N° d'ordre : 1304

THÈSE

présentée devant

L'UNIVERSITÉ DE RENNES I

U.F.R. Structure et Propriétés de la Matière

pour obtenir le titre de

Docteur de l'Université de Rennes I
Mention : Traitement du Signal et Télécommunications

par

Jean-Marc ODOBEZ

Estimation, détection et segmentation du mouvement: une approche robuste et markovienne

Soutenue le 21 décembre 1994, devant la commission d'examen composée de :

M.	Jean-Jacques	FUCHS	Président
MM.	Bernard	CHALMOND	Rapporteurs
	Hans-Hellmut	NAGEL	
MM.	Jean-Marc	BOUCHER	Examinateurs
	Mike	BRADY	
	Bertrand	ZAVIDOVIQUE	
	Patrick	BOUTHEMY	

N° d'ordre : 1304

THÈSE

présentée devant

L'UNIVERSITÉ DE RENNES I

U.F.R. Structure et Propriétés de la Matière

pour obtenir le titre de

Docteur de l'Université de Rennes I
Mention : Traitement du Signal et Télécommunications

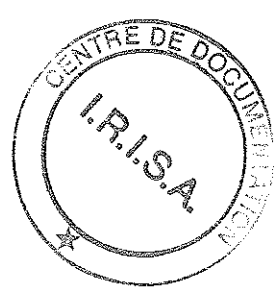
par

Jean-Marc ODOBEZ

Estimation, détection et segmentation du mouvement: une approche robuste et markovienne

Soutenue le 21 décembre 1994, devant la commission d'examen composée de :

M.	Jean-Jacques	FUCHS	Président
MM.	Bernard	CHALMOND	Rapporteurs
	Hans-Hellmut	NAGEL	
MM.	Jean-Marc	BOUCHER	Examinateurs
	Mike	BRADY	
	Bertrand	ZAVIDOVIQUE	
	Patrick	BOUTHEMY	



Remerciements

Ce travail de recherche a été réalisé à l'Institut de Recherche en Automatique et Systèmes Aléatoires (IRISA) de Rennes, au sein du projet Traitement, Exploitation et Modélisation d'Images Séquentielles (TEMIS). Je tiens à remercier Monsieur Claude LABIT, directeur de recherche INRIA, qui m'a accueilli avec bienveillance dans son équipe, ainsi que tous les autres membres de celle-ci.

Je remercie Monsieur Jean-Jacques FUCHS, Professeur à l'Université de Rennes I d'avoir bien voulu accepter de présider le jury de cette thèse.

J'adresse tous mes remerciements à Monsieur Patrick BOUTHEMY, directeur de recherche INRIA, pour la confiance qu'il m'a accordée et les nombreux conseils qu'il m'a prodigués pendant toute la durée de mes travaux.

Je remercie Monsieur Bernard CHALMOND, Professeur à l'Université de Cergy Pontoise, pour avoir accepté sans hésiter d'être rapporteur de cette thèse.

Je tiens à exprimer toute ma gratitude à Monsieur Hans-Hellmut NAGEL, Professeur à l'Université de Karlshruhe (Allemagne), qui s'est penché avec attention et perspicacité sur mon travail.

Messieurs Jean-Marc BOUCHER, Professeur à l'École Nationale Supérieure des Télécommunications de Bretagne, Mike BRADY, Professeur à l'Université d'Oxford (Angleterre), et Bertrand ZAVIDOVIQUE, Professeur à l'Université d'ORSAY ont gentillement accepté de participer à mon jury. Je leur en suis très reconnaissant.

Je tiens ici à manifester toute la joie et le plaisir que m'ont procurés Messieurs Hans-Hellmut NAGEL et Mike BRADY en acceptant de faire l'effort de lire ce document, malgré la difficulté supplémentaire que représente pour eux la lecture du français.

Enfin, toute mon affection va à Catherine, dont les encouragements et le soutien moral constant me furent si précieux lors de la rédaction de cette thèse.

Table des matières

c

Chapitre 1

Introduction

Des cinq sens de l'être humain, la vue est sans conteste celui qui nous est le plus utile: plus de quatre vingts pour cent des informations dont nous disposons sont canalisées par la vue, alors que le rôle d'autres sens, surtout celui du goût, de l'odorat et du toucher, devient de plus en plus marginal dans l'environnement fabriqué par l'homme. Dans le règne animal, la répartition entre les différents sens est plus équilibrée. On distingue les récepteurs chimiques (odorat, goût) et tactiles, par exemple les antennes des crustacés ou plus simplement les moustaches du chat, les récepteurs sensibles aux vibrations mécaniques, dont fait partie l'ouïe mais aussi le système de localisation des cétacés dont le squelette constitue un véritable sonar, ou celui des chauves-souris; enfin les récepteurs sensibles aux ondes électromagnétiques. Dans cette dernière catégorie, nous trouvons les systèmes d'orientation de certains oiseaux comme les pigeons, qui sont sensibles au magnétisme terrestre, et bien sûr la vue. Cette dernière peut prendre de multiples formes. Elle peut par exemple être en couleur, comme chez l'homme, ou en "noir et blanc" comme chez les animaux nocturnes (hiboux, chouettes) dont la rétine est dépourvue des capteurs (les cônes) responsables de la vision en couleur.

Alors que pour un certain nombre d'espèces animales le système de vision ne sert que pour des besoins très spécifiques, le système visuel humain possède un caractère général qui fait sa richesse et sa complexité. Ses multiples aspects le placent à la croisée de nombreuses disciplines scientifiques: optique, géométrie, anatomie, biologie, neurophysiologie, psychophysique. Avec l'étude des systèmes de vision artificiels s'ajoutent à cette longue liste l'informatique, l'électronique, les mathématiques appliquées, c'est-à-dire la vision par ordinateur, qui constitue le cadre général de cette thèse.

En partie par anthropomorphisme, les systèmes de vision développés actuellement tendent souvent à reproduire l'équivalent du système de vision humain. L'utilisation de deux caméras permettant d'assurer une vision stéréoscopique, et donc d'appréhender à chaque instant la profondeur, est courante. Cependant, de tels systèmes sont onéreux et complexes, et ne peuvent pas être employés dans toutes les situations. Une autre approche consiste à employer un système de vision monoculaire. Ce dernier ne donne pas accès

instantanément à la profondeur des différents points d'une scène. Cependant, lorsque la caméra se déplace et tourne par exemple autour d'un objet statique, les images acquises à différents instants (et sous différents angles de vue) nous renseignent tout de même sur la forme de cet objet. L'information utile à la reconstruction du monde tridimensionnel (3D) est donc accessible au système de vision monoculaire, à condition que la caméra ou les objets à reconstruire soient en mouvement. Ce lien qui unit intimement le mouvement dans l'image et la perception de l'univers tridimensionnel semble être en grande partie à l'origine de la faculté qu'ont pratiquement tous les êtres vivants de se déplacer et de se mouvoir dans un environnement quelconque. C'est donc tout naturellement que l'essentiel des premières recherches en vision dynamique, domaine de la vision par ordinateur qui traite du mouvement dans une séquence d'images, s'est porté vers le paradigme de la reconstruction du monde 3D et de l'estimation du mouvement 3D à partir du mouvement 2D observé entre deux images, que l'on nomme le mouvement apparent. Un tel paradigme permettrait en principe d'obtenir un système de vision général et suffisant pour mettre en oeuvre le large éventail des tâches que l'on peut demander à un système robotique par exemple.

L'intérêt de l'analyse du mouvement apparent 2D est évoqué ci-dessus. Ce champ contient intrinsèquement suffisamment d'informations pour réaliser des tâches importantes comme par exemple la détermination de la direction de déplacement du capteur (extraction du foyer d'expansion dans l'image [Sun92, HW88], c'est-à-dire le point où s'intersectent les supports des directions des vecteurs de vitesse 2D dans le cas d'une translation 3D), ou plus généralement l'estimation du torseur cinématique de la caméra [HW88, HT81]. L'obtention de la carte des profondeurs (relatives) des objets dans la scène [Adi85, Han91], la détection d'obstacles fixes sur la trajectoire d'un robot [Anc92], la détection d'objets mobiles dans la scène [AKM93, BL90] et leur suivi [IRP92], sont d'autres objectifs qui peuvent être atteints à l'aide de l'analyse du mouvement. Par ailleurs celle-ci joue également un rôle important dans des domaines plus particuliers, comme ceux de l'imagerie biomédicale ou satellitaire.

De nombreux modèles et méthodes théoriques ont été proposés pour résoudre le problème de l'estimation du mouvement 3D à partir du mouvement 2D. Cependant des difficultés importantes on été rencontrées dans leur mise en œuvre pratique. En effet, la réussite de ces méthodes dépend fortement de la qualité des informations de mouvement extraites de l'image, et la précision requise pour obtenir des solutions correctes est extrêmement difficile à atteindre. De plus, elles supposent généralement que ces informations de mouvement n'appartiennent qu'à un seul objet rigide en mouvement. Or la gestion de la présence de plusieurs mouvements différents dans la scène s'est avérée être un problème très ardu. Comme nous le verrons dans la suite de cette thèse, une façon d'éviter ces difficultés consiste à ne pas rechercher à reconstruire le monde 3D pour ensuite l'analyser,

mais à formuler les tâches de vision dynamique directement en termes d'objectifs d'analyse du mouvement 2D.

Une autre façon d'aborder l'analyse du mouvement apparent dans une séquence d'images acquises par un capteur vidéo consiste donc à positionner le problème dans l'espace même de la séquence d'images. Une telle approche présente un certain nombre d'avantages importants: l'utilisation complète du signal image, l'exploitation efficace des outils de modélisation et d'estimation à notre disposition dans ce cadre, et l'économie des calculs de structures de représentation 3D. Elle permet de traiter de nombreuses situations rencontrées en pratique. Cette thèse se situe dans cette perspective, et essaiera de répondre à un objectif important de la vision dynamique, qui est de détecter des objets en mouvement dans une scène, alors que la caméra elle-même peut être mobile. Cet objectif est motivé par de nombreuses applications en pratique. Il intervient directement pour le suivi d'objets dans l'image ou dans la scène, dans la mesure où ceux-ci doivent avoir été préalablement détectés. Il permet également, en focalisant l'attention du système de vision sur des régions d'intérêt de la scène, de réduire la quantité de données à traiter par des phases de reconnaissance ultérieures. De la même façon, il sert en codage à concentrer l'effort de transmission sur les zones qui le nécessitent vraiment. Pour atteindre cet objectif dans le cas général, il faut répondre à des questions importantes de la vision dynamique, qui sont l'estimation et la segmentation du mouvement. Nous aborderons celles-ci dans cette thèse, ainsi que de manière implicite le suivi de régions au cours du temps.

Il est intéressant de noter que la plupart des systèmes de vision biologique possèdent des mécanismes de suivi. Ceux-ci opèrent généralement de la façon suivante [DGV92]: après avoir détecté l'apparition dans le champ de vision d'un objet mobile, un mouvement rapide de rotation oculaire (une saccade) permet la focalisation sur cet objet. Celui-ci se retrouve alors au centre de la rétine (fovéa), où l'acuité visuelle est la plus élevée. Ensuite, le suivi est effectué par stabilisation du regard, en maintenant l'objet sur la fovéa par compensation visuelle du mouvement de la cible. Cette phase de stabilisation est essentielle si l'observateur doit reconnaître l'objet, évaluer sa forme, sa vitesse, etc. La transposition d'un schéma biologique de ce type à un système artificiel suppose que l'on contrôle au moins en partie le mouvement de la caméra ou/et de certains de ses paramètres, comme la focale. Cette approche constitue en fait un axe de recherche à part entière que l'on désigne sous le terme de vision active [Baj88, AWB87, ECR92], par opposition à l'analyse dynamique classique, qui, elle, considère la séquence d'images comme une donnée dont les résultats de l'analyse ne permettent pas de modifier le processus d'acquisition. Le paradigme de la vision active est très prometteur dans la mesure où, comme Aloimonos *et al.* [AWB87] l'indiquent, "un problème qui est mal posé, non linéaire ou instable pour un observateur passif, devient bien posé, linéaire ou stable pour un observateur actif". Cependant, dans certaines applications, le contrôle du capteur n'est

pas toujours possible. De plus, l'approche que nous avons décrite ci-dessus ne permet de suivre qu'une seule cible à la fois, ce qui peut s'avérer insuffisant dans le cas où un grand nombre d'objets sont présents dans la scène. Dans ce dernier cas, l'établissement de la trajectoire dans l'image de chacun d'entre-eux, par exemple à l'aide d'un filtrage récursif [MB94a], doit se faire de manière simultanée. Ceci peut-être réalisé si l'on dispose d'une segmentation au sens du mouvement apparent.

Le mouvement apparent peut être considéré comme un critère pour effectuer une partition de l'image. Le but est alors d'obtenir des régions à l'intérieur desquelles le champ des vitesses est continu. De fait, ce problème est intrinsèquement lié à celui de l'estimation du mouvement. En effet, pour résoudre ce dernier, l'information locale de mouvement dans l'image est en général insuffisante. Il faut alors faire l'hypothèse d'une certaine continuité spatiale des vecteurs de vitesse. Cette continuité n'est évidemment pas vérifiée aux frontières entre deux objets ayant des mouvements différents, ou sur les contours correspondant à des ruptures de profondeur importantes dans la scène observée. Le rôle de la segmentation consiste donc à repérer ces frontières pour éviter le lissage du champ des déplacements à travers celles-ci. Une segmentation préliminaire correcte basée sur le mouvement 2D mettra en évidence les différentes composantes dynamiques de la scène, ce qui facilitera l'analyse ultérieure de leur mouvement et de leur trajectoire dans l'image, et, si cela est nécessaire, dans l'environnement 3D.

Enfin dans de nombreuses situations, les systèmes de vision sont appelés à fournir uniquement une interprétation qualitative de la scène. Celle-ci peut servir à définir des données symboliques utilisables par un processus de décision de plus haut niveau [Nag88a]. En analyse du mouvement, des études ont montré que le champ des déplacements 2D contient en soit une information suffisamment riche pour caractériser un certain nombre de situations 3D pertinentes [VP89, KD75]. L'analyse directe de ce champ peut donc fournir une description symbolique robuste (c'est-à-dire moins sensible à la précision des vecteurs de mouvement du champ apparent) de la scène, sans avoir à résoudre le délicat problème de la reconstruction et de l'estimation du mouvement 3D. Ces propriétés ont été utilisées par exemple dans [FB90] pour classer différents types de mouvements, ou dans [Nel91] pour détecter en temps réel des objets mobiles dans une scène lorsque la caméra est elle-même en mouvement.

L'analyse du mouvement est donc une source d'information importante. Malheureusement, le traitement de cette information est généralement coûteux en temps de calcul, ce qui est incompatible avec la nature même de cette information, qui suppose qu'elle puisse être exploitée au rythme de son acquisition, c'est à dire en temps réel. De ce point de vue, peu d'algorithmes d'analyse du mouvement sont actuellement opérationnels. Cepen-

dant l'évolution constante vers des machines de plus en plus rapides et avec des capacités de plus en plus importantes, la recherche d'architectures spécialisées pour le traitement d'images, notamment sous la forme d'architectures parallèles [Mem93], ou sous la forme de rétines artificielles [BNDZ93] qui incluent non seulement le processus d'acquisition des images mais également des traitements bas niveaux de type neuronal [Zav92, Fra92], et enfin, l'émergence à plus long terme de calculateurs optiques, laissent présager une exploitation beaucoup plus conséquente du potentiel que représente l'analyse du mouvement. Les domaines dans lesquels elle occupe déjà une place importante sont les suivants [Hua83, Nag88b, AN88, MB94b]:

- La télévision ainsi que des services de télécommunications tels que le visiophone ou la vidéo-conférence. L'analyse du mouvement a pour but principal dans ce cas d'extraire la redondance temporelle entre les images d'une même séquence (codage par compensation de mouvement) pour comprimer l'information à un taux compatible avec la bande passante des systèmes de transmission concernés, tout en préservant la qualité des images reconstruites au récepteur [TL94, TL91, TB89].

- L'imagerie aérienne et satellitaire, pour estimer les champs de déplacement des nuages [LNC71, SN87, CQB92], et analyser les phénomènes météorologiques à partir de ceux-ci.

- La robotique. Comme nous l'avons déjà indiqué, il s'agit de réaliser des robots capables de se déplacer dans un environnement connu ou non [Hor86, WBBH92, Fau93].

- L'imagerie biomédicale (échographie, angiographie) [MLSB89, BMM*89, TMCZ80] [BMDF89, HA92]: on peut citer l'analyse automatique des mouvements déformables du cœur pour l'aide au diagnostique, ainsi que l'analyse du mouvement humain.

- Trafic routier et surveillance de sites: estimation et contrôle du trafic routier ou urbain, détection d'intrusion [Koz89, FMRS94].

- Les applications militaires, notamment pour la détection et le suivi de cibles.

Dans cette thèse, nous abordons le problème de la détection d'objets en mouvement dans le cas général où la caméra est elle-même mobile. Sa résolution implique en partie la réalisation d'autres objectifs, à savoir ceux de l'estimation et de la segmentation du mouvement.

Plutôt que de recourir à l'estimation d'un champ dense des déplacements apparents, notre approche consiste à employer des modèles paramétriques comme descripteurs du mouvement 2D de régions de l'image. Pour calculer ceux-ci, nous avons retenu un estimateur robuste qui permet de s'affranchir (partiellement) de la contrainte de la présence d'un unique mouvement dans la région considérée.

L'estimateur robuste que nous avons défini est mis à profit dans un schéma de détection des régions de l'image correspondant à des éléments en mouvement dans une scène observée par une caméra mobile. Le principe de cette méthode est de compenser dans un premier temps le mouvement induit par le déplacement de la caméra. Cela est réalisé à l'aide de l'estimateur robuste multirésolution, qui permet de mesurer le mouvement dominant dans l'image sans être affecté par la présence d'autres mouvements. Le problème posé se réduit à la détection des zones mal compensées, c'est-à-dire non conformes au mouvement dominant, dans la séquence ainsi reconstruite. Celui-ci est alors formulé comme un problème d'étiquetage dans un cadre bayesien s'appuyant sur une modélisation statistique. Plus précisément, nous avons utilisé des modèles de Markov, dont l'intérêt pour traiter des problèmes d'analyse d'image a été largement démontré par ailleurs [GG84, MMP87, Cha88, Hei93].

La méthode de segmentation d'une séquence d'images en régions de mouvement homogène constitue une extension de la méthode de détection du mouvement. Celle-ci est également considérée comme un problème d'étiquetage statistique en n classes dans un cadre markovien. La gestion des nouvelles régions apparaissant dans l'image est prise en charge par un sous-module similaire à la technique de détection du mouvement définie, ce qui permet d'estimer "en ligne" le nombre de mouvements n présents dans l'image à un instant donné.

Le plan général de la thèse est donc le suivant:

- Dans le chapitre deux, nous exposons plus en détail les problèmes que nous venons de soulever et dressons un état de l'art non exhaustif des approches existantes pour les résoudre.

- Le troisième chapitre s'ouvre sur une discussion portant sur l'opportunité de l'utilisation des modèles de mouvement paramétriques. Nous décrivons ensuite la technique d'estimation aux moindres-carrés multirésolution qui sert de base à notre méthode, et à laquelle sera comparé notre algorithme. Après un bref rappel sur les estimateurs robustes, notre méthode d'estimation robuste multirésolution de modèles paramétrés de mouvement est alors décrite.

- La méthode de détection des objets mobiles d'une scène est présentée dans le chapitre quatre. Nous commençons tout d'abord par définir des mesures de compensation du mouvement adaptées à notre problème. La fiabilité de celles-ci est alors étudiée et mise à profit dans la définition du terme d'énergie d'attache aux données. La prise en compte d'une intégration temporelle des observations et des cartes d'étiquettes dans les termes d'énergie, ainsi que l'utilisation d'un modèle markovien multiéchelle sont ensuite décrits.

- Le schéma de segmentation du mouvement apparent fait l'objet du chapitre cinq. L'approche que nous avons retenue est d'abord exposée. Le synoptique complet de

l'algorithme est présenté et ses différentes phases sont explicitées, notamment la détermination du nombre de régions en mouvements et la segmentation en régions proprement dite, les modèles de mouvement étant identifiés.

Ces trois chapitres sont illustrés par de nombreux résultats obtenus sur des séquences réelles représentant des scènes complexes de type varié. Enfin, dans la conclusion générale, nous tirons un bilan synthétique de ces travaux et proposons quelques perspectives de recherche associées à cette étude.

Chapitre 2

Éléments d'état de l'art sur l'analyse du mouvement 2D

Le mouvement relatif entre la caméra et les objets d'une scène induit un mouvement apparent dans une séquence d'images. La vision dynamique concerne l'analyse et l'interprétation de ce mouvement. L'objectif est en général de retrouver le mouvement 3D à partir de l'information dynamique contenu dans la séquence. Cependant, comme nous l'avons souligné dans l'introduction, cet objectif, difficile à atteindre, n'est pas toujours nécessaire à l'accomplissement de certaines tâches. La détection du mouvement, c'est-à-dire la détection et la localisation dans l'image des projections des objets mobiles de la scène, constitue l'une d'entre elles.

Suivant que la caméra est fixe ou mobile lors de l'acquisition de la séquence d'images, les moyens à mettre en œuvre pour effectuer la détection du mouvement seront plus ou moins complexes.

La détection d'objets mobiles dans une scène à l'aide d'une caméra fixe ne requiert pas nécessairement l'estimation du champ des déplacements 2D de l'image. En l'absence de mouvement dans la scène, les images successives acquises par le capteur vidéo sont identiques. Les objets mobiles, lorsqu'il y en a, peuvent donc se caractériser à partir de la variation du signal intensité au cours du temps. La détection de ces changements temporels contribue alors à la localisation des objets mobiles dans l'image.

Dans notre étude, nous nous sommes intéressés au cas plus général d'une caméra mobile. Le mouvement de celle-ci induit dans l'image un mouvement apparent pour –presque– tous les points des surfaces visibles de la scène. Pour réaliser la détection des objets mobiles, il est alors généralement nécessaire de procéder à des mesures de mouvement. Pour pouvoir les exploiter à des fins de détection du mouvement, celles-ci doivent être estimées de manière fiable, notamment bien sûr dans le cas de la présence de plusieurs objets mobiles dans la scène. Dans ce chapitre, nous exposerons donc tout d'abord les hypothèses

générales utilisées pour estimer le mouvement, puis nous décrirons quelques méthodes d'estimation en relation avec la suite de ce document.

Suivant les mesures de mouvement estimées dans l'image, on distingue trois grandes approches pour effectuer la détection d'objets mobiles:

- Dans la première approche, un champ dense de déplacements est tout d'abord estimé. La connaissance (ou l'estimation) de certains paramètres du torseur cinématique de la caméra permet d'appliquer sur ce champ des contraintes, qui, si elles ne sont pas vérifiées, indiquent la présence d'un objet mobile.

- La seconde repose sur l'hypothèse suivante: le mouvement apparent (2D) des zones statiques de la scène peut être appréhendé par un modèle de mouvement 2D. Celui-ci est alors estimé, puis utilisé pour compenser dans l'image le mouvement apparent dû au déplacement de la caméra. On est alors ramené à la détection de changement temporel dans la séquence compensée.

- Dans la troisième approche, on effectue la segmentation de l'image en régions de mouvement homogène. Une phase d'interprétation du mouvement de chacune de ces régions permet alors de faire la distinction entre régions de mouvement correspondant à des parties statiques de la scène observée, et projections d'objets mobiles.

Soulignons que la segmentation du mouvement dans l'image a une portée plus générale que la détection des objets mobiles. En effet, lorsque la caméra se déplace dans un environnement contenant un ou plusieurs objets mobiles, le champ des déplacements qui en résulte se composera alors de la juxtaposition de plusieurs champs distincts correspondant aux projections des mouvements rigides des différents objets. Avant ou pendant une phase de calcul et d'interprétation du mouvement 3D, le champ apparent doit être segmenté en ses différentes composantes, sur lesquelles les méthodes de reconstruction sont applicables.

Enfin, notons que les phases de détection et de segmentation du mouvement sont essentielles en vision dynamique, car elles facilitent la focalisation sélective des capacités calculatoires d'un système de vision –souvent limitées comparativement à la quantité de données à traiter– sur les régions qui présentent vraiment un intérêt pour les tâches que ce système doit accomplir. Dans ce chapitre, nous proposons une classification des approches traitant plus particulièrement de la détection du mouvement, et de celles traitant de la segmentation du mouvement, en fonction des méthodologies utilisées.

2.1 Estimation du mouvement apparent

Dans cette partie, nous nous intéressons au problème de l'estimation du mouvement apparent dans l'image. La plupart des méthodes d'estimation du mouvement apparent

exploitent deux hypothèses: la conservation de l'intensité lumineuse du point en mouvement et la continuité spatiale du mouvement. La continuité temporelle est une troisième hypothèse que l'on rencontre également. Ces hypothèses constituent une simplification des phénomènes observés et seront donc mises en défaut dans certaines situations que nous décrirons. Pour traiter ces dernières, nous avons alors la possibilité d'utiliser soit des contraintes et des modèles plus réalistes (pour l'estimation de mouvement, cela consiste par exemple à rechercher explicitement les frontières de mouvement), soit des techniques ou des algorithmes qui resteront utilisables même lorsque les hypothèses seront mises en défaut (ceci constitue l'approche robuste). En fait, les deux approches sont complémentaires dans la mesure où, quelle que soit la complexité des modèles, ils seront toujours susceptibles de n'être pas vérifiés. Par ailleurs, notons qu'il est non seulement intéressant de pouvoir traiter les situations ne respectant pas les hypothèses retenues, mais aussi de savoir les détecter et les localiser car elles contiennent souvent une information importante sur la scène. Nous examinerons donc tout d'abord les différentes hypothèses et leurs mises en défaut avant de considérer certaines méthodes d'estimation du mouvement qui permettent de remédier à la non-validité des hypothèses.

2.1.1 Hypothèses et validité des hypothèses

Dans cette partie, nous commencerons par considérer les hypothèses faites par la plupart des algorithmes d'estimation du mouvement, à savoir la conservation de l'intensité, la continuité spatiale, la continuité temporelle, puis nous commenterons leur validité.

Conservation de l'intensité

Notons $I(\mathrm{p}(t), t)$ la valeur de l'intensité à l'instant t en un pixel de l'image. Pour estimer le mouvement apparent, on suppose que cette intensité $I(\mathrm{p}(t), t)$ est une grandeur physique, attachée ou reliée à un point P d'une surface tridimensionnelle visible, dont la projection dans l'image est le point p. On suppose d'autre part que cette intensité reste constante sur la trajectoire de ce point P. On a alors [HS81]:

$$\frac{\mathrm{d}I}{\mathrm{d}t}(\mathrm{p}(t), t) = 0 \tag{2.1}$$

Notons dès maintenant que l'on peut considérer d'autres quantités invariantes par rapport au mouvement. Par exemple, on peut considérer le gradient de l'intensité [Nag83, TP84, BPT88, KT94a], ou plus généralement des opérateurs agissant sur l'image originale, comme le contraste ou l'entropie, ainsi que des grandeurs acquises dans différentes bandes de fréquences radiométriques (R, V, B, IR, etc.) [MWA87, MF90].

L'équation de conservation n'est pas vérifiée dans un nombre important de situations. Tout d'abord, ce modèle ne peut bien évidemment pas prendre en compte les variations d'illumination qui peuvent survenir, particulièrement dans les scènes d'extérieur, ni les phénomènes d'ombrage liés à l'orientation des surfaces des objets et qui ne sont donc pas

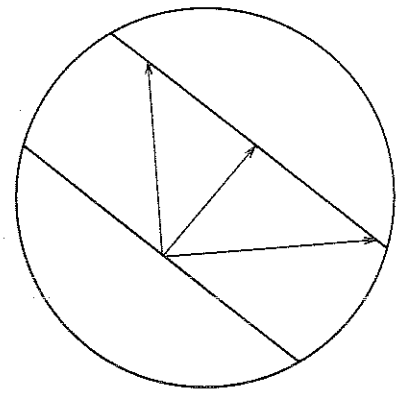


FIG. 2.1 - *Localement, le déplacement d'un segment de droite est ambiguë. Seule la composante de déplacement normale à celui-ci est déterminable localement.*

invariants par rotation 3D. Les reflets spéculaires ne se déplacent généralement pas de la même façon que la surface sous-jacente. Le bruit d'acquisition est aussi une source de changement de l'intensité dans l'image, mais qui peut se modéliser plus facilement. Par ailleurs, il est évident que l'intensité des points d'une région n'est pas conservée si cette région est recouverte ou cachée par une autre surface en mouvement. Cette situation peut être particulièrement gênante et difficile à gérer, par exemple lorsque l'on regarde à travers une grille ou un feuillage. Finalement, dans le cas de la transparence, c'est-à-dire lorsque l'intensité en un point de l'image est la combinaison d'informations relevant de deux (ou plusieurs) points de l'espace appartenant à des surfaces différentes (vision à travers une vitre peinte, ou vision simultanée d'un objet au-delà d'une vitre et de la réflection d'un objet sur cette vitre), il est indispensable de modifier le modèle [BBHP92, SM91].

Continuité spatiale

L'hypothèse de conservation de l'intensité n'est pas suffisante pour déterminer entièrement et précisément le mouvement apparent. La figure 2.1 illustre l'ambiguïté qui existe lorsque l'on veut calculer le déplacement d'un segment de contour ("problème dit de l'ouverture"). Pour surmonter cette difficulté, on suppose généralement que des points voisins de l'image appartiennent à la même surface 3D, et ont de ce fait des déplacements apparents similaires. Cette hypothèse est ensuite utilisée explicitement par l'intermédiaire d'une contrainte de lissage [HS81, Nag83]. Plus implicitement, elle peut aussi être mise en œuvre à travers des modèles de mouvement paramétriques, qui peuvent prendre en compte –plus ou moins localement– un champ constant, un champ affine, voire un champ d'ordre supérieur.

Une hypothèse souvent retenue est que le champ des déplacements est constant à l'intérieur d'une région (hypothèse implicite dans [HS81]). Or celle-ci ne sera pas vérifiée lorsque le champ comprend des composantes significatives de rotation ou de divergence. Cependant, la situation la plus fréquente où l'hypothèse de continuité n'est pas valide se

rencontre aux frontières de mouvement. Celles-ci se produisent lorsque deux objets voisins ont des mouvements 3D différents, mais également lorsque deux points voisins de l'image sont les projections de deux points de l'espace tridimensionnel dont le déplacement obéit au même mouvement rigide mais dont les profondeurs (par rapport à la caméra) sont différentes (la discontinuité en profondeur se répercute sur le mouvement apparent). De l'utilisation de l'hypothèse dans de telles conditions résultera alors soit un champ dense de mouvement qui lisse de manière erronée les vitesses aux frontières de mouvement dans le cas d'une approche non-paramétrique, soit un modèle de mouvement estimé qui "moyenne" des mouvements différents. Dans les deux cas, les informations contenues dans la scène sont mélangées spatialement, ce qui ne facilitera pas la tâche de modules d'interprétation placés en aval de la phase d'estimation du mouvement.

Continuité temporelle

La continuité temporelle suppose que les déplacements mesurés dans l'image varient de manière continue dans le temps sur les trajectoires des points [Bla94], et parfois même en un point fixe de l'image [GS94]. Elle n'est généralement valide que sur des durées limitées. Son exploitation dépend en grande partie de la qualité de l'acquisition, particulièrement de la fréquence temporelle d'échantillonnage, des vibrations éventuelles de la caméra, etc... Par ailleurs, le mouvement propre des objets n'est pas toujours prévisible. Ainsi le mouvement des êtres vivants change fréquemment de direction. Nelson nomme ces mouvements des "mouvements animés", et propose une méthode particulière pour les détecter [Nel91]. Enfin, les frontières de mouvement sont une fois de plus des régions critiques où l'hypothèse est mise en défaut.

Plusieurs algorithmes utilisent plus ou moins implicitement cette hypothèse, particulièrement les méthodes par transformées qui requièrent des supports temporels étendus [FJ90, AB85, Hee88]. Cette continuité dans le temps est parfois appliquée sur des blocs d'images au travers de modèles spatio-temporels de mouvement [AS94], mais le moyen d'assurer la continuité entre blocs d'analyse n'est pas abordé. En fait, la continuité temporelle est généralement utilisée dans une seconde phase. La plupart des algorithmes d'estimation se basent sur deux images uniquement, et sont amenés à effectuer la minimisation de fonctionnelles avec des méthodes d'optimisation itératives. La qualité de la solution obtenue par ces méthodes itératives dépend en grande partie d'une estimée initiale. Cette dernière est alors fixée en fonction des estimations passées, par simple utilisation de l'estimée à l'instant précédent, ou à l'aide d'une prédiction obtenue par filtrage temporel. Mais les estimées précédentes peuvent être utilisées également comme "contraintes" dans le processus d'estimation à l'image courante. Le champ à estimer ne devra pas alors s'écarter de manière trop importante du champ prédit, à moins que des occlusions n'aient été détectées [Bla94].

2.1.2 Méthodes d'estimation du mouvement apparent

Nous ne présenterons ici que les méthodes d'estimation du mouvement dites différentielles et les méthodes par mesure de similarité. Les approches par mise en correspondance ou par transformée, dont il n'est pas question dans la suite de ce document, seront volontairement omises.

Les méthodes différentielles

En utilisant la dérivée particulaire dans la formule (2.1), l'hypothèse de conservation de l'intensité se traduit par:

$$\frac{\mathrm{d}I}{\mathrm{d}t}(\mathrm{p}(t),t) = \vec{\nabla} I(\mathrm{p}(t),t).\frac{\mathrm{dp}}{\mathrm{d}t}(\mathrm{p}(t),t) + \frac{\partial I}{\partial t}(\mathrm{p}(t),t) = 0 \tag{2.2}$$

Cette équation, appelée couramment équation de conservation du mouvement apparent (ECMA) relie le mouvement apparent $\frac{\mathrm{dp}}{\mathrm{d}t} = \vec{\mathrm{v}}(\mathrm{p})$ aux gradients spatio-temporels de l'intensité I: $(\vec{\nabla} I, I_t) = ((\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}), \frac{\partial I}{\partial t})$. De cette équation, seule la composante parallèle au gradient spatial de l'intensité peut être calculée. Si l'on note $\vec{u} = \frac{\vec{\nabla} I}{\|\vec{\nabla} I\|}$ le vecteur unitaire parallèle au gradient d'intensité, et d'autre part $v_n = -\frac{I_t}{\|\vec{\nabla} I\|}$, l'équation précédente peut s'écrire sous la forme:

$$\vec{u}(\mathrm{p}).\vec{\mathrm{v}}(\mathrm{p}) - v_n(\mathrm{p}) = 0 \tag{2.3}$$

Ce problème de l'indétermination du mouvement apparent en chaque point est connu sous le nom du "problème de l'ouverture" (*"aperture problem"* en anglais [HS81]). L'indétermination peut être levée si l'on dispose de sources d'information multiples [MWA87], ou de plusieurs équations de conservation [Tis94]. Par exemple, l'hypothèse de conservation du gradient spatial de l'intensité fournit deux équations, mais qui sont fortement corrélées. De plus, ces équations, qui font intervenir des dérivées d'ordre deux de la fonction intensité, généralement très bruitées, ne sont pas très fiables.

L'addition de la contrainte de continuité spatiale est donc préférable, et se fait généralement sous la forme d'une contrainte de lissage qui restreint l'ensemble des solutions admissibles. Celle-ci aura pour but de favoriser les solutions variant peu [HS81, HB93]. Pour résoudre le problème de l'estimation de champ dense, on fait alors appel à des méthodes d'optimisation globale basées sur l'analyse variationnelle ou sur l'estimation bayésienne associée à des modèles Markoviens.

Dans la première catégorie de méthodes [HS81, NE86, CK83], l'idée est de rechercher les solutions qui satisfont l'ECMA tout en respectant une certaine forme de continuité, c'est-à-dire qui minimisent une fonction faisant intervenir de manière additive les deux aspects du problème, la conservation de l'intensité et la continuité spatiale. La minimisation est effectuée de manière déterministe (algorithme de Gauss-Seidel par exemple). Pour prendre en compte les discontinuités de mouvement, [NE86] propose un terme de lissage "orienté": celui-ci n'est effectué que sur la composante perpendiculaire au gradient

spatial de l'intensité, la vitesse dans la direction de ce gradient étant fournie par l'ECMA. [Coh93] propose une méthode variationnelle non-linéaire pour suivre le mouvement de structures déformables.

Il est possible d'obtenir une formulation markovienne des schémas d'estimation précédents. La prise en compte naturelle d'observations et de primitives à estimer de natures différentes, ainsi que l'introduction aisée d'une gamme de modèles plus large (en particulier non-linéaires) constituent deux avantages de cette approche. De plus, les interactions entre les différentes variables se décrivent facilement avec des potentiels locaux. Ainsi, pour répondre au problème des discontinuités de mouvement, des étiquettes binaires de discontinuité ("line process" en anglais) , caractérisant la présence ou l'absence de frontière de mouvement, sont placées entre chaque couple de pixels voisins de l'image. Ces étiquettes sont alors estimées en même temps que les vecteurs de mouvement [HB93, KD90]; le critère du Maximum a Posteriori (MAP) est généralement utilisé. En outre, dans [HB93], un test de vraisemblance permet de détecter les pixels pour lesquels l'ECMA n'est pas valide. Le terme de conservation de l'intensité pour ces pixels n'est alors pas pris en compte dans le processus d'optimisation globale.

L'emploi d'estimateurs robustes constitue une alternative à la détection explicite des lieux de l'image où l'une ou l'autre des hypothèses n'est pas vérifiée. Celle-ci consiste à remplacer la norme quadratique, qui accorde une trop grande influence aux erreurs importantes, par un estimateur plus "tolérant" vis-à-vis de celles-ci. Ces erreurs importantes, ou mesures "aberrantes", qui se produisent quand les hypothèses ne sont pas vérifiées, sont dénommées *"outliers"* en anglais. Parmi les estimateurs robustes, la classe des estimateurs redescendants, pour laquelle la norme est bornée, a été principalement employée. Cette approche a été utilisée avec succès par Black [Bla92, BA91], qui montre par ailleurs que l'approche robuste est équivalente à l'emploi d'un processus ligne continu (à valeurs entre 0 et 1) [Bla92, BR94]. Notons que l'approche robuste a été récemment exploitée pour estimer des modèles paramétrés de mouvement [BA93b, OB94, BK94]. Nous présenterons ces méthodes dans le prochain chapitre.

Le traitement des régions de découvrement et de recouvrement, dans lesquelles l'hypothèse de conservation de l'intensité n'est pas vérifiée, est implicitement pris en compte par l'approche robuste, mais peut également faire l'objet d'un traitement spécifique [HB93, TB89].

Des tentatives ont également été faites pour prendre en compte les variations de l'illumination qui ne sont pas dues au mouvement. Par exemple, dans [FM91], un modèle affine de la variation de l'illumination est utilisé localement autour d'un point. Dans [CK83], une séquence d'images radiographiques est traitée. Dans ce contexte, la grandeur $\frac{dI}{dt}$ a une signification particulière: elle peut s'interpréter comme la contraction ou l'expansion (du cœur par exemple) dans la direction perpendiculaire au plan image. Cette grandeur est alors considérée comme une inconnue supplémentaire, supposée varier de manière continue dans l'image. Un algorithme similaire à [HS81] permet de résoudre le problème. Enfin, dans [KT94b], l'hypothèse de conservation du gradient spatial de l'intensité, qui est une

grandeur très peu sensible aux variations d'illumination, est utilisée directement dans un schéma de régularisation identique à celui présenté dans [KD90]. Les résultats sont bons et identiques aux résultats obtenus en utilisant la conservation de l'intensité, mais dénotent comme prévu une robustesse plus importante face à des variations d'illumination non uniformes.

Enfin, notons que la validité de l'ECMA dépend des fréquences spatiales de l'intensité. Dans le cas de déplacements importants, il est généralement nécessaire d'employer une technique multirésolution. L'idée sous-jacente à ces méthodes est que les basses fréquences spatiales de l'intensité sont mieux adaptées pour évaluer les grands mouvements, alors que les hautes fréquences le sont pour estimer les petits déplacements. La stratégie descendante "du plus grossier au plus fin" (*"coarse-to-fine"*) dans laquelle les mouvements estimés à une résolution -une fréquence spatiale- donnée sont projetés sur la résolution supérieure pour être estimés plus finement [Enk88, BAHH92], est la plus fréquemment employée. Dans le cas des méthodes par mesure de similarité que nous allons maintenant décrire brièvement, les algorithmes multigrilles ont surtout pour but de réduire la complexité calculatoire [Ana89].

Méthodes par mesure de similarité

Ces méthodes utilisent une version discrétisée de l'équation (2.1). On cherche alors les champs de déplacement qui vérifient:

$$I(\mathrm{p} + \vec{\mathrm{v}}(\mathrm{p})\delta t, t + \delta t) - I(\mathrm{p}, t) = 0 \tag{2.4}$$

L'intérêt de cette expression, communément appelée différence d'images déplacées, est qu'elle ne fait pas intervenir les dérivées de la fonction intensité, ce qui la rend moins sensible au bruit. Le plus fréquemment, le mouvement est supposé constant dans un bloc autour du point p. On recherche de manière exhaustive dans un ensemble de déplacements discrets prédéfinis celui qui minimise une norme (L_2 ou L_1) du terme de gauche de l'équation (2.4) sommée sur le bloc. C'est une méthode très classiquement utilisée en codage d'image, pour laquelle le déplacement mesuré est le plus souvent attribué à tous les points du bloc de la mise en correspondance (ou *"block matching"* en anglais). La reconstruction de l'image après décodage présente alors parfois un effet de bloc, visuellement gênant, et que de nombreuses études cherchent à réduire [NL91].

Dans [Ana89], les valeurs du critére d'appariement calculées pour chaque déplacement discret constituent une surface, dont l'analyse permet de prendre en compte la fiabilité du minimum mesuré ainsi que dans une certaine mesure les frontières de mouvement. Par exemple, ces dernières peuvent être repérées lorsque la surface exhibe deux minima locaux importants, à condition que la mesure n'ait pas été affectée par la frontière de mouvement[1]. Pour obtenir une précision inférieure au pas de discrétisation, le champ des

1. L'emploi d'un estimateur robuste dans la fonction de corrélation permet généralement de préserver ces minima [Bla92]

déplacements est obtenu par régularisation du champ discret tenant compte de la fiabilité de ces déplacements discrets et des discontinuités. [Sin92, GD94] proposent une approche utilisant des mesures de fiabilité similaires en interprétant la surface d'erreur comme une surface de densité de probabilité.

2.2 Détection du mouvement

L'un des problèmes centraux en vision par ordinateur concerne l'élaboration de systèmes de détection d'objets mobiles dans une scène. Ceux-ci ont longtemps été restreints au cas où la caméra est fixe. A partir d'une séquence d'images, il s'agit alors de déterminer les masques binaires caractérisant la présence ou l'absence de mouvement en chaque point de l'image. Ceux-ci peuvent s'obtenir en examinant, à l'aide par exemple de tests statistiques, les différences temporelles de l'intensité.

Cependant, lorsque le capteur est mobile, le problème de détection de mouvement est beaucoup plus complexe et ne peut se réduire à l'analyse de ces différences. Le mouvement de la caméra induit en chaque point de l'image un déplacement apparent. D'autres méthodes plus sophistiquées doivent alors être considérées.

2.2.1 Détection du mouvement avec caméra fixe

Dans de nombreux cas, la détection du mouvement est réduite à la détection de changements temporels. En effet, dans une séquence d'images acquises avec une caméra fixe, le mouvement d'un objet génère dans les régions texturées de cet objet des changements temporels de l'intensité. Cependant, les notions de détection de changement temporel et détection de mouvement ne sont pas équivalentes. Dans les régions uniformes de l'objet en mouvement, aucun changement n'est observé. Inversement, des variations de l'intensité peuvent avoir d'autres causes que le mouvement, comme le changement des conditions d'illumination.

Malgré tout, le simple seuillage de la carte des différences d'intensité entre deux images successives permet d'obtenir un masque approximatif de l'objet. Afin de diminuer les taux de fausses alarmes, on peut considérer des tests statistiques. [HNR84, Lal90, ML78, Let93] utilisent des tests paramétriques basés sur une fonction de vraisemblance et sur une modélisation locale de la fonction intensité (constante, linéaire, quadratique); [ML78] introduit aussi un test non paramétrique basé sur la statistique du rang; enfin, [AKM93] considère des tests dits significatifs (*"significance tests"* en anglais), portant sur la distribution locale de la carte des différences temporelles de l'intensité. Par ailleurs, une régularisation basée sur un modèle markovien permet également de réduire les fausses alarmes, et améliore les masques de détection dans les régions mobiles comprenant des régions uniformes [Lal90, AKM93, Let93].

La localisation précise des objets en mouvement à un instant donné est un problème en soi. Pour l'illustrer, prenons le cas simple d'un objet mobile dont les surfaces projetées dans l'image à deux instants différents ne se chevauchent pas. La détection de changement temporel effectuée entre ces deux images produira deux régions "mobiles": la région découverte par l'objet, correspondant à la position de l'objet dans la première image, et la région recouverte, correspondant à sa position dans la seconde image. Pour résoudre ce problème, il faut prendre en compte plusieurs images successives [Lal90, Jai85], ou disposer d'une image de référence qui caractérise "le fond" de la scène en l'absence d'objets mobiles. Dans ce dernier cas, la détection réalisée entre chaque image courante et l'image de référence donne directement la position de l'objet mobile [KvG90, BAD93]. Le problème inhérent à ces méthodes réside dans la mise à jour de l'image de référence pour la prise en compte d'éventuels changements dans la composition du fond de la scène ou des conditions d'illumination.

Les méthodes précédentes sont relativement sensibles à des vibrations de faible amplitude du système d'acquisition. Ceci s'explique aisément. Un déplacement minime d'un contour très contrasté génère des différences temporelles importantes. Dans [LRB93, Let93], l'utilisation d'un horizon temporel très étendu à l'aide d'une décomposition fréquentielle temporelle du signal de l'intensité, permet de faire la distinction entre les mouvements "parasites" dans l'image, qui seront localisés dans les hautes fréquences temporelles, et les mouvements de petits objets mobiles que l'on retrouve à toutes les fréquences. Des bancs de filtres spatio-temporels [AB85] peuvent également être utilisés en détection. Il est intéressant de noter que le cerveau contient des jeux de neurones spécialisés dans la détection de mouvements d'amplitude et de direction données. Les bancs de filtres constituent en fait une modélisation possible de ces cellules biologiques de détection du mouvement. Par ailleurs, notons que dans [SJ89], deux méthodes sont proposées pour améliorer la robustesse de la détection vis-à-vis cette fois-ci des variations d'illumination.

Pour finir, il faut dire que la détection de petits objets ayant généralement des mouvements de faible amplitude a fait l'objets d'études particulières. Dans [CSR83] la transformée de Hough est utilisée pour déterminer, dans une image de différences cumulées, des droites s'interprétant comme les trajectoires des objets. Dans [SB91], des tests d'hypothèses hiérarchiques sont utilisées, alors que dans [LRB93], l'accent est mis sur une décomposition fréquentielle.

2.2.2 Détection du mouvement dans le cas d'une caméra mobile

Dans le cas d'un capteur mobile, la détection est évidemment plus complexe que celle décrite précédemment. Elle s'apparente alors au problème de la segmentation au sens du

mouvement. Ce dernier peut se définir en partie comme la recherche des discontinuités de mouvement. Or, ces dernières correspondent à des frontières séparant les projections dans l'image d'objets ayant des mouvements différents, et/ou à des ruptures de profondeur entre deux surfaces animées du même mouvement rigide et ayant des projections voisines dans l'image. Cette deuxième cause de discontinuités du champ des déplacements ne se produit que si la caméra est en translation et n'est vraiment perceptible que s'il existe des variations de profondeur significatives dans la scène. Si le mouvement de la caméra est une rotation, ou si les surfaces visibles de la scène sont quasiment situées dans un plan, la détection de l'enveloppe des objets mobiles peut se faire en recherchant à l'aide de détecteur de contour les discontinuités d'un champ des déplacements préalablement extrait [TBM85]. Sinon, la détection de mouvement avec caméra mobile devra alors faire la distinction entre les deux types de discontinuités. Pour éviter la reconstruction 3D explicite, certaines hypothèses sont généralement nécessaires et conduisent aux deux grandes classes d'approches que nous décrivons maintenant.

Contraintes sur le champ des déplacements apparents

La première mise en œuvre possible consiste à exploiter des propriétés locales que vérifie le champ des déplacements projetés[2] lorsque la scène est statique. Ces propriétés constituent alors autant de contraintes, qui, lorsqu'elles ne sont pas validées, indiquent la présence d'un objet mobile ou d'une frontière de mouvement.

Par exemple, si la composante rotationnelle du mouvement de la caméra est connue (ou nulle), la détection peut se ramener au traitement d'une séquence dans laquelle le mouvement apparent des zones statiques de la scène n'est dû qu'à la translation de la caméra, et semble provenir d'un point fixe de l'image: le foyer d'expansion (voir figure 2.2). La connaissance de la position de celui-ci dans l'image permet alors d'imposer une contrainte sur la direction du déplacement en chaque point de l'image. Si cette contrainte n'est pas vérifiée, on a nécessairement affaire à un objet mobile. Dans [TP90], elle est appliquée à un champ de déplacements préalablement obtenu par une méthode de mise en correspondance, alors que [Nel91] utilise une approche qualitative basée sur des observations partielles (les composantes de vitesse parallèles au gradient d'intensité, cf. formule (2.3)) et propose un algorithme fonctionnant à une cadence de dix images par seconde. Une méthode pour détecter les mouvements d'êtres animés, qui se caractérisent par des accélérations apparentes dans l'image beaucoup plus rapides que celles dues à la caméra, est également proposée dans ce dernier article.

Insistons ici sur le fait que la méthode illustrée par la figure 2.2 ne procure qu'une condition nécessaire. Une caméra montée sur l'avant d'un véhicule constitue un cas important où cette contrainte sera en partie inopérante. En effet, le mouvement de voitures circulant devant le véhicule portant la caméra –et dans le même sens– pourra valider la

2. En pratique, ces propriétés seront appliquées sur le champ des déplacements apparents.

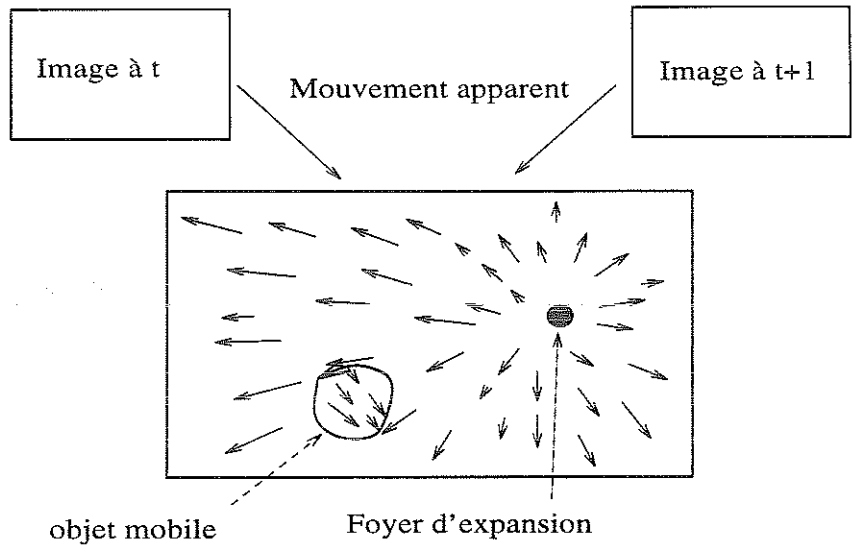


FIG. 2.2 - *Exemple de contrainte sur le mouvement apparent. Dans le cas où le mouvement de la caméra est un mouvement translationnel pur, tous les vecteurs de déplacement apparent des régions statiques semblent émaner d'un seul point, le foyer d'expansion (FOE) (ou converger vers celui-ci).*

contrainte à tout instant. D'autres critères doivent alors être utilisés pour détecter ces objets mobiles [SB94].

Si l'on dispose d'un banc d'acquisition stéréoscopique, on peut estimer la disparité en chaque point de l'image. On montre alors que, entre deux points voisins de l'image correspondant à la projection de deux points 3D soumis au même mouvement rigide, la variation de la disparité et la variation de mouvement sont proportionnelles. Si localement le coefficient de proportionnalité évolue trop rapidement, on est alors en présence d'un objet mobile [TP90].

Le principal inconvénient de ces méthodes est qu'en général elles ne sont efficaces que lorsque les déplacements sont importants, ce qui n'est pas le cas autour du foyer d'expansion et surtout, lorsque la caméra s'immobilise momentanément. En effet, dans ce cas l'algorithme est inexploitable. On peut alors recourir à une autre méthode (par exemple, trouver les points tels que $\vec{V} \neq 0$), mais cela suppose que l'on sache en pratique distinguer le cas où la caméra est mobile, et celui où elle est fixe.

[NSKO94] utilisent un test statistique pour détecter localement les frontières de mouvement. Plus précisément, le champ des vitesses est estimé en minimisant un critère qui tient compte de la corrélation entre les mesures de gradients spatio-temporels de l'intensité, et repose sur l'hypothèse d'une continuité locale[3] du champ des vitesses, exprimé par la prise en compte d'un modèle au premier ordre du mouvement. Lorsque cette dernière

3. La continuité du mouvement est exigée sur un voisinage spatial et temporel.

est vérifiée, le critère suit une distribution du $\mathcal{X}^2$. Ainsi, pour déterminer les frontières de mouvement, il suffit de vérifier si la valeur du critère après minimisation suit localement cette distribution.

Notons que l'on peut utiliser les estimateurs robustes dans le contexte de la détection, dans la mesure où ils sont capables de détecter, dans un système d'équations surdéterminé celles qui ne sont pas consistantes, avec le modèle estimé satisfaisant la majorité de ces équations. Dans [TLS93], ce principe est utilisé. Le problème de la reconstruction 3D de la scène -supposée *a priori* statique- à partir de la position et du déplacement apparent dans l'image est considéré. La projection orthographique des points 3D dans le plan image est utilisée[4]. Cependant, l'objectif étant uniquement de détecter les objets mobiles, la résolution complète du problème de reconstruction n'est pas effectuée. En fait on ne dérive en chaque point de l'image qu'une seule équation ayant pour inconnues certaines composantes du torseur cinématique de la caméra, et faisant intervenir le déplacement apparent dans l'image. La résolution de ce système avec l'estimateur des moindres-carrés médian fait ressortir les points dont le mouvement apparent n'est pas conforme au mouvement rigide de la majorité des points –statiques– de la scène. Il est important de noter ici que les paramètres estimés n'ont pas d'intérêt en soi, et que leurs valeurs sont d'ailleurs éloignées de leurs valeurs réelles connues. Dans [TM93], deux tests statistiques sont utilisés pour atteindre le même but. Cependant, dans la mesure où un modèle de mouvement 2D est utilisé, cette méthode s'apparente plus aux approches que nous allons maintenant décrire.

Détection après compensation du mouvement dominant

Cette seconde catégorie de méthodes est illustrée par la figure 2.3. Elle consiste à modéliser le champ des déplacements apparents induit par le mouvement de la caméra entre deux images I_1 et I_2 par un jeu de paramètres Θ. Celui-ci est alors estimé et utilisé pour générer une image compensée I_2^* définie par:

$$I_2^*(\mathrm{p}) = I_2(\mathrm{p} + \vec{V}_{\widehat{\Theta}}(\mathrm{p})) \tag{2.5}$$

où $\vec{V}_{\widehat{\Theta}}(\mathrm{p})$ désigne le déplacement au point p paramétré par $\widehat{\Theta}$. Le problème de l'extraction des objets mobiles dans la scène se ramène alors en première approximation à la détection du mouvement avec caméra fixe dans la séquence compensée. Cependant, dans la mesure où la compensation n'est pas nécessairement parfaite et inclut des opérations qui ne sont pas effectuées dans le cas fixe (interpolation par exemple), les outils développés dans ce dernier cas sont rarement adaptés au problème présent.

Dans [BBH*89], le processus de compensation se fait à l'aide de deux modules. Le premier extrait le champ des déplacements apparents estimés localement, qui est utilisé par le second module dans l'estimation des coefficients d'un modèle affine Θ de ce champ.

4. Il est intéressant de noter que pour la projection orthographique, les discontinuités liées aux ruptures de profondeur ne proviennent plus de la translation de la caméra, mais de sa rotation.

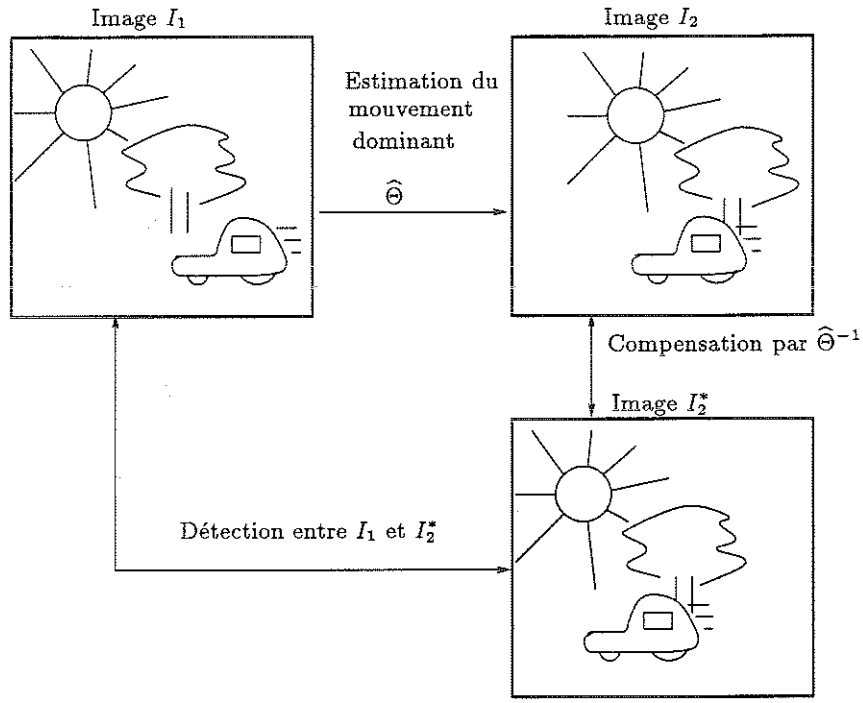


FIG. 2.3 - *Principe de la détection par compensation du mouvement dominant.*

Ce dernier sert alors à déplacer la première image vers la seconde. Les deux modules sont alors à nouveau employés pour calculer les incréments des déplacements, ce qui permet d'affiner l'estimation. Ce schéma est intégré au sein d'une stratégie descendante multirésolution. À la fin du processus, les déplacements locaux estimés qui diffèrent du déplacement paramétré par $\widehat{\Theta}$ indiquent la présence d'objets mobiles.

Dans [IRP92], une stratégie multirésolution et incrémentale est également utilisée, mais le modèle est directement estimé à partir des gradients spatio-temporels de l'intensité. Pour améliorer cette estimation ainsi que la détection, une image de référence constituée de plusieurs images moyennées dans le sens du mouvement estimé est construite. Dans celle-ci, la partie de l'image à laquelle correspond le mouvement estimé (supposé correspondre à celui induit par le déplacement de la caméra), reste nette, tandis que les autres régions dont font partie les objets mobiles deviennent floues. La mesure utilisée pour diffencier ces deux classes de régions, est constituée des déplacements normaux (formule (2.3)) moyennés spatialement et calculés entre l'image de référence et la nouvelle image.

Il est intéressant de noter que les deux types d'approches que nous venons de présenter peuvent être complémentaires. Dans le cas où le mouvement de la caméra est quelconque et inconnu, la méthode par compensation fournit un moyen de nous replacer dans des conditions suffisantes pour appliquer le principe de détection présenté sur la figure 2.2.

Supposons que l'on choisisse le modèle de mouvement (2D) décrivant le mouvement apparent d'un plan 3D, et que le modèle estimé par la méthode ci-dessus [IRP92] soit celui d'une surface plane statique de la scène. Les autres parties statiques n'appartenant pas à ce plan apparaîtront mobiles dans la séquence compensée. Cependant, les coefficients du modèle estimé contiennent implicitement la composante du mouvement apparent due à la rotation de la caméra, qui est globale et indépendante du plan particulier sur lequel sont estimés les paramètres. Cette composante sera donc éliminée (du mouvement de toutes les régions) dans la séquence compensée, et il ne restera alors pour les parties statiques de la scène qu'une composante reliée à la translation de la caméra [IRP94].

2.3 Segmentation du mouvement

Le mouvement apparent dans une séquence d'images constitue une information visuelle essentielle pour structurer une image. Cet aspect est souvent mis en évidence au moyen d'expériences psychovisuelles dans lesquelles les autres facteurs humains de reconnaissance (forme ou contraste, par exemple) sont éliminés. Par exemple, dans une image formée uniquement de pixels d'intensité aléatoire, nous ne distinguons aucune structure. Cependant, si une partie de cette image se déplace de façon homogène, nous groupons les points de même mouvement, ce qui nous permet d'induire la forme de la région en mouvement [Bra74, Nak85].

L'objectif de la segmentation au sens du mouvement consiste donc à fournir une partition de l'image en régions visuellement importantes ayant des caractéristiques ou des propriétés de mouvement différentes. De plus, il est intéressant d'obtenir une segmentation qui reste stable et cohérente au cours du temps.

Le choix des propriétés influence beaucoup la résolution de ce problème. Par exemple, dans [TB89], il s'agit de diviser l'image en quatre classes: les régions statiques, les régions en mouvement, les zones découvertes et recouvertes. Dans le cas de la détection que nous avons présentée, le but est de rechercher les régions mobiles de la scène, ce qui peut être fait en localisant certaines discontinuités du champ des déplacements apparents [TP90, Nel91]. Cependant, le plus souvent, l'objectif à atteindre est une partition complète en régions à l'intérieur desquelles le champ des vitesses est continu. Une possibilité consiste à exploiter une segmentation spatiale de chaque image de la séquence, puis à effectuer la mise en correspondance des régions ainsi segmentées pour analyser le mouvement. Dans [Hea93], une segmentation hiérarchique basée sur des modèles d'intensité est utilisée. La mise en correspondance de ces régions peut se faire entre les différentes échelles des segmentations hiérarchiques aux deux instants considérés, ce qui rend l'algorithme plus robuste face à l'instabilité de la segmentation spatiale au cours du temps, et permet de prendre en compte des déplacements très importants, supérieurs à la taille de la région mobile. Dans [PS94], la segmentation hiérarchique est obtenue au moyen d'opérateurs de morphologie mathématique. Cependant, ces méthodes fournissent généralement une

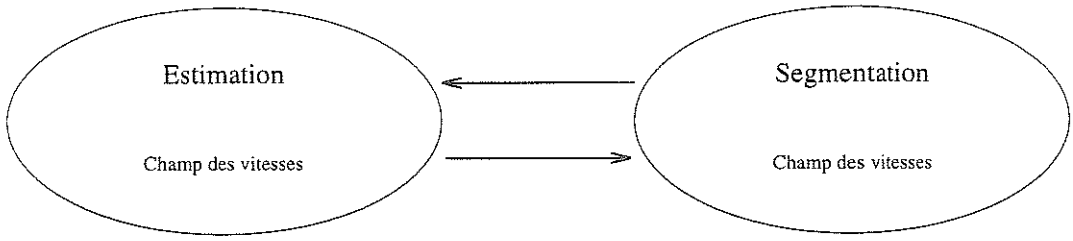


FIG. 2.4 - *Interdépendance entre estimation du champ des vitesses et segmentation du champ des vitesses. Une bonne estimation du mouvement requiert la connaissance d'une bonne segmentation, et réciproquement.*

sursegmentation de l'image. Même si le contenu lumineux de l'image est essentiel pour estimer les déplacements, il n'est pas aisé de faire coopérer cette information avec le mouvement pour segmenter l'image.

Dans l'état de l'art sur l'estimation du mouvement, nous avons présenté des schémas [KD90, HB93], qui permettent de déterminer les discontinuités de mouvement en même temps que l'estimation des déplacements, et ceci pour éviter le lissage intempestif du champ des déplacements. Les discontinuités obtenues par ces méthodes sont bruitées, généralement non fermées, et sont donc insuffisantes pour donner une description de la scène (ou tout au moins de l'image) en termes de régions de mouvement cohérent.

La segmentation au sens du mouvement requiert donc une approche spécifique différente de l'estimation. La façon la plus simple d'introduire la notion de régions est alors de considérer un modèle de mouvement qui sera attaché à chaque région, et assurera de manière implicite la continuité du champ des déplacements dans la région. Cependant, le problème des discontinuités reste entier. L'estimation du mouvement, à travers l'identification des modèles, sera d'autant meilleure que les frontières de mouvement, implicitement présentes dans la partition en régions, seront exactement déterminées. Réciproquement, la segmentation sera d'autant plus précise que les modèles seront estimés correctement. Dès lors, il existe une interdépendance profonde entre estimation et segmentation, qui se traduit fréquemment par des itérations successives entre ces deux processus (figure 2.4).

Dans de nombreux articles, l'accent est mis sur le fait que la segmentation doit fournir explicitement [Adi85], ou implicitement [MB87], une interprétation 3D de la scène. Le modèle de mouvement apparent choisi doit alors refléter au mieux la projection du mouvement rigide d'une surface 3D. L'expression de ce modèle pour une surface quelconque animée d'un mouvement rigide fait intervenir la profondeur de chaque point de cette surface. En supposant que cette profondeur est connue [PR90, LF94], ou que la surface est un plan (approximation par facettes) [Adi85, MB87], on obtient des modèles de mouvement du second ordre (quadratique) en x et y. Les termes d'ordre deux, pour être correctement estimés, nécessitent de se placer sur des régions de grande taille de l'image, ce que l'on rencontre peu fréquemment. Pour gagner en robustesse sans trop perdre en repésentativité, le modèle affine est préférable [NL92, BF93, BS87]. Même si le lien direct

avec l'interprétation 3D n'existe plus, il faut souligner que dans le cas d'un plan, toute l'information sur le mouvement (rigide) et l'orientation de ce plan est contenue dans le modèle affine. Il est alors possible de fournir une interprétation qualitative 3D de la scène *après* la phase de segmentation [BF93], ou d'estimer les paramètres de mouvement et d'orientation 3D en utilisant les coefficients du modèle ainsi que leurs dérivées temporelles premières [Sub89, Mey93], ou les modèles estimés sur plusieurs surfaces planaires différentes du même objet rigide [NL92]. Le modèle constitue de plus une bonne approximation de la projection de nombreux mouvements rencontrés en analyse dynamique. Il est utilisé dans de nombreux schémas de segmentation [ASB94, BF93, WK93, AW94]. On rencontre également des modèles partiels: affine à quatre paramètres [NL91, Hoe89], second-ordre à trois paramètres [DP91].

Une différence importante entre les méthodes concerne les données à partir desquelles est effectuée la segmentation. Dans [Adi85, AW94, BBH*89], les vecteurs de déplacement estimés préalablement sont utilisés. Ceci est un handicap de ces méthodes: si les vecteurs sont estimés localement, ils risquent d'être fortement bruités. Si une régularisation importante est employée, les vecteurs risquent alors d'être moyennés aux frontières de mouvement. Ces dernières seront alors "brouillées", ce qui ne facilitera pas la tâche de la segmentation. Les approches qui utilisent directement la projection du déplacement sur les gradients spatio-temporels de l'intensité (l'ECMA), ou la différence d'image déplacée (relation(2.4)), seront plus robustes [BF93, Sti93, CST94, IRP92, PR90].

Enfin, le cadre méthodologique adopté pour conjointement estimer le mouvement et former les différentes régions de la segmentation permet de regrouper les algorithmes en trois catégories.

La première est basée sur des méthodes de regroupement ou de croissance de régions. Dans [Adi85], le but est de reconstruire et de segmenter la scène 3D en objets de mouvements rigides distincts. Chaque objet est supposé se décomposer en facettes planaires. En partant d'un champ dense estimé, les pixels sont regroupés à l'aide d'une transformée de Hough en régions correspondant chacune à une surface planaire. Ensuite, les différentes régions de mouvement 3D identiques sont fusionnées au moyen d'un test d'hypothèses. Les valeurs des paramètres de mouvement (torseurs cinématiques 3D) estimées par la méthode sont très bruitées. Ceci est dû à l'ambiguïté inhérente au champ des déplacements apparents estimé, plusieurs mouvements (3D) très différents étant susceptibles de produire des champs de déplacements projetés théoriques similaires [Adi89]. Dans [AW94], un champ initial de déplacement est calculé localement à partir d'une méthode de moindres-carrés. Ce champ est alors divisé en blocs sur lesquels un modèle de mouvement affine est estimé. Ces modèles sont ensuites regroupés à l'aide d'un algorithme de classification (*"K-mean clustering"*) modifié. Enfin, ils sont réestimés sur les supports résultant de l'aggrégation des blocs, et utilisés pour segmenter le champ initial. Le critère retenu pour la segmentation est l'écart entre le vecteur vitesse mesuré et le vecteur vitesse modélisé. Dans [BS87], un modèle affine est estimé au départ sur des petits blocs de l'image, et la fusion des blocs est effectuée à l'aide d'un test de vraisemblance généralisé. Ces méthodes présentent

l'inconvénient de fournir généralement une sursegmentation, l'estimation de modèles de mouvement sur de petites régions n'étant pas fiable. Ce dernier problème est généralement résolu par élimination en cours de traitement des régions de taille trop petite. Un autre point faible parfois rencontré concerne l'instabilité des segmentations obtenues dans le temps ainsi que d'une expérience à l'autre (importance de l'ordre de fusion),

La deuxième classe d'approches est constituée par les méthodes qui procèdent par extraction un à un des différents mouvements et des régions qui leurs sont rattachées [PR90, IRP92, BBH*89, WK93]. Le mouvement apparent dans la séquence induit par le mouvement de la caméra sur les régions statiques de la scène est supposé être dominant, les régions de mouvement indépendant n'occupant qu'une faible partie de l'image. Après avoir estimé, généralement par une technique de moindres-carrés, les paramètres décrivant ce mouvement dominant, un seuillage sur des mesures de mouvement ou d'intensité permet de détecter les régions de mouvement non conforme au mouvement dominant. Le processus d'estimation et de détection est alors réitéré sur les régions pour lesquelles aucun des mouvements estimés jusqu'à l'étape courante n'est satisfaisant. Dans [PR90], la segmentation est un peu plus sophistiquée, et se fait en comparant les différences d'images déplacées, moyennées localement et calculées avec les deux modèles de mouvement estimés fournissant la meilleure compensation. L'hypothèse portant sur la présence d'un mouvement dominant est très forte, et représente le principal handicap de ces méthodes dans lesquelles le problème de la segmentation n'est pas explicitement formulé et où les différentes régions ne sont pas traitées de manière équivalente.

Dans la troisième approche, la segmentation est posée comme un problème d'estimation conjointe des modèle de mouvement et de la segmentation associée à ces modèles, c'est-à-dire comme un problème d'étiquetage statistique non supervisé. Elle repose sur une modélisation markovienne des propriétés *a priori* des masques de segmentation et des relations entre observations et modèles de mouvement [MB87, BF93, Sti93, CST94]. L'optimisation se fait généralement en deux étapes qui sont itérées jusqu'à convergence. La première consiste à effectuer l'étiquetage des pixels avec les différents modèles de mouvement estimés, et la seconde à estimer les paramètres de mouvement, la segmentation étant figée. Cette méthodologie, qui sera utilisée dans cette thèse, sera décrite plus en détail dans le chapitre sur la segmentation.

2.4 Conclusion

Dans ce chapitre, nous avons présenté un état de l'art de trois sujets importants de l'analyse de séquences d'images qui seront traités dans cette thèse: l'estimation du mouvement, la détection du mouvement, et la segmentation au sens du mouvement. Nous avons tout d'abord exposé les problèmes ainsi que les hypothèses nécessaires généralement admises pour les résoudre. Nous avons discuté la validité de celles-ci et souligné les difficultés qu'elles entraînaient lorsqu'elles ne sont pas vérifiées.

Lors de cette présentation, nous avons pu nous rendre compte que les trois sujets sont extrêmement liés, dans le cas où la caméra est mobile notamment. Néanmoins, la spécificité de chaque problème a été dégagée. Nous avons classé les différentes approches en catégories qui répondaient à un problème particulier (variations d'illuminations, détection des petits objets), ou faisaient appel à une même méthodologie.

Il existe bien sûr d'autres sujets de vision dynamique que nous n'avons fait qu'évoquer. On peut rappeler le problème général de l'estimation de la structure et du mouvement 3D à partir du mouvement apparent, ou des problèmes plus spécifiques, comme le calcul du temps à collision, c'est-à-dire le temps que mettrait le capteur pour atteindre un point de la scène si sa vitesse restait la même [MB92, Sub90, TS91], ou le calcul du torseur cinématique de la caméra [Sun92, IRP94, NL92, HW88]. Le suivi dans les images de caractéristiques (points, contours) et l'analyse de leurs trajectoires 2D ou 3D est un sujet de plus en plus abordé du fait de l'importance qu'il revêt, que ce soit en vision dynamique [MRM94], ou en vision active [CBBJ94]. Le suivi peut également s'effectuer sur des régions 2D de l'image [MB94a, KWM94]. Dans [MB94a], le mouvement est estimé sur la région suivie avec une méthode différentielle, ce qui évite le problème difficile de la mise en correspondance, notamment dans les régions fortement texturées. Enfin, nous citerons en dernier lieu l'analyse de mouvement de structures déformables [KH94, PH91], dont l'imagerie biomédicale constitue l'un des principaux champs d'applications.

Dans la suite de ce manuscrit, nous aborderons le problème de l'estimation, de la détection et de la segmentation du mouvement à l'aide de modèles de mouvements paramétrés. Nous justifierons l'emploi de tels modèles dans le chapitre qui va suivre, et nous exposerons les motivations qui ont conduit aux différents algorithmes que nous avons définis.

Chapitre 3

Estimation robuste multirésolution de modèles paramétriques de mouvement

Dans ce chapitre, nous abordons le problème de l'estimation de mouvement entre deux images. Le principe général de l'estimation du mouvement apparent 2D consiste à estimer en chaque point de l'image un vecteur vitesse. On obtient alors un champ dense. Comme nous l'avons souligné précédemment, le problème de l'estimation du mouvement est un problème mal posé. Les méthodes globales de calcul du champ des déplacements apparents qui utilisent des contraintes de régularisation [HS81, Nag83, HB93, BA93a] reposent sur une notion "générale" de continuité spatiale (complète ou par morceaux suivant les méthodes) dans l'image, mais sans lien explicite avec les propriétés "physiques" (réelles) du champ des déplacements projetés. Les solutions obtenues avec ces méthodes n'ont donc pas de signification précise, des contraintes de régularisation différentes fournissant des solutions différentes. Par exemple, dans [EC84], plusieurs termes de lissage sont proposés pour contraindre les solutions au problème de l'estimation des déplacements sur un contour. Celui retenu possède de bonnes propriétés mathématiques vis-à-vis du problème de convergence. Cependant, les champs estimés sur des exemples synthétiques ne correspondent pas nécessairement aux déplacements réels, bien que selon [EC84], elles soient perceptuellement valides. D'autres méthodes pour obtenir des champs denses utilisent explicitement ou implicitement un modèle de mouvement constant localement [FJ90, Ana89]. Ces méthodes souffrent de ne pas pouvoir prendre en compte la variabilité locale des vecteurs vitesses. Des modèles locaux plus complexes (affines) ont alors été pris en compte dans [FM91, Nag87]. Dans [ON94], une méthode est proposée pour estimer directement le champ des déplacements et ses dérivées spatio-temporelles. Même si la méthode requiert le calcul de dérivées d'ordre trois de la fonction intensité sur un voisinage restreint, les résultats obtenus sont bons. De manière générale cependant, la localité de ces méthodes les rendent peu robustes aux violations des contraintes évo-

quées dans le chapitre précédent. Les résultats présentés dans [BFB94] montrent que les algorithmes pour calculer des champs denses ne sont pas toujours très précis. Des progrès significatifs ont cependant été réalisés ces dernières années [Bla94, ON94].

Par ailleurs, il faut noter que les articles dans lesquels une analyse locale du champ dense est effectuée sont surtout à caractère théorique [LP80, WU85, Sub87]. [Adi85] représente l'un des articles où un champ dense des déplacements mesuré est effectivement utilisé. Les résultats montrent que la qualité du champ estimé est déterminante pour reconstruire la scène. Généralement, un champ dense de mouvement 2D constitue l'entrée de modules d'interprétation qui se basent sur des modèles plus globaux [BBH*89, AW94]. Bien que [BBH*89] soutient qu'il est préférable d'estimer un modèle de mouvement à partir du champ des déplacements car tous les pixels sont alors traités de façon équivalente, nous pensons au contraire avec [WK93] qu'il est plus robuste d'extraire les paramètres d'un tel modèle directement à partir de l'image. Le choix d'un modèle de mouvement dépend normalement de trois facteurs:

1. du modèle utilisé pour décrire la projection d'un point de l'espace 3D dans le plan image. Le modèle orthographique constitue une bonne approximation de la projection lorsque les objets sont éloignés ou que l'angle de vue est petit. Dans les autres cas, la projection centrale est un modèle plus exact du système optique de la caméra.

2. du modèle de mouvement tridimensionnel de l'objet. L'hypothèse d'un mouvement rigide est souvent retenue.

3. de la description analytique de la surface visualisée par la caméra. Par exemple, lorsque la variation de la profondeur d'un objet est relativement faible par rapport à la distance séparant cet objet de la caméra, un plan constitue (du moins dans un but d'analyse de la scène) une bonne approximation de sa surface.

Une première approche consiste à conserver dans l'analyse les paramètres des modèles 3D (torseurs cinématiques, profondeurs). Cependant, cela revient à traiter le problème général, mais très complexe, de la reconstruction 3D.

La seconde approche que nous avons retenue consiste donc à faire dans un premier temps une analyse 2D de l'image. Plus précisément, nous supposerons que la scène se décompose en différentes régions dans lesquelles le mouvement peut être décrit à l'aide de modèles paramétrés 2D, en l'occurrence des modèles polynomiaux fonction des coordonnées (x, y) des points dans l'image. Le premier intérêt de ces modèles est qu'ils constituent une représentation compacte du mouvement apparent, ce qui est intéressant par exemple dans les problèmes de codage bas débit. De plus, cette représentation compacte contient implicitement l'essentiel de l'information relative au mouvement et à la structure de la surface 3D qui se projette sur la région où est calculé le modèle. Par exemple, dans le cas de la projection orthographique, le mouvement apparent d'un plan correspond à un

modèle affine 2D, le mouvement apparent d'une surface quadratique à un modèle du second ordre. Avec la projection centrale, le mouvement apparent d'un plan rigide revient à un modèle quadratique à huit paramètres libres. Une segmentation du mouvement dans l'image à l'aide de ces modèles fournit donc indirectement une interprétation 3D de la scène. De ce fait, le choix de ces modèles s'est déjà avéré judicieux aussi bien pour segmenter l'image en régions homogènes au sens du mouvement apparent [BF93, WK93, AW94], pour extraire et coder l'information temporelle dans des schémas de codage à compensation de mouvement [Hoe89, NL91], pour fournir une mesure du mouvement apparent [ZQY89], pour suivre des objets en mouvement le long de la séquence [MB94a], pour modéliser le mouvement global d'objets par ailleurs déformables [BBDM94], pour calculer le temps avant collision [MB92, Sub90] ou pour caractériser le mouvement 3D des objets dans la scène, que ce soit qualitativement, [BF93], ou quantitativement, [NL92, Mey93]. Le deuxième intérêt de ces modèles est que l'on peut les estimer de façon peu coûteuse, ce qui est particulièrement intéressant si l'on veut les introduire dans des schémas de vision active [SBC94]. Cependant, le point crucial de ces modèles est d'en obtenir une estimation fiable et précise.

Il est désormais acquis en analyse du mouvement dans une séquence d'images par des approches différentielles, c.à.d. utilisant les gradients spatio-temporels de l'intensité, que la mesure du mouvement est grandement améliorée si l'on adopte une technique d'estimation multi-résolution. On ne considère pas seulement la séquence d'images à sa résolution d'acquisition, mais on construit à partir de chaque image une pyramide d'images successivement filtrées et sous-échantillonnées. Ceci a été étudié et validé pour la détermination de champs denses de vitesses apparentes, [BAK91, Enk88, KD88]. On obtient ainsi des mesures correctes même en présence de mouvements de grande amplitude, ou de distribution d'information de gradients d'intensité irrégulière dans l'image. Plus récemment, des techniques d'estimation aux moindres-carrés multirésolution de modèles de mouvement ont été proposées [CV90, BAHH92, MB92]. Cependant cela ne suffit pas, car pour être opérationnel, ce calcul multi-résolution doit être effectué sur des zones de taille suffisante, le nombre de données étant divisé par quatre à chaque passage à une résolution plus grossière. Dans [CV90] par exemple, il est recommandé d'utiliser des régions de taille 70×70. Or, les modèles de mouvement utilisés ne peuvent rendre compte que d'un seul mouvement, et dans les scènes complexes qui nous intéressent, plusieurs mouvements peuvent alors être simultanément présents sur le support d'estimation choisi arbitrairement. L'estimation globale sur toute l'image (ou sur des blocs de taille convenable) d'un modèle de mouvement s'en trouve très vite perturbée. Dans [BAHH92], il est postulé que le mouvement global perçu résulte du déplacement du capteur, les projections des objets mobiles dans la scène ne couvrant qu'une portion très faible de l'image. Une latitude sensiblement plus grande sur ce point est autorisée dans [BBHP90] pour des situations de transparence, mais les résultats présentés ne mettent en œuvre que des modèles de mouvement

constants. Dans [MB92], l'estimation est effectuée par région, ces régions étant issues d'une phase préalable de segmentation au sens du mouvement garantissant la présence d'un mouvement unique dans chacune des régions considérées.

La question du support d'estimation est donc centrale, si l'on ne veut pas introduire de restrictions fortes sur les situations traitées, ou ajouter une phase préalable de segmentation explicite (c.à.d., une partition de l'image), étape relativement lourde et sujette à des performances variables sur des scènes complexes. D'un coté, une région importante est nécessaire pour d'une part contraindre suffisamment la solution et répondre ainsi au problème de l'"ouverture" évoqué au chapitre précédent, et d'autre part fournir une solution moins sensible au bruit. D'un autre coté cependant, plus la taille de la région est grande, moins le modèle 2D est susceptible de décrire correctement le mouvement de tous les points de la région. En particulier, la probabilité que la région contienne plusieurs mouvements sera plus importante. Dans [JB93], ce problème lié au choix de la taille du support d'estimation est appelé le problème de l'ouverture généralisé (*"Generalized aperture problem"*).

Le problème posé est donc celui de la recherche d'un mouvement dominant en présence de mouvements dits secondaires. Lors de l'estimation de ce mouvement dominant, les mesures locales correspondant aux autres mouvements présents dans le support, peuvent être considérées comme des mesures aberrantes ou *"outliers"*, c'est à dire des observations qui ne suivent pas le modèle de bruit que l'on s'est fixé sur les données. Ceci nous a donc conduit à considérer la classe des estimateurs robustes [Hub81, Rou84] récemment introduite dans le domaine de l'analyse d'image, [MMR91, Bla92, JMB91]. Parallèlement à nos recherches, d'autres méthodes plus au moins similaires à celle que nous allons présenter ont été proposées.

Dans [PR90], l'estimation de mouvement se fait par affinements successifs des valeurs estimées. À chaque étape, l'inverse du résiduel de l'équation de mesure en chaque point est calculé et sert à pondérer la confiance accordée aux observations en ce point pour l'estimation de l'incrément suivant. Le principe commun à l'ensemble des autres méthodes robustes se retrouve ainsi dans cet article, bien qu'aucune formulation robuste du problème d'estimation ne soit proposée. En effet, la plupart des algorithmes reposent sur la classe des M-estimateurs [DP91, BA93b, BK94, LF94], dont le but est de réduire l'influence des *"outliers"*. Dans [DP91], l'estimation de mouvement se fait conjointement à la segmentation. Une description de l'image en couches est proposée. À chacune d'elle est associé un modèle de mouvement. La contribution de chaque pixel à une couche donnée est évaluée par un coefficient qui reflète l'adéquation des gradients spatio-temporels de l'intensité au modèle de mouvement de cette couche. Ce coefficient sert également à pondérer l'estimation aux moindres-carrés du mouvement. Un critère d'information est alors utilisé pour trouver le nombre de couches et de modèles qui décrira au mieux le mouvement apparent dans la scène. Bien que l'introduction de la notion de couches permette de palier le défaut des algorithmes d'estimation basés sur la recherche des frontières de mouvement en cas d'occlusions "fragmentées" (vision à travers une barrière par exemple),

cette méthode ne prend cependant en compte qu'un modèle de translation 3D et suppose la profondeur constante, soit en fait un modèle de mouvement 2D du premier ordre mais réduit à trois paramètres; de plus elle n'opère qu'en monorésolution et souffre d'un problème d'initialisation, les mouvements des initiales de départ étant initialisés de manière aléatoire (ou répartis sur une grille discrète) dans l'espace des paramètres. Dans [BK94], le critère du M-estimateur est minimisé par un algorithme de descente de gradient. Dans [BA93b], ce même critère est minimisé par une technique de sur-relaxation. Enfin, dans [LF94], une méthode de type Gauss-Newton très similaire à notre algorithme est proposée. Notons enfin que d'autres types de méthodes robustes ont également été exploitées, comme celui des moindres-carrés médian [AS94].

Dans la suite de ce chapitre, nous définirons tout d'abord les modèles de mouvements que nous utilisons. Ensuite, nous présenterons la technique d'estimation multirésolution basée sur les moindre-carrés à laquelle sera comparé l'algorithme que nous avons développé. Après quelques rappels sur les estimateurs robustes, nous décrirons notre méthode d'estimation de modèles paramétriques de mouvement à l'aide d'un estimateur robuste, dont deux versions seront proposées. Enfin, des résultats sur des exemples synthétiques et réels seront montrés et commentés.

3.1 Modèles de mouvement

Nous considérons la classe des modèles de mouvement 2D polynomiaux. En utilisant des notations matricielles, ces modèles peuvent s'écrire sous la forme générale suivante (linéaire en les paramètres (a_j)):

$$V_A(\mathrm{p}_i) = \begin{bmatrix} u_A(\mathrm{p}_i) \\ v_A(\mathrm{p}_i) \end{bmatrix} = B(\mathrm{p}_i)A \tag{3.1}$$

où $A^T = (a_j)$ représente les coefficients des polynômes, $\mathrm{p}_i = (x_i, y_i)$ désigne un point de l'image (ou de la région considérée), $V_A(\mathrm{p}_i)$ le vecteur vitesse au point p_i associé au modèle paramétré par A. B est une matrice dont la forme dépend du modèle choisi, mais dont les coefficients ne dépendent que des coordonnées du point considéré.

Dans notre approche, toute forme de modèles de cette classe peut être envisagée. On peut utiliser un simple modèle constant, c'est-à-dire translationnel $V_A(\mathrm{p}_i) = (a_0, a_1)$, comme des modèles affines et quadratiques, complets ou partiels. Le choix d'un modèle de mouvement peut dépendre de l'application, de la structure des objets et de la taille de leur projection dans l'image, et de la nature des mouvements présents dans la scène (y compris le mouvement du capteur). Cependant, nous considérerons en général le modèle affine complet défini par:

$$\begin{cases} u_A(\mathrm{p}_i) = a_1 + a_2\, x_i + a_3\, y_i \\ v_A(\mathrm{p}_i) = a_4 + a_5\, x_i + a_6\, y_i \end{cases} \tag{3.2}$$

On a dans ce cas :

$$B_i = B(\mathrm{p}_i) = \begin{bmatrix} 1 & x_i & y_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x_i & y_i \end{bmatrix}$$

Ce modèle représente en fait un bon compromis entre représentativité et complexité. Il peut rendre compte d'une large gamme de mouvement 2D (translation, similitude plane, déformation linéaire –cisaillement–), et comme nous l'avons souligné dans l'état de l'art sur la segmentation et précédemment, ce modèle contient l'essentiel de l'information sur l'orientation et le mouvement 3D d'une surface de la scène, [BF93, NL92]. Le principal problème pour introduire des termes quadratiques dans le modèle est que ces derniers requièrent des surfaces importantes pour être estimés correctement. Par exemple, considérons le modèle à huit paramètres qui décrit exactement le champ des vitesses apparentes d'un plan de l'espace 3D subissant un mouvement rigide:

$$\begin{cases} u_A(\mathrm{p}_i) = a_1 + a_2\, x_i + a_3\, y_i + a_7\, x_i^2 + a_8\, x_i y_i \\ v_A(\mathrm{p}_i) = a_4 + a_5\, x_i + a_6\, y_i + a_7\, x_i y_i + a_8\, y_i^2 \end{cases} \tag{3.3}$$

Dans ce modèle, les termes quadratiques a_7 et a_8 sont environ $\frac{1}{f}$ fois plus petit que les termes linéaires, eux mêmes $\frac{1}{f}$ fois plus petit que les termes constants, où f est le rapport entre la distance focale et la taille d'un pixel[1]. Par exemple, la distance focale de la caméra utilisée dans le laboratoire de l'équipe TEMIS est de $12,5$mm, et la valeur de f est approximativement 1000. La forme quadratique du champ des déplacements ne sera donc perceptible que si l'on considère des régions étendues. L'estimation des paramètres quadratiques sera donc très sensible au bruit, comme il est montré dans [NL92], et surtout, non robuste en cas de mouvements multiples. Précisons également que ce modèle décrit le mouvement apparent instantané du plan. Dans la mesure où nous disposons uniquement d'une séquence échantillonnée, lorsque la période temporelle d'échantillonnage est importante ou que le mouvement est rapide, ces formules ne constitueront qu'une approximation du déplacement apparent du plan entre deux images. Pour avoir une description exacte du champ des déplacements (d_x, d_y), il serait nécessaire de considérer le modèle homographique suivant:

$$\begin{cases} d_x(\mathrm{p}_i) = \frac{a'_1 + a'_2\, x_i + a'_3\, y_i + a'_7\, x_i^2 + a'_8\, x_i y_i}{1 + a'_7 x + a'_8 y} \\ d_y(\mathrm{p}_i) = \frac{a'_4 + a'_5\, x_i + a'_6\, y_i + a'_7\, x_i y_i + a'_8\, y_i^2}{1 + a'_7 x + a'_8 y} \end{cases} \tag{3.4}$$

Par la suite, nous noterons p le nombre de paramètres du modèle A que nous aurons choisi.

1. Bien sûr, les rapports exacts dépendent du mouvement particulier du plan et de son orientation. Seule une indication de l'ordre de grandeur dans le cas général est proposée ici.

3.2 Estimation aux moindres-carrés multirésolution

Nous allons tout-d'abord décrire dans ses grandes lignes la technique d'estimation moindres-carrés multirésolution [MB92, BAHH92]. Elle est basée sur l'algorithme de Gauss-Newton, qui conduit à une estimation incrémentale du modèle de mouvement. Pour prendre en compte les grands déplacements, une stratégie descendante classique (*"coarse-to-fine"*) à travers les niveaux de résolution d'une pyramide gaussienne des images est utilisée.

3.2.1 Critère de minimisation

L'hypothèse d'invariance de l'intensité d'un point sur sa trajectoire, soit $\frac{dI}{dt}(\mathrm{p}_i(t), t) = 0$, [HS81], conduit à l'équation bien connue de contrainte du mouvement apparent (ECMA):

$$\vec{V}(\mathrm{p}_i(t), t).\vec{\nabla} I(\mathrm{p}_i(t), t) + I_t(\mathrm{p}_i(t), t) = 0 \tag{3.5}$$

où $\vec{V} = (\frac{dx}{dt}, \frac{dy}{dt})^T$ désigne le vecteur vitesse d'un point[2], $\vec{\nabla} I = (I_x, I_y)^T$ et I_t représentent respectivement le gradient spatial de la fonction intensité I et sa dérivée temporelle partielle. Cependant, des changements globaux d'illumination peuvent survenir, par exemple dans des scènes d'extérieur, ou bien dans des séquences d'images satellitaires (où la fréquence d'acquisition des images est de une image toutes les demi-heures pour Météosat par exemple), aussi bien dans le canal du visible que dans le canal infrarouge (dans ce dernier cas, le changement global est dû à des variations diurnes et interdiurnes de température d'illumination, [SM89]). Pour pouvoir en tenir compte, nous avons choisi d'autoriser une variation globale d'intensité sur la zone de calcul considérée, soit :

$$\frac{dI}{dt}(\mathrm{p}_i(t), t) = -\xi \tag{3.6}$$

où ξ représente donc un paramètre supplémentaire à estimer. On introduit alors la variable $r_1(\mathrm{p}_i)$ suivante (en supprimant la variable temps t pour simplifier les notations):

$$\begin{aligned} r_1(\mathrm{p}_i) &= \frac{dI}{dt}(\mathrm{p}_i) + \xi = \vec{V}(\mathrm{p}_i).\vec{\nabla} I(\mathrm{p}_i) + I_t(\mathrm{p}_i) + \xi \\ &= I_x(\mathrm{p}_i)u(\mathrm{p}_i) + I_y(\mathrm{p}_i)v(\mathrm{p}_i) + I_t(\mathrm{p}_i) + \xi \end{aligned} \tag{3.7}$$

En considérant non pas une fonction $\vec{V}$ quelconque, mais le modèle de mouvement $\vec{V_A}$, on obtient finalement l'expression suivante de $r_1(\mathrm{p}_i)$ en utilisant cette fois des notations matricielles:

$$r_1(\mathrm{p}_i) = \mathcal{X}_i\,\Theta - \mathcal{Y}_i \tag{3.8}$$

2. L'exposant T désigne ici l'opérateur de transposition.

$$\text{où} \begin{cases} \Theta^T = (A^T, \xi) \\ \mathcal{Y}_i = -I_t(\mathrm{p}_i) \quad \text{et} \\ \mathcal{X}_i = \mathcal{X}(\mathrm{p}_i, \nabla I(\mathrm{p}_i)) = (\nabla I(\mathrm{p}_i)^T B_i, 1) \quad \text{avec } B_i = B(\mathrm{p}_i) \end{cases} \tag{3.9}$$

En supposant que les $r_1(\mathrm{p}_i)$ sont des variables aléatoires gaussiennes de moyenne nulle et de même variance, l'estimation du paramètre Θ suivant le maximum de vraisemblance se réduit à une estimation suivant les moindres-carrés. Cependant, la validation de l'équation (3.5) implique que le mouvement apparent soit en relation avec les fréquences spatiales et temporelles de l'intensité. Pour une fréquence temporelle donnée, plus on a de hautes fréquences spatiales, plus l'amplitude mesurable sera petite, compte-tenu de l'approximation au premier ordre faite pour obtenir (3.5). Il faut alors lisser l'image (exploiter les basses-fréquences) pour appréhender les mouvements de grande amplitude. Ceci sera effectué dans l'approche multirésolution proposée plus loin. Néanmoins, dans l'image originale, un meilleur critère que (3.5) lorsque les déplacements sont importants consiste à utiliser la différence d'image déplacée (*"Displaced Frame Difference (DFD)"* en anglais), adaptée à notre cas:

$$DFD(\mathrm{p}_i) = I(\mathrm{p}_i + \vec{V}(\mathrm{p}_i)\delta t, t + \delta t) - I(\mathrm{p}_i, t) = -\xi \delta t \tag{3.10}$$

où δt représente l'intervalle de temps entre deux images. Dans ce qui suit, nous choisirons $\delta t = 1$ pour simplifier les notations. Nous chercherons donc à minimiser la fonction d'erreur suivante:

$$E(\Theta) = \sum_{\mathrm{p}_i \in F} r(\mathrm{p}_i)^2 \tag{3.11}$$

où F désigne le support d'estimation considéré, et avec:

$$r(\mathrm{p}_i) = DFD_A(\mathrm{p}_i) + \xi = I(\mathrm{p}_i + B_i A, t + 1) - I(\mathrm{p}_i, t) + \xi \tag{3.12}$$

Cependant, ce critère n'est plus linéaire vis-à-vis des paramètres à estimer. Pour minimiser $E(\Theta)$, nous allons faire appel à la méthode de Gauss-Newton, qui conduit à une estimation incrémentale des paramètres.

3.2.2 Estimation incrémentale

Ce développement concerne aussi bien le passage d'un niveau de résolution au suivant, que l'estimation incrémentale au sein d'un niveau donné de résolution. La méthode de Gauss-Newton consiste à utiliser à chaque étape une approximation linéaire du résiduel à minimiser en chaque point. Pour minimiser E, supposons que l'on ait une estimation courante $\widehat{\Theta}_k^T = (\widehat{A}_k^T, \widehat{\xi}_k)$ de Θ. Dans ce cas, on peut écrire:

$$\Theta = \widehat{\Theta}_k + \Delta\Theta_k, \text{ soit } \begin{cases} A = \widehat{A}_k + \Delta A_k \\ \xi = \widehat{\xi}_k + \Delta\xi_k \end{cases} \tag{3.13}$$

et l'on a donc :

$$V_A(\mathrm{p}_i) = B_i A = B_i \widehat{A}_k + B_i \, \Delta A_k \tag{3.14}$$

Pour chacun des résiduels $r(\mathrm{p}_i)$, on effectue un développement limité au premier ordre de I à l'instant $t+1$ au point $\mathrm{p}_i + B_i \widehat{A}_k$, soit:

$$\begin{aligned} r(\mathrm{p}_i) &= I(\mathrm{p}_i + B_i \widehat{A}_k + B_i \Delta A_k, t+1) - I(\mathrm{p}_i, t) + \widehat{\xi}_k + \Delta \xi_k \\ &= I(\mathrm{p}_i + B_i \widehat{A}_k, t+1) + \nabla I^T(\mathrm{p}_i + B_i \widehat{A}_k, t+1) \, B_i \Delta A_k + O^2 \\ &\quad - I(\mathrm{p}_i, t) + \widehat{\xi}_k + \Delta \xi_k \end{aligned} \tag{3.15}$$

On obtient alors une approximation E' de la fonction d'erreur E:

$$E(\Theta) = E(\Delta\Theta_k) \simeq E'(\Delta\Theta_k) = \sum_{\mathrm{p}_i \in F} (r'(\mathrm{p}_i))^2 \quad \text{avec} \tag{3.16}$$

$$\begin{aligned} r'(\mathrm{p}_i) &= I(\mathrm{p}_i + B_i \widehat{A}_k, t+1) - I(\mathrm{p}_i, t) + \widehat{\xi}_k \\ &\quad + \nabla I^T(\mathrm{p}_i + B_i \widehat{A}_k, t+1) \, B_i \Delta A_k + \Delta \xi_k \\ &= \mathcal{X}'_i \, \Delta\Theta_k - \mathcal{Y}'_i \end{aligned} \tag{3.17}$$

où:

$$\begin{cases} \mathcal{Y}'_i = I(\mathrm{p}_i, t) - I(\mathrm{p}_i + B_i \widehat{A}_k, t+1) - \widehat{\xi}_k \\ \mathcal{X}'_i = \mathcal{X}(\mathrm{p}_i, \nabla I(\mathrm{p}_i + B_i \widehat{A}_k, t+1)) \end{cases} \tag{3.18}$$

La minimisation de E' par rapport à $\Delta\Theta_k$ donne la solution évidente suivante:

$$\widehat{\Delta\Theta}_k = \left[\sum_{\mathrm{p}_i \in F} \mathcal{X}_i'^T \, \mathcal{X}'_i \right]^{-1} \sum_{\mathrm{p}_i \in F} \mathcal{X}_i'^T \, \mathcal{Y}'_i \tag{3.19}$$

Les dérivées de la fonction intensité sont calculées avec l'opérateur de Sobel[3], et dans les expressions de (3.18), les valeurs de I, I_x et I_y aux points de coordonnées non entières sont calculées par interpolation bilinéaire. Notons que ces deux opérations (dérivation et interpolation) ne correspondent pas à la même modélisation de la surface de la fonction intensité. Pour être cohérent, il faudrait utiliser un seul modèle, par exemple en considérant un interpolateur bicubique [Key81] (qui est cependant plus coûteux en temps de calcul).

3.2.3 Stratégie descendante complétée

Ces estimations incrémentales sont itérées selon une stratégie multirésolution descendante, c'est à dire du niveau de résolution le plus grossier au plus fin. Nous utilisons une pyramide gaussienne de chaque image comportant L niveaux, obtenue selon le schéma de [Bur84], à partir de la résolution initiale d'acquisition (niveau 0). Ainsi les déplacements

3. Nous avons également considéré les masques de [VF92], mais sans noter de réelle amélioration en général.

entre deux images successives sont chaque fois divisés par deux en passant au niveau plus grossier suivant. Normalement, il serait préférable de choisir le nombre de niveaux de sorte que, au niveau le plus grossier, le déplacement maximal soit de l'ordre du pixel. Comme ce déplacement nous est inconnu, nous choisirons en fait comme niveau le plus élevé celui pour lequel la région après sous-échantillonnage reste suffisamment significative. Nous avons choisi le critère suivant: la taille d'une région –à un niveau donné– exprimée en pixels ne doit pas contenir moins de six fois le nombre de paramètres à estimer.

L'estimation se déroule alors de la manière suivante[4]. Au niveau de résolution le plus grossier, $L-1$, une valeur nulle est prise pour l'initialisation de Θ. Avec cette valeur initiale de Θ, l'équation (3.17) est évidemment similaire à l'équation de contrainte du mouvement apparent de la formule (3.5). Or, à ce niveau de résolution, les déplacements sont petits, ce qui reconditionne correctement l'ECMA.

En partant de cette valeur, des raffinements successifs de Θ sont opérés à l'aide de (3.19) au même niveau. Lorsque l'incrément estimé $\widehat{\Delta\Theta}^{L-1}$ devient trop faible ou qu'un nombre prédéfini d'itérations ont été effectuées, les paramètres $\widehat{\Theta}^{L-1}$ sont transmis au niveau directement inférieur, où le processus de raffinement reprend avec, comme estimée initiale, la valeur de $\widehat{\Theta}^{L-1}$ projetée au niveau $L-2$. Ceci est répété de niveau l en niveau $l-1$. L'estimée finale de Θ, $\widehat{\Theta}_{est}$, est alors la valeur $\widehat{\Theta}^0$ obtenue après la dernière itération au niveau 0, le niveau le plus fin. Cet algorithme est schématisé à la figure 3.1.

Dans cet algorithme, on pourrait choisir comme "norme" ($\|\widehat{\Delta\Theta^l}\|_F$) la moyenne des quantités $\|\vec{V}_{\widehat{\Delta A}}(\mathrm{p}_i)\|$ sur le support d'estimation F, où $\vec{V}_{\widehat{\Delta A}}(\mathrm{p}_i)$ est le champ des déplacements incrémental induit par la variation $\widehat{\Delta a_j}$ des paramètres a_j. Cependant, cette opération est très coûteuse en temps de calcul, et nous préférons choisir à la place une combinaison linéaire des termes $\widehat{\Delta a_j}$:

$$\kappa = \sum_{j=1}^{p} s_j \, |\widehat{\Delta a_j}| \tag{3.20}$$

Les poids s_j peuvent s'obtenir de la façon suivante. Si nous notons $\vec{V}_{\widehat{\Delta a_j}}$ le champ modélisé uniquement par $\widehat{\Delta a_j}$, les autres paramètres étant placés à zéro, on peut écrire:

$$\frac{1}{T_F} \sum_{\mathrm{p}_i \in F} \|\vec{V}_{\widehat{\Delta a_j}}(\mathrm{p}_i)\| = s_j \, |\widehat{\Delta a_j}| \tag{3.21}$$

où T_F représente la taille du support F. Si l'on considère par exemple le modèle affine, on obtient:

$$s_1 = s_4 = 1, \;\; s_2 = s_5 = \frac{1}{T_F} \sum_{\mathrm{p}_i \in F} |x_i - x_F|, \;\; s_3 = s_6 = \frac{1}{T_F} \sum_{\mathrm{p}_i \in F} |y_i - y_F| \tag{3.22}$$

4. En fait, dans de nombreux articles [IRP92, BBH*89, PR90, Bla92], une présentation similaire mais différente basée sur la notion de *"warping"* est proposée. Une première estimation de Θ est effectuée à l'aide de l'ECMA (résiduel (3.8)). Celle-ci sert alors à former une image "déplacée" (*"warped image"*) I_w de la première image I_1, qui se rapproche de la deuxième image I_2. L'ECMA est alors à nouveau utilisée entre I_w et I_2, et ainsi de suite. La convergence est atteinte lorsque $I_w \simeq I_2$.

$\widehat{\Theta^{L-1}} \longleftarrow 0$

POUR l = niveau $L-1$ au niveau 0 FAIRE

iter $\longleftarrow 0$

FAIRE

$\widehat{\Delta\Theta^l} \leftarrow$ Moindres-Carrés (résiduel(3.17))

$\widehat{\Theta^l} \longleftarrow \widehat{\Theta^l} + \widehat{\Delta\Theta^l}$ et iter $\longleftarrow$ iter $+1$

TANT QUE (iter $< \lambda$ et $\|\widehat{\Delta\Theta^l}\|_F > \frac{d}{2^l}$)

Si $l \neq 0$: $\widehat{A^{l-1}} \longleftarrow Pr(\widehat{A^l})$, $\widehat{\xi^{l-1}} \longleftarrow \widehat{\xi^l}$

FIN POUR

$\widehat{\Theta_{est}} \longleftarrow \widehat{\Theta^0}$

FIG. 3.1 - *Algorithme moindres-carrés multirésolution (MCMR)*

où (x_F, y_F) est le centre de gravité de la région F. En choisissant des régions de forme simple (le plus petit rectangle "englobant" le support par exemple), l'expression du coefficient s_j est très simple. Bien que κ ne puisse rendre compte de la corrélation entre les différents paramètres dans la modélisation du champ, il est suffisant pour prendre en compte le fait que les coefficients linéaires par exemple ont un impact sur le champ estimé qui dépend de la taille de la région, et donc du niveau de résolution. C'est pourquoi nous avons adjoint un indice F à cette norme. De ce fait, le critère d'arrêt utilisé dans l'algorithme précédent ($\|\widehat{\Delta\Theta^l}\|_F > \frac{d}{2^l}$) teste si les valeurs estimées des incréments apportent une modification significative au champ des déplacements estimé. La quantité d s'assimile donc à un déplacement (on choisit en pratique $d = 0{,}1$), et le dénominateur 2^l rend le test homogène aux différentes résolutions.

Par ailleurs, les différents termes de (3.19) utilisés pour calculer un incrément sont considérés au niveau auquel on se trouve : I devient I^l, X_i devient X_i^l, la zone F devient F^l, etc. . Quant à l'opérateur de projection Pr, il permet de transformer les paramètres de mouvement d'un niveau donné au niveau directement inférieur. Il s'agit d'un changement de résolution. Reconsidérons le modèle des déplacements et les unités des termes a_j: les termes constants sont homogènes à des déplacements, les termes linéaires sont sans unité, (les termes quadratiques, si on les avait introduits, seraient homogènes à l'inverse d'une distance, ...). Lorsque l'on descend d'un niveau, l'unité de distance est multipliée par deux.

La transformation P se résume donc en :

$$Pr \; : \begin{cases} a^{l-1}_{const} \longleftarrow 2a^{l}_{const} \\ a^{l-1}_{lin} \longleftarrow a^{l}_{lin} \\ a^{l-1}_{quad} \longleftarrow \frac{1}{2}a^{l}_{quad} \end{cases} \tag{3.23}$$

3.3 Estimation robuste multirésolution

Dans cette section, nous commencerons par faire quelques rappels sur les méthodes d'estimation robuste. Nous présenterons ensuite les deux algorithmes que nous avons mis en œuvre pour améliorer la robustesse de l'estimateur aux moindres-carrés. Ensuite, nous apporterons quelques modifications à ces algorithmes pour prendre en compte quelques particularités du problème de l'estimation de mouvement.

3.3.1 Estimation robuste

En analyse statistique, l'estimation robuste a pour but de trouver le vecteur Θ de p paramètres qui ajuste au mieux un modèle $(M(X_i, \Theta))_{i\in 1,\ldots,n}$ aux observations $(y_i)_{i\in 1,\ldots,n}$, dans le cas où des données ne correspondent pas à la statistique du modèle d'erreur choisi, c'est-à-dire dans le cas où un certain nombre de données se comportent comme des mesures aberrantes ou *"outliers"*. Dans le cas de l'estimation de mouvement, il s'agira donc de trouver le modèle de mouvement qui satisfait la majorité des données, tout en identifiant et en éliminant les observations qui correspondent soit à d'autres mouvements soit à la mise en défaut de l'hypothèse de conservation de l'intensité.

Les estimateurs robustes sont souvent caractérisés par trois indices [MMR91]:

1. *L'efficacité relative*: elle compare la variance sur les paramètres estimés obtenue par la méthode proposée avec la variance minimale qu'il est possible d'atteindre (la borne de Cramer-Rao), soit:

$$E_{\text{ff}} = \frac{\text{VAR}_{\text{Cramer-Rao}}(\widehat{\Theta})}{\text{VAR}_{\text{Estimateur}}(\widehat{\Theta})} \tag{3.24}$$

   L'efficacité dépend bien sûr des hypothèses sur le modèle de bruit, et du nombre d'échantillons. On considère souvent par exemple l'efficacité asymptotique (le nombre d'échantillons tend vers l'infini) avec des erreurs distribuées suivant une loi normale.

2. *Le point de rupture*: C'est le plus faible pourcentage de contamination des données qui peut faire prendre au vecteur estimé une valeur arbitrairement élevée. Par contamination, on entend le remplacement de données "saines", c'est à dire qui suivent le modèle d'erreur, par des valeurs arbitraires. Par exemple, le point de rupture de l'estimateur des moindres-carrés est $\frac{1}{n}$ puisque un seul *"outlier"* important peut complètement perturber le résultat, et son point de rupture asymptotique est

donc 0. Généralement, un gain en robustesse s'accompagne d'une perte en efficacité et réciproquement [HRRS86]. Il est donc nécessaire de faire un compromis entre ces deux notions.

3. *La complexité algorithmique*: en traitement d'images où le nombre de données à manipuler est très élevé, et particulièrement en analyse du mouvement, où le but est d'atteindre un traitement en temps réel, la complexité algorithmique doit rester aussi faible que possible.

On trouve dans la littérature statistique un grand nombre de classes d'estimateurs robustes. Dans ce mémoire, nous n'en présenterons succinctement que deux: les moindres-carrés médian, et les M-estimateurs. Plus de détails mathématiques peuvent être obtenus dans [Hub81, HRRS86, RL87].

Estimateur des moindres carrés médian

La solution optimale pour cet estimateur satisfait:

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \ \operatorname{Med}_{i\in 1,\ldots,n} \left(y_i - M(\Theta, X_i) \right)^2 \tag{3.25}$$

c'est-à-dire la solution est celle qui minimise la valeur médiane de l'ensemble des carrés des résiduels pour la valeur du vecteur de paramètres considérée (estimée généralement avec un nombre de mesures égal au nombre de paramètres). Son avantage principal est bien sûr sa très grande robustesse théorique, l'estimateur restant fiable jusqu'à un taux de 50% d'*"outliers"*. Il présente cependant plusieurs défauts, notamment:

- si le modèle est non linéaire par rapport aux paramètres, il devient difficile de calculer la solution.
- si le modèle est linéaire, l'estimation des paramètres requiert un temps de calcul très important, grandissant très vite avec le nombre de données considéré (en $O(n^{p+1}\log n)$ où p est le nombre de paramètres à estimer. Une technique de Monte Carlo [MMR91], qui tolère une certaine erreur dans la recherche de l'estimée[5] , permet de réduire la complexité algorithmique en $O(mn\log n)$, $p \ll m \ll n$, où m dépend de l'erreur maximale que l'on s'est fixée.
- son efficacité en cas de bruit gaussien sur les données est faible puisque à chaque itération, un nombre de données égal au nombre de paramètres est tiré pour fournir une estimée des paramètres. Cependant, elle peut être améliorée en effectuant une estimation suivant les moindres-carrés après avoir éliminé les *"outliers"*.

5. Plus précisément, seul un échantillon de l'ensemble des p-uplets de données utilisés pour calculer la solution du problème, est considéré. Plus la taille de cet échantillon est réduite, plus on a de "chance" que tous les p-uplets de cet échantillon contiennent des *"ouliers"*, et donc que la solution optimale ne soit pas déterminée.

Dans notre cas, les données (gradients spatio-temporels de l'intensité) peuvent être relativement bruitées, le modèle affine introduit ne constitue qu'une approximation du mouvement réel et l'hypothèse de conservation de l'intensité permettant de dériver l'équation de mesure est loin d'être toujours vérifiée en pratique. Ce sont des points auxquels l'estimateur des moindres carrés médian est sensible [MMR91]. De plus, comme le nombre de données dans le support de calcul F est en général très grand, l'estimateur des moindres carrés médian sera très coûteux. Nous avons donc choisi de privilégier la méthode des M-estimateurs que nous présentons maintenant.

Les M-Estimateurs

Le principe de cet estimateur consiste à minimiser une somme de résiduels:

$$\widehat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(y_i - M(\Theta, X_i), \sigma) \tag{3.26}$$

où σ correspond à un facteur d'échelle des résiduels, et ρ est l'estimateur. Par exemple, si les résiduels sont distribués suivant une loi gaussienne, l'estimateur optimal est:

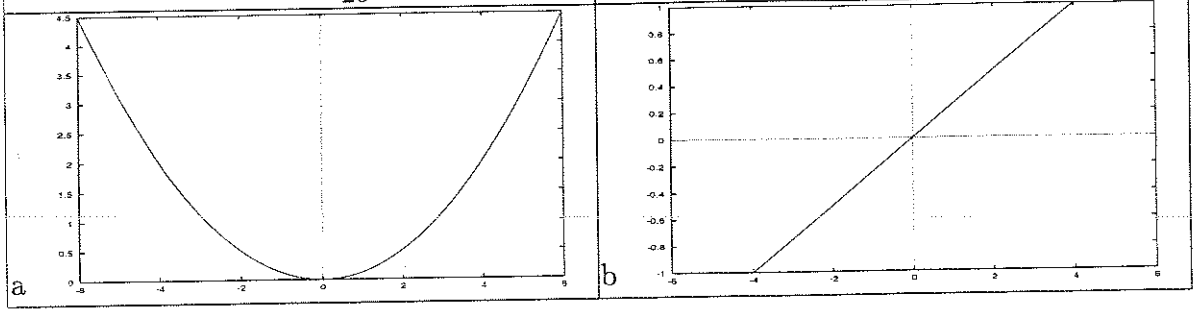
$$\rho(y_i - M(\Theta, X_i), \sigma) = \frac{(y_i - M(\Theta, X_i))^2}{2\sigma^2} \tag{3.27}$$

qui se réduit à l'estimation classique suivant les moindres-carrés. La fonction ρ est appelée un M-estimateur puisqu'il correspond à une estimation suivant le maximum de vraisemblance, si l'on interprète ρ comme étant l'opposé de la log-vraisemblance des observations conditionnellement au modèle.

Un estimateur sera alors robuste si la solution de (3.26) n'est pas trop modifiée lorsqu'une partie des données s'écartent des hypothèses. Pour analyser la robustesse des M-estimateurs, on peut se baser sur les courbes d'influence introduites dans ([HRRS86]). Ces courbes caractérisent le biais ou l'influence que peut introduire une erreur ponctuelle sur l'estimation, et est proportionnelle, dans le cas d'un estimateur continu, à la dérivée ψ de cet estimateur [HRRS86]. Si l'on considère par exemple l'estimateur des moindres-carrés, l'influence des erreurs augmente linéairement sans limite (figure 3.2-1b), ce qui explique la faible robustesse de cet estimateur. Si l'on remplace la norme quadratique L_2 par la norme L_1: $\rho(x) = |x|$, $\psi(x) = sign(x)$, l'influence des erreurs est certes plus faible, mais le point de rupture asymptotique reste tout de même 0. Dans notre cas nous souhaitons que la contribution des mesures aberrantes (par rapport au mouvement dominant) tende vers zéro, voire s'annule complètement. L'utilisation d'un estimateur fortement "redescendant" est nécessaire. La fonction quadratique tronquée bien connue correspond à un tel estimateur (figure 3.2-2). Cependant, pour éviter les discontinuités de celle-ci, nous avons préféré choisir l'estimateur polynomial *"biweight"* de Tukey (figure 3.2-3), dont le comportement est également quadratique pour les faibles erreurs. L'utilisation d'estimateurs fortement redescendants permet d'obtenir des points de rupture strictement supérieurs à 0, mais qui atteignent au maximum $\frac{1}{p+1}$, p étant le nombre de paramètres du modèle à estimer.

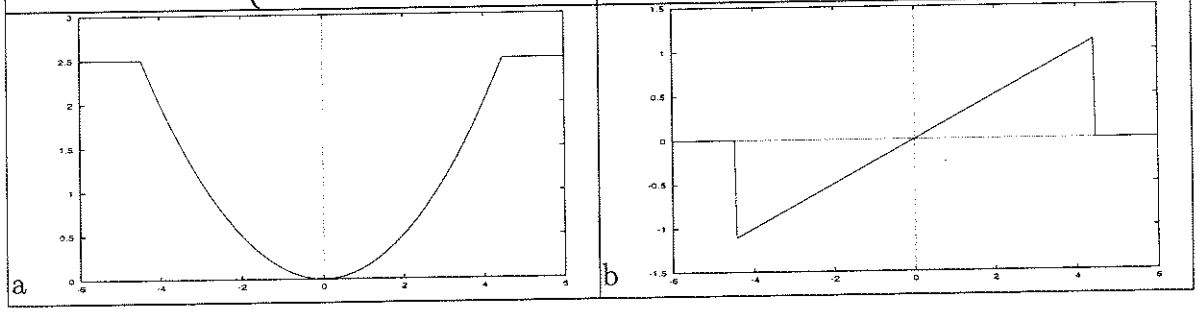
1) Estimateur quadratique:

$$\rho(r,\sigma) = \frac{r^2}{2\sigma^2} \qquad \psi(r,\sigma) = \frac{r}{\sigma^2}$$



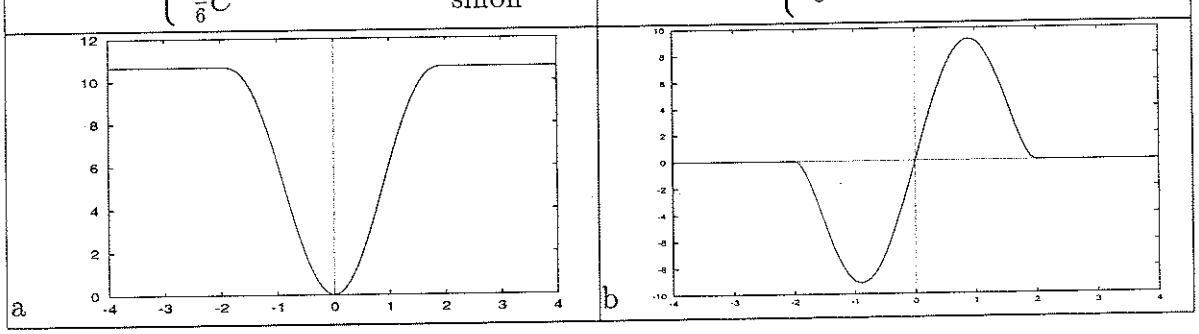
2) Estimateur quadratique tronqué:

$$\rho(r,\alpha,\lambda) = \begin{cases} \lambda r^2 & \text{si } |r| < \frac{\sqrt{\alpha}}{\sqrt{\lambda}} \\ \alpha & \text{sinon} \end{cases} \qquad \psi(r,\alpha,\lambda) = \begin{cases} 2\lambda r & \text{si } |r| < \frac{\sqrt{\alpha}}{\sqrt{\lambda}} \\ 0 & \text{sinon} \end{cases}$$



3) Estimateur *"biweight"* de Tukey:

$$\rho(r,C) = \begin{cases} \frac{r^6}{6} - \frac{C^2 r^4}{2} + \frac{C^4 r^2}{2} & \text{si } |r| < C \\ \frac{1}{6}C^6 & \text{sinon} \end{cases} \qquad \psi(r,C) = \begin{cases} r(r^2 - C^2)^2 & \text{si } |r| < C \\ 0 & \text{sinon} \end{cases}$$



4) Estimateur de Geman et McLure:

$$\rho(r,\sigma) = \frac{r^2}{\sigma + r^2} \qquad \psi(r,\sigma) = \frac{2r\sigma}{(\sigma + r^2)^2}$$

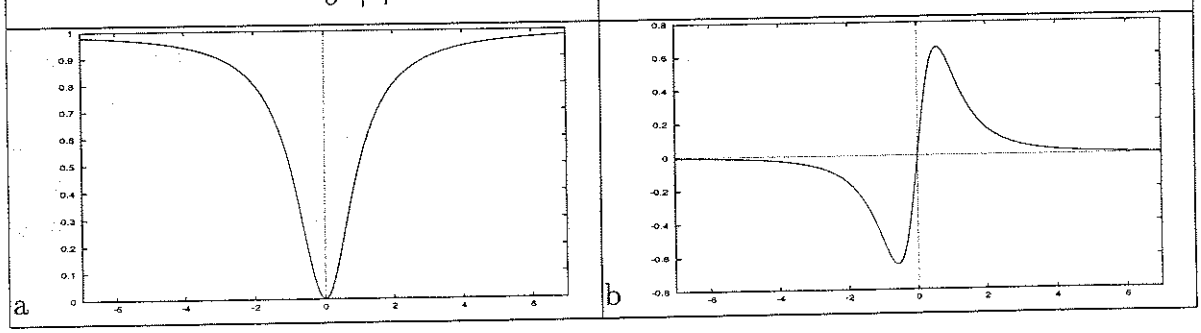


FIG. 3.2 - *Différents estimateurs. Sur chaque rangée: a)l'estimateur ρ; b) la fonction d'influence ψ correspondante.*

Une méthode de minimisation des M-estimateurs: les MCPI

La technique des moindres-carrés pondérés et itérés (MCPI) est une méthode classique pour minimiser la fonctionnelle (3.26) [HW77], particulièrement dans le cas où le modèle M est linéaire par rapport aux paramètres. L'idée de cette méthode est d'attribuer un poids w_i à chaque résiduel, ces poids contrôlant l'estimation des paramètres. Des poids élevés sont attribués aux bonnes données, et des poids faibles aux *"outliers"*. Cette méthode consiste donc à reformuler le problème de l'estimation robuste comme une estimation suivant les moindres carrés pondérés, en posant:

$$\sum_i \rho(\, r_i\,) = \sum_i \frac{1}{2}\, w_i\, r_i^2 \quad \text{avec} \quad r_i = y_i - M(\Theta, X_i) \tag{3.28}$$

Une condition nécessaire pour atteindre un minimum est que les dérivées partielles par rapport à chaque paramètre Θ_j soient nulles, ce qui donne:

$$\sum_i \psi(r_i)\, \frac{\partial r_i}{\partial \Theta_j} = \sum_i w_i\, r_i\, \frac{\partial r_i}{\partial \Theta_j} = 0 \tag{3.29}$$

En un minimum de la fonctionnelle, les coefficients w_i de pondération des moindre-carrés s'expriment comme suit:

$$w_i = \frac{\psi(r_i)}{r_i} \tag{3.30}$$

Le principe est alors le suivant. Si l'on dispose d'une estimation initiale des paramètres, les poids w_i sont tout d'abord évalués pour caractériser l'influence de chaque donnée; sinon, ils sont placés à 1. Ensuite, une estimation du vecteur de paramètres Θ est obtenue en résolvant le problème des moindres-carrés pondérés pour les valeurs courantes des poids w_i. Dans le cas où le modèle est linéaire par rapport aux paramètres, la résolution est immédiate. Les deux phases –calcul des poids et estimation– sont alors itérées jusqu'à atteindre la convergence. En ce qui nous concerne, nous allons utiliser ce schéma avec l'approche multirésolution présentée dans la section 3.2.

3.3.2 Méthodes d'estimation robuste multirésolution proposées

Dans la section 3.2 nous avons utilisé la formulation standard en termes de moindres-carrés pour estimer le modèle de mouvement. Or les hypothèses sur lesquelles repose cette estimation, à savoir le support contient *un seul* modèle de mouvement qui *décrit correctement* l'ensemble des déplacements apparents caractérisés par des vecteurs qui *conservent l'intensité* d'une image à l'autre, ne sont jamais toutes vérifiées en pratique et rendent les moindres-carrés inadaptés. Pour accroître la robustesse de l'estimation, sans remettre en cause les hypothèses, c'est-à-dire sans chercher directement à remédier

aux cas particuliers où elles ne sont pas vérifiées, nous reposons le problème à l'aide de l'estimateur robuste *"biweight"* de Tukey que nous avons décrit précédemment, soit:

$$E_r(\Theta) = \sum_{\mathrm{p}_i \in F} \rho\left(r(\mathrm{p}_i), C\right) \tag{3.31}$$

où $r(\mathrm{p}_i)$ est défini par la formule (3.12). Lorsque l'estimateur est dérivable deux fois, la minimisation de cette fonction peut se faire à l'aide d'algorithmes de descente de gradient, comme la méthode de descente suivant la ligne de plus grande pente (*"steepest descent"*) [BK94], ou une méthode de sur-relaxation [BA93b].

Algorithme RMR

Pour notre part, nous avons conservé le schéma d'estimation multirésolution et incrémentale de la section 3.2. Celui-ci consiste donc à remplacer l'expression de E_r pour le calcul d'un incrément par une expression approchée E'_r:

$$E_r(\Theta) = E_r(\Delta\Theta_k) \simeq E'_r(\Delta\Theta_k) = \sum_{\mathrm{p}_i \in F} \rho(\, r'(\mathrm{p}_i), C\,) \tag{3.32}$$

où $r'(\mathrm{p}_i)$ est donné par la formule (3.17). Ce résiduel étant linéaire par rapport à l'incrément, la méthode des MCPI s'applique facilement. Comme à l'itération k, $\widehat{\Theta}_k$ est supposé être proche de la solution optimale, 0 constitue une estimée initiale de $\Delta\Theta_k$ et peut être utilisée pour calculer les premiers poids dans les MCPI.

Cependant, du fait de l'introduction de l'estimateur robuste, les minima locaux de la fonction E_r sont plus nombreux. Notamment, la fonction E'_r à minimiser à chaque itération n'est plus convexe. Si à une itération donnée, la valeur $\widehat{\Theta}_k$ est trop éloignée de la valeur optimale, le calcul de l'incrément par l'intermédiaire de E'_r restera bloqué dans un minimum local de E'_r. Par la suite, les raffinements successifs feront converger l'estimation vers un minimum local de E_r. L'initialisation du processus d'estimation, notamment à la première itération (au niveau le plus grossier de la pyramide) est donc capitale.

Pour éviter en partie les minima locaux, nous avons adopté une méthode de type GNC (*"Graduated Non Convexity"*) [Bla89]. L'idée générale consiste à former une approximation convexe de la fonction à minimiser, dont on obtient alors facilement un minimum global. Ensuite, des approximations successives de la fonction convergeant vers cette dernière sont minimisées tour à tour en partant de la solution calculée avec l'approximation précédente.

Dans notre cas, nous avons procédé de la façon suivante. Lors du calcul de la toute première estimation de Θ (au niveau de résolution le plus grossier donc), les premiers poids w_i ne sont pas calculés avec la solution initale 0 (pour Θ), mais sont fixés à 1. La première itération des MCPI se fait donc en utilisant les moindres-carrés. Par ailleurs l'élimination graduelle des *"outliers"* se fera à l'aide de la constante C, caractérisant "l'échelle" des résiduels, qui sera utilisée pour contrôler l'allure de l'estimateur et l'introduction de la

non-convexité. Cette valeur sera choisie très élevée au départ, puis diminuée à chaque calcul d'un nouvel incrément, suivant la formule: $C_{t+1} = 0{,}9 \times C_t$, jusqu'à atteindre une valeur finale soit prédéfinie, soit calculée de manière robuste.

Plus précisément, puisque le résiduel correspond à une différence d'images déplacée, C est donc directement identifiable à une variation d'intensité entre deux images. De manière générale, un écart de 2 à 5 niveaux de gris est considéré comme correct dans un recalage global (avec un seul modèle de mouvement). Comme C représente la valeur limite du résiduel, au delà de laquelle la contribution d'un point devient nulle, on retiendra comme valeurs finales de C des valeurs au-delà desquelles le recalage est considéré comme douteux, c'est-à-dire de l'ordre de 8 à 20. En fait, la valeur de ce paramètre dépend du bruit d'acquisition, et surtout de l'adéquation du modèle de vitesse au mouvement réel dans l'image. Si le modèle utilisé est bien approprié, C peut être assez faible, sinon il est préférable de le prendre plus élevé. Pour rendre l'algorithme adaptatif, une possibilité consiste à estimer, parallèlement à la minimisation, la variance du bruit (pour les données conformes au modèle). Dans la mesure où nos données contiennent des *"outliers"*, la valeur médiane de la déviation absolue [MMR91] représente un estimateur robuste de l'écart type de ce bruit:

$$\hat{\sigma} = 1{,}48 \times \mathrm{Med}_i(|r_i - \mathrm{Med}_j(r_j)|) \tag{3.33}$$

où le coefficient 1,48 permet d'avoir un biais nul (sur l'estimation de σ lorsque la distribution des résiduels est gaussienne). Cet écart type "robuste" peut être relié empiriquement au terme C. Par exemple, dans [HW77], des expériences de type Monte Carlo montrent qu'il est préférable de prendre un facteur de proportionnalité proche de 4,7 entre C et $\hat{\sigma}$ pour assurer une meilleure efficacité en cas de bruit gaussien. Cependant, l'estimation de $\hat{\sigma}$ à chaque calcul d'incrément serait très coûteux. Pour fixer une valeur finale de C de manière adaptative, nous avons utilisé la méthode suivante. À l'issue de l'estimation incrémentale au niveau le plus grossier de la pyramide, la formule (3.33) est utilisée pour calculer $\hat{\sigma}$. On prend alors comme valeur finale à atteindre pour C, $4{,}7\,\hat{\sigma}$, en conformité avec ce que nous avons noté au début de ce paragraphe. Des essais menés avec les deux options de valeur finale pour le paramètre C (valeur prédéfinie, typiquement 8, ou valeur robuste), sur des exemples synthétiques et réels, ont donné des résultats très similaires, mais avec un coût calcul évidemment plus élevé dans le second cas. Ces essais ont également montré qu'une certaine latitude existe pour fixer la valeur finale de C.

Dans la plupart des cas, la valeur du nombre maximal λ d'itérations effectuées à un même niveau de résolution sera faible (égale à 5 ou 6). De la même façon, les MCPI se réduiront à quelques itérations (5 ou 6). L'algorithme ainsi défini sera dénommé RMR par la suite (Robuste Multi-Résolution) et est synthétisé sur la figure 3.3.

$C \longleftarrow Max(|I_t^{L-1}|)$

$\widehat{\Theta^{L-1}} \longleftarrow$ MCPI(résiduel (3.17),C,X)[a]

POUR l = niveau L-1 au niveau 0 FAIRE

iter $\longleftarrow$ 0

FAIRE

$C \longleftarrow f(C)$

$\widehat{\Delta\Theta^l} \longleftarrow$ MCPI(résiduel(3.17),C,0)

$\widehat{\Theta^l} \longleftarrow \widehat{\Theta^l} + \widehat{\Delta\Theta^l}$ et iter $\longleftarrow$ iter +1

TANT QUE (iter $< \lambda$ et $\|\widehat{\Delta\Theta^l}\|_{F^l} > \frac{d}{2^l}$)

Si $l \neq 0$: $\widehat{A^{l-1}} \longleftarrow P(\widehat{A^l})$, $\widehat{\xi^{l-1}} \longleftarrow \widehat{\xi^l}$

FIN POUR

$\widehat{\Theta_{est}} \longleftarrow \widehat{\Theta^0}$

[a] MCPI(r,C,vi) indique qu'il faut considérer le résiduel r dans la procédure des MCPI, avec la valeur C qui intervient dans la définition de la fonction ψ, et en utilisant vi comme valeur initiale. Si vi=X, il n'y a pas de valeur initiale et tous les poids initiaux sont pris égaux à 1.

FIG. 3.3 - *Algorithme Robuste Multirésolution (RMR).*

$w_i \longleftarrow 1.0$, $k \longleftarrow 0$ et $\widehat{\Theta} \longleftarrow 0$
FAIRE
$\widehat{\Theta}_p \longleftarrow \widehat{\Theta}$
$\widehat{\Theta} \longleftarrow \mathrm{MCMR}(w_i)$
$w_i \longleftarrow \frac{\psi(q_{\widehat{\Theta}}(\mathrm{p}_i), C)}{q_{\widehat{\Theta}}(\mathrm{p}_i)}$ et $k \longleftarrow k+1$
TANT QUE $\|\widehat{\Theta} - \widehat{\Theta}_p\|_F > d$ et $k < \lambda_1$
$\widehat{\Theta_{est}} \longleftarrow \widehat{\Theta}$

FIG. 3.4 - *Algorithme Pseudo M-estimateur (PSM).*

Algorithme PSM

Nous avons dérivé un deuxième algorithme robuste, qui reprend globalement les mêmes principes. Il repose sur l'application directe de la méthode des MCPI à la fonction E_r. Plus précisément, nous avons donc:

$$E_r(\Theta) = \sum_{\mathrm{p}_i \in F} \rho\left(r(\mathrm{p}_i)\,,\, C\right) = \sum_{\mathrm{p}_i \in F} w_i\, r^2(\mathrm{p}_i) \tag{3.34}$$

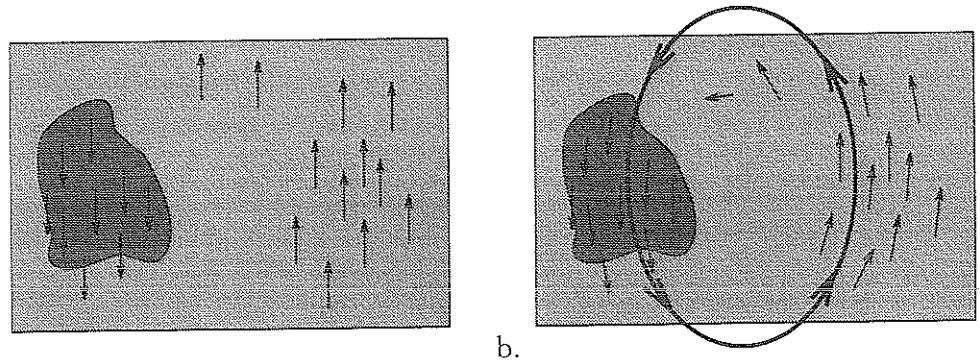
Comme au départ nous ne disposons pas d'estimée, tous les poids sont fixés à 1. La fonction $E_r(\Theta)$ est alors égale à $E(\Theta)$ (relation (3.11)) et peut donc être minimisée avec l'algorithme multirésolution MCMR. On obtient alors l'estimée $\widehat{\Theta}$, à l'aide duquel on calcule les poids $w(\mathrm{p}_i)$ suivant la formule (3.30), le résiduel étant ici la quantité $q_{\widehat{\Theta}}$:

$$q_{\widehat{\Theta}}(\mathrm{p}_i) = I(\mathrm{p}_i + B_i\widehat{A}, t+1) - I(\mathrm{p}_i, t) + \widehat{\xi} \tag{3.35}$$

La pondération est propagée à tous les niveaux de façon appropriée (on utilise les mêmes filtrage gaussien et sous-échantillonnage que ceux utilisés pour construire la pyramide des données), et on relance alors une nouvelle estimation à travers la pyramide. La structure de l'algorithme de minimisation à chaque descente est du type de celle du MCMR, mais en considérant des moindres carrés pondérés. Cette version permet d'éviter le calcul des MCPI pour chaque estimation d'un incrément, mais elle nécessite d'effectuer plusieurs descentes de la pyramide. Nous l'avons dénommée *PSM* (Pseudo M-estimateur). Elle est résumée à la figure 3.4.

3.3.3 Étapes complémentaires

Nous n'avons pas tenu compte dans les paragraphes précédents de certains problèmes de l'estimation de mouvement, notamment ceux relatifs aux données. Ces problèmes spé-



a. b.

FIG. 3.5 - *a) Exemple de problème d'estimation ambiguë: l'objet dominant se déplace vers le haut, tandis que le second objet est animé d'une translation vers le bas (les vecteurs de mouvement ne sont tracés qu'au point de gradient spatial d'intensité supposé significatif). b) le mouvement estimé résultant peut être une rotation.*

cifiques, qui peuvent produire des effets indésirables dans les algorithmes que nous avons définis précédemment, sont de trois ordres:

1. le mouvement réel que l'on souhaite estimer pourrait être décrit par moins de paramètres que le nombre retenu dans notre modèle; de manière équivalente, il peut arriver que le mouvement réel et la distribution spatiale des gradients d'intensité ne contraignent pas suffisamment l'estimation de tous les paramètres du modèle.

2. la minimisation initiale est basée sur les moindres carrés; alors, les mouvements des objets secondaires peuvent être "récupérés" pour contraindre suffisamment les degrés de liberté laissés indéterminés ou mal conditionnés par les observations correspondant au mouvement dominant. Cet effet peut être d'autant plus marqué que les régions supportant le mouvement dominant sont très peu texturées, et qu'à l'opposé, les régions correspondant aux mouvements secondaires le sont fortement.

3. notre estimation est grandement dépendante des paramètres d'estimation initiaux; si l'un des cas précédents se produit, notre algorithme sera susceptible de converger vers un minimum local éloigné de la solution optimale.

La figure 3.5 présente un exemple type. La figure 3.5a nous montre les vecteurs vitesses réels dans l'image, qui ne sont tracés que pour des pixels ayant un gradient spatial d'intensité supposé important, tandis que la figure 3.5b contient le champ estimé que l'on pourrait obtenir dans une telle situation. Comme on peut le constater, les vecteurs de mouvement aux points de mesure possibles sont en fait peu distordus, bien que les deux translations soient "fusionnées" au sein d'un même modèle de rotation qui n'a pas grand chose à voir avec les mouvements réels. Une première possibilité serait de faire confiance à un estimateur plus robuste, du moins pour la première itération, comme les moindres carrés médian. Cette solution souffrira cependant des inconvénients évoqués au paragraphe

Nombre de paramètres $\longleftarrow$ 2
$C \longleftarrow Max(|I_t^{L-1}|)$ et $l \longleftarrow L-1$
$\widehat{\Theta^{L-1}} \longleftarrow$ MCPI(résiduel (3.17),C,X)[a]
TANT QUE $l \geq 0$ FAIRE
  iter $\longleftarrow$ 0
  FAIRE
    $C \longleftarrow f(C)$
    $\widehat{\Delta\Theta^l} \longleftarrow$ MCPI(résiduel(3.17), C , 0)
    $\widehat{\Theta^l} \longleftarrow \widehat{\Theta^l} + \widehat{\Delta\Theta^l}$ et iter $\longleftarrow$ iter +1
  TANT QUE (iter $< \lambda$ et $\|\widehat{\Delta\Theta^l}\|_{F^l} > \frac{d}{2^l}$)
  SI $l = L_a$ et Nombre de paramètres=2
    Nombre de paramètres $\longleftarrow$ p
  SINON
    SI $l \neq 0$: $\widehat{A^{l-1}} \longleftarrow P(\widehat{A^l})$, $\widehat{\xi^{l-1}} \longleftarrow \widehat{\xi^l}$, $l \longleftarrow l-1$
FIN TANT QUE
$\widehat{\Theta_{est}} \longleftarrow \widehat{\Theta^0}$

[a] voir algorithme RMR

FIG. 3.6 - *Algorithme Robuste Multirésolution modifié (RMRmod).*

3.3.1. En fait, une solution plus simple et plus efficace à ce problème et qui fonctionne dans un grand nombre de cas, consiste à commencer le processus d'estimation en considérant un modèle de translation, et à introduire les modèles plus complexes uniquement par la suite, après quelques itérations. La figure 3.6 présente l'algorithme RMR modifié dans ce sens. Nous commençons par estimer un modèle constant du niveau $L-1$ au niveau L_a inclus, puis nous considérons du niveau L_a au niveau 0 le modèle complet retenu, qui est fréquemment le modèle affine.

Dans nos applications, nous choisissons généralement pour L_a une valeur de 2, alors que le nombre de niveaux L est de 3 ou 4. Notons que dans le cas d'une séquence, il est possible à l'aide des estimations effectuées dans les images précédentes de déterminer une valeur opportune du niveau L_a auquel il est souhaitable de commencer l'estimation des paramètres linéaires, voire quadratiques.
L'avantage essentiel de cette méthode repose sur le fait que l'algorithme d'estimation incrémentale est très efficace lorsque un modèle de mouvement constant est utilisé (voir

Nombre de paramètres $\longleftarrow$ 2, $w_i \longleftarrow$ 1.0 et $k \longleftarrow 0$

$\widehat{\Theta} \longleftarrow \mathrm{MCMR}(w_i)$

$w_i \longleftarrow \frac{\psi(q_{\widehat{\Theta}}(\mathrm{p}_i), 2\times C)}{q_{\widehat{\Theta}}(\mathrm{p}_i)}$

Nombre de paramètres $\longleftarrow$ p

FAIRE

$\widehat{\Theta}_{pr} \longleftarrow \widehat{\Theta}$

$\widehat{\Theta} \longleftarrow \mathrm{MCMR}(w_i)$

$w_i \longleftarrow \frac{\psi(q_{\widehat{\Theta}}(\mathrm{p}_i), C)}{q_{\widehat{\Theta}}(\mathrm{p}_i)}$ et $k \longleftarrow k+1$

TANT QUE $\|\widehat{\Theta} - \widehat{\Theta}_{pr}\|_F > d$ et $k < \lambda_1$

$\widehat{\Theta}_{est} \longleftarrow \widehat{\Theta}$

FIG. 3.7 - *Algorithme Pseudo M-estimateur modifié (PSMmod).*

[BHK91]), et facilite la discrimination entre les mouvements de différents objets ou entre le fond et des objets mobiles.

L'algorithme PSM peut également être modifié suivant ce schéma, en estimant un mouvement constant lors de la première estimation multirésolution. Cette nouvelle version PSMmod (voir figure 3.7) devient en fait similaire à la méthode présentée dans [IRP92]. Cependant, dans cet article, après chaque estimation multirésolution, un algorithme de détection explicite (plus complexe qu'une simple procédure de pondération) est utilisé. Il élimine les points pour lesquels le mouvement calculé n'est pas satisfaisant, et attribue un poids de 1 pour tous les autres. Cette décision binaire peut être trop catégorique lorsqu'il n'existe pas de carte de détection explicite ou dans le cas d'entités "floues", comme dans les images météorologiques. De plus, après l'estimation du modèle constant, cette décision peut aussi déboucher sur le choix d'un nombre limité de points formant une région réduite qui sera vraiment susceptible d'avoir un mouvement purement translationnel. Les itérations suivantes avec un modèle affine n'auront d'autre possibilité que de rester bloquées dans une telle configuration. Il ne sera donc pas facile de traiter des situations impliquant des mouvements plus complexes.

Dans notre cas, nous avons choisi de calculer les poids après l'estimation du modèle de mouvement constant effectuée avec une valeur de la constante C dans la fonction Ψ plus importante que lors des itérations suivantes (nous avons choisi une valeur deux fois plus

grande)[6]. Ceci permet de prendre partiellement en compte le fait que le modèle constant est sans doute trop fruste. De cette façon, nous conserverons suffisamment de points pour estimater correctement les paramètres affines dans les phases suivantes.

Enfin, mentionnons ici que dans les deux algorithmes, un pixel dont le poids est nul à une itération donnée n'est pas définitivement écarté, et peut donc de ce fait récupérer un poids non nul si l'affinement de l'estimation joue en "sa faveur". Ceci nous permet d'estimer des mouvements qui ne sont pas nécessairement constants, comme le montreront les résultats présentés dans la section suivante.

3.4 Résultats

Des expériences ont été menées sur des images réelles animées de mouvements synthétiques pour obtenir une évaluation quantitative précise du comportement des algorithmes. Des séquences réelles ont été également traitées et nous présentons ici les résultats pour deux scènes d'extérieur. En pratique, de nombreuses autres expérimentations ont été menées sur des types d'images variés, mais ne peuvent toutes être évoquées ici. La comparaison entre les deux versions d'algorithmes que nous avons introduites (RMR et PSM) sera faite au paragraphe 3.5. Des résultats sur une séquence météorologique seront également présentés en annexe de ce chapitre, dans la sous-section 3.7 traitant du problème particulier de la détection de points singuliers (les vortex notamment) dans une séquence d'images.

3.4.1 Expérimentations de type Monte Carlo sur une image réelle animée de mouvements synthétiques

Pour évaluer quantitativement les performances de nos algorithmes, nous avons effectué une série de N_{exp} expériences sur lesquelles les différents algorithmes ont été testés. Chaque expérience est construite de la façon suivante. Nous avons pris l'image de la figure 3.11a, et nous lui avons appliqué un mouvement synthétique pour construire une seconde image (en utilisant une interpolation bilinéaire pour les points obtenus ayant des coordonnées non entières). Ce mouvement est en fait composé de deux modèles affines différents, l'un, A_1, appliqué sur une fenêtre carrée de l'image (zone 1, Z_1, voir figure 3.8a), l'autre A_2, sur le reste de l'image (zone 2, Z_2). Les expériences diffèrent en fait par les modèles affines qui sont appliqués, ces derniers étant sélectionnés aléatoirement. Plus précisément, les coefficients constants de ces modèles sont tirés au hasard sur l'intervalle $[-3, 3]$ suivant une loi uniforme, et les termes linéaires dans l'intervalle $[-0{,}05\,,\,0{,}05]$ également de manière uniforme. Le centre du repère de ces modèles est le centre du carré formant la zone 1. Le champ des vitesses de la figure 3.8b est l'un des champs synthétiques ainsi générés.

6. De manière générale, nous aurions pu dans l'algorithme PSM, modifié ou non, faire évoluer la constante C en fonction de l'itération, comme pour l'algorithme RMR.

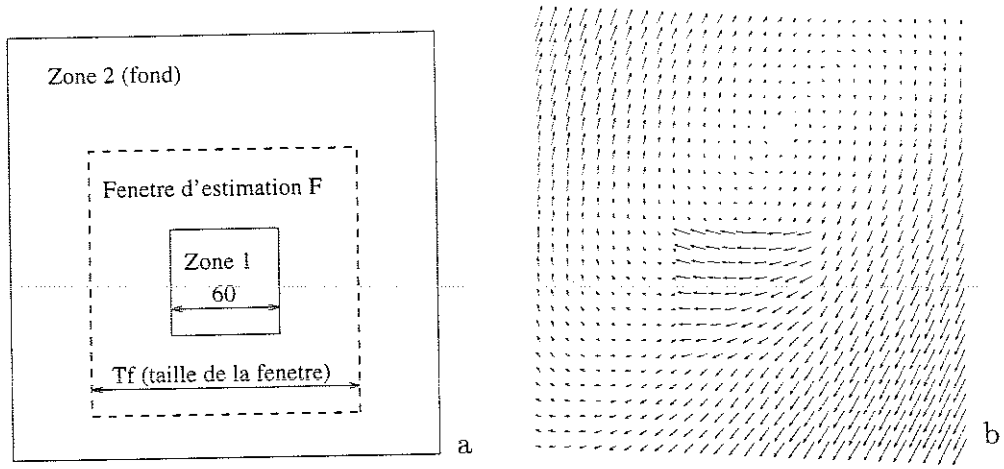


FIG. 3.8 - *a) schéma d'expérimentation b) un exemple de champ synthétique des vitesses appliqué sur l'image;*

Généralement, les modules des déplacements aux points de l'image se situent entre 0 et une quinzaine de pixels.

Chaque expérience consiste simplement à étudier le comportement de l'estimateur lorsque la proportion des deux zones présentes dans la fenêtre ou support d'estimation F varie. Pour cela, les six paramètres du modèle affine ont été estimés sur un support carré F de largeur T_F variable (Fig. 3.8a). Nous avons retenu comme indice d'adéquation des paramètres estimés dans chaque zone, l'erreur moyenne (en norme) sur le champ des vitesses, donnée par la relation suivante ($n = 1, 2$):

$$err_n(t_n) = \frac{\sum_{\mathrm{p}_i \in (F \cap Z_n)} \| \vec{V}_{\hat{A}}(\mathrm{p}_i) - \vec{V}_{A_n}(\mathrm{p}_i) \|}{\sum_{\mathrm{p}_i \in (F \cap Z_n)} \| \vec{V}_{A_2}(\mathrm{p}_i) - \vec{V}_{A_1}(\mathrm{p}_i) \|} \tag{3.36}$$

où t_n est le taux d'occupation de la fenêtre Z_n dans le support d'estimation, c'est à dire:

$$t_n = \frac{\mathrm{Card}(F \cap Z_n)}{\mathrm{Card}(F)} \tag{3.37}$$

Ainsi, lorsque $\hat{A}$ correspond à A_n, l'erreur est proche de zéro. Le terme au dénominateur dans la formule a son intérêt lorsque le modèle estimé correspond à l'autre modèle affine. Dans ce cas, l'erreur moyenne mesurée par (3.36) tend vers 1. Sans ce terme de normalisation, l'erreur dépendrait de la taille de la fenêtre d'estimation, et surtout de la "différence" entre les champs générés par les deux modèles affines, et donc de l'expérience particulière traitée. L'interprétation ensuite des valeurs moyennes des err_n correspondant à l'ensemble des expériences se serait avérée difficile.

L'étude d'autres erreurs (comme l'erreur angulaire) ou même l'étude directe de l'évolution des paramètres donne des résultats similaires et permet d'aboutir aux mêmes conclusions. Les différents paramètres intervenant dans chacun des algorithmes sont donnés dans le tableau 3.1.

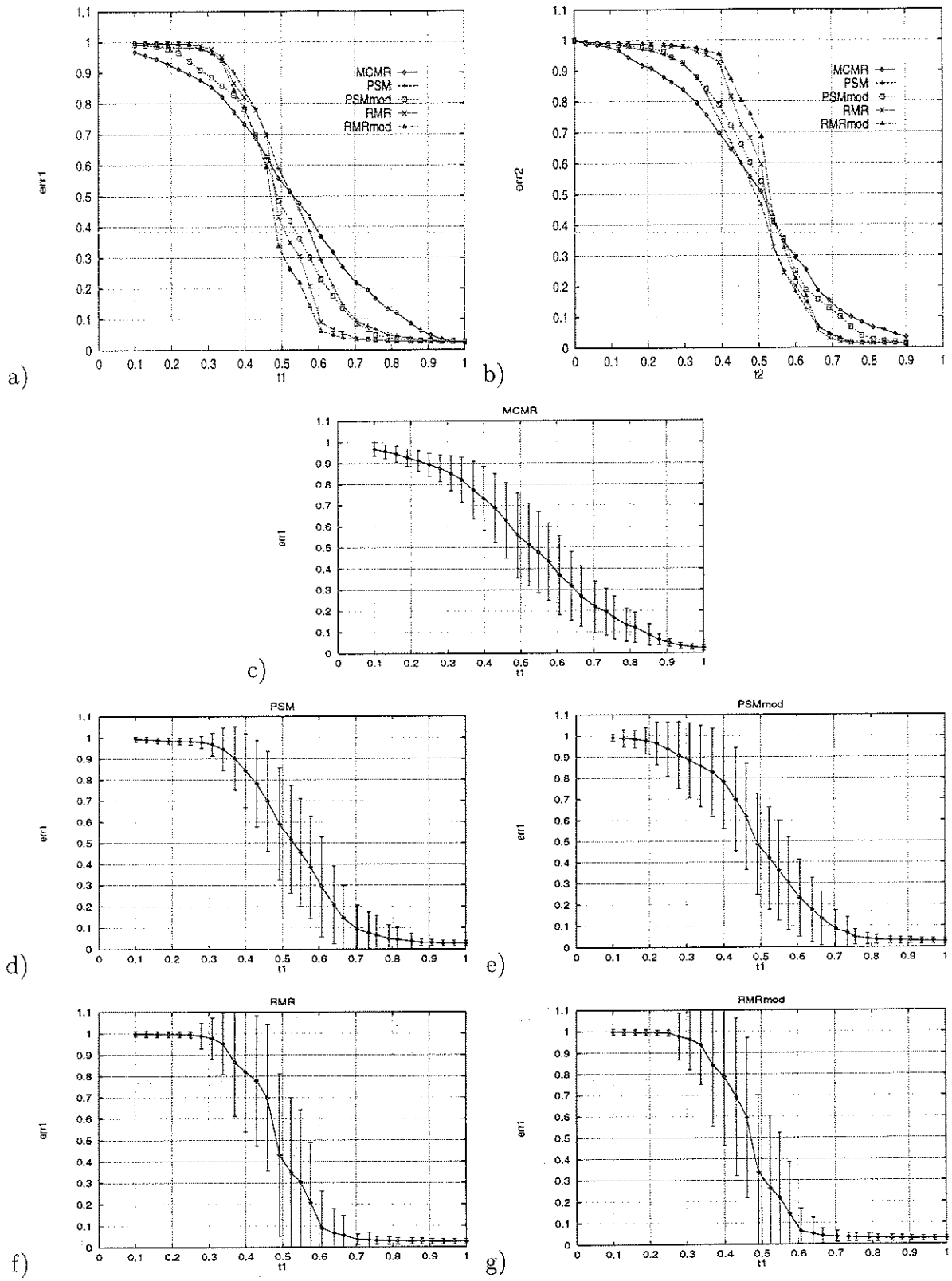


FIG. 3.9 - *a) et b) Erreur moyenne err_i en fonction de la proportion t_i de points de la région i dans le support pour tous les algorithmes: a) $i = 1$ et b) $i = 2$. c) à g) Erreur moyenne err_1 et écart type σ_{err_1} en fonction de t_1 obtenues avec les algorithmes: c) MCMR, d) PSM e) PSMmod f) RMR et g) RMRmod.*

Algorithmes	L	λ	d	C	Nb itération MCPI	L_a	λ_1
MCMR	4	6	0,1	-	-	-	-
RMR et RMRmod	4	6	0,1	8,0[a]	4	2	-
PSM et PSMmod	4	6	0,1	8,0	-	-	4

TAB. 3.1 - *Valeurs des paramètres des différents algorithmes. Le nombre N_{exp} d'expériences différentes est 150.*

[a] Il s'agit de la valeur finale

Les graphiques de la figure 3.9 présentent les résultats obtenus. Remarquons tout d'abord que, compte tenu de l'erreur choisie, les courbes $err_1(t_1)$ (figure 3.9a) et $err_2(t_2)$ (figure 3.9b) sont quasiment symétriques par rapport à la première bissectrice. Par la suite nous ne parlerons que de $err_1(t_1)$. Comme prévu, l'estimateur des moindres carrés multirésolution MCMR effectue un moyennage entre les deux mouvements (figure 3.9c), alors que les estimateurs robustes fournissent beaucoup plus souvent une estimation correcte (figures 3.9d-g). Sur la figure 3.9a nous pouvons constater la supériorité du premier type d'algorithme RMR sur le second PSM, notamment tant que la région 1 reste majoritaire dans le support. Globalement, si nous considérons que le modèle affine estimé ne correspond à aucun des deux modèles A_1 et A_2 des régions 1 et 2 lorsque l'erreur est située entre 0,1 et 0,9, nous avons les résultats suivants: la longueur de la plage de transition séparant l'estimation correcte du modèle 1 de celle du modèle 2, est de 0,6 pour les moindres-carrés, 0,41 pour l'algorithme PSMmod, 0,33 pour le PSM, et d'environ 0,24 pour les deux algorithmes RMR et RMRmod. Si l'on examine maintenant les écarts-types sur l'erreur, on peut remarquer que ceux-ci sont beaucoup plus importants à l'intérieur de la plage de transition pour les algorithmes RMR que pour les algorithmes PSM, dénotant une plus grande variabilité de comportement d'un exemple à l'autre pour les premiers. En effet, à l'intérieur de la zone de transition, pour une même valeur de t_1, les algorithmes RMR estiment souvent correctement l'un ou l'autre des modèles A_1 et A_2 suivant l'expérience particulière traitée, ou parfois moyennent totalement les deux mouvements. A l'opposé, les algorithmes PSM s'écartent peu de l'estimation aux moindres-carrés multirésolution initiale, qui moyenne systématiquement les deux mouvements.

Par ailleurs, dans ces expériences, où la partie linéaire du modèle de mouvement est très marquée, nous pouvons constater que les modifications que nous avons apportées aux algorithmes initiaux, en considérant dans les premières estimations un modèle translationnel, n'apportent ici aucun gain notoire. On peut même remarquer que l'algorithme PSM modifié donne de mauvais résultats lorsque la région 2 est majoritaire. Ceci s'explique aisément en observant le champ des vitesses de la figure 3.8b qui correspond à l'une des expériences. Le mouvement dans la région extérieure n'est pas du tout translationnel. Ainsi, commencer par estimer une translation jusqu'à la résolution la plus fine ne permet

de prendre en compte qu'une partie du champ. Les estimation suivantes du modèle affine se feront sur cette partie, et n'exploiteront donc pas toute l'information de la région 2. On peut remarquer que cet effet n'existe pas sur l'algorithme RMR modifié, qui n'estime le modèle constant qu'aux basses résolutions.

Nous avons également étudié le comportement des algorithmes en fonction du niveau de bruit introduit. Pour cela, l'image artificiellement construite à chaque expérience à partir des deux modèles de mouvement affines est systématiquement bruitée par un bruit blanc gaussien centré d'écart-type σ_G. La figure 3.10a présente les résultats lorsque ce bruit est important ($\sigma_G = 11$). Alors que les algorithmes PSM n'arrivent plus à donner de meilleurs résultats que l'algorithme MCMR (figures 3.10a et 3.10b), les algorithmes RMR restent relativement insensibles au bruit (figures 3.10c et 3.10d). Ceci peut s'expliquer par le fait que dans ces derniers, la sélection de l'un ou l'autre des modèles par l'intermédiaire des coefficients de pondération se fait dès les basses résolutions alors que le bruit est filtré, ce qui n'est pas le cas pour les algorithmes PSM.

Enfin, soulignons qu'il est difficile de dire, pour une taille de fenêtre donnée, quel est, ou quel doit être, le mouvement dominant. En effet, ce n'est pas particulièrement la proportion du nombre de points d'une région contenus dans le support qui permet de déterminer le mouvement dominant, mais plutôt l'information de mouvement que l'on peut extraire de cette région. Or la mesure de mouvement est basée sur l'équation de contrainte du mouvement apparent, dans laquelle le gradient de l'intensité joue un rôle notable. Les zones uniformes, qui occupent une surface non négligeable, ne procurent donc aucune information. La "transition" dans nos expériences synthétiques entre l'estimation du modèle de mouvement de la région Z_1 et celle du mouvement de la région Z_2 devrait plutôt se produire pour une fenêtre F telle que :

$$\sum_{\mathrm{p}\in F\cap Z_1} \|\vec{\nabla} I(\mathrm{p})\| = \sum_{\mathrm{p}\in F\cap Z_2} \|\vec{\nabla} I(\mathrm{p})\| \tag{3.38}$$

Les figures 3.10e et 3.10f présentent les résultats obtenus avec les modèles de mouvement correspondant au champ des vitesses de la figure 3.8b, lorsque la région Z_1 est placée au centre de l'image 3.11a, et lorsqu'elle est placée sur les véhicules à gauche de cette même image. Dans le premier cas, la zone Z_1 est située sur une zone de feuillage où les gradients spatiaux de l'intensité sont relativement uniformes comparativement à ceux de la région Z_2. L'équilibre (3.38) est atteint pour une valeur de t_1 égale à 0,58. Dans le deuxième cas, les gradients spatiaux dans la zone Z_1 sont plus conséquents. Le mouvement 1, toujours d'après (3.38) calculé au niveau de résolution le plus fin, devrait alors rester dominant jusqu'à $t_1 = 0{,}42$. Comme on peut le constater sur les courbes, les transitions pour les algorithmes PSM semblent effectivement se produire autour de ces valeurs, ce qui n'est pas le cas pour les algorithmes RMR (dans la deuxième expérience notamment). Ceci peut s'expliquer par le fait que pour ces derniers, la seule minimisation suivant les moindres-carrés, qui prend en compte tous les gradients, se fait à basse résolution. Notons par ailleurs que pour les deux algorithmes RMR, les transitions sont très bien marquées.

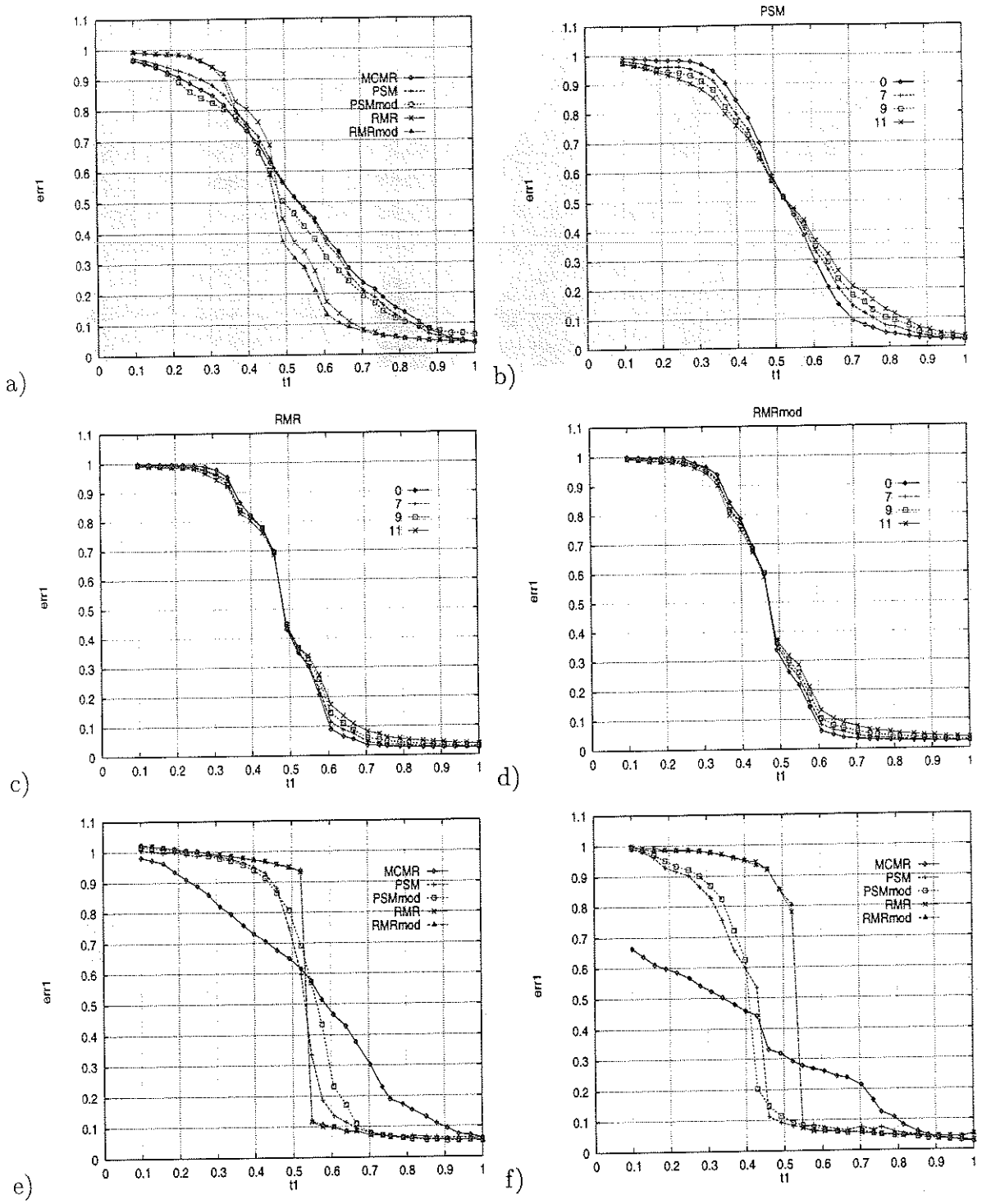


FIG. 3.10 - *a) Erreur moyenne err_1 en fonction de la proportion t_1 de points de la région 1 dans le support, avec un bruit gaussien centré d'écart-type $\sigma_G = 11$ ajouté aux images synthétisées. b) à d) Erreur moyenne err_1 en fonction de t_1 pour différentes valeurs de l'écart-type σ_G du bruit gaussien, pour les algorithmes b) PSM, c) RMR et d) RMRmod. e)f) Courbes d'erreur obtenues pour l'expérience particulière correspondant au champ des vitesses proposé à la figure 3.8b, lorsque la région Z_1 est placée: e) au centre de la première image de la séquence "voiture"; f) sur les véhicules à gauche de cette même image.*

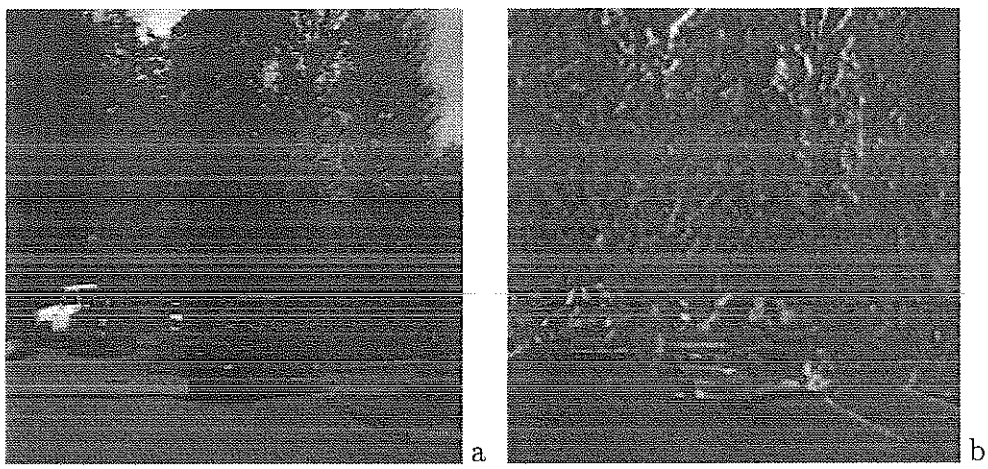


FIG. 3.11 - *Séquence "voitures": a) première image ; b) image des différences temporelles entre les deux images considérées.*

3.4.2 Expérimentations avec des séquences réelles

Séquence "voitures".

Dans de nombreux cas d'analyse dynamique, il est utile et nécessaire de déterminer ou de compenser dans un premier temps le mouvement de la caméra pour ensuite détecter les objets mobiles. La figure 3.11.a montre la première image de la séquence réelle traitée, dont le contenu dynamique est souligné par l'image (Fig.3.11.b) des différences temporelles entre les deux images considérées, auxquelles un "offset" de 128 a été ajouté (Une valeur grise correspond ainsi à une différence nulle, et, plus le point est soit noir, soit blanc, suivant le signe de la différence, plus cette dernière est importante). Ce contenu dynamique se partage en trois composantes: un panoramique de la caméra de la droite vers la gauche induisant une translation apparente, un mouvement un peu erratique du feuillage dû au vent surtout au centre, et enfin les déplacements de deux voitures. Pour illustrer plusieurs cas typiques, l'image a été divisée en quatre blocs. Sur chacun d'eux, on estime le mouvement principal et on calcule la différence DFD_{comp} (donnée par (3.35)) correspondante (plus l'offset de 128). Le nombre de niveaux dans la pyramide est de 4 (nous utilisons un critère simple pour fixer le nombre de niveaux en fonction de la taille du support d'estimation F).

La figure 3.12 présente les résultats obtenus avec les algorithmes MCMR et PSM modifié. Dans les deux blocs du haut, les mouvements du feuillage sont globalement incohérents et faibles. Le mouvement dominant est donc clairement le panoramique que l'algorithme MCMR et l'algorithme PSM modifié sont en mesure d'estimer comme le montrent les champs $\vec{V}_{\hat{A}}$ de la figure 3.13. Les vecteurs de vitesse sont tracés aux points où le mouvement dominant est considéré comme adéquat selon la valeur de w_i fonction de $q_{\widehat{\Theta}}(\mathrm{p}_i)$, l'erreur de recalage (3.35). En revanche, dans le bas, les mouvements rigides des deux voitures constituent des perturbations cohérentes importantes, et l'image d'erreur 3.12.a

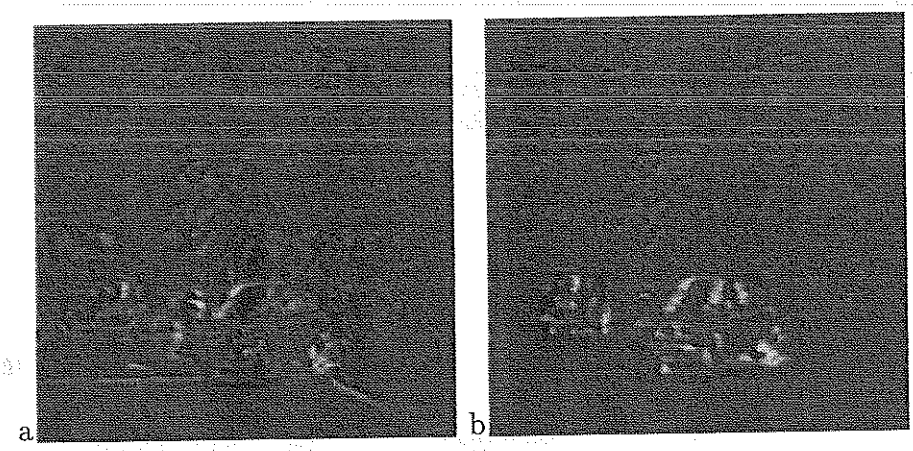


FIG. 3.12 - *Image des différences compensées DFD_{comp} : a) MCMR ($\lambda = 5, d = 0.1$); b) PSM modifié ($C = 9$, $\lambda = 5, d = 0{,}1, \lambda_1 = 12$);*

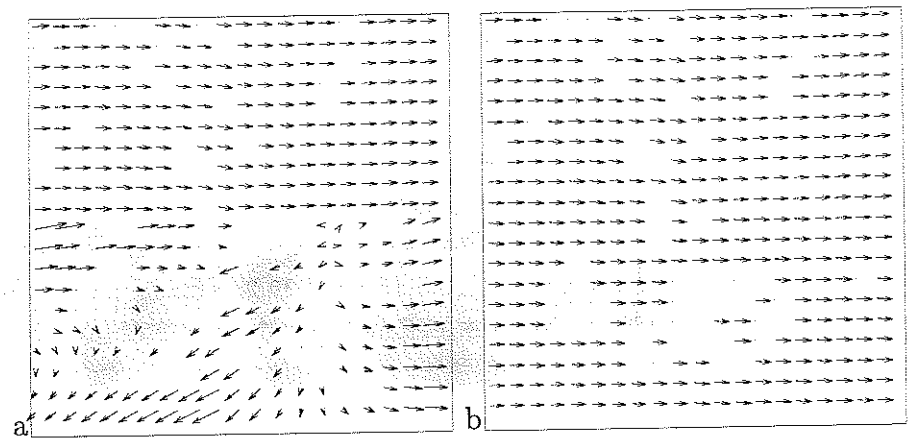


FIG. 3.13 - *Champ de vitesses associé au modèle de mouvement estimé : a) MCMR; b) PSM modifié ($C = 9$);*

ainsi que le mouvement estimé résultant 3.13.a mettent clairement en évidence l'effet de moyennage de l'algorithme MCMR. La DFD_{comp} n'est correcte ni sur la partie fixe de la scène, ni sur les voitures en mouvement et les mouvements estimés sont très erronés. Par contre, les figures 3.12.b et 3.13.b montrent bien que le mouvement panoramique a été parfaitement déterminé par l'algorithme PSM modifié alors que les voitures couvrent une surface non négligeable dans chaque bloc. Non seulement le mouvement de panoramique de la caméra est très bien compensé, aucun élément de la partie fixe de la scène n'apparaissant dans l'image d'erreur de la figure 3.12.b, mais simultanément les zones correspondant à des objets mobiles dans la scène sont beaucoup mieux mises en évidence que dans le cas de l'algorithme MCMR. On peut bien sûr envisagé une seconde étape, où pour les zones ainsi détectées, on relance l'algorithme PSM modifié pour déterminer les mouvements secondaires dans l'image, et ainsi de suite.

Séquence ROND-POINT

Les figures 3.14a, 3.14c et 3.14e représentent trois images de la séquence ROND-POINT[7]. Ici, les déplacements dominants dans l'image sont dus au mouvement de la caméra, qui est montée sur le côté gauche d'une voiture approchant un rond-point, et qui pointe perpendiculairement à l'axe longitudinal de ce véhicule. Cependant, comme les différences entre les profondeurs des objets statiques dans la scène sont importantes, le modèle du mouvement dominant correspond uniquement au mouvement du fond (c.à.d., les maisons principalement). Nous reviendrons sur ce point dans le chapitre suivant consacré à la détection du mouvement.

Dans cette expérience, nous avons estimé un modèle affine entre chaque couple d'images consécutives avec les algorithmes MCMR, RMR et PSM modifiés. Nous avons utilisé quatre niveaux dans la pyramide ($L = 4$) et retenu une valeur de 2 pour L_a. Nous avons choisi 8 pour valeur finale de C. Les modèles calculés sont alors utilisés pour générer les séquences compensées 3.14b-d-f, 3.15a-c-e et 3.15b-d-f. Si le déplacement d'une région est bien compensée par le mouvement estimé, celle-ci doit alors restée fixe dans les images successives compensées.

Une fois encore, les images 3.14b-d-f indiquent clairement l'effet de moyennage de l'algorithme MCMR, puisque toute l'image se déforme au cours du temps. A l'opposé, les figures 3.15a-c-e montrent que l'algorithme RMR modifié estime de manière cohérente le mouvement apparent du fond sur toute la durée de la séquence: les maisons restent fixes dans la séquence compensée. Enfin, l'algorithme PSM modifié estime au départ correctement le mouvement du fond (de t_{62} à t_{63}), passe par une phase où il effectue un moyennage à l'image 64 (voir figures 3.15b-d-f: entre 3.15b et 3.15d, les maisons ne sont pas fixes; le panneau se déplace à droite), et ensuite se cale sur le mouvement du panneau

7. Pour présenter les résultats de l'estimation, nous ne considérons ici que 10 images de la séquence ROND-POINT, entre les instants t_{62} et t_{72}. Une durée plus importante sera retenue dans les chapitres suivants.

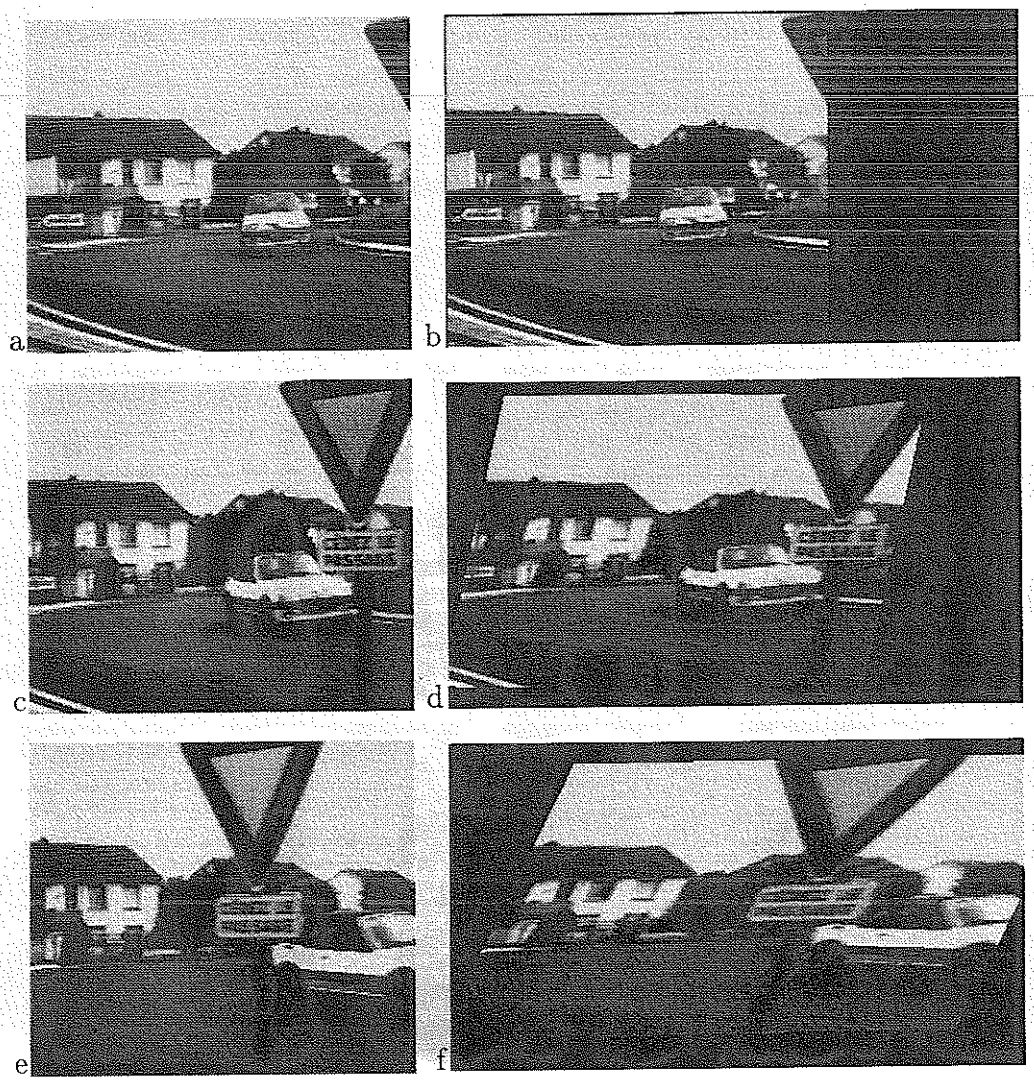


FIG. 3.14 - *a) c) e) Trois images de la séquence ROND-POINT aux instants a) t_{62}, c) t_{67} et e) t_{72}.*
b) d) f) Images aux instants t_{62}, t_{67} et t_{72} compensées avec les modèles de mouvement affines calculés par l'algorithme MCMR.

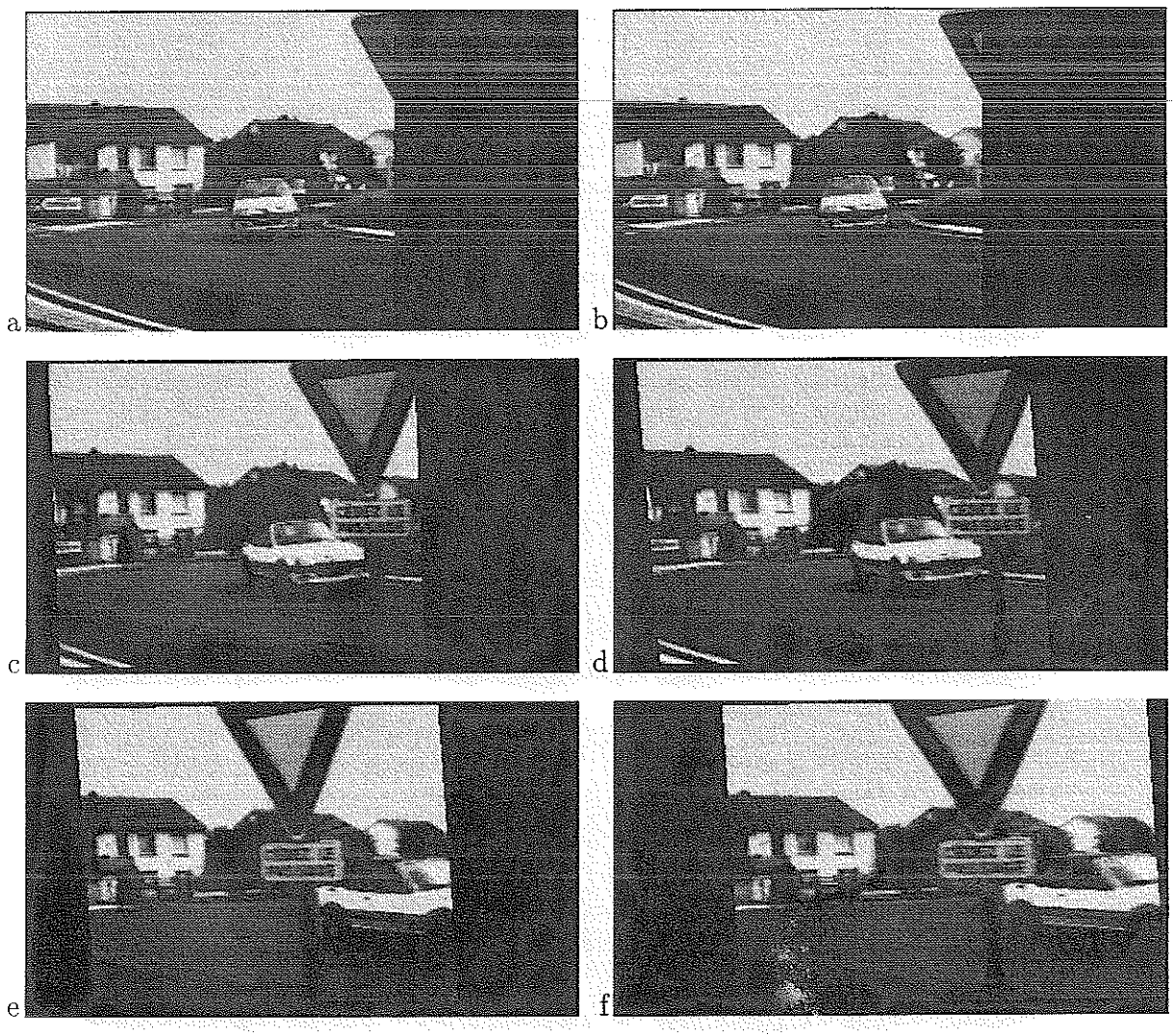


FIG. 3.15 - *a) c) e) Images aux instants t_{62}, t_{67} et t_{72} compensées avec les modèles de mouvement calculés par l'algorithme RMR modifié.*
b) d) f) Images aux instants t_{62}, t_{67} et t_{72} compensées avec les modèles de mouvement calculés par l'algorithme PSM modifié.

de t_{65} à t_{72}. La différence entre les performances des deux algorithmes est commentée dans le paragraphe suivant.

3.5 Comparaison entre les algorithmes RMR et PSM modifiés

Ces deux algorithmes procurent généralement des résultats très similaires, particulièrement lorsqu'il existe de manière évidente une région dominante dont le mouvement peut se décrire à l'aide du modèle paramétrique choisi. Cependant, certaines différences, expliquant en partie les différents comportements observés sur les résultats synthétiques et la séquence ROND-POINT, peuvent être identifiées:

1. supposer que le mouvement est constant uniquement sur les niveaux de résolution les plus grossiers est une hypothèse plus faible que supposer cela sur une première estimation multirésolution complète. Cela implique que l'algorithme RMR modifié permet de récupérer une classe plus large de modèles affines que l'algorithme PSM modifié.

2. avec l'algorithme PSM, tous les points sont considérés de manière équivalente lors de la première estimation multirésolution, ce qui produit une estimation moyennée, même lorsqu'un modèle constant est utilisé. En revanche, l'algorithme RMR élimine les erreurs importantes dès les premières itérations dans l'estimation multirésolution.

Considérons maintenant l'aspect calculatoire. L'algorithme PSM nécessite plusieurs descentes à travers la pyramide. Sa complexité est approximativement celle de l'algorithme MCMR multipliée par le nombre de passes. Dans l'algorithme RMR, il n'y a qu'une descente du plus grossier au plus fin, mais le calcul de chaque incrément $\Delta\Theta_k$ implique une minimisation par la méthode des MCPI.
En fait, le tableau 3.2 indique (pour notre implémentation sans optimisation particulière)

Algorithme	MCMR	RMRmod	PSMmod
Sparc 2	11,66	11,85	38,20
Sparc 10	4,05	4,05	12,60

TAB. 3.2 - *Temps cpu moyen (en secondes) de l'estimation. Les dix images de taille 224×256 de la séquence ROND-POINT sont considéreés; le nombre d'incréments à un niveau est toujours limité à 8; le nombre d'itérations des MCPI est limité à 6, et le nombre de descentes de pyramide est également limité à 6 pour l'algorithme PSM modifié.*

que la complexité des algorithmes MCMR et RMR est équivalente. Ceci s'explique de la

façon suivante. D'une part les MCPI convergent très rapidement, et donc, cette minimisation ne consomme pas trop de temps cpu, et d'autre part, le nombre d'incréments à calculer à un niveau pour atteindre la stabilité est moins important avec l'algorithme RMR qu'avec les algorithmes MCMR ou PSM grâce aux MCPI (c.à.d., les incréments sont mieux estimés). Ainsi, comme chaque calcul d'incrément implique l'interpolation de l'image des intensités en chaque point (nous effectuons une interpolation bilinéaire), l'algorithme RMR épargne de ce fait du temps cpu comparativement aux deux autres algorithmes.

3.6 Conclusion

Nous avons présenté dans ce chapitre une méthode d'estimation robuste multirésolutions de modèles paramétriques de mouvement. [BA93b, BK94, LF94] ont proposé des schémas similaires. En fait, deux (voire quatre) variantes ont été décrites et ont été comparées favorablement à une technique de moindres-carrés multirésolution, à travers des évaluations relativement conséquentes sur des exemples synthétiques et des données réelles correspondant à des scènes complexes. Les résultats obtenus montrent que ces algorithmes sont capables d'estimer le mouvement global dans l'image ou dans une zone de l'image, sans que la présence de mouvements secondaires éventuellement significatifs ou de zones où la mesure est mal conditionnée ne vienne perturber cette estimation. L'utilisation de cet estimateur de mouvement, qui ne requiert pas de segmentation préalable de l'image, sera d'un intérêt évident pour effectuer la détection d'objets mobiles dans le cas d'une caméra en mouvement, comme il est décrit dans le chapitre qui suit.

3.7 Annexe : utilisation des modèles paramétriques pour la localisation de points singuliers dans une image

L'un des buts principaux en vision par ordinateur est de fournir à un système des informations lui permettant de comprendre son environnement et d'interagir avec celui-ci. Par exemple, un robot peut avoir besoin d'une représentation 3D du monde qui l'entoure pour se mouvoir. Cependant, pour décider effectivement de son déplacement, le robot peut également appréhender de manière qualitative la situation dans laquelle il se trouve vis-à-vis de son environnement avant d'exploiter éventuellement (si nécessaire) des mesures 3D quantitatives [Nag88a]. Ainsi, dans un certain nombre d'applications comme par exemple l'évitement d'obstacles, une information qualitative peut être obtenue de manière plus fiable et plus rapide sans passer par une phase de reconstruction 3D explicite [NA89, BF93]. Le passage d'une représentation numérique à une description symbolique permet de réduire considérablement la quantité des données. Néanmoins, cette réduction ne doit

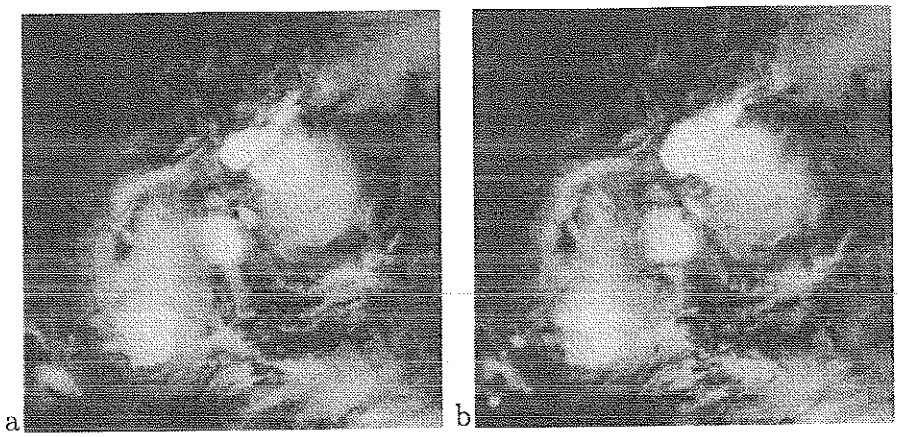


FIG. 3.16 - *Séquence météo : a) première image ; b) deuxième image ;*

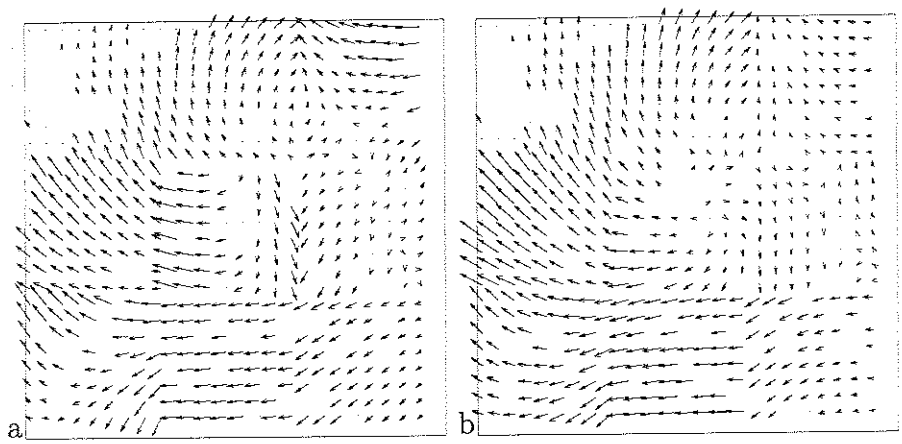


FIG. 3.17 - *Champ des vitesses associé au modèle de mouvement obtenu dans chaque bloc: a) MCMR ; b) PSM modifié* ($C = 18$)*;*

pas s'accompagner d'une perte d'information. Le choix des primitives à extraire dépend donc en partie de l'application traitée. Dans ce qui suit, nous considérons des images météorologiques. Nous montrerons tout d'abord les résultats d'estimation du mouvement apparent des nuages (utiles pour fournir des cartes des vents) à l'aide d'un modèle affine 2D. En se basant sur la notion de portrait de phase des équations différentielles ordinaires, nous indiquerons comment cette mesure du modèle affine de mouvement peut alors être utilisée pour obtenir une description symbolique du champ des vitesses dans ces séquences. Celle-ci passe par la détection et la caractérisation de points singuliers de ce champ, notamment les vortex. Cette méthode constitue ainsi une alternative à la détermination explicite du champ dense des vecteurs vitesses.

0,5	2,0	-1,1
3,4	2,3	2,6
0,7	-0,6	4,0

TAB. 3.3 - *Valeurs du paramètre ξ estimées par bloc.*

3.7.1 Estimation de champ de déplacement dans une séquence d'images météorologiques

Considérons les figures 3.16.a et 3.16.b. Elles correspondent à deux images météorologiques Météosat successives acquises à une demi-heure d'intervalle dans le canal infrarouge. Il s'agit en fait d'une partie de ces images couvrant une zone équatoriale. La scène est constituée principalement d'un développement ascentionnel de forte convexion entraînant un mouvement de "rotation spiralée" dont le centre se situe au cœur du développement (coin supérieur droit du bloc central). Cependant, sur la droite de l'image se trouve des nuages de moyenne altitude tournant globalement en sens inverse, et un second développement dans le bas à gauche dévie certains nuages vers le bas. De plus, de nombreux phénomènes de transparence entre couches sont présents, l'augmentation de la température de l'atmosphère entre les deux acquisitions dissipe une partie des nuages (surtout de la couche moyenne) et l'ascension des nuages provoque des variations de niveau de gris importantes. L'image a été divisée en neuf blocs de 75×75 sur lesquels le modèle affine de mouvement peut constituer une approximation à peu près valable. Cette approximation étant cependant plus fruste que dans les cas précédents, on utilise ici une variance plus élevée ($C = 18$). Les champs de vitesse $\vec{V}_{\hat{A}}$ obtenus sont présentés tels quels, sans post-traitement pour éliminer l'effet de bloc. L'algorithme MCMR obtient des résultats qui sont dans l'ensemble convenables (Fig. 3.17.a), mais montre ses limites dans le bloc central du fait de la présence du mouvement ascentionnel. En revanche, l'algorithme PSM modifié, de même que l'algorithme RMR modifié qui donne des résultats similaires, retrouve tout à fait le mouvement de rotation dans ce bloc central et l'on peut remarquer que les liaisons entre blocs sont plus cohérentes. Les valeurs du paramètre ξ sont données dans le tableau 3.3, et prouvent qu'il peut être opportun d'introduire ce terme complémentaire dans le modèle. Un $\xi > 0$ indique en fait une diminution moyenne des niveaux de gris sur le bloc. Les valeurs estimées le sont presque toutes, ce qui est cohérent avec la dissipation constatée de certains nuages sur la séquence complète.

Plaçons nous maintenant selon le point de vue du météorologue, qui effectue une analyse qualitative de cette séquence particulière. Le champ des vitesses dans ce cas sera pour lui une information secondaire, contrairement à la variation de la surface des taches blanches qui déterminera l'évolution et l'importance des précipitations [Arn92], et éventuellement la détermination de la trajectoire de la position du vortex que l'on peut

distinguer dans le bloc central du champ estimé avec la méthode robuste (figure 3.17.b). En effet, pour les vortex dépressionnaires[8], ces trajectoires peuvent servir d'indices pour déterminer l'intensité et la persistance des dépressions. Plus précisément, une trajectoire qui s'incurve indique généralement que la dépression est en train de faiblir.

La localisation de ces vortex pourrait se faire en estimant tout d'abord la carte des vents pour en extraire ensuite leur position. Une alternative à cette méthode consiste à utiliser le modèle de mouvement affine. En l'estimant sur des supports suffisamment importants, nous obtiendrons une mesure fiable de ce modèle qui, comme l'indique la figure 3.17.b, nous permet par la suite de localiser les vortex. Cette méthode évite ainsi la complexité et la difficulté de l'estimation d'un champ dense des vecteurs vitesses.

3.7.2 Interprétation qualitative d'un champ de vecteurs à l'aide des portraits de phase

Une méthode pour analyser qualitativement un champ de vecteurs $V(\mathrm{p})$, où p est un point du plan, est de ne considérer que l'orientation de ces vecteurs. Celles-ci procurent en général une information caractéristique sur le mouvement. Par exemple, dans l'image 3.17.b, c'est l'allure du champ des vitesses par l'intermédiaire de l'orientation des vecteurs plutôt que de leur module, qui nous renseigne sur la présence du vortex. Pour analyser cette orientation, il est alors possible de faire appel à la notion de portraits de phase. Ceux-ci sont des représentations géométriques qualitatives des solutions d'un système d'équations différentielles [AP82]. Pour analyser le champ V, supposé continu, considérons le comme le second membre de l'équation différentielle suivante:

$$\dot{\mathrm{p}}(\tau) = V(\mathrm{p}) \tag{3.39}$$

dans laquelle p est considérée comme dépendant de la variable τ. Le portrait de phase est alors constitué par les courbes intégrales des solutions de cette équation. D'après cette équation, on constate que ce sont des courbes telles qu'en chaque point, la tangente est dirigée par le vecteur du champ V en ce point. Par exemple, dans le cas de la mécanique des fluides, ces courbes correspondent aux lignes de courant et caractérisent à un instant donné t l'écoulement du fluide. Les solutions constantes de ce système sont des solutions particulières qui jouent un rôle important. Une solution constante $\mathrm{p}(\tau) = \mathrm{p}_0$ n'existe que si $V(\mathrm{p}_0) = 0$. Si cette condition est vérifiée, on dit alors que p_0 est un point singulier, ou point fixe du système dans la mesure où p reste en p_0 pour toute valeur de τ. Une grande partie de l'information qualitative sur la structure du champ V se trouve au voisinage de ces points singuliers. De plus, nous pouvons, comme il est indiqué plus loin, considérer une approximation au premier ordre du champ V. Pour cette approximation, le système (3.39) s'écrit sous la forme:

$$\dot{\mathrm{p}}(t) = \mathcal{A}\mathrm{p} + b \tag{3.40}$$

8. Ceci est en fait surtout valable pour les vortex dépressionnaires des zones tempérées

où $\mathcal{A}$ est la matrice des coefficients que nous supposerons régulière. L'unique point fixe est alors donné par:

$$p_0 = -\mathcal{A}^{-1}b \tag{3.41}$$

En effectuant le changement de variable $q = p - p_0$, on arrive au système linéaire suivant:

$$\dot{q}(t) = \mathcal{A}q \tag{3.42}$$

Pour analyser qualitativement un tel système, il suffit d'étudier les valeurs propres de la matrice $\mathcal{A}$ et la dimension de leur sous-espace propre. Suivant la nature de ces valeurs propres, on détermine alors la forme de Jordan, qui est une matrice réduite équivalente à $\mathcal{A}$, ainsi que le portrait de phase qualitatif qui lui est associé. On obtient alors des critères pour classer les différents champs de vitesse obtenus autour du point critique dans le cas linéaire (ou affine). Les différentes classes sont présentées dans le tableau (3.4), en fonction de la nature des valeurs propres.

Si l'on revient maintenant au champ V différentiable, on peut considérer comme on l'a dit une approximation linéaire de ce champ en un point p_0. On a alors les deux théorèmes importants suivants cités dans [AP82]:

1. si p_0 est un point fixe, alors dans un voisinage de ce point, les portraits de phase du champ original et de sa linéarisation sont qualitativement équivalents[9], à condition que la linéarisation ne corresponde pas à une rotation pure.

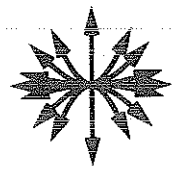
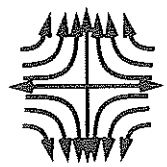
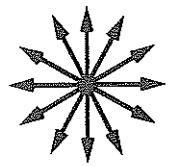
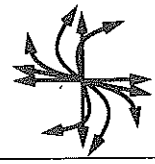
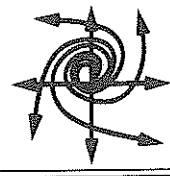
2. tous les portraits de phase en des points ordinaires –i.e. qui ne sont pas des points fixes– sont qualitativement équivalents.

Le premier résultat indique clairement qu'en général, au voisinage d'un point critique, la classification associée à l'approximation linéaire du champ (tableau (3.4)) permet également de caractériser le champ réel. En revanche, le second précise qu'en théorie, en dehors des points fixes, cette même classification fournira une information qualitative instable et donc non pertinente. L'analyse du champ global fait alors intervenir d'autres notions, comme celle de *directions principales* ou de *lignes séparatrices*, qui sont les courbes intégrales particulières qui passent par les points fixes (dans le cas de valeurs propres réelles, voir tableau (3.4)).

L'analyse précédente ne fait intervenir que la notion de champ de vecteurs et plus précisément celle d'orientation, et a de ce fait été utilisée dans plusieurs domaines.

En mécanique des fluides, les chercheurs et les ingénieurs se trouvent confrontés à l'interprétation d'une énorme quantité de données obtenues à partir d'intenses simulations ou

9. Par équivalence qualitative, il est entendu que les deux portraits de phase sont homéomorphes, c'est-à-dire qu'il existe (localement) une bijection, continue dans les deux sens, qui transforme un portrait de phase en l'autre portrait.

valeurs propres	forme de Jordan	type	portrait de phase
réelles distinctes λ_1 et λ_2	$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ λ_1 et λ_2 de même signe	nœud	
	$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ λ_1 et λ_2 de signes différents	selle	
égales à λ_1	$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_1 \end{pmatrix}$	nœud en étoile	
	$\begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix}$	nœud impropre	
complexes $\lambda_1 = \alpha + i\beta$ $\lambda_2 = \alpha - i\beta$	$\begin{pmatrix} 0 & -\beta \\ \beta & 0 \end{pmatrix}$	rotation pure	
	$\begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$	rotation en spirale	

TAB. 3.4 - *Classification des différents portraits de phase en fonction de la nature des valeurs propres et de la forme de Jordan associée.*

d'expérimentations réelles. De plus, les variables qui interviennent dans les phénomènes simulés ou observés sont de nature très diverse, et les lois qui régissent leur évolution sont très complexes. [HH91] ont développé une représentation des écoulements basée sur les points singuliers et les courbes séparatrices qui fait ressortir la topologie du champ des vitesses et aide à la compréhension des phénomènes observés. Dans [ZHA94], les valeurs des invariants du tenseur de déformation, exploitées dans [PC87], sont comparées et utilisées pour localiser les turbulences et les points de selle dans des champ de vecteurs obtenus par simulation numérique ou par expérimentation [IO83].

Les textures orientées que l'on rencontre sur des exemples aussi différents que des images de visualisation d'écoulements fluides, de coupes longitudinales de troncs d'arbres, d'empreintes digitales, ou de coupes géologiques, représentent en fait directement des portraits de phase [KW87]. Le schéma d'analyse présenté plus haut peut alors s'appliquer directement sur le champ des "orientations des textures" préalablement extrait de ces images. Dans [RJ92], un portrait de phase local est calculé par une méthode de minimisation non-linéaire sur une grille régulière de l'image. Les points singuliers obtenus à partir de ces portraits sont alors validés par une méthode de vote similaire à une transformée de Hough. [SF93] utilisent l'index de Poincaré pour détecter les points singuliers dans un champ d'orientations correspondant à des images (statiques) de visualisation d'écoulement de fluide. En modélisant les noeuds comme des sources (ou des puits) d'écoulement, et les spirales comme des tourbillons, les puissances de ces derniers et des sources sont estimées et employées pour resynthétiser le champ des déplacements. [DJ93] proposent un algorithme d'estimation aux moindres-carrés pour calculer le portrait de phase linéaire, dans lequel la sensibilité de l'estimation vis-à-vis des données est prise en compte. La classification des portraits de phase est basée dans cet article sur les invariants du premier ordre de la matrice $\mathcal{A}$ introduite plus haut –la divergence, le rotationnel et la déformation linéaire– et conduit à un même schéma de classification qu'avec les valeurs propres.

Il est intéressant de remarquer que l'étude de ces invariants a été menée dans le cas de l'analyse et de l'interprétation du mouvement dans des séquences d'images [KD75, FB90]. Dans ce même domaine, [VGT89], en faisant appel à la théorie des équations différentielles présentée plus haut, montrent que la trajectoire des points singuliers dans l'image et la nature de leur portrait de phase contiennent une grande partie de l'information sur le mouvement 3D.

3.7.3 Localisation et caractérisation des points singuliers dans une séquence d'images

Les articles précédemment cités qui recherchent les points singuliers dans des champs de vecteurs utilisent des champs denses d'orientation extraits d'images statiques [RJ92, SF93, DJ93] ou directement à partir du champ à analyser [HH91, ZHA94].

Dans le cas de l'analyse des séquences météorologiques, pour éviter le calcul du champ dense des déplacements, nous sommes passés directement par l'estimation robuste du modèle affine. Dans l'exemple traité précédemment, l'analyse des modèles affines estimés fait ressortir la présence d'un point fixe correspondant à une rotation en spirale dans le bloc central. Cependant, dans cet exemple, nous avons déterminé par avance la taille des blocs. Par ailleurs, la position figée des blocs, support de l'estimation des modèles affines, peut ne pas être idéale par rapport à la localisation du point fixe. L'idée est alors d'utiliser des approximations linéaires successives du champ des vitesses. En partant d'une fenêtre rectangulaire, le modèles de mouvement affine est estimé, et nous renseigne alors sur la position éventuelle d'un point fixe par l'intermédaire de la formule (3.41). Le support est alors déplacé en direction de ce point, et une nouvelle approximation est effectuée. Dans le cadre d'un stage de DEA [Mau94], ce schéma a été testé, en faisant particulièrement attention aux trois points suivants:

1. le déplacement vers le point fixe: il ne doit pas excéder la taille du support (i.e. de la fenêtre), dans la mesure où l'approximation est essentiellement valide sur celle-ci.

2. la taille du support: elle doit être suffisamment importante pour pouvoir effectuer une estimation correcte en présence de bruit, et pas trop pour que l'approximation linéaire reste valable. Dans l'algorithme, cette taille est choisie de manière adaptative. Plus précisément, un test "emboité", basé sur l'estimation d'un paramètre de nuisance [Iou94] est effectué. Il permet de comparer l'effet d'un accroissement de la taille du support à la fois sur la localisation du point fixe et sur l'incertitude de cette localisation. L'accroissement du support est validé quand l'écart de localisation du point singulier entre deux tailles de support n'excède pas l'incertitude sur cette localisation.

3. la décision pour qualifier le portrait de phase: le choix du portrait de phase équivalent au modèle de mouvement estimé nécessite de tester la nature des valeurs propres (voir le tableau 3.4), ou plus directement, la nullité de la trace et du déterminant de la matrice des coefficients linéaires. Ces tests sont effectués en prenant en compte la variance des estimés des paramètres de mouvement.

Les résultats obtenus sont très encourageants. Des résultats sur images réelles avec des mouvements synthétiques complexes ont validé l'approche. Dans les séquences réelles traitées, certains points fixes détectés ne semblent pas correspondre à des points singuliers réels. L'utilisation de plus de deux images de la séquence devrait permettre d'éliminer ces fausses détections.

Chapitre 4

Détection du mouvement dans le cas d'une caméra mobile

4.1 Introduction et choix de l'approche

L'objet de ce chapitre est la détection dans une séquence d'images des éléments mobiles de la scène lorsque la caméra est elle-même en mouvement. Après avoir introduit le problème et présenté les grandes lignes de notre approche, nous nous intéresserons dans un premier temps à un schéma de détection s'appuyant sur deux images uniquement, schéma qui sera aussi utilisé dans le cadre de la segmentation spatio-temporelle décrite au chapitre suivant. Nous considérerons ensuite un support temporel plus étendu pour améliorer la détection, puis nous évoquerons les problèmes calculatoires liés à l'algorithme. Enfin, des résultats sur séquences réelles seront présentés et l'on discutera de l'influence des différents paramètres introduits dans notre modélisation.

4.1.1 Présentation du problème

Dans ce chapitre, nous supposerons que l'estimateur décrit dans le chapitre précédent a permis d'obtenir le paramètre $\widehat{\Theta}_t$ modélisant le mouvement dominant sur le support d'estimation F entre les instants t et $t+1$. Il s'agit alors de savoir dans quelle mesure le champ $\vec{V}_{\widehat{\Theta}_t}$ estimé rend bien compte du déplacement réel (figure 4.1). Le problème de détection que nous traitons ici est donc celui de la détermination des points du support F dont le mouvement est non-conforme au champ $\vec{V}_{\widehat{\Theta}_t}$, c'est-à-dire la détermination des points p dont le mouvement 2D réel s'écarte du vecteur vitesse $\vec{V}_{\widehat{\Theta}_t}(\mathrm{p})$ calculé à l'aide du vecteur de paramètres $\widehat{\Theta}_t$. Par la suite, nous dirons que ces points forment les régions (de mouvement) non-conformes, ou les régions "mobiles". Cependant, le terme de région "mobile", qui recouvre une notion qualitative sur les objets tridimensionnels observés, est à employer dans notre cas avec prudence. En effet, considérons les cas suivants:

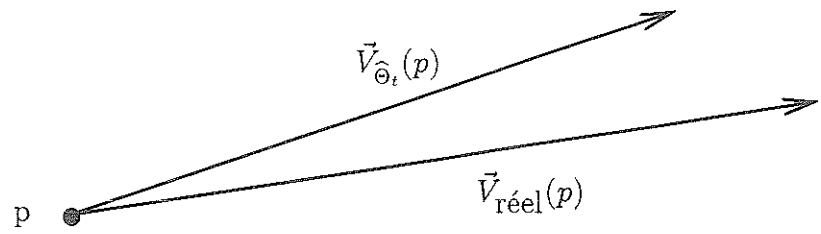


FIG. 4.1 - *Problème de la détection de mouvement par compensation: le vecteur* $\vec{V}_{\widehat{\Theta}_t}(\mathrm{p})$ *est-il un bon estimé de* $\vec{V}_{réel}(\mathrm{p})$ *?*

- F est l'image entière, et l'on fait l'hypothèse que le champ $\vec{V}_{\widehat{\Theta}_t}$ rend compte du mouvement de la caméra, c'est-à-dire modélise le déplacement dans l'image induit par le mouvement de la caméra, de l'ensemble des régions fixes de la scène observée. Dans ces conditions, régions mobiles et régions non-conformes sont équivalentes;

- cependant, dans le cas précédent, si le mouvement de la caméra comporte une composante translationnelle importante relativement à la distance séparant la caméra des objets observés, le champ $\vec{V}_{\widehat{\Theta}_t}$ ne permet pas de modéliser le déplacement apparent de l'ensemble des régions statiques. Notamment, dans le cas où la scène présente des éléments situés à des profondeurs bien distinctes (par exemple avant plan et arrière plan), l'estimation va produire un modèle de mouvement qui décrira correctement le mouvement de l'un de ces éléments, mais sans rendre compte du déplacement des autres. Dans ce cas, la notion de régions non-conformes recouvre à la fois celle de régions mobiles et celle de régions situées à des profondeurs différentes de celles dont les projections ont produit le mouvement dominant estimé;

- enfin, si le support F correspond à une région réellement mobile que l'on suit dans la scène, les régions non-conformes correspondront aux autres régions mobiles et, également, aux régions statiques du monde 3D observé!

En conclusion, la notion de "région mobile" est toute relative au paramètre $\widehat{\Theta}_t$ estimé, et nous nous efforcerons dans ce qui suit d'utiliser de préférence le terme de "région non-conforme", plus approprié à notre problème.

4.1.2 Notations - Rappels

Nous rappelons et précisons ici quelques notations que nous utiliserons dans ce chapitre. Nous désignerons toujours par I^k une image de la séquence à l'instant k, $\widehat{\Theta}_k^{k+1}$ le jeu de paramètres de mouvement estimés entre les instants k et $k+1$, et $\vec{V}_{\widehat{\Theta}_k^{k+1}}(\mathrm{p})$ le vecteur de vitesse (ou de déplacement) au point p calculé à l'aide du vecteur de paramètres $\widehat{\Theta}_k^{k+1}$. Lorsqu'il n'y aura pas d'ambiguïté, nous écrirons $\widehat{\Theta}_k = \widehat{\Theta}_k^{k+1}$. De plus, nous noterons par

$\text{comp}_{\Theta_j^i}()$ l'application:

$$X^i \longmapsto \tilde{X}^j = \text{comp}_{\Theta_j^i}(X^i) \tag{4.1}$$

qui transforme par compensation une image ou une partie d'image X à l'instant i en une image à l'instant j selon le mouvement de j à i paramétré par $\widehat{\Theta}_j^i$, par la relation:

$$\tilde{X}^j(\mathrm{p}) = X^i(\mathrm{p} + \vec{V}_{\widehat{\Theta}_j^i}(\mathrm{p})) \tag{4.2}$$

où les valeurs $X^i(\mathrm{p}+\vec{V}_{\widehat{\Theta}_j^i}(\mathrm{p}))$ sont obtenues par interpolation bilinéaire lorsque $\mathrm{p}+\vec{V}_{\widehat{\Theta}_j^i}(\mathrm{p})$ n'est pas sur la grille d'échantillonnage. Les paramètres $\widehat{\Theta}_j^i$ sont obtenus par estimation et "composition" des mouvements associés aux paramètres $\widehat{\Theta}_k^{k+1}$. Ainsi, nous définirons par exemple $\widehat{\Theta}_t^{t+2}$ comme étant le vecteur de paramètres du modèle de mouvement vérifiant:

$$\vec{V}_{\widehat{\Theta}_t^{t+2}}(\mathrm{p}) = \vec{V}_{\widehat{\Theta}_t^{t+1}}(\mathrm{p}) + \vec{V}_{\widehat{\Theta}_{t+1}^{t+2}}(\mathrm{p} + \vec{V}_{\widehat{\Theta}_t^{t+1}}(\mathrm{p})) \tag{4.3}$$

Notons ici que la "composition" est "interne" à l'ensemble des modèles affines, c'est-à-dire la composition de deux modèles affines est un modèle affine. Pour des modèles d'ordre plus élevé, ce n'est plus le cas. Avec ces notations, nous définirons:

$$\tilde{I}^t = \text{comp}_{\Theta_t^{t+1}}(I^{t+1}) \tag{4.4}$$

et la différence d'images déplacée (par le mouvement dominant) à l'instant t, $\text{DFD}_{\Theta_t^{t+1}}$ (*"Displaced Frame Difference"*) par:

$$\text{DFD}_{\Theta_t^{t+1}} = \tilde{I}^t - I^t \tag{4.5}$$

4.1.3 Approche choisie

Comme nous l'avons souligné dans l'état de l'art, la méthode que nous considérons, basée sur la compensation par le mouvement dominant, nous replace en quelque sorte dans l'hypothèse de caméra fixe. Dans ces conditions, on pourrait penser qu'il suffit de choisir l'une des nombreuses méthodes de détection définies dans ce cadre pour résoudre notre problème (ceci est suggéré dans [AKM93]). Cependant, comme nous avons pu le constater par exemple sur la séquence ROND-POINT(voir les images de la figure 3.15), la compensation n'est pas toujours parfaite, et il faut souligner que le modèle de mouvement utilisé ne constitue qu'une approximation du mouvement réel. Par conséquent, les simples mesures de changement temporel d'intensité, qui sont employées dans la majorité des articles traitant de la détection de mouvement avec caméra fixe, ne sont plus adaptées comme nous le montrons dans le paragraphe 4.2.1 sur le choix des mesures. Dès lors, nous privilégierons l'emploi de mesures locales et partielles de mouvement, plus appropriées à notre problème, pour approcher au mieux le principe de détection présenté sur la figure 4.2. Sur ce schéma, les paramètres estimés $\widehat{\Theta}_t$ servent donc (par l'intermédiaire de $\vec{V}_{\widehat{\Theta}_t}$) à générer la séquence compensée. Dans cette dernière, il nous faut faire la distinction entre

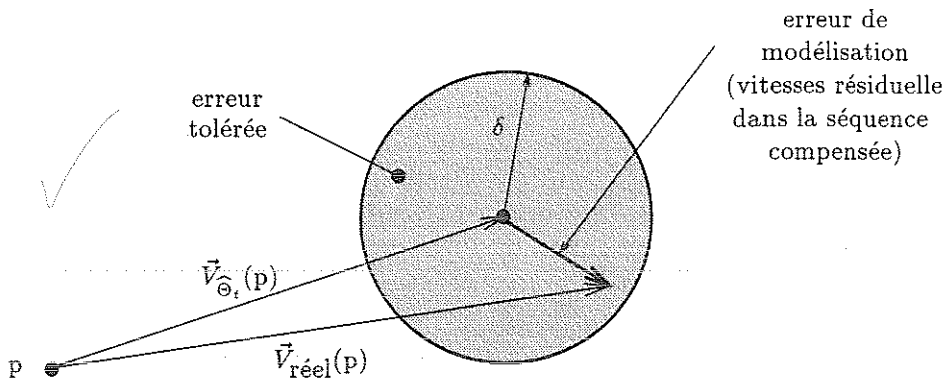


FIG. 4.2 - *Principe de la détection.*

d'une part les déplacements résiduels, dus par exemple à des erreurs de modélisation, et d'autre part les déplacements correspondant vraiment à des objets en mouvement (non pris en compte par le modèle de mouvement estimé $\widehat{\Theta}_t$). L'objectif à atteindre sera le suivant: en dessous d'un seuil δ, le déplacement résiduel est considéré comme une erreur de recalage; au dessus, comme celui d'un objet "mobile". Cependant, comme nous ne disposons pas du déplacement résiduel réel, mais de sa projection sur le gradient spatial, le critère que nous définirons ne constituera qu'une approximation de cet objectif.

Par ailleurs, la détection du mouvement est souvent considérée comme une première phase (de focalisation d'attention notamment, [ZP90, Nel91, GP93]) qui doit indiquer de manière approximative (spatialement) la présence d'objets mobiles. Cependant, dans la mesure où elle précède des traitements de plus haut niveau appliqués sur la (les) région(s) détectée(s), il nous semble important que les masques de ces régions soient les plus complets et les plus précis possibles. Ceci peut éviter une phase de localisation supplémentaire, comme dans [TM93], où des contours actifs (ou *"snakes"* en anglais) sont utilisés (après détection) pour délimiter les régions mobiles. Dès lors, pour combattre le bruit et palier l'absence d'informations fiables dans de nombreuses zones de l'image, l'emploi d'une technique de régularisation s'avère indispensable. Nous avons choisi la formulation markovienne, qui permet, dans un cadre mathématique cohérent, de spécifier simplement des interactions locales non linéaires entre des primitives qui peuvent être de nature différente, et qui a largement été utilisée dans le contexte de la détection avec caméra fixe, [AKM93, LB90, BL90, LPC94].

4.1.4 Modélisation Markovienne et estimation Bayesienne (critère du MAP)

Soit $S = \{s_1, s_2, \ldots, s_N\}$ l'ensemble des sites de l'image (en pratique l'ensemble des pixels), $\mathcal{V}$ un système de voisinage et $\mathcal{C}$ l'ensemble des cliques d'ordre deux associées à celui-ci. Soit également $D^t = \{D_s^t, s \in S\}$ le champ aléatoire des étiquettes de détection à

l'instant t. Les étiquettes de détection sont binaires, c'est-à-dire à valeur dans l'ensemble Λ à deux éléments: $\Lambda = \{c, nc\}$, où c signifie "conforme" et nc "non-conforme". Quant au champ des observations à l'instant t, $O^t = \{O_s^t, s \in S\}$, il comprend un ou plusieurs termes, suivant que l'on s'intéresse à la détection entre deux images uniquement, ou que l'on considère une séquence complète. De manière générale cependant, notre problème s'exprime comme la recherche de la carte de détection $\hat{d}^t$ qui a la plus grande probabilité *a posteriori* d'avoir produit le champ des observations à l'instant t, o^t, soit:

$$\hat{d}^t = \underset{d^t}{\operatorname{argmax}}\ p(D^t = d^t | O^t = o^t) \tag{4.6}$$

Après utilisation de la règle de Bayes, on obtient de façon équivalente:

$$\hat{d}^t = \underset{d^t}{\operatorname{argmax}}\ p(o^t|d^t)\, p(d^t) \tag{4.7}$$

où $p(o^t|d^t)$ représente la vraisemblance conditionnelle des observations et $p(d^t)$ la probabilité *a priori* du champ des étiquettes. Nous supposerons que le champ d'étiquettes est markovien. Alors, comme nous le montrons dans l'annexe A, le champ estimé $\hat{d}^t$ est également le minimum d'une fonction d'énergie $U(d^t, o^t)$:

$$\hat{d}^t = \underset{d^t}{\operatorname{argmin}}\ U(d^t, o^t) = \underset{d^t}{\operatorname{argmin}}\ U_1(d^t, o^t) + U_2(d^t) \tag{4.8}$$

où $U_1(d^t, o^t)$ est le terme d'énergie liant le champ des observations et le champ des étiquettes, et $U_2(d^t)$ le terme d'énergie associée à la probabilité *a priori* des étiquettes qui se décompose en somme de fonctions de potentiels V_c où chaque V_c ne dépend que des sites de c. La définition du terme U_1, explicitée plus loin dans la suite de ce chapitre, fera également intervenir des potentiels locaux.

Dans le paragraphe suivant, nous nous intéresserons à la détection de mouvement entre deux images. Nous spécifierons les observations de mouvement dont nous nous servirons, et nous indiquerons comment nous modéliserons les termes d'énergies U_1 et U_2. La prise en compte plus complète de l'axe temporel, par la considération d'un plus grand nombre d'images, sera décrite au paragraphe 4.3.

4.2 Détection de mouvement entre deux images

Au-delà du point de vue adopté pour traiter un problème, la réussite d'un algorithme dépend avant tout des observations utilisées. De plus, l'analyse de la fiabilité des observations et de leur pertinence permet généralement d'obtenir de meilleurs résultats. Lors de la phase d'optimisation du critère choisi, qui se fait par relaxation, les observations les plus "sûres" et les plus complètes seront "propagées" vers les régions où les observations ne procurent qu'une information parcellaire, voire inexistante. Ainsi, dans un domaine proche de la détection du mouvement, celui de l'estimation de champs denses de vitesse 2D, l'utilisation de mesures de fiabilité des observations [Ana89, Sin90], a donné de

meilleurs résultats [BFB92, BFB94] que les méthodes avec régularisation "isotropique", comme dans [HS81, LK81], qui ne tenaient pas compte de cet aspect.

Dans les paragraphes qui suivent, nous justifierons les observations que nous avons retenues, et nous montrerons comment l'information qu'elles apportent est prise en compte dans notre modélisation par l'intermédiaire des fonctions de potentiel définies.

4.2.1 Choix des observations

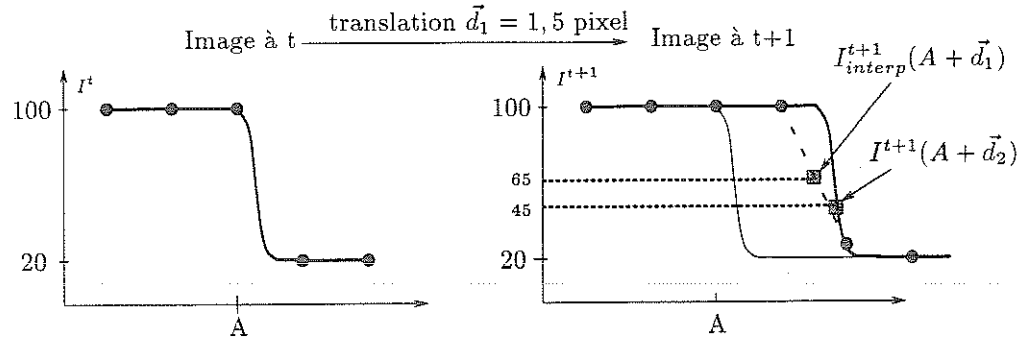
Dans le domaine de la détection d'objets mobiles, un certain nombre de mesures ont été définies. Dans [HJ83], les auteurs désignent par la quantité $|\mathrm{DFD}| \times \|\vec{\nabla} I\|$ la mobilité d'un contour. Cette mesure, utilisée pour détecter les contours en mouvement, s'avère en fait peu efficace et trop bruitée. Dans [LRB93], l'utilisation de simples différences temporelles entre images mais à différentes fréquences temporelles permet de faire la distinction entre "mouvements parasites" (généralement situés dans les hautes fréquences) et petits objets animés d'un mouvement lent (qui se retrouvent à toutes les fréquences, et en particulier dans les basses fréquences). Les auteurs de [LJ89] ont construit un module de détection basé sur le gradient spatio-temporel de l'intensité dans une image. Le comportement de ce gradient est cependant très proche de celui du vecteur vitesse normal que nous examinerons. Enfin, l'observation qui a été la plus fréquemment retenue est sans conteste la différence inter-image ou la différence d'images déplacée dans le cas d'une caméra en mouvement. Voyons maintenant pourquoi cette dernière n'est pas adaptée à notre problème.

Utilisation de la DFD

La figure 4.3 illustre dans le cas monodimensionnel les deux problèmes majeurs liés à l'emploi de la DFD . Dans cette figure, nous avons représenté un profil d'intensité qui s'est déplacé de $\vec{d}_1 = 1{,}5$ pixel entre les instants t et $t+1$, et nous avons considéré un point A situé près d'un contour.

Le premier inconvénient de la DFD (cas 1 de la figure 4.3) vient du fait que son calcul nécessite d'interpoler l'image à $t+1$. Or cette interpolation génère toujours une erreur, qui est surtout importante dans le voisinage des contours très contrastés de l'image. Par exemple, sur la figure 4.3, bien que le vrai déplacement soit utilisé pour effectuer la compensation au point A, l'emploi d'un interpolateur linéaire génère une DFD de l'ordre de -35, largement indicatrice d'un changement temporel! Bien sûr, des interpolateurs d'ordre supérieur permettent de limiter cette erreur, mais ils seront en revanche généralement plus sensibles au bruit d'acquisition et induiront une charge de calcul plus grande. Il est également possible d'atténuer cet effet en filtrant fortement l'image. Ce sera cependant au prix d'un effet de flou sur toute l'image qui réduira aussi la réponse sur les zones réellement à détecter, notamment à l'intérieur des objets mobiles de distribution d'intensité trop uniforme.

Le deuxième problème, plus critique, est intrinsèque au choix de la DFD comme observa-



Cas 1 : $DFD(A) = I^{t+1}_{interp}(A + \vec{d}_1) - I^t(A) = -35$

Cas 2 : $DFD(A) = I^{t+1}(A + \vec{d}_2) - I^t(A) = -55$ avec $\vec{d}_2 = 1,8$ pixel

FIG. 4.3 - *Problèmes liés à l'utilisation de la différence d'images déplacée dans un schéma de détection avec compensation de mouvement. Le cas 1 pose le problème de l'interpolation pour les déplacements non-entiers, et le cas 2 celui de l'influence sur la DFD d'une erreur faible du déplacement estimé.*

tion de mouvement. C'est le cas 2 de la figure 4.3, où l'interpolation est considérée comme parfaite, mais où l'on suppose que le déplacement réel $\vec{d}_1 = 1{,}5$ a été légèrement surestimé ($\vec{d}_2 = 1{,}8$). Il en résulte une DFD très importante de -55, qui ne semble pas en rapport avec l'erreur commise (0,3 pixel). Or, de telles erreurs de mesure (ou de modélisation) du vecteur vitesse risquent fort de se produire si l'on souhaite utiliser notre algorithme dans un contexte où les modèles de mouvement utilisés ne permettront pas d'obtenir un recalage exact de la séquence (ce qui est le cas par exemple des figures 3.15). Ceci nous incite donc à choisir une observation qui soit plus explicitement et plus exclusivement indicatrice de mouvement.

Utilisation de la vitesse normale

Comme nous l'avons vu dans la partie consacrée à l'estimation du mouvement dans l'état de l'art (relation (2.3)), l'équation de contrainte du mouvement apparent nous procure la mesure du vecteur vitesse dans la direction du gradient spatial de l'intensité, c'est-à-dire dans la direction perpendiculaire aux isophotes[1], mesure que nous appellerons vitesse normale ou déplacement normal. En assimilant $\frac{\partial I}{\partial t}$ dans la séquence compensée avec une diffence finie, nous noterons:

$$v_n(\mathrm{p}) = \frac{|\tilde{I}^t(\mathrm{p}) - I^t(\mathrm{p})|}{\|\vec{\nabla} I(\mathrm{p})\|} = \frac{|\mathrm{DFD}_{\Theta^{t+1}_t}(\mathrm{p})|}{\|\vec{\nabla} I^t(\mathrm{p})\|} \tag{4.9}$$

Ainsi, $v_n(\mathrm{p})$ s'interprète comme étant le module du déplacement normal calculé au point p

1. Dans le cas où on se trouve sur un point de contour spatial et si l'on assimile un contour à une isophote, il s'agira de la composante perpendiculaire au contour.

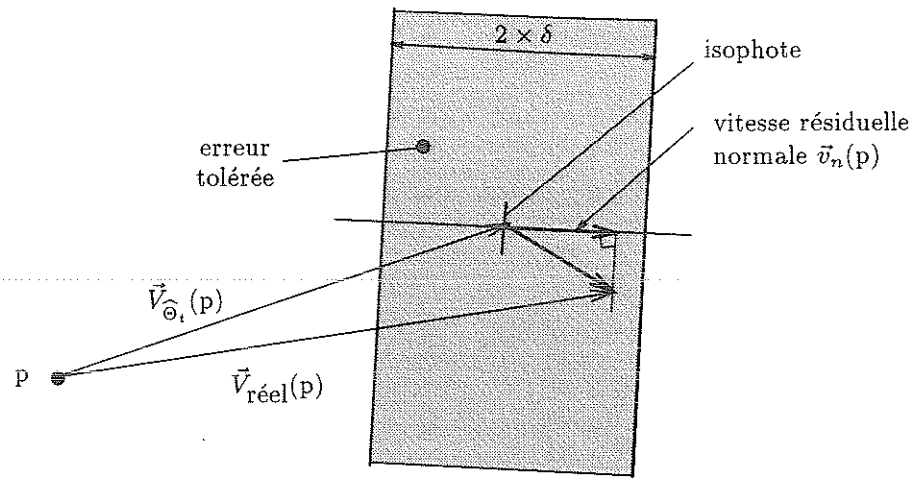


FIG. 4.4 - *Principe de la détection de mouvement: contrainte appliquée sur la vitesse normale.*

dans la séquence compensée, ou comme la mesure de la composante normale $\vec{v}_n(\mathrm{p})$ du vecteur vitesse résiduel. Imposer une contrainte sur v_n du type $v_n \leq \delta$ revient à tester si le déplacement résiduel se situe dans la bande grisée de la figure 4.4. Localement, un tel critère n'est donc pas satisfaisant dans la mesure où, suivant les directions relatives du déplacement et du gradient spatial de l'intensité, un vecteur résiduel important, indiquant donc un point de mouvement non-conforme, pourra malgré tout être considéré comme une simple erreur de recalage. La plus mauvaise configuration est obtenue lorsque les deux directions sont orthogonales, et correspond au cas où le contour glisse sur lui-même. Cette configuration particulière pose donc un problème, mais il est peu probable de la rencontrer en tous les points d'un voisinage. En effet, alors que le déplacement évolue généralement lentement en fonction de la position dans l'image, la direction du gradient spatial, elle, peut varier beaucoup plus rapidement en fonction de cette position. La figure 4.5 présente un cas où les contraintes appliquées en des points voisins permettent de détecter des erreurs de recalage. La première raison du choix de l'observation que nous retiendrons est justement qu'elle combinera en une seule mesure les multiples contraintes locales évoquées sur la figure 4.5, et que l'on pourra en déduire des mesures de fiabilité prenant en compte la distribution locale des gradients spatiaux d'intensité. La deuxième raison est explicitée ci-dessous.

Observation retenue: moyenne pondérée des vitesses normales

Dans [HJ83, LJ89], les auteurs rejettent l'utilisation de la vitesse normale comme mesure de détection en relevant sa grande sensibilité au bruit. De fait, v_n, étant formé d'un rapport, est par nature très sensible aux erreurs numériques lorsque numérateur et dénominateur tendent simultanément vers zéro, c'est à dire ici, dans les régions uniformes. Dans [BFB92, BFB94], Barron *et al.* font remarquer, après analyse des distributions d'erreurs

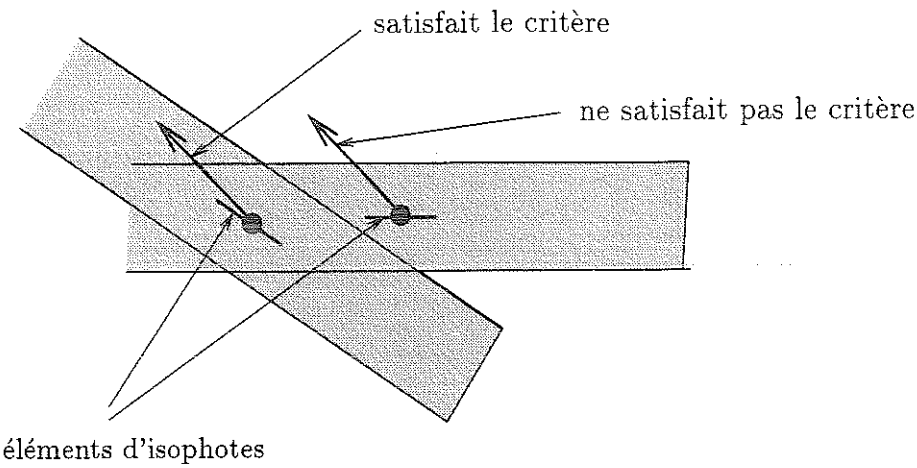


FIG. 4.5 - *Contraintes sur les vitesses résiduelles normales, appliquées en des pixels voisins. Pour le pixel de gauche, la configuration particulière de la vitesse résiduelle et du gradient de l'intensité semble indiquer que le déplacement résiduel correspond à une simple erreur de recalage. En revanche, l'orientation différente du gradient spatial du pixel voisin à droite permet d'indiquer que le mouvement n'est en fait pas conforme.*

de diverses méthodes d'estimation de champs de vitesse 2D, que le module du gradient spatial $\|\vec{\nabla}I\|$ constitue en fait une bonne mesure de fiabilité de la mesure de la vitesse normale: plus le gradient est important, plus la mesure de v_n est précise[2]. Ainsi, plutôt que d'utiliser un simple moyennage des vitesses normales, nous avons préféré utiliser la moyenne des vitesses normales pondérées par le carré du module du gradient spatial de l'intensité, comme le propose [IRP92], c'est-à-dire:

$$\begin{aligned} \mathrm{Mes}_{\Theta_t^{t+1}}(\mathrm{p}) &= \frac{\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})}\left(\|\vec{\nabla}I^t(\mathrm{q})\|^2 \times v_n(\mathrm{q})\right)}{\zeta + \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})}\|\vec{\nabla}I^t(\mathrm{q})\|^2} \\ &= \frac{\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})}\left(\|\vec{\nabla}I^t(\mathrm{q})\| \times |\mathrm{DFD}_{\Theta_t^{t+1}}(\mathrm{q})|\right)}{\zeta + \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})}\|\vec{\nabla}I^t(\mathrm{q})\|^2} \end{aligned} \tag{4.10}$$

où $\mathcal{F}(\mathrm{p})$ est une fenêtre de taille $(2T+1)\times(2T+1)$ autour de p, et ζ est une constante qui évite que le dénominateur ne s'annule. Nous reviendrons sur ce terme ζ dans le paragraphe suivant, et nous le considérerons comme absent dans l'analyse qui suit. Si l'on impose comme dans les cas précédents une contrainte sur la mesure, $\mathrm{Mes}_{\Theta_t^{t+1}}(\mathrm{p}) \leq \delta$, la région correspondant à l'erreur tolérée prend alors, qualitativement, l'allure d'une ellipse comme

2. Ceci est vrai tant que l'équation de contrainte du mouvement apparent reste vérifiée, ce qui n'est pas le cas en des points de contraste trop important. Cependant, ce point n'est pas trop problématique car les images peuvent être filtrées avant d'être traitées –notamment si l'on adopte une approche multirésolution–, ce qui élimine les hautes fréquences et donc ce type de contraste.

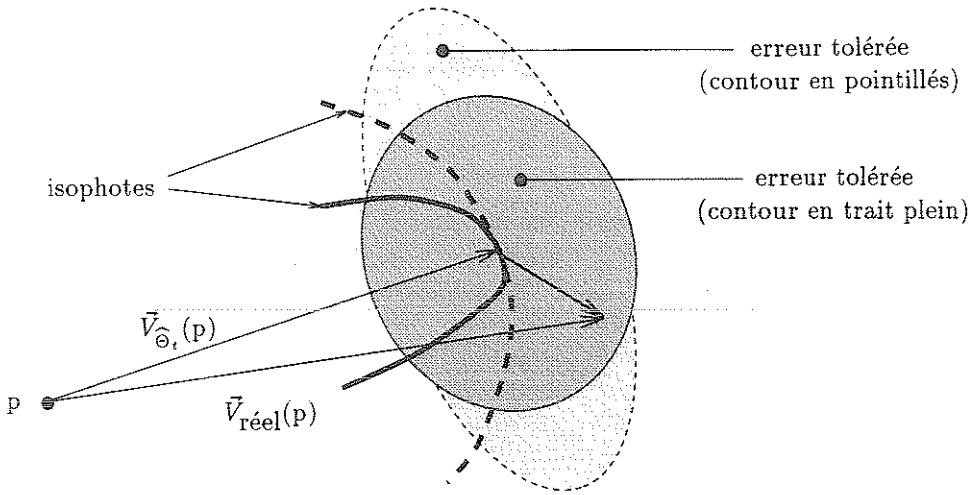


FIG. 4.6 - *Principe de la détection appliqué à l'observation des vitesses normales pondérées et moyennées (exemples de différentes courbures d'isophote).*

le montre la figure 4.6, et l'on se rapproche donc de notre objectif présenté sur la figure 4.2. Cependant, cette ellipse est plus ou moins allongée, suivant la courbure locale du contour[3] (la figure 4.6 présente deux exemples). Dans le cas limite où tous les gradients ont la même direction, l'ellipse dégénère en une bande, comme lorsque l'on utilise une seule vitesse normale (figure 4.4). Il serait donc intéressant de tenir compte de la de l'excentricité de ces "ellipses", ce qui sera évoqué au paragraphe suivant.

4.2.2 Fiabilité des observations

Pour étudier la fiabilité de l'observation choisie, nous allons nous placer dans la situation suivante: supposons que le pixel p et son voisinage subissent (dans la séquence compensée) une translation $\vec{\delta}$ entre les instants t et $t+1$. Qu'observera-t-on alors? Suivant la direction du déplacement $\vec{\delta}$, l'observation variera. Elle atteindra un minimum pour une ou plusieurs directions particulières, et de même pour le maximum. Si une expression simple de ce minimum et de ce maximum n'existe pas (il faut trouver le maximum et le minimum d'une fonction $f(\theta) = \sum_i \alpha_i |\cos(\theta - \theta_i)|$, où les α_i et θ_i dépendent du gradient spatio-temporel de l'intensité en chaque point q du voisinage), on peut en revanche en obtenir une minoration et une majoration.
D'une part, en chacun des pixels q du voisinage, on aura bien sûr:

$$v_n(q) \leq \delta \quad \text{où} \quad \delta = \|\vec{\delta}\|. \tag{4.11}$$

D'où, comme $\|\vec{\nabla} I^t(q)\|^2 \geq 0$,

$$\sum_{q \in \mathcal{F}(p)} \|\vec{\nabla} I^t(q)\|^2 \, v_n(q) \;\leq\; \delta \sum_{q \in \mathcal{F}(p)} \|\vec{\nabla} I^t(q)\|^2 \tag{4.12}$$

3. Pour facilité l'exposé, nous parlerons parfois de contour, mais il s'agit en fait d'une suite de points de gradients, qui ne sont pas nécessairement de "vrais" contours de l'image.

et donc

$$\mathrm{Mes}_{\Theta_t^{t+1}}(\mathrm{p}) = \frac{\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \left(\|\vec{\nabla} I^t(\mathrm{q})\|^2 \; v_n(\mathrm{q}) \right)}{\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \|\vec{\nabla} I^t(\mathrm{q})\|^2} \leq \delta \tag{4.13}$$

D'autre part, un calcul moins immédiat (voir l'annexe B), permet d'obtenir la minoration suivante proposée dans [IRP91]:

$$\mathrm{Mes}_{\Theta_t^{t+1}}(\mathrm{p}) \geq l(\mathrm{p}) \tag{4.14}$$

où $l(\mathrm{p})$ est donné par la formule:

$$l(\mathrm{p}) = \delta \times \lambda'_{min} \quad \text{avec} \quad \lambda'_{min} = \frac{\lambda_{min}}{\lambda_{max} + \lambda_{min}}, \tag{4.15}$$

dans laquelle λ_{max} et λ_{min} sont respectivement les valeurs propres maximales et minimales de la matrice M suivante (où $\nabla I^t(\mathrm{q}) = (I_x^t(\mathrm{q}), I_y^t(\mathrm{q}))$):

$$M = \begin{pmatrix} \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} (I_x^t)^2(\mathrm{q}) & \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} (I_x^t \times I_y^t)(\mathrm{q}) \\ \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} (I_x^t \times I_y^t)(\mathrm{q}) & \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} (I_y^t)^2(\mathrm{q}) \end{pmatrix} \tag{4.16}$$

On a donc l'encadrement suivant:

$$0 \leq l(\mathrm{p}) \leq \mathrm{Mes}_{\Theta_t^{t+1}}(\mathrm{p}) \leq \delta \tag{4.17}$$

dont l'intérêt est présenté sur la figure ci-après. Les observations possibles au pixel p

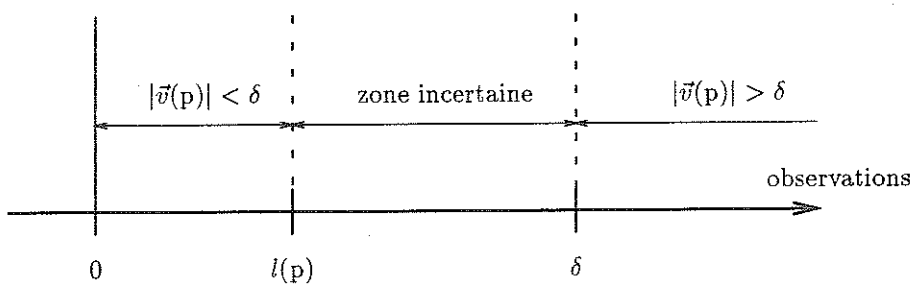


FIG. 4.7 - *Classement des observations.*

peuvent être classées en trois catégories: la première, constituée des observations inférieures à $l(\mathrm{p})$, indique théoriquement avec certitude que l'amplitude du déplacement réel $\vec{v}(\mathrm{p})$ dans la séquence compensée est inférieure à δ; la deuxième (observations supérieures à δ), indique que ce déplacement est supérieur à δ. Enfin, la troisième zone, qui regroupe les observations à l'intérieur de l'intervalle $[l(\mathrm{p}), \delta]$, ne permet pas de conclure de façon catégorique.

FIG. 4.8 - *Deux configurations locales extrêmes.*

Revenons à la borne $l(\mathrm{p})$. La figure 4.8 schématise les deux cas extrêmes suivants. Le premier cas est celui d'un pixel qui est situé sur une isophote rectiligne sur le voisinage considéré. La borne inférieure théorique est 0, et il existe bien une direction où cette borne sera atteinte (déplacement $\vec{\delta}_0$ dans la direction du contour). De même, la valeur supérieure δ est atteinte pour le déplacement $\vec{\delta}_1$.
A l'opposé, dans le deuxième cas, l'isophote est très courbée ce qui indique que la distribution locale des directions des gradients d'intensité sur cette isophote est mieux répartie entre les deux directions principales. Une telle isophote peut être schématisée par un coin (figure 4.8), et correspond au cas où la borne inférieure $l(\mathrm{p})$ est maximale ($\lambda_{min} = \lambda_{max}$ d'où $l(\mathrm{p}) = \frac{\delta}{2}$). Elle est ici atteinte lorsque le pixel p se déplace suivant les directions $\vec{\delta}'_0$ et $\vec{\delta}''_0$ des deux "segments" formant l'isophote. En revanche, quelle que soit la direction du déplacement (de module δ), l'observation sera théoriquement toujours bien inférieure à la borne supérieure δ que nous avons choisie. Ceci est un cas fréquent compte tenu des approximations grossières faites pour obtenir cette borne, et pourrait nuire à la qualité de la détection. Pour resserrer l'encadrement, nous avons alors considéré un second modèle plus "sophistiqué" de l'isophote en un pixel, constitué de deux segments dont on fait varier l'écart angulaire (voir annexe C). Ceci nous a permis de déterminer les bornes effectives de l'observation (avec ce modèle), toujours en fonction des valeurs propres de la matrice M (formule (4.16)). Ces bornes, minimale l_m et maximale L_m, sont données par:

$$\begin{cases} l_m(\mathrm{p}) = \delta \sqrt{\lambda'_{min}(1 - \lambda'_{min})} \\ L_m(\mathrm{p}) = \delta \sqrt{1 - \lambda'_{min}} \end{cases} \tag{4.18}$$

où λ'_{min} est définie par l'équation (4.15).

Pour valider nos encadrements, nous devons vérifier si la modélisation de l'isophote n'est pas trop grossière et si la formule (4.9) qui permet de mesurer effectivement le déplacement normal et qui correspond à l'équation de contrainte du mouvement apparent (dans notre analyse, nous avons supposé que le vecteur normal était exact) n'entache pas les observations d'une erreur trop importante rendant les bornes inutilisables. Nous avons mesuré, pour des mouvements synthétiques, la répartition des observations entre les différentes bornes. Plus précisément, à partir d'une image réelle I, nous avons généré une seconde image $\tilde{I}$ définie par $\tilde{I}(P) = I(P - \vec{\delta})$, qui n'est autre que I déplacée par une

translation $\vec{\delta}$. Cette image a été bruitée par un bruit gaussien additif centré de variance σ^2_{bruit}. Les images I et $\tilde{I}$ sont alors filtrées avec un filtre gaussien de variance σ^2_{gauss}. Nous avons ensuite calculé l'observation en chaque point de l'image I et regardé entre quelles bornes elle se situait. Nous avons réalisé la moyenne de ces observations sur l'ensemble de l'image, et nous avons répété ces opérations sur chaque image d'une séquence entière. Le tableau 4.1 donne les résultats que nous avons obtenus pour les séquences ROND-POINT et INTERVIEW (voir figure 4.25), en utilisant différentes orientations du déplacement synthétisé (le paramètre G_m apparaissant dans ce tableau est défini plus loin). Dans tous les tableaux, les valeurs des bornes correspondent à des valeurs moyennes.
A partir de ce premier tableau, nous pouvons faire quatre constats:

1. la borne inférieure l est de très bonne qualité, mais généralement assez faible et donc un peu trop lâche.

2. la borne inférieure issue de la seconde modélisation de l'isophote, l_m, est beaucoup plus élevée (presque 2,5 fois plus grande que l), mais "laisse passer" environ 10% des observations.

3. les expériences sur la séquence ROND-POINT mettent en évidence une anisotropie importante de la distribution des gradients d'intensité. Celle-ci est due sans doute à l'utilisation d'un capteur de qualité moyenne qui fait ressortir la structure de balayage de l'acquisition et l'entrelaçage.

4. les bornes supérieures L_m et δ ne sont ici pas fiables, étant dépassées par 15 à 35% des observations. Ces résultats empirent dramatiquement lorsque l'on ajoute un peu de bruit pour simuler le bruit d'acquisition (comparer par exemple les deux premières lignes du tableau 4.2).

Ce quatrième point est vraisemblablement dû, comme nous l'avons déjà mentionné, à l'indétermination de l'observation (4.10) lorsque numérateur et dénominateur sont proches de 0. Pour remédier à ce problème, nous avons imposé un "gradient minimum moyen" G_m en chaque pixel, mais ceci uniquement pour le calcul du dénominateur. L'observation que nous retiendrons finalement est donc:

$$\mathrm{Mes}_{\Theta_t^{t+1}}(\mathrm{p}) = \frac{\sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \left(\|\vec{\nabla} I^t(\mathrm{q})\| \times |\mathrm{DFD}_{\Theta_t^{t+1}}(\mathrm{q})| \right)}{N \times \max\left(G_m^2 \ , \ G_{moy}^2(\mathrm{p})\right)} \tag{4.19}$$

où N est le nombre de pixels de la fenêtre $\mathcal{F}$ ($N = (2T+1)^2$), et G_{moy} est le gradient moyen calculé:

$$G_{moy}(\mathrm{p}) = \sqrt{\frac{1}{N} \sum_{\mathrm{q}\in\mathcal{F}(\mathrm{p})} \|\vec{\nabla} I^t(\mathrm{q})\|^2} \tag{4.20}$$

Plutôt qu'une constante additive comme évoquée par la relation (4.10), nous avons préféré en fait cette formulation pour le dénominateur de $\mathrm{Mes}_{\Theta_t^{t+1}}(\mathrm{p})$. Ce choix a pour conséquence

$\sigma^2_{gauss} = 0{,}8$ $\delta = 1$ $G_m = 0{,}01$									
Séquence ROND-POINT									
déplace^t	$0 < \% <$	l	$< \% <$	l_m	$< \% <$	L_m	$< \% <$	δ	$< \%$
horiz.	1,17	0,12	12,99	0,26	69,82	0,94	6,23	1,0	9,80
diag.	2,52	0,12	6,39	0,26	64,02	0,94	5,45	1,0	21,62
vert.	0,74	0,12	3,55	0,26	53,26	0,94	9,34	1,0	33,11
Séquence INTERVIEW									
horiz.	1,28	0,11	8,28	0,26	67,91	0,94	6,77	1,0	15,76
diag.	1,87	0,11	6,54	0,26	66,22	0,94	5,50	1,0	19,87
vert.	1,21	0,11	6,68	0,26	66,03	0,94	7,69	1,0	18,40

TAB. 4.1 - *Répartition (en pourcentage %) des observations et moyenne des différentes bornes pour différentes directions de déplacement (images non bruitées).*

$\sigma^2_{gauss} = 0{,}8$ $\delta = 1$ direction diagonale										
Séquence ROND-POINT										
G_m	σ^2_{bruit}	$0 < \% <$	l	$< \% <$	l_m	$< \% <$	L_m	$< \% <$	δ	$< \%$
0,01	0,0	2,52	0,12	6,39	0,26	64,02	0,94	5,45	1,0	21,62
0,01	4,0	0,07	0,12	0,92	0,26	40,84	0,94	2,62	1,0	55,54
1,00	0,0	2,52	0,09	6,39	0,21	72,76	0,94	3,54	1,0	14,78
1,00	4,0	0,07	0,09	0,92	0,21	47,87	0,94	5,18	1,0	45,96
3,00	0,0	2,52	0,04	6,39	0,11	82,55	0,94	1,30	1,0	7,25
3,00	4,0	0,07	0,04	0,92	0,11	87,92	0,94	1,46	1,0	9,63
5,00	0,0	2,52	0,03	6,39	0,08	85,01	0,94	0,92	1,0	5,17
5,00	4,0	0,07	0,03	0,92	0,08	92,09	0,94	0,95	1,0	5,96
5,00	9,0	0,05	0,03	0,66	0,08	91,83	0,94	1,01	1,0	6,44
10,00	9,0	0,05	0,02	0,66	0,05	95,00	0,94	0,58	1,0	3,71

TAB. 4.2 - *Répartition (en pourcentage %) des observations et moyenne des différentes bornes pour différentes valeurs du bruit et du gradient d'intensité minimum imposé.*

de diminuer arbitrairement les observations dans les zones uniformes. Pour que nos encadrements restent valides, nous devons donc adapter dans le même rapport les bornes inférieures. Les formules définitives sont donc:

$$\begin{cases} l(\mathrm{p}) = \eta \times \delta \times \lambda'_{min} \\ l_m(\mathrm{p}) = \eta \times \delta \sqrt{\lambda'_{min}(1 - \lambda'_{min})} \end{cases} \tag{4.21}$$

où η est défini par:

$$\eta = \frac{G^2_{moy}}{\max\left(G^2_m\,,\, G^2_{moy}\right)} \tag{4.22}$$

Le tableau 4.2 récapitule les résultats obtenus pour différentes valeurs de G_m et de bruit σ^2_{bruit}. On peut noter l'amélioration importante obtenue en utilisant la nouvelle observation par le pourcentage nettement plus élevé de mesures qui se situent entre les bornes l_m et L_m, et surtout par le nombre beaucoup plus réduit d'observations qui se situent au-delà des bornes maximales. Bien sûr, il n'est pas bon d'accroître trop G_m car les bornes minimales moyennes diminuent également, ce qui agrandit les zones d'incertitude et donc accroît le niveau d'indécision. Cependant dans certaines séquences très bruitées (par exemple dans ROND-POINT), nous pourrons utiliser une valeur de $G_m = 10$ qui aura pour effet de réduire de façon très bénéfique les observations de mouvement dues à un repliement de spectre spatio-temporel, et d'effectuer la détection du mouvement en se basant de manière prépondérante sur les contours les plus importants de l'image, et les moins sujets à être corrompus par les divers artefacts d'acquisition.
L'effet de l'introduction du gradient minimum est encore plus visible sur les tableaux 4.3 et 4.4 dans lesquels les répartitions des observations sont présentées cette fois-ci en fonction des plages de gradient d'intensité auquelles appartient le gradient moyen calculé G_{moy} défini par (4.20). Nous avons indiqué chacune de ces plages par un intervalle situé à gauche dans la première colonne de ces tableaux. Le deuxième chiffre de cette première colonne donne le pourcentage de pixels dans la séquence dont le gradient moyen est situé dans la plage de gradient concernée. Comme on peut le constater, pour les plages de gradient faibles, l'accroissement du facteur G_m de 0,01 à 5 (tableaux 4.3 et 4.4) concentre les observations entre les bornes, ce qui est logique dans la mesure où les bornes inférieures sont devenues quasiment nulles.

Concernant les régions uniformes il est bon de rappeler les deux points suivants:

1. une variation temporelle d'intensité nulle n'est pas forcément un signe d'absence de mouvement, mais peut aussi refléter l'absence de gradient spatial;

2. une variation temporelle d'intensité n'est pas nécessairement un signe de mouvement, mais peut être le fait du bruit temporel ou du changement des conditions d'illumination (auquel l'œil est d'ailleurs très sensible dans ces régions).

$\sigma^2_{gauss} = 0{,}8$, $\delta = 1$, $G_m = 0{,}01$, $\sigma^2_{bruit} = 4$										
Séquence ROND-POINT										
G_{moy}	(%)	$0 < \% <$	l	$< \% <$	l_m	$< \% <$	L_m	$< \% <$	δ	$< \%$
0-1	(28)	0,01	0,22	0,02	0,39	1,21	0,88	0,77	1,0	98,00
1-2	(20)	0,01	0,14	0,09	0,32	14,21	0,93	4,03	1,0	81,66
2-4	(13)	0,06	0,08	0,58	0,24	48,51	0,96	4,74	1,0	46,11
4-6	(6)	0,12	0,06	1,59	0,20	67,39	0,97	3,47	1,0	27,44
6-10	(08)	0,23	0,07	2,48	0,19	74,31	0,96	3,49	1,0	19,49
10-	(25)	0,09	0,04	1,83	0,14	82,90	0,98	1,97	1,0	13,21
Total	(100)	0,06	0,12	0,86	0,26	40,52	0,94	2,60	1,0	55,96

TAB. 4.3 - *Répartition (en pourcentage %) des observations et moyenne des différentes bornes pour différentes plages de gradient moyen d'intensité. Le déplacement simulé est diagonal, et le gradient minimum imposé est quasiment nul ($G_m = 0{,}01$).*

$\sigma^2_{gauss} = 0{,}8$, $\delta = 1$, $G_m = 5{,}0$, $\sigma^2_{bruit} = 4$										
Séquence ROND-POINT										
G_{moy}	(%)	$0 < \% <$	l	$< \% <$	l_m	$< \% <$	L_m	$< \% <$	δ	$< \%$
0-1	(28)	0,01	0,00	0,05	0,01	99,93	0,88	0,00	1,0	0,00
1-2	(20)	0,02	0,01	0,16	0,02	99,83	0,93	0,00	1,0	0,00
2-4	(13)	0,07	0,03	0,67	0,08	98,89	0,96	0,09	1,0	0,28
4-6	(6)	0,12	0,06	1,60	0,17	78,29	0,97	2,68	1,0	17,31
6-10	(08)	0,23	0,07	2,48	0,19	74,31	0,96	3,49	1,0	19,49
10-	(25)	0,09	0,04	1,83	0,14	82,90	0,98	1,97	1,0	13,21
Total	(100)	0,06	0,03	0,89	0,08	92,03	0,94	0,96	1,0	6,05

TAB. 4.4 - *Répartition (en pourcentage %) des observations et moyenne des différentes bornes pour différentes plages de gradient moyen d'intensité. Même expérience que pour le tableau précédent, mais le gradient minimum imposé est 5.*

$\sigma^2_{gauss} = 0{,}8$, $\delta = 1$, $G_m = 5{,}0$, $\sigma^2_{bruit} = 4$, Variation $\Delta I = 5$										
Séquence ROND-POINT										
G_{moy}	(%)	$0 < \% <$	l	$< \% <$	l_m	$< \% <$	L_m	$< \% <$	δ	$< \%$
0-1	(28)	0,00	0,00	0,00	0,01	100,0	0,88	0,00	1,0	0,00
1-2	(20)	0,00	0,01	0,02	0,02	99,98	0,93	0,00	1,0	0,00
2-4	(13)	0,01	0,03	0,25	0,08	96,27	0,96	0,78	1,0	2,69
4-6	(6)	0,19	0,06	1,71	0,17	67,26	0,97	2,11	1,0	28,73
6-10	(08)	0,34	0,07	3,11	0,19	62,80	0,96	2,59	1,0	31,16
10-	(25)	0,13	0,04	2,14	0,14	78,86	0,98	1,73	1,0	17,15
Total	(100)	0,07	0,03	0,94	0,08	89,10	0,94	0,88	1,0	9,01

TAB. 4.5 - *Répartition (en pourcentage %) des observations et moyenne des différentes bornes pour différentes plages de gradient moyen d'intensité. Même expérience que pour le tableau précédent, mais on a simulé une variation d'intensité de 5 niveaux de gris entre chaque couple d'images.*

Les observations dans de telles régions ne sont donc pas fiables, et nous verrons comment prendre en compte cet aspect lors de la définition des énergies associées aux champs de Markov. Nous pouvons cependant déjà remarquer que, pour le calcul de l'observation (4.19), nous avons imposé un gradient minimum au dénominateur, mais pas au numérateur. Ainsi, dans ces régions uniformes, quelle que soit la valeur de la différence temporelle (donc même s'il y a une variation d'illumination), la mesure sera toujours presque nulle, le numérateur étant constitué du produit de la différence temporelle des intensités par le gradient spatial de l'intensité qui est proche de zéro. Comme la borne inférieure sera elle aussi très faible, les mesures dans de telles régions se situeront essentiellement dans la zone d'incertitude, et c'est la prise en compte du contexte, comme nous le verrons plus loin, qui permettra de décider si oui ou non il y a bien du mouvement.
Si l'on considère maintenant le cas des mesures dans les régions aux contrastes d'intensité importants, par exemple de gradient moyen supérieur à 10, ce qui représente 25% de la surface des images de la séquence ROND-POINT, on peut noter qu'elles sont peu sensibles aux variations des conditions d'illumination. En effet, notre mesure étant une moyenne pondérée de modules de vecteurs de déplacement normaux, une éventuelle variation d'illumination sera normalisée par le gradient et ne perturbera donc que faiblement la mesure. Ceci est en accord avec notre expérience visuelle: notre œil est peu sensible aux variations d'illumination près d'un contour dans la mesure où ces dernières n'affectent que très faiblement la valeur du contraste local.
Le tableau 4.5 présente les résultats obtenus lorsque l'on a simulé une variation d'illumination en ajoutant à la seconde image non pas un bruit centré, mais un bruit avec une moyenne de 5. Comme on peut le constater en comparant les tableaux 4.4 et 4.5, les répartitions des observations restent quasiment inchangées pour les faibles et les forts gradients spatiaux, conformément aux arguments invoqués plus haut.

En conclusion, au vu des tableaux 4.4 et 4.5, et de résultats similaires obtenus sur la séquence INTERVIEW, on constate que les observations se situent en grande majorité à l'intérieur des encadrements définis. De plus, nous choisirons de préférence l'encadrement obtenu avec les bornes l_m et L_m issues de la modélisation de l'isophote évoquée précédemment, notamment en raison de la borne inférieure l_m, qui est nettement plus stricte que la borne inférieure l du premier encadrement (voir tableau 4.4 par exemple). L'utilisation de cet intervalle dans la définition du champ de Markov est présentée dans le paragraphe suivant.

4.2.3 Modélisation de l'énergie liant les étiquettes aux observations

Rappelons ici que l'ensemble S des sites considérés est l'ensemble des pixels p de l'image, et nous noterons avec un indice s toute grandeur associée au site s. Ainsi, quel que soit le type d'encadrement des observations choisi ((4.17) ou (4.18)), nous noterons par exemple l_s et L_s les bornes inférieure et supérieure de celui-ci au site s.

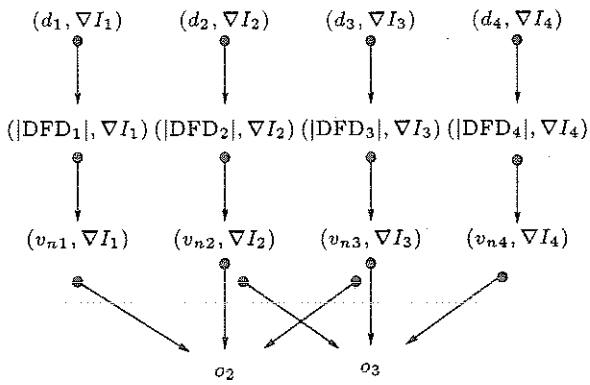


FIG. 4.9 - *Inter-dépendance entre observations et étiquettes.*

Dans le cas de la détection du mouvement entre deux images, les observations retenues sont les mesures décrites dans la partie précédente (nous ometterons ici d'indiquer t en exposant). Nous aurons donc en chaque site: $o_s = \mathrm{Mes}_{\Theta_i^{t+1}}(s)$. Nous avons choisi pour modéliser l'énergie liant les observations conditionnellement aux étiquettes, la forme suivante:

$$U_1(o, d) = \sum_{s \in S} V_1(o_s, d_s) \tag{4.23}$$

Dans cette modélisation, nous faisons donc dépendre la valeur du terme d'énergie lié à l'observation o_s en un site s de la seule connaissance de l'étiquette d_s en ce site. Or, comme l'indique la figure 4.9, puisque le calcul de l'observation o_s se fait à partir de mesures définies sur la fenêtre[4] $\mathcal{F}$, elle devrait dépendre des étiquettes de cette fenêtre si l'on considère celles-ci comme la réalisation d'un processus caché. Cependant une telle dépendance est difficile à modéliser en pratique, dans la mesure où le lien entre l'observation et les étiquettes est d'ordre qualitatif. Par souci de simplicité, nous avons donc retenu la forme (4.23). De plus, nous pouvons remarquer qu'elle est valide à l'intérieur des deux types de régions, conformes et non-conformes, et qu'elle ne sera mise en défaut qu'aux frontières. Elle aura pour conséquence une imprécision sur la localisation de ces dernières, imprécision qui dépendra de la taille de la fenêtre $\mathcal{F}$. C'est pourquoi nous nous sommes restreints au voisinage 3×3 pour le calcul des observations. Notons cependant que l'on peut aisément savoir où se localiseront les frontières de mouvement. En effet, les séparations entre régions conformes et non-conformes sont principalement situées aux frontières d'objets, et donc sur des gradients spatiaux *a priori* importants de l'image. Comme le calcul de l'observation fait intervenir le module au carré du gradient spatial comme poids dans le moyennage, il est clair que tous les voisins du contour subiront son influence (ce qui est l'effet désiré pour l'intérieur des régions!). La frontière de mouvement sera alors placée à l'extérieur de l'objet (ou de la région) duquel relève le gradient spatial, comme l'indique la figure 4.10. Cet effet, peu visible dans le cadre de la détection du mouvement

4. Sans compter la taille du filtre utilisé pour calculer les gradients spatiaux de l'intensité.

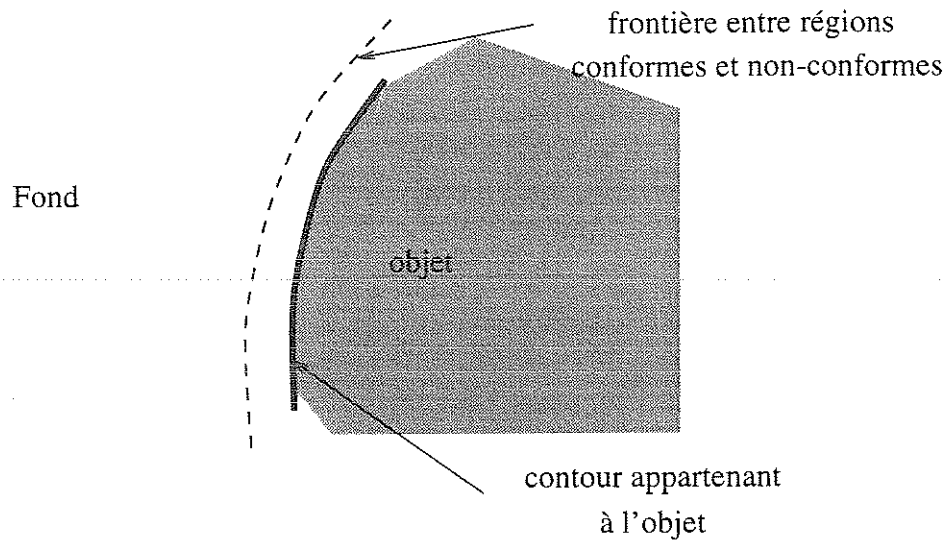


FIG. 4.10 - *Position attendue de la frontière entre région conforme et non-conforme, compte-tenu du voisinage utilisé pour le calcul de l'observation. L'objet mobile (c-à-d. non-conforme) est situé en avant d'un fond uniforme.*

(l'intégration temporelle des observations mettant en jeu plusieurs images successives aura en effet tendance à le gommer), sera plus perceptible pour la segmentation de mouvement.

Définition des fonctions de potentiel

Tout d'abord, introduisons la fonction $A_{\mathrm{tr},k}(x)$ que nous utiliserons par la suite, et qui permet de générer une transition plus douce qu'un simple échelon. Nous prendrons cette fonction croissante à valeurs dans $[0,1]$. Le lieu de transition est tr et vérifie:

$$A_{\mathrm{tr},k}(tr) = 0{,}5 \quad \text{et} \quad \frac{\mathrm{d}A_{\mathrm{tr},k}}{\mathrm{d}x}(tr) = k$$

Comme exemple de fonction A, on trouve la fonction sigmoïde:

$$A_{\mathrm{tr},k}(x) = \frac{1}{1 + e^{-4k(x-\mathrm{tr})}}$$

ou bien la fonction arc-tangente normalisée sur $[0,1]$:

$$A_{\mathrm{tr},k}(x) = \frac{1}{\pi}\arctan(k\pi(x - tr)) + 0{,}5$$

Nous avons représenté ces deux courbes sur la figure 4.11. On peut remarquer que, pour une même rapidité de transition k fixée, la fonction sigmoïde atteint plus -trop- vite la saturation. Ainsi, l'utilisation de celle-ci ne permettra pas de faire la distinction entre des mesures qui se situent légèrement au delà de la transition et celles qui en sont très éloignées. C'est pourquoi nous lui préférerons la fonction arctangente, dont le comportement

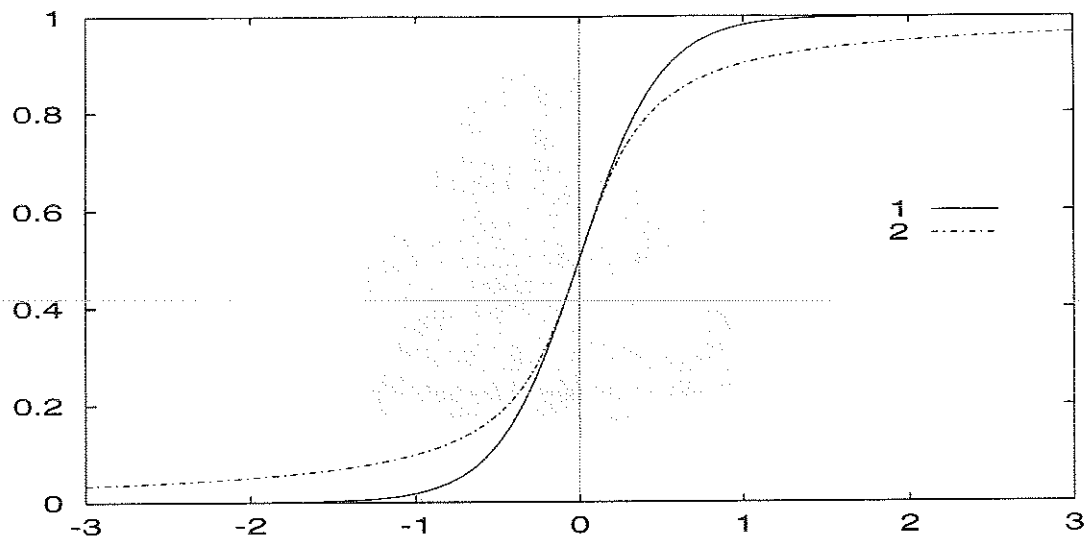


FIG. 4.11 - *Différentes fonctions A (k=1): courbe 1, la fonction sigmoïde, et courbe 2, la fonction arctangente normalisée sur* $[0, 1]$.

asymptotique est plus doux.

Si notre but est de détecter une non-conformité du mouvement résiduel supérieure à une valeur δ, la prise en compte des inégalités entre l'observation o_s et les bornes l_s et L_s (voir la figure 4.7 par exemple) nous conduit à définir un potentiel de la forme (rappelons ici que c signifie conforme, et nc non-conforme):

$$V_1(o_s, d_s) = \begin{cases} \alpha_c \times F_s \times A_{l_s,k_c}(o_s) & \text{si} \quad d_s = c \\ \alpha_{nc} \times F_s \times (1 - A_{L_s,k_{nc}}(o_s)) & \text{si} \quad d_s = nc \end{cases} \tag{4.24}$$

dans lequel:

- α_c et α_{nc} sont les facteurs d'amplitude maximale des potentiels. On prend α_{nc} légèrement supérieur à α_c.
- k_c et k_{nc} sont les facteurs permettant de régler la rapidité de transition autour des bornes minimales et maximales.
- $F_s = F(G_{moy,s}) = \max\left(A_{G,k_a}((G_{moy,s})), \mathrm{At}_{max}\right)$ est un facteur qui permet d'atténuer l'amplitude du potentiel en fonction de la présence ou non de gradient spatial au site s (voir la définition de G_{moy} donnée par la relation (4.20)). Comme nous l'avons déjà fait remarquer (voir page 95), les observations sont peu fiables dans les régions uniformes. En diminuant l'amplitude de l'énergie dans ces régions, on restreint l'information apportée par les observations; *a contrario*, la contribution relative de la régularisation deviendra alors plus importante. Nous définirons les régions uniformes

comme étant celles dans lesquelles G_{moy} est inférieur au paramètre G, paramètre qu'il nous faudra choisir. Le paramètre At_{max} correspond à l'atténuation maximale que l'on s'autorise, ceci pour éviter qu'un site ne porte plus aucune information de mouvement.

Sur la figure 4.12 sont représentées les deux courbes de potentiel suivant la valeur de l'étiquette. La courbe associée à l'étiquette non-conforme est celle dont l'énergie locale diminue lorsque l'observation croît. La forme de cette courbe est choisie de telle sorte que, tant que l'observation se situe au-delà de la borne supérieure L_s, le potentiel est très faible. En revanche, quand l'observation devient inférieure à L_s, nous ne sommes plus sûrs que le pixel a vraiment un mouvement non-conforme, et donc le potentiel augmente indiquant que le choix de l'étiquette est moins approprié. Dans le cas de l'étiquette conforme, le comportement est inversé et prend en compte la borne inférieure de notre information.

On peut remarquer que les potentiels dont nous nous servons sont bornés, notamment pour les fortes observations, ce qui est à rapprocher du comportement des estimateurs robustes dont le but est de modérer voire d'annuler l'influence des observations très erronées ou appartenant à une autre distribution. Ici, la saturation du potentiel d'attache aux données aura pour effet d'éviter qu'une information trop "bruitée" en un site n'impose une mauvaise étiquette si tous les voisins ont une étiquette différente, ou d'un autre point de vue, n'oblige pas à renforcer "artificiellement" le terme de régularisation, ce qui aurait des effets négatifs par ailleurs. Notons ici que la saturation affectera également les observations importantes dues à la présence réelle d'un mouvement de grande amplitude. Ceci n'est pas gênant, dans la mesure où d'une part, dans une telle situation, nous nous attendons à ce que les observations voisines confirment la présence de mouvement, et d'autre part nous nous intéressons uniquement à la détection de mouvement, et non pas à la quantification de celui-ci.

La figure 4.12 représente les courbes de potentiel en un site dont la borne λ'_{min} est maximale, c'est-à-dire que le site correspond en fait à une isophote en forme de coin (voir dessin 4.8). On peut constater que la séparation entre mouvement conforme et non-conforme est très nette. L'écart entre la valeur du potentiel si le site est étiqueté comme conforme et celle s'il est étiqueté comme non-conforme est partout important, à l'exception de la zone de transition (réduite dans ce cas) entre l_s et L_s. Dans le cas d'un contour (ou d'une isophote) rectiligne, cet intervalle d'incertitude est maximal puisque, comme nous l'avons indiqué, dans cette situation, des observations faibles ne caractérisent pas nécessairement l'absence de mouvement mais peuvent être dues à un glissement du contour sur lui-même. Cette incertitude se traduit par des courbes de potentiel proches l'une de l'autre, comme l'indique la figure 4.13. La figure 4.14 met encore plus en évidence cet effet. Sur celle-ci, on voit très bien que pour le cas d'un site situé sur un contour rectiligne, la différence entre les deux potentiels suivant l'étiquette proposée est très faible sur une plage importante des observations. Ainsi, pour un tel site, le contexte jouera un rôle primordial si l'observation se trouve sur cette plage. Autrement dit pour une même

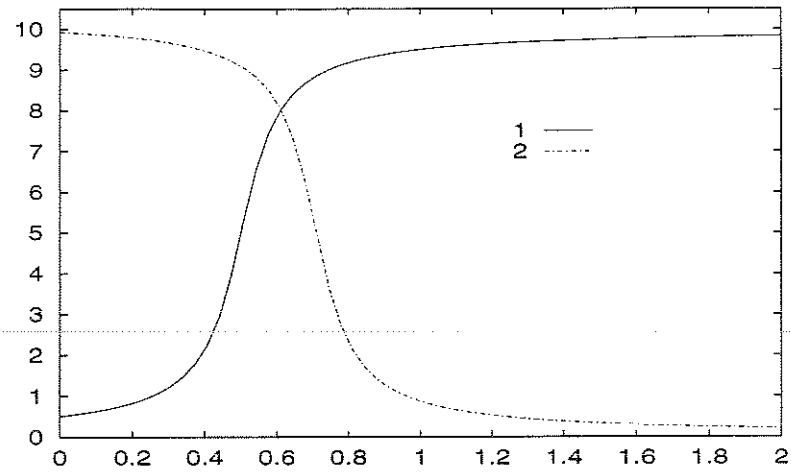


FIG. 4.12 - *Potentiel $V_1(o_s, d_s)$ d'attache aux données en fonction de l'étiquette (courbe 1 pour l'étiquette conforme, courbe 2 pour l'étiquette non-conforme). Cas où λ'_{min} est maximale et vaut $\delta/2$ (correspond à une isophote en "coin" dans l'image).*
Paramètres: $\delta = 1$, $l_s = 1/2$, $L_s = \sqrt{2}/2$, $k_c = k_{nc} = 4$

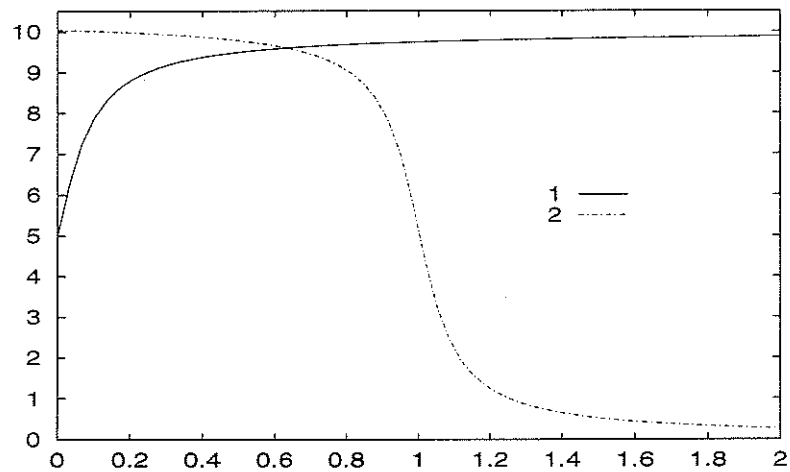


FIG. 4.13 - *Potentiel $V_1(o_s, d_s)$ d'attache aux données en fonction de l'étiquette (courbe 1 pour l'étiquette conforme, courne 2 pour l'étiquette non-conforme). Cas où λ'_{min} est minimale et vaut 0 (correspond à une isophote rectiligne dans l'image).*
Paramètres: $\delta = 1$, $l_s = 0$, $L_s = 1$, $k_c = k_{nc} = 4$

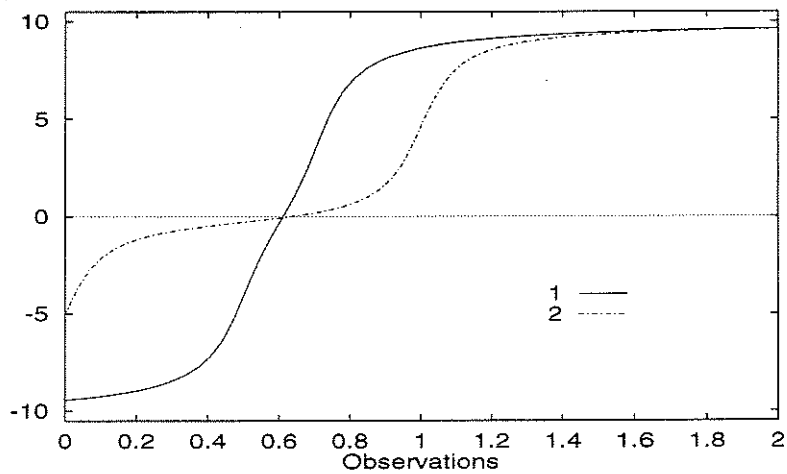


FIG. 4.14 - *Différence des potentiels correspondant aux étiquettes conforme et non-conforme, dans le cas d'une isophote en "coin", courbe 1 (cf figure 4.12) et dans le cas d'une isophote rectiligne, courbe 2 (cf figure 4.13).*

observation, faible (disons 0,3 par exemple), il sera plus facile pour le contexte d'imposer une étiquette non-conforme à un site situé sur une isophote rectiligne qu'à un site localisé sur une isophote en "coin". Les potentiels modélisés reflètent donc bien la confiance que l'on peut accorder à l'observation pour indiquer la présence de mouvement, en fonction de la distribution locale des gradients spatiaux de l'intensité.

4.2.4 Définition du terme de régularisation U_2

Ayant supposé que le champ D des étiquettes est markovien relativement à un système de voisinage d'ordre 2 (8-voisinage), l'énergie U_2 se décompose en une somme de potentiels sur les cliques engendrées par ce voisinage:

$$U_2(d) = \sum_{c \in \mathcal{C}} V_c(d)$$

Ces potentiels permettent de modéliser des propriétés spatiales sur le champ des étiquettes. Dans le cas de la détection du mouvement, nous souhaitons obtenir des masques compacts pour les régions de mouvement conforme et non-conforme. Pour cela, nous recourrons à des simples potentiels à niveaux, définis sur les seules cliques binaires, et indépendant de l'orientation de ces dernières, suivant le modèle:

$$U_2(d) = \sum_{\{s,u\} \in \mathcal{C}} V_2(d_s, d_u) \tag{4.25}$$

avec

$$V_2(d_s, d_u) = \begin{cases} -\beta_c & \text{si } d_u = d_s = \text{ conforme} \\ -\beta_{nc} & \text{si } d_u = d_s = \text{ non-conforme} \\ +\beta_d & \text{si } d_u \neq d_s \end{cases} \tag{4.26}$$

β_d est un coût à payer pour avoir des voisins avec des étiquettes différentes, et β_c et β_{nc} sont des valeurs du potentiel qui permettent de favoriser la sélection de, respectivement, l'étiquette conforme et non-conforme. De manière générale, nous fixerons $\beta_c = 0$, et nous prendrons $\beta_{nc} \ll \beta_d$. L'introduction du paramètre supplémentaire β_{nc} sera profitable pour favoriser l'étiquette non-conforme dans des zones uniformes de l'image correspondant à à l'intérieur des projections d'objets mobiles.

4.3 Détection de mouvement dans une séquence

Il est bien connu en estimation de mouvement, en reconstruction 3D, en segmentation du mouvement, en reconnaisance d'objets ou de mouvements dans une séquence, etc..., que l'information que l'on peut extraire de deux images est généralement ambiguë, celle-ci pouvant résulter de plusieurs interprétations différentes de la scène et du mouvement 3D. Ceci rend les algorithmes basés uniquement sur deux images souvent instables. L'accumulation d'informations au cours du temps permet généralement de résorber cette instabilité

[AD93]. Dans le cas de la détection de mouvement, les ambiguïtés se situent surtout dans les régions peu contrastées. Pour les atténuer, nous allons donc définir de nouvelles observations qui viendront compléter la précédente (c'est à dire o_s^t donné par la formule (4.19)) et dont le but est d'asseoir l'estimation sur une information plus importante.
Le rôle de l'aspect temporel est double:

- d'une part il doit assurer la cohérence des masques de détection à différents instants;
- d'autre part, il doit filtrer les observations pour éliminer les différents bruits issus des mesures: bruit et artefacts d'acquisition, erreurs sur le calcul des gradients spatiaux, interpolation, variations d'illumination.

Examinons maintenant ces deux points.

4.3.1 Utilisation de la carte de détection estimée à l'instant précédent

Dans le monde tridimensionnel, le mouvement des objets est généralement continu. Dès lors, les projections dans le plan image d'un même objet à des instants successifs se recouvrent partiellement. De plus, même si ces objets restent constamment en mouvement au cours du temps, il peut se produire pour certaines configurations particulières de leur déplacement et de l'orientation de la caméra (déplacement le long de l'axe optique) que les projections passent par une phase de mouvement très faible, voire nul en certains points. Il est donc nécessaire que les observations conservent en quelque sorte la mémoire des objets en mouvement aux instants précédents. Nous ferons donc l'hypothèse que les cartes de détection varient peu au cours du temps, et nous emploierons donc la carte de détection estimée à l'instant précédent. Il existe plusieurs façons d'exploiter cette hypothèse:

- une première repose sur le fait que, pour des raisons calculatoires évidentes, les algorithmes d'optimisation utilisés sont déterministes et convergent généralement vers un minimum local de la fonction d'énergie proche de l'initialisation. Cette dernière est donc importante. Comme on s'attend à ce que les cartes de détection n'évoluent que lentement au cours du temps, la carte obtenue à l'instant précédent peut donc servir d'initialisation dans les algorithmes de relaxation.
- dans une seconde façon, les cartes de détection sont considérées comme étant des processus temporels. On peut par exemple considérer que la détection en chaque site est indépendante des autres sites de l'image, et utiliser alors des algorithmes de filtrage récursifs, comme l'algorithme de Kalman, pour estimer les cartes de détection [KvG90]. On peut également considérer les cartes de détection pour un bloc d'images comme un champ de Markov spatio-temporel. Cependant, dans ce cas, l'espace d'état Ω du champ Markovien devient gigantesque. La minimisation de l'énergie globale sera donc très complexe, très coûteuse en temps calcul, et requerra

un espace mémoire important. De plus cette modélisation introduit un délai dans l'obtention des cartes de détection.

- enfin, une troisième façon d'exploiter l'hypothèse consiste à considérér les champs d'étiquettes $(d^1, \ldots, d^{t-1}, d^t, \ldots)$ comme une chaîne de Markov[5]. On recherche alors la carte de détection d^t vérifiant:

$$\max p(d^t|\tilde{d}^{t-1}, \ldots, \tilde{d}^1, o^t) = \max p(d^t|\tilde{d}^{t-1}, o^t) \tag{4.27}$$

C'est cette troisième approche que nous avons retenue. Nous définissons alors un nouveau terme d'énergie U_3 de même type que U_2:

$$U_3(d^t, \tilde{d}^{t-1}) = \sum_{s \in S} V_3(d_s^t, \tilde{d}^{t-1}) \quad \text{avec } V_3(d_s^t, \tilde{d}^{t-1}) = \sum_{u \in \{s\} \cup \mathcal{G}_s} W_3(d_s^t, \tilde{d}_u^{t-1}) \tag{4.28}$$

où $\mathcal{G}_s$ désigne le voisinage du site s et W_3 est un potentiel similaire à V_2:

$$W_3(d_s^t, \tilde{d}_u^{t-1}) = \begin{cases} 0 & \text{si } \tilde{d}_u^{t-1} = d_s^t = c \\ -\beta_{nct} & \text{si } \tilde{d}_u^{t-1} = d_s^t = nc \\ +\beta_{dt} & \text{si } \tilde{d}_u^{t-1} \neq d_s^t \end{cases} \tag{4.29}$$

Si nous considérions la caméra comme étant fixe, $\tilde{d}^{t-1}$ correspondrait tout simplement à la carte de détection estimée à l'instant précédent. L'équivalent, dans le cas d'un capteur mobile, est de projeter la carte obtenue à l'instant précédent dans le sens du mouvement dominant, soit:

$$\tilde{d}^{t-1} = \text{comp}_{\Theta_t^{t-1}}(\hat{d}^{t-1}) \tag{4.30}$$

Dans la mesure où les déplacements ne sont pas nécessairement entiers, il n'est pas possible de déterminer avec exactitude la carte $\tilde{d}^{t-1}$, et une indétermination subsistera donc aux frontières. Bien que cette indétermination n'ait de conséquence que localement, pour les sites proches de ces frontières, nous avons quand même essayé de la réduire en considérant dans la formule (4.28) que la "mémoire" d'un site n'est pas constituée uniquement de l'étiquette à ce même site dans la carte de détection $\tilde{d}^{t-1}$, mais également des étiquettes de ses sites voisins dans $\tilde{d}^{t-1}$, comme l'indique la figure 4.15.

De manière générale, nous choisirons une valeur de β_{dt} de l'ordre de $\beta_d/9$. Par ailleurs, le choix d'une valeur non nulle pour le terme β_{nct} facilitera le suivi des régions non-conformes, notamment lorsque celles-ci seront petites.

Nous aurions pu considérer une mémoire plus longue, en utilisant plusieurs cartes de détection passées, mais cela aurait été au détriment de la précision des cartes obtenues. Déjà avec l'utilisation d'une seule carte passée, nous avons pu constater que l'effet "booléen" de cette contrainte supplémentaire pouvait avoir des conséquences ennuyeuses. Par exemple, lorsqu'un objet mobile se déplace sur un fond uniforme, ce dernier peut conserver pendant un temps important l'étiquette mobile après le passage de l'objet. Ceci est

5. On a donc un champ de Markov spatio-temporel très particulier.

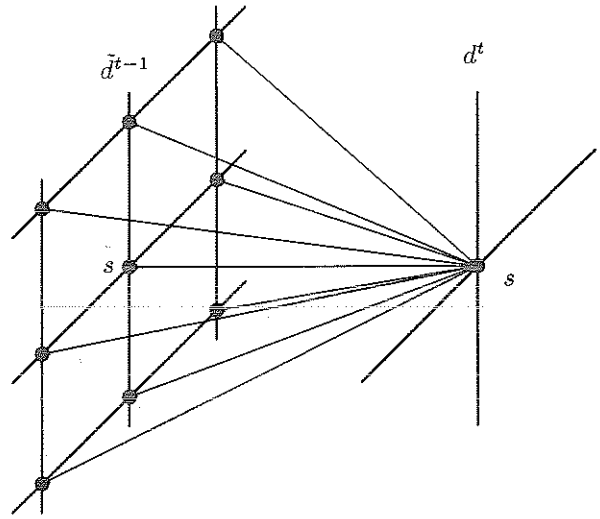


FIG. 4.15 - *La "mémoire" d'un site de la grille à l'instant t est constituée des neufs "voisins" de ce site dans la carte de détection à l'instant* $t-1$, *c-à-d* $\tilde{d}^{t-1}$.

compréhensible. En effet, dans la mesure où ce fond se singularise par une absence d'informations, l'algorithme a recours au passé pour désigner l'état courant de cette zone. On pourrait d'ailleurs imaginer que cette zone corresponde en fait à une partie uniforme postérieure de l'objet en mouvement. C'est parce que ce dernier cas est beaucoup moins fréquent que celui d'un fond uniforme statique (du ciel par exemple) que "l'effet booléen" est indésirable. L'observation suplémentaire que nous allons définir maintenant est moins "radicale" que $\tilde{d}^{t-1}$ et permet de prendre en compte plus souplement un passé plus lointain.

4.3.2 Observations de mouvement filtrées temporellement

La "réaction en chaîne" possible (d'une image à l'autre, en l'absence d'information, les étiquettes auraient tendance à s'auto-reproduire) décrite dans le paragraphe précédent peut s'interpréter de la manière suivante. L'utilisation de la carte de détection à $t-1$ induit implicitement la prise en compte:

1. des observations à $t-1$;
2. de la régularisation spatiale à $t-1$;
3. du masque de détection à $t-2$; et donc des observations et de la régularisation spatiale à $t-2$, de la carte de détection à $t-3$, etc....

Il nous faudrait donc au minimum s'affranchir du troisième point. Cela pourrait se traduire par l'emploi des cartes de détections aux instants $t-1$, $t-2$,... $t-T$ obtenues en ne minimisant que l'énergie formée de U_1 et U_2 (i.e. détection entre deux images). C'est une

possibilité intéressante mais coûteuse en temps de calcul. Il faudrait accomplir à chaque instant, à la fois la détection entre deux images (pour fournir les cartes aux détections futures), et la détection complète utilisant le passé. Une méthode plus simple consiste à n'utiliser que les observations de mouvement aux instants précédents. C'est ce que nous avons fait. Cependant, nous avons tout de même conservé la fonction d'énergie U_3 faisant appel à la carte de détection obtenue à l'instant précédent. En effet, grâce à l'introduction des observations temporelles, d'une part l'effet "booléen" que nous avons décrit est adouci, et d'autre part, il n'est plus nécessaire d'accorder beaucoup d'importance au potentiel U_3 dans l'énergie globale pour prendre en compte le passé.

Si nous désignons par o^{t-q} les champs d'observations précédant l'instant t, il nous faut les projeter dans le sens du mouvement induit par la caméra pour qu'ils se rapportent à l'instant t, soit:

$$\tilde{o}^{t-q} = \mathrm{comp}_{\Theta_t^{t-q}}(o^{t-q}) \ , \ q \in \{0, \ldots, T\} \tag{4.31}$$

Par souci de simplicité et d'homogénéité des notations, nous noterons également o^t par $\tilde{o}^t$.

Si l'on suppose que ces observations sont indépendantes, on peut alors remplacer l'énergie U_1 par U_1':

$$U_1'(\tilde{o}^{t-q}, q \in \{0, \ldots, T\};\ d^t) = \sum_{q=0}^{T} U_1(\tilde{o}^{t-q}, d^t) \tag{4.32}$$

Cependant, outre l'indépendance des observations, cette relation suppose aussi leur équivalence du point de vue de l'information fournie, ce qui n'est bien entendu pas le cas compte-tenu de l'évolution des zones mobiles. Il est préférable de choisir:

$$U_1'(\tilde{o}^{t-q}, q \in \{0, \ldots, T\};\ d^t) = \sum_{q=0}^{T} \gamma^q \times U_1(\tilde{o}^{t-q}, d^t) \tag{4.33}$$

où γ représente un facteur d'amortissement compris entre 0 et 1. Bien entendu, plus ce coefficient est faible, moins on a recours aux observations précédentes, et vice-versa. Enfin, soulignons que si l'amplitude maximum des potentiels de U_1 est α_{nc}, celle de U_1' vaut:

$$\alpha_{nc} \times \frac{1 - \gamma^{T+1}}{1 - \gamma} \qquad (\text{ pour } \gamma \neq 1) \tag{4.34}$$

Par conséquent, pour que la contribution relative de l'énergie liée aux observations de mouvement vis-à-vis des termes de régularisation reste indépendante du filtrage temporel effectué (caractérisé par T et surtout γ), nous choisirons en définitive:

$$U_1'(\tilde{o}^{t-q}, q \in \{0, \ldots, T\};\ d^t) = \frac{1 - \gamma}{1 - \gamma^{T+1}} \sum_{q=0}^{T} \gamma^q \times U_1(\tilde{o}^{t-q}, d^t) \tag{4.35}$$

qui peut également s'écrire sous la forme:

$$U_1'(\tilde{o}^{t-q}, q \in \{0, \ldots, T\};\ d^t) = \sum_{s \in S} V_1'(\tilde{o}_s^{t-q}, q \in \{0, \ldots, T\};\ d_s^t) \tag{4.36}$$

avec:

$$V_1'(\tilde{o}_s^{t-q}, q \in \{0,\ldots,T\};\, d_s^t) = \frac{1-\gamma}{1-\gamma^{T+1}} \sum_{q=0}^{T} \gamma^q \times V_1(\tilde{o}_s^{t-q}, d_s^t) \qquad (4.37)$$

où V_1 est défini par la formule (4.24).

Dans la mesure où l'algorithme proposé par Irani, Rousso et Peleg [IRP91, IRP92] possède certaines similarités avec le nôtre, il nous semble opportun de le décrire dans ses grandes lignes et de souligner les principales différences entre les deux méthodes.

4.3.3 Comparaison avec l'algorithme de Irani, Rousso et Peleg

Le principe de leur algorithme repose sur l'utilisation d'une image de référence, I_{av}^t, comme dans le cas statique [BAD93], qui est mise à jour de la façon suivante:

$$I_{av}^{t+1} = (1-w) \times I^{t+1} + w \times \mathrm{comp}_{\Theta_{t+1}^t}(I_{av}^t) \;,\; w \in [0,1] \quad \text{et} \quad I_{av}^0 = I^0 \qquad (4.38)$$

La seconde différence importante avec notre approche est que le mouvement employé pour effectuer la compensation n'est pas estimé entre I^t et I^{t+1}, mais entre I_{av}^t et I^{t+1}. L'intérêt est le suivant. Si le mouvement estimé constitue une bonne approximation du mouvement de la région suivie (région de mouvement conforme), cette région restera nette et précise dans l'image de référence, alors qu'au contraire, dans les régions mal compensées, l'image de référence aura tendance à se brouiller. Par conséquent, d'une part, comme les gradients spatiaux de l'intensité à l'extérieur de la région suivie seront substantiellement réduits, l'estimation du mouvement correspondant à la région suivie en sera facilitée, et d'autre part, il sera plus aisé d'exécuter la détection de mouvement entre l'image de référence et la nouvelle image qu'entre deux images successives. Quant à la détection proprement dite, elle résulte simplement d'un seuillage sur une image de données provenant de l'addition d'observations identiques aux nôtres, calculées entre I_{av}^t et I^{t+1} à plusieurs résolutions.

Nous avons implémenté leur méthode, hormis la partie "détection" où nous avons conservé notre algorithme de détection de mouvement entre deux images, ce qui apporte un aspect régularisant substantiel que ne possède pas la méthode originelle de IRANI *et al.* . Nous avons pu constater qu'il n'y a pratiquement aucune différence avec notre algorithme de détection n'utilisant que deux images (qui correspond à $w = 0$) lorsque les valeurs de w restent inférieures à 0,3. Au-delà de cette valeur, les résultats s'améliorent un peu mais sont nettement moins bons que ceux fournis par notre algorithme complet. Nous avons également pu constater que:

1. d'une part l'algorithme est sensible aux variations d'illumination, le moyennage retardant la prise en compte de celles-ci dans l'image de référence, ce qui accentue le différentiel avec les nouvelles images. Ceci est un défaut général des méthodes avec image de référence [BAD93].

2. d'autre part, le recalage pour les régions conformes doit être quasiment parfait. Dans le cas contraire, les défauts observés vont de la fausse alarme (régions conformes détectées comme non-conformes) à l'échec complet de l'algorithme (séquence DAMIER, ROND-POINT).

Ce dernier point nécessite quelques explications. Il est aisé de comprendre que si le recalage n'est pas exact, même les régions conformes auront tendance à devenir floues dans l'image de référence rendant l'estimation du mouvement difficile, voire impossible. En fait, l'hypothèse essentielle qui doit être faite pour éviter le brouillage de l'image de référence est que le recalage doit *rester précis* durant *toute la durée de l'intégration temporelle.* Par exemple, si l'on suppose qu'une image ne joue un rôle important dans le filtrage temporel que si son coefficient de pondération dans l'image de référence est supérieur à 0,2 fois celui de l'image la plus représentée dans l'image de référence (c-à-d. l'image courante, qui a un coefficient de $1 - w$), cette durée temporelle est de 3 à 4 images pour $w = 0{,}5$, 5 à 6 pour $w = 0{,}7$.

A contrario, notre algorithme n'implique d'autres contraintes que d'avoir un recalage *plus précis* que l'amplitude des mouvements que l'on souhaite détecter, et ceci, *entre deux images* uniquement, ce qui est nettement plus faible comme hypothèse. C'est pourquoi avec notre méthode, nous pourrons tolérer des erreurs de recalage de 1 à 2 pixels, dans la mesure où les objets à détecter ont un mouvement plus rapide (pour des erreurs de recalage plus importantes, il est préférable de passer à une résolution plus faible pour que les mesures sur lesquelles est basé notre algorithme, et notamment la comparaison avec les bornes, restent valides).

4.4 Aspects calculatoires

Nous sommes donc amenés à minimiser une fonction d'énergie donnée par:

$$U(o^t,\, d^t) = U_1(o^t,\, d^t) + U_2(d^t) \tag{4.39}$$

dans la configuration "deux images", et par:

$$U(\tilde{o}^{t-q}, q \in \{0,\ldots,T\};\, \tilde{d}^{t-1};\, d^t) = U_1'(\tilde{o}^{t-q}, q \in \{0,\ldots,T\};\, d^t) + U_2(d^t) + U_3(d^t;\, \tilde{d}^{t-1}) \tag{4.40}$$

pour le traitement d'une séquence, où les différents potentiels sont fixés par les expressions (4.23), (4.25), (4.28), et (4.35). Deux questions importantes se posent alors: quel algorithme utiliser pour réaliser la minimisation, et quel jeu de paramètres employer.

4.4.1 Minimisation de la fonction d'énergie

En annexe A, nous présentons plusieurs méthodes de minimisation d'une fonction d'énergie associée à un modèle markovien. Dans notre cas, nous avons retenu l'approche

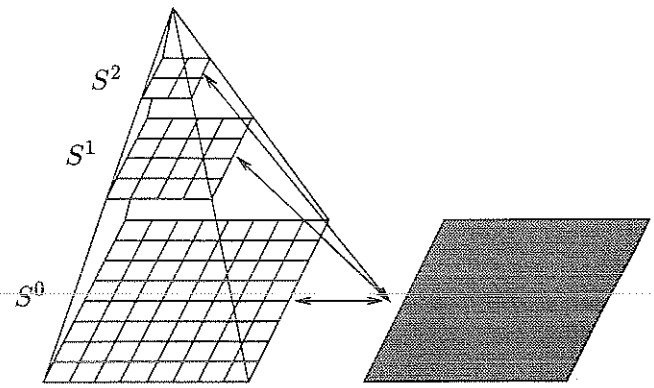


FIG. 4.16 - *Structure multigrilles des étiquettes, et relation avec les observations.*

multiéchelle proposée par Pérez *et al.* [PH93, PHB94], avec à chaque échelle une minimisation basée sur l'algorithme "Highest Confidence First" (HCF) de Chou et Brown [CR87, CB90]. Cette méthode multiéchelle, quoique déterministe, permet d'obtenir des solutions proches de celles fournies par les algorithmes de minimisation stochastiques, en un temps pourtant généralement inférieur à ceux obtenus avec les algorithmes déterministes.

La méthode consiste en l'exploration de sous-espaces de configurations correspondant à des cartes de détection de plus en plus fines. C'est à dire, au niveau de "résolution" i, les configurations sont postulées constantes sur des blocs $B_1^i, B_2^i, \ldots$ de taille $2^i \times 2^i$ partitionnant l'ensemble S des sites de l'image. Ceci est équivalent à l'estimation descendante d'une pyramide d'étiquettes, connaissant les seules observations à la résolution la plus fine (cf. figure 4.16). En effet, une configuration d^t de Ω^i (i.e. constante sur les blocs B_k^i) pourra être assimilée à une configuration $d^{t,i} = \{d_k^{t,i},\ k \in S^i\}$, définie sur la grille réduite d'un facteur 2^i, S^i, par:

$$\forall s \in B_k^i\,,\ \ d_s^t = d_k^{t,i} \tag{4.43}$$

On en déduit alors naturellement une énergie U^i sur l'espace de ces configurations réduites, qui est entièrement déterminée par la connaissance de l'énergie U au niveau le plus fin [PH93]. Nous avons formulé cette énergie dans notre cas et les expressions résultantes sont présentées dans le tableau (4.6). Par exemple, $U_1^{t,i}$ s'obtient en transformant la somme sur les sites s de S des potentiels V_1' par une sommation de potentiels $W_1^{t,i}$ sur les sites k de S^i, où $W_1^{t,i}$ est naturellement constitué de l'addition des V_1' sur le bloc B_k^i rattaché à k. Quand à V_2^i, il se décompose en la somme de potentiels V_c^i définis sur les cliques au niveau i. En annexe, on montre que le 8-voisinage sur S induit un 8-voisinage sur S^i. Dans notre cas, les cliques au niveau i ne sont formées que de un ou deux éléments. Pour une clique unaire $c = \{k\}$, le potentiel V_c^i est égal au produit du potentiel V_2 par le nombre p^i de cliques de $\mathcal{C}$ qui sont incluses dans le bloc B_k^i. Dans le cas des cliques binaires $\{k_1, k_2\}$,

$$\begin{aligned} U^i(\tilde{o}^{t-q}, q \in \{0,\dots,T\};\, \tilde{d}^{t-1};\, d^{t,i}) &= U_1'^i(\tilde{o}^{t-q}, q \in \{0,\dots,T\};\, d^{t,i}) \\ &+ U_2^i(d^{t,i}) \\ &+ U_3^i(\tilde{d}^{t-1};\, d^{t,i}) \end{aligned} \tag{4.41}$$

S^i est la grille support au niveau i.

$d^{t,i} = \{d_k^{t,i},\ k \in S^i\}$ est une réalisation du champ d'étiquettes, à l'instant t et à l'échelle i.

$\mathcal{C}^i$ est l'ensemble des cliques au niveau i, qui se divisent en l'ensemble des singletons $\{k\}$ de S^i, et les ensembles des cliques binaires verticales $\mathcal{C}_v^i$, horizontales $\mathcal{C}_h^i$ et diagonales $\mathcal{C}_d^i$.

B_k^i est le bloc de taille $2^i \times 2^i$ naturellement associé au site k de S^i.

$V_1'(), V_2(), V_3()$ sont donnés par les formules (4.35), (4.26), (4.28).

- $\begin{cases} U_1'^i(\tilde{o}^{t-q}, q \in \{0,\dots,T\};\, d^{t,i}) = \sum_{k \in S^i} W_1'^i(\tilde{o}^{t-q}, q \in \{0,\dots,T\};\, d_k^{t,i}) \\ \text{avec} \quad W_1'^i(\tilde{o}^{t-q}, q \in \{0,\dots,T\};\, d_k^{t,i}) = \sum_{s \in B_k^i} V_1'(\tilde{o}_s^{t-q}, q \in \{0,\dots,T\};\, d_k^{t,i}) \end{cases}$

- $\begin{cases} U_2^i(d^{t,i}) = \sum_{c \in \mathcal{C}^i} V_c^i(d^{t,i}) \quad \text{avec} \\ \quad \triangleright \text{ pour } k \in S^i, \quad V_{\{k\}}^i(d^{t,i}) = \begin{cases} -p^i \times \beta_c & \text{si } d_k^{t,i} = \text{ conforme} \\ -p^i \times \beta_{nc} & \text{si } d_k^{t,i} = \text{ non-conforme} \end{cases} \\ \quad \triangleright \text{ pour } \{k_1,k_2\} \in \mathcal{C}_h^i \cup \mathcal{C}_v^i, \quad V_{\{k_1,k_2\}}^i(d^{t,i}) = q_{hv}^i \times V_2(d_{k_1}^{t,i}, d_{k_2}^{t,i}) \\ \quad \triangleright \text{ pour } \{k_1,k_2\} \in \mathcal{C}_d^i, \quad V_{\{k_1,k_2\}}^i(d^{t,i}) = q_d^i \times V_2(d_{k_1}^{t,i}, d_{k_2}^{t,i}) \\ \quad \triangleright \begin{cases} p^i = 2(2^i-1)(2^{i+1}-1) \\ q_{hv}^i = 3 \times 2^i - 2 \\ q_d^i = 1 \end{cases} \end{cases}$

- $\begin{cases} U_3^i(\tilde{d}^{t-1};\, d^{t,i}) = \sum_{k \in S^i} W_3^i(\tilde{d}^{t-1};\, d_k^{t,i}) \\ \text{avec} \quad W_3^i(\tilde{d}^{t-1};\, d_k^{t,i}) = \sum_{s \in B_k^i} V_3(d_k^{t,i}, \tilde{d}_s^{t-1}) \end{cases}$

(4.42)

TAB. 4.6 - *Énergie U^i au niveau i de la pyramide.*

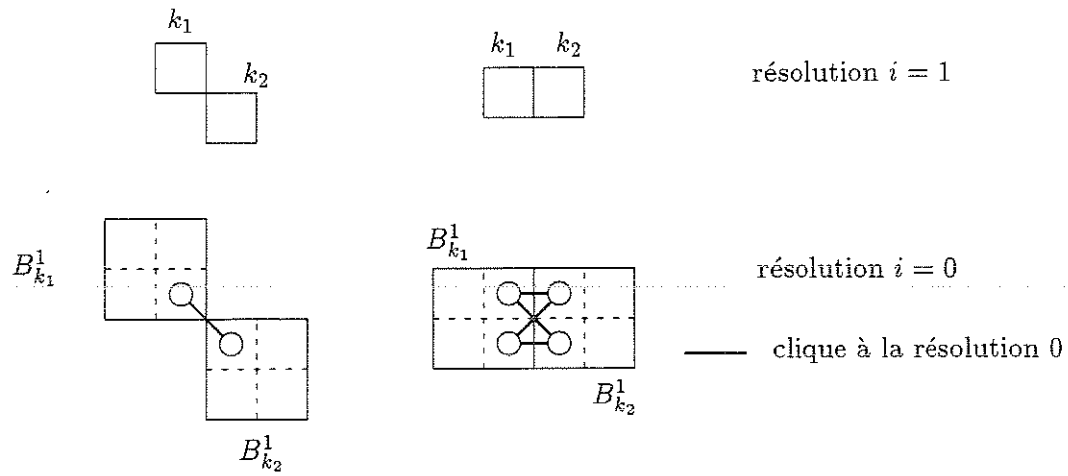


FIG. 4.17 - *Exemple de calcul du nombres de cliques de $\mathcal{C}$ (résolution 0) à cheval sur deux blocs correspondant à une clique de sites voisins à la résolution 1. Si cette clique est horizontale, il y a quatre cliques de $\mathcal{C}$ à cheval sur les deux blocs. Si cette clique est diagonale, il n'y en a qu'une seule.*

V_2 est multiplié par le nombre de cliques de $\mathcal{C}$ à cheval sur les deux blocs $B^i_{k_1}$ et $B^i_{k_2}$. Pour une clique diagonale (au niveau i), il n'y a évidemment qu'une seule clique de $\mathcal{C}$ qui permet de joindre les deux blocs, alors que ce nombre est beaucoup plus élevé pour une clique verticale ou horizontale (figure 4.17). L'anisotropie intuitive de voisinage entre des blocs qui ne se "touchent" que par le coin et ceux qui se touchent par le côté se trouve ici mise en évidence et naturellement intégrée au sein du modèle multiéchelle.

Les énergies U^i ainsi définies sont utilisées pour mettre en œuvre l'algorithme d'estimation multiéchelle avec une stratégie descendante classique (figure 4.18). Nous noterons L_{det} le nombre d'échelles considérées. Au niveau le plus grossier $L_{\text{det}}-1$, nous avons retenu comme carte initiale de l'algorithme celle qui maximise la vraisemblance locale de toutes les observations, c'est à dire la carte obtenue par minimisation (en chaque site de la grille à l'échelle $L_{\text{det}}-1$) de $U_1'^{L_{\text{det}}-1} + U_3^{L_{\text{det}}-1}$. Nous avons également fait des tests en prenant celle qui minimise $U_1'^{L_{\text{det}}-1}$ uniquement (prise en compte des observations de mouvement uniquement), sans noter de différence.

Concernant l'implantation effective de l'algorithme, on peut évoquer les points suivants:

1. À chaque étape, l'algorithme de relaxation implique le calcul de $\Delta U = U(.,c) - U(.,nc)$ en un site. À partir des observations au niveau 0, nous avons précalculé la pyramide des $\Delta W_1'^i$ et ΔW_3^i, pour éviter les calculs redondants de la relaxation à chaque niveau.

2. Calcul de l'énergie U_1': au lieu de compenser à chaque instant le mouvement entre $t-q$ et t pour obtenir les cartes d'observation $\tilde{o}^{t-q}$ à partir des cartes o^{t-q}, puis

- $i \leftarrow L_{\text{det}} - 1$
- $d^{t,i}_{\text{init}} \leftarrow$ argmin $(U_1'^i + U_3^i)$ (seuillage en chaque site de l'échelle i)
- Tant que $i \geq 0$
  - $\hat{d}^{t,i} \leftarrow$ HCF $(d^{t,i}_{\text{init}}, U^i)$
  - si $i \neq 0$
    - $d^{t,i-1}_{\text{init}} \leftarrow$ projection de $\hat{d}^{t,i}$
    - $i \leftarrow i - 1$

FIG. 4.18 - *Algorithme multiéchelle avec stratégie descendante. La minimisation à chaque niveau est effectuée avec l'algorithme "Highest Confidence First" de Chou et Brown. La projection correspond simplement à la recopie d'une étiquette au niveau i sur ses quatre descendants au niveau* $i - 1$.

de calculer et moyenner les expressions $U_1(\tilde{o}^{t-q}, d^t)$, nous avons préféré adopter le schéma récursif suivant:

$$\Delta U_1'^t = (1 - p_p) \times \Delta U_1(o^t) + p_p \times \widetilde{\Delta E}^{t-1}_{\text{ref}} \tag{4.44}$$

$$\Delta E^t_{\text{ref}} = \frac{1}{y_t}\left[p_n \times \Delta U_1(o^t) + (1 - p_n) \times y_{t-1} \times \widetilde{\Delta E}^{t-1}_{\text{ref}}\right] \tag{4.45}$$

$$\text{avec } y_t = 1 - (1 - p_n)^{t+2} \tag{4.46}$$

où p_p et p_n sont deux coefficients choisis comme nous le verrons ci-après, et y_t sert à prendre en compte l'effet de bord temporel dans les premières images. ΔE^t_{ref} (le Δ joue le même rôle que dans le point 1) représente ainsi l'image des différences d'énergies associées aux observations, moyennées temporellement. Par ce biais, lorsque l'on projette dans le sens du mouvement $\Delta E^{t-1}_{\text{ref}}$ entre $t-1$ et t pour obtenir $\widetilde{\Delta E}^{t-1}_{\text{ref}}$, toutes les observations se retrouvent compensées, ce qui réduit considérablement les calculs. Nous économisons également ainsi un espace mémoire important: d'une image à l'autre, seule ΔE^t_{ref} a besoin d'être conservée, au lieu des $T - 1$ cartes o^{t-q}. En revanche, cette méthode ne nous permet de prendre en compte que les deux cas suivants:

- $T = \infty$, en choisissant $p_p = 1 - p_n = \gamma$;
- $T = 2$, avec $p_p = \gamma$ et $p_n = 1$;

Les résultats obtenus montrent que ces deux cas sont suffisants.

3. Le calcul des potentiels V_1' implique celui de termes $A_{l,k}(x)$. Actuellement, le calcul de ceux-ci se fait avec la forme analytique (par appel de la fonction arc-tangente), et représente environ 40% du temps complet de la relaxation. Comme les fonctions $A_{l,k}(x)$ sont choisies par avance, on pourrait, après discrétisation de l'espace des bornes et des observations, les précalculer et les stocker dans des tables. Ceci réduirait le temps de calcul des termes V_1'.

4. Pour réduire le coût calculatoire, il est intéressant de manipuler des images d'entiers plutôt que des termes réels. Pour obtenir une précision suffisante, tous les termes énergétiques ont été multipliés par un facteur K=20.

5. L'utilisation de la méthode multiéchelle permet globalement d'accélérer les calculs. Nous avons consaté un gain de 10% par rapport à la monorésolution. De plus, comme nous le notons en annexe A, l'algorithme HCF permet d'obtenir des temps de calcul relativement faibles. Nous avons mesuré en moyenne un temps calcul de 1, 3 seconde cpu sur un SPARC 10 pour une image 256×256 (le calcul des observations brutes o_s^t et des bornes non-compris), ce temps ne fluctuant que très peu pour les différentes images de la séquence (au plus 10%). En revanche, on peut observer que le gain obtenu en portant notre algorithme sur une machine spécialisée (parallèle) ne serait pas considérable. En effet, par nature, l'algorithme HCF n'est pas "massivement" parallélisable. Enfin, notons que le temps de relaxation est quasiment proportionnel à la taille de l'image, et qu'il peut être réduit si l'on ne souhaite pas avoir une précision de l'ordre du pixel, par exemple en s'arrêtant à un niveau intermédiaire de la pyramide.

Nous donnons dans la partie consacrée aux résultats quelques exemples de temps de calcul obtenus avec notre implémentation.

4.4.2 Choix des paramètres

Le choix des valeurs des paramètres est un point important d'une méthode. Les résultats ne doivent pas exhiber une trop grande sensibilité à leur égard. Dans le cadre d'une application précise, il est généralement possible et préférable de réaliser un apprentissage sur les paramètres. Cet apprentissage peut-être soit empirique [MZ92], soit formalisé comme un véritable problème d'estimation [Cha88, You88]. Cependant, l'analyse *a priori* de l'influence des paramètres doit permettre d'en déduire des valeurs adéquates pour une séquence donnée, avant toute expérimentation. Nous allons tout d'abord rappeler et commenter tous les paramètres que nous avons introduits, puis nous présenterons succintement la méthode des boîtes qualitatives [Aze87, ACJ90] qui est utilisée pour calibrer –en fait encadrer– certains des paramètres pondérant l'influence des différents termes d'énergie.

1. δ, G_m: nous reviendrons plus loin sur le choix de la valeur de δ. Pour le calcul des observations et des bornes, nous devons choisir le paramètre G_m introduit dans la for-

mule 4.19. Rappelons que diminuer ce paramètre accentue les fausses alarmes lorsque la séquence est très bruitée, et que l'augmenter de manière trop importante tend à concentrer les informations de mouvement effectivement utilisées sur les points de contraste d'intensité important des images. Les valeurs retenues vont de 3 pour une séquence avec peu de bruit (INTERVIEW) à 8 pour une séquence très bruitée. En fait, prendre une valeur trop faible peut être très critique, alors que prendre une valeur trop importante réduit simplement la réponse de détection dans les zones uniformes. Rappelons en effet ici que, de par notre modélisation, dans les régions uniformes (caractérisées par un gradient spatial de l'intensité inférieur à G_m en module), l'observation o_s^t et la borne inférieure l_s sont proches de 0 (en l'absence de bruit). Dans ces régions, la différence entre les énergies locales liées aux observations dans le cas où le site est étiqueté comme conforme et dans celui où il est classé non-conforme sera donc généralement: $\Delta V_1 = V_1(o_s^t, c) - V_1(o_s^t, nc) = \alpha_c/2 - \alpha_{nc} \simeq -\alpha_c/2$ (si l'on ne prend pas en compte le facteur d'atténuation). La modélisation privilégie donc l'étiquetage suivant: les régions uniformes ont un mouvement conforme. On peut noter ici que, dans le cas de la détection avec une caméra statique, l'emploi de différences temporelles aboutit à la même constatation [MB94b]. Cet étiquetage *a priori* est globalement valide lorsque l'image entière est compensée par le mouvement dominant dû au capteur, car les grandes plages uniformes correspondent généralement à des zones du fond de l'image, statiques dans la scène, et donc de mouvement apparent conforme. Ce n'est en revanche pas du tout le cas lorsqu'il s'agit de suivre un objet. C'est pour réduire cet effet que nous avons introduit un facteur d'atténuation dans le potentiel V_1.

2. $G, k_a, \mathrm{At}_{max}, \alpha_{nc}, \alpha_c, k_c, k_{nc}, T$ et γ

   - Dans le facteur d'atténuation F_s, dont nous venons de donner une justification, interviennent les paramètres G, k_a, et At_{max}. G permet de séparer les sites de l'image en deux catégories: ceux dont le module du gradient spatial de l'intensité est inférieur à G, pour lesquels on atténue l'énergie liée aux observations, et ceux pour lesquels il n'y a pas d'atténuation. k_a permet de spécifier la rapidité de la transition entre ces deux populations, et est fixé à 1. At_{max} correspond à l'atténuation maximale, ceci pour éviter qu'un site ne porte plus aucune information de mouvement.
     G est plus faible que G_m, car il s'agit simplement de faciliter le basculement des régions complètement uniformes et totalement dépourvues d'information vers l'une ou l'autre des étiquettes, en fonction du contexte. On prend en général G entre 0 et 4, suivant le niveau de bruit dans ces régions uniformes. Nous verrons plus loin comment déterminer la valeur de At_{max}.

   - Pour les fonctions A modélisant le potentiel associé aux observations en l'absence d'atténuation (formule (4.24)), on a choisi, $k_c = k_{nc} = 4$. Par ailleurs,

nous avons décidé de fixer l'amplitude de ce potentiel pour toutes les expérimentations, et de n'utiliser la méthode des boîtes qualitatives que pour avoir un encadrement des paramètres de régularisation. Nous avons donc choisi des valeurs de 10,3 pour α_{nc}, et 10 pour α_c. Le choix de α_{nc} légèrement supérieur à α_c n'est fait ici que pour éviter que, dans les régions peu texturées et bruitées, l'étiquetage initial obtenu sans tenir compte de la régularisation (c-à-d. basé sur U_1' uniquement) ne favorise trop l'étiquette non-conforme.

- Les paramètres T et γ apparaissent dans l'intégration temporelle. Comme nous l'avons dit, notre mise en œuvre ne permet de choisir qu'entre $T = 2$ ou $T = \infty$, et nous retiendrons généralement ce deuxième cas. Quant à γ il dépend de "l'activité temporelle" dans la scène, c'est-à-dire de la rapidité d'évolution des masques des régions non-conformes comparée à la taille de ces mêmes régions. Par exemple, dans le cas où un objet mobile ne se recouvre jamais d'une image sur l'autre, il est impératif, pour le détecter, que les observations courantes comptent pour plus de 50% dans l'énergie U_1', ce qui nécessite que γ soit inférieur à $0,5$. Il en va de même si l'on souhaite éviter les effets de traînée (une région reste étiquetée non-conforme alors que l'objet de mouvement non-conforme est déjà passé) que peut générer l'intégration temporelle. Inversement, une valeur de γ supérieure à $0,5$ aura pour effet de retarder la détection d'une région non-conforme, en repoussant la décision à l'image suivante, qui doit alors apporter une confirmation du mouvement détecté.

3. $\beta_c, \beta_{nc}, \beta_d, \beta_{nct}, \beta_{dt}$: on choisit $\beta_c = 0$, et de même $\beta_{nct} = 0$. On choisira le poids de la régularisation passée comme étant du même ordre que la régularisation spatiale, soit: $\beta_{dt} = \beta_d/9$. Nous allons maintenant fixer des contraintes pour donner des valeurs indicatives pour les paramètres restant β_{nc} et β_d.

La méthode des boîtes qualitatives constitue une approche effective pour encadrer les valeurs des paramètres pondérant l'influence de plusieurs types d'observations ou de termes de régularisation [Aze87, Let93]. Elle correspond en une sorte d'apprentissage de ces paramètres par examen de situations particulières. Si c_p correspond à une telle configuration particulière du champ d dans le voisinage du site s, la méthode revient à comparer les probabilités d'avoir les étiquettes λ_1 et λ_2 pour ce contexte c_p, en exprimant par exemple que λ_1 est au moins "z fois plus probable" que l'étiquette λ_2:

$$p(d_s = \lambda_1 | d_{\mathcal{G}_s} = c_p) \geq z \times p(d_s = \lambda_2 | d_{\mathcal{G}_s} = c_p) \tag{4.47}$$

ce qui, en utilisant les distributions de Gibbs nous amène à:

$$U(d^{s,\lambda_1}) - U(d^{s,\lambda_2}) + \ln(z) \leq 0 \tag{4.48}$$

où d^{s,λ_1} et d^{s,λ_2} sont deux champs d'étiquettes vérifiant:

$$\begin{cases} d_s^{s,\lambda_1} = \lambda_1 \qquad \text{et} \qquad d_s^{s,\lambda_2} = \lambda_2 \\ d_u^{s,\lambda_1} = d_u^{s,\lambda_2} \ \forall u \neq s, \qquad (\text{en particulier:} \quad d_{G_s}^{s,\lambda_1} = d_{G_s}^{s,\lambda_2}) \end{cases} \tag{4.49}$$

Dans la plupart des cas, l'énergie U est linéaire vis à vis des paramètres que l'on souhaite déterminer. Le jeux de contraintes similaires à (4.47) que l'on s'impose permet de déterminer un polyhèdre que l'on désigne sous le nom de boîte qualitative et à l'intérieur duquel doit se trouver le jeu de paramètres à déterminer. Une solution du système d'équations ainsi généré, auquel on ajoute une contrainte supplémentaire pour obtenir l'unicité (par exemple la solution qui minimise la somme des paramètres), s'obtient par programmation linéaire, avec la méthode du simplexe par exemple.

Notons que dans le cas où le nombre d'étiquettes se limite à deux, on peut remplacer la contrainte (4.47) par une contrainte portant directement sur la probabilité d'occurence d'une étiquette étant donné son voisinage:

$$p(d_s = \lambda_1 | d_{G_s} = c_p) \geq p \tag{4.50}$$

qui nous donne:

$$U(d^{s,\lambda_1}) - U(d^{s,\lambda_2}) \leq \ln(\frac{1}{p} - 1) \tag{4.51}$$

Pour déterminer l'encadrement des paramètres de régularisation présenté sur la figure 4.19, nous allons nous fixer quatre contraintes locales, deux pour imposer une régularisation suffisamment importante du champ des étiquettes, et deux autres pour éviter la sur-régularisation de ce champ. Nous utiliserons alors ces encadrements pour choisir nos paramètres en fonction de la qualité de la séquence traitée et de la taille des objets à détecter.

Pour toutes les contraintes introduites, nous ferons les hypothèses suivantes. Dans les configurations locales du site s que nous définirons, les sites voisins de s dans la carte de détection précédente (voir figure 4.15) seront considérés comme étant équivalent à un seul site voisin (spatialement) de s. Le site s aura donc neuf voisins. On supposera également que l'on se trouve en un point qui porte de l'information c-à-d. qu'il n'y aura pas d'atténuation[6] ($F_s = 1$). Rappelons également que nous nous sommes fixés les amplitudes maximales des potentiels liés aux observations: $\alpha_c = 10$ et $\alpha_{nc} = 10{,}3$.

1. cette première contrainte favorise la régularisation des étiquettes conformes. Le voisinage est constitué de neuf sites étiquetés "conformes". L'observation indiquant que le mouvement est "non-conforme", on souhaite tout de même imposer l'étiquette "conforme" avec une probabilité $p = 0{,}6$. D'après (4.51), on a donc:

$$U(d^{s,c}) - U(d^{s,nc}) \ \leq \ \nu \quad \text{avec} \quad \nu = \ln(\frac{1}{p} - 1)$$

6. Les potentiels sont donc ceux définis à la figure 4.12.

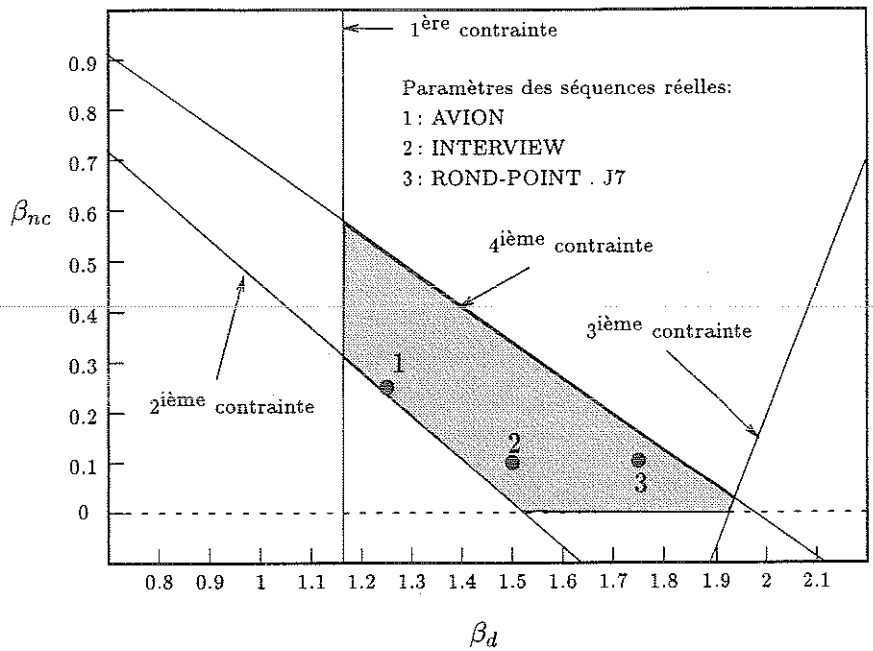


FIG. 4.19 - *Boîte qualitative obtenue pour les paramètres de régularisation β_d et β_{nc}. Les numéros indiquent les paramètres qui ont effectivement été retenus dans les quatre séquences réelles.*

$$\begin{aligned} (\alpha_c + 0 \times \beta_d) - (0 + 9 \times \beta_d) &\leq \nu \ , \quad \text{soit} \\ \frac{\alpha_c - \nu}{9} &\leq \beta_d \end{aligned} \tag{4.52}$$

2. on souhaite régulariser plus facilement les étiquettes non-conformes (dans la mesure où la modélisation favorise l'étiquetage "conforme" dans les zones uniformes). Ainsi, nous souhaitons que la présence de huit sites non-conformes (il y a donc un site conforme) impose (avec la même probabilité) l'étiquette non-conforme à un pixel dont l'observation de mouvement est pourtant nulle, ce qui conduit à:

$$\beta_{nc} \geq -\frac{7}{8}\beta_d + \frac{\alpha_{nc} - \nu}{8} \tag{4.53}$$

3. pour éviter cette fois-ci une sur-régularisation, un voisinage constitué de sept pixels conformes et de deux pixels non-conformes ne doit pas forcer un site à être étiqueté conforme si son observation indique une présence de mouvement résiduel, d'où:

$$\beta_{nc} \geq \frac{5}{2}\beta_d - \frac{\alpha_c + \nu}{2} \tag{4.54}$$

On peut remarquer sur la figure 4.19 que cette contrainte est moins restrictive que la suivante.

4. ce cas est le symétrique du point précédent. Le voisinage est constitué de sept pixels non-conformes et de deux pixels conformes, et l'observation indique que le mouvement résiduel est nul. Cela nous mène à:

$$\beta_{nc} \leq -\frac{5}{7}\beta_d + \frac{\alpha_{nc} + \nu}{7} \tag{4.55}$$

Rappelons également ici que β_{nc} est un terme positif. La valeur de ν pour $p = 0{,}6$ est de $-0{,}405$. On obtient alors la boite qualitative de la figure 4.19, dans laquelle nous avons également indiqué par un numéro les couples de valeurs (β_c, β_{nc}) retenues dans les expériences réelles.

Il nous faut maintenant fixer l'atténuation maximale pour éviter qu'un site (ou plutôt une zone) de l'image ne possède plus d'information suffisante pour changer d'étiquette. Pour cela, nous chercherons à éviter l'effet de "réaction en chaîne" présenté à la page 104. Supposons donc que toutes les étiquettes passées d'un site donné d'une région uniforme soient celle "non-conforme". La région étant uniforme, on rappelle qu'en l'absence de bruit, l'observation *et* la borne inférieure sur l'observation sont toutes les deux nulles. On souhaite que l'initialisation de l'algorithme d'optimisation à l'échelle la plus grossière, qui se fait sans faire intervenir les termes de régularisation spatiale, se fasse avec l'étiquette conforme. Ceci impose alors:

$$\begin{aligned}
U(d^{s,c}) &\leq U(d^{s,nc}) \\
U_1(d^{s,c}) + U_3(d^{s,c}) &\leq U_1(d^{s,nc}) + U_3(d^{s,nc}) \\
(\mathrm{At}_{max} \times \frac{\alpha_c}{2}) + (9 \times \beta_{td}) &\leq (\mathrm{At}_{max} \times \alpha_{nc}) + (0 \times \beta_{td}) \\
\frac{9\beta_{td}}{\alpha_{nc} - \frac{\alpha_c}{2}} &\leq \mathrm{At}_{max}
\end{aligned}$$

Enfin, le paramètre le plus important est sûrement δ. Il n'est pas difficile de le fixer pour une application donnée. Comme nous l'avons indiqué depuis le début, il modélise la séparation entre les déplacements résiduels (après compensation) qui seront considérés comme des erreurs de recalage, et les déplacements d'objets réellement en mouvement. Il dépend donc directement de la validation de l'hypothèse sous-jacente à la détection des objets mobiles dans une séquence d'images: notre modèle de mouvement doit approcher au mieux le mouvement 2D, induit par le déplacement de la caméra, des projections des objets statiques du monde 3D. C'est la qualité de cette approximation, que nous commenterons dans les exemples qui suivent, qui nous permettra de fixer δ.

4.5 Résultats

Dans un premier temps, pour vérifier et évaluer l'influence de nos paramètres sur la détection de mouvement, nous avons considéré une animation synthétique. Nous présentons ensuite des résultats détaillés sur trois séquences de types différents, et finalement sur

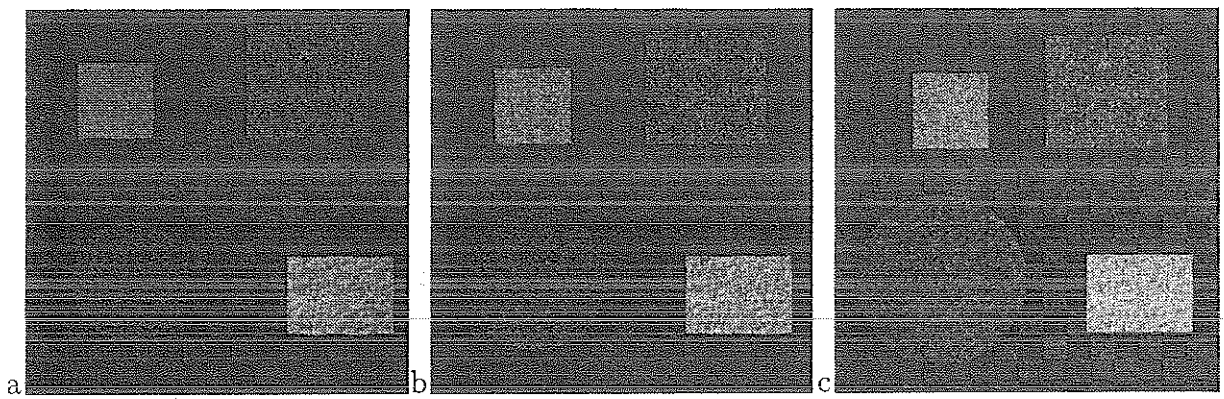


FIG. 4.20 - *Séquence originale DAMIER aux instants t_0, t_4, t_8.*

une séquence où le but sera cette fois-ci de suivre un objet mobile de la scène. Dans toutes les séquences, nous avons utilisé l'algorithme RMR modifié avec le modèle de mouvement affine et une valeur 8 pour C. Cependant, à chaque itération, nous avons considéré comme support d'estimation soit toute l'image, soit la carte de détection précédente projetée dans le sens du mouvement, $\tilde{d}^{t-1}$. Par ailleurs, pour tous les résultats que nous présentons, nous masquerons toujours les régions d'étiquette conforme (en blanc ou en noir) et ferons apparaître dans les régions non-conformes le contenu de l'image originale. Enfin, indiquons que dans les exemples qui suivent, tous les traitements de l'agorithme de détection ne sont pas effectués sur les images originales, mais sur des versions obtenues après un filtrage gaussien de variance 0,8 sur les séquences INTERVIEW et J7, et de variance 1,0 sur les séquences DAMIER, ROND-POINTet AVION.

4.5.1 Séquence synthétique DAMIER

La séquence DAMIER (figure 4.20) a été générée avec un utilitaire de l'IRESTE de Nantes. Elle comprend 10 images de taille 256 × 256. Sur un damier uniforme et fixe se déplacent un carré (translation $\vec{V} = (2, 1)$, i.e. 2 pixels vers la droite et 1 pixel vers le bas par image), deux rectangles (déplacements de $\vec{V} = (-1, 1)$ et $\vec{V} = (-1, 0)$ pour celui du haut et du bas respectivement), et un disque avec un mouvement divergent d'amplitude 0,045 (1,8 pixels de déplacement sur les bords dans la première image). Ces objets sont tous les quatre peu texturés. De plus, une variation d'illumination de 4 niveaux de gris est introduite entre deux images successives, et les images sont bruitées avec un bruit blanc gaussien d'écart type 12, ce qui nous place dans des conditions de détection particulièrement défavorables.

Examinons tout d'abord les caractéristiques indépendantes du mouvement, à savoir les bornes et l'atténuation. Les images 4.21a-b d'une part, 4.21c-d d'autre part présentent les deux types de borne minimale des observations pour un déplacement d'amplitude donnée (formules (4.21)), calculées sur la première image de la séquence pour deux valeurs de G_m: 4 (a et c) et 8(b et d). On peut constater qu'elles sont effectivement élevées sur

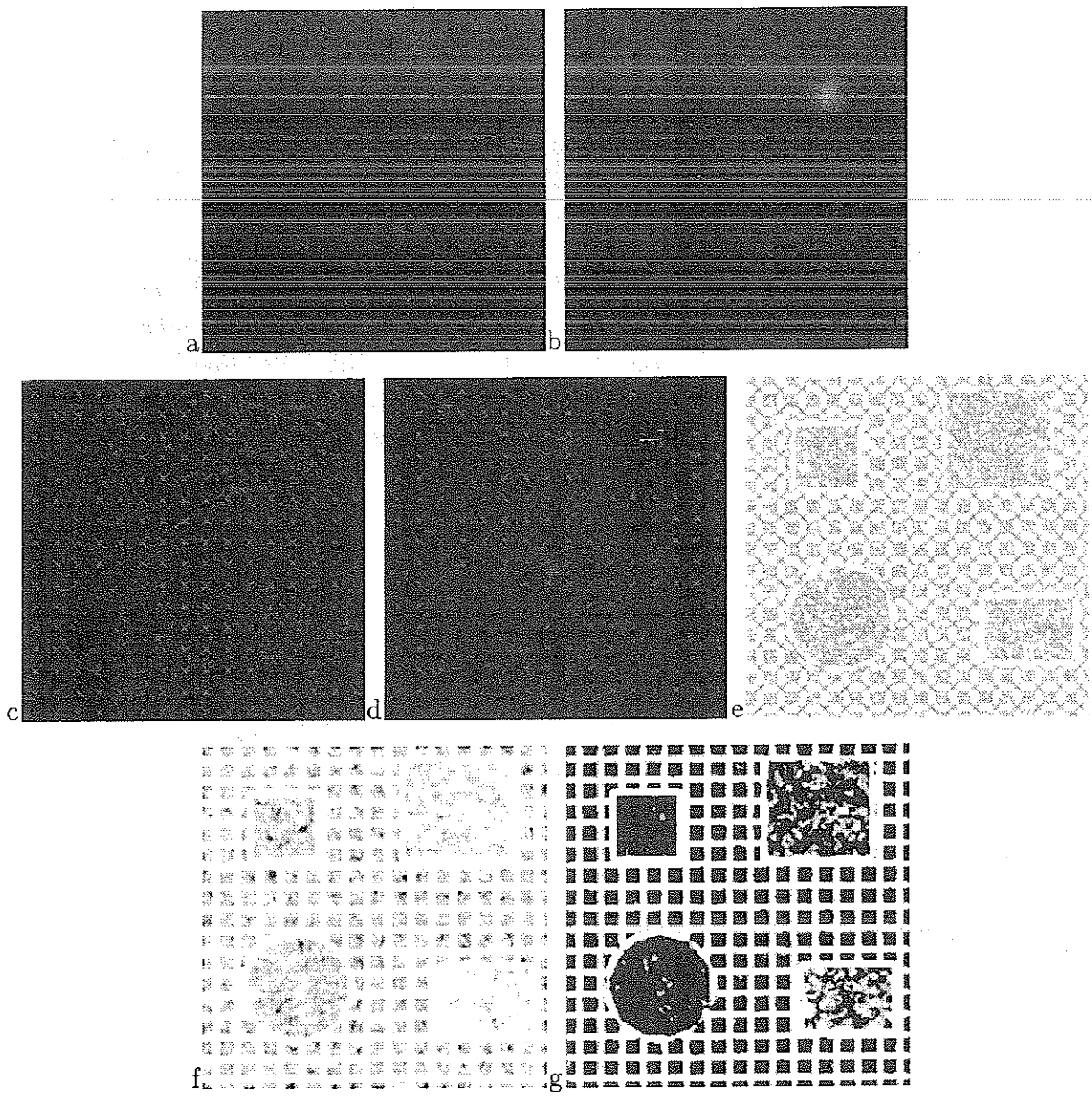


FIG. 4.21 - *Première rangée: borne minimale de l'observation notée l dans les formules du paragraphe sur les observations, pour deux valeurs de G_m: a) 4 et b) 8. Seconde rangée: borne minimale l_m (c,d) et maximale L_m (e) de l'observation utilisant la modélisation de l'isophote passant par un pixel. c) et d) correspondent à deux valeurs différentes de G_m: c) 4 et d) 8. Troisième rangée: coefficient d'atténuation calculé avec f) $k_a = 1$ et $G = 1$, et g) $k_a = 1$ et $G = 4$.*

les coins dans l'image, et nulles sur les contours rectilignes (voir 4.21c notamment). Par contre, étant donné le bruit important, on remarquera que ces bornes ne sont pas nulles à l'intérieur des carrés du damier, pourtant de niveau de gris uniforme à l'origine, lorsque $G_m = 4$. En élevant G_m jusqu'à 8, ne demeurent importantes que les bornes des coins correspondant à des lieux de forte transition (localisés principalement sur les intersections du damier). La borne l_m issue de la modélisation locale de l'isophote passant en un pixel est plus élevée que l. Il suffit de comparer l'intérieur des carreaux du damier entre les figures 4.21a et 4.21c. De faibles observations permettront donc de valider plus facilement l'étiquette "conforme" (correspondant à une région statique ici) avec l_m qu'avec l dans ces zones, ce qui est préférable dans la mesure où effectivement, compte tenu du bruit, on ne peut espérer obtenir des mesures nulles. En revanche, dans ces mêmes zones, la borne maximale L_m (figure 4.21) est plus faible que la valeur δ du premier encadrement que nous avons introduit à la page 89 (cette borne, qui est donc indépendante de la position, serait représentée par une image blanche), indiquant que des observations plus faibles (dues uniquement au bruit dans ces régions) favoriseront donc plus facilement l'étiquette "non-conforme", ce qui ne correspond pas à la réalité. Notons enfin que cette borne supérieure est élevée sur les contours rectilignes (quadrillage blanc du damier) et faible (donc plus sombre sur la figure) aux intersections de ce damier, ce qui est en accord avec notre modélisation. Dans toutes les expériences, nous utiliserons les bornes l_m et L_m issues de la modélisation de l'isophote.

Finalement, les figures 4.21f et 4.21g montrent le coefficient d'atténuation de l'énergie d'attache aux données. Dans ces images, un coefficient d'atténuation de 1 (c'est à dire qu'il n'y a pas d'atténuation) est représenté en blanc, et plus l'image est sombre, plus le coefficient est petit et donc plus l'atténuation est importante. Constatons que pour $G = 1$ (figure 4.21f) –grosso modo, l'amplitude de l'énergie d'attache aux données sera atténuée uniquement aux pixels dont le module du gradient spatial de l'intensité est inférieur à 1–, les objets apparaissent très peu, alors que dans la figure 4.21g où G vaut 4, la structure de l'image originale ressort nettement, ce qui permettra de prendre en compte l'information sur les régions "sûres" de l'image, c'est-à-dire les contours.

Dans les expériences avec cette séquence, l'estimation multirésolution du mouvement est effectuée avec trois niveaux de résolution. Le support d'estimation est constitué à chaque instant de l'image entière. La compensation n'est pas parfaite: on estime un déplacement quasi translationel de 0,1 à 0,2 pixels au lieu du mouvement nul. Les poids w_i calculés à la fin de l'estimation de mouvement, présentés sur la figure 4.22a, sont particulièrement bruités.

Dans le tableau 4.7, nous donnons les valeurs des différents paramètres qui resteront identiques pour toutes les expériences sauf indication contraire.

La figure 4.22b présente pour la première image la détection obtenue en minimisant U_1 uniquement. On peut noter les masques incomplets et les nombreuses fausses alarmes. Les cartes de détection 4.22c et 4.22d montrent, que avec peu ou beaucoup d'atténuation (respectivement $G = 1$, figure 4.22c et $G = 4$, figure 4.22d), il est indispensable de prendre

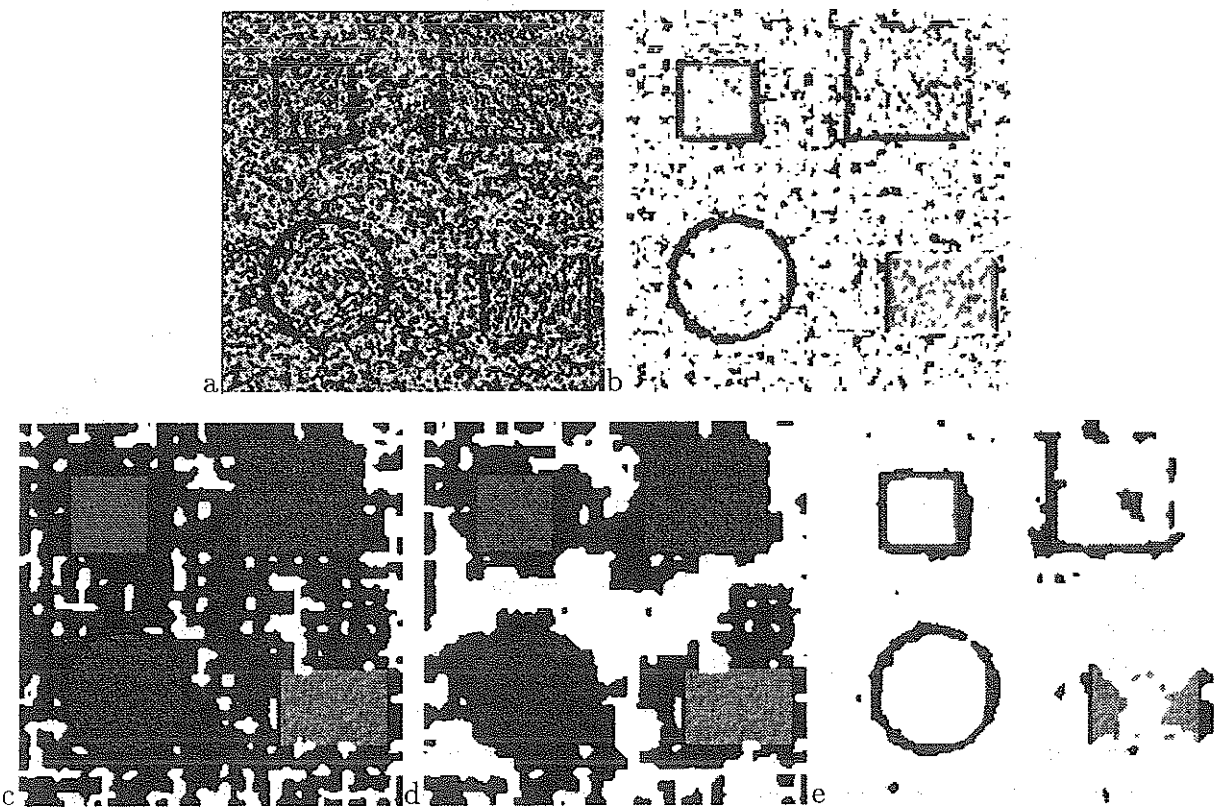


FIG. 4.22 - *Toutes les cartes sont relatives à t_0. a)poids w_i issus de l'estimation de mouvement. b) détection obtenue avec U_1 uniquement (maximum de vraisemblance local appliqué aux observations) ($G_m = 8$). c) détection avec $G_m = 4$ et $G = 1$. d) détection obtenue avec $G_m = 4$ et $G = 4$. e) détection obtenue avec $G_m = 8$ et $G = 1$.*

Paramètre	G_m	δ	G	At_{max}	T	γ	β_{nc}	β_d	β_{td}	L_{det}
Valeur	8,0	0,8	4,0	0,2	∞	0,4	0	1,5	0,1	5

TAB. 4.7 - *Valeurs par défaut des paramètres utilisés dans toutes les expériences, sauf indication contraire.*

une valeur de G_m supérieure à la valeur 4 utilisée ici pour éviter les fausses observations de mouvement dues au bruit. La figure 4.22e montre ce que l'on obtient avec une valeur de 8, mais avec peu d'atténuation (la carte initiale de détection correspond dans ce cas à 4.22b). Les fausses alarmes ont presque disparu, mais l'intérieur des régions en mouvement est considéré comme statique.

C'est pourquoi nous devons conjuguer une valeur forte de G_m et un facteur d'atténuation important. Les figures 4.23a-c montrent les résultats obtenus dans ce cas aux instants t_0, t_2, t_8. On peut voir également l'effet bénéfique de l'intégration temporelle. Les fausses alarmes sont inexistantes, et les masques détectés ont très rapidement une forme proche des objets. Notons ici que comme le disque est en expansion, son intérieur a un mouvement très faible voire nul en son centre, ce que retrouve bien l'algorithme. Remarquons également que les côtés horizontaux du rectangle inférieur sont étiquetés comme "conformes" car ils glissent sur eux-mêmes et ne génèrent de ce fait absolument pas d'observations pouvant indiquer leur déplacement. De plus, comme ils constituent des régions fiables, ils ne se laissent pas "influencer" par les bords verticaux, contrairement à l'intérieur de ce même rectangle. Les résultats sur la rangée suivante proviennent exactement du même modèle, mais la minimisation est effectuée avec un seul niveau d'échelle au lieu des cinq utilisés précédemment avec l'algorithme multiéchelle. Notons tout de même qu'il est possible de choisir dans le cas monoéchelle des paramétrages donnant de meilleurs résultats, notamment avec plus de régularisation, mais toujours de qualité inférieure à ceux utilisant la minimisation multiéchelle.

L'expérience aboutissant aux cartes de détection 4.23g-i diffère de celle de la première rangée par l'emploi d'une valeur de β_{nc} non nulle qui favorise le formation des régions mobiles (non-conformes) compactes. Cependant, compte tenu de l'atténuation importante, les étiquettes mobiles se propagent trop facilement à l'extérieur des objets dès qu'une observation favorable au mouvement est détectée sur le damier, créant des situations pratiquement irréversibles (les positions passées du carré restent étiquetées mobiles). La dernière rangée de résultats, figures 4.23j-l, montre qu'avec une atténuation plus modeste ($G = 1$), β_{nc} remplit son rôle et son utilisation conduit à des résultats similaires à ceux de la première rangée.

Enfin, les neuf images 4.24a-i montrent l'évolution des masques lorsque le paramètre δ passe de 0,8 (la valeur par défaut), à 2. Pour interpréter ces résultats, il est important de rappeler que les objets ont une texture d'amplitude équivalente au bruit, et que ce

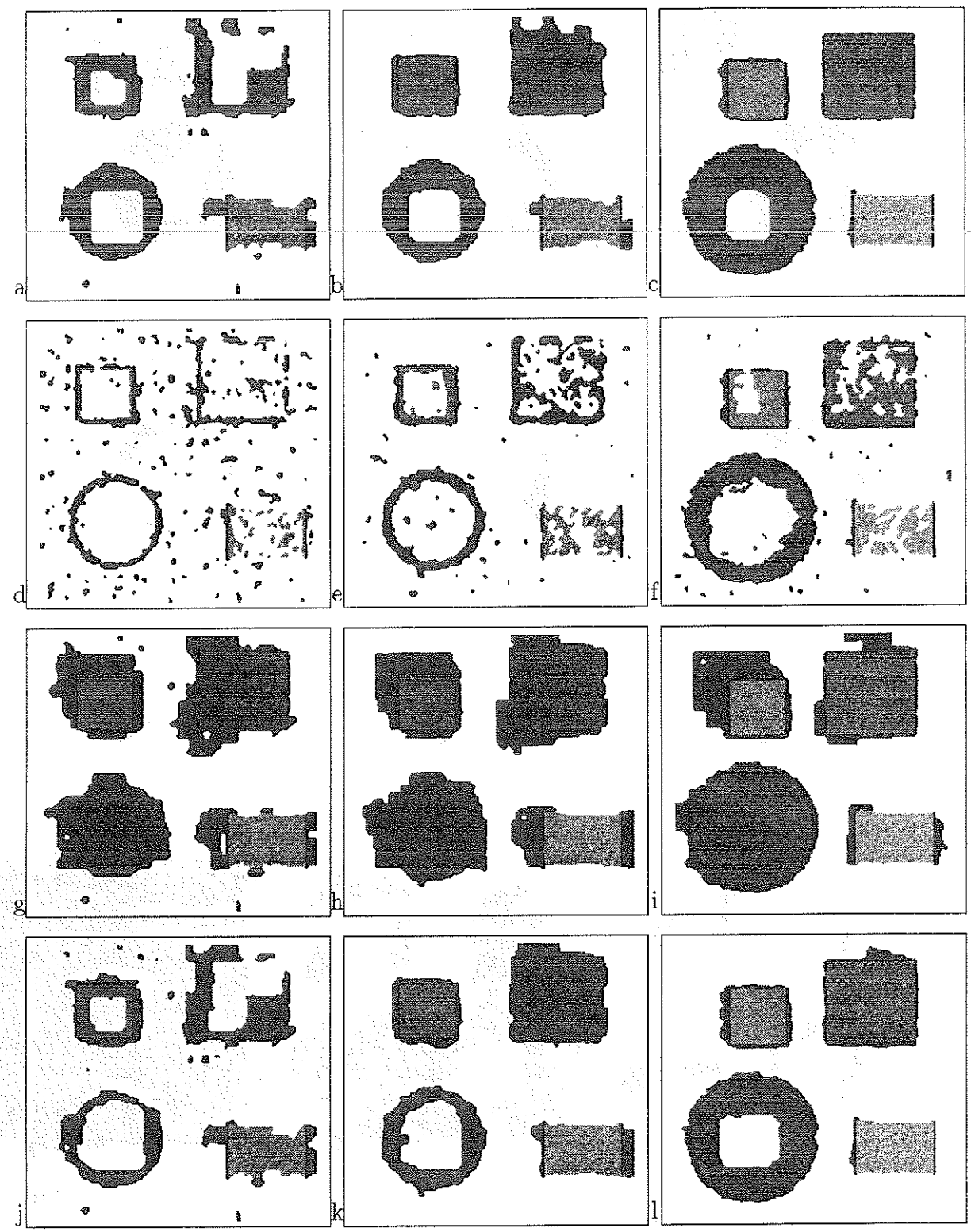


FIG. 4.23 - *Résultats aux instants t_0, t_2, t_8 de quatre expériences. Première rangée, a,b,c: paramètres par défaut. Seconde rangée, d,e,f: minimisation monoéchelle ($R = 1$). Troisième rangée, g,h,i: on favorise les régions mobiles en choisissant $\beta_{nc} = 3$. Dernière rangée, j,k,l: $\beta_{nc} = 3$ et $G = 1$ (facteur d'atténuation plus faible).*

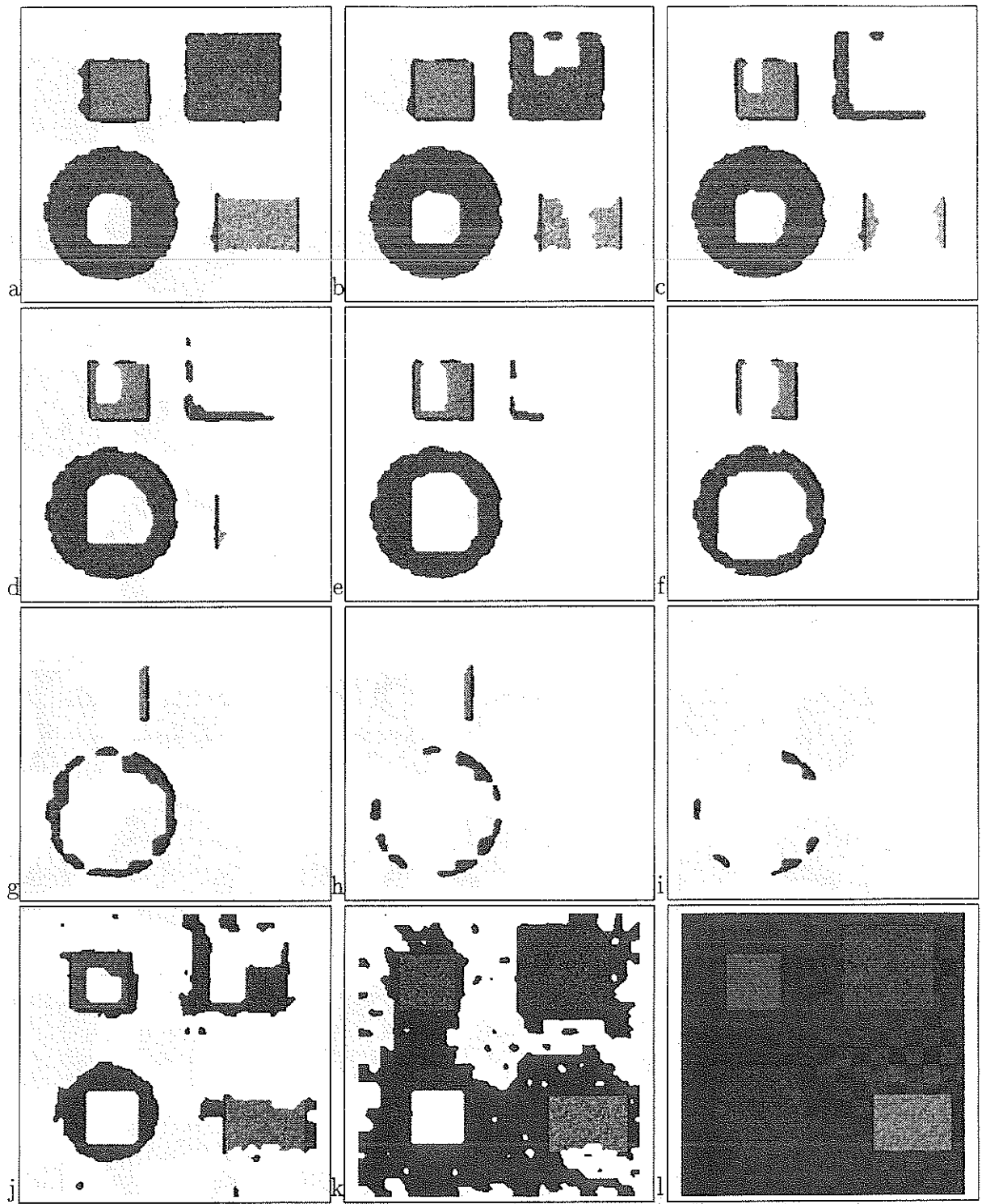


FIG. 4.24 - *a-i: cartes de détection à t_8 pour différentes valeurs de δ (les autre paramètres sont les paramètres par défaut). Valeur de δ: a)* 0,8*; b)* 0,9*; c)* 1*; d)* 1,1*; e)* 1,2*; f)* 1,4*; g)* 1,6*; h)* 1,8 *et i)* 2*. Dernière rangée: résultats obtenus avec une carte de référence $w = 0{,}3$; pas d'intégration temporelle des observations ($T = 1, \beta_{dt} = 0$). j,k: cartes de détection à t_0, t_1. l: image de référence à t_2.*

Paramètre	G_m	δ	G	At_{max}	T	γ	β_{nc}	β_d	β_{td}	L_{det}
Valeur	3,0	0,5	1,0	0,3	∞	0,4	0,1	1,5	0,15	5

TAB. 4.8 - *Valeurs des paramètres utilisés pour la séquence INTERVIEW.*

sont donc les côtés de ces objets qui portent l'essentiel de l'information de mouvement. Ces derniers se déplacent de 1 pixel au plus pour les deux rectangles, et le côté vertical du carré de deux pixels. Lorsque δ croît, on observe alors d'une part que l'information de mouvement se propage plus difficilement à l'intérieur des objets, et d'autre part que les déplacements de ces objets sont progressivement assimilés à des erreurs de recalage. L'exemple du cercle est significatif: la tache centrale se développe au fur et à mesure de l'accroîssement de δ, conformément à la signification de ce paramètre.

Finalement, nous présentons les résultats obtenus avec le schéma de Irani *et al.* [IRP92]. Dans cette expérience, nous employons donc une image de référence moyennée temporellement. La détection se fera en utilisant notre méthode de régularisation (qui est beaucoup plus sophistiquée que le simple seuillage réalisé dans [IRP92]) appliquée entre deux images, l'image de référence à l'instant t et l'image de la séquence à l'instant $t+1$. Nous ne prenons donc pas en compte la carte de détection à l'instant précédent ($\beta_{td} = 0$), et n'utilisons pas d'intégration temporelle des observation ($\gamma = 0$), ceci devant être effectué dans la méthode de Irani par l'intermédiaire de la carte de référence. Nous avons pris pour les autres paramètres les valeurs par défaut (nous avons également testé d'autres jeux de paramètres sans succès). Comme on peut le constater, l'utilisation d'une carte de référence, même avec une faible intégration temporelle ($w = 0{,}3$) conduit tout de suite à des résultats catastrophiques dès la seconde image (figure 4.24k à l'instant t_1). Le problème vient ici de l'estimation de mouvement qui n'est pas parfaite et brouille toute l'image de référence, y compris le damier (figure 4.24l), rendant plus difficiles les estimations futures.

4.5.2 Séquences réelles

Séquence INTERVIEW

La séquence INTERVIEW provient de la BBC et est fréquemment utilisée pour l'évaluation d'algorithmes de codage de séquence d'images. Nous utilisons une version sous-échantillonnée spatialement (figure 4.25a-d) de 50 images de taille 256 × 320, qui est de bonne qualité et nous conduit à choisir des valeurs faibles pour G_m et G (respectivement 3 et 1). Un exemple d'images présentant la borne inférieure l_m en chaque point, ainsi que le coefficient d'atténuation est proposé sur les figures 4.25e et 4.25f respectivement.

Au cours de cette séquence, la femme sur la partie droite de l'image dont la main gauche est initialement cachée par un bouquet de fleurs, se redresse et se lève. La caméra

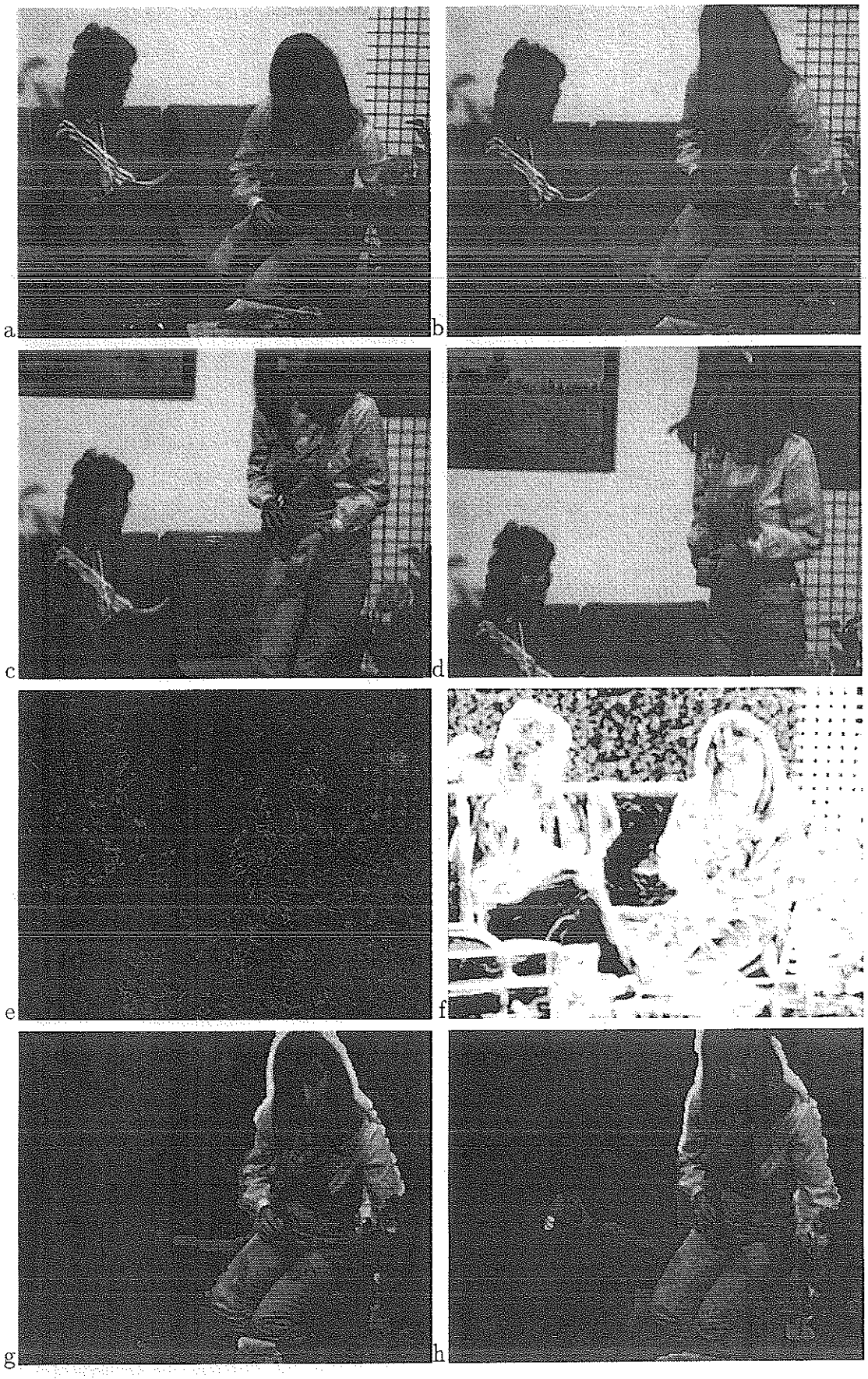


FIG. 4.25 - *Séquence originale INTERVIEWaux instants a) t_1, b) t_{17}, c) t_{33} et d) t_{49}. e) Borne inférieure l_m en chaque point de l'image 1 ($G_m = 3$). f) Coefficients d'atténuation en chaque point de l'image 1 ($G = 1$). g) h) résultats aux instants t_{13} et t_{19} obtenus avec une carte de référence ($w = 0{,}3$) et sans d'intégration temporelle des observations ($T = 1, \beta_{dt} = 0$).*

Paramètre	G_m	δ	G	At_{max}	T	γ	β_{nc}	β_d	β_{td}	L_{det}
Valeur	10,0	1,0	1,0	0,4	2	0,4	0,1	1,75	0,2	5

TAB. 4.9 - *Valeurs des paramètres utilisés pour la séquence ROND-POINT.*

suit son mouvement ascendant, ce qui entraîne dans la séquence un déplacement apparent vers le bas de la partie statique de la scène. Par ailleurs, des ombres portées (correspondant à la femme en mouvement) glissent sur le canapé, tandis qu'au début de la scène, le coussin comprimé sur lequel était assise la femme qui se lève, retrouve sa forme naturelle.

L'estimation du mouvement se fait ici en considérant comme support d'estimation à chaque itération, la carte de détection précédente projetée dans le sens du mouvement, soit $\vec{d}^{t-1}$, hormis pour la première image où toute l'image sert de support. La compensation est quasi parfaite, sauf sur les carreaux à droite où un léger glissement est observable. Les six paramètres du modèle de mouvement affine dominant (formule 3.2, où le centre du repère est le centre de l'image), estimés à l'instant t_{40} sont les suivants: $a_1 = -0{,}37$, $a_2 = -0{,}00004$, $a_3 = 0{,}00014$, $a_4 = 3{,}40$, $a_5 = -0{,}00004$ et $a_6 = -0{,}0001$. L'axe des x est orienté de gauche à droite, l'axe des y de haut en bas.

Les paramètres utilisés se trouvent dans le tableau 4.8. Les images 4.25g et 4.25h présentent deux images de la séquence obtenue en s'inspirant de la méthode Irani *et al.* (que nous avons présentée pour l'exemple précédent) pour une valeur de w égale à 0,3, sur lesquelles on peut remarquer l'imprécision de la détection. Des étiquettes "non-conformes" sont retenues sur les bords des régions immobiles de la scène (revues sur la table, bouquet de fleur). En fait, plus w augmente, plus les résultats se dégradent, montrant que l'intégration temporelle des images n'est pas une bonne chose. *A contrario*, les résultats de notre algorithme, figures 4.26a-h, donnent des cartes de détection plus précises (les revues et les fleurs sont bien étiquettées, figures 4.26c et 4.26d). On peut vérifier sur cette séquence un aspect que nous avions évoqué: les frontières se trouvent principalement à l'extérieur des régions mobiles. De plus, le masque des régions détectées comme mobiles (non-conformes) s'étend d'autant plus facilement sur les régions immobiles que ces dernières sont uniformes. Il suffit par exemple de comparer la frontière à gauche de la femme qui se lève, et à droite au niveau de son bras, où se situent les carreaux du mur du fond.

Séquence ROND-POINT

Nous avons retenu 34 images (de t_{46} à t_{79}) de la séquence ROND-POINTque nous avons déjà présentée dans le chapitre sur l'estimation de modèles de mouvements. L'estimation de mouvement est effectuée ici en prenant à chaque instant l'image complète comme support. Par exemple, les six paramètres du modèle de mouvement affine dominant (formule 3.2, où le centre du repère est le centre de l'image), estimés entre les deux premières images, sont les suivants: $a_1 = -4{,}56$, $a_2 = 0{,}0018$, $a_3 = -0{,}0105$, $a_4 = -0{,}35$, $a_5 = 0{,}0018$ et

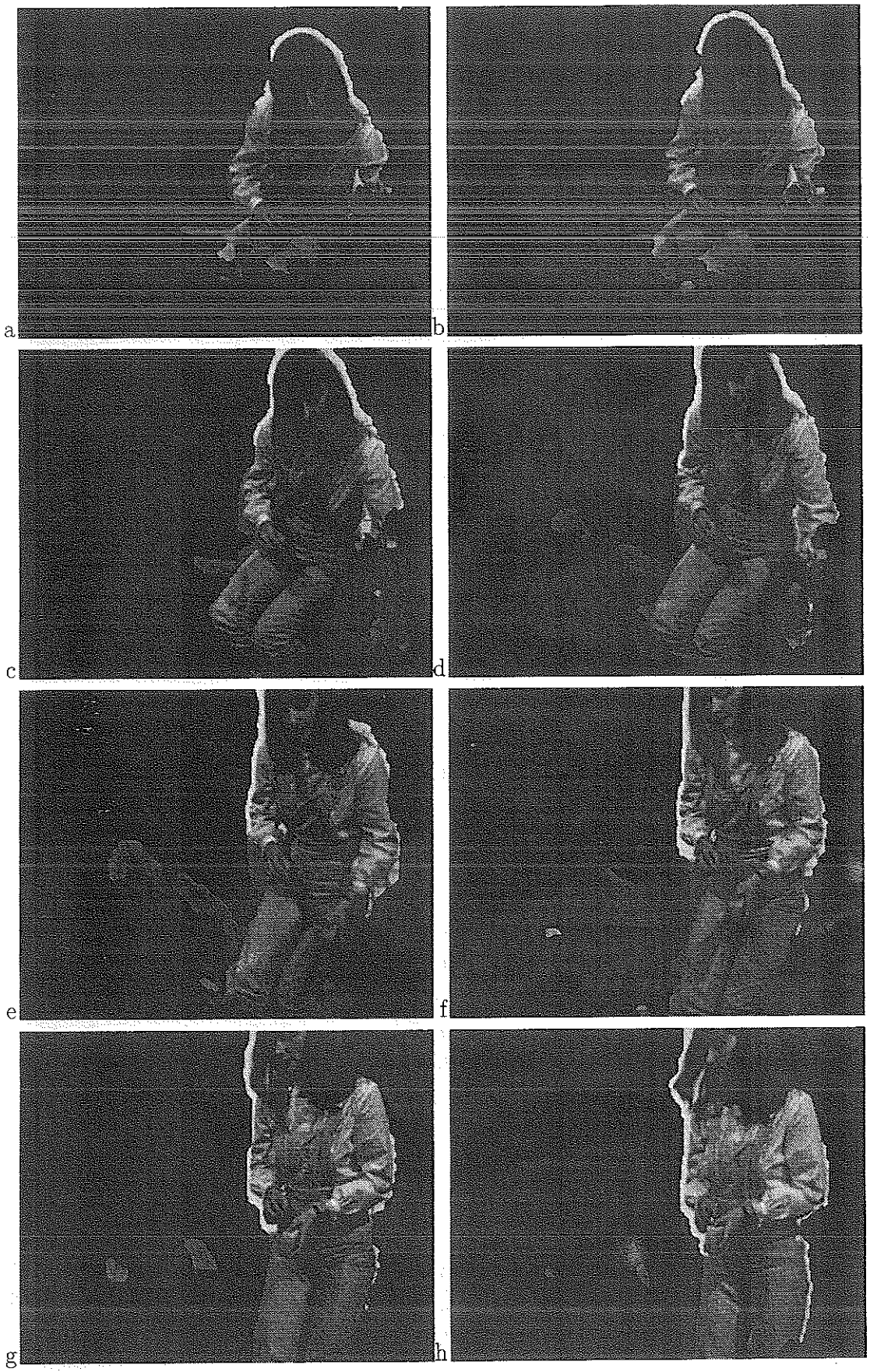


FIG. 4.26 - *Cartes de détection obtenues aux instants a) t_1, b) t_7, c) t_{13}, d) t_{19}, e) t_{25}, f) t_{31}, g) t_{37} et h) t_{43}.*

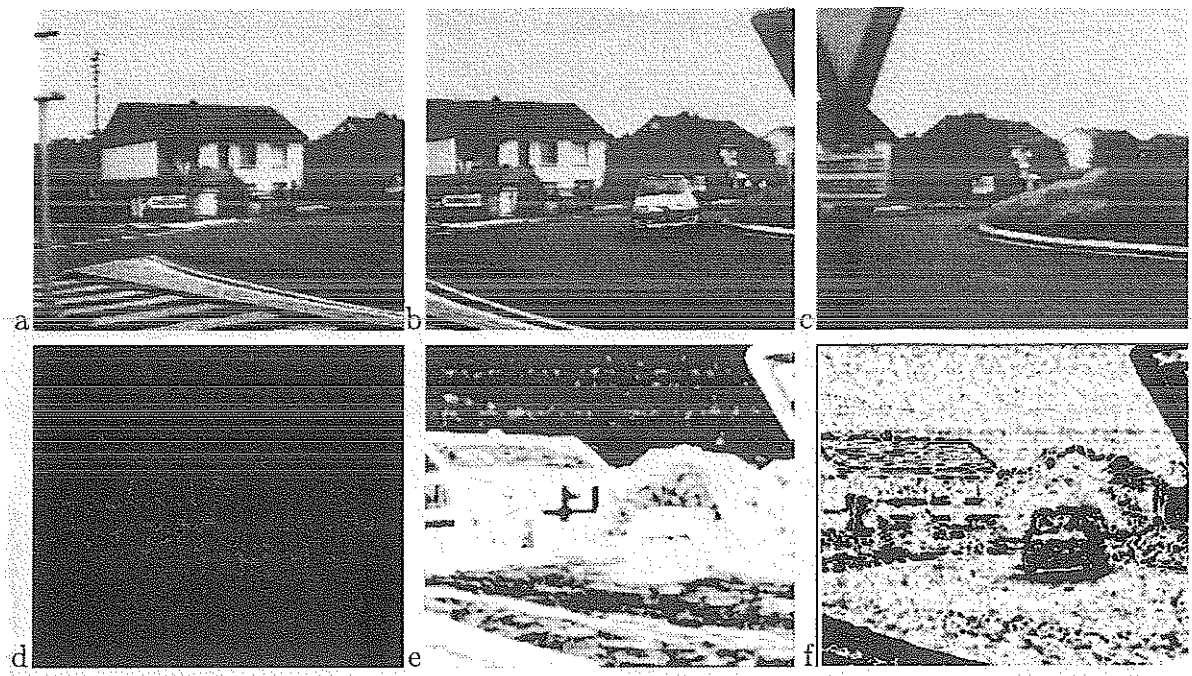


FIG. 4.27 - *Séquence originale ROND-POINT aux instants a) t_{46}, b) t_{62}, et c) t_{78}. d) e) f) Pour l'image à t_{62}: d) valeur de la borne inférieure l_m en chaque point ($G_m = 10$). e) coefficient d'atténuation en chaque point ($G = 1$). f) poids w_i calculés à l'issue de l'estimation de mouvement ($C = 8$).*

$a_6 = 0{,}0030$. La figure 4.27f présente les poids w_i en chaque point de l'image à t_{62} calculés à l'issue de l'estimation de mouvement. Pour cette même image, nous proposons également la carte des valeurs de la borne inférieure l_m et la carte des coefficients d'atténuation (figures 4.27d et 4.27e respectivement) calculés en utilisant les valeurs des paramètres du tableau 4.9.

Comme nous avons pu l'observer sur les séquences présentées dans le chapitre sur l'estimation, la compensation n'est pas parfaite, et nous choisissons de tolérer des erreurs de recalage jusqu'à 1 pixel. Du fait du sous-échantillonnage temporel de la séquence originale, les masques des régions "non-conformes" évoluent très rapidement. C'est pourquoi nous considérons de filtrer uniquement les observations calculées à deux instants ($T = 2$). Les autres paramètres sont en accord avec les considérations générales que nous avons exposées (avec une régularisation plus importante compte-tenu du bruit). Notons que plusieurs jeux de paramètres différents ont donné des résultats tout à fait similaires ($T = \infty$, $G = 1$ à 4, β_d de 30 à 40, δ de 0,8 à 1,5). Une valeur inférieure à 10 pour G_m (6 à 8) donne également des résultats comparables à la différence près suivante: le repliement de spectre spatio-temporel qui existent sur les tuiles des toits des maisons engendre une sorte d'onde sinusoïdale qui se déplace au cours du temps; avec une valeur plus faible de G_m, ce mouvement apparent (réellement présent dans l'image) est alors détecté.

On peut constater, sur les résultats de la figure 4.28 l'excellente qualité de la détection

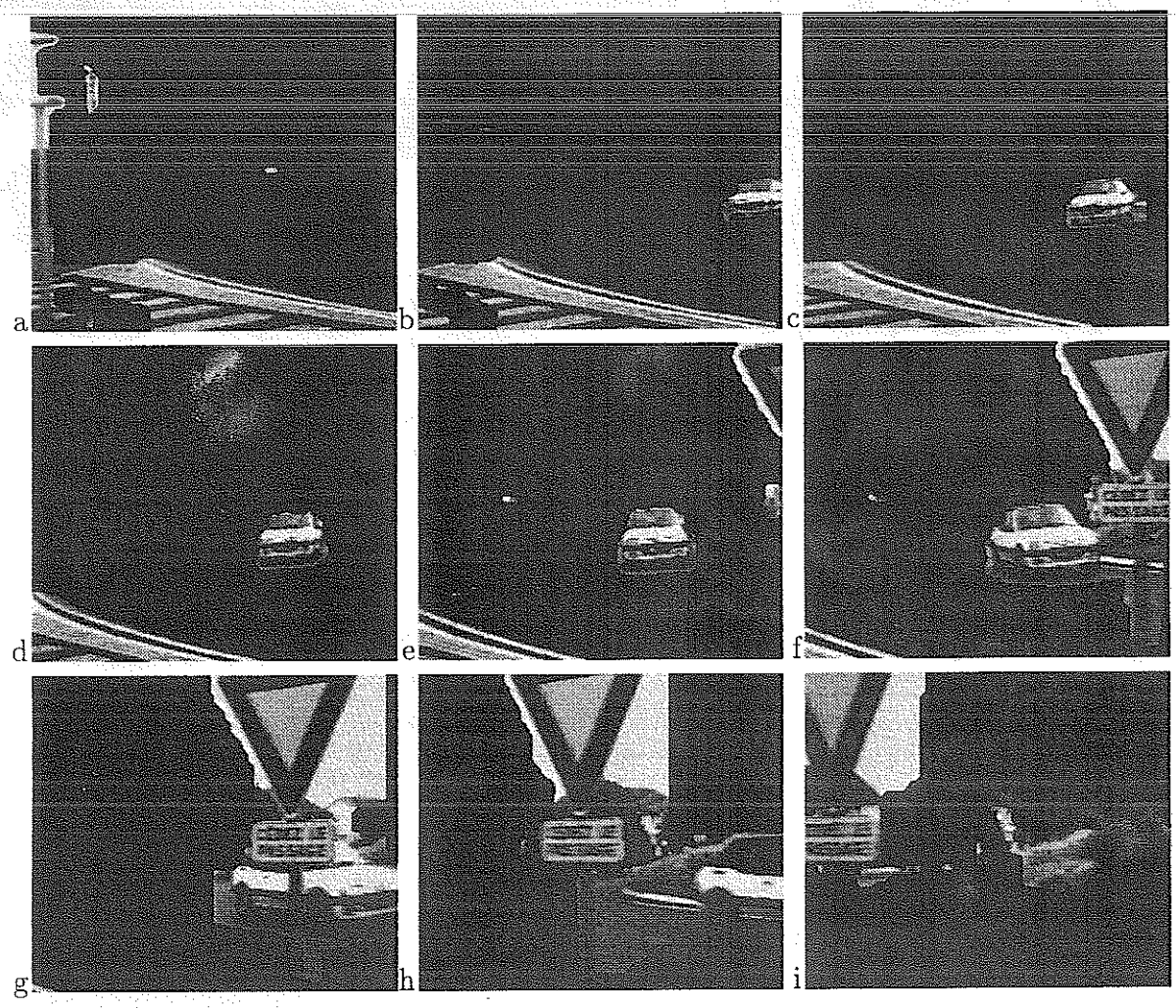


FIG. 4.28 - *Cartes de détection obtenues aux instants a)* t_{46}, *b)* t_{50}, *c)* t_{54}, *d)* t_{58}, *e)* t_{62}, *f)* t_{66}, *g)* t_{70}, *h)* t_{74} *et i)* t_{78}.

Paramètre	G_m	δ	G	At_{max}	T	γ	β_{nc}	β_d	β_{td}	L_{det}
Valeur	4,0	0,3	1,0	0,3	∞	0,6	0,25	1,25	0,15	3

TAB. 4.10 - *Valeurs des paramètres utilisés pour la séquence AVION.*

sur cette séquence extrèmement bruitée. Rappelons ici que le mouvement estimé correspond principalement au déplacement de l'arrière plan (la rangée de maisons) et que les objets, même statiques, situés à une profondeur bien inférieure (comme les marques sur la route, ou le panneau de signalisation) sont justement détectés comme ayant un mouvement non-conforme au mouvement dominant estimé.

La méthode de Irani *et al.* présente ici les mêmes handicaps que dans la séquence DAMIER. Très rapidement, et ce même pour des valeurs faibles de w, l'image de référence se brouille, l'estimateur de mouvement n'arrive plus à calculer le mouvement dominant, et l'image entière est détectée comme ayant un mouvement non-conforme.

Enfin, donnons ici quelques indications de temps de calcul pour cette séquence. Nous avons mesuré en moyenne pour les 34 images, une durée de 8,1 secondes de temps cpu par image sur une machine SPARC10. Celle-ci se décompose en: 3,3 secondes pour le calcul du modèle de mouvement, 2,4 secondes pour le calcul des observations et des bornes, 1,15 seconde pour la partie minimisation de l'énergie (dont 0,4 seconde environ pour le seul calcul des potentiels V_1, en utilisant la formule analytique, alors que des tables permettraient de considérablement réduire ce temps de calcul), le reste se décomposant en opérations diverses (calcul des pyramides gaussiennes, des dérivées, de $\tilde{d}^{t-1}$, etc.).

Séquence AVION

La troisième séquence considérée se compose de 28 images de taille 200×200. Trois avions se déplacent relativement au fond. L'un d'eux (au milieu) s'apprète à aterrir. Alors qu'il se détache légèrement dans le ciel dans la première partie de la séquence, il se confond ensuite partiellement avec la ligne de séparation entre le ciel et les frondaisons. La caméra effectue un léger panoramique pour suivre les avions. Le deplacement total du sol entre la première et la dernière image est à peu près de 16 pixels horizontalement. Relativement à ce sol, les mouvements moyens des avions sont indiqués dans le tableau 4.11.

Comme on peut le constater, l'amplitude moyenne totale des déplacements des avions est relativement importante. Cependant, le déplacement est essentiellement horizontal. Les projections des avions glissent sur elles-mêmes dans l'image et ne génèrent que peu d'observations de mouvement (hormis au bout des ailes!).

L'estimation de mouvement est effectuée sur toute l'image. Par exemple, les six paramètres du modèle de mouvement affine dominant (formule (3.2), dans laquelle le centre

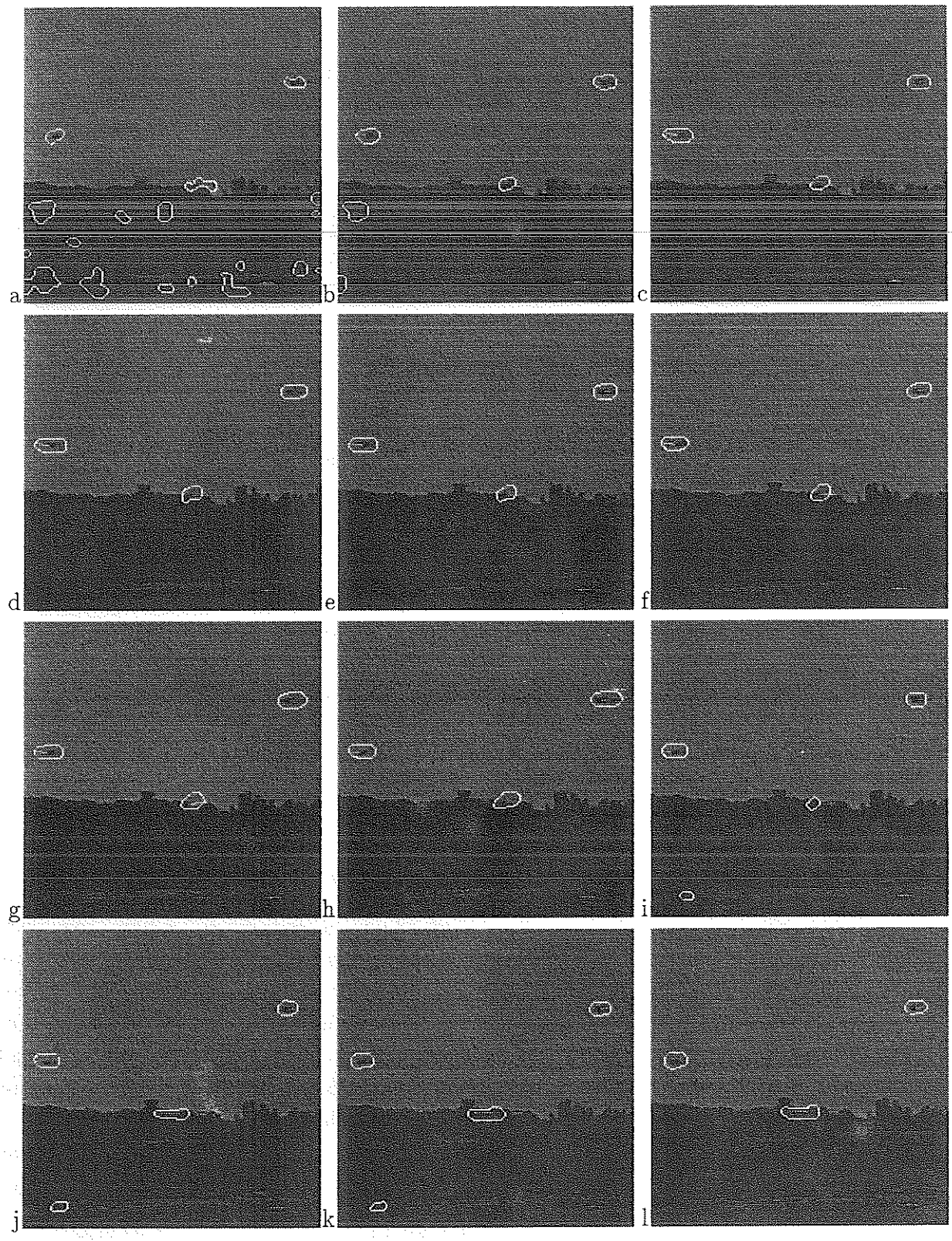


FIG. 4.29 - *Cartes de détection obtenues aux instants a)* t_1, *b)* t_2, *c)* t_5, *d)* t_7, *e)* t_9, *f)* t_{11}, *g)* t_{13}, *h)* t_{15}, *i)* t_{17}, *j)* t_{19}, *k)* t_{21} *et l)* t_{23}.

Avion	gauche	milieu	droite
Déplacement horizontal/image (en pixel)	0,81	0,85	0,74
Déplacement vertical/image (en pixel)	0,14	0,07	0,14

TAB. 4.11 - *Deplacements des avions par rapport au fond.*

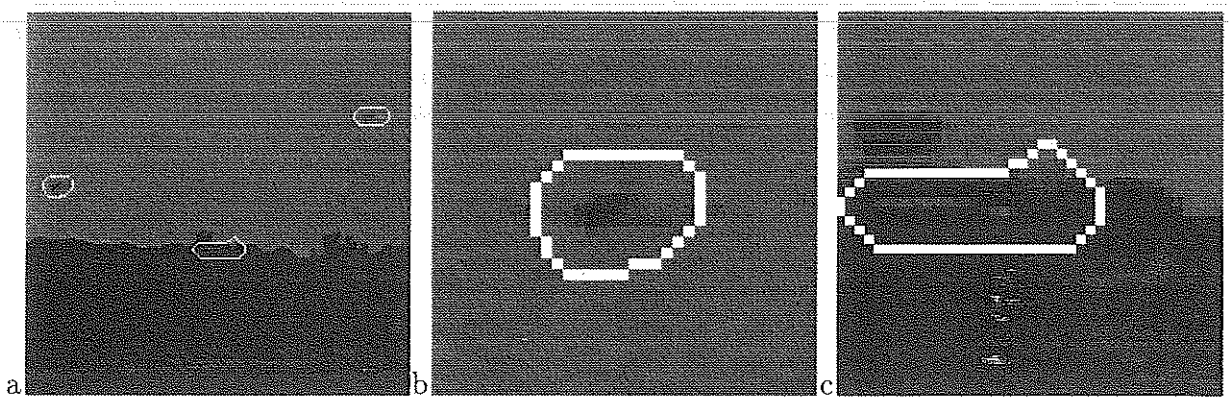


FIG. 4.30 - *a) carte de détection à l'instant t_{27}. b) détail autour de l'avion à gauche de l'image (le contour en blanc n'appartient pas à la zone non-conforme). c) détail au niveau de l'avion du milieu.*

du repère est le centre de l'image), estimés à l'instant t_6, sont les suivants: $a_1 = 0{,}54$, $a_2 = -0{,}0012$, $a_3 = 0{,}0020$, $a_4 = 0{,}07$, $a_5 = 0{,}00002$ et $a_6 = 0{,}0004$. La compensation est de très grande qualité sur le fond de l'image, mais légèrement imprécise au premier plan (de 0,1 à 0,2 pixels). Nous avons donc choisi $\delta = 0{,}3$. Le paramètre de régularisation spatial, étant donné la taille des objets à détecter, est plus faible ($\beta_d = 25$) que dans les autres cas. Par ailleurs, pour favoriser les groupements d'étiquettes "non-conformes", nous avons choisi $\beta_{nc} = 5$. Les résultats sont présentés sur la figure 4.29, et l'on peut remarquer la qualité de la détection. Très rapidement, l'intégration temporelle élimine les détections intempestives initiales. Nous obtenons sur le reste de la séquence deux fausses alarmes, une qui persiste sur cinq images, et l'autre présente sur une image uniquement.

Au milieu de la séquence (image 15 à 19), le masque de l'avion du milieu se réduit au cockpit. Le mouvement qui était perceptible surtout sur son aile gauche au début de la séquence le devient surtout sur l'aile droite qui quitte la zone de séparation ciel/végétation. Du fait de l'intégration temporelle des observations, l'algorithme met un certain temps à réagir. Nous présentons sur la figure 4.30 l'image de détection obtenue à la fin de la séquence ainsi que des vues plus détaillées des régions détectées.

Nous avons également effectué la détection sur des versions sous-échantillonnées temporellement de la séquence. Par exemple, si l'on prend une image sur deux, l'expérience menée avec les mêmes paramètres, hormis une valeur de δ plus élevée $(0,4)$ donne des résultats sans fausses alarmes et avec un masque plus complet de l'avion du milieu lorsqu'il

se confond avec l'horizon. L'explication est la suivante. Alors que le mouvement relatif des avions par rapport au fond double du fait du sous-échantillonnage, l'erreur de recalage, elle, n'est pas nécessairement le double de ce qu'elle est entre deux images consécutives; en effet, l'erreur de recalage mesurée résulte à la fois de l'erreur de modèle (le mouvement affine ne peut pas modéliser parfaitement à la fois le déplacement du fond et celui du premier plan) et de l'erreur due à l'interpolation des intensités (dans notre cas, nous utilisons une simple interpolation bilinéaire). Or, cette dernière est indépendante de l'amplitude des mouvements. Il est donc vraissemblable que l'emploi d'un meilleur interpolateur ainsi que de mesures à différentes échelles temporelles (comme dans [LRB93, Let93]) permettront d'obtenir des résultats plus robustes dans le cas de mouvements aussi faibles.

Séquence J7

Cette dernière expérience présente un exemple de suivi d'objet dans une séquence. La séquence J7 est composée de 66 images de taille 288×332 pixels acquises par une caméra montée sur l'avant d'une voiture. Celle-ci roule derrière un fourgon J7. Le tableau 4.12 rassemble les valeurs des paramètres que nous avons utilisées. Les résultats sont présentés

Paramètre	G_m	δ	G	At_{max}	T	γ	β_{nc}	β_d	β_{td}	L_{det}
Valeur	6,0	1,0	2,0	0,2	∞	0,6	0,1	1,75	0,2	5

TAB. 4.12 - *Valeurs des paramètres utilisés pour la séquence J7.*

sur la figure 4.31. Pour lancer l'algorithme, le premier support d'estimation du mouvement est une fenêtre positionnée sur le fourgon dans la première image (à t_5; la figure 4.32e donne une idée de la position de cette fenêtre). Les six paramètres du modèle de mouvement affine (formules 3.2, où le centre du repère est le centre de l'image), estimés à cet instant, sont les suivants: $a_1 = -0{,}37$, $a_2 = 0{,}0444$, $a_3 = 0{,}0062$, $a_4 = -1{,}55$, $a_5 = -0{,}0033$ et $a_6 = 0{,}0440$. Ensuite, l'algorithme s'exécute comme dans les cas précédents. Le support d'estimation à chaque instant suivant t_5 est constitué des régions conformes de la carte $\tilde{d}^{t-1}$.

Au début de la séquence, la voiture et le fourgon ralentissent en raison d'un feu rouge (figures 4.31a-c). Le mouvement apparent divergent croissant du fourgon (i.e. la voiture se rapproche du fourgon suivant un mouvement axial) devient similaire au mouvement divergent du reste de l'image. Le masque des régions détectées comme ayant un mouvement conforme au mouvement apparent estimé du fourgon s'étend alors à l'extérieur du fourgon (figure 4.31c). Par la suite, le fourgon accélère et s'éloigne de la voiture. Son mouvement se distingue alors de celui du fond, et l'on peut récupérer une carte de détection plus précise (figure 4.31e). A la fin de la séquence, le fourgon est éloigné et son mouvement apparent est de faible amplitude. Les régions qui l'entourent, situées près du foyer d'expansion, sont également animées d'un petit déplacement qui est donc proche (à $\delta = 1$ pixel près) de celui du modèle estimé. Elles se retrouvent donc incluses dans la région des points ayant

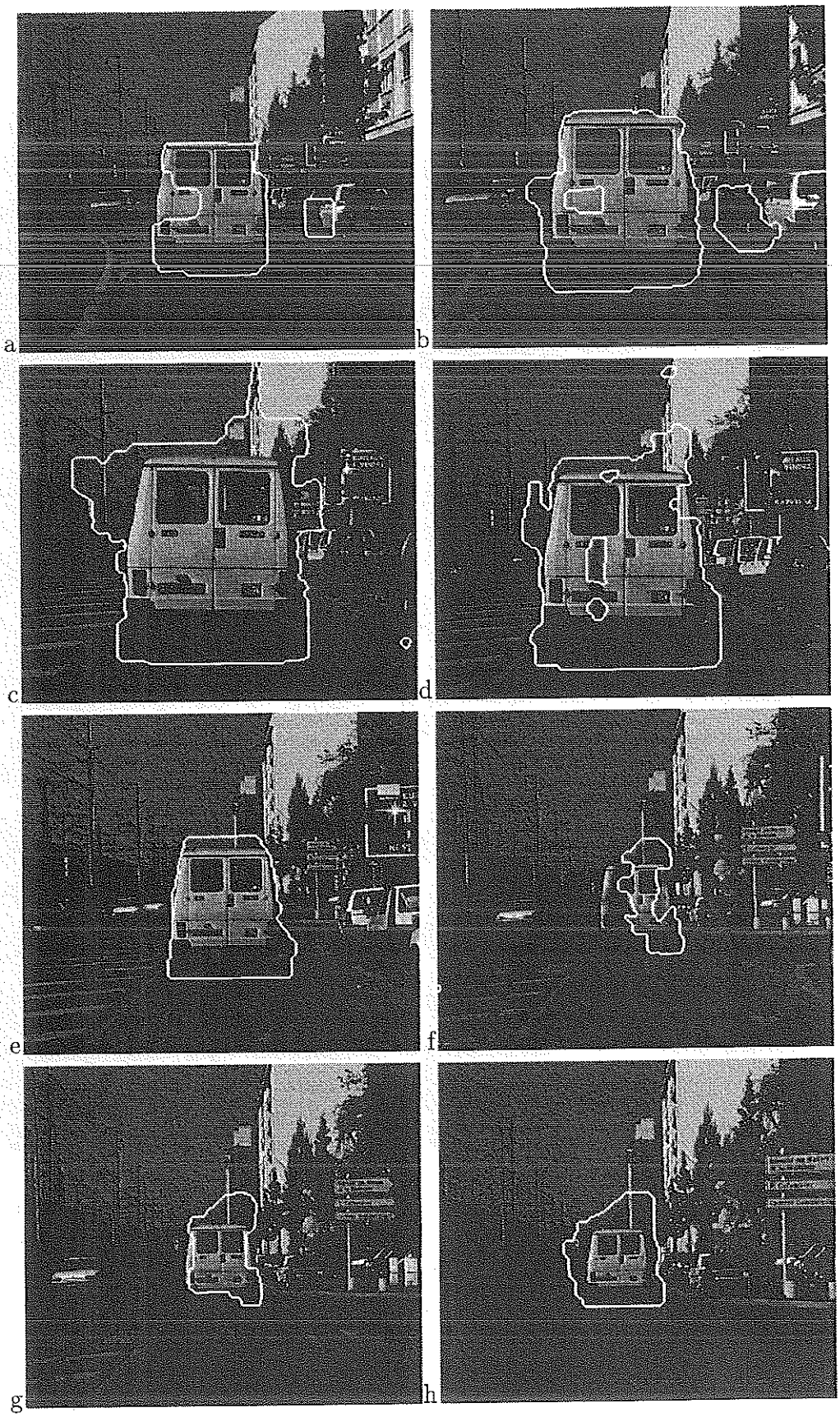


FIG. 4.31 - *Cartes de détection obtenues aux instants a) t_5, b) t_{11}, c) t_{20}, d) t_{27}, e) t_{40}, f) t_{55}, g) t_{58} et h) t_{65}. Les régions d'étiquette "conforme" sont entourées en blanc.*

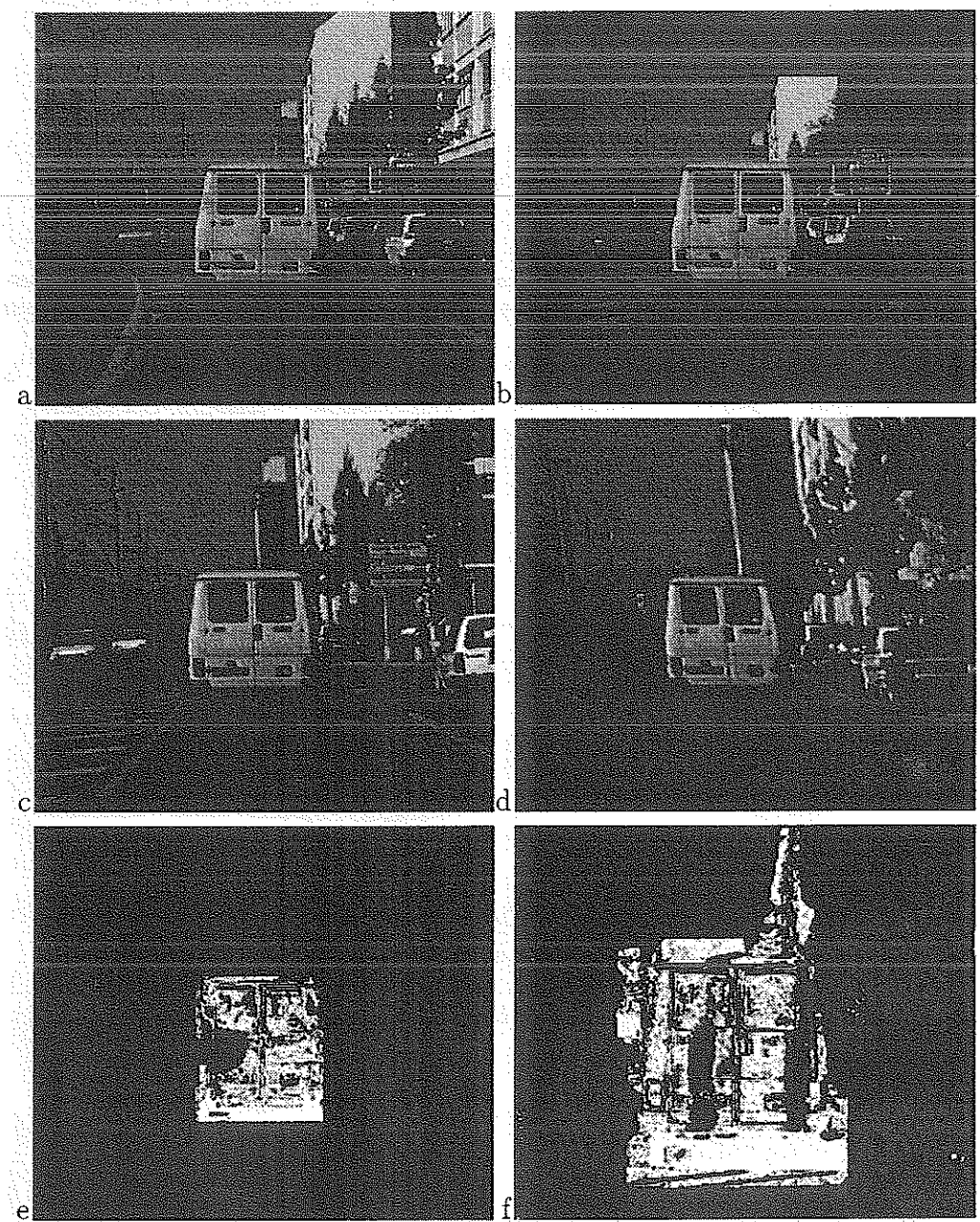


FIG. 4.32 - *a) Image à l'instant t_5. b) c) d) images aux instants b) t_{25}, c) t_{45} et d) t_{65} compensées à l'aide de l'ensemble des mouvements estimés entre ces instants et t_5. e) poids w_i à l'issue de l'estimation de mouvement pour l'instant t_5. f) poids w_i à l'instant t_{27}.*

un mouvement "conforme". Durant toute cette séquence, l'estimation et la détection sont perturbées par l'apparition d'ombres portées de taille importante sur l'arrière du fourgon (images 4.31a, 4.31b, 4.31d et 4.31f). Les figures 4.32e et 4.32f correspondant aux instants t_5 et t_{27}, qui représentent les poids w_i issus de l'estimation de mouvement, montrent la robustesse de notre algorithme d'estimation. Alors que, dans les deux cas, le support d'estimation comprend entièrement le fourgon, l'algorithme écarte les ombres du processus d'estimation (ombre à gauche à t_5, et à t_{27}, il s'agit de l'ombre d'un poteau qui se situe sur la gauche du fourgon dans l'image à l'instant t_{27} et sur la droite à l'instant t_{28} -d'où les deux marques noires verticales-). Cette robustesse est également mise en évidence sur la séquence compensée dont les figures 4.32a à 4.32d présentent quatre images. Le fourgon reste à peu près de la même taille sur toute la durée de cette séquence reconstruite, ce qui prouve la qualité du recalage effectué.

4.6 Conclusion

Nous avons décrit dans ce chapitre un nouvel algorithme pour détecter les régions dont le mouvement n'est pas conforme à un mouvement estimé. Cet algorithme peut donc être utilisé pour détecter des objets mobiles dans une séquence dans le cas d'une caméra en mouvement, et également pour suivre un objet mobile comme le montre le dernier exemple. Il est important de souligner également que ces objectifs d'analyse de scène dynamique sont atteints sans phase préalable de segmentation explicite en régions de mouvement différents. Notons que l'algorithme peut également être employer tel quel lorsque la caméra est fixe, ce qui n'est pas toujours le cas des schémas prévus pour le cas d'un capteur mobile [TP90].

Dans cet algorithme, des observations de mouvement adaptées à la détection sont intégrées au sein d'un schéma qui permet de prendre en compte au mieux l'information présente en chaque point de l'image. De plus, l'intégration temporelle de ces observations rend l'algorithme robuste vis-à-vis du bruit et de l'adéquation éventuellement imparfaite du modèle utilisé pour décrire le mouvement apparent de l'objet à suivre ou le mouvement induit par la caméra, comme le montrent les résultats expérimentaux et la comparaison avec une autre méthode.

Le schéma comprend un nombre certes important de paramètres à régler. Nous pensons cependant que cela ne constitue pas un réel handicap. Au contraire, cela représente plutôt une richesse car ainsi, l'algorithme peut s'adapter facilement à différents types de séquences, de qualités d'image, et de mouvements, dans la mesure où les valeurs des paramètres se fixent aisément pour une application donnée.

Rappelons qu'avec l'algorithme que nous proposons, nous détectons les régions dont le mouvement n'est pas conforme au mouvement estimé. Ainsi, les masques obtenus peuvent englober, en plus des objets mobiles de la scène, des entités statiques du monde tridimensionnel. Suivant le contexte, ceci ne constitue pas nécessairement une limitation

de la méthode. Si l'on considère l'exemple de la séquence ROND-POINT, dans une problématique de détection d'obstacle, il apparait tout aussi important de détecter l'entrée du rond-point (le marquage au sol) et le panneau de signalisation, que le véhicule qui passe sur ce rond-point. Néanmoins, si l'on souhaite distinguer plus finement ces différents obstacles, on doit avoir recours, par exemple, à une phase plus sophistiquée de segmentation.

Chapitre 5

Segmentation du mouvement apparent dans une séquence d'images

En analyse de séquences d'images, la segmentation au sens du mouvement, qui consiste à partitionner l'image en régions homogènes au sens d'un critère de mouvement donné, est un problème important. Elle intervient dès lors que la caméra est en mouvement et/ou que la scène contient un ou plusieurs objets mobiles. Une bonne segmentation, en évitant notamment le "mélange" de l'information aux frontières de mouvement, constitue un apport essentiel dans de nombreux domaines où il est généralement préférable de séparer la scène en ses différentes composantes cinématiques pour pouvoir entreprendre par exemple des phases de suivi, d'interprétation ou de reconstruction 3D, c'est-à-dire pour pouvoir mesurer ou identifier une grandeur sur des données homogènes. C'est le cas entre autres si l'on veut estimer un champ dense de mouvement [CST94], calculer le temps avant collision [MB92] ou de manière équivalente éviter des obstacles [NA89], interpréter qualitativement [BF93] ou quantitativement [NL92] le contenu dynamique d'une scène 3D. Dans [MB94a], une segmentation au sens du mouvement obtenue par la méthode décrite dans [BF93] sert de support à l'analyse et au recollement de trajectoires d'objets qui peuvent être occultés dans un nombre conséquent d'images. Cependant, malgré son importance, le problème de la segmentation (au sens de l'obtention d'une partition complète de l'image) n'a vraiment été abordé que dans peu de travaux de recherche en analyse de scènes dynamiques [Adi85, MB87, PR90, BF93, WK93]. Les premières études en vision dynamique, qui étaient surtout théoriques, supposaient généralement que la scène n'était constituée que d'un seul objet. Par la suite, l'effort a tout d'abord porté en premier lieu sur la mise au point d'algorithmes d'estimation du mouvement 3D plus efficaces et moins sensibles à l'incertitude sur le mouvement 2D mesuré effectivement dans une séquence réelle. L'étude de cette incertitude et de ses conséquences sur les solutions de diverses problématiques a par ailleurs fait l'objet d'analyses particulières [TK87, Adi89, YC90].

Enfin, dans les études actuelles, il semble que l'accent soit mis sur la robustesse des algorithmes vis-à-vis bien sûr de mesures erronées –par exemple, mauvaise mise en correspondance de points caractéristiques–, mais surtout vis-à-vis de mesures provenant d'une distribution minoritaire dans les données [LHZ89]. Ainsi par exemple, l'utilisation d'estimateurs robustes pour effectuer des tâches d'analyse de séquences d'images est devenue plus fréquente, et constitue en fait une alternative à la détermination d'une segmentation explicite. Néanmoins, cette approche n'est généralement pas suffisante pour obtenir une description cinématique de l'image entière.

Actuellement, le besoin d'une segmentation au sens du mouvement se fait également ressentir en codage d'image pour les codeurs dits de seconde génération. C'est donc dans ce domaine que se situent les principales recherches sur ce sujet [Die91, Sti93, CST94, AW94]. En effet, jusqu'à présent, la méthode de compensation de mouvement très souvent utilisée en codage (en particulier dans la recommandation MPEG 1 et la norme MPEG 2) consiste à choisir une partition en blocs de l'image et à estimer un modèle de mouvement constant à l'intérieur de chacun d'eux. Malgré les multiples améliorations apportées à ce schéma, comme par exemple l'utilisation de tailles de blocs adaptatives et de modèles de mouvement hiérarchiques [NL91], la qualité visuelle de l'image reconstruite au récepteur se détériore rapidement lorsque le débit devient très faible et les mouvements complexes. En fait, avec les très bas débits que l'on rencontre par exemple en visiophonie, il devient extrêmement difficile, voire illusoire, de pouvoir reconstruire exactement l'image enregistrée au codeur. Il peut s'avérer alors judicieux de décomposer l'image en objets ayant divers attributs, comme un contour, un jeu de paramètres de mouvements et de déformations, une texture. Ainsi, il est possible de hiérarchiser la qualité des données à transmettre non seulement entre les différents objets (focalisation), mais également entre les attributs d'un même objet, ce qui permet de s'adapter aux différents débits. En d'autres termes, il est possible d'introduire une distorsion importante dans l'image reconstruite, si l'on sait où et comment la placer. Par ailleurs, notons qu'en devenant plus sémantique, l'information à transmettre pourra également être plus concise. Dans ce contexte, la segmentation au sens du mouvement joue deux rôles essentiels. En premier lieu, elle est utile pour analyser la scène, l'interpréter, et en extraire les différents objets visuellement importants [Die91, NLO94]. Le choix du mouvement comme critère pour former les objets permet de limiter leur nombre, contrairement aux approches qui se basent sur une segmentation spatiale de l'image [PS94]. En second lieu, la segmentation au sens du mouvement constitue en elle-même une méthode de compensation directement adaptée au codage. Avec l'approche par objets ou régions, l'effet de bloc observé en utilisant les schémas traditionnels se trouve considérablement réduit. Dans ce chapitre, nous essayerons de montrer, notamment dans la partie résultat, en quel sens l'algorithme de segmentation que nous avons développé est susceptible de jouer ces deux rôles.

Dans la partie qui va suivre, nous décrirons l'approche que nous avons retenue. Ensuite, l'algorithme mis en œuvre sera décrit plus en détails, avant de présenter les résultats obtenus sur des séquences réelles.

5.1 Approche retenue pour la segmentation du mouvement

La segmentation du mouvement constitue une étape supplémentaire dans l'analyse d'une séquence d'images lorsque l'information fournie par la détection du mouvement devient trop ambiguë, difficilement interprétable, ou tout simplement insuffisante pour la tâche fixée. Considérons l'algorithme décrit dans le chapitre précédent. Il est basé sur l'hypothèse selon laquelle le mouvement dominant, censé correspondre aux zones statiques de la scène, peut être décrit par un modèle polynomial. Cette hypothèse sera valide dans les cas particuliers importants suivants:

- le mouvement de la caméra est purement rotationnel;
- le mouvement de la caméra comporte une composante translationnelle, mais la partie statique de la scène est approximativement plane et/ou est éloignée de la caméra.

Si ces conditions ne sont pas respectées, le mouvement dominant estimé peut ne correspondre qu'à une seule partie des régions statiques, comme nous l'avons vu avec l'exemple ROND-POINT, ou à aucune région en particulier si la structure de la scène est trop complexe. Par ailleurs, une même région connexe de la carte des régions détectées peut correspondre à plusieurs entités mobiles, lorsque ces dernières sont de taille importante, nombreuses, ou se croisent. Dans ces conditions, les régions détectées non-conformes seront difficiles à analyser et à interpréter dans une phase ultérieure.

Une première possibilité pour lever les ambiguïtés et pour effectuer la segmentation consiste à décomposer la scène en procédant par éliminations successives [IRP92, BBH*89, WK93, Die91]. Le modèle de mouvement dominant Θ_0 est tout d'abord estimé dans la séquence. Les régions R_0 dont le mouvement est conforme à ce modèle sont déterminées, retirées de l'image, et le processus reprend sur les zones restantes. On obtient alors le schéma de segmentation hiérarchique présenté à la figure 5.1. Notons qu'éventuellement l'extraction du mouvement secondaire Θ_1 (et des suivants) peut se faire non pas sur $\overline{R_0}$ tout entier ($\overline{R_0}$ désignant le complément de R_0), mais simplement sur la composante connexe la plus importante de $\overline{R_0}$. L'avantage d'une telle approche, qui est une conséquence du schéma hiérarchique, est qu'il est possible d'arrêter l'analyse de la scène à tout instant, en fonction du nombre de composantes de mouvement déjà extraites, ou si les régions restantes n'ont pas de mouvement cohérent. De plus, si une région R_i, une fois détectée et déterminée, peut-être suivie sans l'intervention de la phase de segmentation hiérarchique, comme il est présenté sur la figure 5.1, il sera possible de ne considérer et de ne suivre qu'une seule région dont des critères prédéfinis (forme, intensité, mouvement, etc.) auront montré l'intérêt de le faire.

Cependant, un effet pervers de cette approche réside dans le fait que, au niveau i de la hiérarchie, toutes les régions de l'image (n'appartenant pas déjà à une région de mouvement précédente de cette hiérarchie), dont le mouvement est représentable par Θ_i sont

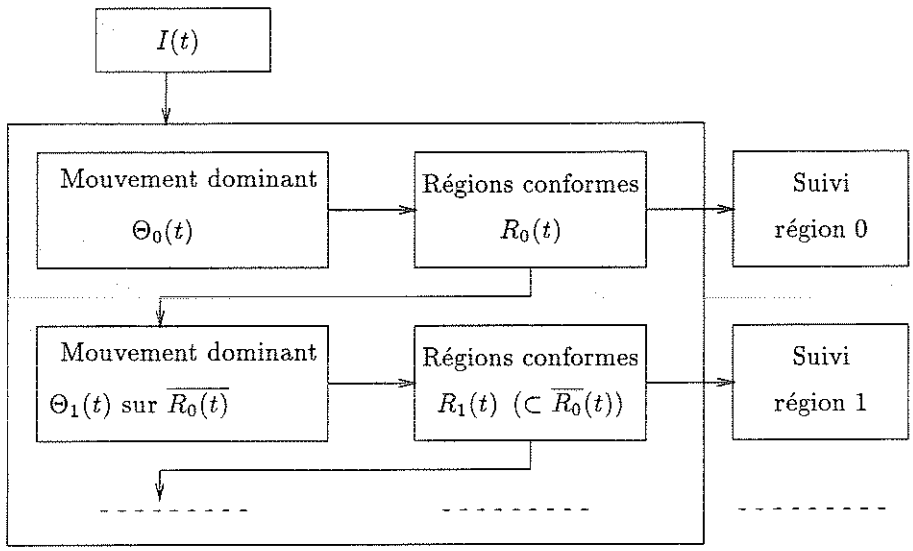


FIG. 5.1 - *Segmentation hiérarchique du mouvement dans une séquence d'images.* $\overline{R_0}(t)$ *représente le complément de* $R_0(t)$.

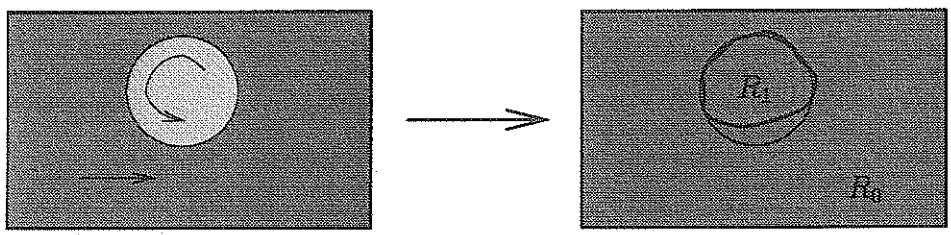


FIG. 5.2 - *Segmentation hiérarchique du mouvement. On suppose que la translation est le premier modèle estimé (modèle 0). Les régions conformes à ce mouvement,* R_0, *incluent également une partie du disque.*

attribuées à la région R_i. Celle-ci contiendra donc la région "particulière" dont Θ_i permet de décrire le mouvement global, et également dans certains cas, des parties d'autres régions dont le mouvement est représentable par un modèle différent. Ce cas est illustré sur la figure 5.2, dans laquelle un disque tourne devant un fond qui se déplace vers la droite. Autour de la partie inférieure de ce disque, les mouvements sont similaires. Si par exemple la translation est estimée en premier, les régions conformes à ce modèle ont toutes les chances d'inclure également le bas du disque. Ainsi, avec ce schéma de segmentation, deux modèles de mouvement estimés distincts ne doivent pas être en mesure de décrire le mouvement d'une même zone. Dans le cas contraire, la zone en question pourrait être (définitivement) affectée au mauvais modèle de mouvement, suivant l'ordre dans lequel les deux modèles concurrents sont placés dans la hiérarchie d'identification des modèles. De telles situations se rencontrent fréquemment lorsque l'on traite uniquement deux images

consécutives, mais sont plus rares au fur et à mesure que l'on augmente le support temporel de l'analyse [AD93, ASB94]. Néanmoins, elles subsisteront généralement aux frontières entre deux régions qui correspondent à la projection d'une même surface 3D continue de la scène mais dont l'orientation varie abruptement (par exemple, deux plans jointifs formant un angle droit). Cette formulation de la segmentation ne permet pas de choisir le modèle de mouvement qui décrit le mieux le déplacement d'une région donnée de l'image, mais conduit à prendre le premier dans la hiérarchie qui le décrit correctement, au sens d'un critère prédéfini. Ainsi, les différents modèles de mouvement ne sont pas traités de manière équivalente. Par ailleurs, il est important de souligner que cette méthode ne permet pas d'assurer l'obtention d'une partition complète de l'image en régions de mouvement homogène[1], ce qui est une limitation forte dans certaines applications telles que le codage. Lorsque celle-ci est requise, il est alors préférable d'utiliser une formulation comme celle décrite ci-dessous.

Bien que l'algorithme que nous avons décrit dans le chapitre précédent sur la détection du mouvement eût pu s'intégrer dans le schéma hiérarchique, nous préférons poser le problème de la segmentation sous la forme de l'estimation conjointe des modèles de mouvement $\{\Theta_k\}_{k\in\{1,\ldots,N_r\}}$ et de la partition associée à l'instant t: $\{R_k^t\}_{k\in\{1,\ldots,N_r\}}$. Le nombre N_r de régions de l'image est également à estimer. Il s'agit d'une formulation pertinente de la segmentation qui a déjà été utilisée plusieurs fois fois [MB87, DP91, BF93, CST94, AW94]. Notons par ailleurs que si l'on considère le modèle quadratique à huit paramètres (formule 3.3), on peut ainsi obtenir directement une interprétation de la scène en termes de facettes planaires animées de mouvements rigides. Cependant, l'estimation fiable des modèles de mouvement étant directement liée à la pertinence de la partition, et réciproquement, la partition optimale en régions dépendant des paramètres de mouvement courants, l'estimation jointe de $\{\Theta_k,\ R_k^t\}$ est très difficile. Toutes les méthodes précédemment citées procèdent donc en deux étapes naturelles, présentées à la figure 5.3, qui sont itérées jusqu'à convergence. La première consiste à estimer les paramètres de mouvement (en général à partir des gradients spatio-temporels de la fonction intensité), la partition en régions étant figée; la seconde revient à déterminer la carte de segmentation optimale, les modèles Θ_k restant fixés. Or, ce schéma itératif entre les deux procédures peut s'avérer long et coûteux d'un point de vue calculatoire, surtout lorsqu'une procédure de minimisation stochastique est utilisée [MB87], et peut conduire à des segmentation de qualité moyenne. En effet, considérons le scénario suivant que nous avons pu observer. Supposons qu'à l'instant t nous disposions d'une segmentation correcte, et qu'aux instants suivant apparaisse progressivement un objet dans la scène. Au départ, étant de taille réduite, il risque d'être fusionné avec l'une des régions existantes, même s'il perturbe "légèrement" l'estimation du mouvement de cette région. Au fur et à mesure de l'acquisition de nou-

1. Il faudrait, pour avoir une partition complète, classer dans une même région tous les pixels non affectés à une région R_i.

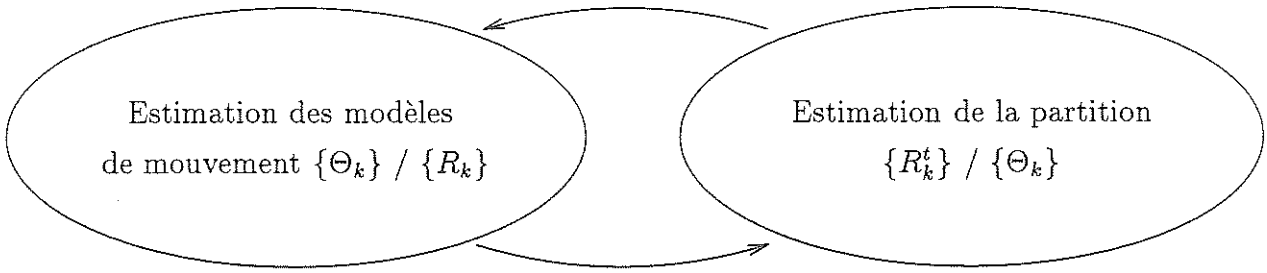


FIG. 5.3 - *Segmentation au sens du mouvement "classiquement" utilisée: Estimation conjointe des modèles de mouvement et de la partition en régions. Les étapes sont généralement itérées jusqu'à ce que le nombre de pixels qui changent d'étiquette dans la carte de segmentation soit faible.*

velles images et de l'accroissement en taille du nouvel élément, estimation et segmentation pourront s'influencer mutuellement pour garder l'objet au sein de la région dans laquelle il s'est retrouvé initialement classé. Plus généralement, il peut se produire les deux classes d'erreurs suivantes. La segmentation converge vers une solution instable et fragmentée, donc inexploitable; ou bien, comme cela était illustré précédemment, la segmentation arrive à fournir une partition apparemment cohérente, mais en fait incorrecte si le modèle de mouvement peut "absorber" un mélange de mouvements donnés.

En fait, les situations précédentes sont susceptibles de se produire si l'une des conditions suivantes est vérifiée:

1. l'estimateur de mouvement n'est pas robuste (estimation suivant les moindres-carrés par exemple), et peut donc être perturbé par la mise en défaut localement de l'hypothèse de conservation de l'intensité, ou par la présence d'un objet secondaire dans la région, [MB87, BF93, CST94, AW94].

2. aucune procédure particulière n'est prévue pour détecter explicitement l'apparition de nouvelles régions, ou plus généralement, pour vérifier l'adéquation du modèle [MB87, DP91, CST94] avec les déplacements dans l'image des pixels de la région auquel il est associé. Ceci signifie implicitement que le nombre de régions n'est pas déterminé automatiquement, ce qui limite donc l'intérêt de l'approche.

Les chapitres sur l'estimation robuste et la détection du mouvement contiennent les réponses à ces deux difficultés, et constituent donc deux phases importantes de la méthode de segmentation que nous avons adoptée. En remarquant que la partition recherchée $\{R_k, \Theta_k\}$ modélise en fait un champ dense de vitesses dans l'image (discontinu ou non), l'idée générale de notre approche sera d'en estimer un (par l'intermédiaire de la segmentation) qui constitue une approximation du mouvement apparent réel, *à la précision* δ_{seg} *près*. Pour atteindre ce but, nous utiliserons le même concept d'erreur résiduelle de compensation introduit dans le chapitre précédent.

Le problème de la segmentation du mouvement sera également traité comme un problème d'étiquetage statistique dans un cadre Bayesien. La notion de régularisation statistique introduite sera à nouveau formalisée à l'aide de modèles Markoviens et donc concrétisée par la définition d'une fonction d'énergie appropriée.
Les observations seront semblables à celles que nous avons utilisées dans le schéma de détection du mouvement. Avant de décrire plus en détail cette phase particulière, examinons tout d'abord le synoptique complet du schéma de segmentation que nous avons retenu, et qui est représenté à la figure 5.4.

Nous cherchons donc à l'instant t la carte de segmentation que nous noterons $e(t)$ définie sur l'ensemble des sites de l'image (en pratique, les pixels) et dont les valeurs correspondent aux numéros où étiquettes k des régions R_k^t.

De plus, il s'agira de même d'estimer l'ensemble des modèles de mouvement correspondant $\Xi_t^{t+1} = \{(\Theta_k)_t^{t+1}\}$, en reprenant la notation définie au début du chapitre 4. Supposons alors que nous disposons de la carte de segmentation $e(t-1)$ et des paramètres de mouvement estimés $\widehat{\Xi}_{t-1}^t$. Les différentes phases de l'algorithme sont alors les suivantes:

Prédiction de la segmentation

La première opération à effectuer concerne la prédiction de la carte initiale de segmentation à l'instant t, $e^{\text{init}}(t)$. Cette première étape réalise la liaison temporelle entre les partitions successives de l'image, et assure par là-même qu'une région de mouvement cohérent tout au long de la séquence reste étiquetée de la même façon. Pour chaque point p d'une région R_k^{t-1}, nous avons propagé l'étiquette k de la région aux points de la grille d'échantillonnage situés autour de $\mathrm{p} + \vec{V}_{(\Theta_k)_{t-1}^t}(\mathrm{p})$. À l'issue de cette phase, un point de la grille qui n'a pas reçu d'étiquette correspond à une région découverte, et un point qui en a reçu plusieurs, à une région recouverte ou d'occlusion. La figure 5.5b présente le résultat de la prédiction appliquée à la segmentation de la figure 5.5a.

Estimation robuste des modèles de mouvement pour les régions existantes

Les paramètres $(\Theta_k)_t^{t+1}$ sont alors estimés sur cette carte prédite. Comme nous employons un estimateur robuste, une imprécision sur la prédiction provenant par exemple d'une mauvaise segmentation à l'instant précédent, ou de l'apparition d'une nouvelle région, n'affecteront pas l'estimation des paramètres de mouvement. Il est important de souligner ici que ces paramètres ne seront plus modifiés par la suite. Ceci permet d'éviter les nombreuses itérations des méthodes basées sur le schéma présenté à la figure 5.3, et dont nous avons évoqué les inconvénients plus haut.

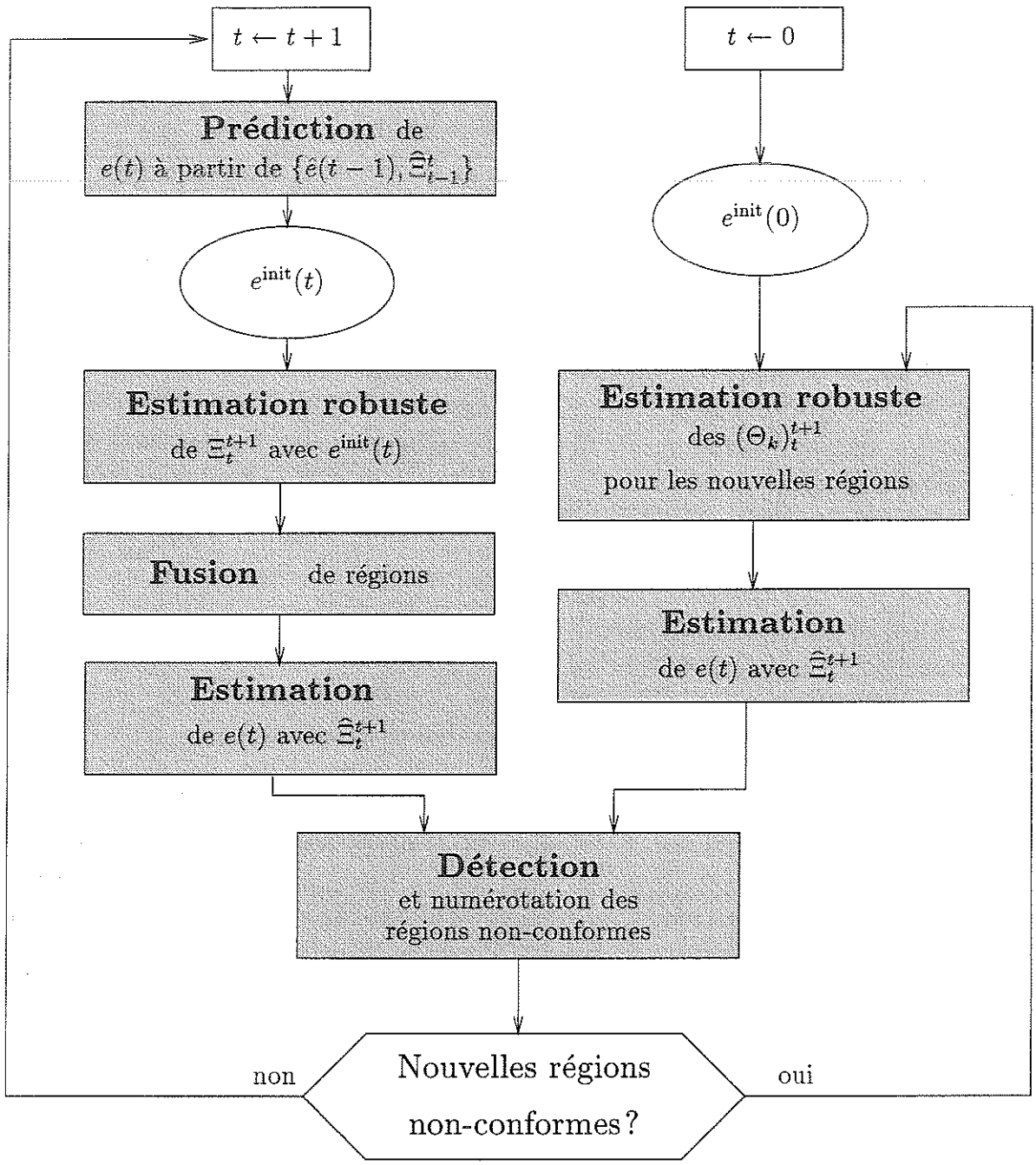


FIG. 5.4 - *Schéma de segmentation complet adopté. $e(t)$ désigne la carte de segmentation à l'instant t, et Θ_t^{t+1} est l'ensemble des modèles de mouvement associé à $e(t)$ décrivant le champ des déplacements entre t et $t+1$.*

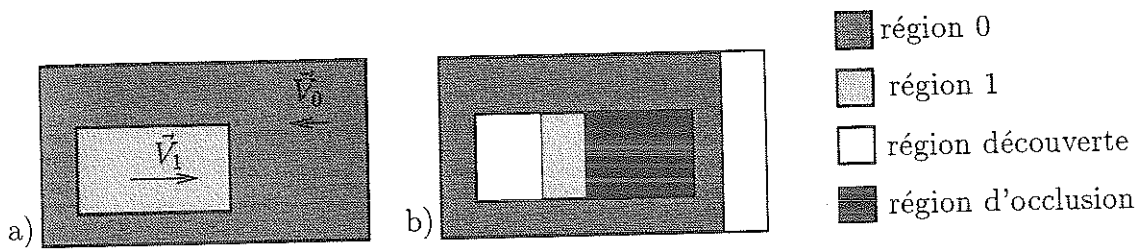


FIG. 5.5 - *Exemple de carte de prédiction obtenue (figure b)) à l'aide de la segmentation à l'instant précédent a).*

Fusion des régions de mouvement redondant

A l'issue de l'étape précédente, un module permet de fusionner de manière très simple les régions ayant des mouvements semblables. Ce module repose uniquement sur la connaissance des modèles de mouvement et des "rectangles" englobant chacune des régions. Le principe en est le suivant. On souhaite savoir si le modèle $\widehat{\Theta}_k$ peut remplacer le modèle $\widehat{\Theta}_l$ pour décrire le mouvement de la région R_l^t. Nous utilisons pour cela la norme que nous avons introduite dans le chapitre sur l'estimation (voir la formule (3.20) page 44). Puisque $\|\widehat{\Theta}_k - \widehat{\Theta}_l\|_{R_l^t}$ représente une borne maximale sur l'erreur moyenne (en module) des vecteurs vitesses à l'intérieur du "rectangle" englobant R_l^t, commise en utilisant $\widehat{\Theta}_k$ au lieu de $\widehat{\Theta}_l$, nous fusionnerons la région l à la région k si:

$$\|\widehat{\Theta}_k - \widehat{\Theta}_l\|_{R_l^t} \leq \delta_{\text{seg}} \tag{5.1}$$

où δ_{seg} est l'erreur que l'on s'autorise sur la description du champ des vitesses sous-jacent. Lorsque le test est validé, le nombre N_r de régions de la segmentation est réactualisé, et le modèle de mouvement de la "nouvelle" région k est estimé. En pratique, ce test n'est effectué que pour des régions dont les rectangles englobants ont une intersection non vide. De plus, dans la mesure où les autres étapes de l'algorithme produisent une segmentation où généralement chaque région a son mouvement propre, les cas où des régions sont fusionnées sont limités, et correspondent à des situations bien particulières que nous évoquons ci-dessous.

La phase de fusion permet en fait de palier deux problèmes que nous avons rencontrés et qui génèrent une sur-segmentation de l'image. Le premier, d'ordre algorithmique, vient du fait qu'à un instant donné, l'étape de création de régions qui sera décrite plus loin peut étiqueter différemment deux nouvelles régions correspondant à un même objet. Ceci se produit par exemple lorsqu'un objet non-convexe apparait dans la scène, comme dans la séquence ROND-POINT, où dans un premier temps deux zones de la projection de cet objet dans l'image peuvent ne pas apparaître liées. La coexistence de ces deux régions par la suite n'est alors pas souhaitable dans la mesure où leurs mouvements sont "redondants" et qu'elles forment en fait un même objet connexe. Le second problème a pour origine directe la variabilité de la complexité des mouvements au cours d'une séquence. En effet, à

certains moments d'une séquence, le mouvement d'une région, correspondant par exemple à celui d'une personne, peut devenir plus compliqué. Comme nous le verrons, l'algorithme crée alors de nouvelles régions pour pouvoir tenir compte de cette complexité (en fonction bien sûr de la nature du modèle de mouvement choisi). Si cette phase s'achève est que l'on revient à un mouvement global plus "simple", ces régions pourront subsister, alors qu'un nombre beaucoup plus restreint de régions (et donc de modèles de mouvement) suffirait à décrire le mouvement de l'objet.

Notons que pour éviter des fusions intempestives, nous aurions pu attendre que lorsque le critère défini ci-dessus est validé à un instant donné, il soit confirmé dans les instants suivants avant de prendre effectivement la décision de fusionner les régions. Par ailleurs des tests plus performants pourraient être mis en place, par exemple si les paramètres de mouvement étaient filtrés à l'aide d'un filtrage de Kalman, comme dans [Mey93], ou en exprimant un critère s'appuyant sur les données observées, c'est-à-dire les gradients spatio-temporels de l'image, par l'utilisation de tests de vraisemblance comme dans [BS87] [2].

Phase de segmentation proprement dite

À l'issue de la fusion de régions, la carte de segmentation $e^{\text{init}}(t)$ est alors mise à jour, à l'aide des modèles de mouvement préalablement estimés, suivant une procédure de relaxation que nous décrirons dans la section suivante. Au cours de cette phase, les régions déclarées découvertes ou recouvertes lors de la phase de prédiction sont censées recevoir les étiquettes qui conviennent, et les frontières entre deux régions k et k' s'ajustent en fonction des nouvelles estimation des modèles de mouvement. La définition de potentiels d'énergie permettant de réaliser une segmentation de l'image de bonne qualité est d'une importance capitale et sera abordée dans la section suivante. Notons cependant ici que lors de cette relaxation, les régions dont le nombre de pixels devient inférieur à un seuil (faible) N_{elim} sont éliminées. En effet, celles-ci correspondent généralement à des régions dont le mouvement est soit mal estimé, soit ne contribue plus que de manière marginale à la description du champ des déplacements. Fréquemment, ces régions disparaissent d'elles mêmes du fait de la présence d'un terme de régularisation spatiale dans la relaxation.

A l'issue de cette phase, le partitionnement de l'image en régions de mouvement homogènes, pour le jeux de modèles de mouvement intervenant dans la segmentation précédente, est effectuée. Il est important de noter que, dans la mesure où les modèles de mouvement ne sont plus estimés par la suite, les frontières entre ces régions sont établies lors de cette phase, et ne seront donc pas modifiées ensuite. Les étapes suivantes de l'algorithme servent alors à tester:

1. si des nouveaux objets mobiles sont apparus dans la scène;
2. si le nombre courant de modèles de mouvement introduits est toujours suffisant pour décrire correctement le mouvement apparent dans l'image.

2. Dans ce dernier cas, le coût calcul est beaucoup plus important.

Les éventuelles phases de relaxation supplémentaires ne contribueront alors qu'à la détermination des frontières séparant ces régions de celles qui pourront être nouvellement créées.

Détection des zones non-conformes pour chaque région k

À l'intérieur de chaque région R_k^t, les zones de mouvement non-conformes au modèle $(\widehat{\Theta}_k)_t^{t+1}$ sont détectées à l'aide d'un schéma de relaxation très similaire à celui présenté dans le chapitre précédent sur la détection de mouvement. Dans celui-ci, seuls les termes d'énergie U_1 et U_2, c'est-à-dire ceux liés aux observations traduisant l'aptitude du modèle $(\widehat{\Theta}_k)_t^{t+1}$ à décrire le mouvement des pixels concernés, et le terme de régularisation spatiale, sont utilisés. La réunion des sites décidés non-conformes au sein de R_k^t est notée Z_k^{nc}.

Création de nouvelles régions

Cette étape, avec la précédente, est cruciale. Tout d'abord, elle est nécessaire pour extraire de la toute première image de la séquence traitée les différentes entités mobiles présentes dans l'image. D'autre part, elle permet à l'algorithme de s'adapter aux évolutions "brutales" du contenu dynamique des nouvelles images. Plus précisément, l'ensemble des zones Z_k^{nc} de l'image où les modèles de mouvement existants ne rendent pas compte de manière satisfaisante du mouvement apparent réel, sont regroupées au sein d'une même région R_{nc}. Ces zones peuvent avoir pour origine l'apparition de nouveaux objets dans la scène. Elles peuvent également indiquer comme nous l'avons évoqué plus haut, que le (ou les) modèle(s) de mouvement(s) d'un objet[3] n'est (ne sont) plus suffisant(s) pour décrire le mouvement apparent à l'intérieur de celui-ci, car son mouvement devient plus complexe. Dans les deux cas, nous devons créer de nouvelles régions.

Les composantes connexes de R_{nc} sont alors extraites. Celles dont le nombre de pixels dépasse N_{nc} sont numérotées, et le nombre N_r est remis à jour en conséquence. Les autres régions sont alors considérées comme négligeables et rejetées. Rappelons ici que, si deux zones non-conformes correspondant à un seul objet sont numérotées différemment, elles seront vraisemblablement regroupées par la suite au sein d'une même entité grâce à la phase de fusion décrite précédemment. Si aucune nouvelle région n'a été générée, la carte $e(t)$ obtenue lors de la dernière relaxation est considérée comme la segmentation finale de l'algorithme. Dans le cas contraire, les modèles de mouvement correspondant à l'ensemble des régions créées à l'instant t sont estimés selon l'approche robuste. Puis la carte de segmentation est mise à jour par une nouvelle phase de relaxation, faisant intervenir le nouvel ensemble $\widehat{\Xi}_t^{t+1}$ des modèles de mouvement. Le processus est alors itéré tant que des régions sont créées, et qu'un nombre d'itérations prédéfini n'a pas été atteint. Dans la

3. Par objet, nous entendons ici une partie de la scène dont le mouvement apparent est continu à l'intérieur de sa projection dans l'image. À l'évidence, malgré cette continuité, plusieurs modèles de mouvement peuvent être nécessaires pour décrire son champ des vitesses.

plupart des cas, une à deux itérations sont au plus nécessaires, excepté pour la première image de la séquence. Soulignons que les calculs à effectuer à chaque passage dans la boucle sont peu nombreux, se résumant à l'estimation de quelques modèles de mouvement sur des régions assez petites généralement, et à une relaxation n'impliquant de nouveaux calculs quant à l'évaluation des variations locales de la fonction d'énergie que dans le voisinage de ces régions.

Initialisation

La phase d'initialisation, c'est à dire celle correspondant à la recherche de la segmentation entre les deux premières images de la séquence, diffère peu du déroulement normal de l'algorithme. Nous partons d'une carte de segmentation initiale $e^{\text{init}}(0)$. Dans toute les expériences, nous avons considéré que nous partions d'une seule région, soit l'image complète. Nous pourrions également utiliser une partition initiale de l'image sous forme de blocs si la scène s'avérait être vraiment très complexe. Cependant, quel que soit le choix effectué, les régions de cette segmentation initiale seront toujours considérées comme étant nouvellement créées, ce qui permettra au cours des boucles suivantes d'affiner l'estimation de leurs modèles de mouvement. Le mécanisme d'estimation des modèles de mouvement, d'optimisation par relaxation déterministe de la fonction d'énergie définie, de détection des zones non-conformes, et de création de nouvelles régions, que nous avons décrit précédemment, permet alors d'obtenir une segmentation initiale de bonne qualité.

5.2 Segmentation de l'image en régions de mouvement homogène

Dans cette section, nous décrivons plus particulièrement l'étape appelée "phase de segmentation proprement dite" de la méthode de segmentation de l'image, phase effectuée à l'aide de modèles de mouvements préalablement estimés. Il s'agit d'une étape fondamentale de l'algorithme sur laquelle reposent toutes les autres phases du schéma général. Celle-ci correspond à un processus d'étiquetage des pixels de l'image, dans lequel nous introduisons des contraintes d'homogénéité d'ordre spatial et temporel pour obtenir une partition de l'image en régions régulières, compactes, et de taille suffisante. Comme nous l'avons déjà indiqué, nous utiliserons le formalisme markovien qui permet d'intégrer aisément les contraintes liées aux observations et à la régularisation.

Cependant, notre but ne sera pas nécessairement en un pixel donné de déterminer les paramètres de mouvement qui modélisent le mieux le déplacement de ce pixel, mais de trouver un jeu de paramètres qui permet de décrire ce déplacement correctement, c'est-à-dire *à une erreur δ_{seg} près*. Ainsi, si en un site plusieurs modèles de mouvement satisfont cette condition "qualitative", celui-ci sera étiqueté en fonction de l'homogénéité locale des régions associées à ces modèles.

Précisons maintenant les notations et le problème. Nous disposons donc de N_r modèles de mouvement estimés $\widehat{\Xi}_t^{t+1} = \{(\widehat{\Theta}_k)_t^{t+1}\}_{k\in\{1,\ldots,N_r\}}$. Nous souhaitons alors partitionner l'image en N_r régions, c'est-à-dire obtenir à l'instant t la carte d'étiquettes de segmentation $e(t) = \{e_s(t), s \in S\}$ où les sites s de notre grille sont les pixels de l'image, et les étiquettes e_s sont à valeur dans l'ensemble $\Lambda_{\text{seg}} = \{1, \ldots, N_r\}$ des numéros de régions. Il nous reste à définir le champ des observations. Celui-ci sera constitué des images aux instants t et $t+1$, soit l'ensemble d'observation $o(t) = \{I_t, I_{t+1}\}$ et par extension $\tilde{e}(t) = \{\tilde{e}_s(t), s \in S\}$ la carte issue de la projection dans le sens du mouvement de la segmentation à l'instant $t-1$[4], qui inclue donc des étiquettes à valeur dans Λ_{seg} auquelles s'ajoutent les deux étiquettes correspondant aux régions découvertes et recouvertes.

Nous conserverons ici pour caractériser l'adéquation des étiquettes avec les observations, c'est-à-dire les cartes d'intensité, la notion de compensation et d'erreur résiduelle introduite dans le chapitre sur la détection du mouvement. Nous définirons alors en chaque site s pour chaque modèle de mouvement k, les quantités suivantes:

$$\epsilon_s(k) = \text{Mes}_{(\widehat{\Theta}_k)_t^{t+1}}(s) \tag{5.2}$$

où l'expression $\text{Mes}_\bullet$ est donnée par la formule (4.19) page 91. Par la suite, nous omettrons les indices et exposants des modèles $(\widehat{\Theta}_k)_t^{t+1}$ lorsqu'il n'y aura pas d'ambiguïté, de même que la variable t dans $e(t)$, $o(t)$ et $\tilde{e}(t)$.

En utilisant la formulation markovienne, le problème que nous traitons se ramène à la minimisation d'une énergie U_{seg} qui se décomposera en trois termes, le premier $(U_{2\text{seg}})$ correspondant à la modélisation *a priori* des étiquettes exprime l'homogénéité spatiale attendue des régions, les deux autres ($U_{1\text{seg}}$ et $U_{3\text{seg}}$) caractérisent l'adéquation des étiquettes aux observations (au sens large):

$$U_{\text{seg}}(e, o, \tilde{e}) = U_{1\text{seg}}(e, o) + U_{2\text{seg}}(e) + U_{3\text{seg}}(e, \tilde{e}) \tag{5.3}$$

5.2.1 Définition de l'énergie liant étiquettes et observations de mouvement

Le terme d'énergie $U_{1\text{seg}}$ est décomposé en une somme de potentiels locaux comme suit:

$$U_{1\text{seg}}(e, o) = \sum_{s\in S} V_{1\text{seg}}(e_s, o) \tag{5.4}$$

Cette décomposition, qui ne prend pas en compte la dépendance d'une observation en un site s vis-à-vis des étiquettes voisines de s, souffrira des mêmes défauts aux frontières entre les régions de la partition que celle que nous avons utilisée pour le terme d'énergie U_1 de la méthode de détection du mouvement (voir discussion page 96). Le potentiel $V_{1\text{seg}}$

4. $\tilde{e}$ est noté $e^{\text{init}}(t)$ dans le schéma général de la segmentation (figure 5.4).

doit indiquer dans quelle mesure l'étiquette e_s, par l'intermédiaire du modèle $\widehat{\Theta}_{e_s}$ qu'elle désigne, permet d'obtenir une bonne description de la vitesses apparente $\vec{V}_{\text{réel}}(s)$ au site s. Plus précisément, est-ce-que le vecteur $\vec{V}_{\widehat{\Theta}_{e_s}}(s)$ constitue une bonne approximation de $\vec{V}_{\text{réel}}(s)$? Ceci nous renvoie directement au chapitre sur la détection du mouvement, et à l'utilisation des bornes l_s et L_s sur les observations, qui ne dépendent que des gradients spatiaux de l'intensité et de l'erreur maximale δ_{seg} que l'on tolère sur le mouvement résiduel entre $\vec{V}_{\text{réel}}(s)$ et $\vec{V}_{\widehat{\Theta}_{e_s}}(s)$. Par la suite, bien que ces bornes dépendent du paramètre δ_{seg} –qui pourrait être différent du paramètre δ utilisé en détection du mouvement–, nous continuerons à les désigner par l_s et L_s. Ainsi, en nous basant sur la signification de ces bornes, nous souhaitons que le potentiel $V_{1\text{seg}}$ soit minimal lorsque la quantité $\epsilon_s(e_s)$ est inférieure à l_s (le déplacement résiduel est inférieur à δ_{seg}), maximal lorsque $\epsilon_s(e_s)$ est supérieure à L_s (le déplacement résiduel est alors supérieur à δ_{seg}), et moyenne lorsque $\epsilon_s(e_s)$ se situe entre les deux bornes. Une telle fonction peut alors se décrire à l'aide du potentiel V_1 déjà défini pour la détection du mouvement (formule 4.24), selon:

$$V_{1\text{seg}}(e_s, o) = V_1(\epsilon_s(e_s), c) - V_1(\epsilon_s(e_s), nc) \tag{5.5}$$

où c et nc indiquent respectivement un modèle de mouvement conforme ou non-conforme aux observations. Après développement, nous obtenons:

$$V_{1\text{seg}}(e_s, \epsilon_s) = F_s \times [\, \alpha_{c\text{seg}} \times A_{l_s, k_{c\text{seg}}}(\epsilon_s(e_s)) - \alpha_{nc\text{seg}} \times A_{L_s, k_{nc\text{seg}}}(\epsilon_s(e_s)) + \alpha_{nc\text{seg}} \,] \tag{5.6}$$

où la fonction $A_{t,k}$ est une fonction échelon "adoucie" (voir exemples page 98). Les différents paramètres intervenant dans ce potentiel correspondent à ceux décrits dans le chapitre sur la détection du mouvement: $\alpha_{c\text{seg}}$ et $\alpha_{nc\text{seg}}$ sont des facteurs d'amplitude, $k_{c\text{seg}}$ et $k_{nc\text{seg}}$ sont des coefficients modélisant la rapidité de la transition de la fonction A autour de l_s et L_s. F_s quant à lui est un coefficient d'atténuation relié à la présence de gradient spatial d'intensité dans le voisinage de s; il indique en quelque sorte la quantité d'information fournie par les observations. Les potentiels ainsi définis (en fonction de l_s et L_s) correspondent alors directement aux courbes présentées à la figure 4.14 et traduisent bien l'aspect qualitatif attendu.

5.2.2 Définition du terme de régularisation

Le champ des étiquettes étant supposé markovien relativement à un système de voisinage d'ordre 2 (8-voisinage), l'énergie $U_{2\text{seg}}$ se décompose en une somme de potentiels sur les cliques engendrées par ce voisinage:

$$U_{2\text{seg}}(e) = \sum_{c \in \mathcal{C}} V_c(e) \tag{5.7}$$

Ce terme sert à modéliser l'information *a priori* sur les étiquettes. Il peut servir par exemple à favoriser la présence de régions (focalisation) en fonction d'indices obtenus à

partir des résultats d'analyse de ces régions dans le passé. Ainsi, on peut souhaiter favoriser la conservation des régions qui existent depuis une durée importante (et ainsi éviter qu'elle ne disparaisse si le mouvement est momentanément mal estimé) plutôt que celles qui n'ont été créées que récemment. Ou bien, dans une problématique de détection d'obstacles, privilégier les régions dont le mouvement indique un rapprochement par rapport à la caméra. En l'absence d'indices particuliers, nous modéliserons $U_{2\text{seg}}$ par:

$$U_{2\text{seg}}(e) = \sum_{\{s,u\} \in C} V_{2\text{seg}}(e_s, e_u) \tag{5.8}$$

avec

$$V_{2\text{seg}}(e_s, e_u) = \beta_{d\text{seg}}(1 - \delta_{e_u = e_s}) = \begin{cases} 0 & \text{si } e_u = e_s \\ \beta_{d\text{seg}} & \text{si } e_u \neq e_s \end{cases} \tag{5.9}$$

où δ désigne ici le symbole de Kronecker, et $\beta_{d\text{seg}}$ est un paramètre positif. Ce potentiel favorise l'homogénéité du champ d'étiquettes, en décourageant les configurations d'étiquettes voisines différentes. Les potentiels ainsi définis sont simples et ne mettent en jeu qu'un seul paramètre. De plus, d'un point de vue calculatoire, ils n'impliquent localement que le dénombrement des voisins appartenant aux autres régions. Il serait bien sûr possible de faire appel à des étiquettes de frontières en introduisant par exemple des processus-ligne (*"line process"*) [GG84]. Cette politique conduite dans [MB87], permet de favoriser des configurations particulières de frontière (par exemple des contours rectilignes), mais au prix d'une complexité beaucoup plus élevée. En effet, ces processus lignes sont des nouvelles variables à estimer qu'il convient donc d'introduire dans la phase d'optimisation. De plus, il faut déterminer des valeurs de potentiel adéquates pour chaque configuration, ce qui n'est pas aisé. Enfin, lors de la minimisation de la fonction d'énergie, l'identification des configurations impliquent des calculs locaux plus complexes qu'un simple comptage de sites.

5.2.3 Définition de l'énergie de conservation temporelle de la segmentation

Les fonctions de potentiel $V_{1\text{seg}}$ et $V_{2\text{seg}}$ définies précédemment permettent d'effectuer la segmentation du mouvement en régions entre deux images. Comme dans le cas de la détection du mouvement, il semble raisonnable d'exploiter la cohérence temporelle entre les projections d'une même surface de la scène dans les différentes images de la séquence. Nous utilisons à cet effet la carte $\tilde{e}$ prédite à partir de la segmentation finale à l'instant précédent, et définissons un terme d'énergie "conservateur":

$$U_{3\text{seg}} = \sum_{s \in S} V_{3\text{seg}}(\tilde{e}_s, e_s) \tag{5.10}$$

où $V_{3\text{seg}}$ s'exprime de manière similaire à $V_{2\text{seg}}$ par:

$$V_{3\text{seg}}(\tilde{e}_s, e_s) = F_s \times \beta_{td\text{seg}}(1 - \delta_{\tilde{e}_s = e_s}) \tag{5.11}$$

où F_s est le même coefficient d'atténuation lié à la présence de gradient que dans (5.6), et β_{tdseg} est un paramètre positif.

Le but de ce terme d'énergie est d'utiliser l'information de mouvement acquise dans le passé, en évitant que la nouvelle segmentation ne s'écarte trop de la segmentation prédite, lorsque les modèles de mouvements nouvellement estimés $\widehat{\Theta}_k$ à l'instant t permettent toujours de décrire de manière satisfaisante le mouvement apparent à l'intérieur de leur région R_k^t associée.

D'un autre coté, celui-ci ne doit pas non plus empêcher la segmentation d'évoluer, notamment aux frontières entre les différentes régions qui existaient dans la carte de segmentation à l'instant précédent. Cette adaptation est rendue possible par la façon dont nous construisons la carte $\tilde{e}$, construction qui a été décrite dans la section précédente. La figure 5.5 illustrait cette dernière sur un exemple simple. Dans la carte prédite, en plus des étiquettes de la segmentation présentes à l'instant $t-1$, nous voyons apparaître les deux classes particulières correspondant d'une part aux zones découvertes, et d'autre part aux régions de recouvrement. Dans les deux cas, sans l'utilisation d'une phase d'interprétation de plus haut niveau de l'évolution de la segmentation – par exemple, quelle région "passe" devant une autre, c'est-à-dire quelle région est la projection de la surface 3D la plus proche de la caméra–, nous ne pouvons pas prédire les nouvelles étiquettes de ces régions. Par conséquent, nous attribuerons à celles-ci des étiquettes particulières dans $\tilde{e}$, qui aura pour effet d'après la définition de $V_{3\text{seg}}$ de ne favoriser aucun modèle de mouvement particulier à l'intérieur de ces régions.

Cette formulation est suffisante pour permettre une adaptation au nouveau contenu de l'image à l'instant t. Dans le cas de mouvements non prédictibles, ou fortement accélérés, il est aisé d'agrandir la zone d'incertitude de la prédiction (et donc la zone laissée libre de toute contrainte de continuité temporelle) en procédant par exemple à une dilatation morphologique des zones de recouvrement et de découvrement dans la carte $\tilde{e}$.

Il peut apparaître "redondant", voire dangereux, d'utiliser la carte $\tilde{e}$ à la fois comme contrainte temporelle et comme initialisation de la partition à un instant t. Cependant, comme nous le verrons plus loin, la minimisation de la fonction d'énergie est effectuée selon l'approche multiéchelle décrite dans le chapitre précédent. L'initialisation de la segmentation se faisant à l'échelle la plus élevée est assez frustre. De plus, le schéma décrit ci-dessus permet de n'appliquer de contrainte temporelle qu'à l'intérieur des régions prédites à partir de la partition de l'instant précédent, ce qui se justifie si l'on veut avoir un minimum de confiance dans les partitions que nous extrayons. Enfin, notons que l'évolution de la segmentation étant essentiellement localisée sur des zones où le gradient spatial de l'intensité est en général assez significatif, les potentiels liés à l'observation de mouvement, d'amplitude beaucoup plus élevée que le terme β_{tdseg}, devraient permettre à eux seuls de faire la différence entre deux modèles de mouvement en compétition.

La description précédente concerne l'ajustement des frontières de mouvement entre régions lors de la première phase de relaxation du schéma général présenté à la figure 5.4, c'est-à-dire lorsque seules les régions existant à l'instant précédent[5], et dont on a définitivement estimé le mouvement, interviennent. Le terme d'énergie $U_{3\text{seg}}$ joue un second rôle lors de la création de régions. Supposons qu'une zone non-conforme significative soit détectée dans une région k. D'après le schéma général de la segmentation du mouvement que nous avons présenté, une nouvelle région l, dont le modèle de mouvement Θ_l est estimé, sera générée et introduite dans la phase de relaxation de la segmentation (le modèle $\widehat{\Theta}_k$ restant inchangé par ailleurs). Cette nouvelle région, n'existant pas bien entendu dans $\tilde{e}$, devra donc apporter un "gain" significatif –à travers $\widehat{\Theta}_l$– dans la description du mouvement de cette zone pour compenser la pénalisation introduite par le terme $V_{3\text{seg}}$. Ce gain devra se traduire par une valeur plus faible du potentiel $V_{1\text{seg}}$ lié aux observations de mouvement. Plus précisément, en faisant abstraction du terme de régularisation spatiale, on devra avoir en moyenne sur les sites s de la région créée:

$$V_{1\text{seg}}(e_s = k, o) \geq V_{1\text{seg}}(e_s = l, o) + F_s \times \beta_{td\text{seg}} \tag{5.12}$$

Comme le coefficient d'atténuation F_s intervient déjà dans la définition de $V_{1\text{seg}}$, il est nécessaire de le prendre en compte dans le potentiel $V_{3\text{seg}}$. En effet, si ce n'était pas le cas, cela signifierait que nous exigeons un gain en énergie plus important dans les zones uniformes que dans les zones texturées[6], alors qu'au contraire, dans les régions (complètement) uniformes, les observations étant susceptibles de valider la plupart des modèles de mouvement, le gain attendu devrait être plus faible.

Lorsque le gain (5.12) n'est pas effectif, c'est-à-dire si $\widehat{\Theta}_l$ ne permet pas de mieux décrire le mouvement que ne le faisait $\widehat{\Theta}_k$, l'algorithme préfèrera alors conserver la région existante k à cet endroit. En ce sens, le terme $U_{3\text{seg}}$ peut être considéré comme un critère d'information permettant de choisir la représentation complète du mouvement apparent global la plus simple possible. Notons ici que l'effet que nous venons de présenter est particulièrement utile lorsque des "mouvements" d'ombres portées, difficiles à prendre en compte par un modèle de mouvement, sont à l'origine des zones non-conformes.

Enfin, nous souhaitons souligner que dans la phase de détection des zones non-conformes, aucune contrainte temporelle n'est utilisée. Ainsi, nous n'empêcherons pas les tentatives de création de nouvelles régions lorsque le modèle de mouvement courant est inadapté pour décrire le mouvement apparent au sein de la région qu'il représente.

5. Certaines de ces régions ont éventuellement été regroupées lors de la phase de fusion.

6. À la limite, en un site s sans gradient spatial d'intensité, où F_s serait égal à 0, le modèle l, pour être choisi par l'algorithme, devrait vérifier d'après (5.12), $0 \geq \beta_{td\text{seg}}$, ce qui est bien sûr impossible, $\beta_{td\text{seg}}$ étant un paramètre positif. Il est donc bien nécessaire de prendre en compte le terme F_s dans (5.11).

5.2.4 Choix des paramètres - Minimisation de l'énergie

Choix des valeurs des paramètres

Les paramètres que nous avons introduits dans les fonctions de potentiel précédentes sont en fait similaires à ceux utilisés pour la détection du mouvement. Les commentaires concernant leur influence restent valables. Cependant, pour fixer leurs valeurs, nous avons le choix entre deux approches possibles:

1. Soit nous considérons les phases de segmentation proprement dite et de détection (des zones non-conformes) comme des parties bien distinctes de l'algorithme de segmentation général, auquel cas nous pouvons déterminer les paramètres de l'une indépendamment de ceux de l'autre. Nous pourrions alors également utiliser pour ces deux phases des observations différentes, ou des potentiels liant les observations aux étiquettes différents[7]. Notons tout de même que dans ce cas, il doit exister une certaine cohérence dans les choix faits pour les deux phases, c'est-à-dire que la phase de détection ne doit pas créer des régions là où la phase de segmentation proprement dite considère que les modèles existants sont suffisants[8].

2. Soit la détection est conçue comme une phase de vérification de la qualité de la segmentation, qui se fait par examen des valeurs des potentiels liées aux observations [Fra91]. Le principal intérêt est alors le faible coût calculatoire supplémentaire qu'elle induit, les potentiels étant déjà calculés.

Nous avons retenu la deuxième option, qui permet d'assurer pleinement la cohérence dont il est question dans le premier point. Dans ce cas, la détermination des paramètres intervenant dans la définition des observations et des potentiels liant celles-ci aux étiquettes résulte alors d'un compromis. Ce dernier permet de prendre en compte la différence essentielle entre segmentation et détection, qui est que dans un cas on compare des modèles entre eux, alors que dans l'autre on teste uniquement la validité d'un modèle.

En effet, dans le cas de la segmentation, nous pourrons parfois être amenés, en un pixel donné, à devoir faire un choix entre des modèles dont aucun ne permet vraiment de décrire le mouvement de ce pixel, ou inversement, entre plusieurs modèles le décrivant correctement (du point de vue du critère que nous nous sommes fixé, c'est-à-dire à δ_{seg} près). Pour pouvoir faire la distinction entre ces modèles, et éviter que celle-ci ne repose uniquement sur les termes de régularisation, les potentiels doivent pouvoir répercuter sur

7. Nous avons par exemple testé avec succès une fonction de potentiel basée simplement sur l'utilisation du vecteur résiduel $v_{n(\Theta_k)}(\mathrm{p}) = DFD_{\widehat{\Theta}_k}(\mathrm{p})/\|\vec{\nabla}I(\mathrm{p})\|$. L'intérêt de cette observation est qu'elle n'est pas le résultat d'un moyennage local. Le modèle d'énergie (5.4) est alors mieux validé.

8. Par exemple, baser la segmentation sur un critère utilisant la DFD , alors que la détection utilise le déplacement résiduel v_n pourrait causer quelques difficultés, certaines régions créées étant systématiquement éliminées ensuite par la phase de relaxation (du fait de la régularisation) associée à la segmentation proprement dite.

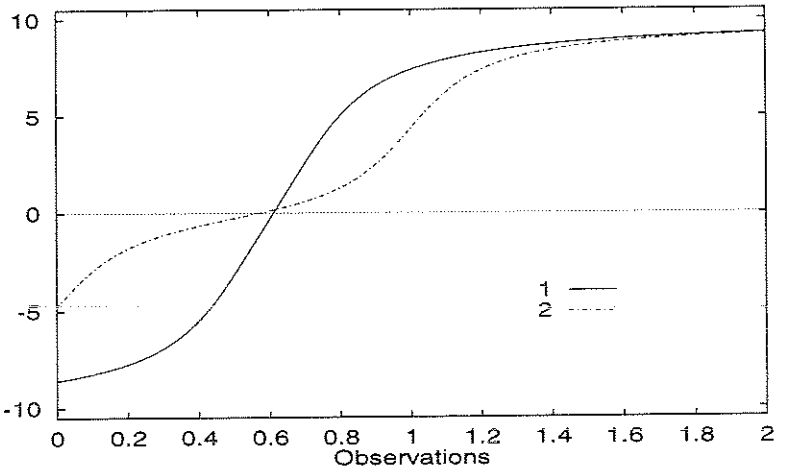


FIG. 5.6 - *Potentiels $V_{1\text{seg}}$ dans le cas d'un site situé en un "coin" d'un contour (courbe 1), et dans le cas d'un site situé sur un contour rectiligne (courbe 2).*
Valeurs des paramètres: $\alpha_{\text{cseg}} = 10$, $\alpha_{\text{ncseg}} = 10{,}3$, $k_{\text{cseg}} = k_{\text{ncseg}} = 2$, ($F_s = 1$; $\delta_{\text{seg}} = 1{,}0$; pour la courbe 1, $l_s = 1/2$ et $L_s = \sqrt{2}/2$; pour la courbe 2, $l_s = 0$ et $L_s = 1$).

l'énergie globale toute différence sensible entre deux observations. Ainsi, dans le potentiel $V_{1\text{seg}}$ de la formule (5.6), nous retiendrons pour k_{cseg} et k_{ncseg}, les paramètres déterminant la rapidité de transition des "échelons" autour des bornes liées à l'observation, une valeur de 2, c'est-à-dire plus faible que dans le cas de la détection du mouvement (où elle était de 4 pour les deux paramètres). La figure 5.6 présente alors ce potentiel pour deux configurations différentes de la fonction intensité.

Concernant les autres paramètres définissant $U_{1\text{seg}}$, notons les points suivants:

- Le paramètre G_m, intervenant dans le calcul de l'observation et des bornes, est lié à la fiabilité de celles-ci, et donc au "bruit" dans l'image (bruit d'acquisition, variations d'illuminations, effet *"d'aliasing"* pour des séquences à l'échantillonnage temporel insuffisant, cf exemple ROND-POINT). Dans le cas de la segmentation du mouvement, on pourrait choisir une valeur plus faible qu'en détection du mouvement, en considérant que si localement la fonction intensité est bruitée, elle le sera pour tous les modèles de mouvement. Cependant, dans la phase de détection, nous risquons alors d'avoir plus souvent de fausses alarmes, et de créer des régions inutiles. Nous conserverons donc pour ce paramètre la valeur utilisée en détection du mouvement.

- Les zones uniformes posent dans la phase de segmentation du mouvement un problème similaire (mais moins important) à celui rencontré en détection. Les observations ont tendance à indiquer que presque tous les modèles $\widehat{\Theta}_k$ décrivent bien le mouvement dans ces régions. Pour faire la distinction entre ceux-ci, nous ne devons donc pas trop atténuer dans ces régions les potentiels liés aux observations, qui permettent

de prendre en compte le peu d'informations qu'elles contiennent. Nous avons donc retenu, pour toutes les expériences, les valeurs suivantes pour le calcul du coefficient d'atténuation F_s défini page 98: $G = 1$, $k_a = 1$ et $\mathrm{At}_{max} = 0{,}5$. Notons qu'ainsi, la phase de détection de régions non-conformes procurera des zones plutôt incomplètes. Ceci n'est pas ennuyeux dans la mesure où elles seront suffisantes pour estimer les modèles de mouvement correspondants. La phase de segmentation proprement dite complètera alors le "travail" de cette phase, en "propageant" l'étiquette associée au nouveau modèle estimé aux régions uniformes correspondant effectivement à ce nouveau modèle.

- nous avons retenu pour α_{cseg} et α_{ncseg} les mêmes valeurs que pour la détection du mouvement, soit: $\alpha_{cseg} = \alpha_c = 10$ et $\alpha_{ncseg} = \alpha_{nc} = 10{,}3$.

- le rôle de la régularisation spatiale est un peu moins important dans le cas de la segmentation du mouvement. En effet, dans la mesure où nous disposons des modèles de mouvement pour l'ensemble des régions, les observations permettent pour une large partie de faire la séparation entre ces modèles. Le terme de régularisation spatiale servira ici à assurer la cohésion des régions. En revanche, comme nous l'avons souligné précédemment, la régularisation temporelle est importante pour éviter la création de régions n'apportant rien à la description du mouvement apparent, notamment pour les mouvements d'ombres portées. En pratique, nous avons constaté que le choix d'une valeur de β_{tdseg} deux fois plus élevée que β_{dseg} était approprié. Pour la détermination des paramètres de régularisation, nous avons fait en sorte que, en un site, la somme des termes régularisants soit de l'ordre de la moitié de l'amplitude totale du potentiel $V_{1\mathrm{seg}}$, soit:

$$8\beta_{dseg} + \beta_{tdseg} \simeq \frac{\alpha_{ncseg} + \alpha_{cseg}}{2} \tag{5.13}$$

Nous obtenons alors une valeur de 1 pour β_{dseg} et de 2 pour β_{tdseg}. Nous souhaitons insister sur le fait que les résultats de la segmentation sont peu dépendants de ces paramètres. Nous avons traité de nombreuses séquences, dont celles présentées dans la partie consacrée aux résultats, avec une régularisation temporelle plus faible, et avec des valeurs de β_{dseg} s'échelonnant entre 0,5 et 1,5, sans constater de différences notables. Il est bien évident qu'une valeur faible de β_{dseg} permet d'obtenir des contours plus précis, mais au détriment de la persistance de petites régions inutiles, et réciproquement pour une valeur élevée de β_{dseg}. De la même façon, une valeur trop élevée de β_{tdseg} rend l'adaptation de la partition aux nouvelles observations plus difficile.

Les paramètres intervenant dans les autres phases du schéma général de la segmentation du mouvement seront donnés lors de la présentation des résultats.

Minimisation de l'énergie

Nous avons également adopté ici la technique d'optimisation multigrille de Pérez et Heitz [PH93, PHB94], avec à chaque niveau une minimisation basée sur l'algorithme HCF de Chou et Brown [CB90]. Cependant, lors du calcul de l'instabilité en un site donné s, seules les étiquettes d'un voisinage de ce site sont considérées[9]. Les raisons suivantes justifient ce choix:

1. si toutes les étiquettes sont considérées, le coût calcul induit par chaque remise à jour d'un site –qui nécessite le calcul de la nouvelle énergie de ces huit voisins– risque d'être relativement important.

2. chaque relaxation se fait à partir d'une carte de segmentation initiale proche de la solution finale. Il s'agit le plus souvent de modifier les régions aux frontières, en fonction des nouveaux estimés des modèles de mouvement a l'instant t, ou d'identifier les frontières des régions nouvellement créées.

3. la fonction d'énergie faisant intervenir un terme de régularisation, les étiquettes absentes du voisinage de s sont fortement défavorisées.

4. la phase de fusion de régions permet de regrouper les régions de mouvement identiques ou similaire.

5. par l'intermédiaire de la minimisation multiéchelle, les étiquettes de mouvement sont propagées –ou du moins proposées– à des sites relativement éloignés.

Ajoutons la remarque suivante concernant la présence de régions uniformes (au sens de la distribution des d'intensités). Ces dernières constituent *a priori* des "barrières" pour la propagation des étiquettes, dans la mesure où la valeur du potentiel V_{lseg} d'adéquation des étiquettes aux observations sera inchangée pour pratiquement toutes les étiquettes, y compris l'étiquette courante de ces régions. Ce sont donc des régions "conservatrices". Or dans l'approche multiéchelle, ce potentiel est calculé pour tous les sites d'un bloc de taille $2^i \times 2^i$, où i est l'échelle considérée. Il existe un niveau d'échelle où ce bloc est plus susceptible de contenir des sites porteurs de gradient spatial d'intensité (et donc d'information). Le potentiel V_{lseg} rejouera alors son rôle pour départager les modèles de mouvement candidats (i.e. étiquettes). Grâce à l'approche multiéchelle, la présence de grandes zones uniformes est donc moins pénalisante.

Comme nous le faisons remarquer dans le second point ci-dessus, la minimisation est toujours effectuée en partant d'une carte de segmentation initiale "évoluée", que ce soit la prédiction $\tilde{e}$ lors de la première relaxation effectuée à un instant donné, ou bien la partition courante dans laquelle sont ajoutées les régions créées pour les éventuelles

9. Ce voisinage n'est pas nécessairement identique au voisinage utilisé pour la modélisation markovienne. Néanmoins, les résultats obtenus avec celui-ci sont largement satisfaisants.

relaxations ultérieures. La partition courante est sous-échantillonnée jusqu'à l'échelle la plus grossière à laquelle commence la minimisation. Les étiquettes des (petites) régions qui n'apparaissent pas à cette échelle sont conservées, en attente d'une échelle plus adéquate[10]. A l'issue de la minimisation au niveau le plus grossier, la carte de segmentation obtenue à ce niveau est projetée au niveau suivant. A ce niveau, les étiquettes dont nous venons de parler sont éventuellement introduites et remplacent automatiquement celles provenant de la projection. La minimisation est alors à nouveau effectuée, et le processus se poursuit jusqu'au niveau le plus fin.

Le nombre de niveaux L_{seg} utilisé dans la minimisation multiéchelle dépend de la taille attendue des régions les plus petites, notamment celles qui sont nouvellement créées. En effet, pour qu'une étiquette k puisse subsister lors de la première relaxation à l'échelle $L_{\text{seg}} - 1$, il faut qu'au moins l'un des blocs de taille $2^{L_{\text{seg}}-1} \times 2^{L_{\text{seg}}-1}$ de cette échelle ne contienne que des sites de mouvement conforme à $\widehat{\Theta}_k$. Comme ces blocs ont une position figée dans la grille, cela implique que la région, si elle était carrée, soit au moins quatre fois plus grande que la taille d'un bloc, c'est-à-dire contienne $2^{L_{\text{seg}}} \times 2^{L_{\text{seg}}}$ pixels. Ce nombre peut alors servir d'indication pour le choix de la taille minimale N_{nc} requise pour qu'une région non-conforme connexe soit acceptée comme nouvelle région.

5.3 Commentaires - Modification des observations

Le schéma général de segmentation du mouvement que nous avons développé repose sur des observations locales de mouvement calculées entre t et $t+1$ uniquement. Pour obtenir une segmentation précise, il est donc important qu'entre ces deux instants, les déplacements soient distincts de part et d'autre de la frontière entre objets ayant des modèles de mouvement différents. Néanmoins, s'il existe une ambiguïté locale entre deux modèles de mouvement, mais qu'elle soit passagère (deux à trois images), la frontière entre les deux régions se maintiendra du fait de l'utilisation de $\tilde{e}$. Malheureusement, si l'ambiguïté persiste sur un intervalle de temps plus important, la segmentation risque d'être imprécise, et notamment de dépendre en partie de la position particulière des blocs du niveau le plus grossier de la minimisation multiéchelle. Il peut alors arriver que l'une des deux régions s'étende sur l'autre[11].

L'utilisation d'un support temporel plus important permet de réduire ces ambiguïtés [AD93, ASB94]. Nous aurions pu alors considérer des observations de mouvement calculées entre plusieurs images consécutives, et les intégrer au cours du temps comme nous l'avons fait pour la détection du mouvement. Cependant, ceci nécessiterait de conserver

10. Cet événement ne se produit que de manière très exceptionnelle en pratique.

11. Ce phénomène dépend beaucoup du cas particulier considéré. Si l'ambiguïté existe entre deux régions de taille moyenne, elle n'affectera la frontière de mouvement que durant la durée de cette ambiguïté. Lorsque l'une des régions est très petite, c'est généralement l'autre qui "prendra le dessus" et s'étendra sur la plus petite. En effet, l'imprécision de la segmentation perturbera surtout l'estimation du mouvement de la très petite région.

pour chaque région les quantités $\epsilon_s(k)$ calculées entre les images précédentes, ou, à chaque remise à jour de l'étiquette d'un site, de recalculer toutes les quantités $\epsilon_s(k)$ passées pour les étiquettes en compétition. La complexité augmente alors très vite avec l'accroisement du nombre de régions. Une autre façon de faciliter la distinction entre les mouvements des différentes régions consiste à considérer des séquences sous-échantillonnées temporellement: on accroît ainsi l'amplitude du mouvement apparent entre deux images successives, et donc les différences entre les modèles de mouvement. Cependant, si l'on augmente trop l'intervalle de temps entre deux images successives, les modèles de mouvement employés sont susceptibles de ne plus être valides.

Les séquences dont nous disposions présentaient généralement des mouvements suffisamment importants ou distincts pour éviter les situations particulières décrites précédemment. De fait, il est apparu que l'imprécision des cartes de segmentation obtenues venaient surtout des régions de recouvrement, pour lesquelles entre t et $t+1$ aucun modèle de mouvement n'est valide[12] Leur étiquetage se faisait alors de manière "aléatoire", en fonction de la régularisation ou de configurations particulières des mouvements et des textures[13]. Un élément plus gênant réside dans le fait que la phase de détection crée systématiquement de nouvelles régions sur ces zones. Bien sûr, l'estimation du mouvement dans ces régions étant impossible entre t et $t+1$, les régions ainsi générées subsistent rarement dans la phase de relaxation suivant leur création. Néanmoins, dû à la fréquence de ces créations, il arrive que pour des situations particulières, la région continue d'exister durant quelques instants, ce qui provoque un phénomène de sur-segmentation et d'instabilité de la carte de segmentation au cours du temps.

Pour pouvoir étiqueter ces régions, nous devons alors prendre en compte l'image à $t-1$. Ceci a été fait par exemple dans [Die91]. Dans cet article, des critères heuristiques utilisant les segmentations spatiales des images (à différents instants) sont employés pour résoudre le problème. Dans notre cas, nous pouvons régler ce problème de façon élégante et simple en ajoutant l'image I_{t-1} aux observations, et en modifiant le terme d'adéquation des étiquettes aux observations V_{lseg}. Nous avons donc considéré pour chaque région k, en plus du modèle de mouvement de t vers $t+1$, $(\Theta_k)_t^{t+1}$, celui de t vers $t-1$, $(\Theta_k)_t^{t-1}$. Dans le cas du modèle affine (que nous avons utilisé dans toutes les expériences), $(\Theta_k)_t^{t-1}$ s'obtient directement à partir du modèle de mouvement $(\Theta_k)_{t-1}^{t}$ calculé à l'instant précédent. Si un modèle de mouvement plus complexe est employé, $(\Theta_k)_t^{t-1}$ peut être estimé en prenant comme support la carte de segmentation prédite $\tilde{e}$. Nous avons alors introduit en un site

12. Nous souhaitons attirer l'attention sur le fait qu'il s'agit là des régions de recouvrement *entre* t *et* $t+1$. Celles-ci différent de celles apparaissant dans la carte $\tilde{e}(t)$, dans laquelle les régions de recouvrements entre les *instants* $t-1$ *et* t sont prises en compte.

13. Ajoutons ici qu'un problème important se posait pour les régions disparaissant dans l'image entre t et $t+1$. Dans ces régions, il n'est pas possible de calculer la quantité ϵ_s pour le "bon" modéle de mouvement. Ce probléme n'était pas marginal dans plusieurs séquences traitées, notamment dans la séquence ROND-POINT ou du fait de la présence de déplacements d'amplitude importante, jusqu'à un dixième de l'image se trouvait concerné.

s les quantités suivantes, ϵ'_s, définies par:

$$\epsilon'_s(k) = \mathrm{Mes}_{(\Theta_k)_t^{t-1}}(s) \tag{5.14}$$

où $\mathrm{Mes}_\bullet$ est donné par la formule (4.19). En un site s d'une zone d'occlusion appartenant à la région k, le terme $\epsilon'_s(k)$ est alors approprié pour caractériser l'adéquation de l'étiquette k aux observations, alors que $\epsilon_s(k)$ ne l'est pas puisque le point p de l'image I_t attaché au site s n'a pas de correspondant dans l'image suivante. Or, à l'instant t, nous ne connaissons pas la localisation des zones d'occlusions entre t et $t+1$. Une étude spécifique des régions non-conformes pourrait peut-être permettre de faire la distinction entre ces zones de recouvrement et les régions dont la détection est due à l'apparition d'une nouvelle région en mouvement. Plutôt que de rajouter une telle phase dans l'algorithme général, nous avons préféré faire le choix d'une modification des quantités caractérisant l'adéquation d'une étiquette aux observations en un site s. Elle consiste simplement à ne retenir entre $\epsilon_s(k)$ et $\epsilon'_s(k)$ que la plus faible, soit:

$$\epsilon''_s(k) = \min(\epsilon_s(k), \epsilon'_s(k)) \tag{5.15}$$

Ainsi, en un site s d'une région de recouvrement, le terme $\epsilon''_s(l)$ sera faible pour l'étiquette correspondant à la véritable région à laquelle appartient le site, et élevée pour les autres.

Nous tenons compte de cette modification en remplaçant dans le terme d'énergie V_{1seg} (formule (5.6)) les quantités $\epsilon_s(e_s)$ par $\epsilon''_s(e_s)$. Pour éviter des incohérences, dans la phase de détection du mouvement, nous utiliserons également $\epsilon''_s(e_s)$.

La validation de l'algorithme de segmentation du mouvement ainsi défini a été effectuée sur de nombreuses séquences réelles de nature variée. Nous présentons dans la section suivante les résultats de quatre expériences significatives.

5.4 Résultats

L'algorithme de segmentation complet a été testé avec succès sur de nombreuses séquences réelles. Nous présentons quatre expériences qui permettent de mettre en évidence les différents aspects de la méthode. Dans la première, la caméra est fixe; l'évaluation de l'algorithme portera sur sa capacité à extraire et à suivre les projections des objets mobiles de la scène, notamment celles qui apparaissent au cours de la séquence. Dans la deuxième expérience, la caméra est mobile et les mouvements sont plus complexes. La troisième expérience illustrera l'influence du paramètre δ_{seg}. Enfin, une dernière expérience montrera la performance de l'algorithme dans les conditions très difficiles de la séquence ROND-POINT.

Dans toutes les séquences, nous avons opté pour le modèle affine 2D comme modèle de mouvement et nous avons recouru, pour l'estimation de ces différentes instances, à

Paramètre	G_m	G	k_a	At_{max}	δ	α_c	α_{nc}	k_c	k_{nc}	β_d	β_{nc}	L_{det}
Valeur	$\diamond$	1,0	1,0	0,5	δ_{seg}	10,0	10,3	2,0	2,0	1,5	0,1	L_{seg}

TAB. 5.1 - *Valeurs par défaut des paramètres utilisés dans toutes les expériences, pour la phase de détection des zones non-conformes. Le paramètre G_m est fixé à chaque expérience (il est identique à celui employé dans la phase de segmentation).*

Paramètre	G_m	G	k_a	At_{max}	δ_{seg}	α_{cseg}	$\alpha_{nc\mathrm{seg}}$	k_{cseg}	$k_{nc\mathrm{seg}}$	$\beta_{d\mathrm{seg}}$	$\beta_{td\mathrm{seg}}$	L_{seg}
Valeur	$\diamond$	1,0	1,0	0,5	$\diamond$	10,0	10,3	2,0	2,0	1,0	2,0	4

TAB. 5.2 - *Valeurs par défaut des paramètres utilisés dans toutes les expériences, pour la phase de segmentation proprement dite . Le paramètre δ_{seg} est fixé pour chaque expérience, de même que G_m.*

l'algorithme RMR modifié. Nous avons choisi C égal à 8, et un nombre maximal de niveaux dans la pyramide gaussienne égal à 5.

Les tableaux 5.1 et 5.2 donnent les valeurs des paramètres employés pour toutes les séquences, sauf indication contraire, dans les phases de détection et de segmentation. On peut remarquer que tous les paramètres liés aux observations et aux potentiels correspondants sont identiques. Par ailleurs, notons que la taille minimale exigée pour le maintien d'une région à l'issue de la phase de relaxation associée à la segmentation est de $N_{\mathrm{elim}} = 80$ pixels (ce qui représente environ un millième d'une image de taille 256×256). Au dessous de cette valeur, la région est éliminée. Comme nous avons retenu quatre niveaux dans la minimisation multiéchelle, le nombre de points minimum considéré comme suffisant pour qu'une région soit créée est de 256 (voir la discussion à ce sujet dans la partie consacrée à la minimisation). Pour être un peu plus strict, nous avons choisi $N_{nc} = 300$. Enfin, notons que le nombre de passages autorisés dans la boucle de création de régions est égal à six pour la première image. Ce nombre est diminué de une unité lors du traitement de chaque nouvelle image, jusqu'au nombre minimum de un passage par image. Pour chaque expérience, les seuls paramètres qu'il nous faut fixer sont donc G_m et δ_{seg}.

5.4.1 Séquence CROISEMENT

Cette première séquence est constituée de 66 images réelles de taille 288×344 pixels, obtenues par numérisation et sous-échantillonnage d'une bande vidéo VHS. La qualité de l'image est moyenne, et nous avons choisi une valeur de 6 pour G_m. Cette séquence pose le problème de la détection de nouvelles régions apparaissant dans l'image. La scène se passe au croisement de deux rues, comme le montre la figure 5.7a. La caméra est placée

sur un véhicule qui marque un arrêt. Un fourgon blanc arrive de la gauche, et se dirige vers la droite. Il est suivi par une voiture noire qui roule quasiment à la même allure. Venant en sens inverse, une voiture blanche traverse l'image tout en se rapprochant de la caméra, avant de disparaître derrière le fourgon pendant une vingtaine d'images environ. Les mouvements, légèrement divergents, étant d'amplitude moyenne (quatre pixels environ pour tous les véhicules), le paramètre δ_{seg} est fixé[14] à 0,8 .

Sur la figure 5.7a, nous avons entouré les zones considérées comme non-conforme au mouvement dominant estimé entre les deux premières images de la séquence. Comme on part d'une seule région –l'image entière– le tout premier modèle de mouvement estimé par la méthode robuste est bien le mouvement dominant. L'image 5.7b montre alors la segmentation obtenue à ce même instant, après création des régions. Nous pouvons constater, que, en utilisant le seul mouvement de t à $t+1$ (qui est le seul possible à l'instant t_1), nous avons obtenu une segmentation imprécise à l'avant du fourgon du fait de la zone d'occlusion. Sur les images suivantes, notamment lors de la disparition de la voiture blanche dernière le fourgon (images 5.7d-e et 5.7f, qui est un aggrandissement de 5.7d), les régions d'occlusion sont par contre correctement étiquetées.

Dans la deuxième partie de la séquence, la voiture noire apparaît. Comme celle-ci roule à la même vitesse que le fourgon et que leur projections se touchent, l'avant de cette voiture est englobé dans la région du fourgon. D'autre part, cette voiture noire, passant devant un fond sombre, est difficilement perceptible, avant que n'apparaisse sa partie arrière. Sur cette dernière, une nouvelle région est créée (image 5.8a), qui s'étend peu à peu à l'ensemble du véhicule. Leurs mouvements étant similaires, les régions correspondant au fourgon et à la voiture sont fusionnées à l'instant suivant[15]. Les figures 5.8c, 5.8d et 5.8f mettent en évidence la formation d'une nouvelle région lors de la réapparition de la voiture blanche. À l'instant t_{53}, cette dernière, tout juste visible (image 5.8c), est détectée par l'algorithme, mais sa taille n'est pas suffisante pour qu'une nouvelle région soit créée. À l'instant suivant, après la phase de segmentation effectuée avec les régions existantes de l'instant précédent, la phase de détection indique précisement la présence d'une région non-conforme (carte de détection, image 5.8f). Une nouvelle région est alors validée, et la phase de relaxation suivante associée à la segmentation permet d'obtenir la segmentation finale présentée à la figure 5.8d (avec un agrandissement), dans laquelle les frontières sont très bien localisées. Notons ici que l'algorithme proposé dans [Fra91] ne parvenait pas à traiter cette apparition de la voiture blanche. En effet, les mouvements opposés (dans la scène) des véhicules peuvent être –partiellement– pris en compte par un modèle de mouvement divergent dans l'image. La projection de la voiture blanche se trouvait alors être intégrée à la région du fourgon. Cette derniére s'étirait donc de plus

14. Nous obtenons des résultats équivalents avec des valeurs de δ_{seg} entre 0,5 et 1,5.

15. Le fait de fusionner ces deux régions pourrait être contesté, dans la mesure où deux objets "autonomes" sont ainsi regroupés au sein d'une même entité. Notons cependant que la région située derrière le fourgon aurait pu tout aussi bien correspondre à une remorque. Dans ce dernier cas, on aurait effectivement souhaité le regroupement des régions !

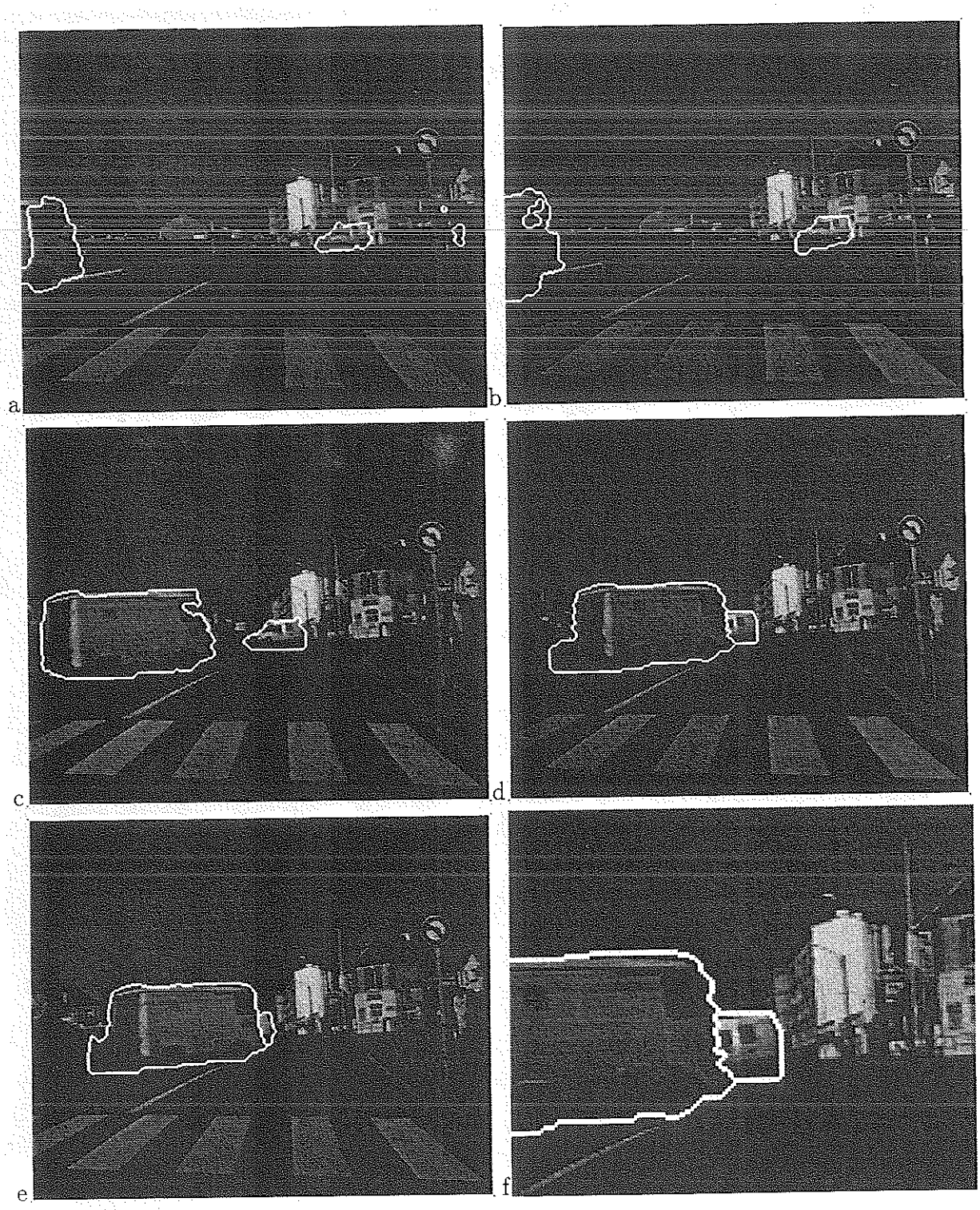


FIG. 5.7 - *a) Régions non-conformes au mouvement dominant à l'instant t_1, avant création des régions. b)c)d)e) Cartes de segmentation obtenues aux instants b) t_1, c) t_{26}, d) t_{34}, e) t_{36}. f) Détail de la segmentation à l'instant t_{34}.*

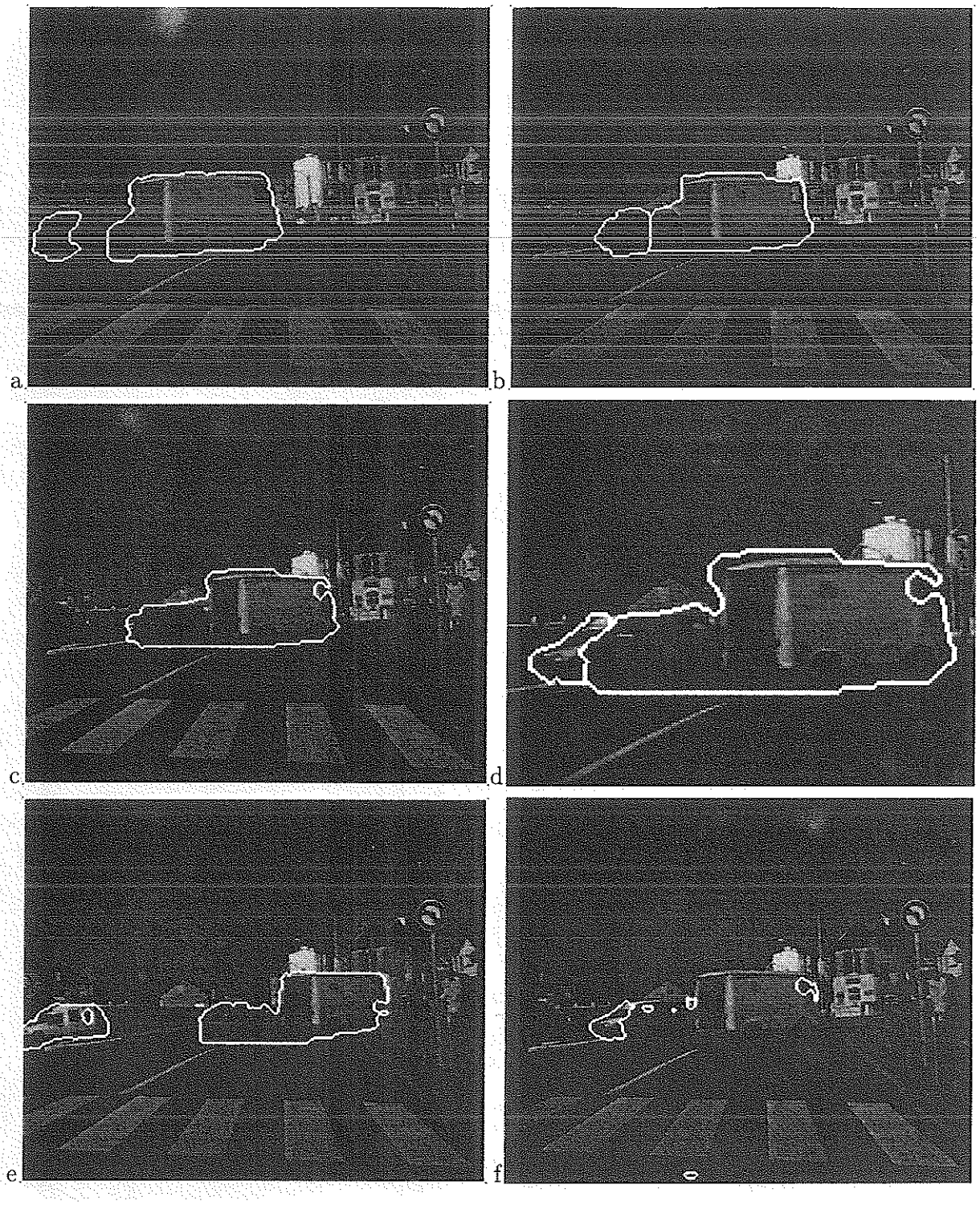


FIG. 5.8 - *Séquence CROISEMENT: cartes de segmentation obtenues aux instants a)* t_{40}, *b)* t_{51}, *c)* t_{53}, *d)* t_{54}*(détail) et e)* t_{67}. *f) Régions non-conformes aux mouvements présents après la première relaxation, à l'instant* t_{54}, *avant création de la région correspondant à la voiture blanche se déplaçant de la droite vers la gauche.*

en plus –pour prendre en compte les deux mouvements– jusqu'à la fin de la séquence. Plus généralement, notre algorithme fournit des localisations de frontières de régions plus précises que dans [Fra91]. De plus, il a un meilleur pouvoir de "résolution" (de séparation) de deux mouvements proches (par exemple dans la séquence J7, dont les résultats en segmentation ne sont pas présentés ici, les mouvements de la fourgonnette et du fond sont tous deux divergents et ne se distinguent que par leur amplitude), ou dans le traitement de scènes très complexes (par exemple pour la séquence ROND-POINTprésenté plus loin).

Enfin, cette séquence met en jeu des mouvements transparents. En effet, nous pouvons voir l'arrière de la scène à travers les vitres des véhicules venant de la gauche, notamment au travers du pare-brise avant du fourgon. Suivant la présence (figure 5.7c) ou l'absence (5.7d-e par exemple) de texture visible à travers ces fenêtres, la segmentation attribue ces régions soit au fond, soit au fourgon.

5.4.2 Séquence MOBI

Les dix images de cette deuxième séquence proviennent du CCETT et sont fréquemment employées pour évaluer des méthodes de compression d'images, notamment avec compensation du mouvement. La qualité de la numérisation est bien meilleure que dans le cas précédent, et nous pouvons utiliser une valeur de 3 pour G_m. La scène (voir figure 5.9a) est constituée de quatre régions aux mouvements apparents différents. La caméra effectue un panomarique de droite à gauche. Celui-ci induit un mouvement translationnel apparent vers la droite, qui est donc le mouvement apparent de la tapisserie. Le calendrier se déplace en outre dans la scène vers le haut, tandis qu'un ballon roule devant un train électrique se déplaçant vers la gauche. L'amplitude des mouvements étant relativement faible, nous présentons les résultats obtenus avec une valeur de δ_{seg} égale[16] à 0,6.

Parmi les tailles des quatre régions évoquées, aucune ne dépasse la moitié de celle de l'image. Il est donc difficile de prédire quel sera le mouvement dominant. La première estimation de ce mouvement sur toute l'image produit en fait un modèle qui permet de prendre en compte le mouvement de toute la tapisserie ainsi que celui d'une partie du calendrier. En effet, les deux mouvements translationnels se trouvent être partiellement "intégrés" au sein d'un même modèle affine. Les régions non-conformes à ce modèle de mouvement sont entourées en blanc dans l'image 5.9a. Visiblement, celles-ci n'incluent qu'une partie des régions de mouvement apparent différent de celui de la tapisserie. En particulier, le calendrier est très incomplètement détecté comme une zone non-conforme, puisque son mouvement a été pris en compte lors de l'estimation. Néanmoins, l'estimation des mouvements sur les régions ainsi obtenues est largement suffisante pour produire, lors de la phase de relaxation liée à la segmentation, la première carte de segmentation de la séquence (figure 5.9b)[17]. L'image 5.9c présente la carte de segmentation obtenue à

16. Nous avons obtenu des résultats complètement similaires avec des valeurs de δ_{seg} s'échelonnant entre 0,4 et 1.

17. Rappelons que lors du traitement de la première image, toutes les régions sont considérées comme

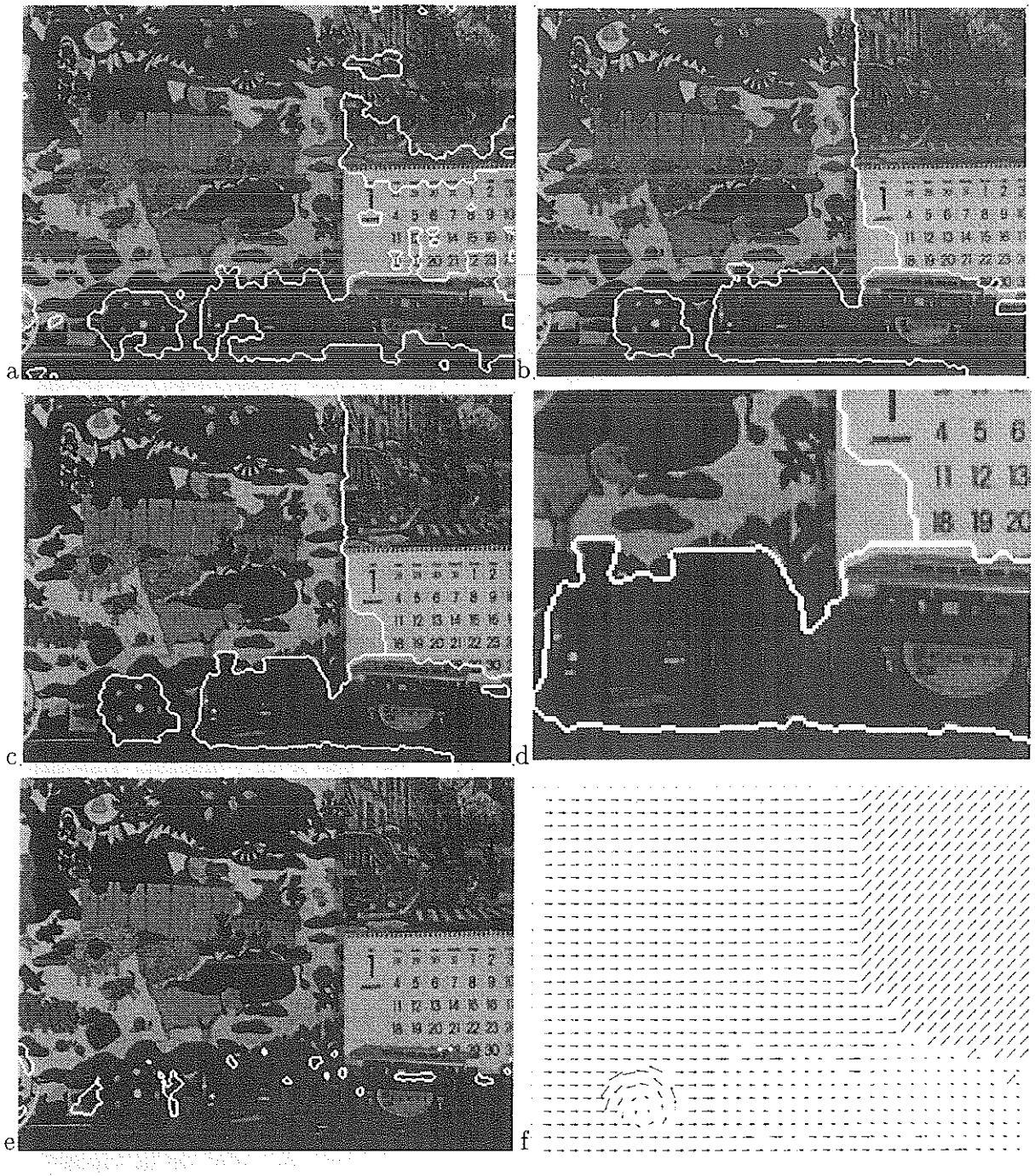


FIG. 5.9 - *Séquence MOBI: a) régions non-conformes au mouvement dominant estimé dans la première image (les régions sont entourées par des traits blancs). b) Carte de segmentation obtenue à l'instant t_1. c) carte de segmentation obtenue à l'instant t_4. d) agrandissement d'une partie de l'image c). e) Régions non-conformes aux modèles de mouvements présents dans la carte de segmentation à l'instant t_4. f) champ des vitesses correspondant aux modèles de mouvement estimés dans les différentes régions de l'image t_4 (le champ est sous-échantilloné par un facteur 9, et l'amplitude des vecteurs est multipliée par 7).*

l'instant t_4, et la figure 5.9d correspond à un agrandissement de cette dernière. On peut remarquer la précision de la localisation des frontières séparant les différentes régions[18]. La figure 5.9e localise les régions non-conformes de la segmentation précédente. On peut remarquer par exemple la détection du bord inférieur du calendrier à travers les vitres du premier wagon du train électrique, ainsi que les mouvements d'ombres autour du ballon. Le champ des vitesses obtenu à ce même instant, correspondant à l'ensemble des modèles estimé dans les différentes régions, est reproduit à la figure 5.9f.

5.4.3 Séquence INTERVIEW

Cette séquence, que nous avons présentée dans le chapitre précédent (figure 4.25a-d), illustre l'évolution de la segmentation lorsque le mouvement d'une région se complexifie. En effet, lorsque la femme située sur la droite de l'image se redresse, les différentes parties de son corps, notamment les membres, bougent de façon distincte. Le modèle de mouvement que l'on estime sur une région prédite à partir de la segmentation précédente n'est alors plus susceptible de décrire correctement les mouvements de sa propre région à l'instant courant. Si la finesse de l'analyse, qui est fixée par le paramètre δ_{seg}, est suffisante, les mouvements "indépendants" du corps seront détectés comme tels, et de nouvelles régions seront créées.

La valeur de G_m que nous avons choisie est la même que celle employée dans le chapitre sur la détection du mouvement, soit 3. Les figures 5.10, 5.11 et 5.12 montrent les résultats obtenus sur l'ensemble de la séquence pour deux valeurs de δ_{seg} différentes. Sur la gauche de ces figures, nous avons placé les cartes de segmentation obtenues ainsi que les champs de vecteur vitesse des modèles estimés pour un choix de δ_{seg} égal à 1,25. Sur la droite, ceux produits avec une valeur de δ_{seg} égale à 0,75. Dans la mesure où le mouvement de la partie gauche de la séquence n'est dû qu'au mouvement de la caméra, nous ne montrons des résultats que pour la zone de l'image correspondant à la femme qui se lève.

Au début de la séquence, les partitions sont presque identiques (figures 5.10a-b). Cependant, en se redressant, la femme déplie ses jambes, qui ont alors un mouvement différent de celui du haut du corps. Celui-ci est pris en compte avec la valeur la plus faible de δ_{seg} (voir figures 5.10d et 5.10f à l'instant t_{13}), mais pas avec la valeur plus élevée de δ_{seg} (figures 5.10c et 5.10e).
À l'instant t_{25} (5.11a et 5.11b) la différence subsiste entre les deux analyses. Cependant, dans les deux cas, le mouvement important du bras gauche a été détecté. On peut remarquer ici que dans le second cas ($\delta_{\text{seg}} = 0{,}75$), le bras a été partitionné en deux régions. Ceci

nouvelles. Ainsi, le modèle de mouvement de la tapisserie a donc également été réestimé avant la phase se relaxation liée à la segmentation.

18. Les cartes de segmentation sont construites en positionnant un pixel de contour (blanc) de part et d'autre de la frontière séparant deux régions. L'épaisseur des contours est donc de deux pixels.

vient du fait qu'il est déclaré dans les images précédentes zone non-conforme, au niveau de la main et de l'épaule, alors que l'avant-bras est encore caché par les fleurs. Ces deux régions sont fusionnées ultérieurement, comme le montre le résultat 5.11d correspondant à l'instant t_{31}. Dans cette figure, et celle correspondant à l'instant t_{43} (figure 5.12b) nous pouvons remarquer la création d'autres régions: celles correspondant à la main droite de la personne, aux cheveux, à l'ombre de ces mêmes cheveux sur le buste, ainsi que l'ombre du bras gauche sur les jambes. Avec la valeur 1,25 de δ_{seg}, le nombre de régions crées est beaucoup plus faible. On peut d'ailleurs remarquer la moins bonne segmentation au niveau du fessier à l'instant t_{43}. Enfin à la fin de la séquence, le mouvement propre de la personne diminue d'amplitude. Le mouvement apparent de celle-ci ressemble de plus au mouvement du fond, comme l'indique les champs de déplacements obtenus à l'instant t_{49} (figures 5.12e et5.12f). Par conséquent, le nombre de régions à cet instant est plus réduit que dans les phases précédentes.

Ainsi, suivant la valeur de δ_{seg}, la partition de l'image est plus ou moins fine. Une valeur faible permet de mieux refléter la complexité des mouvements apparents, mais en produisant généralement une segmentation difficile à exploiter pour l'interprétation du mouvement (de nombreux mouvements d'ombres par exemple sont pris en compte). Avec une valeur élevée, ce sont surtout les mouvements globaux des différentes parties de la scène qui sont pris en compte. Le choix d'une valeur de δ_{seg} dépend donc directement de notre objectif. Si l'on souhaite interpréter le contenu dynamique de la scène d'une façon assez globale, une valeur "importante" de δ_{seg} est préférable. Si l'on se place dans un contexte de codage par compensation de mouvement, une description précise du mouvement est requise. Il faut alors sélectionner une valeur faible de δ_{seg}.

Notons cependant qu'il n'est pas possible de choisir une valeur trop petite de ce paramètre. Ceci est illustré sur les figures 5.13 et 5.14 qui présentent les résultats à l'instant t_{37} obtenus avec quatre valeurs du paramètre δ_{seg}: 1,25, 1,0, 0,75 et 0,5. On peut remarquer que, à cet instant, la partition n'est pas de meilleure qualité avec la valeur de 0,5 qu'avec celle de 0,75[19], et ceci pour trois raisons :

1. Pendant une dizaine d'images, les ombres portées, au niveau du pantalon notamment, ainsi que les mouvements réels de la personne, génèrent des zones non-conformes importantes qui forment souvent une seule région connexe. Dans celle-ci, le mouvement désordonné perçu ne peut pas être estimé correctement avec un modèle de mouvement affine.

2. comme nous l'avons indiqué au début de cette section, dans toutes les expériences, un seul passage par la boucle de création de régions est autorisé. Ainsi, ce facteur, lié au précédent, fait que les régions créées sont moins nombreuses et que modèle de mouvement est moins bien estimé.

19. Insistons sur le fait qu'à d'autres moments de la séquence la qualité est tout de même meilleure.

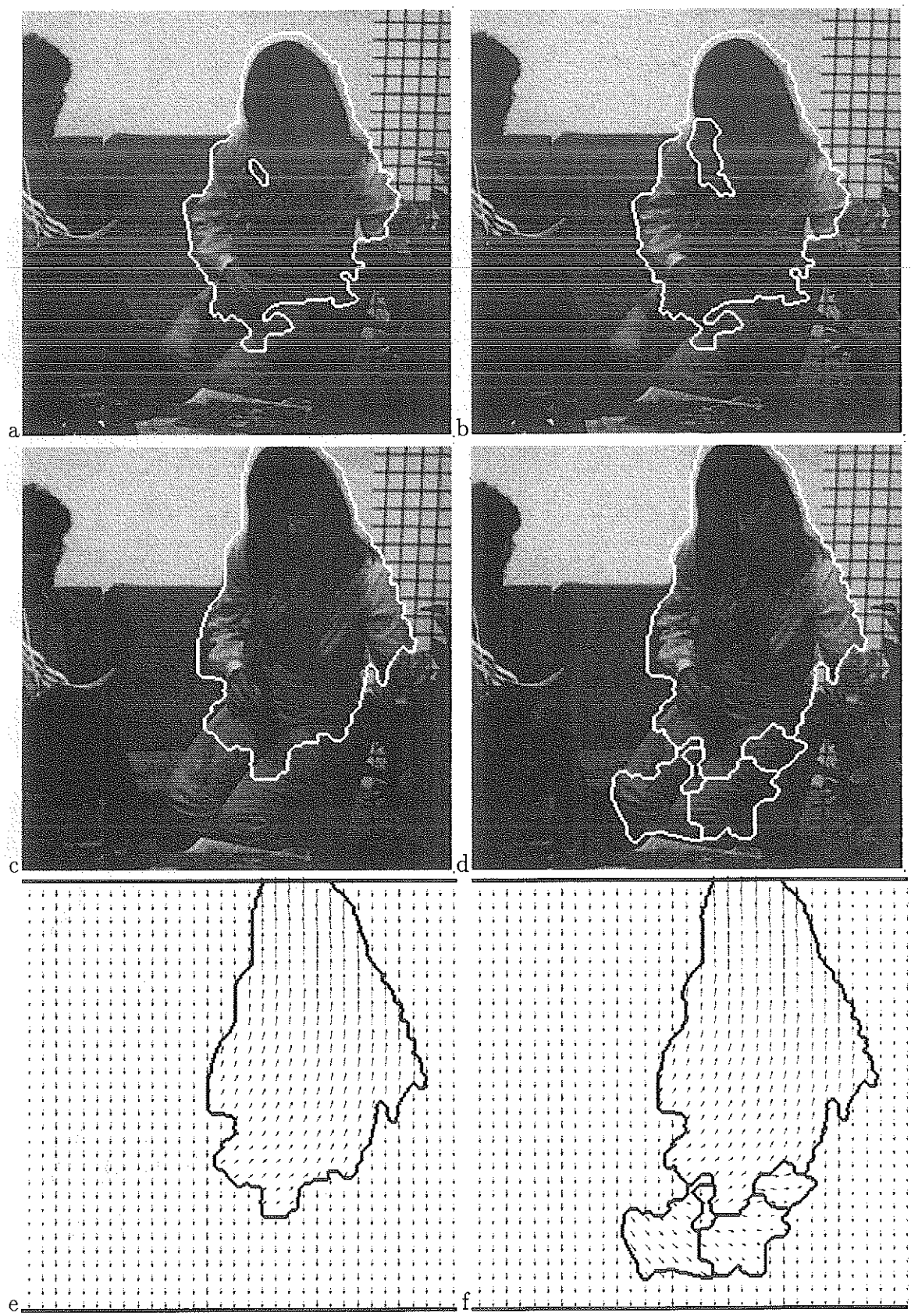


FIG. 5.10 - *Séquence INTERVIEW: cartes de segmentation obtenues aux instants a) b)* t_1*, et c) d)* t_{13}*. La colonne de gauche correspond à une valeur de* δ_{seg} *égale à* 1,25*, celle de droite à une valeur de* 0,75*. e) f) champs des vitesses correspondant aux modèles de mouvement estimés dans les différentes régions de l'image à l'instant* t_{13}*, respectivement associés aux partitions de c) et d). (les champs de vecteurs sont sous-échantillonnés par un facteur 8, et l'amplitude des vecteurs est multipliée par un facteur 3).*

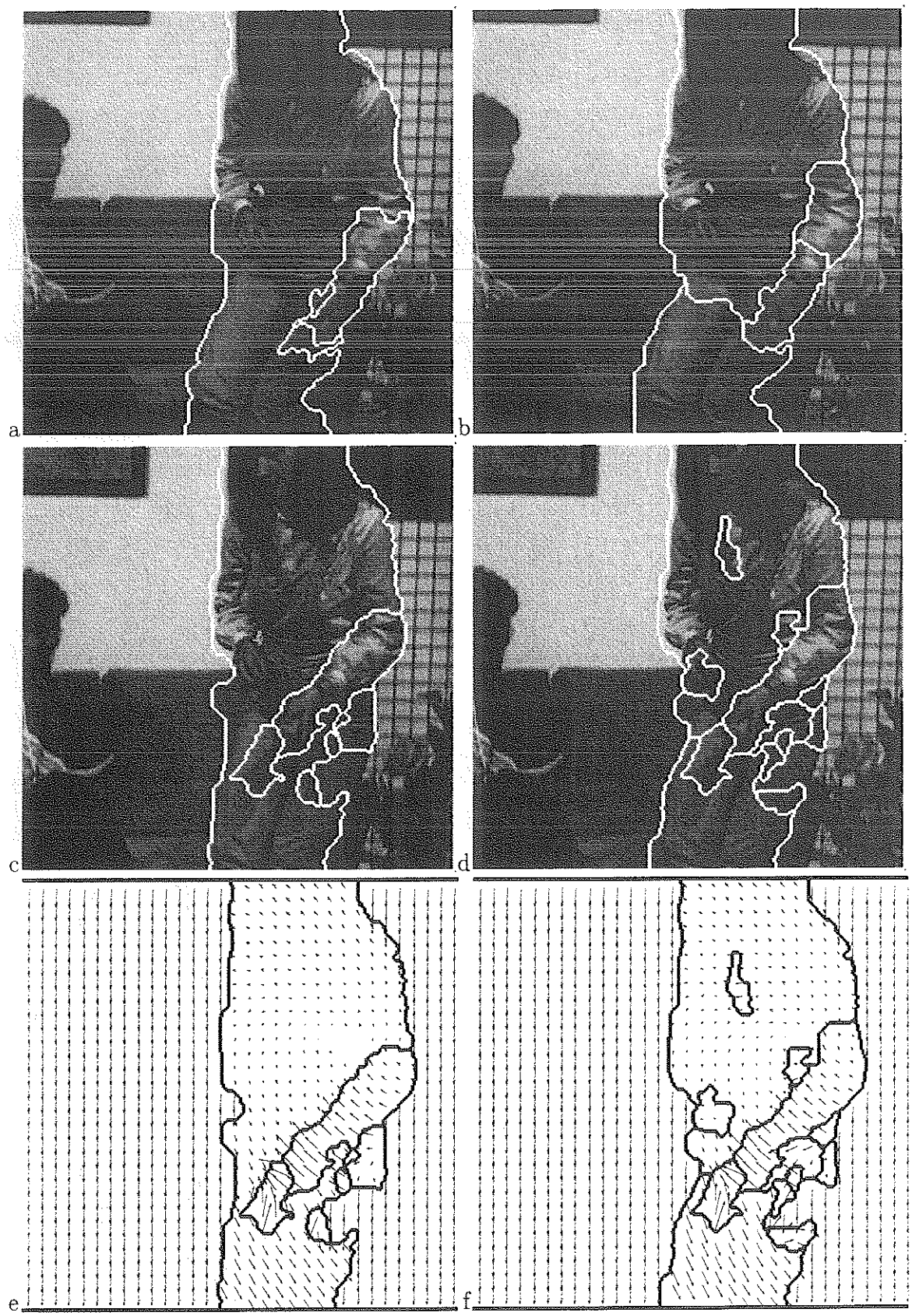


FIG. 5.11 - *Séquence INTERVIEW: cartes de segmentation obtenues aux instants a) b)* t_{25}, *et c) d)* t_{31}. *La colonne de gauche correspond à une valeur de* δ_{seg} *égale à* 1,25, *celle de droite à une valeur de* 0,75. *e) f) champs des vitesses correspondant aux modèles de mouvement estimés dans les différentes régions de l'image à l'instant* t_{31} *respectivement associés aux partitions de c) et d). (les champs de vecteurs sont sous-échantillonnés par un facteur 8, et l'amplitude des vecteurs est multipliée par un facteur 3).*

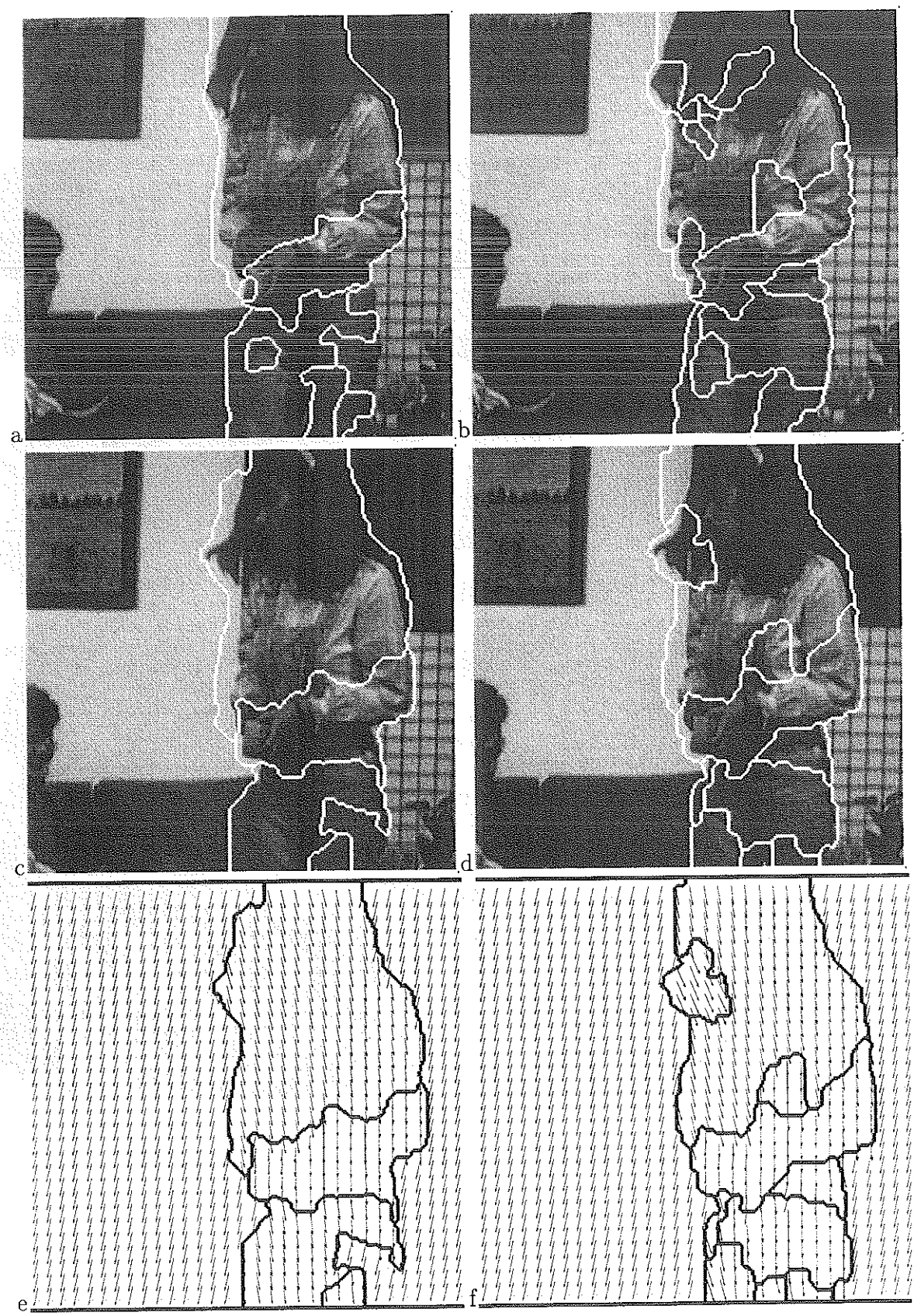


FIG. 5.12 - *Séquence INTERVIEW: cartes de segmentation obtenues aux instants a) b)* t_{43}*, et c) d)* t_{49}*. La colonne de gauche correspond à une valeur de* δ_{seg} *égale à* 1,25*, celle de droite à une valeur de* 0,75*. e) f) champs des vitesses correspondant aux modèles de mouvement estimés dans les différentes régions de l'image à l'instant* t_{49} *respectivement associés aux partitions de c) et d). (les champs de vecteurs sont sous-échantillonnés par un facteur 8, et l'amplitude des vecteurs est multipliée par un facteur 3).*

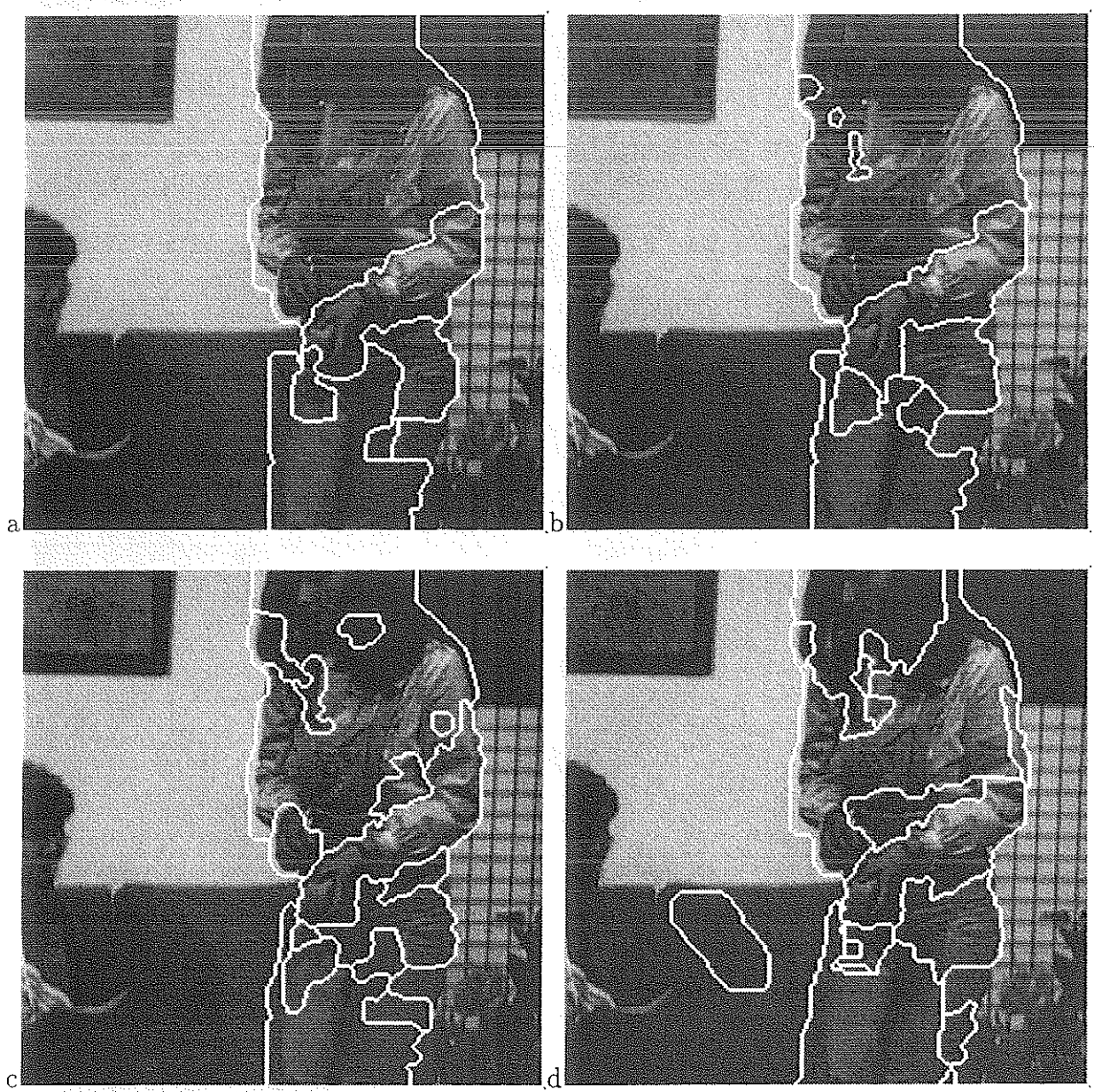


FIG. 5.13 - *Séquence INTERVIEW: cartes de segmentation obtenues à l'instant t_{37} pour différentes valeurs de δ_{seg}: a) $\delta_{\text{seg}} = 1{,}25$; b) $\delta_{\text{seg}} = 1{,}0$; c) $\delta_{\text{seg}} = 0{,}75$ et d) $\delta_{\text{seg}} = 0{,}5$.*

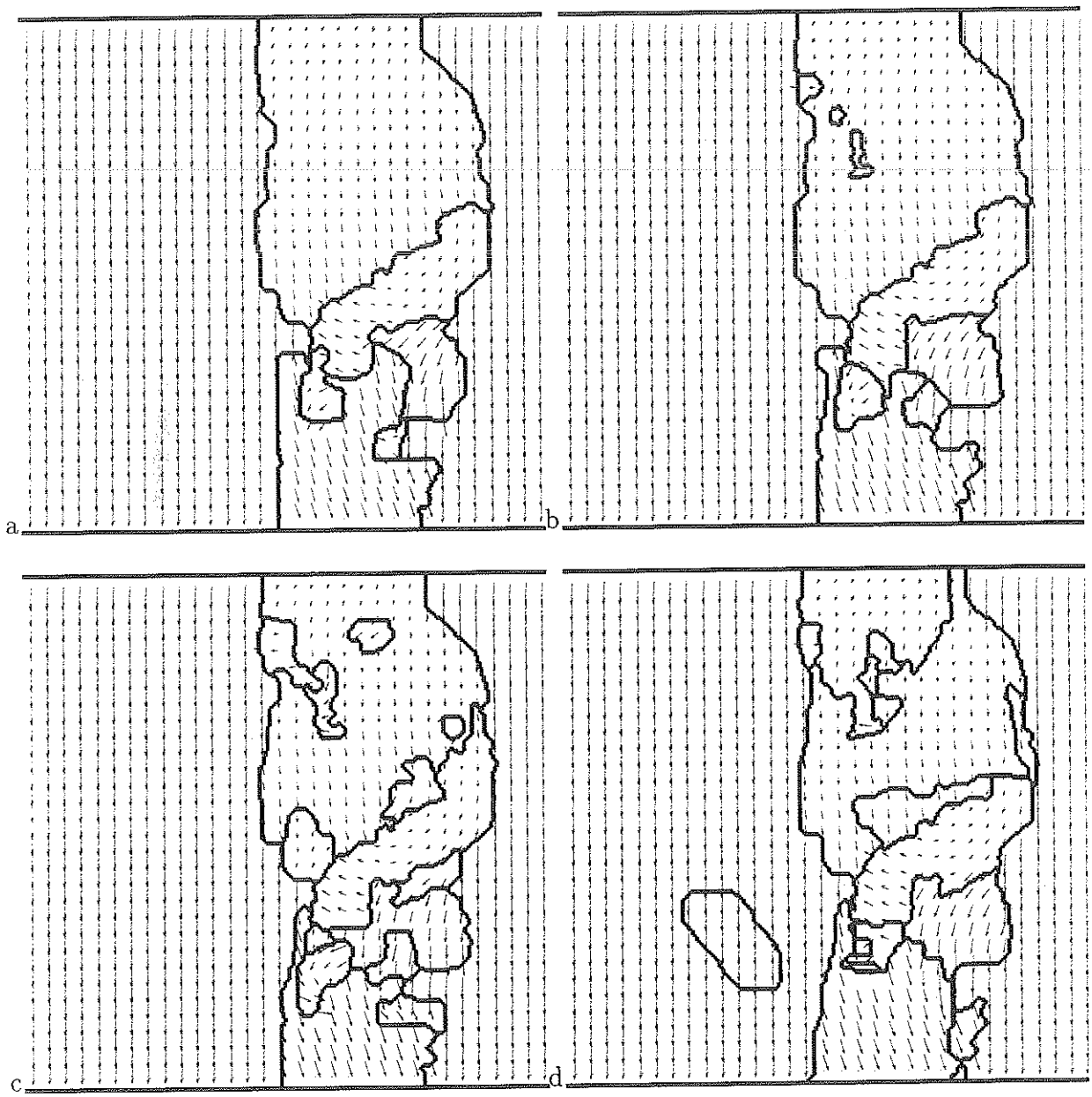


FIG. 5.14 - *Séquence INTERVIEW: champs des vitesses correspondant aux modèles de mouvement estimés dans les différentes régions de l'image à l'instant t_{37} associés aux partitions obtenues pour différentes valeurs de δ_{seg}: a) $\delta_{seg} = 1{,}25$; b) $\delta_{seg} = 1{,}0$; c) $\delta_{seg} = 0{,}75$ et d) $\delta_{seg} = 0{,}5$. (les champs de vecteurs sont sous-échantillonnés par un facteur 8, et l'amplitude des vecteurs est multipliée par un facteur 3).*

3. enfin, compte tenu des deux points précédents, les terme β_{dseg} de la régularisation spatiale et β_{tdseg} de la régularisation temporelle sont trop élevés.

En effet, pour qu'une région soit créée, il faut qu'elle procure une meilleure description du champ des vitesses, "meilleure" dépendant directement du facteur δ_{seg} et de β_{tdseg}. Ainsi, pour une valeur de β_{tdseg} donnée, on est plus exigeant sur la qualité des modèles de mouvement estimés dans les nouvelles régions correspondant aux zones non-conformes lorsque δ_{seg} est faible. Dans un contexte de codage, et avec une valeur faible de δ_{seg}, on recherche une adaptativité "instantanée". Il serait alors préférable de permettre plusieurs itérations lors de la phase de création de région et, surtout, de réduire les paramètres de régularisation temporel et spatial.

5.4.4 Séquence ROND-POINT

Cette séquence, que nous avons déjà présentée dans les chapitres précédents, permet de mettre en évidence à la fois la robustesse de l'algorithme, et certaines de ses limites. Comme nous l'avons souligné auparavant, cette séquence a été enregistrée (de façon analogique sur un support VHS) à une cadence cinq fois inférieure à la cadence vidéo (pour des raisons de multiplexage de cinq sources d'images simultanées), puis numérisée sur le banc de l'IRISA. Les différentes régions de l'image ont donc un mouvement de forte amplitude: de 4 à 8 pixels par image pour l'arrière-plan (maisons), une dizaine pour les marques au sol au premier plan, de 0 à une vingtaine de pixels pour la voiture, et de 15 à 25 pixels pour le panneau de signalisation. Par ailleurs, des phénomènes de repliement de spectre spatio-temporel sont présents dans la séquence (toits des maisons, avant de la voiture). Ces conditions extrêmes nous conduisent à choisir une valeur de 8 pour G_m, et 1,25 pour δ_{seg}.

Étant donné l'amplitude des déplacements, notamment dans la deuxième partie de la séquence, les régions de découvrement et de recouvrement occupent une surface non négligeable de la carte de segmentation initiale $\tilde{e}$ à l'instant t. Les supports de certaines régions de taille faible sont alors insuffisants pour effectuer une estimation fiable du mouvement. Nous avons donc introduit une phase de relaxation supplémentaire avant l'étape d'estimation des mouvements. Celle-ci est réalisée en utilisant $\tilde{e}$ comme carte de segmentation initiale, et à l'aide des modèles de mouvement suivants[20]:

$$(\Theta_k)_t^{t+1} = (\widehat{\Theta}_k)_{t-1}^t \quad \text{et} \quad (\Theta_k)_t^{t-1} = [(\widehat{\Theta}_k)_{t-1}^t]^{-1} \tag{5.16}$$

Nous supposons ainsi que les modèles de mouvement restent constants d'un instant à l'autre. Il aurait été sans doute plus judicieux, dans cet exemple où les mouvements des différentes régions de la scène évoluent très rapidement, de prendre en compte l'accélération. Ceci aurait pu être réalisé à l'aide d'un filtrage de Kalman des paramètres du modèle

20. Rappelons qu'il n'est possible de déterminer "l'inverse" de $(\Theta_k)_{t-1}^t$ que dans le cas du modèle affine (ou de modèles plus simples).

de mouvement [Mey93]. Néanmoins, la phase de relaxation supplémentaire utilisant les modèles indiqués ci-dessus permet d'étiqueter convenablement les régions découvertes et recouvertes. On obtient alors des supports plus complets pour estimer les modèles de mouvement[21].

Les résultats de la segmentation sont présentés sur les images 5.15 et 5.16 à différents instants de la séquence. À l'instant t_{46} (figure 5.15a), l'image est principalement divisée en deux régions. Celles-ci correspondent approximativement à deux "plans" de la scène: le plan vertical formé par les deux maisons, et celui horizontal correspondant à la chaussée. Comme on peut le constater sur les images 5.15d et 5.15f, le champ des vitesses des modèles estimés est quasiment continu à la frontière séparant ces deux régions.

À l'instant t_{50}, la région correspondant au véhicule est créée. Elle est correctement suivie jusqu'à l'instant t_{65}. Cette région se divise alors en deux parties (figure 5.16a), puis en plusieurs autres régions (figure 5.16c). En effet, le changement d'attitude du véhicule est trop rapide pour pouvoir être pris en compte par le modèle affine. Ensuite, la voiture est partiellement cachée par le panneau de signalisation, générant des occlusions importantes. Néanmoins, malgré la petite taille des régions comparativement à l'amplitude des mouvements, et compte-tenu des phénomènes d'occlusion évoqués, le mouvement apparent est globalement bien estimé sur les régions correspondant à ce véhicule (figure 5.16d). Lorsque celui-ci disparaît, le mouvement de son ombre derrière lui, encore visible, est pris en compte dans l'estimation du mouvement (figure 5.16f).

Quant au panneau de signalisation, son apparition dans l'image est détectée deux fois. La première fois à l'instant t_{61} (figure 5.15c) par l'intermédiaire de son triangle supérieur. La seconde, à l'instant t_{64}, lors de l'apparition de l'indication "cédez le passage". Les deux régions ainsi créées coexistent alors durant huit images (figures 5.16a et 5.16c), puis sont fusionnées (image 5.16e). On peut constater dans ces images, que la région correspondant au panneau est parfaitement délimitée, malgré l'amplitude extrêmement importante du déplacement de celui-ci[22], et le peu de texture qu'il contient. Bien sûr, le ciel est englobé dans cette région, mais comme nous l'avons déjà évoqué, celui-ci ne contient aucune information, pas même des nuages!

5.5 Conclusion

Dans ce chapitre, nous avons décrit un schéma complet de segmentation du mouvement apparent dans une séquence d'images. Ce schéma a été validé sur de très nombreuses séquences correspondant à des scènes complexes d'extérieur et d'intérieur. Nous avons

21. La relaxation initiale que nous venons de décrire pourrait être également utilisée dans les autres séquences. On obtient généralement des cartes de segmentation de qualité légèrement meilleure, mais au prix d'un coût calcul plus élevé.

22. Il suffit d'observer l'inscription "cédez le passage": celle-ci est encore lisible dans l'image 5.16a, mais devient complètement brouillée dans 5.16c et 5.16e.

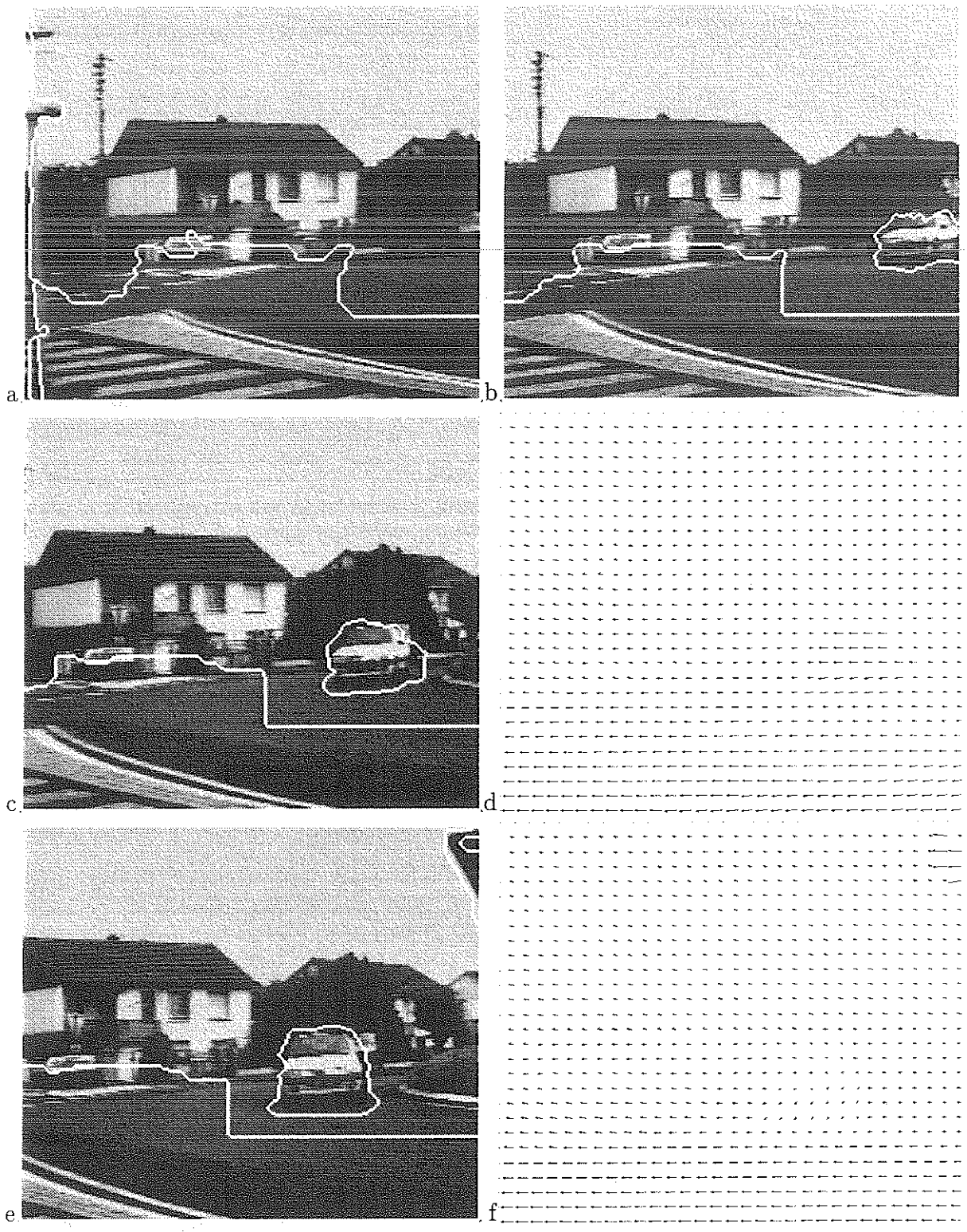


FIG. 5.15 - *Séquence ROND-POINT: cartes de segmentation obtenues aux instants a)* t_{46}, *b)* t_{51}, *c)* t_{56}, *e)* t_{61}. *d) et f): champs des vitesses correspondant aux modèles de mouvement estimés dans les différentes régions de l'image aux instants d)* t_{56} *et f)* t_{61} *(les champs de vecteur sont sous-échantillonnés par un facteur 8, et l'amplitude des vecteurs est multipliée par un facteur* 0,75*).*

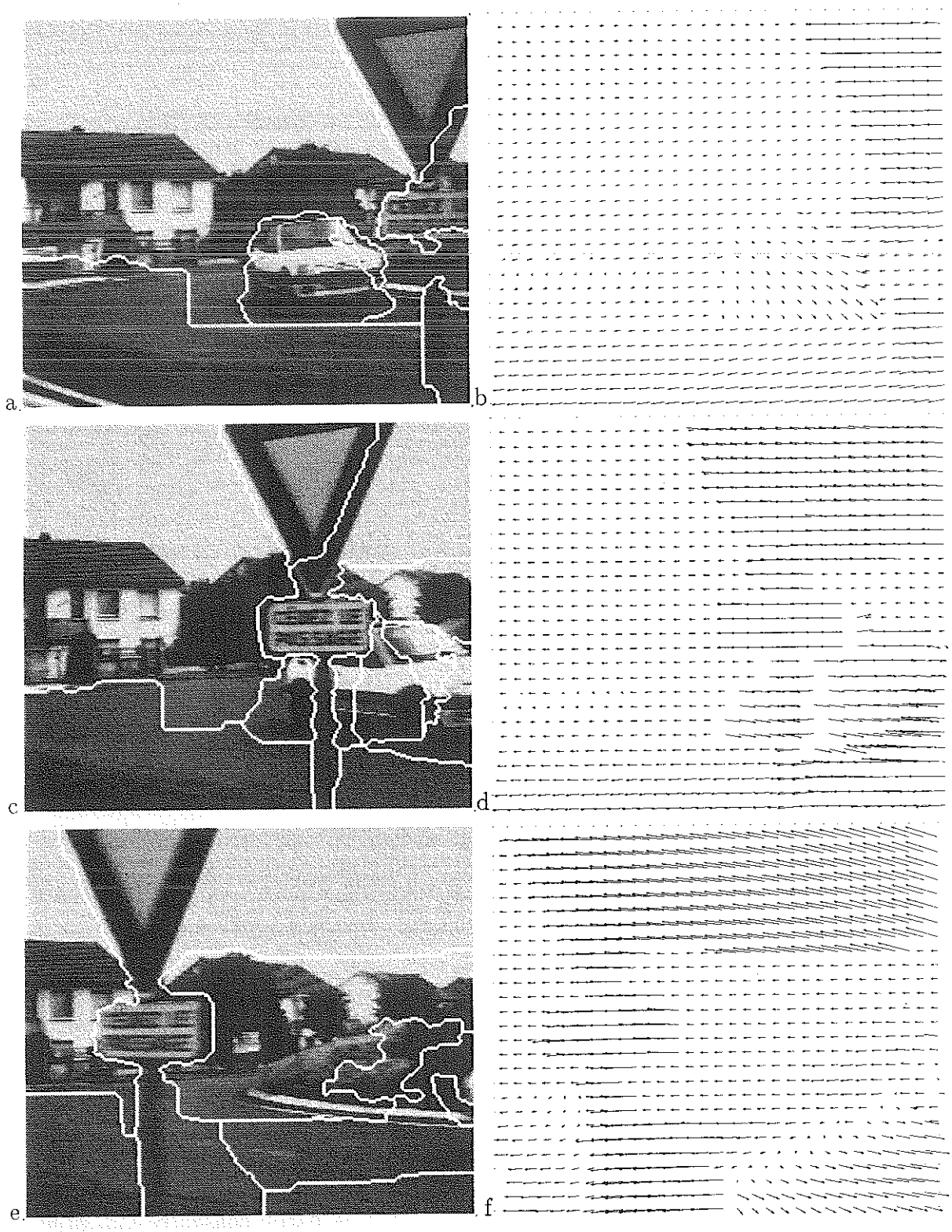


FIG. 5.16 - *Séquence ROND-POINT: cartes de segmentation obtenues aux instants a)* t_{66}, *c)* t_{71} *et e)* t_{76}. *b) d) f): champs des vitesses correspondant aux modèles de mouvement estimés dans les différentes régions de l'image aux instants b)* t_{66}, *d)* t_{71} *et f)* t_{76}. *(les champs de vecteur sont sous-échantillonnés par un facteur 8, et l'amplitude des vecteurs est multipliée par facteur* 0,75*).*

présenté les résultats obtenus sur quatre d'entre-elles. Ceux-ci prouvent le bien fondé de l'approche retenue ainsi que les choix qui ont été faits. L'originalité de la méthode repose sur cinqpoints importants:

- Le premier porte sur l'utilisation de modèles de mouvement 2D et des seules valeurs d'intensité. Ainsi, aucune mesure 3D explicite, ni estimation préalable d'un champ dense de déplacements ne sont requises.

- Le second correspond à l'emploi d'un estimateur de mouvement robuste. Grâce à celui-ci, il suffit d'estimer les modèles de mouvement des différentes régions une seule fois à chaque instant. En conséquence, une seule phase de relaxation est suffisante pour ajuster les frontières entre les régions déjà existantes. Ceci diffère des schémas traditionnels, qui requièrent généralement plusieurs itérations entre les phases d'estimation et de relaxation.

- D'autre part, le problème de segmentation est formulé à l'aide d'une modélisation statistique bien posée, à savoir celle des champs de Markov. Une attention toute particulière a été portée à la définition d'une fonction d'énergie adaptée aux problème de segmentation du mouvement. Le choix de grandeurs calculées homogènes à un mouvement dans entre deux images, et non pas à une différence d'intensités interimage comme c'est le cas avec la DFD ou l'équation de contrainte du mouvement apparent, s'est montré très judicieux. Ainsi, le paramètre essentiel (δ_{seg}) influant sur la méthode de segementation du mouvement possède une signification directe en terme de mouvement dans l'image.

- L'utilisation d'une méthode de minimisation multiéchelle mathématiquement cohérente et performante, qui permet d'éviter de nombreux minima locaux de la fonction d'énergie définie. Par ailleurs, l'utilisation d'une pile d'instabilité (méthode HCF) dans la phase de minimisation, ainsi que la stratégie particulière pour le choix des étiquettes à considérer localement, permettent d'obtenir une complexité calculatoire réduite.

- Enfin, le cinquième apport est dû à la définition d'une phase explicite de détection des zones où le mouvement n'est pas conforme aux modèles de mouvement estimés. Celle-ci permet d'obtenir une segmentation correcte dès le début de la séquence et de gérer adéquatement et efficacement l'apparition de nouveaux objets dans la scène, ainsi que le passage par des phases de mouvement plus complexes des régions présentes dans l'image.

Ces différents points ont été mis en évidence dans la partie consacrée aux résultats. Les différentes composantes dynamiques de la scène observée sont bien délimitées et reliées temporellement. La qualité des partitions obtenues permettrait ainsi d'aborder une seconde phase de suivi 2D [MB94a] ou d'interprétation 3D [FB90, NL92] de la scène dans de très bonnes conditions.

Chapitre 6

Conclusion

Nous résumons tout d'abord les principaux aspects et contributions de notre étude, et proposons ensuite des extensions possibles à ces travaux.

6.1 Contributions

Dans cette thèse, nous avons abordé le problème général de la détection d'objets mobiles dans une séquence d'images acquises par une caméra elle-même en mouvement.

La méthode que nous avons retenue consiste à modéliser le mouvement apparent des zones statiques de la scène par un modèle de mouvement 2D. Celui-ci est alors exploité pour compenser le mouvement dû à la caméra, ce qui nous replace en quelque sorte dans le cas d'une caméra fixe et permet de poser le problème d'extraction des éléments mobiles de la scène comme un processus d'étiquetage binaire, c'est-à-dire comme un problème de détection.

La première phase de la méthode de détection consiste donc à effectuer une estimation correcte du modèle de mouvement ainsi introduit. Puisque l'on s'attend à ce que la scène contienne des objets mobiles, dont les projections dans l'image peuvent être de taille significative comparativement aux régions statiques, l'emploi d'un estimateur robuste s'avère nécessaire. Nous avons donc développé une méthode multirésolution basée sur un tel estimateur et sur une approche incrémentale pour minimiser le critère retenu. Cette méthode permet de calculer de manière fiable le modèle de mouvement 2D dominant (supposé correspondre au mouvement apparent du fond statique de la scène), même en présence de mouvements secondaires importants.

La détection des régions de mouvement non-conforme au modèle de mouvement estimé est ensuite formulée dans un cadre bayesien à l'aide de modèles markoviens. Dans la fonction d'énergie à minimiser associée à cette formulation, le terme exprimant l'adéquation des étiquettes aux observations a été défini avec soin. Plus précisément, nous avons cerné et exploité explicitement l'information partielle de mouvement réellement disponible

dans le voisinage de chaque point de l'image. Nous avons notamment pris en compte la variété des distributions locales de la direction du gradient spatial de l'intensité, ainsi que la présence ou non localement de ce même gradient. Ces différents aspects, conjugués à l'utilisation d'informations de compensation de mouvement sur une plage temporelle étendue et le choix d'une méthode d'optimisation multiéchelle performante, permettent d'obtenir des taux de fausses alarmes très faibles ainsi que des masques complets des objets mobiles. Notons ici que la complexité calculatoire est réduite. Soulignons également que, dans son principe, cette méthode permet également de suivre des objets dans une séquence d'images.

Enfin, le fait que l'algorithme développé ne requiert pas une compensation quasi parfaite sur plusieurs images, comme c'est le cas d'autres algorithmes utilisant le procédé de compensation, mais seulement un recalage plus précis que l'amplitude des mouvements à détecter, consitue un atout important de la méthode.

La méthode de détection ainsi définie donne de très bons résultats dans un nombre de cas pratiques importants. Cependant, pour analyser plus finement une séquence d'images, il peut s'avérer nécessaire, voire crucial dans certains problèmes d'analyse de scène dynamique, de partitionner la séquence en régions de mouvements homogènes. Nous avons donc étendu l'algorithme d'étiquetage binaire de la détection (et de "suivi") au cas n-aires. Dans le schéma de segmentation complet que nous avons défini, une phase de détection explicite détermine les zones dans lesquelles le mouvement est mal décrit par les modèles de mouvements existants. Comme les résultats ont pu le mettre en évidence, cette phase permet de s'adapter au contenu dynamique de la scène. Notamment, de nouvelles régions sont créées lors de l'apparition de nouveaux objets dans la scène ou lorsque le mouvement d'une région donnée devient plus complexe.

6.2 Extensions possibles, et futures recherches

Un certain nombre de questions relatives aux travaux effectués restent à explorer. Nous en proposons ici quelques unes qui constituent des voies d'investigation intéressantes dans la suite des travaux que nous avons présentés.

Un premier problème relatif à l'estimation "hiérarchique" des modèles de mouvement concerne le choix des niveaux de résolution où doit débuter l'estimation des paramètres constants, linéaires, voire même quadratiques, du modèle de mouvement choisi. Actuellement, seul un critère de taille ainsi que des niveaux prédéfinis sont pris en compte. Il pourrait être intéressant de tenir compte des estimés obtenues aux instants précédents pour sélectionner ces niveaux, et d'évaluer l'impact d'un choix inadéquat sur les coefficients estimés.

Par ailleurs, comme l'estimation des paramètres de mouvement se fait de manière ité-

rative, il pourrait être judicieux d'utiliser les mesures passées pour initialiser l'estimation courante. En fait, le problème est plus complexe qu'il n'y paraît. En effet, s'il est préférable de filtrer les coefficients estimés lorsqu'on les utilise pour calculer d'autres quantités comme le temps-à-collision, ou des paramètres d'un mouvement 3D rigide (voir [Mey93]), nous avons pu constater en revanche que l'initialisation de l'estimation du modèle de mouvement avec des valeurs prédites par un filtre de Kalman[1] ou même l'utilisation des valeurs filtrées pour effectuer la compensation d'images, pouvaient conduire à des résultats désastreux. L'utilisation de la cohérence temporelle des paramètres, qui est très importante pour suivre les petites régions animées d'une très grande vitesse, n'est donc pas évidente à exploiter pour estimer un modèle de mouvement, et nécessiterait une recherche approfondie.

Un axe de recherche attractif concerne l'utilisation de la multirésolution, à la fois spatiale et temporelle, dans la détection et la segmentation du mouvement. Comme nous l'avons souligné sur l'exemple AVION, l'utilisation d'observations calculées entre deux instants non consécutifs peut améliorer la détection. De même, l'utilisation de différentes résolutions spatiales devrait permettre de discriminer plus aisément différents mouvements. On peut imaginer plusieurs façons d'exploiter la multirésolution. Par exemple, il serait possible de rajouter aux fonctions d'énergie que nous devons minimiser, des termes énergétiques faisant intervenir les observations à différentes résolutions; ou bien de sélectionner le niveau de résolution spatial et temporel le plus approprié pour distinguer des modèles de mouvement différents ou apprécier plus précisément les erreurs de compensation de mouvement. En fait, le problème posé est très général puisqu'il concerne directement celui de l'adaptativité d'un algorithme à la diversité des signaux à traiter, c'est-à-dire la recherche des échelles spatio-temporelles dans lesquelles un phénomène (dans notre cas le mouvement) s'observe le mieux pour un signal donné.

Enfin, rappelons que dans l'algorithme de détection du mouvement, les régions de mouvement non-conforme au modèle de mouvement dominant appartiennent en fait à deux classes distinctes: celle des régions qui correspondent réellement à des objets mobiles, et celle des régions statiques situées à une profondeur (ou suivant une orientation) bien différente de la région statique dominante[2]. La différentiation de ces régions est un problème intéressant, et peut sans doute être réalisée en exploitant la partition fournie par l'algorithme de segmentation du mouvement apparent. En fait, deux alternatives sont

1. Avec cette méthode, le problème suivant se pose: les mesures effectuées ne sont plus du tout indépendantes de la prédiction du filtre de Kalman. Si le processus d'estimation est mal initialisé par la prédiction, l'estimateur de mouvement tombe dans un minimum local très proche de la mauvaise initialisation. La prédiction s'en trouve ainsi "confirmée". Ceci peut conduire à une dérive aberrante des paramètres de mouvement.

2. Rappelons que cette dernière classe n'apparait qu'en présence de mouvement translationnel important du capteur.

à notre disposition. La première consiste à adapter et à appliquer sur le champ des vitesses résultant de la partition et des modèles de mouvement associés, les méthodes qui utilisent des contraintes sur le champ des déplacements apparents, comme [TP90, Nel91]. La seconde consiste à s'intéresser directement aux modèles de mouvement et à l'information qu'ils contiennent sur l'environnement tridimensionnel de la scène. En effet, si l'on suppose par exemple que la région associée à un modèle de mouvement correspond à la projection d'une surface plane de la scène, il est théoriquement possible de tester si deux surfaces planes associées à deux modèles de mouvement différents appartiennent en fait au même objet, c'est-à-dire ont le mouvement rigide 3D.

Notons également que la complexité algorithmique des algorithmes proposés étant relativement faible, on peut espérer que l'étude de leur parallélisation et de leur implantation sur des cartes spécialisées débouche sur des versions fonctionnant presque en temps réel. Ainsi, il pourrait être envisagé d'étudier les éventuels avantages des modèles de mouvement 2D et de leur estimation sur des régions pour la réalisation de certaines tâches robotiques, dans un contexte d'asservissement visuel par exemple [SBC94].

Annexe A

Rappel sur les champs de Markov

Dans le cadre général de l'analyse d'images, un grand nombre de problèmes peuvent s'exprimer comme la recherche de primitives ou étiquettes (c'est-à-dire informations cachées à estimer) à partir des données observées ou observations. Cependant, le processus de formation des observations opère généralement une réduction de l'information accompagnée de l'introduction de différents bruits, ce qui rend l'extraction des primitives (ou étiquettes) difficile. On parle alors de problème mal posé. *A Contrario*, un problème bien posé se caractérise par les éléments suivants [BPT88]:

1. il existe une solution, qui est unique,
2. elle dépend continûment des observations.

Pour surmonter cette difficulté, il faut restreindre l'ensemble des solutions admissibles du problèmes en introduisant un certain nombre d'hypothèses sur les propriétés attendues de la solution.

Dans ce contexte, la théorie bayesienne fournit un support mathématique cohérent pour modéliser le lien physique entre observations et étiquettes ainsi que l'information *a priori* sur ces dernières. Dans ce cadre, les champs de Markov, notamment grâce à leur équivalence avec les distributions de Gibbs, se prêtent particulièrement bien à la modélisation de nombreux problèmes en analyse d'images. Les principaux avantages des modèles markoviens sont les suivants:

- ils sont adaptés à la modélisation d'interactions non-linéaires entre les étiquettes et les observations ainsi qu'entres étiquettes voisines qui peuvent être de nature différentes (contours et vecteurs vitesse par exemple);
- la description globale de ces interactions peut se faire de manière simple par la définition de fonctions d'énergie;
- les algorithmes mis en jeu pour atteindre la configuration optimale possèdent très souvent des propriétés intéressantes pour leur mise en œuvre: ils sont *uniformes*

ou réguliers (les opérations sont identiques en chaque point), *locaux* (les calculs en chaque point ne font intervenir que des données locales), et *parallélisables* (les opérations peuvent se faire simultanément sur des sous-ensembles importants et réguliers des points de l'image[1]);

Dans cette annexe, nous rappelerons les propriétés de base des champs de Markov et leur utilisation dans un cadre bayesien . Nous décrirons ensuite différents algorithmes de minimisation permettant d'obtenir une solution, en prêtant plus particulièment attention à l'algorithme multiéchelle proposé par Pérez et Heitz [PH93, PHB94].

A.1 Champs de Markov - Critère du Maximum a Posteriori

Soit O l'ensemble des observations et E l'ensemble des étiquettes à estimer. On suppose que ces deux ensembles forment des champs de variables aléatoires sur une grille S de N sites:

$$O = \{O_s, s \in S\} \quad \text{et } E = \{E_s, s \in S\} \tag{A.1}$$

On notera également $o = \{o_s, s \in S\}$ et $e = \{e_s, s \in S\}$ une réalisation quelconque de ces deux champs. Nous supposerons qu'une étiquette E_s prend ses valeurs dans un ensemble discret Λ, et nous définirons Ω l'ensemble de toutes les configurations possibles de E: $\Omega = \{e, e \in \Lambda^S\}$.
Parmi les différents critères d'estimation, celui du Maximum A Posteriori est le plus souvent retenu. Étant donné une réalisation o de O, on cherche la configuration $\hat{e}$ la plus probable *a posteriori*, soit:

$$\hat{e} = \underset{e \in \Omega}{\operatorname{argmax}}\ p(E = e | O = o) \tag{A.2}$$

ce qui donne après utilisation de la règle de Bayes et élimination de $p(O = o)$, qui ne dépend pas de e:

$$\hat{e} = \underset{e \in \Omega}{\operatorname{argmax}}\ p(O = o | E = e) \times p(E = e) \tag{A.3}$$

Le premier terme de (A.3) s'obtient en modélisant le lien entre les observations et les étiquettes. Le second terme, la probabilité d'occurence du champ des étiquettes, renferme les propriétés *a priori* du champ E. Celles-ci sont modélisées en faisant l'hypothèse que E est un champ de Markov relativement à un système de voisinage $\mathcal{G}$ défini sur S. Rappelons ici que $\mathcal{G}$ est formé de la réunion de sous-ensembles $\mathcal{G}_s$ de S vérifiant les propriétés: 1) $s \notin \mathcal{G}_s$ et 2) $t \in \mathcal{G}_s \Leftrightarrow s \in \mathcal{G}_t$. De plus on notera $\mathcal{C}$ l'ensemble des cliques c, une clique c désignant toute partie de S réduite à un singleton ou dont les sites sont voisins deux à deux. La figure A.1 montre les systèmes de voisinage d'ordre 1 et 2 ainsi que les différents types de cliques associées.

1. Par exemple, dans le cas d'un voisinage d'ordre 1, la remise à jour des étiquettes peut se faire simultanément sur la moitié des sites de l'image.

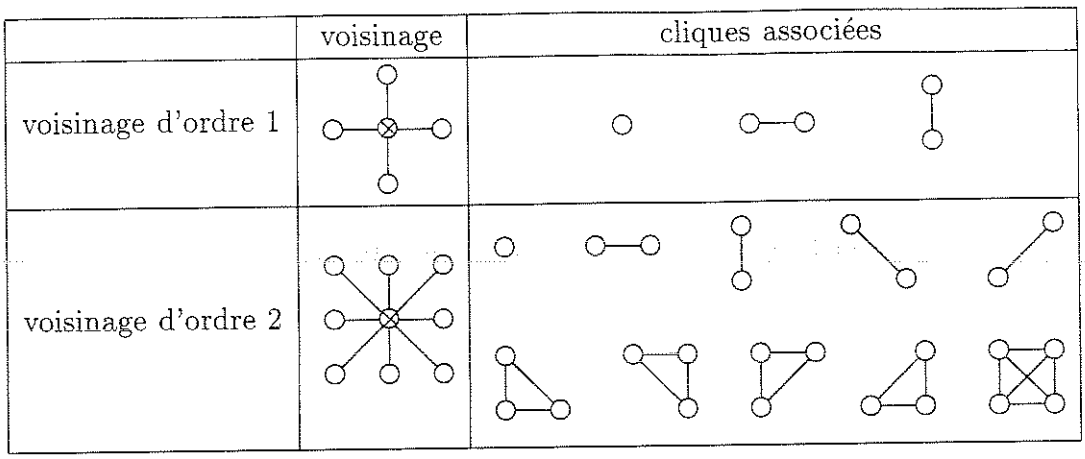


FIG. A.1 - *Exemple de voisinage et de cliques associées.*

E est un champ de Markov par rapport au système de voisinage $\mathcal{G}$ s'il vérifie les deux propriétés suivantes:

$$a) \qquad \forall e \in \Omega, \;\; p(E = e) > 0 \tag{A.4}$$

$$b) \quad \forall s \in S, \;\; p(E_s = e_s | E_r = e_r, r \neq s) = p(E_s = e_s | E_r = e_r, r \in \mathcal{G}_s). \tag{A.5}$$

La première hypothèse signifie qu'aucune configuration n'est interdite *a priori* , et la seconde correspond à l'hypothèse markovienne, à savoir qu'en chaque site le conditionnement par tous les autres sites est équivalent au conditionnement par les seuls voisins. D'après le théorème établi par Hammersley et Clifford [Bes74], E suit une distribution de Gibbs, soit:

$$p(E = e) = \frac{1}{Z} e^{-U_1(e)} \tag{A.6}$$

dans laquelle:

- Z est une constante de normalisation appelée fonction de partition:

$$Z = \sum_{e \in \Omega} e^{-U_1(e)} \tag{A.7}$$

- $U_1(e)$, la fonction d'énergie, est de la forme:

$$U_1(e) = \sum_{c \in C} V_c(e) \tag{A.8}$$

où V_c, appelé potentiel d'interaction, ne dépend que des étiquettes des sites de c. C'est la définition de ces potentiels qui permettra d'attribuer des propriétés *a priori* sur le champ des étiquettes.

En supposant que $\forall e, \forall o, p(o|e) > 0$, et en notant $U_2(e,o) = -\ln(p(o|e))$, on voit que $\hat{e}$ définie en (A.3) correspond alors au minimum de la fonction d'énergie $U(o,e) = U_1(e) + U_2(e,o)$:

$$\hat{e} = \underset{e \in \Omega}{\operatorname{argmin}} \ (\ U_1(e) + U_2(e,o) \ ) \tag{A.9}$$

A.2 Algorithmes de minimisation

La minimisation de l'expression précédente (A.9) est un problème d'optimisation en général difficile. Le cardinal de Ω trop important $(card(\Lambda)^{card(S)})$ rend toute recherche exhaustive impossible. De plus, la fonction U étant généralement une fonction non- convexe, il n'est pas possible de déterminer des solutions analytiques. On a donc recours à des méthodes itératives qui se partagent en méthodes stochastiques et méthodes déterministes.

A.2.1 Algorithmes de relaxation stochastiques

L'optimisation par relaxation stochastique, connue sous le nom de recuit simulé, permet théoriquement d'atteindre assymptotiquement un minimum global de la fonction d'énergie [GG84]. De fait, une suite T_n de températures de limite 0 est utilisée pour contrôler le processus. À chaque étape n, un algorithme de type Monte Carlo comme par exemple l'algorithme de Métropolis présenté sur la figure A.2, initialisé avec la configuration terminale de l'étape $n-1$, est appliqué à la distribution de Gibbs:

$$p_{T_n}(e,o) = \frac{1}{Z_{T_n}} e^{-\frac{U(e,o)}{T_n}} \tag{A.10}$$

À haute temperature, cette distribution de probabilité est relativement "plate" et uniforme sur Ω. Au fur et à mesure que la température décroit, les pics correspondant aux modes (solutions les plus probables) s'accentuent et les configurations obtenues avec l'algorithme de Métropolis se concentrent autour de ces modes. On montre que si la descente de température est suffisamment lente, l'algorithme converge vers l'un des minima globaux de la fonction d'énergie, et ceci, indépendamment de la configuration initiale.
Même si les conditions de refroidissement ne sont pas respectées (en pratique, on utilise souvent un recuit simulé par paliers de température décroissant exponentiellement), l'algorithme converge vers des solutions de très bonne qualité, mais dans un temps qui reste rédhibitoire pour un grand nombre d'applications, notamment pour le traitement de séquences d'images.

A.2.2 Algorithmes de relaxation déterministes

Pour éviter le problème de la lenteur de convergence, des algorithmes de relaxation déterministes ont été dérivés des méthodes de recuit simulé comme celle définie précédemment. Seules les transitions de la configuration courante vers une configuration d'énergie

- $x \leftarrow$ configuration initiale
- Répéter
  - tirer aléatoirement un site $s \in S$
  - tirer aléatoirement une étiquette $\lambda \in \Lambda$
  - $x' \leftarrow x$
  - $x'_s \leftarrow \lambda$
  - calculer $\Delta U = \frac{1}{T_n} \sum_{c \in \mathcal{C}: s \in c} (V_c(x') - V_c(x))$
  - si $\Delta U < 0$ réactualiser x avec x'
  - si $\Delta U \geq 0$
    - tirer un nombre aléatoire a dans [0,1]
    - si $a < e^{-\Delta U}$, réactualiser x avec x'

FIG. A.2 - *Algorithme de Métropolis à temperature T_n avec balayage aléatoire des sites. On suppose que la fonction d'énergie peut se mettre sous la forme $U(x) = \sum_{c \in C} V_c(x)$. On note x la configuration courante. Après un grand nombre d'itérations, x constitue une réalisation de la distribution de Gibbs associée à U/T_n.*

inférieure sont autorisées. Aucune "remontée" d'énergie n'étant admise par l'algorithme, l'algorithme convergera plus vite, mais vers une solution stable correspondant simplement à un minimum local de la fonction d'énergie. Vue sous un autre angle, cette approche consiste à remplacer dans la configuration courante x, l'étiquette courante au site s par un des modes de la distribution conditionnelle locale connaissant $x_t, t \in \mathcal{G}_s$, d'où le nom ICM qui lui est donné [Bes86]: Iterated Conditional Modes.

Le principal inconvénient de cet algorithme est sa convergence vers un minimum local proche de la configuration initiale. Il est donc essentiel de bien définir cette dernière. On peut par exemple, dans le cadre de la segmentation au sens du mouvement, fournir à l'algorithme une projection (dans le sens du mouvement) de la carte de segmentation obtenue à l'instant précédent. Dans les approches multirésolutions, la carte obtenue à une résolution sert généralement d'initialisation au niveau plus fin suivant. C'est également cette stratégie qui est retenue dans le cadre de l'approche multiéchelle décrite plus loin. [Bla89] propose également une approche permettant d'éviter en partie le problème de la convergence vers un minimum local. La méthode consiste à construire, à l'aide d'un paramètre de contrôle, une suite d'approximations de l'énergie à minimiser qui converge vers cette dernière. Par l'intermédiaire de cette suite d'approximations, la non-convexité de la fonction à minimiser est ainsi introduite progressivement dans le processus de minimisation (d'où son nom *"Graduated Non-Convexity"*, *GNC*). Plus précisément, la première approximation est choisie convexe, et son minimum global peut donc s'obtenir par des méthodes

d'optimisation classiques. Chacune des approximations est alors minimisée en utilisant comme condition initiale la solution trouvée lors de la minimisation de l'approximation précédente. La principale limitation de cette technique réside dans la construction de la séquence des fonctions d'énergie, qui n'est pas toujours aisée dans un contexte général (plusieurs types de primitives, additions de termes énergétiques, etc).

Stratégie de visite des sites

Bien qu'un simple balayage séquentiel des sites entraîne généralement dans le cas déterministe des effets directionnels qui ne sont jamais totalement éliminés, même lorsque le sens de balayage est périodiquement modifié, c'est ce mode de visite qui est le plus souvent retenu. Par ailleurs, bien que l'on soit assuré d'atteindre un minimum en un temps fini avec l'algorithme ICM, la convergence peut être lente et non significative lors des derniers balayages. Il est alors souvent judicieux d'user de critères d'arrêt. Généralement, on considère que l'algorithme a convergé lorsque le pourcentage de sites ayant changé d'étiquette entre deux ou plusieurs balayages successifs est inférieur à un seuil τ.

Dans [CR87, CB88], Chou et Brown proposent une stratégie de visite par pile d'instabilité, HCF (pour "Highest Confidence First") qui produit généralement de meilleures solutions qu'un simple balayage, et en un temps plus court. L'idée est de remettre à jour les sites par instabilité décroissante. A partir d'une configuration initiale x, l'instabilité est calculée en chaque site s suivant:

$$\text{Instab}(s) = \underset{\lambda \in \Lambda}{\text{argmax}} \sum_{c \in \mathcal{C}: s \in c} (V_c(x) - V_c(x_{s,\lambda})) \tag{A.11}$$

où $x_{s,\lambda}$ est une configuration identique à x en tout point, sauf au site s où elle vaut λ. On construit alors une *pile d'instabilité*, dans laquelle les sites instables (i.e. tels que Instab(s)>0) sont classés par valeur décroissante de leur instabilité. A chaque pas de l'algorithme, le site le plus instable est remis à jour. Ses voisins, dont l'instabilité se trouve alors modifiée, sont retirés de la pile, puis réinsérés (ou non) dans celle-ci conformément à leur nouvelle instabilité. L'algorithme s'arrête lorsque la pile est vide, tous les sites se trouvant alors dans une situation stable localement. Notons ici un inconvénient de cette méthode: de par sa structure, cet algorithme n'est que très faiblement parallélisable. En le transportant sur une machine spécialisée, on ne peut espérer obtenir un gain en rapidité comparable à celui que l'on obtient en portant un algorithme ICM classique sur une machine parallèle.

A.3 Modèles markoviens multiéchelles et relaxation multigrille

La présentation qui suit est largement inspirée de [PH93, Per93, PHB94] auxquels nous renvoyons le lecteur pour un exposé plus complet. Dans cette partie, nous nous

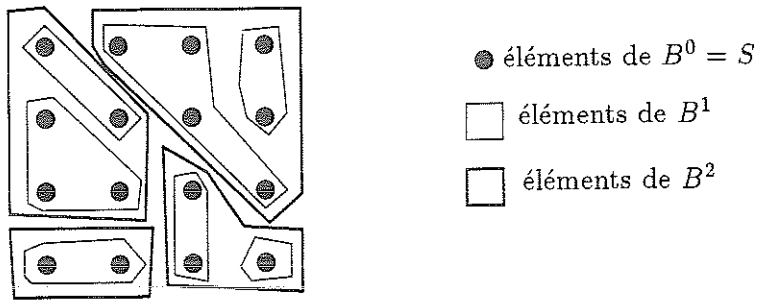


FIG. A.3 - *Exemple de suite de partitions hiérarchisées.*

contenterons d'un système de voisinage d'ordre 2 où seules les cliques à un (cliques unaires) et deux éléments (cliques binaires) sont utilisées, et nous supposerons que $U_2(e, o)$ s'écrit sous la forme:

$$U_2(e, o) = \sum_{s \in S} W_s(e_s, o_s), \tag{A.12}$$

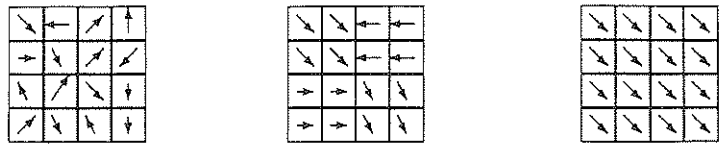
ce qui sera le cas des modèles définis dans cette thèse.
L'idée de l'approche multiéchelle est de résoudre le problème (A.9) dans des espaces de configurations emboîtés $\Omega^n \subset \ldots \subset \Omega^i \subset \ldots \subset \Omega^0 = \Omega$ correspondant à des configurations de plus en plus "fines". Les espaces Ω^i peuvent se définir de la façon suivante. Soit $B^i = \{B^i_k\}_{k=1,\ldots,N_i}$ une suite de partitions hiérarchisée de S, c'est-à-dire telle que chaque élément B^i_k soit formé par la réunion d'éléments de la partition au niveau plus fin précédent B^{i-1}. Par définition, on aura $B^0 = S$. Bien que l'approche qui suit soit indépendante du choix de cette suite (la figure A.3 montre un exemple quelconque de suite de partitions), nous la présenterons dans le cas le plus communément utilisé, celui où les B^i_k sont des blocs réguliers de taille $2^i \times 2^i$.
Le sous-ensemble Ω^i des champs d'étiquettes à l'échelle i est alors défini comme l'ensemble des configurations constantes sur les blocs B^i_k, soit [2]:

$$\Omega^i = \left\{ e \in \Omega \,/\, \forall k \in \{1, \ldots, N_i\},\ \exists e^i_k \in \Lambda \,/\, \forall s \in B^i_k,\ e_s = e^i_k \right\} \tag{A.13}$$

La figure A.4 présente un exemple de sous-espaces de ce type, où les étiquettes sont des vecteurs vitesse. L'approche consiste à reformuler l'énergie U en exploitant la structure particulière des éléments e de Ω^i. Tout d'abord, notons que U_2 peut s'exprimer sous la forme:

$$U_2(e, o) = \sum_{B^i_k \in B^i} W^i_k(e^i_k, o) \quad \text{avec} \quad W^i_k(e^i_k, o) = \sum_{s \in B^i_k} W_s(e^i_k, o_s) \tag{A.14}$$

2. Ceci est le cas le plus simple et le plus naturel dans le cas d'étiquettes symboliques. Lorsque l'on estime un champ numérique, il est possible de considérer des espaces plus complexes, comme par exemple celui des solutions linéaires par blocs.



$e \in \Omega^0 = \Omega$ $e \in \Omega^1$ $e \in \Omega^2$

FIG. A.4 - *Exemples de configurations de Ω^0, Ω^1, Ω^2 (les étiquettes sont ici des vecteurs vitesse).*

Ensuite, observons qu'une clique c de $\mathcal{C}$ se trouve (avec les cliques considérées ici):

- soit à l'intérieur d'un bloc B_k^i. On note alors $\mathcal{C}_k^i$ l'ensemble des cliques incluses dans le bloc B_k^i.
- soit à cheval sur deux blocs. On note de même $\mathcal{C}_{k_1,k_2}^i$ l'ensemble des cliques dont les intersections avec $B_{k_1}^i$ et $B_{k_2}^i$ sont non-vides (et donc, incluses dans $B_{k_1}^i \cup B_{k_2}^i$).

On peut remarquer que les $\mathcal{C}_k^i$ et $\mathcal{C}_{k_1,k_2}^i$ forment une partition de $\mathcal{C}$. De plus, la structure de voisinage $\mathcal{G}$ sur S induit une structure de voisinage $\mathcal{G}_B^i$ sur B^i définie par:

$$B_k^i \text{ et } B_l^i \text{ sont voisins } \Leftrightarrow \exists c \in \mathcal{C}/c \cap B_k^i \neq \emptyset \text{ et } c \cap B_l^i \neq \emptyset \tag{A.15}$$

On note $\mathcal{C}_B^i$ l'ensemble des cliques associées à ce système de voisinage. Avec ces définitions, on peut alors écrire U_1 sous la forme:

$$U_1(e) = \sum_{\{B_k^i\} \in \mathcal{C}_B^i} V_k^i(e) \quad + \sum_{\{B_{k_1}^i, B_{k_2}^i\} \in \mathcal{C}_B^i} V_{k_1,k_2}^i(e) \quad \text{, avec} \tag{A.16}$$

$$V_k^i(e) = \sum_{c \in \mathcal{C}_k^i} V_c(e) \quad \text{, et} \tag{A.17}$$

$$V_{k_1,k_2}^i(e) = \sum_{c \in \mathcal{C}_{k_1,k_2}^i} V_c(e) \tag{A.18}$$

À partir des potentiels (A.8) et (A.12), il est donc possible d'obtenir une expression simplifiée de l'énergie totale U pour les configurations e de Ω^i. Nous allons voir maintenant comment cette expression simplifiée s'interprète en terme de pyramide.

Lien avec l'approche pyramidale

En remarquant d'une part qu'une configuration e de Ω^i ne présente qu'une seule étiquette e_k^i à estimer par bloc B_k^i, et que d'autre part l'énergie U_1 se décompose sur des cliques formées de blocs B_k^i (formule (A.16)), on peut alors identifier la partition B^i à une grille S^i de N_i sites, où chaque site k de S^i représente le bloc B_k^i de B^i. S^i est naturellement dotée d'un système de voisinage $\mathcal{G}^i$ découlant directement de (A.15), et on note $\mathcal{C}^i$ les cliques que ce système engendre. Une configuration e de Ω^i peut alors s'interpréter comme

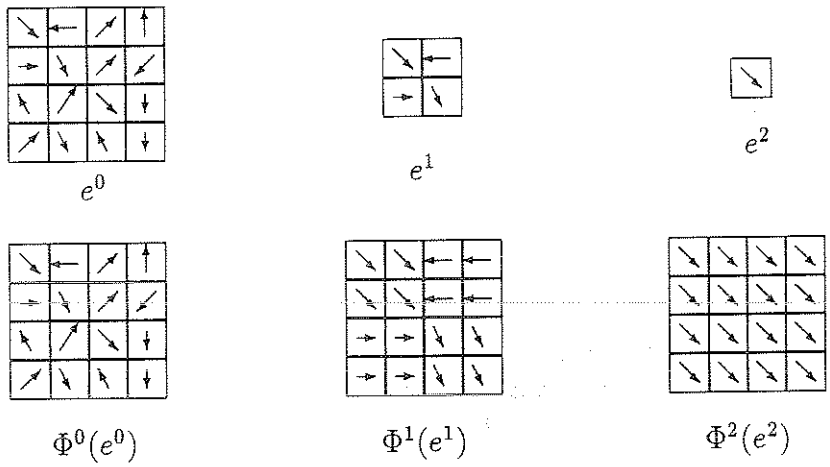


FIG. A.5 - *Exemples de configurations $e^i \in \Gamma^i$, et les $\Phi^i(e^i) \in \Omega^i$ qui leurs sont associées.*

la réalisation e^i d'un champ d'étiquettes E^i défini sur S^i, où e^i_k est l'étiquette du bloc correspondant au site k de S^i. L'ensemble des configurations de E^i est $\Gamma^i = \Lambda^{S^i}$, qui est isomorphe à Ω^i. On note alors Φ^i la bijection canonique de Γ^i sur Ω^i (exemples sur la figure A.5):

$$\begin{aligned} \Phi^i : \quad \Gamma^i &\rightarrow \Omega^i \\ e^i &\mapsto e = \Phi^i(e^i). \end{aligned} \tag{A.19}$$

Les partitions B^i permettent donc bien de définir une structure pyramidale (figure A.6) des étiquettes, où l'écriture de l'énergie U^i correspondant au champ de Markov E^i à chaque échelle i est entièrement déterminée par la définition de l'énergie à la résolution maximale. Cette énergie s'écrit:

$$U^i(e^i, o) = U^i_1(e^i) + U^i_2(e^i, o) \quad \text{où} \tag{A.20}$$

- $U^i_2(e^i, o) = U_2(\Phi^i(e^i), o) = \sum_{k \in S^i} W^i_k(e^i_k, o)$ avec $W^i_k()$ défini en (A.14).
- $U^i_1(e^i) = U_1(\Phi^i(e^i)) = \sum_{c^i \in C^i} V^i_{c^i}(e^i)$, les potentiels $V^i_{c^i}$ se déduisant directement de (A.17) et (A.18).

Grace à cette structure pyramidale, la recherche de solutions dans les espaces Ω^i se trouve simplifiée. En effet, le support S^i sur lequel est défini E^i est beaucoup plus petit que S, et les potentiels $V^i_{c^i}$ liés à son système de voisinage se déterminent analytiquement à partir de la donnée des potentiels V_c. Le nombre de données à manipuler pour calculer U_1 est donc considérablement réduit lorsque l'on travaille dans Γ^i au lieu de Ω^i. Pour ce qui est de U_2, notons que comme l'indique la figure A.6 ou la formule (A.14), les observations utilisées sont toujours celles à la résolution la plus fine, contrairement à ce qui est fait dans les méthodes pyramidales classiques (multirésolution). La réduction de calcul liée au terme U_2 ne sera alors vraiment effective que s'il est envisageable de précalculer les

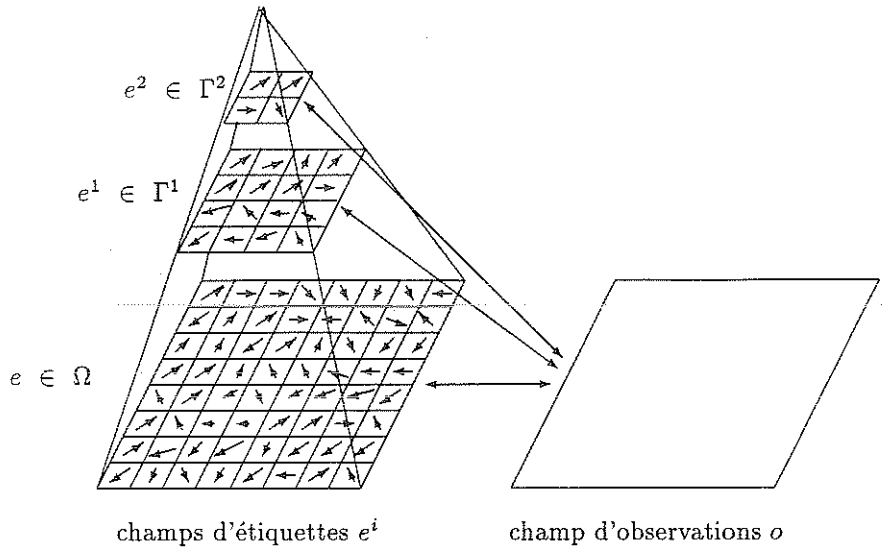


FIG. A.6 - *Structure multigrilles des étiquettes, et relation avec les observations*

pyramides $(W^i)_{i\in\{0,\ldots,n\}}$ avec $W^i = (W^i_k)_{k\in S^i}$ pour toutes les étiquettes de Λ. Si cela est réalisable dans le cas binaire de la détection, il n'en va pas de même pour l'estimation de champs de déplacements ou pour la segmentation au sens du mouvement, où le nombre d'étiquettes est trop important ou/et imprévisible. Chaque modification d'une étiquette e^i_k dans l'un des schémas de relaxation que nous avons présentés, nécéssitera alors le calcul de toutes les observations (et énergies associées) dans un bloc de taille $2^i \times 2^i$.

Enfin, la structure pyramidale est exploitée avec une stratégie descendante pour traiter le problème d'optimisation initial (A.9). À partir d'une configuration de départ au niveau n, choisie arbitrairement ou par minimisation du terme U^n_2, on obtient par relaxation déterministe(ICM) une solution $\hat{e}^n$ à l'échelle n. La projection au niveau inférieur de cette solution sert alors d'initialisation à la relaxation à l'échelle $n-1$. Le processus est alors reproduit d'échelle en échelle comme l'indique la figure A.7, jusqu'à atteindre au niveau 0 la solution finale $\hat{e}$.

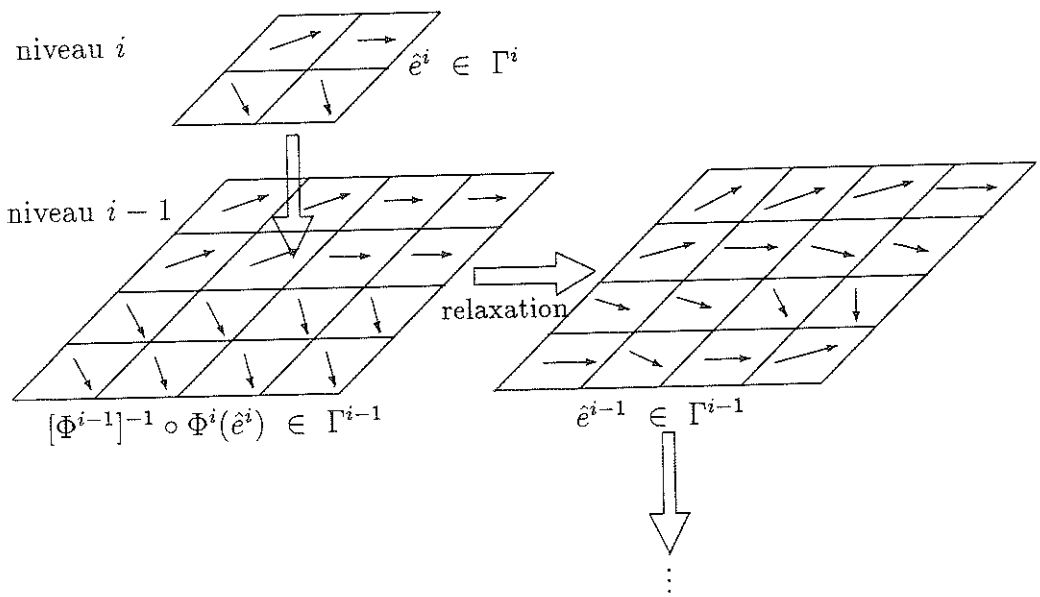


FIG. A.7 - *Stratégie multiéchelle descendante.*

Annexe B

Détermination de la borne minimale de l'observation

B.1 Proposition

- Si $P_i, i = 1, \ldots, n$ sont n pixels d'une région R animée d'un mouvement constant (translation uniforme) $\vec{\delta}$ entre les images I_1 et I_2, alors la propriété suivante est vérifiée:

$$\frac{\sum_{i=1}^{n} \|\vec{\nabla}^i I\| \times |I_t^i|}{\sum_{i=1}^{n} \|\vec{\nabla}^i I\|^2} \geq \delta \times \frac{\lambda_{min}}{\lambda_{min} + \lambda_{max}} \tag{B.1}$$

avec :

- $I_t^i = I_2(P_i) - I_1(P_i)$, et I est par défaut I_1
- $\vec{\nabla}^i I = \frac{\partial I}{\partial x}(P_i)\, \vec{u_x} + \frac{\partial I}{\partial y}(P_i)\, \vec{u_y} = I_x^i\, \vec{u_x} + I_y^i\, \vec{u_y}$ dans tout repère orthonormé $(\vec{u_x}, \vec{u_y})$
- δ est le module $\|\vec{\delta}\|$ de $\vec{\delta}$
- λ_{min} et λ_{max} sont respectivement les valeurs propres minimale et maximale de la matrice M donnée par :

$$M = \begin{pmatrix} \sum_{i=1}^{n} (I_x^i)^2 & \sum_{i=1}^{n} I_x^i I_y^i \\ \sum_{i=1}^{n} I_x^i I_y^i & \sum_{i=1}^{n} (I_y^i)^2 \end{pmatrix} \tag{B.2}$$

B.2 Preuve

Dans la suite, pour alléger les notations, nous omettrons l'indice de sommation i dans les formules lorsqu'il n'existera pas d'ambiguïté.

B.2.1 Préliminaire

Si $I : (x, y) \rightarrow I(x, y)$ est une fonction de dérivée continue et $\Phi : (x', y') \rightarrow (x, y)$ est un $\mathcal{C}^1$ difféomorphisme, alors on a:

$$\begin{pmatrix} I_{x'} \\ I_{y'} \end{pmatrix} = \nabla' I = J^T \nabla I = J^T \begin{pmatrix} I_x \\ I_y \end{pmatrix} \tag{B.3}$$

où :

$$J = \begin{pmatrix} \frac{\partial x}{\partial x'} & \frac{\partial x}{\partial y'} \\ \frac{\partial y}{\partial x'} & \frac{\partial y}{\partial y'} \end{pmatrix}$$

est la matrice jacobienne de Φ. En effet, par exemple:

$$I_{x'} = \frac{\partial I}{\partial x'} = \frac{\partial I}{\partial x} \times \frac{\partial x}{\partial x'} + \frac{\partial I}{\partial y} \times \frac{\partial y}{\partial x'} = \frac{\partial x}{\partial x'} I_x + \frac{\partial y}{\partial x'} I_y$$

Dans le cas où le changement de variable se réduit à un changement de repère orthonormal, Φ s'écrit : $X = PX'$ où $X^T = (x, y)$, $X'^T = (x', y')$ et P est la matrice de passage de la base $\mathcal{B} = (\vec{u_x}, \vec{u_y})$ à la nouvelle base $\mathcal{B}' = (\vec{u_{x'}}, \vec{u_{y'}})$. La jacobienne J n'est donc rien d'autre que la matrice de passage P, donc:

$$\nabla' I = P^T \nabla I \tag{B.4}$$

Le déplacement des n pixels étant $\vec{\delta} = \delta_x \vec{u_x} + \delta_y \vec{u_y}$, chacun d'eux doit vérifier l'équation de contrainte du mouvement apparent $(\nabla I)^T \Delta = -I_t$ avec $\Delta^T = (\delta_x, \delta_y)$. Si l'on note:

$$A = \begin{pmatrix} I_x^1 & I_y^1 \\ \vdots & \vdots \\ I_x^n & I_y^n \end{pmatrix} \text{ et } B = \begin{pmatrix} I_t^1 \\ \vdots \\ I_t^n \end{pmatrix}$$

on a le système suivant:

$$A\,\Delta = -B \tag{B.5}$$

En multipliant chaque membre par A^T, et en remarquant que $A^T A = M$, on arrive à:

$$M\Delta = -A^T B \tag{B.6}$$

Or M est une matrice symétrique positive. De ce fait il existe une matrice de passage orthogonale P et une matrice diagonale D telle que:

$$M = PDP^T \text{ avec } D = \begin{pmatrix} \lambda_{max} & 0 \\ 0 & \lambda_{min} \end{pmatrix} \text{ et } \lambda_{max} \geq \lambda_{min} \geq 0 \tag{B.7}$$

En remplaçant M dans la formule (B.6) par l'expression précédente, on obtient:

$$\begin{aligned} PDP^T\Delta &= -A^T B \\ D(P^T\Delta) &= -(P^T A^T)B = -(P^T\nabla I^1, \ldots, P^T\nabla I^n)\,B \\ &= -(\nabla' I^1, \ldots, \nabla' I^n)\,B \quad \text{d'après (B.4)} \end{aligned} \tag{B.8}$$

donc

$$D\Delta' = -A'^T B \ , \ \text{ou} \ \left\{ \begin{array}{rcccc} \lambda_{max}\delta_{x'} & = & -\sum I^i_{x'} I^i_t & = & -\sum I_{x'} I_t \\ \lambda_{min}\delta_{y'} & = & -\sum I^i_{y'} I^i_t & = & -\sum I_{y'} I_t \end{array} \right. \tag{B.9}$$

dans lesquelles $\Delta'^T = (\delta_{x'}, \delta_{y'})$ sont les coordonnées du vecteur $\vec{\delta}$ dans le nouveau repère $\mathcal{B}' = (\vec{u_{x'}}, \vec{u_{y'}})$.

B.2.2 Démonstration

On a d'une part:

$$\begin{aligned} Q &= \sum_{i=1}^{n} \|\vec{\nabla}^i I\|^2 = \sum (I_x^2 + I_y^2) = \sum I_x^2 + \sum I_y^2 \\ &= \text{trace}(M) \\ &= \lambda_{max} + \lambda_{min} \end{aligned} \tag{B.10}$$

et d'autre part, le module du gradient étant indépendant du repère orthonormé dans lequel on le calcule:

$$\begin{aligned} N &= \sum_{i=1}^{n} \|\vec{\nabla}^i I\| \times |I^i_t| = \sum \left(\sqrt{\nabla^T I \, \nabla I} \right) |I_t| \\ &= \sum \left(\sqrt{(\nabla^T I \, P)(P^T \nabla I)} \right) |I_t| = \sum \left(\sqrt{(P^T \nabla I)^T (P^T \nabla I)} \right) |I_t| \\ &= \sum \left(\sqrt{(\nabla' I)^T \, \nabla' I} \right) |I_t| \quad \text{(d'après(B.4))} \ = \sum (\sqrt{I_{x'}^2 + I_{y'}^2}) |I_t| \\ &= \sum \sqrt{(I_{x'} I_t)^2 + (I_{y'} I_t)^2} \\ &\geq \sqrt{(\sum I_{x'} I_t)^2 + (\sum I_{y'} I_t)^2} = \sqrt{\lambda_{max}^2 \delta_{x'}^2 + \lambda_{min}^2 \delta_{y'}^2} \quad \text{d'après (B.9)} \\ &\geq \sqrt{\lambda_{min}^2 (\delta_{x'}^2 + \delta_{y'}^2)} = \lambda_{min} \sqrt{\delta_{x'}^2 + \delta_{y'}^2} \\ &= \lambda_{min} \times \delta \end{aligned} \tag{B.11}$$

En conclusion, on a donc bien d'après les relations précédentes:

$$\frac{\sum_{i=1}^{n} \|\vec{\nabla}^i I\| \times |I^i_t|}{\sum_{i=1}^{n} \|\vec{\nabla}^i I\|^2} = \frac{N}{Q} \geq \delta \times \frac{\lambda_{min}}{\lambda_{min} + \lambda_{max}} \tag{B.12}$$

Annexe C

Détermination des bornes minimales et maximales de l'observation dans un cas particulier

C.1 Hypothèses et position du problème

Le but est ici de trouver des bornes minimales et maximales effectives aux observations que nous nous sommes définies en (4.19), et ainsi tester dans quelle mesure celles correspondant au premier encadrement (formule 4.17) sont grossières. Pour pouvoir calculer ces nouvelles bornes, nous utiliserons une modélisation simplificatrice de la surface d'intensité[1]. Nous supposerons qu'en un pixel p de l'image passe une isophote qui sépare son voisinage en deux régions de niveaux de gris uniformes, comme l'indique la figure C.1. Plus précisément, cette isophote est modélisée localement par deux demi-segments S_0 et S_1 de longueur $\frac{1}{2}$. Les seuls points du voisinage où le gradient n'est pas nul, sont alors les points de S_0 et S_1, où le gradient a pour valeur ∇I_V. Dans ces conditions, si l'on suppose que tout le voisinage (c'est-à-dire les deux isophotes) se déplace suivant la translation $\vec{\delta} = \delta\,\vec{u}_\theta$, notre observation consiste simplement en la moyenne des amplitudes des déplacements normaux des deux segments, soit:

$$f(\theta) = \frac{\delta}{2}\left(|\vec{u}_0.\vec{u}_\theta| + |\vec{u}_{\pi/2-\theta_1}.\vec{u}_\theta|\right) \tag{C.1}$$

Nous allons maintenant rechercher les extrema de cette expression en fonction de la direction du déplacement, et nous essaierons de les relier aux valeurs propres de la matrice de la distribution des gradients locaux M décrite dans l'annexe précédente (équation (B.2)).

1. Les expériences menées dans le chapitre sur la détection sont réalisées pour vérifier que les bornes ainsi déterminées sont valables.

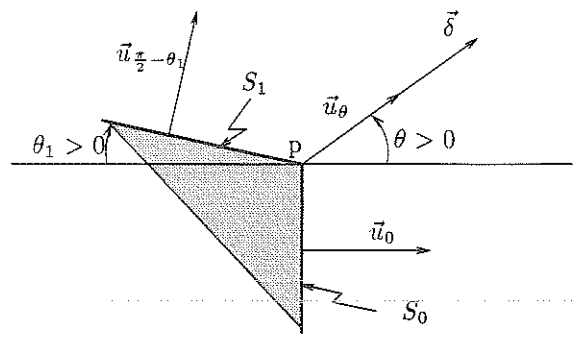


FIG. C.1 - *Modélisation de l'intensité dans le voisinage d'un pixel* p. *Une isophote formée de deux segments S_0 et S_1 sépare le voisinage en deux régions de niveaux de gris uniformes. Les seuls points du voisinage pour lesquels le gradient n'est pas nul sont donc les points de S_0 et S_1.*

C.2 Obtention des bornes

Pour des raisons évidentes de symétrie due à la modélisation, on pourra se contenter de faire l'étude pour θ_1 compris entre $-\frac{\pi}{2}$ et $+\frac{\pi}{2}$ (voir le signe de θ_1 sur la figure C.1). De la même façon, pour chaque valeur de θ_1, on pourra rechercher le minimum et le maximum de f par rapport à Θ sur l'intervalle $[0, \pi - \theta_1]$.

Premier cas : $0 < \theta_1 < \frac{\pi}{2}$ (figure C.1)

- $\theta \in [0, \frac{\pi}{2}]$:

$|\vec{u}_0.\vec{u}_\theta| = \cos\theta$ et $|\vec{u}_{\pi/2-\theta_1}.\vec{u}_\theta| = \cos(\theta - (\frac{\pi}{2} - \theta_1))$ d'où:

$$\begin{aligned} f(\theta) &= \frac{\delta}{2}\left(\cos\theta + \sin(\theta + \theta_1)\right) \\ &= \frac{\delta}{2}\left(\cos\theta\,(1 + \sin\theta_1) + \sin\theta\cos\theta_1\right) \end{aligned}$$

ce qui donne après utilisation de formules trigonométriques classiques:

$$f(\theta) = \frac{\delta}{2}\sqrt{2(1+\sin\theta_1)}\ \cos(\theta - \theta_1') \quad \text{avec} \quad \theta_1' = \frac{\pi}{4} - \frac{\theta_1}{2} \tag{C.2}$$

Sur cet intervalle, le maximum est obtenu ici pour $\theta = \theta_1'$, et le minimum pour $\theta = \frac{\pi}{2}$. Les valeurs de la fonction f en ces extrema sont (on note que $\cos(\frac{\pi}{4} + \frac{\theta_1}{2}) = \sqrt{\frac{1}{2}(1 - \sin\theta_1)}$):

$$\begin{cases} f_{Max}(\theta_1') = \frac{\delta}{2}\sqrt{2(1+\sin\theta_1)} \\ f_{min}(\frac{\pi}{2}) = \frac{\delta}{2}\sqrt{2(1+\sin\theta_1)}\cos(\frac{\pi}{4} + \frac{\theta_1}{2}) = \frac{\delta}{2}\ \cos\theta_1 \end{cases} \tag{C.3}$$

- $\theta \in [\frac{\pi}{2}, \pi - \theta_1]$:

$|\vec{u}_0.\vec{u}_\theta| = -\cos\theta$ et $|\vec{u}_{\pi/2-\theta_1}.\vec{u}_\theta| = \cos(\theta - (\frac{\pi}{2} - \theta_1))$ donc:

$$\begin{aligned} f(\theta) &= \frac{\delta}{2}\left(\cos\theta\,(\sin\theta_1 - 1) + \sin\theta\cos\theta_1\right) \\ &= \frac{\delta}{2}\sqrt{2(1-\sin\theta_1)}\,\cos(\theta - \theta_1'') \quad \text{avec} \quad \theta_1'' = \frac{3\pi}{4} - \frac{\theta_1}{2} \end{aligned} \tag{C.4}$$

Sur cet intervalle, le maximum est atteint en $\theta = \theta_1''$, le minimum en $\theta = \frac{\pi}{2}$ ou $\theta = \pi - \theta_1$, et les valeurs de la fonction sont:

$$\begin{cases} f_{Max}(\theta_1'') = \frac{\delta}{2}\sqrt{2(1-\sin\theta_1)} < f_{Max}(\theta_1') \\ f_{min}(\frac{\pi}{2}) = \frac{\delta}{2}\cos\theta_1 \end{cases} \tag{C.5}$$

Deuxième cas : $-\frac{\pi}{2} < \theta_1 < 0$

En étudiant de la même façon la fonctionnelle f dans ce cas, on montre que les maxima locaux sont atteints de même pour les directions correspondant aux bissectrices des deux segments, et les minima pour les directions des deux segments. Les valeurs extrêmes de la fonction sont alors:

$$\begin{cases} f_{Max} = \frac{\delta}{2}\sqrt{2(1-\sin\theta_1)} \\ f_{min} = \frac{\delta}{2}\cos\theta_1 \end{cases} \tag{C.6}$$

Conclusion

Quel que soit le contour paramétré par θ_1, on montre que l'observation est comprise entre les bornes l_{θ_1} et L_{θ_1}:

$$\boxed{\begin{cases} l_{\theta_1} = \frac{\delta}{2}\cos\theta_1 \\ L_{\theta_1} = \frac{\delta}{2}\sqrt{2(1+|\sin\theta_1|)} \end{cases}} \tag{C.7}$$

C.3 Valeurs propres de M_{θ_1}

Les termes de la formule (B.2) se calculent facilement dans notre cas, les sommes étant remplacées par des intégrales le long des segments S_0 et S_1. En effet, on a en chaque point Q de S_0:

$$\vec{\nabla}_{S_0}(Q) = \nabla I_V\,\vec{u}_0 = \nabla I_V\,\vec{i}$$

et en chaque point de S_1:

$$\vec{\nabla}_{S_1}(Q) = \nabla I_V\,\vec{u}_{\pi/2-\theta_1} = \nabla I_V\,(\sin\theta_1\,\vec{i} + \cos\theta_1\,\vec{j})$$

d'où par exemple:

$$\int_{S_0 \cup S_1} I_x^2 \, dx = (\nabla I_V)^2 (\frac{1}{2} + \frac{1}{2} \sin^2 \theta_1)$$

et donc la matrice M_{θ_1} est donnée par:

$$M_{\theta_1} = \frac{(\nabla I_V)^2}{2} \begin{pmatrix} 1 + \sin^2 \theta_1 & \cos \theta_1 \sin \theta_1 \\ \cos \theta_1 \sin \theta_1 & \cos^2 \theta_1 \end{pmatrix} \tag{C.8}$$

Les valeurs propres de M_{θ_1} sont les racines de l'équation caractéristique:

$$\lambda^2 - \lambda \, \text{trace} M_{\theta_1} + \det M_{\theta_1} = 0, \text{ soit } \lambda^2 - (\nabla I_V)^2 \lambda + \frac{\cos^2 \theta_1}{4} (\nabla I_V)^4 = 0$$

On trouve alors les valeurs propres suivantes:

$$\boxed{\begin{cases} \lambda_{\Theta_1 min} = \frac{1 - |\sin \theta_1|}{2} (\nabla I_V)^2 \\ \lambda_{\Theta_1 max} = \frac{1 + |\sin \theta_1|}{2} (\nabla I_V)^2 \end{cases}} \tag{C.9}$$

C.4 Bornes sur l'observation en fonction des valeurs propres de M

En pratique, nous disposons d'une matrice M calculée sur un voisinage V du pixel p. Pour pouvoir exploiter les formules de (C.7) qui nous donnent les bornes sur l'observation, nous devons réussir à exprimer $\cos \theta_1$ et $|\sin \theta_1|$ en fonction des valeurs propres λ_{min} et λ_{max} de la matrice M. Pour cela nous procédons à l'identification des matrices M et M_{θ_1}. Nous avons alors par exemple (il n'y a que deux paramètres indépendants):

$$\begin{cases} \text{trace} M = \text{trace} M_{\theta_1} \quad \text{d'où } (\nabla I_V)^2 = \lambda_{min} + \lambda_{max} \quad \text{et donc} \\ \lambda_{min} = \lambda_{\Theta_1 min} = \frac{1 - |\sin \theta_1|}{2} (\lambda_{min} + \lambda_{max}) \end{cases} \tag{C.10}$$

En notant:

$$\lambda'_{min} = \frac{\lambda_{min}}{\lambda_{min} + \lambda_{max}} \tag{C.11}$$

on obtient alors:

$$\begin{cases} |\sin \theta_1| = 1 - 2 \lambda'_{min} \\ \cos \theta_1 = \sqrt{1 - |\sin \theta_1|^2} = 2 \sqrt{\lambda'_{min} (1 - \lambda'_{min})} \end{cases}$$

On en déduit alors les bornes minimale et maximale de notre observation en un pixel p de l'image en fonction des valeurs propres de la matrice M en ce même point:

$$\boxed{\begin{cases} l_m(P) = \delta \sqrt{\lambda'_{min} (1 - \lambda'_{min})} \\ L_m(P) = \delta \sqrt{1 - \lambda'_{min}} \end{cases}} \tag{C.12}$$

Bibliographie

[AB85] E.H Adelson and J.R. Bergen. Spatio-temporal energy models for the perception of motion. *Jal Optical Society of America A*, Vol.2, No.2:284–299, Feb. 1985.

[ACJ90] R. Azencott, B. Chalmond, and Ph. Julien. Bayesian 3-d path search and its application to focusing seismic data. P. Barone, A. Frigessi, and M. Piccioni, editors, In Lecture Notes in Statistics, volume 74:*Stochastic Models, Statistical Methods, and Algorithms in Image Analysis*, pages 46–74, Springer-Verlag, 1990.

[AD93] M. Allmen and C.R. Dyer. Computing spatiotemporal relations for dynamic perceptual organisation. *CVGIP: Image Understanding*, 58(3):338–351, November 1993.

[Adi85] G. Adiv. Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Trans. Pattern Anal. Machine Intell.*, Vol 7:384–401, July 1985.

[Adi89] G. Adiv. Inherent ambiguities in recovering 3D motion and structure from a noisy flow field. *IEEE Trans. Pattern Anal. Machine Intell.*, 11(5):477–489, May 1989.

[AKM93] T. Aach, A. Kaup, and R. Mester. Statistical model-based change detection in moving video. *Signal Processing*, 31:165–180, 1993.

[AN88] J.K. Aggarwal and N. Nandhakumar. On the computation of motion from sequences of images- a review. *Proc. of the IEEE*, Vol.76, No.8:917–935, August 1988.

[Ana89] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, Vol 2:283–310, 1989.

[Anc92] N. Ancona. A fast obstacle detection method based on optical flow. In *Second European Conference on Computer Vision, Santa Margherita Ligure, Italy*, pages 267–271, May 1992.

[AP82] D.K. Arrowsmith and C.M. Place. *Ordinary differential equations.* Chapman and Hall, 1982.

[Arn92] Yves Arnaud. *Caractérisation des nuages précipitants en fonction de leur structure spatiale et de leur évolution temporelle en milieu sahelien partir d'images METEOSAT.* Thèse de l'université Paul Sabatier, Toulouse, 1992.

[AS94] S. Ayer and P. Schroeter. Multiple motion estimation by robust parameter estimation over multiple frames. In *Proc 7th European Signal Processing Conference (EUSIPCO)*, pages 700–703, September 1994.

[ASB94] S. Ayer, P. Schoeter, and J. Bigün. Segmentation of moving objects by robust motion parameter estimation over multiple frames. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 316–327, Stockholm, Sweden, May 1994.

[AW94] E.H. Adelson and J.Y.A Wang. Representing moving images with layers. *IEEE Trans. on Image Processing, Special Issue: Image sequence compression*, 5(3):625-638, 1994.

[AWB87] J. Aloimonos, I. Weiss, and A. Bandopadhay. Active vision. In *Proc. of the 1st Int. Conf. on Computer Vision*, pages 35–54, London, June 1987.

[Aze87] R. Azencott. Image analysis and Markov fields. In *ICIAM87: First International Conference on Industrial and Applied Mathematics*, pages 53–61, SIAM, Philadelphia, June 1987.

[BA91] M.J. Black and P. Anandan. Robust dynamic motion estimation over time. In *Proc. Conf. Computer Vision and Pattern Recognition*, pages 296–302, Hawaii, June 1991.

[BA93a] M.J Black and P. Anandan. A framework for the robust estimation of optical flow. In *Proc. 4th Int. Conf. Computer Vision*, pages 231–236, Berlin, May 1993.

[BA93b] M.J. Black and P. Anandan. *The robust estimation of multiple motions: affine and piecewise-smooth flow fields.* Technical Report SPL-93-092, XEROX, Palo Alto research Center, California, December 1993.

[BAD93] J. Bulas-Cruz, A.T. Ali, and E.L. Dagless. Real time motion detection and tracking. In *Proc. 4^{th} Int. Conference on Computer Vision*, pages 512–522, Berlin, May 1993.

[BAHH92] J.R. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proc. European Conf. Computer Vision*, pages 237–252, Springer-Verlag, S.Margherita Ligure, Italy, 1992.

[Baj88] R. Bajcsy. Active perception. *Proceedings of the IEEE*, Vol.76, No.8:996–1005, August 1988.

[BAK91] R. Battiti, E. Amaldi, and C. Koch. Computing optical flow across multiple scales: an adaptative coarse-to-fine strategy. *Intern. J. Comput. Vis.*, 6:2:133–145, 1991.

[BBDM94] B. Bascle, P. Bouthemy, N. Deriche, and F. Meyer. Tracking complex primitives in an image sequence. In *Proc. 12th Int. Conf. on Pattern Recognition*, pages 426–431, Jerusalem, October 1994.

[BBH*89] P.J. Burt, J. R. Bergen, R. Hingorani, R. Kolczynski, W.A. Lee, A. Leung, J. Lubin, and H. Shvaytser. Object tracking with a moving camera. *IEEE Workshop on Visual Motion*, 2–12, March 1989.

[BBHP90] J.R. Bergen, P.J. Burt, R. Hingorani, and S. Peleg. Computing two motions from three frames. In *Proc. 3rd Int. Conf. on Computer Vision*, pages 27–32, Osaka, Dec. 1990.

[BBHP92] J.R. Bergen, P.J. Burt, R. Hingorani, and S. Peleg. A three-frame algorithm for estimating two-component image motion. *IEEE Trans. Pattern Anal. Machine Intell.*, 14(9):886–896, December 1992.

[Bes74] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Royal Statistical Society*, Serie B, Vol.36:192–236, 1974.

[Bes86] J. Besag. On the statistical analysis of dirty pictures. *Journal Royal Statistic Society*, Vol.48, Serie B, No.3:259–302, 1986.

[BF93] P. Bouthemy and E. François. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *Int. Journal of Computer Vision*, Vol.10, No 2:157–182, April 1993.

[BFB92] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. *Performance of Optical Flow techniques.* Technical Report 299, Departement of Computer Science, University of Western Ontario, London, Ontario, Canada N6A 5B7, July 1992.

[BFB94] J.L. Barron, D.J. Fleet, and S.S. Beauchemin. Performance of Optical Flow techniques. *International Journal of Computer Vision*, 12(1):43–77, January 1994.

[BHK91] P.J. Burt, R. Hingorani, and R.J. Kolczynski. Mechanisms for isolating component patterns in the sequential analysis of multiple motion. In *IEEE Workshop on Visual Motion*, pages 187–193, Princeton, October 1991.

[BK94] M. Bober and J. Kittler. Robust motion analysis. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 947–952, Seattle, Washington, June 1994.

[BL90] P. Bouthemy and P. Lalande. Detection and tracking of moving objects based on statistical regularization method in space and time. In *Proc. European Conf. Computer Vision*, pages 307–311, Springer, Antibes, April 1990.

[Bla89] A. Blake. Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. *IEEE Trans. Pattern Anal. Machine Intell.*, Vol.11, No.1:2–12, Jan. 1989.

[Bla92] M. J. Black. *Robust incremental optical flow.* N° 923, Yale University, Computer Science Dept, September 1992.

[Bla94] M.J. Black. Recursive non-linear estimation of discontinuous flow fields. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 138–145, Stockholm, Sweden, May 1994.

[BMDF89] M. Bertrand, J. Meunier, M. Doucet, and G. Ferland. Ultrasonic biomechanical strain gauge based on speckle tracking. In *IEEE Ultrasonic Symposium, Montréal, Canada*, October 1989.

[BMM*89] M. Bertrand, J. Meunier, G. Mailloux, M. Doucet, and R. Petitclerc. Measurement of soft tissue deformation using echographic speckle tracking. In *Computer Assisted Radiology (CAR), Berlin*, June 1989.

[BNDZ93] T.M. Bernard, P.E. NGuyen, F.J. Devos, and B.Y. Zavidovique. A programmable VLSI retina for rough vision. *Machine Vision and Applications*, 7(1):4–11, Fall 1993.

[BPT88] M. Bertero, A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proc. of the IEEE*, Vol.76, No.8:869–889, August 1988.

[BR94] M.J. Black and A. Ramgarajan. The outlier process: unifying line process and robust statistics. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 15–22, Seattle, Washington, June 1994.

[Bra74] O. Braddick. A short-range process in apparent motion. *Vision Research*, 14:519–527, 1974.

[BS87] P. Bouthemy and J. Santillana Rivero. A hierarchical likelihood approach for region segmentation according to motion-based criteria. In *Proc. 1st Int. Conf. on Computer Vision*, pages 463–467, Londres, 1987.

[Bur84] P.J. Burt. The pyramid as a structure for efficient computation. In A. Rosenfeld, editor, *Multiresolution Image Processing and Analysis*, pages 6–35, Springer-Verlag, 1984.

[CB88] P.B. Chou and C.M. Brown. Multimodal reconstruction and segmentation with Markov random fields, HCF optimization. *Proc. Image Understanding Workshop, Cambridge Massachusetts*, 214–221, April 1988.

[CB90] P.B. Chou and C.M. Brown. The theory and practice of Bayesian image modeling. *Int. Jal of Computer Vision*, Vol.4:185–210, 1990.

[CBBJ94] F. Chaumette, S. Boukir, P. Bouthemy, and D. Juvin. Optimal estimation of 3D structures using visual servoing. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 347–354, Seattle, Washington, USA, June 1994.

[Cha88] B. Chalmond. Image restoration using an estimated Markov model. *Signal Processing*, 15(2):115–129, September 1988.

[CK83] N. Cornelius and T. Kanade. Adapting optical flow to measure object motion in reflectance and x-ray image sequences. *ACM SIGGRAPH/SIGART Interdisciplinary Workshop on Motion, Toronto*, 50–58, April 1983.

[Coh93] I. Cohen. Nonlinear variational method for optical flow computation. In *Proc. 8th Scandinavian Conference on Image Analysis*, pages 523–530, Tromso, Norway, June 1993.

[CQB92] C. Collet, A. Quinquis, and J.M. Boucher. Cloudy sky velocity estimation based on optical flow estimation leading with an entropy criterion. In *Proc. IEEE Int. Conf. Pattern Recognition*, pages 160–163, The Hague, 1992.

[CR87] P.B. Chou and R. Raman. *On relaxation algorithms based on Markov random fields.* Technical Report TR. 212, Computer Science department, Univ. of Rochester, Rochester, New-York, July 1987.

[CSR83] A Cowart, W. Snyder, and W. Rudger. The detection of unresolved targets using the Hough transform. *Computer Vision, Graphics and Image Processing*, Vol.CVGIP-21:222–238, 1983.

[CST94] M.M. Chang, M.I. Sezan, and A.M. Tekalp. An algorithm for simultaneous motion estimation and scene segmentation. In *Proc. of ICCASP 94*, pages 221–224, Adelaide, Australia, June 1994.

[CV90] M. Campani and A. Verri. Computing optical flow from an overconstrained system of linear algebraic equations. In *Proc. 3rd Int. Conf. on Computer Vision*, pages 22–26, Osaka, Dec. 1990.

[DGV92] Jacques Droulez, Alexey Grantyn, and Pierre-Paul Vidal. Les bases neuronales du contrôle oculomoteur. *Le courrier du CNRS*, 79:52, May 1992.

[Die91] N. Diehl. Object-oriented motion estimation and segmentation in image sequences. *Signal Processing: Image communication*, 3:23–56, 1991.

[DJ93] C.F. Dhu and R.C. Jain. Direct estimation and errors analysis for oriented patterns. *CVGIP: Image Understanding*, 58(3):383–398, November 1993.

[DP91] T. Darrell and A. Pentland. Robust estimation of a multi-layered motion representation. In *Proc. IEEE Workshop on Visual Motion*, pages 173–178, Princeton, Oct. 1991.

[EC84] Hildreth E.C. Computations underlying the measurement of visual motion. *Artificial Intelligence*, Vol. 23:309–354, 1984.

[ECR92] B. Espiau, F. Chaumette, and P. Rives. A new approach to visual servoing. *IEEE Trans. on Robotics and Automation*, Vol.RA-8, No.3:313–326, June 1992.

[Enk88] W. Enkelmann. Investigations of multigrid algorithms for the estimation of optical flow fields in image sequences. *Computer Vision, Graphics and Image Processing*, Vol.43:150–177, 1988.

[Fau93] O. Faugeras. *Three-dimensional computer vision: a geometric viewpoint.* The MIT Press, 1993.

[FB90] E. François and P. Bouthemy. Derivation of qualitative information in motion analysis. *Image and Vision Computing Journal*, Vol.8, No.4:279–287, Nov. 1990.

[FJ90] D.J. Fleet and A.D. Jepson. Computation of component image velocity from local phase information. *Int. Jal of Computer Vision*, Vol.5, No.1:77–104, 1990.

[FM91] C.S. Fuh and P. Maragos. *Affine Models for Image Matching and Motion Detection.* Technical Report CICS-P-280, Center for Intelligent Control Systems, February 1991.

[FMRS94] G.L. Foresti, P. Matteucci, C.S. Regazzoni, and S. Spaggiari. A visual surveillance system for autonomous vehicle risk avoidance. In *7th European Signal Processing Conference*, pages 1369–1373, Edinburgh, Scotland, UK, September 1994.

[Fra91] Edouard François. *Interprétation qualitative du mouvement à partir d'une séquence d'images.* Thèse de l'université de Rennes I, June 1991.

[Fra92] Nicolas Franceschini. L'intelligence visuo-motrice. *Le courrier du CNRS*, 79:47, May 1992.

[GD94] D. Geiger and K.J. Diamantras. Occlusion ambiguities in motion. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 175–180, Stockholm, Sweden, May 1994.

[GG84] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, Vol.6, No.6:721–741, Nov. 1984.

[GP93] S. Gil and T. Pun. Non-linear multiresolution relaxation for alerting. In H. Dedieu, editor, *European Conference on Circuits Theory and Design*, pages 1639–1644, Elsevier Science Publishers B.V., 1993.

[GS94] F. Germain and T. Skordas. An image motion estimation technique based on combined statistical test and spatiotemporal generalized likelihood ratio approach. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 152–157, Stockholm, Sweden, May 1994.

[HA92] I. Herlin and N. Ayache. Features extraction and analysis methods for sequences of ultrasound images. In *Second European Conference on Computer Vision (ECCV)*, pages 43–57, Santa Margherita Ligure, Italy, May 1992.

[Han91] K.J. Hanna. Direct multi-resolution estimation of ego-motion and structure from motion. In *Workshop on Visual Motion*, pages 152–156, Princeton, NJ, October 1991.

[HB93] F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using Markov random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol 15(12):1217–1232, Dec. 1993.

[Hea93] Glenn Healey. Hierarchical segmentation-based approach to motion analysis. *Image and Vision Computing*, 11(9):570–576, November 1993.

[Hee88] D.J. Heeger. Optical flow using spatio-temporal filters. *Int. Journal of Computer Vision*, Vol.1, No.4:279–302, Jan. 1988.

[Hei93] F. Heitz. Contribution à l'analyse statistique de l'image: modèles, algorithmes et applications. Document d'habilitation à diriger des recherches, IRISA - IFSIC, Université de Rennes I, Octobre 1993.

[HH91] J.L. Helman and L. Hesselink. Visualizing vector field topology in fluid flow. *IEEE Computer Graphics & Applications*, 36–46, May 1991.

[HJ83] S.M. Haynes and R.C. Jain. Detection of moving edges. *Computer Vision, Graphics and Image Processing*, 21:345–367, 1983.

[HNR84] Y. Z. Hsu, H.-H Nagel, and G. Rekers. New likelihood test methods for change detection in image sequences. *Computer Vision, Graphics and Image Processing*, Vol.26:73–106, 1984.

[Hoe89] M. Hoetter. Differential estimation of the global motion parameters zoom and pan. *Signal Processing*, Vol.16:249–265, 1989.

[Hor86] B.K.P. Horn. *Robot vision.* MIT Press, 1986.

[HRRS86] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics : The Approach Based on Influence Functions.* John Wiley and Sons, New York, 1986.

[HS81] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, Vol.17:185–203, 1981.

[HT81] T.S. Huang and R.Y. Tsai. Image sequence analysis : motion estimation. In T.S. Huang, editor, *Image Sequence Processing and Dynamic Scene Analysis*, Springer-Verlag, 1981.

[Hua83] T.S. Huang, editor. *Image Sequence Processing and Dynamic Scene Analysis.* Volume F2 of *NATO-ASI Series*, Springer, New-York, 1983.

[Hub81] P.J. Hubert. *Robust statistics.* Wiley, 1981.

[HW77] P.W. Holland and R.E. Welsch. Robust regression using iteratively reweighted least squares. *Commun. Stat.- Theor. Meth.*, A6:813–828, 1977.

[HW88] B.K.P Horn and E.J. Weldon. Direct methods for recovering motion. *Int. Jal of Computer Vision*, Vol.2:51–76, 1988.

[IO83] K. Imaichi and K. Ohmi. Numerical processing of flow-visualization pictures – measurement of two-dimensional vortex flow. *J. Fluid Mech.*, 129:283–311, 1983.

[Iou94] Anatoli Iouditski. Communication personnelle, 1994.

[IRP91] M. Irani, B. Rousso, and S. Peleg. *Detection of multiple moving objects using temporal integration.* Technical Report 91-14, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel, December 1991.

[IRP92] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In *Proc. of 2nd ECCV-92, S.Margherita Ligure, Italy*, pages 282–287, Springer-Verlag, May 1992.

[IRP94] Michal Irani, Benny Rousso, and S. Peleg. Recovery of ego-motion using image stabilization. In *Proc. of Comp. Vision and Pattern Recognition Conference*, pages 454–460, Seatle (Washington) - USA, June 1994.

[Jai85] J.R. Jain. Dynamic scene analysis. *in Progress in Pattern Recognition 2*, L. Kanal and A. Rosenfeld (eds.):125–167, 1985.

[JB93] A. Jepson and M.J. Black. Mixture models for optical flow computation. In *Proc. of the Computer Vision and Pattern recognition conference, CVPR-93*, pages 760–761, New-York, USA, June 1993.

[JMB91] J.M. Jolion, P. Meer, and S. Batauche. Robust clustering with application in computer vision. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13:791–802, August 1991.

[KD75] J.J. Koenderink and J.J. Van Doorn. Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer. *Optica Acta*, Vol.22, No.9:73–791, 1975.

[KD88] J. Konrad and E. Dubois. Multigrid Bayesian estimation of image motion fields using stochastic relaxation. In *Proc. 2nd Int. Conf. Computer Vision*, pages 354–362, Tarpon Springs, Florida, Dec. 1988.

[KD90] J. Konrad and E. Dubois. Comparison of stochastic and deterministic solution methods in Bayesian estimation of 2d motion. *Image and Vision Computing*, 8(4):304–317, November 1990.

[Key81] R.G. Keys. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-29(6):1153–1160, December 1981.

[KH94] C. Kervrann and F. Heitz. A hierarchical statistical framework for the segmentation of deformable objects in image sequences. In *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pages 724–728, Seattle, USA, June 1994.

[Koz89] E.J. Kozlow. David: advanced developments for the next generation of video intrusion detection. *Proc. Int. Carnahan Conf. on Security Technology, Zurich*, 145–147, April 1989.

[KT94a] Janusz Konrad and Paolo Treves. Estimation of dense 2-d motion based on the constancy of intensity gradient. In *Proc. of the 7th European Signal Processing Conference*, pages 684–687, Edinbourgh, September 1994.

[KT94b] Janusz Konrad and Paolo Treves. Motion estimation and compensation under varying illumination. In *Proc. of the 1st Int. Conf. On Image Processing*, Austin, November 1994.

[KvG90] K.-P. Karmann, A. v.Brandt, and R. Gerl. Moving object segmentation based on adaptive reference images. In *Signal Process. V: Theories and Applications (Proc. Fifth European Signal Process. Conf.)*, pages 951–954, Barcelona, September 1990.

[KW87] M. Kass and A. Witkin. Analysing oriented patterns. *CVGIP*, 37:362–385, 1987.

[KWM94] D. Koller, J. Weber, and J. Malik. Robust multiple car tracking with occlusion reasoning. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 189–198, Stockholm, Sweden, May 1994.

[Lal90] Patrick Lalande. *Détection du mouvement apparent dans une séquence d'images selon une approche markovienne; Application à la robotique sous-marine.* Thèse de l'université de Rennes I, March 1990.

[LB90] P. Lalande and P. Bouthemy. A statistical approach to the detection and tracking of moving objects in an image sequence. *Proc. 5th Conf. Eusipco, Barcelona*, 947–950, Sept. 1990.

[Let93] Jean-Michel Létang. *Intégration temporelle et régularisation statistique appliquées la détection d'objets mobiles dans une séquence d'images.* Thèse de l'Institut National Polytechnique de Grenoble, 1993.

[LF94] H. Li and R. Forchheimer. Two-view facial movement estimation. *IEEE Trans. on Circuits and Systems for Video Technology*, 4(3):276–287, June 1994.

[LHZ89] C. Lee, R.M. Haralick, and X. Zhuang. Recovering 3d motion parameters from image sequences with gross-errors. In *IEEE Workshop on Visual Motion*, pages 46–53, Irving California, March 1989.

[LJ89] S.-P. Liou and R.C. Jain. Motion detection in spatio-temporal space. *Computer Vision, Graphics and Image Processing*, Vol.CVGIP-45:227–250, 1989.

[LK81] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proc. of 7th. IJCAI*, pages 674–679, Vancouver,B.C., Canada, 1981.

[LNC71] J.A. Leese, C.S. Novak, and B.B. Clark. An automated technique for obtaining cloud motion from geosynchronous satellite data using cross correlation. *Journal of applied meteorology*, 10:118–132, 1971.

[LP80] H.C. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. *Proc. Roy. Soc. Lond.*, B-208:385–397, April 1980.

[LPC94] F. Luthon, G.V. Popescu, and A. Caplier. An MRF based motion detection algorithm implemented on analog resistive network. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 167–174, Stockholm, Sweden, May 1994.

[LRB93] J.M. Létang, V. Rebuffel, and P. Bouthemy. Motion detection robust to perturbations: a statistical regularization and temporal integration framework. In *Proc. 4th Int. Conf. Computer Vision*, pages 21–30, Berlin, May 1993.

[Mau94] Mariette Maurizot. Caractérisation, localisation et suivi de points singuliers dans un champ de vecteurs 2d. Rapport de stage - IRISA, Juillet 1994.

[MB87] D.W. Murray and H. Buxton. Scene segmentation from visual motion using global optimization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.9, No.2:220–228, March 1987.

[MB92] F. Meyer and P. Bouthemy. Estimation of time-to-collision maps from first order motion models and normal flows. In *Proc. 11th Intern. Conf. on Pattern Recognition, The Hague*, pages 78–82, 1992.

[MB94a] F.G. Meyer and P. Bouthemy. Region-based tracking using affine motion models in long image sequences. *CVGIP: Image Understanding*, Vol. 60(2):119–140, September 1994.

[MB94b] A. Mitiche and P. Bouthemy. Computation and analysis of image motion: a synopsis of current problems and methods. *Intern. J. Comput. Vis.*, 1994. Accepted for publication.

[Mem93] Etienne Mémin. *Algorithmes et architectures parallèles pour les approches markoviennes en analyse d'images.* Thèse de l'université de Rennes I, June 1993.

[Mey93] François Meyer. *Suivi de régions et analyse des trajectoires dans une séquence d'images.* Thèse de l'université de Rennes I, June 1993.

[MF90] V. Markanday and B.E. Flinchbaugh. Multispectral constraints for optical flow computation. In *Proc. 3rd Int. Conf. on Computer Vision*, pages pp 38–41, Osaka, Dec. 1990.

[ML78] L.B. Milstein and T. Lazicky. Statistical tests for image tracking. *Computer Graphics and Image Processing*, Vol.CGIP-7:413–424, 1978.

[MLSB89] G.E. Mailloux, F. Langlois, P.L. Simard, and M. Bertrand. Restoration of the velocity field of the heart from two-dimensional echocardiograms. *IEEE Trans. on Medical Imaging*, 11(2):143–153, June 1989.

[MMP87] J. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journ. of the American Statistical Association*, Vol.82, No.397:76–89, Mars 1987.

[MMR91] P. Meer, D. Mintz, and A. Rosenfeld. Robust regression methods for computer vision: a review. *International Journal of Computer Vision*, 6(1):59–70, 1991.

[MRM94] P.F. MacLauchlan, I.D. Reid, and D.W. Murray. Recursive affine structure and motion from image sequences. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 217–224, Stockholm, Sweden, May 1994.

[MWA87] A. Mitiche, Y.F. Wang, and J.K. Aggarwal. Experiments in computing optical flow with the gradient-based, multiconstraint method. *Pattern Recognition*, Vol.20, No.2:173–179, 1987.

[MZ92] J.W. Modestino and J Zhang. A Markov random field model-based approach to image interpretation. *IEEE Trans. Pattern Anal. Machine Intell.*, 14(6):606–615, June 1992.

[NA89] C. Nelson and J. Aloimonos. Obstacle avoidance using flow field divergence. *IEEE Trans. Pattern Anal. Machine Intell.*, Vol.11, No.10:1102–1106, Oct. 1989.

[Nag83] H.H. Nagel. Displacement vectors derived from second-order intensity variations in image sequences. *Computer Vision, Graphics and Image Processing*, Vol 21:85–117, 1983.

[Nag87] H.-H. Nagel. On the estimation of optical flow : relations between different approaches and some new results. *Artificial Intelligence*, Vol.33:299–324, 1987.

[Nag88a] H.H. Nagel. From image sequences towards conceptual descriptions. *Image and Vision Computing Jal*, Vol.6, No.2:pp 59–74, May 1988.

[Nag88b] H.H. Nagel. Image sequences - ten (octal) years - from phenomenology towards a theoretical foundation. *Int. Jal of Pattern Recognition and Artificial Intelligence*, 2(3):459–483, 1988.

[Nak85] K. Nakayama. Biological image motion processing: a review. *Vision Research*, Vol.25, No.5, 1985.

[NE86] H.H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.8:565–593, Sept. 1986.

[Nel91] R.C. Nelson. Qualitative detection of motion by a moving observer. *Int. Journal of Computer Vision*, Vol.7, No.1:33–46, 1991.

[NL91] H. Nicolas and C. Labit. Global motion identification for image sequence analysis and coding. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 2825–2828, Toronto, May 1991.

[NL92] S. Negahdaripour and S. Lee. Motion recovery from image sequences using only first order optical flow information. *Intern. J. Comput. Vis.*, 9(3):163–184, 1992.

[NLO94] E. Nguyen, C. Labit, and J-M. Odobez. A ROI approach to hybrid image sequence coding. In *1st Int. Conference on Image Processing*, Austin, Texas, November 1994.

[NSKO94] H.-H. Nagel, G. Socher, H. Kollnig, and M. Otte. Motion boudary detection in image sequences by local stochastic tests. In *Proc. European Conf. Computer Vision*, pages 305–315, Spriger-Verlag, Stockholm, Sweden, May 1994.

[OB94] J.M. Odobez and P. Bouthemy. Estimation robuste multi-échelle de modèles paramétrés de mouvement sur des scènes complexes. In *actes du IX Congrès AFCET de Reconnaissance des Formes et Intelligence Artificielle*, pages 419–430, PARIS, mai 1994.

[ON94] M. Otte and H.-H. Nagel. Optical flow estimation: advances and comparisons. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 51–60, Stockholm, Sweden, May 1994.

[PC87] A.E. Perry and M.S. Chong. A description of eddying motions and flow patterns using critical point concepts. *Ann. Rev. Fluid Mech.*, 19:125–155, 1987.

[Per93] Patrick Pérez. *Champs markoviens et analyse multirésolution de l'image: application à l'analyse du mouvement.* Thèse de l'université de Rennes I, Juin 1993.

[PH91] A. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol-PAMI 13, No. 7:730–742, July 1991.

[PH93] P. Pérez and F. Heitz. Une approche multiéchelle de l'analyse d'images par champs markoviens. *Traitement du Signal*, 9(6):459–472, 1993.

[PHB94] P. Perez, F. Heitz, and P. Bouthemy. Multiscale minimization of global energy functions in some visual recovery problems. *CVGIP : Image Understanding*, 59(1):125–134, January 1994.

[PR90] S. Peleg and H. Rom. Motion based segmentation. In *Proc. 10th IEEE Conf. on Pattern Recognition*, pages 109–113, Atlantic City, 1990.

[PS94] M. Pardas and P. Salambier. Time-recursive segmentation of image sequences. In *Proc. of the 7th European signal processing conference (EUSIPCO)*, pages 18–21, Eurasip, Edinburgh (Scotland), UK, September 1994.

[RJ92] A.R. Rao and R.C. Jain. Computerized flow field analysis: oriented texture fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 14(7):693–709, July 1992.

[RL87] P.J. Rousseeuw and A.M. Leroy. *Robust regression and outlier detection.* John Wiley, 1987.

[Rou84] P.J. Rousseeuw. Least median of squares regression. *Jal of the American Statistical Association*, Vol.79, No 388:871–880, Dec. 1984.

[SB91] T.S. Huang S.D. Blostein. Detecting small, moving objects in image sequences using sequential hypothesis testing. *IEEE Trans. on Signal Processing*, Vol.39, No.7:1611–1629, July 1991.

[SB94] D. Sinclair and B. Boufama. Independent motion segmentation and collision prediction for road vehicles. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 161–166, Stockholm, Sweden, May 1994.

[SBC94] V. Sundareswaran, P. Bouthemy, and F. Chaumette. Active camera self-orientation using dynamic image parameters. In *Proc. of 3rd European Conference on Computer Vision (ECCV)*, pages 111–116, Stockholm, Sweden, May 1994.

[SF93] R.N. Strickland and R.M. Ford. Image models for 2d flow visualization and compression. In *Proc. of the 8th Scandinavian Conference on image analysis*, pages 183–190, Tromso, May 1993.

[Sin90] A. Singh. An estimation-theoretic framework for image-flow computation. *Proc. 3rd Int. Conf. on Computer Vision, Osaka*, 168–177, Dec. 1990.

[Sin92] A. Singh. Incremental estimation of image-flow using a Kalman filter. *Jal of Visual Communication and Image Representation*, Vol.3, No.1:39–57, March 1992.

[SJ89] K. Skifstad and R. Jain. Illumination independent change detection for real world image sequences. *Computer Vision, Graphics and Image Processing*, Vol.46:387–399, 1989.

[SM89] J. Schmetz and M.S. Mhita. Diurnal and interdiurnal variability of IR and WV brightness temperatures from Meteosat. *ESA Journal*, 13:329–341, 1989.

[SM91] M. Shizawa and K. Mase. Principle of superposition: a common computational framework for analysis of multiple motion. In *Proc. IEEE Workshop on Visual Motion*, pages 164–172, Princeton, October 1991.

[SN87] J. Schmetz and M. Nuret. Automatic tracking of high-level clouds in Meteosat infrared images with a radiance windowing technique. *ESA Journal*, 11:275–286, 1987.

[Sti93] C. Stiller. A statistical image model for motion estimation. In *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, pages 193–196, Mineapolis, USA, April 1993.

[Sub87] M. Subbarao. Solution and uniqueness of image flow equations for rigid curved surfaces in motion. In *Proc. 1st Int. Conf. Computer Vision*, pages 687–692, London, England, 1987.

[Sub89] M. Subbarao. Interpretation of image flow: a spatio-temporal approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.11, No.3:266–278, 1989.

[Sub90] M. Subbarao. Bounds on time-to-collision and rotational component from first-order derivatives of image flow. *CVGIP*, Vol.50:329–341, 1990.

[Sun92] V. Sundareswaran. A fast method to estimate sensor translation. In *Second European Conference on Computer Vision (ECCV)*, pages 253–257, Santa Margherita Ligure, Italy, May 1992.

[TB89] R. Thoma and M. Bierling. Motion compensating interpolation considering covered and uncovered background. *Signal Processing : Image communication*, Vol.1(2):191–212, October 1989.

[TBM85] W.B. Thompson, V.A. Berzins, and K.M. Mutch. Dynamic occlusion analysis in optical flow fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.7, No.4:374–383, July 1985.

[Tis94] M. Tistarelli. Multiple constraints for optical flow. In *Proc. of the 3rd European Conference on Computer Vision (ECCV)*, pages 61–70, Stockholm, Sweden, May 1994.

[TK87] W.B. Thompson and J.K. Kearney. Optical flow estimation: an error analysis of gradient-based methods with local optimization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.9, No.2:229–244, 1987.

[TL91] A. Tamtaoui and C. Labit. Constrained disparity and motion estimators for 3D TV image sequence coding. *Signal Processing: Image Communication*, Vol.4:45–54, 1991.

[TL94] G. Tziritas and C. Labit. *Motion analysis for image sequence coding.* Volume 4 of series *Advances in Image Communication*, Elsevier, 1994.

[TLS93] W.B. Thompson, P. Lechleider, and E.R. Stuck. Detecting moving objects using the rigidity constraint. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2(15):162–166, February 1993.

[TM93] P.H.S. Torr and D.W. Murray. Statistical detection of independent movement from a moving camera. *Image and Vision Computing*, 11(4):180–187, May 1993.

[TMCZ80] J.K. Tsotsos, J. Mylopoulos, H.D. Covvey, and S.W. Zucker. A framework for visual motion understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.2, No. 6:563–573, Nov. 1980.

[TP84] O. Tretiak and L. Pastor. Velocity estimation from image sequences with second order differential operators. In *Proc. 7th Int. Conf. on Pattern Analysis and Machine Intelligence*, pages 16–19, Montréal, July 1984.

[TP90] W.B. Thompson and T.-G. Pong. Detecting moving objects. *Int. Journal of Computer Vision*, Vol.4:39–57, 1990.

[TS91] M. Tistarelli and G. Sandini. Direct estimation of time-to-impact from optical flow. In *Proc. of the IEEE Workshop on Visual Motion*, pages 226–233, Princeton, Oct. 1991.

[VF92] T. Viéville and O. Faugeras. Robust and fast computation of unbiased intensity derivatives in images. In *Proc. of 2nd ECCV-92, S.Margherita Ligure, Italy*, pages 203–212, May 1992.

[VGT89] A. Verri, F. Girosi, and V. Torre. Mathematical properties of the two-dimensional motion field: from singular points to motion parameters. *Journal of Optical Society of America*, Vol.6, No.5:698–712, May 1989.

[VP89] A. Verri and T. Poggio. Motion field and optical flow: qualitative properties. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.11, No.5:490–498, May 1989.

[WBBH92] H. Wang, C. Bowman, M. Brady, and C. Harris. A parallel implementation of a structure-from-motion algorithm. In *Proc. European Conf. Computer Vision*, pages 272–276, Springer-Verlag, Santa Margherita Ligure, Italy, May 1992.

[WK93] S. F. Wu and J. Kittler. A gradient-based method for general motion estimation and segmentation. *Jal of Visual Communication and Image Representation*, 4(1):25–38, March 1993.

[WU85] A.M. Waxman and S. Ullman. Surface structure and three-dimensional motion from image flow kinematics. *Int. Journal of Robotics Research*, Vol.4, No.3:72–94, 1985.

[YC90] G.S. Young and R. Chellappa. 3D motion estimation using a sequence of noisy stereo images : models, estimation, and uniqueness results. *IEEE Trans on Pattern Anal. and Machine Intel.*, Vol.12 No 8:735–759, August 1990.

[You88] L. Younes. *Problème d'estimation paramétrique pour les champs de Gibbs Markoviens - Application au traitement d'images.* Thèse de l'Université de Paris-Sud, Orsay, France, 1988.

[Zav92] B. Zavidovique. Des rétines en silicium pour une vision des robots. *Le courrier du CNRS*, 79:54–55, May 1992.

[ZHA94] J. Zhong, T.S. Huang, and R.J. Adrian. Salient structure analysis of fluid flow. In *Proc. of the Computer Vision and Pattern Recognition Conference*, pages 310–315, Seattle, Washington, USA, June 1994.

[ZP90] H. Zabrodsky and S. Peleg. Attentive transmission. *Journal of Visual Communications and Image Representation*, 1:189–198, 1990.

[ZQY89] W.-Z. Zhao, F.-H. Qi, and T.Y. Young. Dynamic estimation of optical flow field using objective functions. *Image and Vision Computing*, Vol.7, No.4:259–267, Nov. 1989.

Liste des publications

• J.M. ODOBEZ ET P. BOUTHEMY, Robust multiresolution estimation of parametric motion models, *Accepté avec révisions mineures pour publication dans la revue Journal of Visual Communication and Image Representation.*

• J.M. ODOBEZ ET P. BOUTHEMY, Estimation robuste multirésolution de modèles paramétrés de mouvement sur des scènes complexes, *Accepté pour publication dans la revue Traitement du Signal.*

• J.M. ODOBEZ ET P. BOUTHEMY, Robust multiresolution estimation of parametric motion models, *Dans les actes de la* $7^{ième}$ *European Signal Processing Conference (EUSIPCO 94)*, pages 411-415, Edimbourg, Septembre 1994.

• J.M. ODOBEZ ET P. BOUTHEMY, Detection of multiple moving objects using multiscale Markov random fields with camera compensation, *Dans les actes de la* $1^{ière}$ *IEEE International Conference on Image Processing (ICIP)*, volume II, pages 257-261, Austin, Novembre 1994.

• E. NGUYEN, C. LABIT ET J.M. ODOBEZ, A ROI approach for hybrid image sequence coding, *Dans les actes de la* $1^{ière}$ *IEEE International Conference on Image Processing (ICIP)*, volume III, pages 245-249, Austin, Novembre 1994.

• J.P. LEDUC, J.M. ODOBEZ ET C. LABIT, Motion-compensated adaptative wavelet filtering for image sequence processing, *À paraître dans les actes de International Conference on Acoustics, Speech and Signal Procecessing (ICASSP)*, Detroit, Mai 1995.

• J.M. ODOBEZ ET P. BOUTHEMY, Estimation robuste multi-échelle de modèles paramétrés de mouvement sur vdes scènes complexes, $9^{ième}$ *conférence AFCET Reconnaisance des Formes et Intelligence Artificielle*, pages 419-430, Paris, France, Janvier 1994. Le prix "jeune chercheur" (en Reconnaissance de Forme) m'a été décerné par l'AFCET pour cette communication.

• M. MAURIZOT, P. BOUTHEMY, B. DELYON, A. IOUDITSKI ET J.M. ODOBEZ, Locating singular points and characterizing deformable flow fields in an image sequence. *soumis à* 5^{th} *International Conference on Computer Vision (ICCV)*, Cambridge (Mass.), juin 1995.

• J.M. ODOBEZ ET P. BOUTHEMY, Robust multiresolution estimation of parametric motion models applied to complex scenes, Publication Interne No. 788, IRISA, Janvier 1994.

• M. MAURIZOT, P. BOUTHEMY, B. DELYON, A. IOUDITSKI ET J.M. ODOBEZ Locating singular points and characterizing deformable flow fields in an image sequence, Publication Interne No. 891, IRISA, décembre 1994.

• J.M. ODOBEZ ET P. BOUTHEMY, Détection et segmentation du mouvement par une approche robuste et markovienne, Actes des journées CNET-CCETT *"Nouvelles techniques pour la compression et la représentation des signaux audiovisuels"*, Rennes, janvier 1995.

Résumé

Cette thèse traite de la détection et de la localisation d'objets en mouvement dans une séquence d'images acquises par une caméra mobile. Nous motivons tout d'abord l'intérêt du problème et rappelons diverses méthodes existantes proposées pour le résoudre. L'approche que nous avons retenue pour la détection consiste à reconstruire dans un premier temps une séquence d'images, dans laquelle le déplacement apparent dans l'image induit par le mouvement de la caméra a été compensé. Pour cela, nous supposons que ce déplacement peut être décrit par un modèle paramétrique 2D. Le troisième chapitre de ce mémoire présente la méthode robuste et multirésolution que nous avons développée, qui permet d'estimer ce modèle de mouvement paramétré (dominant) dans l'image sans être affecté par la présence d'autres mouvements (ceux des objets mobiles notamment). Le problème posé se ramène alors à la détection des zones mal compensées dans la séquence ainsi reconstruite. Dans le chapitre quatre, nous définissons des mesures de compensation du mouvement adaptées à ce problème. Ces mesures et leur fiabilité, calculées à différents instants, ainsi que la carte de détection à l'instant précédent, sont prises en compte au sein d'une régularisation statistique basée sur des modèles de Markov multiéchelles. L'algorithme que nous avons défini est relativement rapide et permet d'obtenir d'excellents résultats dans des situations complexes. Dans le chapitre cinq, l'algorithme de détection précédent est étendu à la segmentation du mouvement dans une séquence d'images (gestion de n étiquettes). Le schéma complet que nous avons défini permet notamment de s'adapter au contenu dynamique de la scène, en créant de nouvelles régions lors de l'apparition de nouveaux objets dans la scène ou lorsque le mouvement d'une région donnée devient plus complexe.

Mots-clés: vision dynamique - séquence d'images - mouvement - estimation - détection - segmentation - modèles paramétrés - régularisation statistique - champ de Markov - robuste.

Abstract

This PhD thesis deals with the detection and location of moving objects in a monocular image sequence acquired by a mobile camera. The two first chapters introduce the issue and review the work related to it. The method we propose is based on the assumption that the apparent flow field of the static background, induced by the camera motion, can be modeled by a 2D parametric model and is the *dominant motion* in the image. Such a motion model is accurately estimated from frame to frame, even in the presence of secondary motions (like those of moving objects), with a multiresolution robust method which is described in chapter three; the estimated models are then used to compute a compensated sequence in which the background then appears as static. Thus, the original problem reduces to the detection of non-static entities in this reconstructed sequence. In chapter four, we propose a method which can tolerate noisy data and imprecise compensation that can occur since the model is only an approximation of the background motion. This method relies on a statistical regularization approach based on multiscale Markov random field (MRF) models. Particular attention has been paid to the definition of appropriate observations, and to the energy function which expresses the adequacy between labels and observations. More precisely, successive observations in time as well as their reliability are explicitly taken into account. In chapter five, the previous algorithm is extended to the resolution of the motion segmentation problem. The whole scheme that we have defined adapts the segmentation to the dynamic content of the sequence by creating new regions when new moving objects appear in the scene, or when the motion of a given region becomes more complex.