEEE Intelligent Vehicles Symposium, San Diego, USA*, pages 369–376, Juin 2010.

[HS88] C Harris and M Stephens. A combined corner and edge detector. *Alvey vision conference*, 15:147–151, Jan 1988.

[HS93] R Hummel and V Sundareswara. Motion parameter estimation from global flow field data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(5):459–476, 1993.

[HS94] B K P Horn and B G Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1994.

[HTFF05] T Hastie, R Tibshirani, J Friedman, and J Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, Jan 2005.

[Hub73] P Huber. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1973.

[HZ03] Richard Hartley and Andrew Zisserman. Multiple view geometry in computer vision. *Cambridge University Press*, page 655, 2003.

[IA96] M Irani and P Anandan. Parallax geometry of pairs of points for 3d scene analysis. *lecture Notes in Computer Science*, 1064:17–30, 1996.

[IKB01] Luca Iocchi, Kurt Konolige, and Max Bajracharya. Visually realistic mapping of a planar environment with stereo. *Lecture Notes in Control and Information Science*, 271:521–532, Jun 2001.

[IRP97] M Irani, B Rousso, and S Peleg. Recovery of ego-motion using region alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):268–272, Jan 1997.

[JU97] S Julier and J Uhlmann. A new extension of the kalman filter to nonlinear systems. *Signal Processing, Sensor Fusion and Target Recognition*, (6):182–193, 1997.

[Kal60] R Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82:35–45, 1960.

[KAS11] K Konolige, M Agrawal, and J Sola. Large-scale visual odometry for rough terrain. *Robotics Research*, 66:201–212, 2011.

[KGL10] B Kitt, A Geiger, and H Lategahn. Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme. *IEEE Intelligent Vehicles Symposium, San Diego, USA*, pages 486–492, Juin 2010.

[KMH95] C Kolb, D Mitchell, and P Hanrahan. A realistic camera model for computer graphics. *SIGGRAPH Conference on Computer Graphics and Interactive Techniques, New York, USA*, Mar 1995.

[KTS98] T Kalinke, C Tzomakas, and W V Seelen. A texture-based object detection and an adaptive model-based classification. *IEEE Intelligent Vehicles Symposium, Stuttgart, Allemagne*, pages 341–346, Juin 1998.

[KZP+08] S Kammel, J Ziegler, B Pitzer, M Werling, T Gindele, D Jagzent, J Schröder, M Thuy, M Goebl, F von Hundelshausen, O Pink, C Frese, and C Stiller. Team annieway's autonomous system for the 2007 darpa urban challenge. *Journal of Field Robotics*, 25(9):615–639, 2008.

[LAC06] S Lefebvre, S Ambellouis, and F Cabestaing. Obstacles detection on a road by dense stereovision with 1d correlation windows and fuzzy filtering. *IEEE Intelligent Transportation Systems Conference, Toronto, Canada*, pages 739–744, Sep 2006.

[LAT02] R Labayrade, D Aubert, and J Tarel. Real time obstacle detection in stereovision on non-flat road geometry through "v-disparity" representation. *Intelligent Vehicle Symposium, Londres, Royaum-Uni*, 2:646–651, Juin 2002.

[LCCG07] B Leibe, N Cornelis, K Cornelis, and L Van Gool. Dynamic 3d scene analysis from a moving vehicle. *IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA*, pages 1–8, Juin 2007.

[LDD+06] M Lhuillier, F Dekeyser, M Dhome, E Mouragnon, and P Sayd. Real-time localization and 3d reconstruction. *IEEE Conference on Computer Vision and Pattern Recognition, New York, USA*, 1:363–370, Juin 2006.

[Lev44] K Levenberg. A method for the solution of certain problems in least squares. *Quarterly of Applied Mathematics*, 1944.

[LGLL10] J Li, W Gong, W Li, and X Liu. Robust pedestrian detection in thermal infrared imagery using the wavelet transform. *Infrared Physics & Technology*, 53(4):267–273, 2010.

[LH95] C L Lawson and R J Hanson. Solving least squares problems. *SIAM's Classics in Applied Mathematics*, 1995.

[LK81] B Lucas and T Kanade. An iterative image registration technique with an application to stereo vision. *International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[Llo82] S Lloyd. Least squares quantization in pcm. *Information Theory*, 28(2):129–137, 1982.

[Low04] D Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LRGA05] R Labayrade, C Royere, D Gruyer, and D Aubert. Cooperative fusion for multi-obstacles detection with use of stereovision and laser scanner. *Autonomous Robots*, 19(2):117–140, 2005.

[LS09] C Lundquist and T B Schön. Estimation of the free space in front of a moving vehicle. *Proceedings of the SAE World Congress, Detroit, USA*, Jan 2009.

[LZ99] C Loop and Z Zhang. Computing rectifying homographies for stereo vision. *IEEE Conference on Computer Vision and Pattern Analysis, Fort Collins, USA*, 1:11–25, Juin 1999.

[LZGR11] P Lenz, J Ziegler, A Geiger, and M Roser. Sparse scene flow segmentation for moving object detection in urban environments. *IEEE Intelligent Vehicles Symposium, Baden Baden, Allemagne*, Juin 2011.

[Mar63] D Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 1963.

[MC95] J Martin and J L Crowley. Experimental comparison of correlation techniques. *Proceedings of the International Conference on Intelligent Autonomous Systems*, 1995.

[MG06] S Munder and D M Gravila. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.

[MMP08] D Muller, M Meuter, and S B Park. Motion segmentation using interest points. *Intelligent Vehicles Symposium, Eindhover, Pays-Bas*, pages 19–24, Juin 2008.

[MTKW02] M Montemerlo, S Thrun, D Koller, and B Wegbreit. Fastslam: A factored solution to the simultaneous localization and mapping problem. *AAAI National Conference on Artificial Intellignece, Edmonton, Canada*, pages 593–598, Juil 2002.

[MTKW03] M Montemerlo, S Thrun, D Koller, and B Wegbreit. Fastslam 2.0: An improved particle filtering algorithm for simultaneous localization and mapping that provably converges. *International Joint Conference on Artificial Intelligence, Accapulco, Mexique*, pages 1151–1156, Aou 2003.

[NBT09] S Nedevschi, S Bota, and C Tomiuc. Stereo-based pedestrian detection for collision-avoidance applications. *IEEE Transactions on Intelligent Transportation Systems*, 10(3):380–391, 2009.

[OM10] S Obdržálek and J Matas. A voting strategy for visual ego-motion from stereo. *IEEE Intelligent Vehicles Symposium, San Diego, USA*, pages 382–387, Juin 2010.

[PBB+06] N Papenberg, A Bruhn, T Brox, S Didas, and J Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67(2):141–158, 2006.

[Per08] M Perrollaz. Détection d'obstacles multi-capteurs supervisée par stéréovision. *PhD thesis, Université Pierre et Marie Curie, Paris*, Sep 2008.

[PF10] D Pfeiffer and U Franke. Efficient representation of traffic scenes by means of dynamic stixels. *IEEE Intelligent Vehicles Symposium, San Diego, USA*, pages 217–224, Juin 2010.

[PKF07] J Pons, R Keriven, and O Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72(2):179–193, 2007.

[Pra79] K Prazdny. Motion and structure from optical flow. *Proceedings of the 6th International Joint Conference on Artificial Intelligence*, 2, 1979.

[RA80] J W Roach and J K Aggarwal. Determining the movement of objects from a sequence of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 554–562, 1980.

[RASP96] B Rousso, S Avidan, A Shashua, and S Peleg. Robust recovery of camera rotation from three frames. *IEEE International Conference on Computer Vision and Pattern Recognition, San Francisco, USA*, Juin 1996.

[RFB09] F Rodriguez, V Frémont, and P Bonnifait. An experiment of a 3d real-time robust visual odometry for intelligent vehicles, st louis, usa. *IEEE International Conference on Intelligent Transportation Systems*, pages 1–6, Oct 2009.

[RFBC10] F Rodríguez, V Frémont, P Bonnifait, and V Cherfaoui. Visual confirmation of mobile objects tracked by a multi-layer lidar. *IEEE International Conference on Intellignent Transportation Systems, Funchal, Portugal*, pages 849–854, Sep 2010.

[RFG07] C Rabe, U Franke, and S Gehrig. Fast detection of moving objects in complex scenarios. *IEEE Intelligent Vehicles Symposium, Istambul, Turquie*, pages 398–403, Juin 2007.

[Ric03] I Richardson. H. 264 and mpeg-4 video compression. *Wiley Online Library*, Jan 2003.

[SB02] J Stoer and R Bulirsch. Introduction to numerical analysis. *Springer*, 2002.

[SBW05] N Slesareva, A Bruhn, and J Weickert. Optic flow goes stereo: A variational method for estimating discontinuity-preserving dense disparity maps. *Pattern Recognition*, 3663:33–40, 2005.

[SFS09] D Scaramuzza, F Fraundorfer, and R Siegwart. Real-time monocular visual odometry for on-road vehicles with 1-point ransac. *International Conference on Robotics and Automation, Kobe, Japon*, pages 4293–4299, Mai 2009.

[SGH04] A Shashua, Y Gdalyahu, and G Hayun. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. *IEEE Intelligent Vehicles Symposium, Parme, Italie*, pages 1–6, Juin 2004.

[Sha76] G Shafer. A mathematical theory of evidence. *Princeton University Press*, 1976.

[Sha85] S Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985.

[SJ89] D Shulman and J.Y.Herve. Regularization of discontinuous flow fields. *Visual Motion*, pages 81–86, 1989.

[SJBK08] J Suhr, H Jung, K Bae, and J Kim. Outlier rejection for cameras on intelligent vehicles. *Pattern Recognition Letters*, 29:828–840, 2008.

[SJHG99] E Simoncelli, Berndt Jahne, H Haussecker, and P Geissler. Bayesian multi-scale differential optical flow. *Handbook of Computer Vision and Applications*, 1999.

[SM07] W Souid-Miled. Mise en correspondence stéréoscopique par approches variationnellles convexes ; application à la détection d'obstacles routier. *PhD thesis, Université Paris Est*, 2007.

[Sme90] P Smets. The combination of evidence in the transferable belief model. *Pattern Analysis and Machine Intelligence*, 12(5), 1990.

[SMS00] G Stein, O Mano, and A Shashua. A robust method for computing vehicle ego-motion. *IEEE Intelligent Vehicles Symposium, Dearborn, USA*, pages 362–368, Juin 2000.

[SP98] K K Sung and T Poggio. Example-based learning for view-based human face detection. *Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[SP07] N Sunderhauf and P Protzel. Stereo odometry–a review of approaches. *tu-chemnitz.de*, 2007.

[SPLA07] N Soquet, M Perrollaz, R Labayrade, and D Aubert. Free space estimation for autonomous navigation. *International Conference on Computer Vision, Rio de Janeiro, Brésil*, Oct 2007.

[SRBB06] F Suard, A Rakotomamonjy, A Bensrhair, and A Broggi. Pedestrian detection using infrared images and histograms of oriented gradients. *IEEE Intelligent Vehicles Symposium, Tokyo, Japon*, pages 206–212, Juin 2006.

[SRR04] D Sazbon, H Rotstein, and E Rivlin. Finding the focus of expansion and estimating range using optical flow images and a matched filter. *Machine Vision and Applications*, 15:229–236, 2004.

[SS02a] D Scharstein and R Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[SS02b] H Y Shum and R Szeliski. Construction of panoramic image mosaics with global and local alignment. *International Journal of Computer Vision*, 48(2):151–152, 2002.

[ST94] Jianbo Shi and C Tomasi. Good features to track. *IEEE Conference on Computer Vision and Pattern Recognition, Seattle, USA*, pages 593 – 600, Juin 1994.

[Suv06] N Suvonvorn. Mise en correspondence d'images pour l'analyse du mouvement et la stéréovision. *PhD thesis, Université Paris Sud XI*, 2006.

[TK92] C Tomasi and T Kanade. Shape and motion from image streams: a factorization. *International Journal of Computer Vision*, 9(2):137–154, 1992.

[TM04] A Talukder and L Matthies. Real-time detection of moving objects from moving vehicles using dense stereo and optical flow. *IEEE*

Conference on Intelligent Robots and Systems, Sendai, Japon, pages 3718–3725, Sep 2004.

[TP91] M A Turk and A P Pentland. Face recognition using eigenfaces. *IEEE International Conference on Computer Vision and Pattern Recognition, Maui, USA*, pages 586–591, Juin 1991.

[TSO+05] T Teshima, H Saito, S Ozawa, K Yamamoto, and T Ihara. Estimation of foe without optical flow for vehicle lateral position detection. *Proceedings of IAPR Conference on Machine Vision Applications*, pages 406–409, Jan 2005.

[Ull79] S Ullman. The interpretation of visual motion. *MIT Press*, Jan 1979.

[Vap99] V Vapnik. An overview of statistical learning theory. *Neural Networks*, 1999.

[VBA08] T Vu, J Burlet, and O Aycard. Grid-based localization and online mapping with moving objects detection and tracking: new results. *IEEE Intelligent Vehicles Symposium, Eindhoven, Pays-Bas*, pages 684–689, Juin 2008.

[VJS05] P Viola, M J Jones, and D Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

[VW95] P Viola and W.M.Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1995.

[WBBN06] J Weickert, A Bruhn, T Brox, and N.Papenberg. A survey on variational optic flow methods for small displacements. *Mathematical models for Registration and Applications to Medical Imaging*, 10(1):103–136, 2006.

[Wil99] Todd A Williamson. A high-performance stereo vision system for obstacle detection. *PhD thesis, Carnegie Mellon University*, May 1999.

[Wol89] L B Wolff. Using polarization to separate reflection components. *IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA*, pages 363–369, Juin 1989.

[WRV+08] A Wedel, C Rabe, T Vaudrey, T Brox, and U Franke. Efficient dense scene flow from sparse or dense stereo data. *European Conference on Computer Vision, Marseille, France*, Oct 2008.

[WTPW09] M Werlberger, W Trobin, T Pock, and A Wedel. Anisotropic huber-l1 optical flow. *Proceedings of the British Machine Vision Conference, Londres, Royaume-uni*, Jan 2009.

[WWH07] F Wu, L Wang, and Z Y Hu. Foe estimation: Can image measurement errors be totally. *Pattern Recognition*, 40(7):1971–1980, 2007.

[XD92] S Xu and P E Danielson. Robust estimation of focus of expansion and depth from high confidence optical flow. *IAPR Workshop on Machine Vision Applications*, pages 105–108, 1992.

[YK04] K J Yoon and I S Kweon. Voting-based separation of diffuse and specular pixels. *Electronics Letters*, 40(20):1260–1261, 2004.

[YPP+10] J Yoder, M Perrollaz, I Paromtchik, Y Mao, and C Laugier. Experiments in vision-laser fusion using the bayesian occupancy filter. *International Symposium on Experimental Robotics, Delhi, Inde*, Dec 2010.

[ZDFL95] Z Zhang, R Deriche, O Faugeras, and Q T Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence*, 78(1-2):87–119, 1995.

[Zha98] Z Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision*, pages 161–198, 1998.

[ZPB07] C Zach, T Pock, and H Bischof. A duality based approach for real-time tv-l optical flow. *Lecture Notes in Computer Science*, 4713:214–223, Jan 2007.

[ZW94] R Zabih and J Woodfill. Non-parametric local transforms for computing visual correspondence. *European Conference on Computer Vision, Stockholm, Suède*, Mai 1994.

[ZYCA06] Q Zhu, M.C Yeh, K.T Cheng, and S Avidan. Fast human detection using a cascade of histograms of oriented gradients. *Computer*

Vision and Pattern Recognition, New York, USA, pages 1491–1498, Juin 2006.