DÉCISION CUMULATIVE POUR LA VISION DYNAMIQUE DES SYSTÈMES

Samia Bouchafa

Laboratoire IBISC, Informatique Biologie Intégrative et Systèmes Complexes, 91025 Evry cedex, France

Résumé

Les travaux présentés dans cette synthèse portent essentiellement sur l'analyse de scènes à partir de caméras mobiles avec pour application immédiate l'apport d'une vision par ordinateur efficace dans les systèmes autonomes. Ils sont le fruit d'une décennie de recherches menées d'abord à l'INRETS (actuellement IFSTTAR : Institut français des sciences et technologies des transports, de l'aménagement et des réseaux) puis à l'Université Paris Sud XI (Institut d'Électronique Fondamentale). L'idée initiale est que l'autonomie d'un système implique, ne serait-ce que pour raisons énergétiques, une faible variété d'opérateurs de perception, dont les algorithmes de vision. Les "primitives" extraites des images seront intrinsèquement robustes et stables vis-à-vis de perturbations variées. Elles doivent de plus anticiper, voire faciliter, un processus de décision à divers niveaux voulu systématique. Les lignes de niveaux répondent parfaitement à ces contraintes : on vérifie sans peine leur robustesse et leur abondance dans une image suggère et alimente un processus de décision cumulatif (manipulant un objet de décision unique : l'histogramme généralisé en espace de vote). Nos efforts se sont alors concentrés sur deux aspects. 1) Le premier concerne la définition d'une méthodologie cohérente dans laquelle un processus primaire d'extraction de lignes de niveaux est enrichi afin de permettre la construction de primitives plus complexes guidée par un modèle de déformation de l'image. Le nombre de composants donc la forme des primitives est fonction directe du nombre de variables caractérisant le mouvement (déformation) à déterminer. 2) Le second intéresse une méthode de décision cumulative unifiée permettant de traiter des thèmes applicatifs de complexité croissante. Nos travaux se déclinent alors en trois niveaux de cumul, chacun associé de manière à un stade particulier de l'analyse d'images. Les thèmes applicatifs traités pour illustrer notre démarche sont de complexité croissante : détection et estimation du mouvement en caméra fixe, recalage d'images en caméra mobile (type de mouvement connu et profondeur des objets contrainte) puis estimation générale du mouvement propre et de la structure de la scène en caméras embarquées sur un véhicule mobile. Les résultats obtenus montrent comment un choix de primitives robustes associé à un processus de décision cumulatif permet la réutilisation des opérateurs dans plusieurs secteurs. Les systèmes proposés ont la particularité d'être compacts et cohérents, propriété recherchée dans les applications considérées.

Mots clés : Vision par ordinateur, Analyse du mouvement 2D et 3D, Ego-mouvement, Recalage d'images, Extraction de primitives, Lignes de niveaux, Décision cumulative.

Abstract

This review deals with scene analysis from mobile cameras in the context of efficient computer vision algorithm design for autonomous systems. They result from a decade of research at INRETS (currently IFSTTAR : the French research institute on transportation) and at University Paris Sud XI. The main idea is that the autonomy of an artificial system should involve, at least for energetic reasons, a small diversity of perception operators (including vision algorithms). Moreover, primitives that are extracted from images should be inherently robust and stable towards various perturbations. They must also anticipate or facilitate a decision-making process at various levels. We show that level-line primitives perfectly meet these constraints : one can easily check their robustness and their abundance in an image suggests a cumulative decision process (handling a single decision object : an histogram or a voting space). Our study focuses on two aspects. 1) The first one concerns the definition of a consistent methodology in which a primary process of level lines extraction is enhanced and enriched to allow the construction of more complex primitives. This construction should be guided by the image distortion model. Therefore, primitive shape is directly related to the number of variables which characterize the image transform to be determined. 2) The second one concerns a unified methodology based on a cumulative decision strategy that could be adapted to several applications of increasing complexity. We define three levels of accumulation, each associated with a particular stage of image analysis. Chosen applications to illustrate our approach are of increasing complexity : motion detection and estimation with fixed camera, image registration with mobile camera (with known transform model or objects depth constraints) and general egomotion and structure estimation from embedded cameras. Results show how a choice of robust primitives associated with a cumulative decision process allows reusing of image processing operators in several steps.

Keywords : *Computer vision, 2D and 3D motion analysis, Egomotion, Registration, Feature extraction, Level line, Cumulative decision strategy.*

1. Introduction

Les stratégies de décision cumulatives font l'objet de nombreuses études dans des domaines très divers : la résolution de problèmes en psychologie cognitive, la prise de décision (intelligence cumulative versus combinée) en finance et management ou encore en Intelligence Artificielle (accumulation d'évidences) en sont quelques illustrations. L'intérêt pour ces stratégies est tout aussi grand en fusion de données par accumulation de "décisions" issues des traitements sur différents capteurs qu'en analyse d'images où l'exemple le plus connu est sans doute la transformée de Hough dont on peut montrer aussi l'analogie avec la transformée de Radon ou le filtrage adapté en traitement du signal (Maître, 1985). C'est à Paul Hough que nous devons l'idée initiale que des points alignés "partagent" des paramètres communs (ceux de la droite les supportant) et qu'en changeant d'espace de représentation ces points sont "transformés" en droites concourantes. La transformée de Hough telle que nous la connaissons est le résultat de plusieurs contributions successives comme celle de Rosenfeld en 1969 - qui écrit les équations algébriques associées et suggère l'utilisation d'un accumulateur - ou celle de Duda et Hart - qui ajoutent un formalisme issu de la Géométrie Intégrale pour représenter les droites et englober ainsi le cas particulier des droites verticales. À partir de là, de nombreuses variantes ont été proposées aussi bien pour détecter d'autres formes paramétriques (cercles, ellipses, etc.) que pour optimiser les calculs ou la mémoire requise pour l'accumulateur. Ce sont ces différentes variantes qui nous ont inspiré lors de l'implémentation pratique de nos approches [1] :

- Pondération des votes pour tenir compte du degré de confiance accordé à certains points (O'Gorman and Clowes, 1973). Le plus souvent la confiance est calculée à partir de l'amplitude du gradient.
- Quantification non régulière des cellules. En effet, il est important de tenir compte du fait que les formes à détecter ne sont pas continues et que l'image comporte une dimension finie conduisant à une inhomogénéïté des paramètres. Certains auteurs proposent alors une quantification au maximum d'entropie permettant d'assurer des comptes égaux pour chaque cellule de l'accumulateur (Cohen and Toussaint, 1977).
- Vote à plusieurs tours (Gerig, 1987).
- Sélections des votants en utilisant l'orientation des gradients par exemple.
- Votes couplés. Cette stratégie appelée "transformation m à 1" permet de tirer profit de connaissances locales pour faire correspondre plusieurs points à un seul point de l'espace des paramètres, ce qui permet un gain important en temps de calcul. Les connaissances locales les plus utilisées sont les dérivées partielles de la courbe en chaque point (Shapiro, 1978).
- Incrémentation des accumulateurs de courbes voisines permettant de tenir compte du bruit et de l'incertitude sur la position des points "votants" (Thrift and Dunn, 1983).
- Stratégie adaptée de sélections des *maxima* dans l'espace de vote (Gorman and Sanderson, 1984).

En plus de la robustesse intrinsèque, les approches cumulatives sont simples et naturelles. En outre, la règle cumulative de base peut être modifiée de façon à intégrer les imperfections des données sous forme de fiabilité. Enfin, la construction de l'espace cumulatif est guidée par la paramétrisation déjà opérée en amont.

Nous avons montré à travers trois exemples applicatifs qu'il était possible de généraliser l'emploi de ce type d'approches à tout système exigeant des décisions robustes. Celles-ci correspondent alors à des *maxima* dans un espace cumulatif bien choisi. Toute la difficulté étant de le définir convenablement. En effet, selon le type de décision à opérer, la définition des entités à cumuler peut ne pas être très intuitive et nécessiter alors un changement d'espace de représentation des données. C'est ainsi que dans le domaine de l'analyse d'images et en particulier en vision dynamique des systèmes autonomes, nous avons proposé une méthodologie unifiée permettant de manipuler des primitives issues de l'accumulation en vue d'une décision elle même cumulative. Cette décision peut se décliner en plusieurs niveaux de cumul selon la complexité de l'application.

- Au plus bas niveau, c'est l'information binaire apparition/disparition d'une primitive dans le temps qui est cumulée. La complexité se situe strictement sur l'axe temporel. Le cumul dans le temps nous permettra ainsi de reconstruire la scène fixe et donc par soustraction du fond, l'image des objets mobiles. Les espaces de vote sont 1D et multiples, associés à chaque primitive.
- Le consensus se voudrait spatio-temporel au 2ème niveau pour identifier le mouvement : des primitives voisines dans l'image s'associent pour former des "pré-objets" contraints exhibant ainsi des invariants exploitables : leur mouvement à instancier doit être cohérent. Le cumul s'opère donc cette fois selon un modèle de mouvement de la caméra. Les primitives votent pour la transformation globale qui les aurait conduites dans leur nouvelle position. L'espace de vote est commun à toutes les primitives et multidimensionnel (une dimension par paramètre de mouvement).
- Au niveau le plus élevé, la sémantique accrue implique des hypothèses à la fois sur les primitives et sur l'origine du mouvement. Les primitives sont supposées par exemple appartenir à un même objet 3D (ex. un plan) présentant, pour un modèle de déplacement du capteur donné, une propriété caractéristique commune des vecteurs vitesse qui

1. Nous ne citerons ici que les articles les plus anciens pour chacune des variantes mentionnées.

permet de l'extraire. Notamment, leurs amplitudes sont constantes le long de courbes image prédéfinies par leurs équations analytiques. Les primitives ne votent plus selon leur structure mais selon leur vitesse. Dans le cas d'une scène 3D approximée par un ensemble de plans et d'une caméra à mouvement majoritairement longitudinal, l'espace de vote (appelé *c-velocity*) a 2 dimensions : une pour la vitesse, l'autre pour le paramètre des courbes iso-vitesse. Chaque vitesse vote pour sa courbe. Les surfaces 3D émergent dans cet espace de vote comme courbes 2D connues (droites ou paraboles).

Dans cette synthèse [2], les approches cumulatives proposées sont classées en fonction de la nature et de la dimensionnalité des espaces de cumuls. Dans les trois approches décrites dans cet article (sections 3, 4 et 5 respectivement), nous allons nous attarder systématiquement sur : la description des entités à cumuler et la stratégie de sélection de ces entités ; la définition de l'espace de cumul, ses paramètres et dimensions ; l'exploitation de cet espace pour établir une décision. Avant de décrire les processus de décision cumulatifs et leur déclinaisons, la section 2 sera consacrée aux primitives images proposées servant à alimenter ces processus de décision. Ces primitives ont l'avantage d'être robustes vis-à-vis de diverses perturbations affectant l'image et sont construites et enrichies selon les besoins en fonction du modèle de déformation des images.

2. Primitives image basées sur les lignes de niveaux

S'inspirant de la théorie d'un phénoménologue gestaltiste, Gaetano Kanizsa, des mathématiciens du CEREMADE et de l'Université des Iles Baléares ont proposé en 1999 une théorie atomique de l'image (Caselles et al., 1999) afin de répondre à une préoccupation majeure des traiteurs d'images : quelles sont les informations atomiques fiables d'où devrait partir tout algorithme d'analyse ? Cette théorie a engendré un algorithme de simplification d'image qui met en évidence sa structure occlusive : il permet de retrouver les jonctions en T ou en X et d'établir une carte topographique qui exhibe toutes les lignes de niveaux de cette image. Les propriétés pertinentes des lignes de niveaux pour l'analyse d'images ont été ainsi parfaitement étudiées. Elles se résument en trois points : 1) Les lignes de niveaux sont insensibles vis-à-vis des variations de contraste globales dont les effets sont un changement des niveaux sans effet sur l'ordre relatif [3]. Les effets des changements de contraste sur les lignes de niveaux ont été étudiés dans (Bouchafa, 1998; Monasse and Guichard, 2000)) La représentation d'une image par lignes de niveaux est complètement inversible : on sait reconstituer l'image à partir de sa représentation. 3) L'existence d'une définition mathématique unique autorise une étude rigoureuse des propriétés et une comparaison claire et sans ambiguïtés avec d'autres primitives éventuelles. A l'opposé, notons par exemple l'absence de définition mathématique de la notion de contour, dont la conséquence immédiate est la multiplicité des détecteurs - tous plus ou moins combinés avec des opérateurs de lissage - rendant la comparaison malaisée.

Notre choix s'est très tôt porté sur l'utilisation des lignes de niveaux dont nous avons montré expérimentalement la robustesse (Bouchafa, 1998). Les applications sur lesquelles nous nous sommes focalisés cette dernière décennie ne pouvaient que tirer bénéfice de ce choix initial. En effet, en vision dynamique, stéréovision, ou recalage d'images, le choix des primitives initiales est crucial : le mouvement ou le recalage ne peut être déterminé qu'en comparant les images d'une séquence. De même, l'étape d'appariement stéréoscopique se base sur la comparaison des images issues de deux caméras, fournissant chacune un point de vue différent d'une même scène. Cette comparaison doit, de ce fait, s'appuyer sur des descripteurs associés à des primitives invariantes dans le temps et l'espace (point de vue). Notre choix pour les lignes de niveaux n'a pas été remis en cause malgré l'apparition et le regain d'intérêt pour de "nouvelles" primitives (SURF, SIFT, etc.). En effet, on constatera que les critères qui nous ont amenés à faire ce choix ne sont pas automatiquement vérifiés pour ces dernières. En particulier : la non dépendance directe aux niveaux de gris (mais la prise en compte plutôt des relations d'ordre entre les niveaux) ; la distinction explicite par le processus d'extraction entre l'information "contraste" et l'information "géométrie" ou "structure" dans l'image ; la non utilisation de seuils d'extraction difficiles à ajuster et dépendants de l'image. En effet, nous avons pris le parti de bannir les seuils lors de l'extraction des lignes de niveaux. De ce fait, toutes les lignes de niveaux sont détectées afin de ne perdre aucune information utile à la compréhension des structures géométriques. Cependant, ce choix a pour conséquence directe le très grand nombre de primitives à manipuler. Ce qui aurait pu constituer un inconvénient majeur à l'utilisation des lignes de niveaux est alors tourné en avantage puisque cela autorise un processus de décision cumulatif où la taille de la population des lignes supporte l'émergence d'une décision dès que nécessaire. En revanche, nous sommes conduits à définir la notion de fiabilité d'une ligne, chacune d'elle étant ainsi utilisée à une étape adéquate de la décision.

Dans cette section, nous évoquerons le processus d'extraction des lignes de niveaux que nous avons proposé. En raison des applications considérées, nous avons fait le choix d'une extraction et d'une description

2. Cette synthèse résumé une partie du document d'Habilitation à diriger des recherches. "Vision fruste revisitée. Contribution à la vision dynamique des systèmes". Samia Bouchafa, Université Paris Sud, nov. 2011.

3. Toute fonction croissante au sens large peut être considérée comme une variation globale de contraste.

locales. Par ailleurs, nous nous intéressons aussi bien à la détection d'objets en mouvement avec caméra fixe ou mobile qu'au cas de deux caméras stéréoscopiques. Dans chacun des cas, nous proposons alors des primitives issues des lignes de niveaux adaptées :

- Dans le cas où la caméra est fixe, l'objectif est la détection du mouvement en chaque point. Nous proposons alors une extraction ponctuelle des lignes. Les descripteurs retracent les orientations locales des lignes extraites (Guichard et al., 2002).
- Dans le cas du recalage d'images (caméra en mouvement de translation planaire avec rotation autour de l'axe optique et zoom éventuel : transformation affine, homographie), le modèle de transformation étant connu et les invariants de cette transformation parfaitement définis. Nous proposons alors de définir des "segments" de lignes de niveaux (faisceau rectilignes de lignes de niveaux) et de les grouper en "V", "X", "Y" ou "Z" afin de pouvoir calculer les invariants adéquats (Almehio and Bouchafa, 2010).
- Lorsque la transformation n'est pas globale (caméra embarquée et/ou mouvement général), lorsque le déplacement d'une primitive dépend de sa profondeur (stéréovision), la primitive ne peut être que ponctuelle (la plus locale possible) mais avec toutefois une nécessité de définir des points les plus stables et robustes possibles à partir des lignes de niveaux. Nous proposons alors les "jonctions" de lignes de niveaux (Suvonvorn et al., 2005).

2.1. Un processus simple et efficace d'extraction des lignes de niveaux

Soit $I(\mathbf{p})$ l'intensité lumineuse du pixel de coordonnées $\mathbf{p}(x, y)$. L'ensemble de niveaux $\aleph_\lambda^I$ est l'ensemble de tous les points de l'image I ayant une luminosité supérieure ou égale à λ : $\aleph_\lambda^I = \{\mathbf{p} / I(\mathbf{p}) \geq \lambda\}$. Les ensembles de niveaux sont une représentation complète de l'image puisqu'il est possible de la reconstituer à partir des $\aleph_\lambda^I$. En effet, $I(\mathbf{p}) = \sup\{\lambda / I(\mathbf{p}) \geq \lambda\}$. La frontière d'un ensemble de niveaux est appelée "ligne de niveaux". La représentation par lignes de niveaux a été étudiée en détails dans (Caselles et al., 1999; Froment, 1999). Notons en particulier sa propriété d'invariance par rapport aux changements de contraste et sa commutation avec une transformation affine (translation, rotation, zoom). Caselles propose alors de considérer les lignes de niveaux comme les atomes de base sur lesquels devrait s'appuyer tout algorithme d'analyse d'images. L'extraction des lignes de niveaux est souvent réalisée à partir des ensembles de niveaux associés en effectuant simplement une série de seuillages. L'approche que nous proposons (Bouchafa, 1998; Bouchafa and Zavidovique, 2006) permet une exploitation immédiate et efficace du résultat en fournissant directement des faisceaux rectilignes de lignes de niveaux superposées. En effet, nous exploitons une propriété élémentaire des ensembles de niveaux que traduit leur inclusion : $\forall \lambda > \mu \;\; \aleph_\lambda^I \subset \aleph_\mu^I$. Par conséquent, les lignes de niveaux peuvent se superposer mais jamais se croiser. Nous proposons un simple processus récursif de suivi de lignes par groupe (se superposant), qui s'arrête lorsque le groupe cesse d'être rectiligne. On peut vérifier que la complexité totale n'est pas supérieure à celle d'une détection puis codage de lignes sur multi-seuils. De plus, notre processus n'utilise à aucun moment les niveaux de gris directement mais seulement leur ordre relatif, ce qui le rend en lui même particulièrement robuste aux variations de contraste. Il démarre en chaque point[4] $\mathbf{p}_o(x, y)$, détermine lequel parmi ses 4 voisins $\mathbf{p}_o(x, y+1)$, $\mathbf{p}_o(x-1, y)$, $\mathbf{p}_o(x, y-1)$ ou $\mathbf{p}_o(x+1, y)$ est le successeur, selon les 4 chemins possibles. Chaque successeur sélectionné devient à son tour le point courant p_k et le processus se répète jusqu'à ce que l'un des critères d'arrêt cesse d'être vérifié. Le processus d'extraction des lignes de niveau fournit directement en sortie un ensemble de **segments** de lignes de niveau, qui constituent les entités de base du groupement. Un "segment" désigne donc un groupe de lignes de niveaux rectilignes obtenu à partir de la procédure de suivi.

2.2. Fiabilité des lignes de niveaux

Les lignes de niveaux n'ont pas toutes la même importance visuelle et ne coïncident pas toujours avec les contours que nous percevons à l'oeil nu. En effet, certaines séparent des ensembles de niveaux en réalité très proches en terme de niveaux de gris. Par ailleurs, les *segments* de faible longueur sont probablement associés à des ensembles de niveaux de taille réduite générés par du bruit. Il existe plusieurs moyens de sélectionner les lignes de niveaux qui correspondent le mieux aux contours perçus (Froment, 1999; Paragios and Deriche, 1998). Nous avons choisi, dans le cadre des applications que nous traitons, de tirer profit des différences perceptuelles entre les lignes extraites afin de définir une notion de fiabilité basée sur la longueur l du *segment* et le contraste moyen c de part et d'autre de chaque segment. Les *segments* jugés peu fiables ne seront pas pour autant complètement écartés du processus de décision : ils y contribuent avec toutefois un plus faible crédit. Afin de préparer cette étape, nous avons choisi de classer les *segments* par ordre décroissant (des plus au moins fiables) du produit $l \times c$, considéré comme étant un indice de fiabilité. Par ailleurs, nous proposons une analyse plus fine de la répartition statistique de l'indice de fiabilité permettant de découper les primitives en plusieurs catégories. Ce procédé sera détaillé en section 4.1.

2.3. Configurations de segments guidées par un modèle de déformation de l'image

Lorsque le modèle de déformation de l'image est connu, le nombre de paramètres de la transformation

4. En réalité, les lignes de niveaux passent **entre** les pixels. Le référentiel est donc placé au niveau de l'espace inter pixel.

recherchée nous permet de déterminer le nombre de points caractéristiques au sein d'une même primitive nécessaires à l'estimation de la transformation. Il correspond de fait au nombre d'équations à résoudre. Ainsi, on définit des primitives pouvant être construites de manière progressive de la plus simple à la plus complexe (Almehio and Bouchafa, 2010). Le système qui les exploite est alors capable de fournir une transformation réponse quel que soit le temps de calcul alloué. Nous devons tenir compte du nombre de paramètres et des invariants de la transformation considérée. Nous focaliserons nos explications dans cette section sur le procédé de groupement proprement dit visant à former des "primitives". Une "primitive" désigne alors un ensemble de p segments configurés de manière pertinente en fonction du modèle de déformation de l'image. Dans le cas des transformations Euclidienne et de Similarité qui admettent respectivement 3 et 4 paramètres, deux points caractéristiques sont nécessaires, chaque point fournissant deux coordonnées. Ces deux points peuvent être fournis simplement par deux segments que nous proposons de grouper en primitive "angle". Un "*angle*" désigne alors un ensemble de $p = 2$ segments. Dans le cas de la transformation affine caractérisée par 6 paramètres, les segments doivent donc être groupés en triplets dont les formes sont construites par une extension de la primitive "angle". Un "Z" ou un "Y" désigne un ensemble de $p = 3$ segments groupés sous la forme de Z ou Y. Ils sont construits à partir d'un angle auquel on ajoute un segment. Enfin, dans le cas de la transformation projective caractérisée par 8 paramètres, 4 points colinéaires sont nécessaires. Ces points sont issus du groupement de 4 segments ou plus précisément d'un Z ajouté à un segment. Le "Z" étendu (en ajoutant un segment supplémentaire) donne un "W" bien défini par 4 segments. Un "W" désigne donc un ensemble de $p = 4$ segments groupés sous la forme de W. Ils sont construits à partir d'un "Z" ajouté à un segment. Le tableau 1 résume le type de primitives obtenu par groupement en fonction de la nature de la transformation recherchée. Elle donne de plus le nombre de paramètres de la transformation, justifiant ainsi le nombre de segments à grouper nécessaires pour former une primitive.

Dans les images de la Figure 1, des exemples de primitives de type "Z", "Y" et "W" extraites sont donnés. Il est à noter que le choix d'enrichir des primitives de base pour "fabriquer" des primitives plus complexes nous a naturellement conduit à grouper des segments connectés. Il aurait été envisageable de relâcher cette contrainte mais au prix d'une complexité plus importante liée à l'augmentation de la zone de recherche des segments candidats au groupement.

2.4. Extraction directe des jonctions de lignes de niveaux

Lorsque les déformations entre images sont locales ou lorsque le modèle de ces déformations n'est pas présupposé, quelle que soit la raison, nous proposons de construire des primitives générales basées sur la topologie locale dans l'image, que nous appelons "jonctions de lignes de niveaux". L'approche d'extraction de ces jonctions est basée sur les groupes de lignes de niveaux superposés précédemment définis, appelés "flux de lignes de niveaux" $F_{p,u,v}$, par souci de concision. Pour renforcer sa robustesse par rapport au bruit, un détecteur de variations d'intensités associé à un modèle appelé "EFLAM" (Extended Flow Laminating Average Milieu) est proposé. Inspiré du détecteur de Girard (1980), utilisé ensuite dans la méthode SUSAN (Smith and Brady, 1997), il est adapté pour détecter des variations d'intensité autours des jonctions et permet de ce fait de sélectionner les jonctions potentiellement fiables et stables. L'approche que nous proposons (Suvonvorn et al., 2005) pour l'extraction des jonctions de lignes de niveaux exploite alors la variation d'intensité sur un voisinage puis l'extraction des flux. Le procédé se déroule en deux étapes.

La *première étape* permet le calcul des variations des jonctions selon le modèle EFLAM. La méthode est appliquée sur tous les points, entre quatre pixels quelconques, dans l'image. Les jonctions potentielles seront formées par les variations positives et non nulles. Le degré de variation donne une indication sur la fiabilité de la jonction. Notons que dans la méthode classique d'extraction des points d'intérêt, le point potentiel est déterminé localement par la suppression des variations non maximales. Dans notre méthode, ces suppressions seront basées sur la fiabilité des jonctions, qui ne dépend pas seulement de la variation, mais aussi des caractéristiques des flux et cela joue un rôle très important dans la robustesse. En effet, nous donnons ainsi une chance aux jonctions ayant une petite variation. Le flux maximal doit être supérieur ou égal à $\mathcal{E}$.

Dans la *deuxième étape*, à chaque jonction potentielle, les intensités moyennes des régions sont déterminées, ainsi que les flux formés par la jonction. Par conséquent, le type de jonction sera déduit automatiquement. Notons que la jonction de type X peut être réduite à une jonction de type Y, et la jonction de type Y à une jonction de type L, par la suppression du plus petit flux, en fusionnant les régions des deux côtés. En général, la jonction de type X est rare et très sensible au bruit, ce qui nous amène alors à nous intéresser uniquement aux jonctions de type L et Y. Celles qui correspondent au type X doivent être déduites automatiquement du type Y. L'automate est alors conçu pour traquer ces flux (Suvonvorn et al., 2005). La figure 2 montre un exemple de résultats obtenus, de même qu'une comparaison avec le détecteur de Harris. Ces résultats soulignent un avantage important de notre méthode : nous pouvons contrôler le nombre de jonctions extraites de l'image. Cela est très important par exemple pour la reconstruction 3D ou l'analyse du mouvement. L'extraction des jonctions sera exécutée comme un processus itératif en diminuant la valeur de $\mathcal{E}$. Au début, des jonctions sont extraites avec une certaine valeur de $\mathcal{E}$. Une zone d'exclusion est définie comme un secteur circulaire autour du point de la jonction extraite. A

Transformations	Euclidienne	Similarité	Affine	Projective
Nombre de paramètres	3	4	6	8
Nombre de points caractéristiques	2	2	4	4
Primitives construites	V	V	Z ou Y	W

TABLE 1 : Groupement des lignes de niveaux en fonction des transformations.

(a) (b) (c)

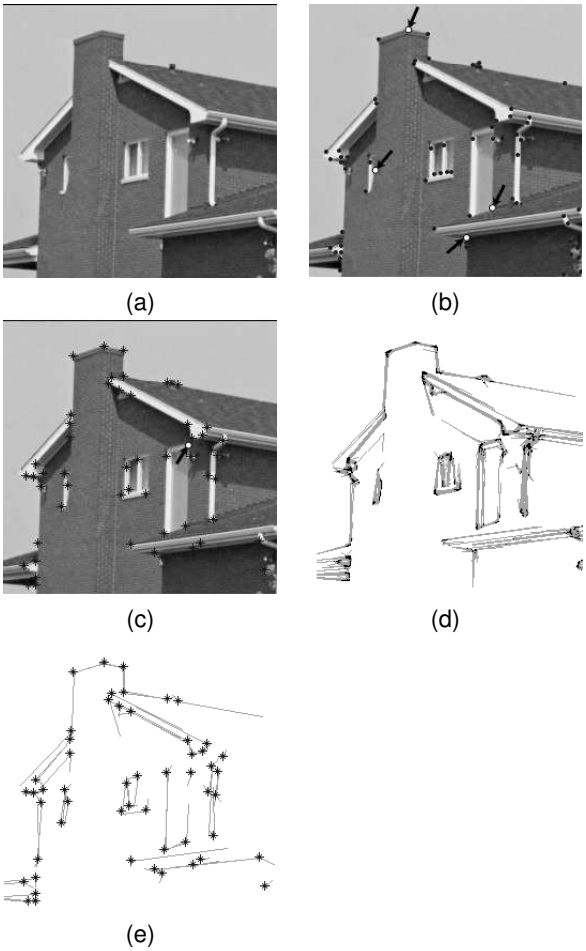
FIGURE 1 : Exemple de primitives de type Z, Y et W extraites.

l'itération suivante, la méthode d'extraction sera encore une fois exécutée en utilisant un $\mathcal{E}$ plus petit, et ajoutant ainsi de nouvelles jonctions extraites seulement dans les zones autorisées. Notons que lorsque $\mathcal{E}$ est plus petit, le nombre de jonctions augmente. Puisque la fiabilité des primitives est calculée en utilisant le flux maximal à travers la jonction $\mathcal{E}$ en respectant la longueur et la variation de contraste, alors les primitives obtenues plus tôt dans le processus sont plus fiables.

2.5. Conclusion sur l'extraction de primitives

Nous avons proposé des primitives image construites à partir des lignes de niveaux, adaptées chacune à une problématique donnée de l'analyse d'images, traitée dans le cadre de nos applications. Nous avons concentré nos efforts sur la définition d'une méthodologie cohérente dans laquelle un processus d'extraction de lignes de niveaux basique est enrichi afin de permettre la construction de primitives plus complexes. Par ailleurs, ces primitives ont d'une part suggéré la nature des processus de décision que nous avons défini par la suite (cumulatif par leur grand nombre). D'autre part, les descripteurs qui leur sont associés sont adaptés aux contraintes des dit-processus.

Au delà de cet exemple "ponctuel", nous nous sommes attachée, en détaillant le processus d'extraction, à montrer comment le choix d'une caractéristique adaptée, la ligne, et d'une procédure d'extraction adaptée, le suivi, conditionnent de manière naturelle un enchaînement logique de procédures élémentaires et de mesures associées (ex. fiabilité) renforçant la robustesse initiale dont elles tirent parti en préjugeant à leur tour des procédures d'appariement *i.e.*, du niveau décision. Il s'agit donc bien, au stade de la description statique d'image, d'un premier exemple d'exploitation d'un ensemble très restreint mais très cohérent d'objets et



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

FIGURE 2 : (a) L'image "house". (b) Les jonctions obtenues par le détecteur de Harris. (c) Les jonctions obtenues par notre méthode avec $\mathcal{E} = 30$ (flux minimum). (d) Les jonctions avant le filtrage avec $\mathcal{E} = 30$. (e) Les jonction après le filtrage avec $\mathcal{E} = 30$.

d'opérateurs au service de la perception minimale nécessaire à une classe d'actions.

3. Décision cumulative binaire

Pour détecter les objets en mouvement à partir d'une caméra fixe, une approche répandue consiste à tenter de reconstituer la scène statique, appelée "fond", à partir d'une analyse statistique ponctuelle de la variation des niveaux de gris ou de tout autre descripteur(s) de primitives. Les objets en mouvement sont alors ceux dont les niveaux (ou descripteurs de primitives) ne concordent pas avec ceux de l'image de "référence". Cette image de référence [5], doit être constamment mise à jour. Les approches s'appuyant sur la construction d'une image de référence comportent donc deux étapes : la mise à jour à chaque nouvelle image acquise de l'image de référence et la comparaison de l'image courante avec l'image de référence afin d'établir la liste des objets mobiles. La principale difficulté de ces approches est de tenir compte des diverses perturbations pouvant affecter les descripteurs choisis. La mise à jour de la référence se doit d'être insensible aux fluctuations et bruits variés, aux variations d'illuminations pouvant affecter l'image, etc.

Afin de "reconstituer" la scène statique, les approches dites "naïves" consistent à déterminer le mode, la moyenne ou la médiane temporelle d'un descripteur de primitive sur une fenêtre temporelle donnée (Lo and Velastin, 2000; Cucchiara et al., 2003). Ces techniques sont très coûteuses puisque une mémorisation des N images constituant la fenêtre temporelle est nécessaire. L'utilisation d'une moyenne exponentielle a ensuite été largement adoptée. Cette technique a été suggérée initialement par Charles C. Holt en 1957 (Holt, 2004) dans le domaine de la prévision des séries temporelles. Elle avait été proposée, sous la forme que nous lui connaissons, en 1963, par Brown (1963). Cette approche appliquée à la construction des images de références souffre non seulement de n'être adaptée qu'au cas d'une distribution statistique mono-modale mais aussi de ne pas fournir de méthodes explicites d'ajustement du seuil d'intégration à la référence. Ces dernières années, de nombreuses approches ont alors été proposées afin de tenir compte de la répartition statistique plus complexe des niveaux de gris et éventuellement de leur corrélation spatiale. Ainsi, dans (Wren et al., 1997), les auteurs utilisent un modèle statistique multi-classes pour modéliser les objets en mouvement. Le modèle de construction de la référence est Gaussien pour chaque pixel. L'approche a été employée pour les scènes d'intérieur présentant peu de variations de contraste. Stauffer and Grimson (2000) proposent une approche dans laquelle la référence est modélisée par un mélange de gaussiennes en chaque point (*mixture of Gaussians*). Dans cette approche, pour chaque point, le mélange de $K = 3, 5$ gaussiennes est effectué. Les points qui ont un modèle statistique de variation non conforme à ceux de la référence sont considérés en mouvement. Les paramètres des gaussiennes (moyenne et variances) et leur contribution dans le mélange sont les informations cumulées dans le temps. Le mélange de gaussiennes est l'approche la plus répandue actuellement. Ces deux dernières approches (Wren et al., 1997; Stauffer and Grimson, 2000) supposent une fonction densité de probabilité dont les paramètres sont obtenus durant une phase de modélisation préalable de la référence. Ces approches ne partent d'aucune hypothèse sur la distribution des données et estiment la fonction densité de probabilité afin de construire le modèle de la référence. La probabilité qu'un point n'appartienne pas à la référence est estimée grâce au modèle construit en utilisant les N précédentes images. Dans (Elgammal et al., 2002), les auteurs proposent une modélisation statistique de la référence basée sur l'estimation non paramétrique de la densité de certains noyaux, ce qui constitue une généralisation de la technique de mélanges de Gaussiennes où chaque échantillon parmi les N considérés est vu à son tour comme étant une distribution Gaussienne, cela permet d'estimer les fonctions densité de probabilité de manière plus précise en se basant sur les informations les plus récentes dans la séquence. Afin d'exploiter les corrélations spatiales, des approches sont fondées sur le principe du "mean shift" (Han et al., 2004) ou de la décomposition en valeurs singulières exploitant ainsi les matrices de co-variances (Oliver et al., 2000).

Nous constatons que la majorité des travaux basés sur la construction d'une image de référence ont majoritairement fait ce choix de considérer des primitives basiques (le point avec pour descripteur le niveau de gris ou la couleur) tout en concentrant tous les efforts sur la stratégie de sélection statistique des pixels "statiques", appartenant donc à la référence. Le lecteur aura remarqué que notre approche fait le pari inverse : nous avons concentré des effort significatifs sur la définition de primitives robustes vis-à-vis des diverses perturbations citées et nous relâchons nos efforts sur la procédure de sélection des primitives permanentes dans le temps. Notre choix des primitives robustes est tel que nous pouvons nous contenter d'un procédé simple de mise à jour par moyennage exponentiel par exemple, en adéquation totale avec l'approche intuitive qui consiste simplement à calculer l'occurrence d'un descripteur.

Dans cette section, nous décrivons de manière succincte notre procédé de mise à jour d'une référence basée sur les directions locales de lignes de niveaux. Pour plus de détails, on pourra se référer à (Guichard et al., 2000). Nous mettons l'accent sur le procédé cumulatif en explicitant la nature des entités cumulées, la dimension de l'espace de cumul et le procédé de sélection des *maxima* de cet espace. Cette section sera illustrée par des résultats ayant pour principal objectif de montrer la pertinence du choix de primitives robustes malgré un processus de mise à jour de la référence très fruste.

5. qui n'est pas forcément une image 2D au sens "collection de niveaux de gris : elle peut être un vecteur d'images si les descripteurs utilisés sont multi-dimensionnels par exemple.

3.1. Définition des entités à cumuler

On considère ici une extraction locale (sur un voisinage spatial donné) de groupes de lignes de niveaux droites passant par chaque point (Guichard et al., 2000). Du point de vue de la primitive considérée (direction locale de lignes de niveaux), seule l'information "présence" ou "absence" d'une direction en un point donné est cumulée dans le temps. Les entités à cumuler sont alors binaires.

3.2. Description de l'espace cumulatif

L'espace de cumul est défini indépendamment en chaque point, il est de dimension η, où η est le nombre de directions de lignes de niveaux au maximum en chaque point (dépendant du degré de discrétisation). L'espace de cumul total est donc de dimension $M \times N \times \eta$ pour une image de dimension $M \times N$.

Introduisons quelques notations avant d'examiner plus en détail le processus de mise à jour de la référence.

Soient :

- $\theta_0, ..., \theta_\eta$: les η directions possibles de lignes de niveaux passant en un point.
- $f_t(p, \theta_k)$: la valeur prise à l'instant t par la direction associée au point p (avec k pouvant varier de 0 à $\eta - 1$). Cette fonction, à valeur dans $\{0, 1\}$, traduit simplement la présence ou l'absence d'une direction donnée au point p.
- $F_{t \leq T}(p, \theta_k) = \sum_{t=1}^{T} f_t(p, \theta_k)$: le nombre d'occurrences (fréquence d'apparition) de la direction θ_k au point p pendant la période d'observation T.

Le moyen le plus immédiat de déterminer si une orientation locale de lignes de niveaux passant par un point est suffisamment permanente pour appartenir à la scène statique est de calculer sa fréquence d'apparition dans une fenêtre glissante. La mise à jour de la référence consiste à réactualiser cette fréquence à chaque nouvelle image acquise.

$$F_{i \leq t}(p, \theta_k) = \sum_{i=1}^{t} f_i(p, \theta_k) = F_{i \leq t-1}(p, \theta_k) + f_t(p, \theta_k),$$

ou bien, sous la forme d'un filtre récursif du premier ordre :

$$F_{i \leq t}(p, \theta_k) = m \times F_{i \leq t-1}(p, \theta_k) + (1 - m) \times f_t(p, \theta_k)$$

avec : $m = \frac{t}{t+1}$

3.3. Décision

Une orientation dont la fréquence d'apparition est supérieure à un seuil T_0 sera considérée comme appartenant au fond et sera intégrée à la référence. Ce seuil peut être choisi de manière empirique, en fonction de la nature des objets en mouvement composant les scènes considérées. En effet, la permanence des directions des lignes de niveaux produites par le passage des objets en mouvement dépend au moins de quatre paramètres : la taille des objets, leur vitesse de déplacement, leur durée d'arrêt et la fréquence d'acquisition des images. Nous pouvons partir de l'hypothèse qu'une direction est permanente si elle apparaît plus de C% de l'intervalle de temps considéré. Si f_{acq} est la fréquence d'acquisition des images (en images/seconde), et T l'intervalle de temps considéré, le nombre d'occurrences requis pour une direction donnée est alors : $T_0 = \frac{C \times T}{100} f_{\text{acq}}$. Nous pouvons aussi, si la durée de présence d'un objet en mouvement sur l'image est connue (ou estimable par exemple à partir des données champs/vitesse moyenne) et égale à d secondes, en déduire le seuil nécessaire afin de ne pas intégrer les objets en mouvement dans la référence : $T_0 > d \times f_{\text{acq}}$.

3.4. Résultats

Nous avons sélectionné ci-dessous les résultats permettant de rendre compte de la robustesse de l'approche, en particulier vis-à-vis des diverses perturbations telles que les changements de contraste (voir figures 3 et 4).

3.5. Conclusion sur la décision cumulative binaire

Dans cette section, nous avons montré comment la robustesse des primitives choisies avait pour conséquence la possibilité de mettre en oeuvre un processus de décision cumulatif élémentaire voire binaire. Nous avons choisi délibérément un processus de mise à jour du fond basique, sans aucun filtrage (aucun lissage ou préfiltrage de l'image initiale, aucun filtrage des primitives, aucun filtrage des détections *a posteriori*). La qualité des résultats obtenus encourage la généralisation à la situation d'un capteur en mouvement, ce qui constitue l'objet des sections suivantes.

4. Décision cumulative multidimensionnelle

Nous étudions dans cette section le cas où la caméra est en mouvement. Nous considérons ici que le modèle de transformation est connu, cette dernière devant être estimée et les images recalées en conséquence. Nous nous limiterons aux cas où l'estimation de la profondeur des objets n'est pas nécessaire, ce qui limite le mouvement du capteur aux transformations suivantes : euclidienne, similarité, affine et projective (homographie). Ces dernières décennies ont vu apparaître une grande variété de techniques de mise en correspondance pour le recalage d'images, chacune adaptée à l'application visée. Le nombre croissant de ces nouvelles approches a eu pour conséquence la publication de synthèses très complètes permettant de comparer - selon des critères très divers - les techniques existantes (Brown, 1992; Maintz and Viergever, 1998; Lester and Arridge, 1999; Penney et al., 1998; Anandan, 2001). Parmi les critères de distinction utilisés, nous citerons :

- La nature (fréquentielle ou spatiale) du domaine considéré. Les approches fréquentielles sont basées sur la corrélation des phases et exploitent

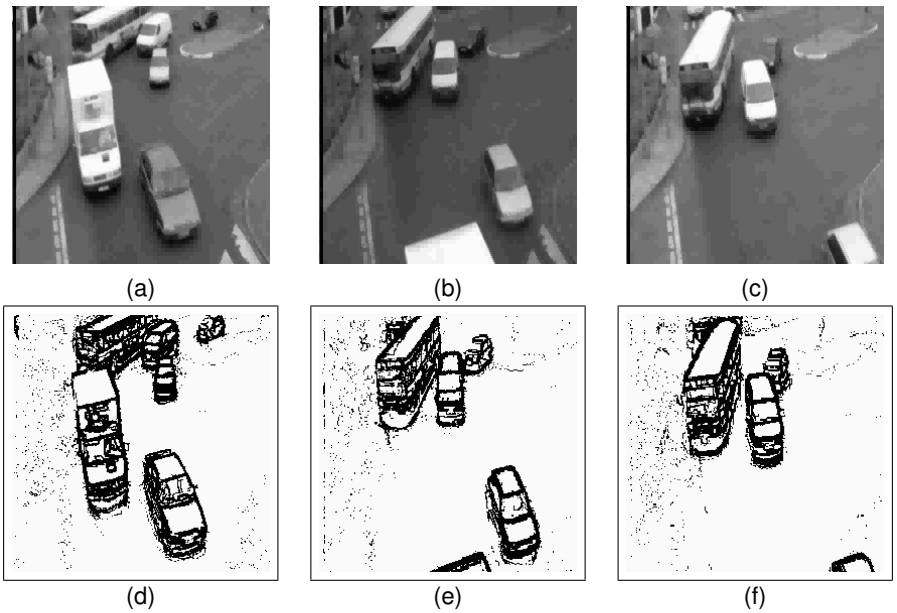


FIGURE 3 : Les images (a), (b) et (c) sont extraites d'une séquence IFSTTAR présentant des variations de contraste assez importantes. Les images (d), (e) et (f) donnent le résultat de la détection sans aucun filtrage additionnel. Seul le processus cumulatif ici sert à juger de la volatilité d'une direction de lignes de niveaux.

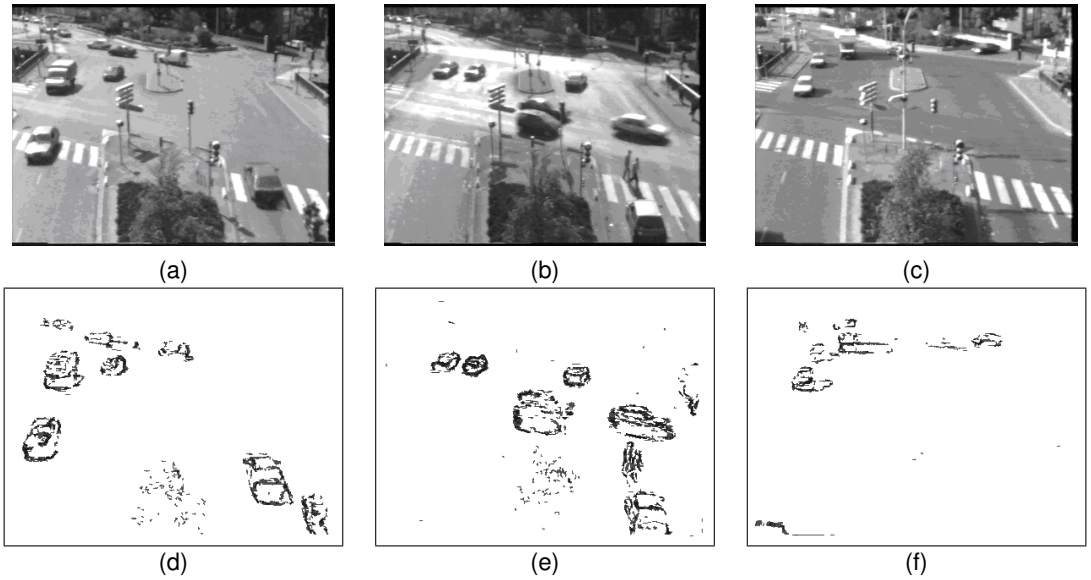


FIGURE 4 : Une image de référence d'un carrefour a été construite. Trois images extraites de la séquence IFSTTAR sont analysées. (a) et (b) ont été enregistrées la même journée pluvieuse mais à des heures différentes. (c) a été enregistrée 6 mois plus tard un jour ensoleillé. (d), (e) et (f) représentent les images détection. Nous remarquons la robustesse et la stabilité vis-à-vis des variations de contraste.

des propriétés basiques de la transformée de Fourier (ou par ondelettes) (Reddy and Chatterji, 1996). Quant aux approches spatiales, elles s'appuient sur une extraction et une comparaison de primitives image.

- La nature de la transformation recherchée (euclidienne, affine, projective ou élastique). Cette transformation dépend en général de l'application considérée. Chaque transformation particulière implique des invariants adéquats, calculés à partir des primitives extraites : longueur, orientation, rapports de longueur, birapport, etc.
- Les primitives choisies. Les plus usuelles sont les points particuliers (Moravec, 1981; Lowe, 2004; Bay et al., 2008), les contours (Nack, 1977; Medioni and Nevatia, 1984), les surfaces (Pelizzari et al., 1989), les régions (Goshtasby et al., 1986), les lignes (Stockman et al., 1982), ou les descripteurs de Fourier (Kuhl and Giardina, 1982).
- La mesure de similarité entre primitives, qui dépend aussi du type de primitives utilisées. Les plus classiques sont : la corrélation croisée (Rosenfeld and Kak, 1982), la somme des différences en valeur absolue (Barnea and Silverman, 1972) ou la distance de Levenstein (Guilloux, 1986).
- L'espace de recherche et la stratégie d'appariement. De manière simplifiée, les stratégies sont basées sur une mise en correspondance de primitives qui est explicite (ex. recherche exhaustive, programmation dynamique (Guilloux, 1986), relaxation (Shapiro and Haralick, 1990), transformée de Hough (Ballard, 1981)), ou implicite (ex. erreur quadratique, minimisation (Szeliski, 1994), programmation linéaire (Baird, 1985)).

Notre étude est partie de trois constatations :
D'abord, les images à recaler sont rarement acquises dans les mêmes conditions d'éclairement. Pourtant, la plupart des primitives utilisées dans la littérature sont sensibles aux variations de contraste. Elles sont soit directement dépendantes des niveaux de gris, ou alors souffrent d'un manque de techniques réellement indépendantes du contraste (et/ou nécessitent des seuils difficiles à ajuster) capables de les extraire (Caselles et al., 1999).
Ensuite, les motifs répétitifs - pourtant fréquents dans les images naturelles - engendrent des ambiguïtés d'appariement qui sont rarement prises en considération dans les stratégies de mise en correspondance existantes.
Enfin, dans les applications temps-réel, la robustesse du système complet tirerait profit d'un processus de décision progressif qui facilite un contrôle rapide par une première estimation des paramètres de la transformation recherchée, puis si nécessaire son raffinement.

Notre approche (Bouchafa and Zavidovique, 2006) tente alors de fournir des éléments de réponse aux problèmes cités ci-dessus. Pour cela, nous avons choisi une classe de primitives adaptées : les lignes de niveaux, robustes vis-à-vis de variations de contraste (Caselles et al., 1999; Monasse and Guichard, 2000) [6]. Nous avons utilisé les lignes de niveaux avec succès en analyse du mouvement, en particulier pour des scènes d'extérieur, là où l'hypothèse d'invariance de la luminosité n'est pas vérifiée (Guichard et al., 2000; Bouchafa, 1998). De plus, elles sont particulièrement adaptées à une stratégie d'appariement basée sur un processus de vote. En effet, le nombre important de lignes de niveaux dans l'image y entraîne de fortes redondances locales et permet d'envisager un processus de mise en correspondance cumulatif où chaque ligne est appelée au vote à une phase donnée en fonction de sa fiabilité. Pour restreindre l'espace de vote, nous exploitons une technique de décision sur graphe bi-partite dite des mariages stables (Gusfield and Irwing, 1989; Khuller et al., 1994). Chaque primitive crée une liste de préférence des primitives dans l'autre image, classées des plus aux moins similaires en fonction de mesures ou métriques adaptées. Ces préférences pondèrent le processus de décision qui se déroule en plusieurs étapes : les primitives les plus fiables participent d'abord au vote, les autres sont sollicitées ensuite pour confirmer ou infirmer le vote précédent. La transformation retenue, issue des couples de primitives, est celle qui recueille le maximum de votes.

Plus précisément, la notion de fiabilité introduite dans la section 2.2 nous amène à envisager une stratégie de vote à plusieurs tours où chaque phase permet à une nouvelle catégorie de votants d'exprimer leur opinion. La technique des mariages stables (Gusfield and Irwing, 1989; Khuller et al., 1994; Zemirli et al., 2000) suggère de construire pour chaque primitive une liste de préférence des $\kappa_{\max}$-plus proches primitives trouvées dans l'autre image. Le *couple* inter images [primitive, correspondant potentiel] permet alors d'obtenir une estimée et donc un vote pour une transformation donnée. Ce vote aura plus ou moins de crédit en fonction de la fiabilité des primitives votant et de la position du candidat dans la liste de préférence.

4.1. Définition des entités à cumuler

Nous considérons que la phase de construction des primitives par groupement de segment de lignes de niveaux a été réalisée, nous fournissant ainsi un ensemble $C^k_{I\,k=1,N}$ de vecteurs descripteurs de primitives pour la première image I et un ensemble $C^k_{J\,k=1,M}$ de vecteurs descripteurs de primitives de l'image J. Rappelons que les descripteurs sont : les invariants calculés ainsi que les contrastes moyens de part et d'autre des segments formant la primitive. Les coordonnées du point milieu de chaque segment formant la primitive seront stockés dans le vecteur caractéristique dans le but d'établir le système d'équation à résoudre pour estimer les paramètres de la transformation. Les primitives ainsi constituées vont

6. Soulignons que d'autres existent : Li et al. (2000) exploitent par exemple une variable de chromaticité par nature indépendante de la luminosité

s'associer à d'autres primitives de l'autre image formant ainsi des **couples** de primitives. En se plaçant toujours du point de vue de la primitive image extraite, celle-ci génère une liste de préférence de primitives de l'image suivante construisant ainsi des "hypothèses" de transformations inter-images. Ces hypothèses vont ou non être confirmées par le vote des autres primitives images liées par la même transformation. La décision cumulative n'est plus binaire puisqu'il ne suffit plus de vérifier la seule présence de la primitive dans l'image suivante, il faut rechercher sa nouvelle position, induisant ainsi une hypothèse de transformation. Toutes les primitives liées par la même transformation participent à l'émergence de la solution recherchée, *i.e.*, l'estimation du modèle de mouvement. Par ailleurs, en fonction de la fiabilité des primitives, celles-ci peuvent être classées en plusieurs catégories, entrainant un processus de décision à plusieurs phases.

Construction des listes de préférence à longueur variable. Nous inspirant de la technique des mariages stables, chaque primitive $u \in \left\{\overrightarrow{C_I^k}\right\}_{k=1,N}$ de l'image I construit une liste de préférence triée de primitives parmi les primitives de l'image J, des plus similaires aux moins similaires, à l'aide d'une mesure de distance euclidienne sur les descripteurs de la primitive. A titre d'exemple, pour des primitives de type "Angle", dont les descripteurs sont le rapport des longueurs, la différence des orientations et le contraste moyen de part et d'autres des lignes, la distance $dist$ entre u et $v \in \left\{\overrightarrow{C_J^k}\right\}_{k=1,M}$ est définie simplement par :

$$dist(u,v) = k_1 \left|\Delta\theta_{\mathbf{u}} - \Delta\theta_{\mathbf{v}}\right| + k_2 \left|\ell_{\mathbf{u}} - \ell_{\mathbf{v}}\right| + k_3 \left(\left|c_{i\mathbf{u}} - c_{i\mathbf{v}}\right| + \left|c_{j\mathbf{u}} - c_{j\mathbf{v}}\right|\right).$$

k_1, k_2 et k_3 permettent d' ajuster l'importance relative de chaque caractéristique. Il est à noter que $\sum k_i = 1$ et $k_3 < k_{i=1,2}$ car le contraste n'est pas une caractéristique invariante mais reste néanmoins à considérer avec un plus faible poids. Bien sûr, chaque différence en valeur absolue est normalisée par la valeur maximale prise afin d'obtenir une distance finale comprise entre 0 et 1 (de la plus à la moins ressemblante).

Le procédé est identique pour les autres types de primitives. Seul le nombre d'invariants et de segments constituant la primitive change.

Chaque primitive construit ainsi sa liste des $\kappa-$plus proches $(\kappa < \kappa_{\max})$ primitives en utilisant la distance définie ci-dessus. Il est important de noter que la taille de la liste de préférence est variable d'une primitive à l'autre. Les nouveaux candidats sont insérés dans cette liste jusqu'à ce que la distance obtenue devienne trop élevée. En fait, le choix de listes de préférences variables est nécessaire afin d'éviter l'élimination arbitraire d'un candidat en raison de tailles de listes fixées *a priori* s'avérant inadaptées *a posteriori*. Dans la pratique, nous avons fixé prophylactiquement $\kappa_{\max} = 10$, qui n'a jamais été atteint. Adapter cette borne automatiquement, si cela s'avérait nécessaire, en fonction de la distribution des *segments* (longueurs, orientations), ne présente pas de difficulté.

Classement des primitives en catégories. Les primitives sont classées en plusieurs catégories pour préparer la décision. Plus précisément, nous considérons les moyennes et écart types des longueurs $(\bar{l},\ \sigma_l)$ et contraste $(\bar{c},\ \sigma_c)$ sur toute l'image. Les primitives les plus fiables sont celles correspondant aux longueurs et contrastes les plus élevés, *i.e.,* plus grand que $\bar{l} + \sigma_l$ et $\bar{c} + \sigma_c$ respectivement. La seconde catégorie est attribuée à celles dont les longueurs sont élevées mais le contraste plus faible, *i.e.*, $(l > \bar{l} + \sigma_l$ et $\bar{c} < c < \bar{c} + \sigma_c)$. Enfin, la dernière catégorie, regroupe les primitives de faible longueur et faible contraste avec : $(\bar{l} < l < \bar{l} + \sigma_l$ et $c > \bar{c} + \sigma_c)$. Les primitives restantes sont écartées.

4.2. Description de l'espace cumulatif

L'espace de vote V est multidimensionnel : à 4 dimensions pour une transformation de similarité (dx, dy, α, ψ) (2 translations, 1 rotation et 1 échelle), à 6 dimensions pour une transformation affine $(a_{11}, a_{12}, a_{21}, a_{22}, t_x, t_y)$ et à 8 pour une transformation projective $(h_i)_{i=1,8}$. Chaque couple de primitives (u, v) vote - si le couple se préfère mutuellement [7] - en fonction de la position de v dans la liste de préférence de u pour une transformation, celle estimée par résolution d'un système d'équations dans lequel les coordonnées du point milieu des segments constituant les primitives du couple apparié apparaissent. La résolution du système peut être soit directe lorsque celui-ci est constitué de peu d'équations soit effectuée à l'aide d'une méthode de résolution aux moindres carrés.

Cas de la transformation de similarité. Le système obtenu pour un couple de primitives et 4 points (chacun fournissant 2 équations) est résolu directement de manière analytique.

Cas de la transformation affine. L'objectif est de trouver les vecteurs **h**, **k** qui minimisent les normes $\left\|\mathbf{x}' - M\mathbf{h}\right\|$ et $\left\|\mathbf{y}' - M\mathbf{k}\right\|$.

Cas de la transformation projective. Chaque primitive fournit 4 points, chaque point générant 2 équations, le système d'équations suivant qui en découle est résolu aux moindres carrés par exemple.

Quel que soit le type de transformation considéré, la contribution $\triangle V$ d'un vote pour une transformation donnée entre 2 primitives u et v est calculée en tenant compte des considérations ci-dessous. Soit $\triangle V = a_1 \times a_2 \times a_3 \times a_4$, où les paramètres a_1, a_2, a_3, a_4 sont ajustés comme suit :

7. Préférence mutuelle = "*u* est dans la liste de préférence de *v* **et** *v* est dans la liste de préférence de *u*".

1. Un correspondant potentiel v_{pos} aura une plus grande contribution s'il se trouve en début de liste de préférence et symétriquement en u. Le coefficient a_1 est alors inversement proportionnel à une fonction des positions pos_v (resp. pos_u) de la primitive v_{pos_v} (resp. u_{pos_u}). Choisissons [8] :
$$a_1 = \frac{1}{\sqrt{pos_v \times pos_u}} \; ; a_1 \in \left\{\frac{1}{\kappa}, ..., 1\right\} \; / \sup(pos_u, pos_v) \leq \kappa < \kappa_{\max}$$
2. À un tour de vote donné, les primitives de catégorie inférieure ou égale à t ($t < W_{\max}$ est l'itération courante du processus de vote) sont autorisées à s'associer à d'autres primitives de la seconde image. Au tour $i, t = i$ et les votants doivent avoir une priorité au moins de i. Les primitives de priorité supérieure vont voter avec plus de poids. a_2 est alors proportionnel à la fiabilité de la primitive :
$$a_2 = \frac{1}{w_{\mathbf{u}}}; \; a_2 = \left\{\frac{1}{W_{\max}}, ..., 1\right\}$$
3. Les primitives les plus ressemblantes (distance faible) votent avec un poids plus grand :
$$a_3 = (1 - dist(\mathbf{u}, \mathbf{v})) \; ; a_3 \in [0...1]$$
4. Enfin, puisque tous les $a_i \; i = 1, 3$ appartiennent à l'intervalle $[0, 1]$, nous pouvons utiliser un facteur d'échelle a_4 afin d'obtenir des incréments entiers.

4.3. Décision

L'estimation de la transformation recherchée est réalisée à travers un processus de vote à plusieurs tours (l'approche proposée est schématisée dans la figure 5). À chaque tour de vote, une nouvelle catégorie de *couples* exprime son opinion (une estimée de la transformation recherchée) en fonction de sa fiabilité. Cette stratégie a deux avantages :

1. La disponibilité à chaque tour d'une approximation même grossière de la transformation recherchée peut être - même si elle n'est pas très précise - exploitée dans certaines applications nécessitant un résultat immédiat.
2. En raison des motifs répétitifs présents dans les images naturelles, conduisant fatalement à des ambiguïtés d'appariement, le processus qui consiste à inviter au vote des primitives moins fiables - mais souvent plus nombreuses - donne une nouvelle chance à l'émergence de la solution exacte.

Les *couples* sont classés en autant de catégories que de tours de votes en divisant simplement la liste de primitives. Au premier tour, la première catégorie vote. Si un pic unique apparaît dans l'espace de vote alors une transformation obtient une majorité absolue : le processus de vote peut donc s'arrêter. Sinon, un second tour peut démarrer autorisant une nouvelle population à voter : les primitives ayant une fiabilité moindre. S pics dans l'espace de vote sont sélectionnés. Les valeurs cumulatives des *maxima* sélectionnés sont amplifiées pour le tour suivant afin de privilégier les décisions des primitives les plus fiables. Le vote s'arrête lorsqu'il n'y a plus de primitives ou lorsqu'un pic émerge, récoltant ainsi la majorité absolue.

8. D'autres fonctions symétriques ou non de u et v peuvent être testées en regard de l'application.

4.4. Résultats

Transformation de similarité. Dans le cadre d'un travail sur le recalage d'images microscopiques (Bouchafa and Zavidovique, 2002) en collaboration avec MINASYS (Institut d'Électronique Fondamentale, Université Paris Sud XI), nous avons comparé notre approche avec 2 techniques classiques : la première est basée sur une corrélation de phases, la seconde sur une minimisation de distance entre intensités (méthode d'optimisation de Levenberg-Marquardt). Aucune n'a donné les résultats escomptés en particulier sur les images contenant de nombreuses répétitions. Nous avons aussi testé une démonstration web en ligne basée sur une approche différentielle (Fedorov et al., 2002). Là encore, la transformation obtenue n'est pas correcte, voir figure 6f. L'algorithme que nous proposons parvient à trouver le bon recalage : voir figure 6g. Nous l'avons aussi évalué sur des images particulièrement difficiles à traiter en raison des variations de contraste et des nombreuse répétitions créant des ambiguïtés (Bouchafa and Zavidovique, 2004).

Transformation Affine. La figure 7 montre quelques résultats obtenus lorsque l'hypothèse d'un mouvement affine était choisie. L'existence d'un plan perpendiculaire à l'axe optique, majoritaire dans l'image, rend cette hypothèse de transformation quasi valide dans les cas choisis pour cet exemple. Le recalage puis la superposition des images montrent l'erreur de recalage plus importante sur les zones ne correspondant pas à ce plan servant de référence au recalage.

Transformation projective. La figure 8 donne quelques résultats dans le cas d'une transformation projective. L'existence d'un plan majoritaire dans l'image (bâtiment) a pour conséquence, à travers le choix d'un seul pic dans l'espace de vote, une estimation de l'homographie associée à ce plan de référence.

4.5. Conclusion sur la décision cumulative multidimensionnelle

L'approche proposée pour recaler des images partant d'un *a priori* sur le modèle de la transformation recherchée a été conçue au départ pour faciliter la réutilisabilité des primitives construites dans l'éventualité où le modèle de transformation se complexifierait. C'est dans cet esprit que nous comptons étendre cette approche à la prise en compte graduelle des transformations de la moins à la plus complexe. Il serait envisageable de mettre en place un processus multi-phases, permettant

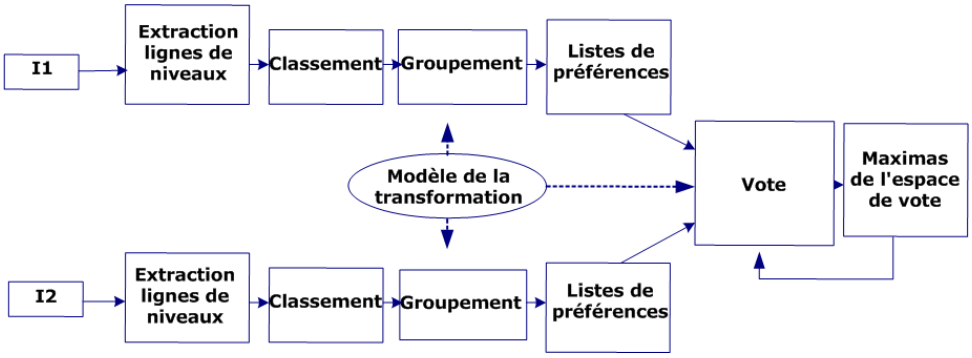


FIGURE 5 : Schéma général du processus de vote.

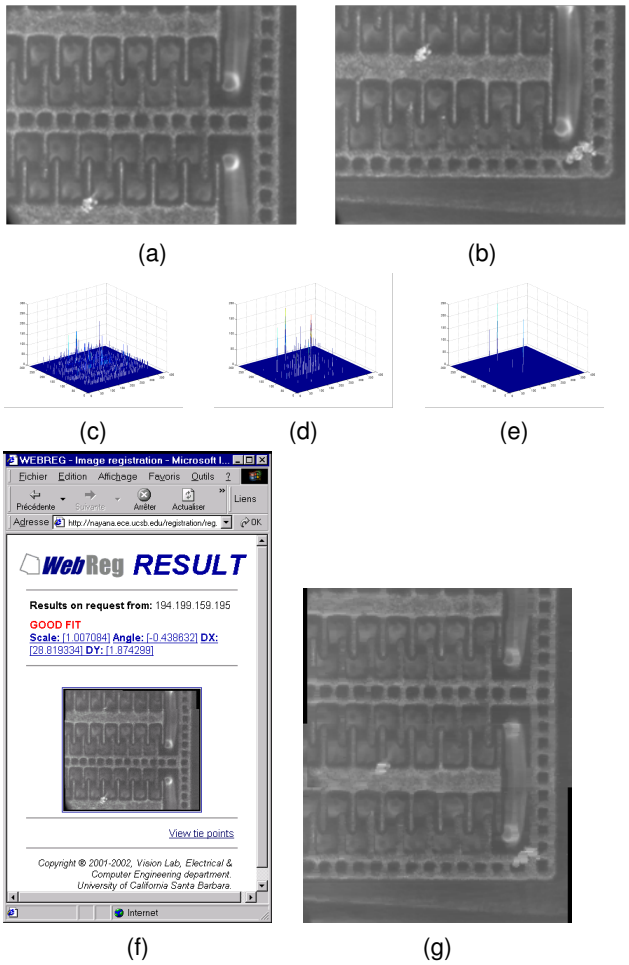


FIGURE 6 : (a), (b) Images obtenues par un microscope électronique à balayage de franges qui correspondent à des profils partiels d'un peigne électrostatique en cuivre d'un micro gyromètre. (c) à (e) Les espaces de vote en translation pour 3 tours de vote. (f) Résultats obtenus en utilisant une approche classique différentielle basée sur une minimisation de distance d'intensités. En raison de la présence de motifs répétitifs dans les images, la transformation obtenue n'est pas correcte. (g) La transformation correcte estimée (translation de -98,3^o, rotation = 0, échelle = 1) malgré les motifs répétitifs. Ce résultat est très satisfaisant pour le recalage d'images de micro-dispositifs obtenues par un microscope électronique dont le champ est limité.

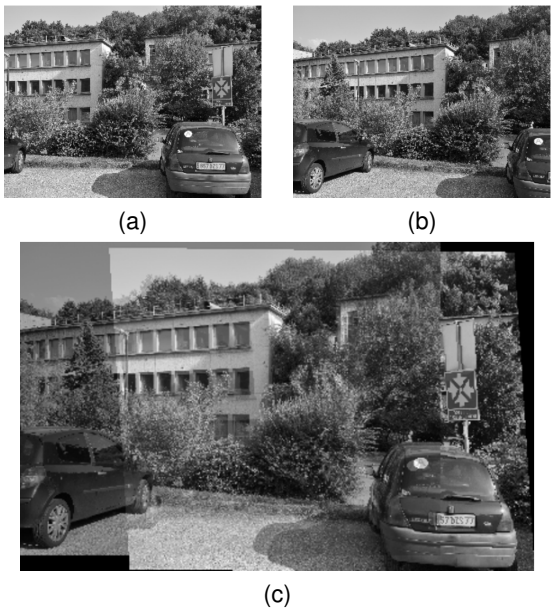


FIGURE 7 : (a) et (b) Deux images successives d'une séquence. (c) Résultat de l'alignement en supposant un modèle de transformation affine.

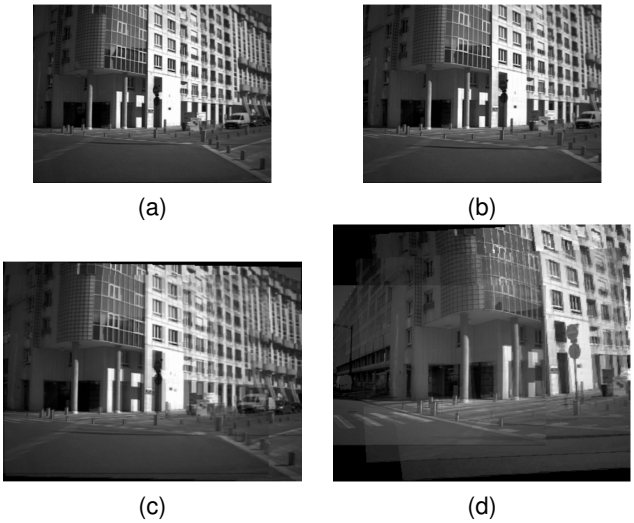


FIGURE 8 : (c) Alignement des deux images (a) et (b) obtenues à partir d'une caméra en mouvement. La scène est essentiellement composée de plans, ce qui rend l'hypothèse de transformation projective valide. (d) Alignements multiples.

d'estimer d'abord les transformations de similarité (modèle le plus simple) associées à des zones différentes de l'image à travers l'extraction de plusieurs pics dans l'espace de vote. Les zones n'ayant pas voté pour les pics choisis sont alors probablement déformées par un autre modèle plus complexe : affine, puis en utilisant le même procédé, projectif. La phase suivante consiste alors à enrichir les primitives construites pour qu'elles soient adaptées au nouveau modèle à estimer. Et ainsi de suite. Cette extension a plusieurs avantages : d'abord l'estimation de plusieurs transformations pour un même modèle mais associées à différentes zones de l'image. Ensuite, l'estimation de plusieurs modèles de transformations pouvant cohabiter au sein d'une même image. Pour résumer : dans le cas d'une même transformation globale liant les primitives, le multi-tour (au niveau du vote) permet de fournir la transformation majoritaire (le plan qui "saute au yeux") puis les suivantes par ordre d'importance visuelle, le multi-phase permettra de gérer plusieurs modèle de transformations au sein d'une même image.

5. Décision cumulative 2D en cascade

Cette section décrit notre contribution au problème de la navigation d'un véhicule autonome doté d'un système de vision. La problématique est largement étudiée dans les domaines de la robotique et des systèmes d'aide à la conduite. En effet, la navigation sécurisée requiert au minimum une détection des obstacles potentiels (fixes ou mobiles) ainsi qu'une reconstitution de l'environnement même fruste. Afin d'accomplir ces tâches, il est acquis que le véhicule ou le robot devra se munir d'une multitude de capteurs (extéroceptifs tels que Radar et Lidar ou proprioceptifs tels que accéléromètres, gyromètres ou odomètres). Malheureusement, la plupart de ces capteurs fournissent des données entachées d'erreur et imprécises et peuvent être ponctuellement défaillants entrainant ainsi des données manquantes. Une idée naturelle est donc de faire coopérer ces capteurs, d'où l'essor considérable des techniques de fusion de données multi-capteurs (Laneurit et al., 2003; Comport et al., 2010). Parmi tous les capteurs envisagées, la "caméra" a sans aucun doute une place particulière : malgré la complexité des processus de vision, la richesse des informations fournies ainsi que son coût en font un moyen privilégié, concentrant ainsi les efforts d'une grande partie de la communauté scientifique dans ces domaines (Dickmanns, 2002). La profusion de publications récentes sur ces sujets en témoigne [9]. Le manque de recul immédiat, la multiplicité et la variété des approches proposées entraine la rareté voire l'inexistance de synthèses complètes récentes de techniques existantes. Sans être exhaustive, nous tenterons une première classification permettant de situer notre approche.

5.1. Etat de l'art

Approches 2D (x, y), basées "modèle". L'image 2D issue d'un capteur contient à elle seule un certain nombre d'informations sur l'environnement perçu. En particulier si l'objectif est d'éviter des obstacles et que ceux ci sont visuellement discriminants – dans le sens où l'on peut définir des signatures / attributs visuels les distinguant d'autres "objets" de l'environnement – alors il est possible de construire des modèles permettant de les détecter. Plusieurs approches ont été proposées, exploitant la symétrie (Bensrhair et al., 2001), la texture (Kalinke et al., 1998), ou la couleur (Buluswar and Draper, 1998). Ces approches assez performantes pour des obstacles de type "véhicule" peinent à détecter ou "reconnaitre" un piéton dont la variabilité et les déformations rendent la tâche plus complexe. C'est tout naturellement que des techniques basées sur de la classification et de la reconnaissance ont été adoptées. Parmi les approches proposées, citons : les techniques de classifications binaires en cascade (Gama and Brazdil, 2000) reposant sur l'utilisation d'un classifieur, généralement linéaire ; l'utilisation des Machines à Vecteurs de Support (Vapnik, 1999) ou le recours à des réseaux de neurones (Munder and Gravila, 2006). L'avantage de ces méthodes est leur capacité à travailler sur des espaces de grande dimension. Citons par ailleurs les techniques de boosting qui consistent à agglomérer plusieurs classifieurs faibles en un classifieur fort. Les résultats obtenus par le classifieur fort sont alors supérieurs à ceux obtenus par chaque classifieur faible. Les méthodes couramment utilisées sont dérivées de la méthode AdaBoost (Hastie et al., 2005). L'utilisation d'un tel classifieur repose sur la définition d'une

9. Des revues et conférences entières sont dédiées à ce sujet : ITSC, IV, IROS, IEEE Trans on Intelligent Transportation Systems, etc.

base de représentation dans laquelle projeter les imagettes sur lesquelles vont s'effectuer les tâches d'entraînement et de détection. Une solution extrêmement populaire pour la détection de piétons est l'utilisation d'histogrammes de gradients orientés (Histograms of Oriented Gradients, HOG) (Dalal and Triggs, 2005). D'autres descripteurs ont été proposés, comme les Joint Ranking of Granules (Huang and Nevatia, 2010), la décomposition en ondelettes de Haar (Viola et al., 2005), ou encore une analyse en composantes principales (Turk and Pentland, 1991). Ces méthodes de détection reposent également sur une phase d'apprentissage hors ligne. Cet apprentissage se déroule en présentant au classifieur des populations de négatifs et de positifs. La représentativité des bases d'apprentissage va alors fortement conditionner l'aptitude du classifieur à discerner différents objets.

Approches 3D (x, y, z), basées "structure". Ces approches se basent sur l'estimation d'informations structurelles caractérisant les obstacles potentiels en exploitant une deuxième caméra qui assure ainsi un processus de vision stéréoscopique estimant la profondeur z des objets perçus. Dans certains travaux, les obstacles sont considérés comme étant des plans fronto-parallèles aisément détectables en particulier si le système de vision stéréoscopique est parfaitement calibré et rectifié. Dans (Labayrade et al., 2002) un espace dans lequel les plans fronto-parallèles sont transformés en droites est défini. Celles-ci sont ensuite extraites à l'aide d'une transformée de Hough. Nous reviendront largement sur cette technique qui a inspiré les travaux de cette section. Au delà de la simple détection de plans fronto-parallèles, la construction de cartes, ou plus précisément de grilles d'occupation connaît un certain succès (Vu et al., 2008; Nedevschi et al., 2009). L'un des principaux intérêts de cette approche est que la collaboration entre plusieurs capteurs est alors immédiate et de type "tableau noir" (*i.e.*, accumulation naturelle pour partage). 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 fournira des hypothèses de détection que la vision viendra ensuite confirmer (Rodriguez et al., 2010). 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 manoeuvrer (Soquet et al., 2007).

Approches 3D (x, y, t), basées "mouvement". Avant de situer notre travail parmi les approches existantes et afin de mettre en évidence les différents points de vue adoptés, formalisons la problématique de la navigation d'un véhicule doté d'une seule caméra embarquée, le formalisme peut être facilement généralisé à plusieurs caméras.

Considérons le système de coordonnées $OXYZ$ calé au centre optique d'une caméra. L'axe OZ coïncide avec l'axe optique. Si l'on considère un mouvement rigide du capteur, caractérisé par sa vitesse translationnelle instantanée $\mathbf{T} = (T_X, T_Y, T_Z)$ et sa vitesse rotationnelle instantanée $\Omega = (\Omega_X, \Omega_Y, \Omega_Z)$. Chaque point $\mathbf{P} = (X, Y, Z)$ appartenant à la scène statique est doté d'un mouvement relatif $\mathbf{V} = -\mathbf{T} - \Omega \times \mathbf{P}$. Si l'on considère que la projection [10] du point $\mathbf{P} = (X, Y, Z)$ dans le plan image est $p = (x, y, z)$, que la distance focale est f alors la vitesse 2D (u, v) en chaque point de l'image est :

$$\begin{cases} u = \frac{xy}{f}\Omega_X - \left(\frac{x^2}{f} + 1\right)\Omega_Y + y\Omega_Z - \frac{fT_X + xT_Z}{Z} \\ v = -\frac{xy}{f}\Omega_Y - \left(\frac{y^2}{f} + 1\right)\Omega_X - x\Omega_Z - \frac{fT_Y + yT_Z}{Z} \end{cases} \quad (1)$$

L'examen de ces équations permet plusieurs constatations :

- Le mouvement 2D dépend de la profondeur ;
- Seule la composante translationnelle du mouvement dépend de la profondeur ;
- Toute discontinuité de mouvement 2D ne peut être due qu'à une variation de profondeur ;
- Le mouvement ne peut être déterminé qu'à un facteur d'échelle près. Un objet situé à une distance Z, se translatant de T produira le même mouvement 2D qu'un objet situé à une distance $2Z$, se déplaçant en translation de $2T$.

Devant ces constatations et la non linéarité des équations, certaines approches partent d'un modèle simplifié de projection (Aloimonos, 1990). Sous les hypothèses de longueur focale importante et d'objets proches de l'axe optique, le modèle orthographique (ou perspective faible) considère que la projection 3D→2D se résume en une projection orthogonale suivie d'un changement d'échelle. Dans le cas où les objets sont éloignés de la caméra ou de petite dimension par rapport à la distance au centre de projection, l'emploi de la projection para-perspective permet aussi de simplifier les équations (Poelman and Kanade, 1992). Sans passer par des hypothèses sur la nature de la scène, il est possible aussi de faire usage de la projection sphérique bien adaptée à la représentation des champs de vecteurs.

A côté de cette tentative de simplifier les équations à travers le modèle de projection, certains travaux préfèrent simplifier le modèle du mouvement lui même, en considérant par exemple un véhicule mobile en translation longitudinale majoritaire pure ou en introduisant un ou deux angles de rotation (souvent le lacet pour la prise en compte des virages) (Bailey and Durrant-Whyte,

10. On considère ici un modèle projectif simple (modèle sténopé).

2006; Scaramuzza et al., 2009). D'autres approches tentent non pas de limiter le nombre de degrés de liberté mais de séparer l'estimation des translations (dépendant de la profondeur) des rotations en exploitant la parallaxe que crée ce mouvement (*motion parallax*, *affine motion parallax*, *plane+parallax*). Ces méthodes exploitent le fait qu'aux discontinuïtés de profondeur, il est possible de distinguer les effets de la rotation de ceux de la translation de la caméra. En particulier pour les approches type "Plane+parallax", le mouvement 2D d'une région de l'image où les variations en profondeur ne sont pas significatives permet de supprimer les effets de la rotation de la caméra. A partir de ce mouvement de parallaxe résiduel obtenu, la translation peut être facilement calculée (Irani et al., 1997; Hanes et al., 2008). Dans le même esprit de simplification du modèle de mouvement, citons les approches se concentrant sur l'estimation du Foyer d'Expansion dont on sait que les coordonnées regroupent les informations liées au mouvement translationnel (Xu and Danielson, 1992; Sazbon et al., 2004; Wu et al., 2007).

Ces dernières années se sont multipliés les travaux sur l'estimation du mouvement 3D (*egomotion*) d'une caméra embarquée sur véhicule mobile et sur la reconstitution de la profondeur de la scène observée (*Structure From Motion*). Il s'en est suivi de nombreuses classifications des méthodes existantes selon divers critères. La classification couramment adoptée distingue 3 catégories principales : les approches discrètes, continues et directes.

- Les approches **discrètes** se basent sur la mise en correspondance de primitives image et sur une expression matricielle, incluant toutes les inconnues (paramètres de mouvement et paramètres intrinsèques de la caméra), reliant les points de correspondance. Le problème se ramène alors à un problème d'algèbre linéaire et les nombreuses approches existantes diffèrent dans le choix des méthodes de résolution adoptées, plus ou moins sensibles aux perturbations des données (Longuet-Higgins and Prazdny, 1980; Hartley, 1995; Luong and Faugeras, 1997).
- Les approches **continues** exploitent le flot optique calculé. La relation entre le flot optique et le mouvement préalablement paramétré permet - par des techniques d'optimisation - d'estimer les paramètres du mouvement ainsi que la profondeur en chaque point. Les résultats obtenus sont alors dépendants de la qualité du flot optique calculé (Negahdaripour and Horn, 1989; Hildreth, 1992; MacLean et al., 1994).
- Dans les approches **directes**, le mouvement est déterminé "directement" à partir de la contrainte d'invariance de la luminosité d'un point au cours de son déplacement sans avoir à calculer explicitement le flot optique. Les paramètres du mouvement sont alors déduits par des approches d'optimisation classiques. Les approches **continues** et **directes** sont toutes deux globales et récoltent des informations sur toute l'image. Leur avantage par rapport aux approches discrètes provient alors du nombre important de données traitées, contribuant ainsi à réduire les erreurs (Irani et al., 1997; Stein et al., 2000).

Approches 4D (x, y, z, t), basées coopération "structure / mouvement". 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. Parmi les travaux les plus significatifs, citons ceux décrits dans (Heinrich, 2002) 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. Par ailleurs, le principe de 6D-Vision, avancé dans (Franke et al., 2005) 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. 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 (Dornaika and Chung, 2000), ou simplement enrichi de l'information temporelle (Braillon et al., 2008; Leibe et al., 2007). Les approches recevant le plus grand intérêt de la communauté sont cependant celles 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 (Lenz et al., 2011) ou d'intégrer la stéréo dans une méthode de calcul du flot optique type Horn & Schunk (Pons et al., 2007; Wedel et al., 2008). A partir de ce champ de correspondances, des techniques de segmentations classiques peuvent être utilisées pour obtenir une représentation de la scène en fonction du mouvement apparent des objets.

5.2. Notre contribution

Nous partons de l'hypothèse qu'une navigation sécurisée du véhicule requiert au moins un étiquetage fruste de la scène. Les objets sont étiquetés en fonction de leur nature en relation avec leur structure (surface horizontale = route ; surface verticale = bâtiment ; surface frontale = obstacle) et en fonction de leur mouvement (mouvement conforme à l'ego-mouvement du véhicule = objet statique ; mouvement non conforme à l'ego-mouvement du véhicule = objets ayant un mouvement indépendant). Nous considérons par ailleurs, que toutes les possibilités de la vision monoculaire, pourtant économe en moyens, n'ont pas encore été exploitées, ce qui classe notre approche parmi les approches 3D (x, y, t) permettant de

fournir une estimation fruste de la structure. Partant de cet objectif, deux études dans la littérature ont retenu notre attention.

- Dans la première (Fermuller and Aloimonos, 1995), les vecteurs vitesse d'amplitude et d'orientation données sont contraints d'appartenir à des courbes dans l'image dont les paramètres dépendent des paramètres du mouvement 3D du capteur. En particulier, si les vecteurs considérés sont issus d'un flot optique ou d'un champ de disparité alors on peut montrer que ces vecteurs sont contraints d'appartenir à des sections de coniques pouvant être définies. En étudiant les propriétés de ces courbes, une estimation de l'ego-mouvement peut être réalisée.
- Dans la seconde étude (Labayrade et al., 2002), les auteurs proposent, en stéréovision, une technique très efficace basée sur le concept de *v-disparité* qui consiste à exploiter la relation entre disparité et lignes image dans le cas particulier où les images stéréoscopiques sont rectifiées. Un nouvel espace de projection cumulatif – l'espace appelé *v-disparité* – composé des histogrammes de disparité sur toutes les lignes image, permet de mettre en évidence la relation de proportionnalité entre disparité et lignes image dans le cas particulier d'un plan horizontal.

Ces deux études peuvent être mises en parallèle d'une manière intéressante : toutes deux exploitent des courbes d'iso-valeurs –vitesse pour l'une, disparité pour l'autre. Notre point de vue est que ce procédé, basé d'une part sur la définition d'iso-courbes et s'appuyant d'autre part sur des statistiques le long de ces courbes, peut être étendu. Nous avons donc naturellement pensé à sa généralisation au cas de la vision monoculaire.

Cette section décrit l'approche proposée. Des surfaces paramétrées peuvent être détectées sans aucune calibration *a priori* de la caméra, ni aucune connaissance de l'ego-mouvement du véhicule. Dans un premier temps, dans le but de contourner l'estimation de la profondeur des objets (non connue), nous partons de l'hypothèse que la scène 3D peut être approximée par un ensemble de plans. Ces plans seront alors détectés et étiquetés en fonction de leur mouvement en exploitant les courbes d'iso-vitesse 2D, les vitesses pouvant être estimées par un flot optique quelconque. Pour tenir compte des imprécisions liées à l'estimation du flot optique et parce que nous sommes convaincue de la robustesse des techniques cumulatives dans ce contexte applicatif, nous définissons un espace cumulatif appelé *c-vélocité* par analogie à la *v-disparité*.

5.3. Définition des entités à cumuler

Dans l'approche que nous proposons, l'estimation des vecteurs vitesse par une méthode de flot optique dense engendre une population de votants de taille conséquente rendant le processus de décision cumulatif représentatif et pertinent. Pour un modèle de mouvement et une structure de nature donnés (un plan d'orientation donnée et situé à une distance donnée) les "primitives" sont associées à une vitesse. Celle-ci se trouve sur une courbe d'iso-vitesse renforçant l'hypothèse du mouvement et de la structure conjecturés. L'espace de cumul est bi-dimensionnel (paramètre de courbe d'iso-vitesse / norme de vitesse) puis monodimensionnel car les vitesses le long de ces courbes sont non seulement constantes pour des structures définies mais elles sont aussi liées par une relation linéaire. Un deuxième espace cumulatif 1D est alors défini permettant d'extraire du premier espace les relations de proportionnalité (droite) exhibées.

Dans (Labayrade et al., 2002), les auteurs prouvent que pour un plan horizontal, le long d'une ligne image, issue d'un couple d'images stéréoscopiques rectifiées, la disparité est constante et varie linéairement en fonction de la profondeur et donc des lignes image. Le plan de la route est alors détecté dans l'espace *v-disparité*, construit en accumulant les disparités le long des lignes image. La route se projette dans cet espace en une droite. Le procédé a été généralisé aussi aux plans verticaux en considérant les colonnes image et définissant par analogie, l'espace *u-disparité*.

Nous montrons dans ce qui suit comment le procédé peut être généralisé à la vitesse (déplacement en analogie à la disparité), cumulée le long de courbes d'iso-vitesse (en analogie aux lignes et colonnes image).

Cas d'un point 3D en mouvement. Considérons dans un premier temps un mouvement translationnel le long de l'axe Z (le véhicule avance). Nous verrons par la suite comment le généraliser à d'autres types de mouvements. En se référant à l'équation du mouvement d'un point (Equation 1) et en posant $\Omega_X = \Omega_Y = \Omega_Z = T_X = T_Y = 0$, la vitesse 2D (u, v) devient :

$$\begin{cases} u = \frac{T_Z}{Z} x \\ v = \frac{T_Z}{Z} y \end{cases} \quad (2)$$

Les équations 2 décrivent le mouvement 2D d'un point (projection dans l'image du mouvement 3D) qui ne doit pas être confondu avec le flot optique. Nous partirons de l'hypothèse communément admise que le flot optique est une approximation correcte du mouvement 2D.

Considérant l'équation 2, la relation entre la vitesse $\|\mathbf{w}\|$ (analogie avec la disparité) et la fonction iso-vitesse c (analogie avec les index ligne v) devient :

$$\|\mathbf{w}\| = \sqrt{u^2 + v^2} = \left| \frac{T_Z}{Z} \right| \sqrt{x^2 + y^2} = K \cdot C(x, y) \quad (3)$$

$$\frac{\|\mathbf{w}\|}{K} = C(x, y) = c \quad (4)$$

Prenons pour illustrer le concept le cas où la profondeur Z est constante en tout point (*i.e.*, le véhicule s'approche d'un mur par exemple), la translation T_Z étant

celle de la caméra, identique pour tous les points statiques. De ce fait, K, défini par $\left|\frac{T_Z}{Z}\right|$ dans l'équation (3) est constant, et les courbes d'iso-vitesse $C(x,y)$ sont des cercles. De plus, le rayon de ces cercles (paramètre des courbes d'iso-vitesse) c varie linéairement avec la norme de la vitesse $\|\mathbf{w}\|$ (voir equation (4)). A côté de ce cas particulier (profondeur constante) décrit ici pour illustrer la démarche, dans le cas général, Z peut être simplement éliminé en considérant des surfaces planaires permettant ainsi d'exploiter la relation entre Z et (X,Y) pour éliminer Z.

Cas d'un plan 3D en mouvement. Supposons à présent que la caméra observe une surface plane d'équation : $\mathbf{n}^T\mathbf{P} = d$, avec $\mathbf{n} = (n_X, n_Y, n_Z)$ le vecteur unitaire normal à la surface, d la distance "plan/origine" et P le point de coordonnées (X,Y,Z). A partir de l'équation (1) et de $Z = \frac{1}{n_Z}(d - n_X X - n_Y Y)$, la vitesse 2D s'écrit alors (Longuet-Higgins and Prazdny, 1980; Verri and Poggio, 1989) :

$$\begin{cases} u = \frac{1}{fd}\left(a_1x^2 + a_2xy + a_3fx + a_4fy + a_5f^2\right) \\ v = \frac{1}{fd}\left(a_1xy + a_2y^2 + a_6fy + a_7fx + a_8f^2\right) \end{cases} \quad (5)$$

$$\begin{aligned} a_1 &= -d\Omega_Y + T_Z n_X \\ a_2 &= d\Omega_X + T_Z n_Y \\ a_3 &= T_Z n_Z - T_X n_X \\ a_4 &= d\Omega_Z - T_X n_Y \\ a_5 &= -d\Omega_Y - T_X n_Z \\ a_6 &= T_Z n_Z - T_Y n_Y \\ a_7 &= -d\Omega_Z - T_Y n_X \\ a_8 &= d\Omega_X - T_Y n_Z \end{aligned}$$

Étudions précisément 4 cas particuliers de plans en mouvement adaptés à l'application visée :

a) Horizontal (route)
b) Latéral (bâtiment)
c) Frontal1 (obstacle fuyant ou approchant)
d) Frontal2 (obstacle traversant)

Le tableau 2 liste, pour chaque cas, le vecteur normal unitaire associé au plan $\mathbf{n}$, le vecteur translation 3D $\mathbf{T}$ et la distance plan-origine d. Le mouvement caméra est supposé translationnel $\mathbf{T} = (0,0,T_Z)$. En conséquence, dans l'équation (5), $T_X, T_Y, \Omega_X, \Omega_Y$ et Ω_Z sont annulés (à l'exception de l'obstacle fuyant/approchant possédant sont propre modèle de mouvement $(0,0,T_Z^o)$ ou $(T_X^o,0,0)$ qui s'ajoute à celui de la caméra $\mathbf{T}$ pour donner : $\mathbf{T}'$).

Les vecteurs vitesse correspondants sont ainsi obtenus en injectant respectivement $\mathbf{T}$ (ou $\mathbf{T}'$) et $\mathbf{n}$ dans a_i for $i = 1,...,8$. L'équation (5) donne alors u et v comme listé dans le tableau 3 pour chaque cas.

Soient $\|\mathbf{w}_o\|$, $\|\mathbf{w}_r\|$, et $\|\mathbf{w}_b\|$, amplitudes des vitesses associées respectivement à un obstacle, une route et un bâtiment. Regroupons les paramètres du mouvement 3D et la focale inconnus dans le paramètre K que nous ne chercherons pas à estimer ici ; son seul intérêt à ce stade est d'être constant.

	$\mathbf{n}$	mouvement 3D	Dist. plan-origine
a)	$(0,1,0)$	$\mathbf{T} = (0,0,T_Z)$	d_r
b)	$(1,0,0)$	$\mathbf{T} = (0,0,T_Z)$	d_b
c)	$(0,0,1)$	$\mathbf{T}' = (0,0,T_Z^o + T_Z)$	d_o
d)	$(0,0,1)$	$\mathbf{T}' = (T_X^o,0,T_Z)$	d_o

TABLE 2 : Paramètres de 4 types de plans.

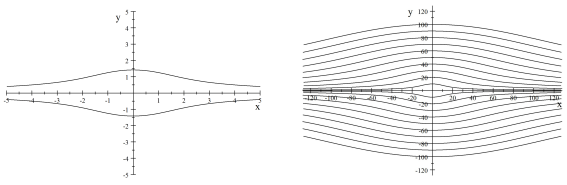
a)	$u = \frac{T_Z}{fd_r}xy$ $v = \frac{T_Z}{fd_r}y^2$	$\|\mathbf{w}_r\| = K\sqrt{y^4 + x^2y^2}$
b)	$u = \frac{T_Z}{fd_b}x^2$ $v = \frac{T_Z}{fd_b}xy$	$\|\mathbf{w}_b\| = K\sqrt{x^4 + x^2y^2}$
c)	$u = \frac{T_Z+T_Z^o}{d_o}x$ $v = \frac{T_Z+T_Z^o}{fd_o}y$	$\|\mathbf{w}_o\| = K\sqrt{x^2 + y^2}$
d)	$u = \frac{T_Z}{d_o}x - \frac{T_X^o f}{d_o}$ $v = \frac{T_Z}{d_o}y$	$\|\mathbf{w}_o\| = \begin{cases} Ksi T_X^o \ggg T_Z \\ K\sqrt{x^2+y^2} sinon \end{cases}$

TABLE 3 : Vecteurs vitesse associés à 4 types de plan.

Mise en évidence de la relation de proportionnalité $c/\|\mathbf{w}\|$. Chaque type de $\|\mathbf{w}\|$ conduit à une expression de c et de ce fait à des courbes d'iso-vitesse différentes résumées dans le tableau 3, de a) à d). Par souci de clarté, nous emploierons le terme $c-courbe$ pour "courbe d'iso-module de vitesse" et le terme $c-valeur$ pour la valeur $c(x,y)$ le long d'une $c-courbe$. En particulier, dans le cas b) correspondant à un plan bâtiment,

$$c = \frac{\|\mathbf{w}\|}{K} = \sqrt{x^4 + x^2y^2} \quad (6)$$

La figure 9a illustre les $c-courbes$ pour une valeur c_0 donnée de c, pour un "plan horizontal" : cas a) dans le tableau 3). Chaque courbe est l'ensemble des pixels dont l'amplitude de vitesse est constante, $\mathbf{w} = Kc_0$, si et seulement si les points appartiennent bien à l'image d'un plan horizontal. Par conséquent, la relation précédente (6) prouve que c, constante le long d'une courbe d'iso-vitesse par définition, est proportionnelle à $\|\mathbf{w}\|$. En réalité, un plan horizontal intersecte dans l'image la famille de ces courbes obtenues en variant c : voir figure (9 (b) où sont affichées les courbes en incrémentant c de 10.



(a) (b)

FIGURE 9 : (a) Un couple de courbes pour une valeur de c donnée dans le cas "plan horizontal". L'accumulation est faite le long de ces courbes qui ne s'intersectent pas en théorie. (b) Un ensemble de $c-courbes$ pour un pas de variation égal à 10.

Rectifications des $c-courbes$. A chaque point $\mathbf{p} = (x, y)$ dans l'image, est associée une valeur c en fonction du modèle de plan choisi. Cette valeur peut être calculée hors ligne une seule fois à l'initialisation puisqu'elle ne dépend que de (x, y). Par ailleurs, il est possible pour faciliter l'implémentation de l'approche et par analogie à la rectification d'images stéréo, de calculer la transformation permettant de redresser les courbes d'iso-vitesse afin de les rendre parallèles aux lignes ou au colonnes image en fonction du modèle considéré. Les images résultantes ne sont autres que $I(c, y)$ pour les modèles route et obstacle et $I(x, c)$ pour le modèle bâtiment. Le détail des calculs est donné dans (Bouchafa and Zavidovique, 2012).

Extraction du Foyer d'expansion. L'origine du repère image – servant notamment au calcul des $c - valeurs$ – est l'intersection du plan de projection et de la direction de translation. Ce point particulier appelé Foyer d'Expansion (FoE) doit être déterminé le plus précisément possible : son estimation est donc une étape clé du processus visuel décrit ici. Dans le cas d'un mouvement translationnel, chaque vecteur vitesse se dirige vers le FoE (Equation 2).

$$\frac{u}{v} = \frac{x}{y} \tag{7}$$

Supposons que (x_0, y_0) sont les coordonnées du FoE dans l'image. Sachant que l'origine du repère image est placé en haut à gauche, alors les équations de l'egomouvement (u, v) translationnel deviennent :

$$\begin{cases} u = \frac{T_Z}{Z}(y_0 - y) \\ v = \frac{T_Z}{Z}(x - x_0) \end{cases} \tag{8}$$

et

$$\theta = \tan^{-1}\left(\frac{v}{u}\right) = \tan^{-1}\left(\frac{x - x_0}{y_0 - y}\right) \tag{9}$$

Cette relation montre que le FoE peut être extrait simplement en déterminant l'intersection des droites supportant les vecteurs vitesse. Plusieurs approches ont été proposées dans la littérature (Negahdaripour and Horn, 1989; Sazbon et al., 2004). Nous avons choisi de maintenir une cohérence avec le point de vue "cumulatif" pour des raisons incluant notamment la réutilisation de modules (Bouchafa and Zavidovique, 2006). Dans l'approche implémentée, les vecteurs vitesse votent pour tous les points appartenant à la droite les supportant. Le FoE résultant est alors le point qui recueille le maximum de votes, c'est-à-dire l'intersection de toutes les droites supportant les vecteurs vitesse (figure 11b). Dans ces images : en rouge sont représentés les points qui recueillent le maximum de votes ; l'intersection des deux droites en blanc donne la position du FoE estimé ; le taux de votant pour chaque point est représenté par une couleur allant du noir (aucun vote) au rouge (nombre de votants maximum) ; les taux intermédiaires sont respectivement colorés en bleu, cyan, vert et jaune.

La position du FoE peut confirmer ou non l'hypothèse de mouvement translationnel du capteur. Par exemple, si le FoE n'est pas au centre de l'image, les images peuvent être rectifiées pour compenser les effets éventuels des roulis ou tangages de la caméra (Bouchafa and Zavidovique, 2011). Notons par ailleurs que l'équation (7) peut aussi servir à ajouter des contraintes additionnelles au moment du vote dans l'espace *c-vélocité*.

5.4. Description de l'espace cumulatif *c-vélocité*

Nous avons montré que la vitesse 2D est constante le long des courbes d'iso-vitesse (les $c - courbes$), caractérisées chacune par un paramètre constant c, correspondant au modèle de plan 3D adéquat.

Si l'on considère toutes les perturbations pouvant affecter les amplitudes des vitesses estimées, il est peu probable d'obtenir des vitesses constantes le long de ces courbes. Nous proposons donc de considérer le mode à travers une analyse de l'histogramme des $\|\mathbf{w}\|$ le long des $c-courbes$. En collectant ainsi tous les modes le long de toutes les $c - courbes$, la relation de proportionnalité entre $\|\mathbf{w}\|$ et c est mise en évidence. Par conséquent, l'espace *c-vélocité*, **bidimensionnel** en $(c, \|\mathbf{w}\|)$, est cumulatif : il est construit en affectant à chaque pixel (x, y) la valeur c ($c - valeur$) correspondant au modèle de plan choisi –a), b), c) ou d) dans le tableau 3– et en incrémentant la valeur au coordonnées $(c, \|\mathbf{w}\|)$, où $\mathbf{w}$ est la vitesse estimée en (x, y). Nous avons choisi dans nos expérimentations d'estimer $\mathbf{w}$ par une approche classique de calcul de flot optique, en l'occurrence celle proposée par Lucas & Kanade (Lucas and Kanade, 1981).

Considérations numériques. Une étude de la fonction $c(x, y)$ pour chaque modèle de plan –en particulier pour le modèle "route" et le modèle "bâtiment"– nous amène aux conclusions suivantes : d'abord, chacune de ces courbes intersecte l'axe des x (pour le modèle route) ou l'axe y (pour le modèle bâtiment) dans le plan image aux coordonnées : $x = \pm\sqrt{c}$ ou $y = \pm\sqrt{c}$, respectivement. Par ailleurs, pour une taille d'images standard, l'intervalle de variation de c est très grand, en l'occurrence égal à 128000 (pour le modèle route) et à 96000 (pour le modèle bâtiment) pour une image de taille 320×240. Par conséquent, pour des raisons d'implémentation, de complexité autant que d'homogénéité, nous choisissons pour ces deux modèles de plans de considérer plutôt la relation entre $\|\mathbf{w}\|$ et $\sqrt{c}$[11]. Un plan est alors représenté dans l'espace *c-vélocité* par une parabole au lieu d'une droite.

11. D'autres fonctions servant à réduire la dynamique peuvent être employées.

5.5. Décision

Nous avons montré comment un plan 3D était représenté dans l'espace *c-vélocité* par une parabole. La détection de ces paraboles dans l'espace *c-vélocité* pour leur retro-projection dans l'image, nous permet de déterminer effectivement les plans 3D associés. La détection des paraboles dans un espace 2D peut se faire de différentes manières ; nous avons encore une fois privilégié le choix d'une méthode cumulative. C'est donc naturellement que nous sommes tournée vers la transformée de Hough. Ici, elle sera 1D car les paraboles passent par l'origine du repère (les $c - courbes$ sont définies à partir de l'origine *i.e.*, le FoE). L'espace de Hough 1D est construit en cumulant le paramètre p de chaque parabole, *i.e.*, la distance p entre chaque parabole et son foyer ou sa directrice (Duda and Hart, 1972). L'ensemble du processus obéit aussi à l'expression fonctionnelle suivante, témoin de la réutilisation de procédé et de l'économie des moyens :

$$(x, y, \|\mathbf{w}\|) \rightarrow (c, \|\mathbf{w}\|, P(c, \|\mathbf{w}\|))$$
$$\rightarrow$$
$$[p(c, \|\mathbf{w}\|), \sum_{[(c,\|\mathbf{w}\|),p(c,\|\mathbf{w}\|)]} P(c, \|\mathbf{w}\|)] \; [i.e.(p(c, \|\mathbf{w}\|), P(p))]$$

où P est la probabilité et p le paramètre de la parabole. L'histogramme 1D obtenu est alors segmenté par une méthode de clustering quelconque, ici un K-means classique (MacKay, 2003).

$$\|\mathbf{w}\| = K\left(\sqrt{c}\right)^2 \Rightarrow p = \frac{1}{4K} = \frac{\left(\sqrt{c}\right)^2}{4\|\mathbf{w}\|} \qquad (10)$$

En théorie, chaque plan 3D correspond à une parabole (une valeur donnée du paramètre p) dans l'espace *c-vélocité*. Il est évident que plusieurs perturbations concourent à transformer une parabole (un Dirac dan l'espace de Hough 1D) en une patatoïde parabolique (une Gaussienne dans l'espace de Hough 1D). Nous les avons étudiées en détail dans (Bouchafa and Zavidovique, 2011). C'est la raison pour laquelle l'espace de Hough 1D doit être segmenté en clusters. On peut objecter que l'utilisation des K-means impose de fixer le nombre de classes *a priori*. Nous pouvions aisement nous orienter vers des méthodes de clustering non supervisées. Cependant, dans le cadre des applications traitées et à nouveau en conformité avec la vision fruste, le nombre de clusters peut être fixé *a priori*. On peut imaginer que le nombre de plans est limité par la structure et la nature de la scène. En particulier, dans les scènes urbaines nous pouvons dénombrer 4 types de plans : 1) horizontal (une route), 2) verticaux (2 bâtiments de part et d'autre du véhicules et éventuellement 2 plans correspondant aux plans des voitures garées de part et d'autre de la route, ce qui donne 4 classes dans cette catégorie), 3) des obstacles fronto traversants et 4) des obstacles frontaux fuyants/approchants . Il faut être conscient des conséquences d'un tel *a priori* :

- Supposons que le nombre de plans est supérieur au nombre de classes fixées dans le K-means clustering. Certains classes d'orientations proches vont donc être fusionnées. La question est alors de savoir si cela entraîne des conséquences sur la précision pour la navigation d'un véhicule ? (ex : distance entre le véhicule et les plans latéraux). Cette question est abordée plus précisément dans (Bouchafa and Zavidovique, 2011).
- Peut-on, indépendamment du nombre de plans, prévoir une correction *a posteriori* ? Nous pouvons certainement envisager de reboucler sur le résultat du K-means en affinant le clustering en fonction de la vraisemblance du résultat. Par exemple dans l'image de la figure 11, prévoir une classe supplémentaire permet de séparer le plan "lampadaire" du plan "voitures garées à droite".
- Ce problème est un dilemme classique entre un clustering non supervisé plus complexe et un plus élémentaire avec connaissance *a priori*. Dans cette étude, nous avons privilégié le second choix, le but étant davantage de valider d'abord le principe de la méthode.

Le schéma de la figure 10 résume les différentes étapes de l'approche proposée.

5.6. Résultats de la détection des plans

5.6.1. Images réelles

Données et paramètres en entrée. Toutes les séquences d'images ci-dessous sont issues de la base de données constituée dans le cadre du projet ANR LOVe (Logiciel d'Observation des VulnérablEs). Ces séquences sont très variées et ont été fournies par les constructeurs automobiles partenaires du projet. L'approche proposée requiert en entrée un flot optique. Nous avons choisi la méthode classique de Lucas and Kanade (1981) avec 9×9 comme taille de fenêtre d'analyse. Nous avons de manière délibérée choisi une approche d'estimation du flot optique la plus basique et classique qui soit, car notre objectif est, rappelons le, de confirmer le maintien de la robustesse de la méthode de détection de plan quelle que soit la qualité du flot en entrée. Les résultats présentés ci-dessous se focalisent sur la détection de plans latéraux qui cadrent bien avec les scènes urbaines considérées et l'application de conduite automatique.

Deux paramètres ont été introduits. Le premier est le taux minimum de votants dans une cellule de l'espace *c-vélocité*. Ce taux minimum est calculé en fonction du nombre maximum de points le long de courbes c. En raison de la discrétisation des courbes, celles-ci peuvent en effet avoir un nombre variable d'éléments. Le second paramètre est le nombre minimum de points votant pour un paramètre de parabole donné p dans l'espace de Hough 1D : trivialement le seuil est ici égal à 3 points. Par ailleurs, afin d'évaluer quantitativement notre approche, deux facteurs de confiance ont été définis. Le

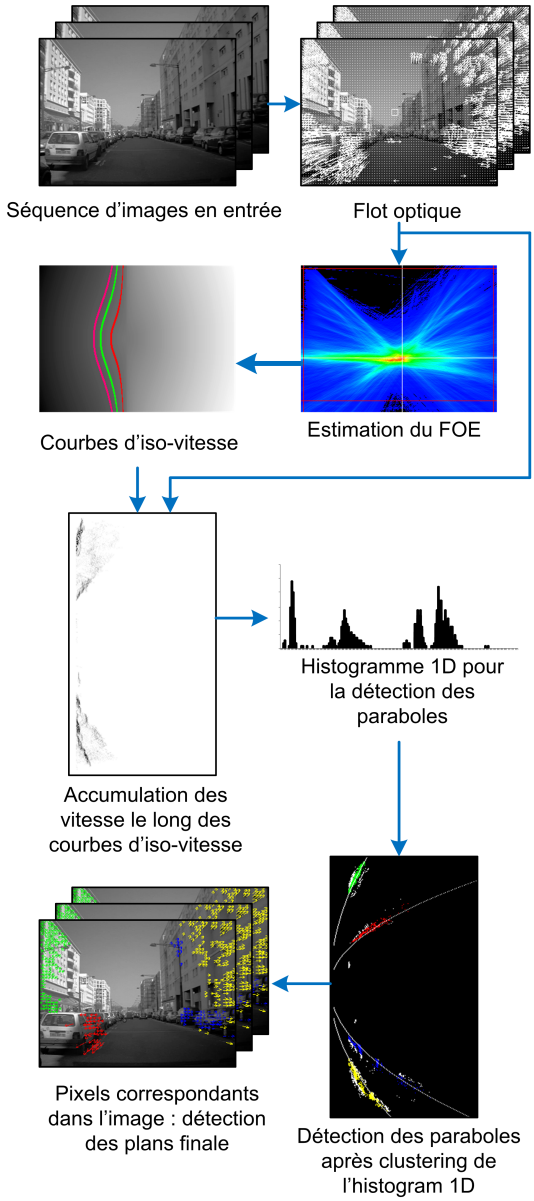


FIGURE 10 : Schéma récapitulatif de l'approche proposée.

premier est lié à l'hypothèse de mouvement translationnel : il s'agit de la différence Δ_{foe} entre la position du FoE trouvé et le centre de l'image. Si Δ_{foe} est grand, le véhicule n'est pas en translation pure. Le second est l'écart type σ des classes trouvées après K-means. Un σ faible ajouté à un pic élevé dans l'espace de Hough 1D confirme la détection d'un plan. En effet, le nombre de points constituant le plan est alors le nombre de votants.

Résultats. Les exemples de la figure 11 illustrent les images de flot optique (images étiquetées "a") échantillonnées par souci de lisibilité, la position du FoE (images "b") et l'espace *c-vélocité* "bâtiment" associé à différentes séquences d'images (images "c"). Les résultats de la détection des paraboles sont donnés dans les images étiquetées "d" avec les plans correspondants dans les images "e".

Interprétation. Dans toutes les séquences considérées, le cercle noir est le centre de l'image. Δ_{foe} est sa distance par rapport au FoE. Dans les images 1 et 2, on peux dénombrer six plans 3D : 2 plans "bâtiment", 2 plans correspondant aux voitures garées de part et d'autre de la route, un plan frontal traversant (une moto) et un plan "route". Dans cette séquence, la plupart des vecteurs vitesse estimés sont situés sur les plans latéraux qui sont plus larges et texturés. Dans les espaces *c-vélocité* "bâtiment" dans (d.1) et (d.2), comme attendu, 4 paraboles correspondant aux 4 plans latéraux sont nettement visibles. Les plans en (e.1) et (e.2) sont étiquetés en fonction du résultat des K-means. La même étiquette (et couleur) est utilisée pour afficher les points correspondants dans l'image. Les points écartés sont en blanc. L'image 3 montre un exemple où la caméra n'est pas en mouvement ; des piétons traversent la rue ; l'espace *c-vélocité* correspondant (c.3) fait apparaître une verticale (vitesse constante).

D'autres expérimentations ont été réalisées en faisant varier le nombre de plans dans l'image, lorsque la caméra est en rotation autour de l'axe Y (virage), lorsque un camion par exemple traverse la route ou lorsqu'un véhicule se déplace avec approximativement le même mouvement relatif. Tous les résultats sont disponibles dans (Bouchafa and Zavidovique, 2012). Dans une séquence de 2300 images, dans laquelle les images 1 et 2 (Figures 11) ont été extraites, où la caméra est en translation longitudinale, où la scène est composée de 4 plans latéraux dans 500 images de cette séquence, le taux de détection des paraboles obtenu est alors de **87%** durant cette phase.

L'analyse de la robustesse de l'algorithme a été détaillée dans (Bouchafa and Zavidovique, 2012). Nous avons étudié de manière précise les effets de différentes perturbations pouvant affecter le processus de détection des surfaces planes. Les cinq sources d'imprécisions ou d'erreur ci-dessous ont été considérées une à une :

- Les approximations numériques dues notamment aux différentes discrétisations réalisées ;

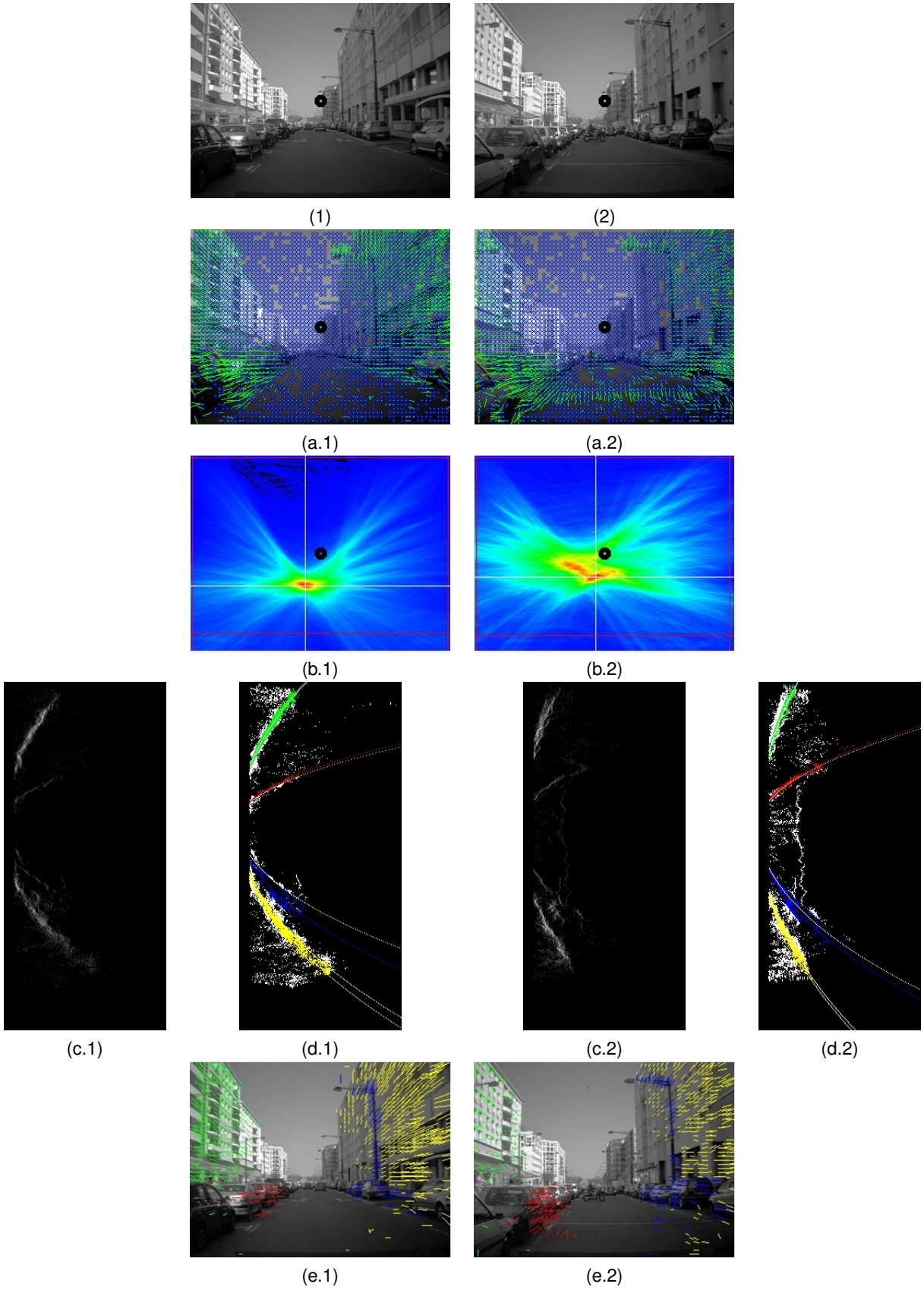


FIGURE 11 : Quelques résultats typiques. Image (1) correspond au cas où la caméra est en translation rectiligne. Image (2) correspond au cas où une moto traverse la rue. Dans l'image (3), le véhicule s'arrête. Pour chaque image, les images (b) donnent le résultat de la détection du FoE, les images (c) l'espace *c-vélocité* "bâtiment" , les images (d) l'espace *c-vélocité* après K-means clustering et les images (e) la détection de plans latéraux finale.

- Un flot optique en entrée bruité ;
- Un axe optique de caméra non nécessairement parallèle au plan de la route ;
- Une erreur dans l'estimation de la position du FoE ;
- Les contaminations diverses des votes de pixels appartenant à d'autres types de plans que ceux considérés (contamination inter-modèles).

L'étude qualitative et quantitative des sources d'incertitude ainsi que les expérimentations réalisées ont confirmé la robustesse du procédé, dans la limite des hypothèses initiales posées. Mais il serait intéressant à présent de considérer des mouvements plus généraux incluant des rotations (virage, suspension des roues, etc.) ou des translations latérales (glissement des roues, etc.). Les premiers calculs effectués dans ce sens montrent que le procédé peut aisément être généralisé à condition de définir des espaces de vote en cascade dédiés chacun à une composante du mouvement.

5.7. Conclusion sur la décision cumulative 2D en cascade

La décision cumulative en cascade présentée dans cette section fournit une détection de plans 3D par cumul de vecteurs vitesse 2D. Deux espaces de vote sont traités de manière séquentielle : l'espace *c-vélocité* et l'espace de Hough 1D découlant de la paramétrisation des paraboles. Les perspectives immédiates apparaissent : la généralisation à des surfaces paramétrées plus complexes qui augmentera probablement la dimension de l'espace de Hough initialement 1D et la prise en compte de mouvements plus généraux incluant les deux autres translations ainsi que les rotations. Un premier examen des équations du mouvement initiales suggère que l'espace *c-vélocité* sera lui même décomposé en plusieurs espaces en cascade chacun associé à un paramètre du mouvement. Ceci fera l'objet à court terme d'une étude détaillée.

6. Conclusion

Nous avons montré comment un choix de primitives robustes (construites à partir des lignes de niveaux) associé à un processus de décision cumulatif (pouvant se décliner en plusieurs niveaux de cumuls) permettait la réutilisabilité des opérateurs dans plusieurs secteurs. Les systèmes proposés ont la particularité d'être compacts et cohérents, propriétés recherchées dans les applications considérées. Nous avons pu traiter différentes phases de la perception dynamique : le cas où la caméra est fixe et les objets mobiles, le cas où la caméra est en mouvement avec connaissance *a priori* sur le modèle de ce mouvement et le cas plus général de l'estimation de la structure à partir du mouvement. Chacune des alternatives traitées ouvre une multitude d'extensions et de perspectives.

Références

Almehio, Y., Bouchafa, S., 2010. Matching images using invariant level-line primitives under projective transformation. In : IEEE Canadian Conference on Computer and Robot Vision. Otawa, Canada, pp. 130–135.

Aloimonos, J. Y., 1990. Perspective approximations. Image and Vision Computing 8 (3), 177–192.

Anandan, P., 2001. Video registration and motion estimation : a retrospective. In : Conference on Recent Advances in 3D Digital Imaging and Modeling, Quebec, Canada.

Bailey, T., Durrant-Whyte, H., 2006. Simultaneous localization and mapping (SLAM). Robotics & Automation Magazine 13 (3), 108–117.

Baird, H., 1985. Model-based image matching using location.

Ballard, D. H., 1981. Generalizing the Hough transform to detect arbitrary shapes. Pattern recognition 13 (2), 111–122.

Barnea, D., Silverman, H., 1972. A class of algorithms for fast digital image registration. IEEE Transactions on Computers 21 (2), 179–186.

Bay, H., Ess, A., Tuytelaars, T., Gool, L. V., 2008. Surf : Speeded up robust features. Computer Vision and Image Understanding 110 (3), 346–359.

Bensrhair, A., Bertozzi, M., Broggi, A., Miche, P., Mousser, S., Toulminet, G., 2001. A cooperative approach to vision-based vehicle detection. In : IEEE Intelligent Transportation Systems, Oakland, USA. pp. 207–212.

Bouchafa, S., 1998. Détection du mouvement insensible aux variations de contraste. application à la détection de comportements de foules anormaux dans le métro par traitement d'images. PhD thesis, Université Paris VI, Paris, France.

Bouchafa, S., Zavidovique, B., 2002. Efficient line voting for mems profile registration. In : Indian Conference on Computer Vision Graphics and Image Processing. Ahmedabab, India.

Bouchafa, S., Zavidovique, B., 2004. Stratégie de vote pour la mise en correspondance de lignes de niveaux. In : Reconnaissance des Formes et Intelligence Artificielle. Vol. 2. pp. 605–614.

Bouchafa, S., Zavidovique, B., 2006. Efficient cumulative matching for image registration. Image and vision computing 24 (1), 70–79.

Bouchafa, S., Zavidovique, B., 2011. Robustness of c-velocity methods for 3d plane detection. Pattern recognition and Image Analysis 21 (2), 233–237.

Bouchafa, S., Zavidovique, B., 2012. C-velocity : a flow-cumulating uncalibrated approach for 3d plane detection. International Journal of Computer Vision 97 (2).

Braillon, C., Pradalier, C., Usher, K., Crowley, J., Laugier, C., 2008. Occupancy grids from stereo and optical flow data. Experimental Robotics 39, 367–376.

Brown, L., 1992. A survey of image registration techniques. ACM computing surveys 24 (4), 325–376.

Brown, R. G., 1963. Smoothing Forecasting and Prediction of Discrete Time Series. Englewood Cliffs, NJ : Prentice-Hall.

Buluswar, S. D., Draper, B. A., 1998. Color machine vision for autonomous vehicles. International Journal for Engineering Applications of Artificial Intelligence 11 (2), 245–256.

Caselles, V., Coll, B., Morel, J., 1999. Topographic maps and local contrast change in natural images. International Journal of Computer Vision 33 (1), 5–27.

Cohen, M., Toussaint, G., 1977. On the detection of structures in noisy pictures. Pattern recognition 9, 95–98.

Comport, A. I., Malis, E., Rives, P., 2010. Real-time quadrifocal visual odometry. International Journal of Robotic Research, Special Issue on Robotic Vision, 486–492.

Cucchiara, R., Grana, C., Piccardi, M., Prati, A., 2003. Detecting

moving objects, ghosts and shadows in video streams. IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10), 1337–1342.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. In : Computer Vision and Pattern Recognition, San Diego, USA.

Dickmanns, E., 2002. The development of machine vision for road vehicles in the last decade. In : IEEE Intelligent Vehicles Symposium, Versailles, France.

Dornaika, F., Chung, R., 2000. Cooperative stereo-motion : Matching and reconstruction. Computer Vision and Image Understanding 79 (3), 408–427.

Duda, R., Hart, P. E., 1972. Use of the hough transformation to detect lines and curves in pictures. Comm. ACM 15, 11–15.

Elgammal, A., Duraiswami, R., Harwood, D., Davis, L., 2002. Background and foreground modeling using non-parametric kernel density estimation for visual surveillance. Proc. of the IEEE 90, 1151–1163.

Fedorov, D., Fonseca, L., Kenney, C., Manjunath, B., 2002. Automatic registration and mosaicking system for remotely sensed imagery. In : SPIE 9th International Symposium on Remote Sensing, Agia Pelagia, Crete, Greece.

Fermuller, C., Aloimonos, Y., 1995. Global rigidity constraints in image displacement fields. In : International Conference on Computer Vision, Boston, USA. pp. 245–250.

Franke, U., Rabe, C., Badino, H., Gehrig, S., 2005. 6d-vision : Fusion of stereo and motion for robust environment perception. Lecture Notes in Computer Science 3663, 216–223.

Froment, J., 1999. A compact and multiscale image model based on level sets. In : Scape Space Theories in Computer Vision. LNCS 1682, pp. 152–163.

Gama, J., Brazdil, P., 2000. Cascade generalization. Machine Learning 41, 315–343.

Gerig, G., 1987. Linking image-space and accumulator space. In : International Conference on Computer Vision, London, UK. pp. 112–117.

Girard, M., 1980. Digital target tracking. In : NATO Group.

Gorman, L., Sanderson, A. C., 1984. The converging squares algorithm : and efficient method for locating peaks in multidimensions. IEEE Transactions on Pattern Analysis Machine Intelligence 6 (3), 280–288.

Goshtasby, A., Stockman, G., Page, C., 1986. A region-based approach to digital image registration with subpixel accuracy. IEEE Transactions on Geosciences and Remote sensing 24 (3), 390–399.

Guichard, F., Bouchafa, S., Aubert, D., 2000. A change detector based on level sets. In : International Symposium on Mathematical Morphology. Palo Alto, USA.

Guichard, F., Bouchafa, S., Aubert, D., 2002. A change detector based on level sets. In : Mathematical Morphology and its Applications to Image and Signal Processing, Computational Imaging and Vision. Vol. 18. Springer US, pp. 321–330.

Guilloux, Y. L., 1986. A matching algorithm for horizontal motion, application to tracking. In : International Conference on Pattern Recognition, Paris, France.

Gusfield, D., Irwing, R., 1989. The stable marriage problem - Structure and Algorithms. MIT Press.

Han, B., Comaniciu, D., Davis, L., 2004. Sequential kernel density approximation through mode propagation : applications to background modeling. In : Asian Conference on Computer Vision, Jeju, Korea.

Hanes, D., Keller, J., McCollum, G., 2008. Motion parallax contribution to perception of self-motion and depth. Biological Cybernetics 98 (4), 273–293.

Hartley, R., 1995. In defense of the 8-point algorithm. In : International Conference on Computer Vision, Boston, USA. pp. 1064–1070.

Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning : data mining, inference and prediction. The Mathematical Intelligencer 27 (2), 83–85.

Heinrich, S., 2002. Fast obstacle detection using flow/depth constraint. In : Intelligent Vehicle Symposium.

Hildreth, E. C., 1992. Recovering heading for visually-guided navigation. Vision Research 32 (6), 1177–1192.

Holt, C. C., 2004. Forecasting trends and seasonal by exponentially weighted averages. International Journal of Forecasting 10 (1), 5–10.

Huang, C., Nevatia, R., 2010. High performance object detection by collaborative learning of joint ranking of granules features. In : Conference on Computer Vision and Pattern Recognition, San Francisco, USA.

Irani, M., Rousso, B., Peleg, S., 1997. Recovery of egomotion using region alignment. IEEE Transactions on Pattern Analysis Machine Intelligence 19 (3), 268–272.

Kalinke, T., Tzomakas, C., Seelen, W., 1998. A texture-based object detection and an adaptive model-based classification. In : IEEE Intelligent Vehicles Symposium, Stuttgart, Germany.

Khuller, S., Mitchell, S. G., Vazirani, V. V., 1994. Online algorithms forweighted bipartite matching and stable marriages. Theorical Computer Science 127 (2), 255–267.

Kuhl, P., Giardina, C., 1982. Elliptic Fourier features of a closed contour. Computer Graphics and Image Processing 18, 236–258.

Labayrade, R., Aubert, D., Tarel, J., 2002. Real time obstacle detection on non flat road geometry through 'v-disparity' representation. In : IEEE Intelligent Vehicles Symposium, New-York, USA. pp. 646–651.

Laneurit, J., Blanc, C., Chapuis, R., Trassoudaine, L., 2003. Multisensorial data fusion for global vehicle and obstacles absolute positioning. In : IEEE Intelligent Vehicles Symposium, Colombus, USA.

Leibe, B., Cornelis, N., Cornelis, K., Gool, L. V., 2007. Dynamic 3d scene analysis from a moving vehicle. In : IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA.

Lenz, P., Ziegler, J., Geiger, A., Roser, M., 2011. Sparse scene flow segmentation for moving object detection in urban environments. In : IEEE Intelligent Vehicles Symposium.

Lester, H., Arridge, S. R., 1999. A survey of hierarchical nonlinear medical image registration. Pattern Recognition. Special issue on Image Registration 32 (1), 129–149.

Li, Z., Tauber, Z., Drew, M., 2000. Local-based object search under illumination change using chromaticity voting and elastic correlation. In : IEEE Conference on Multimedia and Expo, New-York, USA.

Lo, B. P. L., Velastin, S., 2000. Automatic congestion detection system for underground platforms. In : International symposium on intelligent multimedia video and speech processing, Hong-Kong, China. pp. 158–161.

Longuet-Higgins, H., Prazdny, K., 1980. The interpretation of a moving retinal image. In : Proceeding of Royal Society London. Vol. B-208. pp. 385–397.

Lowe, D., 2004. Distinctive image features from scale-invariant key- points. International Journal of Computer Vision 60 (2), 91–110.

Lucas, B., Kanade, T., 1981. An iterative image registration technique with an application to stereovision. In : DARPA Image Understanding Workshop. pp. 121–130.

Luong, Q. T., Faugeras, O. D., 1997. Camera calibration, scene motion and structure recovery from point correspondences and fundamental matrices. International Journal of Computer Vision 22 (3), 261–289.

MacKay, D. J., 2003. Information Theory, Inference, and Lear-

ning Algorithms.
MacLean, W., Jepson, A., Frecker, R., 1994. Recovery of egomotion and segmentation of independant object motion using the EM algorithme. In : British Machine Vision Conference. pp. 13–16.
Maintz, J., Viergever, M., 1998. A survey of medical image registration. Medical Image Analysis 2 (1), 1–36.
Maître, H., 1985. Un panorama de la transformée de Hough. Traitement du signal 2 (2), 305–317.
Medioni, G. G., Nevatia, R., 1984. Matching images using linear features. IEEE Transactions on Pattern Analysis Machine Intelligence 6 (6), 675–685.
Monasse, P., Guichard, F., 2000. Fast computation of a contrast-invariant image representation. IEEE Transactions on Image Processing 9 (5), 860–872.
Moravec, H. P., 1981. Rover visual obstacle avoidance. In : Int. joint conference on Artificial intelligence, Vancouver, Canada. pp. 785–790.
Munder, S., Gravila, D. M., 2006. An experimental study on pedestrian classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 1863–1868.
Nack, M. L., 1977. Rectification and registration of digital images and the effect of cloud detection. Machine Processing of Remotely sensed data, Mexico City, Mexico, 12–23.
Nedevschi, S., Bota, S., Tomiuc, C., 2009. Stereo-based pedestrian detection for collision-avoidance applications. IEEE Intelligent Transportation Systems 10 (3), 380–391.
Negahdaripour, S., Horn, B., 1989. A direct method for locating the focus of expansion. Computer Vision Graphics and Image Processing 46 (3), 303–326.
O'Gorman, F., Clowes, M. B., 1973. Finding picture edges through collinearity of feature points. In : Int. Joint Conf. Art. Int. pp. 543–555.
Oliver, N. M., Rosario, B., Pentland, A. P., 2000. A bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8), 831–843.
Paragios, N., Deriche, R., 1998. A PDE-based level set approach for detection and tracking of moving objects. In : International Conference on Computer Vision, Bombay, India. pp. 1139–1145.
Pelizzari, C., Chen, G., Spelbring, D., Weichselbaum, R., 1989. Accurate three dimensional registration of ct, pet and/or mr images of the brain. Journal of computer assisted tomography 13 (1), 20–26.
Penney, G., Weese, J., Little, J., Desmedt, P., Hill, D., Hawkes, D. J., 1998. A comparison of similarity measures for use in 2d/3d medical image registration. IEEE Transactions on Medical Imaging 17 (4), 586–595.
Poelman, C. J., Kanade, T., 1992. A paraperspective factorization method for shape and motion recovery. Tech. rep., CMU-CS-91-208.
Pons, J., Keriven, R., Faugeras, O., 2007. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. International Journal of Computer Vision 72 (2), 179–193.
Reddy, B., Chatterji, B., 1996. An FFT-based technique for translation, rotation and scale-invariant image registration. IEEE Transctions on Image Processing 5 (8), 1266–1271.
Rodriguez, F., Frémont, V., Bonnifait, P., Cherfaoui., V., 2010. Visual confirmation of mobile objects tracked by a multi-layer lidar. In : IEEE International Conference on Intelligent Transportation Systems, Madeira Island, Portugal.
Rosenfeld, A., Kak, A. C., 1982. Digital picture processing. Academic Press.
Sazbon, D., Rotstein, H., Rivlin, E., 2004. Finding the focus of expansion and estimating range using optical flow images and a matched filter. Machine Vision and Applications 15 (4), 229–236.
Scaramuzza, . D., Fraundorfer, F., Siegwart, R., 2009. Real-time monocular visual odometry for on-road vehicles with 1-point Ransac. In : International Conference on Robotics and Automation.
Shapiro, L., Haralick, R., 1990. Matching relational structures using discrete relaxation. In : Syntactic and Structural Pattern Analysis Theory and Applications. World Scientific Pub. Co., Teaneck, NJ.
Shapiro, S. D., 1978. Transform method of curve detection for textured image data. IEEE Transactions on Computing C-27 (3), 254–255.
Smith, S. M., Brady, J. M., 1997. Susan : A new approach to low level image processing. International Journal of Computer Vision 23 (1), 45–78.
Soquet, N., Perrollaz, M., Aubert, D., 2007. Free space estimation for autonomous navigation. In : International Conference on Computer Vision Systems, Bielefeld, Germany.
Stauffer, C., Grimson, W. E. L., 2000. Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8), 747–757.
Stein, G. P., Mano, O., Shashua, A., 2000. A robust method for computing vehicle egomotion. In : IEEE Intelligent Vehicles Symposium, Dearborn, USA. pp. 362–368.
Stockman, G., Kopstein, S., Benett, S., 1982. Matching images to models for registration and object detection via clustering. Pattern Recognition and Image Analysis 4 (3), 229–241.
Suvonvorn, N., Bouchafa, S., Zavidovique, B., 2005. Marrying level lines for stereo or motion. In : LNCS, Image Analysis and Recognition. Vol. 3656. pp. 391–398.
Szeliski, R., 1994. Image mosaicing for telereality applications. In : IEEE Workshop on Applications of Computer Vision.
Thrift, P., Dunn, S., 1983. Approximating point-set images by line segments using a variation of the Hough transform. Computer Graphics and Image Processing 21 (3), 383–394.
Turk, M. A., Pentland, A. P., 1991. Face recognition using eigenfaces. In : Computer Vision and Pattern Recognition.
Vapnik, V., 1999. An overview of statistical learning theory. Neural Networks 10 (5), 988–999.
Verri, A., Poggio, T., 1989. Motion field and optical flow : Qualitative properties. IEEE Transactions on Pattern Analysis Machine Intelligence 11 (5), 490–498.
Viola, P., Jones, M. J., Snow, D., 2005. Detecting pedestrians using patterns of motion and appearance. International Journal of Computer Vision 63 (2), 153–161.
Vu, T., Burlet, J., Aycard, O., 2008. Grid-based localization and online mapping with moving objects detection and tracking : new results. In : Intelligent Vehicles Symposium, Eindhoven, The Netherlands.
Wedel, A., Rabe, C., Vaudrey, T., Brox, T., Franke, U., 2008. Efficient dense scene flow from sparse or dense stereo data. In : European Conference on Computer Vision, Marseille, France.
Wren, C., Azarbayejani, A., Darrell, T., Pentland, A., 1997. Pfinder : Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7), 780–785.
Wu, F., Wang, L., Hu, Z. Y., 2007. Foe estimation : Can image measurement errors be totally. Pattern recognition 40 (7), 1971–1980.
Xu, S., Danielson, P. E., 1992. Robust estimation of focus of expansion and depth from high confidence optical flow. In : IAPR Workshop on Machine Vision Applications.
Zemirli, K., Seetharaman, G., Zavidovique, B., 2000. Stable matching or selective junction points grouping. In : IEEE Joint Conference on Information Sciences.