Numéro d'ordre : 2889

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEUX 1

ÉCOLE DOCTORALE DES SCIENCES PHYSIQUES ET DE L'INGÉNIEUR

PAR

David IZQUIERDO

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : TRAITEMENT DU SIGNAL ET D'IMAGE

****

CONTRIBUTION AU DÉVELOPPEMENT D'UNE ARCHITECTURE GÉNÉRIQUE DÉDIÉE AU SUIVI D'OBJETS EN TÉLÉSURVEILLANCE : APPLICATION AU SUIVI DE VÉHICULES ET DE VISAGES

****

Soutenue le 12 Novembre 2004

Après avis de :

MM.	**Jean-Michel JOLION,**	Professeur, INSA de Lyon	**Rapporteur**
	Murat KUNT,	Professeur, EPFL de Lausanne	**Rapporteur**

Devant la commission d'examen formée de :

MM.	**Yannick BERTHOUMIEU,**	Maître de conférences, ENSEIRB	**Rapporteur**
	Jean-Michel JOLION,	Professeur, INSA de Lyon	**Examinateur**
	Murat KUNT,	Professeur, EPFL de Lausanne	**Examinateur**
	Philippe MARCHEGAY,	Professeur, Directeur de Thèse, ENSEIRB	**Examinateur**
	Mohamed NAJIM,	Professeur, ENSEIRB	**Président**

A aquellos que más quiero :
a mis padres, a mis hermanos.

Remerciements

Ce travail a été effectué dans le cadre d'une collaboration entre le Laboratoire d'Etude de l'Intégration des Composants et Systèmes Electroniques (IXL, UMR 5818) de l'Université de Bordeaux I et le Laboratoire d'Automatique, Productique et Signal (LAPS, UMR 5131) de l'Université de Bordeaux I.

Je remercie le Professeur Philippe Marchegay pour avoir accepté de m'encadrer en tant que directeur de thèse. Je le remercie pour l'accueil et l'humanité avec lesquelles il m'a dirigé pendant cette thèse.

Je remercie M. Yannick Berthoumieu pour m'avoir accueilli dans son équipe de recherche et pour m'avoir orienté sur un sujet novateur. A travers ces quatre ans de travail et de discussions, j'ai pu beaucoup apprendre sur le traitement du signal et d'image plus spécialement sur la détection et le suivi d'objets en mouvement. Je le remercie pour l'enthousiasme et la joie avec laquelle il m'a dirigé et pour les nombreuses idées qu'il m'a fournies, me permettant de faire évoluer mon sujet de recherche dans des voies intéressantes.

Je remercie très sincèrement mes rapporteurs M. Murat Kunt, Professeur de l'Ecole Polytechnique Fédérale de Lausanne et M. Jean-Michel Jolion, Professeur de l'Institut National des Sciences Appliquées de Lyon, pour avoir eu l'amabilité d'accepter la tâche de rapporteur de ce mémoire.

Je remercie M. Mohamed Najim, Professeur de l'Ecole Nationale Supérieure d'Electronique, Informatique et Radiocommunications de Bordeaux, qui m'a fait l'honneur d'accepter de participer au jury de soutenance.

Je remercie très sincèrement toutes les personnes qui m'ont apporté leur aide essentielle dans la réalisation de cette thèse : M. Pierre Baylou pour ses corrections toujours très pertinentes et M. Mohamed Najim pour ses conseils très judicieux.

Je remercie également toutes les personnes que j'ai côtoyées pendant ces années de laboratoire pour leur aide amicale. J'aimerais remercier plus spécialement M. Yannick Berthoumieu qui, pendant ces quatre ans, a tou-

jours su se montrer patient et disponible lorsque j'avais besoin d'aide ou de conseils, et aussi pour la confiance qu'il a déposée sur moi. J'adresse un remerciement particulier aux anciens membres et étudiants du groupe : Alexander, Tchi, Anouar, Lucie, Benoît, Nicolas, Mohamed, Brahim, Kizito, Marc, Patrick, David, Sophie, Guillaume et Enoch, qui d'une façon ou d'une autre ont contribué à la bonne marche de cette thèse.

Je ne peux pas oublier tous les membres actuels de notre laboratoire, plus spécialement ceux qui ont travaillé de façon plus proche avec moi : Marco, Pierre, Regis, Mounia, Cyprian, et Sorin, en qui j'ai trouvé des collègues et amis sympathiques avec lesquels les discussions ont été très enrichissantes.

Un grand Merci à tous mes grands amis que j'ai rencontrés sur Bordeaux, Sebas, Dani, Gema, Alberto, Isabel et Irene et toute la liste des "Erasmus" qui sont passés par Bordeaux, pour tous les très bons moments que nous avons passés ensemble. Tout particulièrement, je voudrais remercier Victor, Irene, Jose et Javi pour m'avoir encouragé à finir cette thèse et m'avoir soutenu avec tout leur coeur.

Finalement je ne peux finir sans rappeler le soutien essentiel de ma famille à laquelle j'ai dédié tout ce travail. Muchas gracias por todo.

Table des matières

Introduction

Contexte

La perception visuelle, c'est-à-dire la capacité de distinguer les formes, les couleurs et les mouvements, naît dans l'oeil, mais ne prend réellement sens que dans le cerveau. La compréhension des mécanismes sous-jacents de la perception visuelle est au coeur de nombreuses recherches menées conjointement dans les domaines de la neurophysiologie, des sciences cognitives et de la vision artificielle. Pour le neurophysiologiste, l'enjeu est la compréhension anatomique du système de vision. Le cogniticien étudie l'influence de notre mémoire sur notre perception et l'élaboration du sens. Enfin, le chercheur en vision artificielle conçoit les procédés automatiques d'interprétation pour doter la machine d'une perception s'inspirant des processus naturels mis en oeuvre chez l'homme.

Le domaine de la vision artificielle a longtemps consisté à interpréter le contenu d'une image fixe. Bénéficiant de progrès technologiques, les chercheurs s'orientent à présent vers la compréhension de scènes dynamiques et notamment vers le suivi d'entités mobiles à l'instar de la vision humaine. L'implantation d'un procédé automatique de suivi dans un système électronique au sens large est ainsi le problème central de nombreuses applications. Principalement utilisé à l'origine dans un contexte militaire (suivi de cibles, guidage de missile), ce type de traitement est aujourd'hui au coeur de nombreuses applications en multimédia (compression, production vidéo), en télésurveillance et en robotique mobile. Dans les applications multimédia, il s'agit de développer des algorithmes capables de décomposer un flux vidéo en éléments indépendants pour augmenter les performances de méthodes de compression (MPEG4) ou pour faciliter l'indexation de contenus (MPEG7). En télésurveillance, l'objectif est d'extraire des informations permettant une supervision de situations réelles. La tâche de suivi peut être nécessaire en vue d'une simple mise en *alarme* conformément à une nomenclature d'états interdits ou pour alimenter un procédé plus complexe de *contrôle*. De nombreux

secteurs industriels sont très attentifs aux avancés algorithmiques dans ce domaine afin de pouvoir proposer une pléïade de services pour la surveillance de sites, le contrôle du trafic routier, l'aide aux handicapés (interprétation du langage des signes, aide à la conduite) ou pour l'interaction homme-machine (interacteur utilisant le geste humain).

Le suivi d'objets a pour but principal d'assurer la localisation spatio-temporelle d'entités mobiles. La problématique est donc de distinguer des objets dans chaque image d'un flux vidéo afin de suivre leur trajectoire. Plusieurs phénomènes peuvent rendre difficile cette opération. Un problème récurrent est l'occultation qui consiste en un masquage partiel ou total d'un objet à un instant donné de son évolution dans l'image. Le masquage est dû à la projection, inhérente au procédé d'acquisition, de la scène tridimensionnelle sur le plan image du capteur vidéo. L'occultation correspond alors à une rupture d'observabilité de l'objet, rupture qu'il s'agit de compenser. D'autres problèmes sont également sources de difficultés. Dans le cas de scènes extérieures, les conditions d'éclairage de la scène nécessitent une prise en charge des fluctuations photométriques (passages nuageux, cycle diurne). En outre, la présence d'ombre portée, notamment par fort ensoleillement, peut être à l'origine d'instabilités dans le comportement des méthodes proposées. Finalement, il apparaît que le suivi automatique n'est pas une opération triviale d'autant que la plupart des applications imposent un temps de traitement en accord avec la cadence d'acquisition.

La problématique traitée

Dans le cadre de cette thèse, nous nous intéressons à l'étude de techniques numériques en vue du développement d'un système automatique de suivi dans un flux continu d'images. Afin de bien cerner la tâche à réaliser, précisons que le suivi est la reconnaissance visuelle continue d'un objet. Cette tâche implique de *localiser* et de *reconnaître* chacun des objets présents dans chaque image. Dans ce manuscrit, la *reconnaissance* est à prendre au sens de la *mise en correspondance*. Quotidiennement, l'homme est capable de percevoir et de suivre des centaines d'objets, de reconnaître des centaines de visages, d'interpréter de nombreuses informations signalétiques et cela sans effort particulier. Cette capacité ne tient pas à la simplicité de la tâche, mais plutôt à l'extrême efficience de notre système de vision. La question est : comment doter la machine de cette aptitude à reconnaître ? Utilisant les connaissances en neurophysiologie et en psychophysique, Marr [Mar82] propose, dans le contexte de la vision artificielle, une théorie globale de la

reconnaissance s'appuyant sur deux principes fondamentaux : la modularité des traitements et la hiérarchisation de l'information. Le premier rappelle que toute opération complexe doit être divisée en sous-opérations afin d'assurer flexibilité et stabilité au procédé global. Le second principe implique que la reconnaissance d'un objet s'élabore à partir de divers niveaux de description dont les basiques sont connus sous l'appellation *primitive*, principe très exploité en traitement d'images aujourd'hui. L'intérêt de l'approche modulaire proposée par Marr, outre l'accroissement des performances des procédés, est d'offrir une réduction de données afin d'éviter une explosion combinatoire du système de reconnaissance.

Fondée sur l'utilisation des primitives, la théorie de Marr se développe principalement de manière ascendante (bottom-up) en enchaînant les étapes de segmentation, reconstruction et reconnaissance. D'autres variantes [BGG96] [AT91] favorisent un schéma descendant (Top-down) dans lequel il s'agit, non plus d'analyser, mais de reconnaître l'objet recherché à partir d'une représentation structurale et abstraite de celui-ci. Plus récemment, cherchant à modéliser le processus de reconnaissance chez l'homme, des travaux [DRMFT04] montrent que ce processus est régi par une gestion dichotomique entre la perception, système de production abstraités, et la mémoire, système de conservation des abstractions. Plus concrètement, le processus de reconnaissance se caractérise par l'intervention simultanée et interactive de processus ascendants et descendants. En outre, tous les travaux montrent de façon unanime que la reconnaissance s'appuie sur les trois postulats suivants :

1. La reconnaissance nécessite la mémorisation structurale des objets (contour, région, points caractéristiques, ...)
2. L'identification s'appuie nécessairement sur une comparaison de l'instance perçue à la description structurale stockée (distance, corrélation, mise en correspondance, prédiction, ...)
3. Le traitement dédié aux primitives est strictement ascendant.

Finalement, les postulats et le principe d'interaction entre des procédés ascendants et descendants résument de façon très schématique la modélisation admise aujourd'hui de notre système de reconnaissance. Ainsi en vision artificielle, il semble pertinent de concevoir une architecture qui reprenne ce mode de fonctionnement. Comme préconisé par Marr, nous nous intéressons à une architecture modulaire d'implantation de notre système de suivi. Plutôt que de manipuler des primitives, nous utiliserons des *entrées perceptives* car la description de l'objet exploite la perception que l'on a

de lui, mêlant couleur, forme et mouvement. Dans cette architecture, nous mettrons en évidence des interactions entre différents procédés verticaux.

Approche proposée

Le suivi d'entités à partir d'une séquence d'images a suscité une activité scientifique importante depuis une vingtaine d'années. Si un bilan est possible, il montre, en première analyse, que l'élaboration de solutions passe par la résolution de nombreux problèmes (multi vues, caméra fixe ou mobile), concerne une multitude d'environnements (intérieur, extérieur, spécifique ...), nécessite l'utilisation d'informations de types très divers (régions, contours, coins ...) et oblige à la mise en oeuvre de moyens algorithmiques très différents selon l'application (véhicules, personnes, visages ...). Cette multiplicité de cas nuit à l'établissement d'un formalisme unifié des approches proposées. Toutefois, si on considère une granularité fonctionnelle élevée, tous les travaux récents convergent vers une solution impliquant un procédé cyclique modulaire enchaînant les phases de détection, de reconnaissance, d'estimation et de prédiction.

- La *détection* est la mise en oeuvre de méthodes de segmentation générant des entrées perceptives permettant d'initialiser le procédé de suivi. Ces entrées ont pour rôle la mise en évidence de la cible. De part le contexte, le mouvement est la caractéristique la plus évidente, mais d'autres comme la couleur ou des primitives géométriques peuvent venir renforcer le caractère discriminant de la détection. Une prise en compte de plusieurs entrées nécessite naturellement une étape de fusion.
- La *reconnaissance* a pour objectif de comparer les entrées perceptives fournies par la phase de détection à une description abstraite de la cible. La production de la forme abstraite sera faite sur la base d'une prédiction. La difficulté dans cette phase d'appariement est d'obtenir un procédé robuste en cas d'occultation.
- L'*estimation* met à jour les primitives de description de l'objet. Les entrées perceptives, filtrées par la procédure de reconnaissance, sont utilisées à cet effet.
- La *prédiction* permet d'émettre une hypothèse sur l'évolution de position le long de la trajectoire de l'objet. Cette étape est indispensable à la phase de reconnaissance. Un outil de prédiction très populaire est, par exemple, le filtre de Kalman.

Les nombreuses propositions présentées dans la "littérature" sur le sujet s'appuient sur ces différentes étapes, sans forcement respecter le même ordre d'enchaînement. Toutefois, toutes les approches s'appuient sur une architecture en boucle fermée. La rétroaction, découlant de l'étape de prédiction, est en effet indispensable pour garantir une solution stable et rapide au problème de suivi.

Concernant l'étape de reconnaissance, les travaux existants se scindent en deux grandes familles. La première utilise un modèle d'apparence de l'objet fondé généralement sur une description spatiale de celui-ci. Une zone 2D, connexe ou non, lui est affectée à partir des informations issues de la phase de détection (régions de même couleur, regroupement de segments ...). Le procédé est alors purement ascendant [Bré97]. L'autre famille décrit l'objet par un modèle plus proche des caractéristiques naturelles de ce dernier. C'est le cas par exemple lorsqu'un objet rigide est modélisé par une forme géométrique 3D pré contrainte [KWM94]. L'approche modèle est par essence descendante et nécessite généralement des phases complexes de prise en compte de transformations spatiales. Cependant dans le cas d'un modèle simple, elles fournissent des résultats intéressants.

Le problème des approches utilisant un procédé ascendant ou descendant réside dans la difficulté de gérer les cas d'occultation. Dans les schémas verticaux, il n'y a pas de remise en cause a posteriori des étapes de simplification de l'information, ce qui empêche une gestion adaptée des occultations. Pour pallier ce problème, nous proposons de développer une stratégie fondée sur l'interaction de procédés ascendants et descendants :

- Les entrées perceptives sont construites à partir d'un schéma ascendant. Nous nous intéresserons tout particulièrement à la détection de mouvement ainsi qu'à l'utilisation d'une discrimination chromatique adaptée à la détection de visage, l'objectif étant d'obtenir un ensemble de *régions perceptives* représentatives de l'objet.
- Dans la phase de reconnaissance, nous associons à chaque objet un jeu de descripteurs d'apparence. L'ensemble des descripteurs est constitué par défaut d'une zone spatiale et d'un modèle de mouvement. Les descripteurs sont la représentation abstraite de l'objet. Ce sont eux qui vont représenter l'instance courante pour le développement d'une méthode de reconnaissance multi-objets de structuration descendante.

A partir des descripteurs et des régions perceptives, nous développons deux niveaux d'interaction :

1. La gestion des occultations s'appuie sur une utilisation conjointe des deux types de représentation (descripteurs et régions), l'objectif étant

le calcul de cartes d'appartenance à partir d'une modélisation stochastique de la distribution des pixels de l'image courante.

2. La mise à jour est réalisée à partir d'un nouveau partitionnement des régions perceptives guidé par les descripteurs prédits.

La Fig. 1 résume les différents flux d'informations caractérisant le procédé global de suivi.

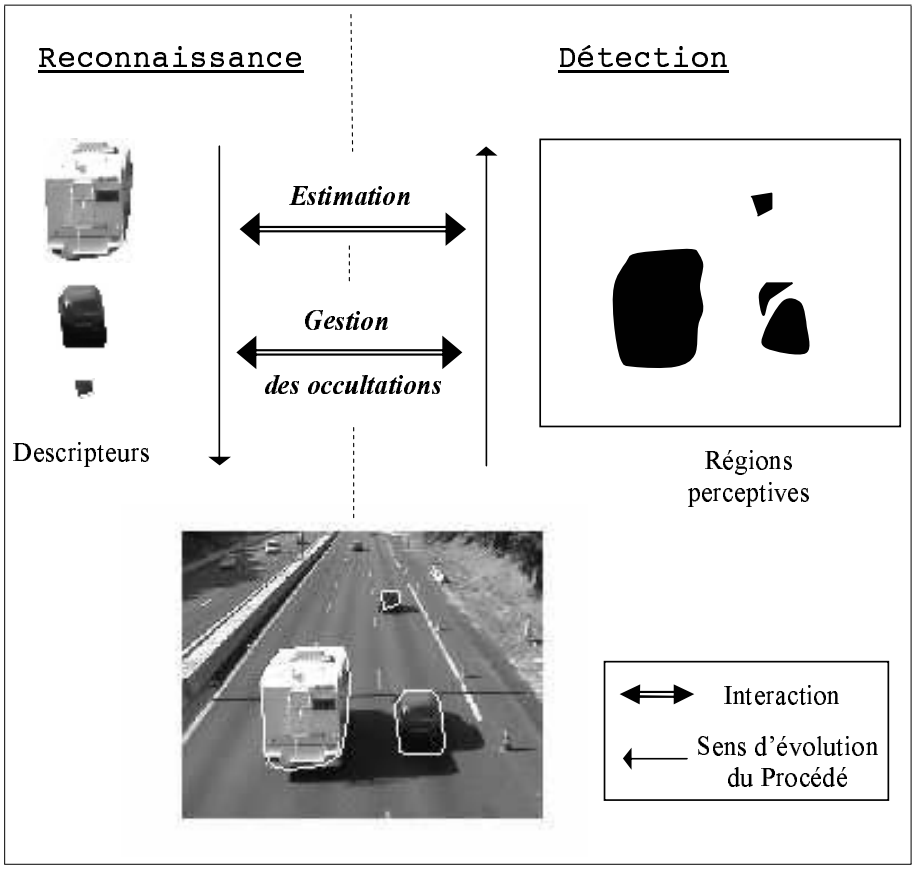


FIG. 1 – Bilan des formats et des flux d'informations constituant le procédé de suivi d'objets.

Structure du mémoire

Dans le ***chapitre 1***, nous présentons un état de l'art concernant l'analyse du mouvement en vision par ordinateur. La caractéristique de mouvement

est en effet essentielle quant à la problématique du suivi d'objets mobiles. Dans ce chapitre, nous rappelons que l'analyse du mouvement s'organise autour de trois grandes tâches : la *détection*, l'*estimation* et la *segmentation* ; l'objectif n'étant pas de faire une présentation de chaque méthode existante, mais plutôt d'extraire, à partir des grandes idées directrices une stratégie adaptée à notre problématique.

Suite à l'état de l'art que nous avons effectué, nous pouvons maintenant présenter l'architecture générique d'implantation du procédé de suivi. C'est l'objet du ***chapitre 2*** où nous spécifions la notion de descripteurs, d'entrées perceptives et les différents modules de traitement constituant l'architecture [IBB+03, IBM03a]. Nous développons une méthode de segmentation haut niveau en utilisant une modélisation stochastique des régions perceptives, le but étant de générer des cartes d'appartenances région-descripteur. Pour la prédiction, nous avons choisi d'utiliser un filtre de Kalman. Enfin, nous nous sommes intéressés à une sélection adaptée de l'espace d'état en privilégiant la simplicité.

Le mouvement étant par essence une entrée perceptive majeure, nous proposons dans le ***chapitre 3*** une nouvelle méthode de détection de mouvement bas niveau fondée sur une construction adaptative d'une image de référence. Pour contrôler la mise à jour de l'image de référence, nous développons une fonction d'adaptation fondée sur une mesure de la variationnelle sur la base du profil d'intensité pour chaque pixel pris le long de l'axe temporel. Cette méthode est robuste aux variations de la luminosité ambiante et propose en outre une gestion des ombres portées. [IBM02b, IBM02c].

Dans le ***chapitre 4***, nous illustrons la pertinence de notre conception du procédé par l'étude de deux applications. La première concerne la gestion du trafic routier [IBM02a, IBM03b]. Nous développons dans cette section l'approche par modèle stochastique dédiée à la gestion des occultations. Nous montrons en particulier la pertinence de la méthode par des résultats obtenus pour diverses séquences réelles. Nous étudions ensuite une seconde application consacrée au suivi de visages. Cette étude est l'occasion de montrer que les entrées perceptives peuvent être de natures diverses. En outre, nous adapterons les descripteurs pour une prise en compte des spécificités de forme du visage ce qui nous permet d'obtenir une plus grande fiabilité de notre schéma de prédiction.

Chapitre 1

État de l'Art en analyse de mouvement

1.1 Introduction

Le suivi d'objets mobiles est un processus indispensable à de nombreuses applications. La prise en compte de la mobilité passe par l'analyse du mouvement de chaque entité. En télésurveillance, il s'agit de distinguer les mouvements "autorisés" de ceux interdits. La détection d'un véhicule se déplaçant à contre sens sur une voie rapide est une situation critique qu'un système de contrôle doit pouvoir appréhender. L'étude quantitative et qualitative du mouvement revet donc un caractère particulier quant à l'extraction d'informations significatives pour de nombreuses applications.

Ce premier chapitre est l'occasion de faire le point sur les différentes approches algorithmiques dédiées à l'analyse du mouvement. Dans un premier temps, nous nous intéressons à la **détection**, l'**estimation** et à la **segmentation** du mouvement pour aborder ensuite la tâche de suivi. Ces différentes rubriques constituent la base des outils développés dans cette thèse. Ils nous permettent ensuite d'aller graduellement vers une caractérisation individuelle des objets. La détection de mouvement permet d'extraire deux catégories de sites : ceux associés à l'arrière plan (éléments de la scène représentant l'environnement des entités mobiles) et ceux correspondant aux objets (caractérisés souvent par un mouvement différent de celui de la caméra). La segmentation, associée à l'estimation, permet de retrouver les liaisons spatio-temporelles de chacune des apparitions des objets dans la séquence temporelle d'images.

1.2 Détection, Estimation et Segmentation du Mouvement

L'analyse au sens du mouvement est une tâche fondamentale de la perception visuelle. Différentes expériences ont été effectuées dès le début des années 70 [Joh73] pour étudier ce système perceptif. Ces travaux démontrent que le seul fait d'analyser le mouvement à travers de simples sources lumineuses fixées sur un acteur en chambre noire permettait l'identification et l'interprétation de scènes dynamiques.

De la même manière, en vision artificielle, la compréhension de certains phénomènes physiques nécessitent une analyse au sens du mouvement. Nous pouvons citer, par exemple, la mécanique de fluides, la météorologie ou la sismologie.

Suivant le contexte applicatif, les objectifs quant à l'analyse du mouvement varient. Une simple détection peut être suffisante ou peut initialiser une procédé plus complexe. Une mise en alarme par simple détection est utilisée en surveillance de site par exemple. L'estimation du mouvement nous permet de qualifier son amplitude et son orientation. Cette tâche nous donne l'opportunité de développer des procédés réagissant à des classes différentes de formes de mouvement suivant les situations (véhicules à contre sens, zone de turbulence). Plus ambitieuse dans les objectifs, la segmentation a pour but de segmenter le contenu de la scène observée sur un critère utilisant la paramétrisation du mouvement. Sa forme la plus simple est la détection mais elle peut exploiter de façon complète l'information de mouvement en utilisant son estimation pour affiner sa discrimination.

De nombreuses solutions algorithmiques existent dans la littérature pour répondre à ces différents objectifs d'analyse. Avant de rentrer dans les détails des approches, il est obligatoire de borner notre domaine d'étude pour structurer notre état de l'art. Pour cela, intéressons nous tout d'abord aux différents systèmes d'acquisition de l'image qui peuvent être utilisés.

Premièrement, concernant la source vidéo, nous pouvons opter pour deux types de systèmes d'acquisition :

1. **Le mode multi-vues**.- c'est le cas de la stéréovision, qui essaye de reproduire au mieux le système de la vision humaine. Ces systèmes permettent une analyse 3D de la scène. Leur mise en oeuvre reste complexe et ils sont encore peu utilisés en télésurveillance.
2. **Le mode monoculaire**.- il se caractérise par un seul point de vue (une seule caméra). C'est le système le plus utilisé et les données acquises sont uniquement des informations 2D puisque la formation des

images résulte de la projection de la scène 3D sur un plan 2D. Du point de vue de l'information de mouvement, on parlera de *mouvements apparents*, résultant d'un mouvement 3D.

Deuxièmement, le système d'acquisition peut utiliser une caméra fixe ou mobile.

- ***Mouvement et caméra fixe*** : Dans ce cas, la problématique revient à séparer les zones mobiles du fond statique. Pour de nombreuses solutions algorithmiques, l'analyse du mouvement est réduite à l'analyse des changements temporels [JMA79]. Cependant, on ne peut pas toujours utiliser cette hypothèse. Dans les régions uniformes de l'objet en mouvement, aucun changement n'est alors observé. Au contraire, des variations de l'intensité lumineuse peuvent avoir d'autres causes que le mouvement, comme le changement des conditions d'illumination par exemple.
- ***Mouvement et caméra mobile*** : Une estimation préalable du mouvement dominant apparent est nécessaire afin de distinguer les entités en mouvement des changements induits par le mouvement de la caméra. La problématique se ramène à celle de la segmentation au sens du mouvement. L'objectif est de rechercher les frontières qui séparent les objets ayant des mouvement différents. La difficulté est qu'un même objet peut être représenté par plusieurs régions possédant leur mouvement propre. C'est le cas d'une rupture de profondeur entre deux surfaces caractérisées par un même mouvement et ayant des projections voisines dans l'image [OB94]. La tâche est alors beaucoup plus complexe.

Sur le plan applicatif, les systèmes de télésurveillance utilisent pour une grande majorité une capture fondée sur une seule caméra fixe. La raison en est que l'utilisation d'une caméra fixe permet une prise en compte optimale de la dimension temporelle du problème. Nous pouvons en effet exploiter la cohérence temporelle de l'information pour une intégration à long-terme permettant de rendre plus robuste les techniques proposées. C'est dans ce contexte que nous placerons l'étude présentée dans ce mémoire.

Maintenant que nous avons défini notre cadre d'étude, nous allons rappeler les grandes familles de méthodes utilisées en analyse de mouvement.

1.2.1 Estimation du mouvement

L'estimation du mouvement a été très largement traitée dans la "littérature". De nombreuses méthodes fondées sur des approches derivatives (fondées sur les gradients) [Nag89, FT79], sur des formulations spec-

trales [AB85] ou sur des approches de type corrélatoire [Ana89] peuvent être utilisées. Sans rentrer dans les détails des différentes solutions, il est intéressant de connaître les hypothèses de travail exploitées par ces différents algorithmes. Deux principales hypothèses sont à la base de la plupart des méthodes proposées pour l'estimation du mouvement apparent : la première est fondée sur la conservation de l'intensité lumineuse du point en mouvement et la seconde valide une continuité spatiale du mouvement. La continuité temporelle est une troisième hypothèse peu utilisée par ce type d'algorithmes. L'objectif de ces hypothèses est la simplification de la mise en équation du problème d'estimation. Des contraintes additionnelles ont aussi été proposées pour relaxer les hypothèses initiales [HS81] comme une contrainte de lissage sur le champ de déplacement estimé.

Conservation de l'intensité

Notons $I^k(p)$ la valeur de l'intensité à l'instant k sur un pixel p de l'image. Pour estimer le mouvement apparent, on suppose que cette intensité $I^k(p)$ est une grandeur physique, reliée à un point P d'une surface tridimensionnelle visible, dont la projection dans l'image est le point p. On suppose par ailleurs que cette intensité reste constante sur la trajectoire de ce point P. On a alors [HS81] :

$$\frac{dI^k(p)}{dt} = 0 \tag{1.1}$$

Notons 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, KT94], ou 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, ...) [MWA87, MF90].

Cependant, l'équation de conservation ci dessus n'est pas vérifiée dans un nombre important de situations. Cette équation ne prend pas en compte les variations d'illumination qui peuvent survenir lors d'une séquence vidéo de scènes d'extérieur. Elle ne prend pas non plus en compte les phénomènes d'ombrage sur les objets par rapport à leur orientation. Le bruit d'acquisition est aussi une source de changement de l'intensité dans l'image, mais il peut se modéliser plus facilement.

Continuité spatiale

L'hypothèse de conservation de l'intensité n'est pas suffisante pour déterminer entièrement et de façon précise le mouvement apparent. La Fig. 1.1 illustre l'ambiguïté qui existe lorsque l'on veut calculer le déplacement d'un segment de contour (problème de *l'ouverture*). Pour résoudre ce problème, on considère que les points voisins appartiennent à la même surface 3D, et donc qu'ils vont suivre des déplacements apparents similaires. Cette hypothèse est ensuite utilisée sous une contrainte de lissage [HS81, Nag83]. Elle peut aussi être mise en ouvre à travers des modèles de mouvement paramétriques, qui peuvent prendre en compte - plus ou moins de façon locale - un champ constant, un champ affine ou un champ d'ordre supérieur.

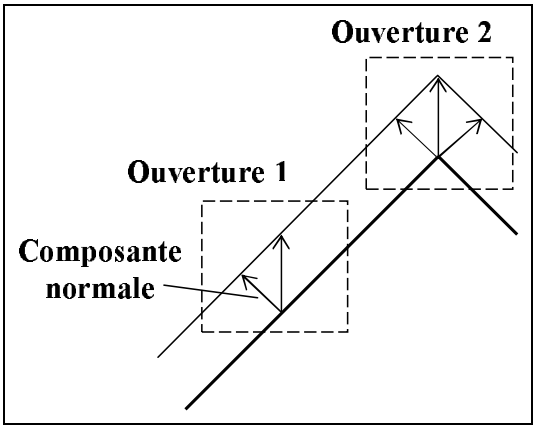


FIG. 1.1 – Seule la composante de déplacement normale au segment de droite est identifiable localement.

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, GS94]. Elle n'est généralement valide que sur des durées temporelles limitées. Il est nécessaire d'avoir une fréquence élevée d'échantillonnage et des faibles vibrations de la caméra (jitter). Par ailleurs, le mouvement propre des objets n'est pas toujours prévisible. Ainsi le mouvement des êtres vivants change fréquemment de direction. Enfin, les frontières de mouvement sont une fois de plus des régions critiques où l'hypothèse n'est pas toujours valide.

1.2.2 Détection de mouvement

La détection de mouvement permet l'extraction des régions de l'image associées aux objets mobiles dans la scène.

Dans le cas général, le processus de détection est formulé comme un problème de classification en termes de régions ou pixels. La phase de détection est fondée sur des techniques de seuillage [IRP92, IA98, TP90, TM93] ou des approches bayésiennes à travers une information contextuelle [OB97, XG94].

La principale limite des algorithmes de détection du mouvement réside dans la capacité du modèle du mouvement dominant de représenter le mouvement produit dans l'image par le déplacement de la caméra. Lorsque un objet mobile occupe une surface trop importante dans l'image (de l'ordre de 50% du support de l'image [Odo94]), le modèle de mouvement estimé ne correspond pas à celui de la caméra, ce qui constitue alors une violation de l'hypothèse sur laquelle s'appuie le principe de détection des objets mobiles. D'autre part, dans le cas où la représentation envisagée pour le mouvement dominant est un modèle 2D paramétrique (affine ou quadratique), cette représentation ne permet pas la modélisation complète de tous les mouvements effectués par la caméra. Ainsi, pour des scènes présentant de fortes variations de profondeur, le modèle de mouvement dominant estimé ne sera valide que pour une partie de la scène observée. Des travaux récents [CB99, IA98] s'intéressent à la différence entre les zones qui ne suivent pas le mouvement dominant.

1.2.3 Segmentation de mouvement

Le mouvement apparent dans une séquence d'images constitue une information visuelle essentielle pour analyser le contenu de l'image. Cet aspect est mis en évidence par de nombreuses expériences psychovisuelles. Ainsi, si on prend 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 pouvons grouper les points de même mouvement, ce qui nous permet de définir la forme de la région en mouvement [Bra74, Nak85].

L'objectif de la segmentation du mouvement est de déterminer les régions homogènes au sens du mouvement contenues dans une image. Le critère d'homogénéité normalement exploité est fondé sur la validité d'un modèle de mouvement en 2D paramétrique qu'il faut aussi estimer lors de la phase de segmentation.

Nous pouvons principalement distinguer deux catégories d'approches. La première est fondée sur l'estimation conjointe du mouvement. Cette tâche peut être formulée comme un problème d'estimation de mélanges de modèles de mouvement [AS96]. Elle peut aussi exploiter des méthodes statistiques d'étiquetage contextuel [ASB94, BF93, OB98]. Il est à noter que l'estimation robuste des modèles de mouvement paramétrique comme c'est le cas dans [OB98] permet de ne plus itérer des étapes alternées d'estimation des modèles de mouvement et de mise à jour de la segmentation associée (détermination des supports des régions) comme proposé dans [BF93]. Cette classe de techniques comprend également la méthode exploitant une estimation dense du champ de vitesses comme dans [BA93, MP98]. L'utilisation d'estimateurs robustes pour la minimisation d'une certaine fonction d'énergie permet la localisation des discontinuités du mouvement réalisant en même temps l'estimation dense et la segmentation du champ des vitesses.

La seconde catégorie de techniques [AET98, GB00, Pat98, WA94, WBPDB96, ZB95] consiste à introduire une segmentation (partition) spatiale initiale de l'image sur des critères d'intensité, de couleur ou de texture par exemple. Une étape de fusion est nécessaire pour regrouper en régions spatiales tous les pixels qui ont des caractéristiques similaires. Ces méthodes s'appuient généralement sur une mesure de similarité de mouvement entre régions, exprimée à partir de modèles 2D paramétriques. Le regroupement des régions spatiales est fondé sur des techniques d'agglomération dans l'espace des paramètres [AET98, WA94], des critères statistiques d'information comme le critère MDL (Minimum Descriptor Length) [Pat98, ZB95, GBR03] ou encore des méthodes d'étiquetage contextuel au niveau d'un graphe de régions [GB00].

Si on compare les méthodes de segmentation de mouvement avec les méthodes de détection de mouvement on peut conclure que les premières nous fournissent des informations plus riches. Le niveau de description de l'image est en effet supérieur dans le cas des méthodes de segmentation au sens de mouvement. Elles permettent ainsi de considérer des scènes contenant plusieurs objets en mouvement même quand ces derniers se recouvrent partiellement (occultation partielle). Cependant, ces méthodes ont un coût de calcul très supérieur par rapport aux méthodes de détection de mouvement. De plus, les objets mobiles non rigides seront généralement sursegmentés. Par contre, il est intéressant de concevoir une méthode exploitant une segmentation en deux temps. L'objectif est de proposer une stratégie de segmentation s'appuyant initialement sur une détection bimodale au sens du mouvement. Cette décomposition de la tâche d'analyse permet de réduire considérablement la complexité calculatoire de la méthode.

1.3 Techniques de détection de mouvement

Comme nous l'avons vu précédemment, la détection de mouvement comme procédé d'extraction primaire est tout à fait en accord avec notre objectif de création d'une entrée perceptive. Dans cette partie, nous allons détailler les solutions algorithmiques proposées dans la "littérature". Nous rappelons que nous nous plaçons dans le cas de l'utilisation d'une caméra fixe.

1.3.1 Différences d'images

La différence d'images est la plus ancienne et la plus simple de toutes les techniques de détection de mouvement. Le calcul de la différence d'images est fait soit en utilisant deux (ou plusieurs) images consécutives (appelée différence inter-images), soit entre l'image courante et une image de référence. L'utilisation d'une image de référence implique une stratégie de mise à jour, par exemple lorsque les conditions d'éclairement changent, ou lorsqu'il est impossible d'obtenir initialement une image de référence complète et qu'il faut la construire au fur et à mesure. Une variation d'intensité supérieure à un seuil θ, pour un pixel donné est considérée comme équivalente à un mouvement. Un seuillage permet d'obtenir une carte binaire d'étiquettes indiquant si le pixel est considéré comme appartenant à une région en mouvement, ou immobile et appartenant donc au fond.

Différence inter-images

Il est possible de réaliser une détection de mouvement en analysant les différences existantes entre deux images successives d'une séquence. Soit I^k la variable qui exprime l'intensité de luminosité à l'image k. On appelle observation O^k, l'image définissant la différence temporelle de deux images consécutives :

$$O^k = I^k - I^{k-1} \tag{1.2}$$

Le résultat O^k ainsi obtenu est nul en tout point où le signal d'entrée I reste constant, tandis que les points où un changement temporel est intervenu peuvent être validés par simple seuillage, [JMA79, YAT81, Jai81, WG87]. Ce seuillage reflète la prise en compte ou non, en tant qu'information suffisamment significative, de la valeur de la différence d'images en chaque pixel. Le seuillage est indispensable car la présence du bruit dans l'image engendre des observations non nulles dans presque la totalité de l'image. La

valeur du seuil est assez facilement ajustée pour les images d'intérieur, ce qui n'est pas le cas pour les scènes d'extérieur où son ajustement peut être délicat.

Cette méthode présente l'avantage d'être simple à mettre en oeuvre et ainsi adaptée à une mise en oeuvre temps réel. Son emploi direct est néanmoins très limité, notamment par son importante sensibilité au bruit, ce qui rend son utilisation problématique sur des scènes réelles acquises en extérieur : ce que l'on gagne sur le plan de l'élimination des détections parasites (en rehaussant la valeur de seuillage) se fait au détriment de la qualité des zones détectées et vice versa.

Il est alors possible d'augmenter la qualité par moyennage sur un bloc ou bien par seuillage par hystérésis. Dans le même but, la décision peut s'appuyer sur un test de Hinkley, plus robuste qu'une simple comparaison à un seuil [Pla85]. Dans certains cas, le seuil est défini à partir de données statistiques (test du χ^2) [AKM93], ou bien il est fixé en fonction de la variance du bruit présent dans les images [Die91].

Le problème majeur de cette procédure est relatif à la représentativité de l'information obtenue. L'image O^k fait apparaître quatre types de zones distinctes, pour le cas le plus fréquent où il y a recouvrement entre les projections successives du mobile entre *(k-1)* et k (voir fig. 1.2) :

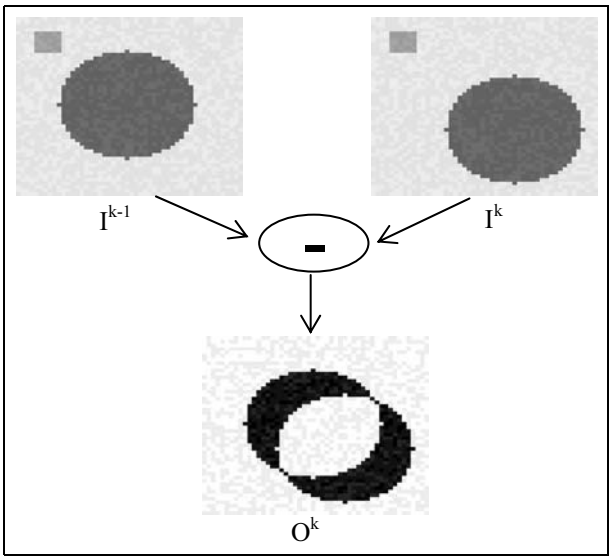


FIG. 1.2 – Zones de changements temporels induites par un objet en mouvement.

- *Zone fixe :* zones de niveau de gris faible correspondant au fond de la scène.
- *Zone de recouvrement :* zones où le fond a été recouvert par les objets en mouvement.
- *Zone de découvrement :* zones où le fond a été découvert par les objets en mouvement.
- *Zone de chevauchement :* zones correspondant aux recouvrement des objets.

Pour retrouver une bonne image de détection de mouvement, nous devons effectuer deux opérations : éliminer *"l'echo"*, c'est à dire, la zone découverte par l'objet et correspondant à sa position précédente, et reconstituer complètement le masque de l'objet en mouvement.

Plusieurs méthodes ont été proposées. Dans [YAT81] les auteurs proposent une identification des zones couvertes et découvertes à partir de la connaissance a priori des niveaux de gris relatifs au fond et aux objets. Le signe de la différence nous permet cette séparation. Cependant, la connaissance a priori des objets dans la scène n'est pas toujours possible.

Dans le cas où cette connaissance n'est pas disponible, l'utilisation de plus de deux images successives peut résoudre le problème. Ainsi dans [LG83] les auteurs proposent de considérer deux différences consécutives d'images. Grâce à une combinaison de ces deux différences, on peut tenter de retrouver les objets en mouvement à l'image k. Pour y parvenir, deux méthodes équivalentes sont considérées. La première nous amène au seuillage des images de différences. Comme résultat du seuillage, deux cartes binaires sont formées. L'application d'un "ET logique" nous permet de retrouver les objets en mouvement à l'instant k. La seconde méthode combine au préalable les images de différences à travers un opérateur de minimum (qui est équivalent au "ET logique"). Un seuillage a posteriori nous permet de retrouver un résultat identique.

La solution n'est pas toujours fiable, car si l'objet se recouvre lui-même, entre deux images consécutives, et que sa texture est uniforme, alors le masque final obtenu peut ne pas correspondre à la forme de l'objet. En plus, si deux objets sont très près l'un de l'autre, il peut arriver que le masque final contienne de "fausses détections", communément appelées *fantômes*, car leur détection ne correspond pas à la présence d'un objet.

Pour pallier ce problème, [WG87] proposent de remplacer le "ET logique" par un "ET conditionnel" : si une région de différence temporelle produit deux masques différents, seul le plus grand est validé.

Pour intégrer toutes les informations de différences et fiabiliser la vali-

dation des zones, Jain et al. ont introduit dans [JN79, Jai84] la méthode des différences inter-images cumulées. Dans la pratique, la différence est effectuée entre chaque image et la première de la séquence. Après seuillage, les points de changements temporels validés sont accumulés dans une image de détection finale.

Ce type d'approche permet de détecter et retrouver la position d'objets en mouvement dans les images, mais il y a de fortes limitations à ce principe. Si la première image contient des objets déjà en mouvement, les différences vont toujours tenir compte de cette initialisation erronée, ce qui pénalise la détection finale.

La première image de la séquence, à laquelle sont comparées toutes les suivantes, joue le rôle d'**image de référence**. Pour éviter les problèmes cités précédemment, l'idéal serait d'obtenir une image de référence en l'absence d'objets en mouvement. La section suivante nous offre une vision globale des méthodes d'extraction fondées sur une image de référence.

Détection à l'aide d'une image de référence

L'utilisation d'une image de référence représentant la scène exempte de tout élément mobile rend immédiate la localisation des objets mobiles à l'instant k [Wen83]. Lorsque l'on extrait une différence entre une image de référence du fond de la scène et l'image courante, l'image résultat ne présente que deux types de régions, contrairement au cas de la différence inter-images :

- Régions correspondant au fond de la scène.
- Régions où le fond a été recouvert par l'objet.

D'une manière générale, la distinction entre ces deux types de régions se fait de façon simple en supposant que la différence de niveau de gris des pixels correspondant aux variations d'éclairage de la scène par rapport à la référence reste limitée. La valeur du seuil, utilisée sur cette différence entre l'image courante et l'image de référence, est en général moins critique que dans le cas d'une différence inter-images. Si on arrive à obtenir une image de référence de bonne qualité (ce qui vient à dire que les conditions d'éclairement sont en accord avec la réalité) un seuil relativement faible va permettre l'extraction des zones des objets en mouvement, même s'ils sont peu texturés ou si le contraste entre le fond et l'objet est assez faible.

Cependant, la difficulté majeure de ce type de techniques reste dans l'obtention d'une image de référence de qualité. Autrement dit, la difficulté de la technique repose sur la fidélité de l'image de référence pour représenter

les conditions d'éclairage ambiant, notamment pour les scènes d'extérieur. Le lever ou le coucher du soleil, les passages nuageux, les ombres ou les reflets vont faire que l'image de référence perd sa pertinence s'il n'y a pas de remise à jour. Cette réactualisation nous amène à un problème de construction de l'image de référence.

La construction d'une image de référence est un problème abordé dans la "littérature" scientifique par de nombreux auteurs [Jai84, Dre78, Tan82, DHA88a, LT90]. La difficulté réside dans le fait qu'une image de référence est "perturbée" par le passage des objets en mouvement dans la scène. Cette extraction de l'image de référence peut être opérée *"manuellement"*, en sélectionnant une image de la séquence exempte d'objets mobiles [WD84, HW87], ou bien, obtenue en utilisant une mise à jour contrôlée par une détection de mouvement ou changements de contenue. Nous exposons dans la suite des techniques existantes :

Mise à jour locale en fonction du mouvement

Certain auteurs utilisent le résultat de la détection de mouvement pour construire une image de référence. La mise à jour de la référence se fait dans les régions non affectées par le mouvement des objets. Cependant, l'erreur entre l'image de référence et l'image courante augmente si la détection de mouvement devient bruitée, c'est à dire lorsque le détecteur indique la présence de mouvement dans des zones correspondant au fond. Le comportement de ce type de réactualisation est difficilement contrôlable, surtout lorsque la détection de mouvement est sensible aux variations d'éclairage.

Moyenne pondérée

La valeur à chaque pixel, p, de l'image de référence correspond au niveau de gris moyen des N dernières images de la séquence :

$$B^{k+1}(p) = \frac{1}{N} \sum_{j=k-N}^{k} I^j(p) \tag{1.3}$$

où B représente l'image de référence et I l'image acquise.

Cette utilisation d'une moyenne pour construire l'image de référence reste limitée car il faut stocker un ensemble de N images en mémoire. De manière à faciliter le calcul de l'image de référence on utilise dans la pratique une forme recursive. L'équation suivante est utilisée dans [Mec89, Mak96, Van97] pour calculer la valeur de l'image de référence en un pixel p :

$$B^{k+1} = \alpha \cdot B^k + (1 - \alpha) \cdot I^k \tag{1.4}$$

où α désigne un coefficient permettant d'adapter la rapidité de la réponse à une nouvelle valeur de la référence. Cette équation représente une somme, pondérée exponentiellement, des images précédentes et de la dernière image de référence, sur toute l'image [KB90] ou bien sur les régions statiques de l'image comme dans [WD84]. Le choix du paramètre α est un compromis entre l'effet de mémoire de la valeur de la référence et l'inclusion d'une nouvelle valeur suite à des changements de luminosité, par exemple. Ce compromis limite l'utilisation de la méthode, car lorsqu'un objet se déplace très lentement, les pixels du fond occultés momentanément par l'objet vont acquérir au fur et à mesure la valeur d'intensité de l'objet. De la même façon, lorsqu'un objet appartenant au fond de l'image commence à se déplacer, les pixels situés à la position originale de l'objet vont être déclarés comme étant en mouvement le temps de les inclure à nouveau dans la référence. Les zones déclarées improprement par erreur en mouvement sont appelées *"fantômes"*.

Analyse de l'histogramme des niveaux de gris

Pour chaque pixel dans l'image, on peut construire un histogramme représentant le nombre d'occurrences de chacun des niveaux de gris qui apparaissent au cours de la séquence [KSUS]. La valeur du niveau de gris correspondant au maximum de l'histogramme est considérée comme le niveau de gris le plus fréquent au pixel p de l'image de référence. Le nombre d'images prise en compte pour le calcul de l'histogramme doit être suffisamment grand pour que le maximum de l'histogramme soit significatif. Par contre, le coût calculatoire reste directement dépendant du nombre d'images.

Utilisation d'un filtre de Kalman

L'utilisation d'un filtre de Kalman peut permettre d'estimer l'intensité lumineuse de chaque pixel de l'image de référence à partir des valeurs des niveaux de gris du pixel aux instants précédents [KBG90]. L'état du filtre est représenté par les valeurs des niveaux de gris des pixels du fond.

1.3.2 Détection par test de vraisemblance

Les méthodes fondées sur le *maximum de vraisemblance* (Maximum Likelihood) sont des méthodes similaires à une différence d'image, mais plus robustes, car elles font intervenir le voisinage du pixel considéré dans la prise de décision. Ce test de vraisemblance avait été développé initialement pour

la segmentation d'images fixes [Yak76, Nag76] mais a été ensuite étendu pour la détection des changements temporels par Jain, Nagel et Hsu dans [JMN77, Nag78, HNR84].

Dans un voisinage local les auteurs considèrent que la distribution d'intensité résulte de l'addition d'un bruit blanc gaussien à un modèle a priori. Ce modèle peut être constant, linéaire ou quadratique. Soit A_1 la fenêtre de référence centrée sur le pixel p de l'image I^k en étude, et soit une seconde fenêtre A_2 à l'image I^{k+1}. Deux hypothèses sont comparées et mises en compétition :

- H_0 : les distributions d'intensités lumineuses possèdent le même paramétrage $\Rightarrow$ pas de variation temporelle.
- H_1 : les distributions suivent des paramétrages différents sur A_1 et A_2 $\Rightarrow$ ce qui se traduit en un changement temporel.

A chacune des hypothèses est associée une fonction de vraisemblance. La comparaison du rapport de ces fonctions à un seuil permet de decider en faveur d'une hypothèse ou de l'autre.

1.3.3 Algorithmes de relaxation

Pour améliorer la qualité des cartes de détection du mouvement, qui apparaissent souvent bruitées et délicates à exploiter, les études se sont orientées vers l'emploi d'approches incluant un traitement spatial pour rendre le résultat plus homogène. Dans ce cadre, on trouve les méthodes à base d'algorithmes de relaxation, comme les champs de Markov (Markov Random Fields, MRF) [DLC97].

Si leurs avantages majeurs sont la robustesse et la qualité des résultats de détection, leur principal inconvénient est leur lenteur d'exécution due au grand volume de calcul.

1.3.4 Mise en correspondance de blocs ou de points

Cette méthode consiste à comparer les intensités sur une fenêtre (ou bloc de référence) prise dans la première image avec une fenêtre de même taille dans la seconde image, mais translatée d'un vecteur *(u, v)*. Un voisinage de recherche est défini dans la seconde image. La fenêtre la plus semblable, selon un critère de similitude donné, est considérée comme correspondante de la fenêtre de référence. Cette méthode est souvent désignée par son nom anglais : *Block Matching*.

Un compromis doit être trouvé non seulement pour le choix de la fonction de similitude, mais aussi pour le choix de la taille du bloc et de la fenêtre de

recherche. L'utilisation d'une fonction de corrélation est plus précise, mais nécessite plus de calculs, la complexité des calculs étant proportionnelle à la taille du bloc et de la fenêtre de recherche.

1.4 Identification et suivi d'objets en mouvement

Après l'extraction des différentes régions en mouvement présentes dans une séquence d'images, la question qu'on peut se poser est la suivante : quelle est la correspondance de ces régions d'une image à l'autre. Chez l'homme la réponse à ce problème est automatiquement obtenue grâce à la spécificité de certains groupes de neurones permettant le regroupement perceptif (faculté de regrouper les points connexes qui ont les mêmes propriétés de vitesse ou de couleur par exemple). Pour un processeur, cette réponse n'est pas si facile à trouver.

Jusqu'à présent, nous avons introduit le terme **région** qui définit l'ensemble des pixels connexes détectés par l'analyse du mouvement. Si la détection de mouvement était parfaite, les régions peuvent correspondre directement aux objets réels. Cette configuration n'est malheureusement pas souvent vérifiée car de nombreux facteurs font que les régions different des objets. Les principaux problèmes que nous pouvons citer sont : le bruit de segmentation, la sur-segmentation de régions, la fusion de régions et les occultations. Tous ces problèmes font qu'une stratégie commune doit être définie lors de la conception des différents algorithmes de suivi d'objets. L'objectif de cette stratégie est bien sûr d'éviter la perte d'identification d'un objet. Cette idée d'identification d'objet est reliée à l'idée de **primitive** que nous avions ennoncée. Une *primitive* est un élément structurant caractéristique d'un objet. Les primitives suivies peuvent être par exemple des points d'intérêt de l'image (ou *feature points* en anglais), qui peuvent être des coins [CH96], points singuliers du champ de vitesse [MBD95], des segments [ZF92, DF90], des régions de l'image [MB94], ou des regroupements de ces primitives.

Le suivi d'une primitive d'un instant k vers un instant $k+1$ est généralement fondé sur une prédiction de la position de celle-ci recherchée à l'instant $k+1$, construite à partir de ses positions précédentes. Pour reproduire la nature dynamique de l'évolution de l'objet, un modèle peut être utilisé (comme par exemple un modèle AR). La prédiction de la primitive va permettre l'initialisation de la phase d'identification de la primitive à l'image suivante.

Nous décrivons brièvement ci-dessous quelques représentations courantes

et les méthodes de suivi qui leur sont associées :

- **Techniques fondées sur les contours.** Le contour est la frontière définie par un contraste d'intensité. Les contours peuvent être utilisés comme modèles représentatifs de l'objet à suivre. Les modèles employés sont essentiellement déformables, permettant leur adaptation au cours de l'évolution de l'image. Parmi ces modèles, les contours actifs ont été largement exploités [BI98]. Dans [KH95], un état de l'art est dédié aux modèles de contours déformables. L'auteur s'intéresse tout particulièrement à la décomposition modale de ses déformations, pour la segmentation et le suivi de structures 2D. Le terme d'énergie définissant le contour actif ou snake est choisi pour s'adapter au mieux au contour observé de l'objet. Cependant, la convergence correcte du contour actif vers cette frontière dépend fortement d'une bonne initialisation. Cette initialisation peut être obtenue par l'application d'une transformation spatiale du modèle précédent. La transformation utilisée dans [BBDM94] est un modèle affine appliqué aux gradients spatio-temporels de l'intensité sur la région intérieure du contour déformable. Une mise en correspondance entre les contours de deux images consécutives est utilisée dans [GMP96]. Dans [BD95], les auteurs ont établi un critère de corrélation avec un modèle paramétrique global. Un modèle de la dynamique du contour suivi peut être défini et fournit alors une initialisation satisfaisante. Des travaux importants ont été effectués par Blake et al. concernant l'apprentissage du modèle dynamique [RWBM96] et le suivi de contour dans des conditions difficiles (fond texturé, dynamique complexe) sur la base de techniques dites de *Condensation* [IB96].
  L'application des techniques à base de contours actifs classiques au suivi simultané de plusieurs régions est limitée à des objets non connexes (sans occultations entre les objets). Les contours actifs géodésiques [CK97], appliqués à la détection et au suivi de zones mobiles [PD98], ne nécessitent pas une initialisation proche de la solution, et sont adaptés pour suivre plusieurs régions simultanément, même si les régions suivies sont perturbées par des changements de forme ou de taille. Ces approches souffrent cependant d'une complexité calculatoire élevée.

- **Suivi de partition par carte d'étiquettes**. Dans [MM97], les auteurs proposent une technique de suivi d'une zone quelconque de l'image. La région délimitée est segmentée au sens de l'intensité par

une méthode morphologique. Chacune de ces régions spatiales est projetée dans le sens du mouvement estimé sur cette région. Les régions de cette carte prédite servent de *marqueurs* pour initialiser une nouvelle segmentation, c'est à dire dans la pratique une mise à jour rapide. Dans [MDK96], par contre, les cartes de segmentation successives sont construites indépendamment, et une phase postérieure d'association temporelle des régions est nécessaire. Dans [BF93, OB98] la détection de nouvelles régions est incluse dans la technique de segmentation.

- **Utilisation de maillages**. Ces techniques utilisent l'information de mouvement contenue à l'intérieur de la région suivie. Ces techniques permettent aussi des déformations locales de la région. Un partitionnement régulier ou adapté au contenu est effectué, et le mouvement d'ensemble est représenté par des mouvements paramétriques (affines, homographiques, ...) estimés localement au niveau des mailles. Ces techniques sont employées par exemple dans [AT97, TEST96].

1.5 Conclusion

Grâce à ce chapitre dédié à l'état de l'art en analyse du mouvement, nous pouvons introduire les lignes directrices du travail présenté dans ce mémoire. Tout d'abord, les outils algorithmiques proposés vont principalement s'attacher au traitement d'une série d'images acquises par une seule caméra statique dont le flux est analysé de façon continue. La raison en est que l'application principale que nous intéresse est le suivi de véhicules pour la surveillance du trafic routier. La conséquence directe de l'utilisation d'une source statique est de disposer d'une continuité naturelle dans la perception des entités mobiles. En comparant les approches de détection et de segmentation, il nous semble plus pertinent de nous doter d'une première phase de détection du fait de cette stabilité temporelle. Il est alors plus simple, suite à la réduction de l'information à traiter, d'enchaîner ensuite avec une phase plus discriminante. Comme nous l'avons énoncé dans l'introduction, nous fondons notre détection sur une génération d'entrée perceptive orientée région. Il semble intéressant d'utiliser dans la phase de reconnaissance une gestion de carte d'étiquettes ou d'appartenance fondée sur un critère de vraisemblance. Comme nous le verrons dans le chapitre suivant, nous développerons notre architecture autour d'un schéma ***détection-classification*** utilisant le mouvement comme source discriminante.

Chapitre 2

Suivi d'objets en vidéo surveillance

2.1 Introduction

Le suivi d'objets vidéo est d'une grande utilité pour de nombreuses applications en particulier pour la transmission numérique des images avec ou sans manipulation du contenu. La vidéo surveillance, la vidéo assistance et l'indexation en sont quelques déclinaisons. Sur le plan du contenu, chacune de ces applications propose des scenarii de scène de complexité très variable. Cette multiplicité des contextes fait qu'il n'existe pas de solution algorithmique unique au problème du suivi. Dans le cadre de la vidéo surveillance, notre étude va s'appuyer sur deux hypothèses fortes afin de proposer une méthode de suivi adaptée à ce secteur applicatif. Tout d'abord, nous considérons que la prise de vue est assurée par une caméra monoculaire statique ou très lentement mobile. Puis, nous supposons que le contenu mobile de la scène correspond à un flux régulier et organisé d'objets en mouvement. Ces deux hypothèses nous permettent de profiter d'une grande stabilité du contenu des scènes le long de l'axe temporel. Fort de ce constat, notre contribution consiste en la conception d'une architecture algorithmique en boucle fermée constituée d'une méthode de détection bas-niveau et d'une étape de mise en correspondance haut-niveau permettant de traiter individuellement chaque objet. La technique de mise en correspondance proposée est fondée sur l'exploitation d'un ensemble de descripteurs permettant d'obtenir une gestion individuelle de chaque objet. Cet ensemble peut regrouper des caractéristiques comme la forme, le mouvement et la couleur. L'intérêt de la méthode proposée est sa capacité de gérer les problèmes d'occultation par-

tielle ou totale des objets grâce à une boucle de rétroaction sur la base des descripteurs. En résumé, la méthode de suivi se décompose en deux étapes :

1. **Pré-segmentation de l'image**. Cette étape est dédiée à l'extraction des régions à suivre dans l'image. Ces dernières sont des groupements de pixels connexes associés à l'objet, à un sous-ensemble de l'objet ou à un regroupement d'objets. Cette étape ne réalise qu'une détection binaire orientée selon l'application. Elle ne permet donc pas d'individualiser chacun des objets. Le contenu à localiser peut être de nature très variée. Dans le chapitre suivant nous exploitons le mouvement comme *"la caractéristique"* discriminante du processus de pré-segmentation. D'autre types d'attributs peuvent être utilisés (couleur, gradients, etc). Le dernier chapitre présente un exemple de suivi de visage dont la pré-segmentation est fondée sur les propriétés de chrominance de la peau. Cette étape a cependant l'intérêt de réaliser une première extraction permettant de simplifier l'étape de mise en correspondance. Cette simplification a un impact direct sur le coût calculatoire global de la méthode et sur la probabilité d'avoir un appariement correct entre deux images consécutives.
2. **La mise en correspondance**. L'étape précédente permet d'extraire des régions associées aux objets. Il s'agit donc ici d'assurer une mise en correspondance temporelle permettant la poursuite. Dans ce but, la méthode d'appariement proposée s'appuie sur les régions détectées et sur les descripteurs. Une prédiction de ces derniers permet de fusionner ou de segmenter les régions, en extrayant ce que nous appelons les *zones*, qui sont des représentations surfaciques univoques des objets. La zone contient l'ensemble des régions (ou des sous-régions) associées à un même objet et va donc permettre la mise à jour des descripteurs. Plus riche sera le contenu des descripteurs, plus probable sera la mise en correspondance correcte. Cette hiérarchie de regroupement permet de mieux gérer les cas de sur-segmentation ou d'occultation.

Dans la "littérature" dédiée au suivi, deux groupes de méthodes se dégagent : celui fondé sur les méthodes exploitant une caractéristique géométrique et celui représentant les méthodes qui utilisent la notion de groupement spatial ou "région". Le terme d'*imagette* est souvent utilisé (en anglais *template*). Il faut noter que ce découpage caractéristiques/imagette n'exprime pour l'instant qu'un cloisonnement relatif à l'information utilisée pour représenter l'objet. La justification de l'existence de ces deux groupes ne s'arrête pas là car l'utilisation du mouvement n'est pas identique pour chacun d'eux. Pour toutes ces méthodes, la compensation du mouvement est

l'outil utilisé pour assurer la mise en correspondance. Cependant, le mouvement, pour le premier groupe, est vu que comme une quantité à estimer et pour le second, il est exploité comme outil de prédiction. Pour ce dernier, l'identification de l'objet est alors formulée sur la base d'un test de superposition spatiale. Pour que cette superposition soit valide, il faut que le modèle de mouvement soit prédictible. Une variation forte, estimée de façon incorrecte, débouche sur une perte de suivi. Par exemple, les images de la Fig. 2.1 montrent clairement l'importance du modèle de mouvement. Dans le cas où l'objet se déplace le long de l'axe de visée de la caméra, les images montrant deux projections consécutives présentent un effet marqué de contraction spatiale. La région représentant l'objet dans l'image subit donc de fortes modifications de contenu. Pour assurer une mise en correspondance effective, il est donc nécessaire d'assurer une bonne mise à jour du modèle de mouvement.

(a) Objet au debut de la scène. (b) Objet quelques instants après.

FIG. 2.1 – Modification de la taille de l'objet due à un effet de zoom engendré par une translation 3D.

Pour le premier groupe, dans [Kol], les auteurs exploitent les coins et les points de contour et dans [AMYT00] ce sont des points caractéristiques issus des zones texturées qui sont utilisés. Les contours actifs dans [FET00] sont aussi étudiés. Pour le deuxième groupe, dans [LFP98], les auteurs prennent comme imagette la première apparition de l'objet. Elle est par la suite recherchée dans les images suivantes sur la base de la corrélation. Pour rendre plus robuste cette approche, il est nécessaire de mettre à jour cette référence. Dans [Mag02], le fond de l'image et l'ensemble des objets en mouvement sont séparés. Le suivi est alors réalisé par prédiction grâce à l'utilisation d'un filtre de Kalman directement appliqué sur les positions associées aux objets. Une mesure de distance combinant la taille et la position permet l'assignation

finale de chaque pixel à un modèle. Une prédiction similaire est établie dans [WADP97] mais les auteurs proposent d'utiliser un paramètre de vraisemblance comme base de décision pour la mise en correspondance et cela pour chaque pixel.

Dans ce chapitre, nous présentons une méthode exploitant les points positifs des différentes techniques des deux groupes précités. L'objet est qualifié non plus à partir de sa seule distribution d'amplitude associée aux canaux couleurs mais par un ensemble plus discriminant utilisant des caractéristiques géométriques, de couleur et de mouvement. Le mouvement est une des caractéristiques utilisées pour la reconnaissance. Grâce à la prédiction, l'ensemble des caractéristiques est extrapolé puis mis à jour pour tester la vraisemblance de l'appariement. D'un coté, nous exploitons la simplification apportée par l'utilisation de certaines caractéristiques géométriques comme le contour et le barycentre et de l'autre, nous étendons la notion d'imagette à celle de région. Les caractéristiques géométriques permettent une première identification de l'objet par test de recouvrement. La région alimente un test statistique d'appartenance. En résumé, la méthode, décomposée en deux phases, utilise :

1. Dans un premier temps, une prédiction des descripteurs pour chacun des objets est effectuée pour les comparer à ceux des régions issues de la phase de pré-segmentation de l'image. La prédiction est obtenue grâce à l'utilisation d'un filtre de Kalman par exemple, oeuvrant sur l'espace d'état des paramètres du modèle de mouvement.

2. Dans un second temps, la méthode met en compétition tous les couples régions-descripteurs. L'objectif est de gérer les cas d'occultation. Cette compétition est développée sur la base d'une approche de type espérance-maximisation plus connue sous le sigle EM pour *"Expectation-Maximization"* [SG99]. Un formalisme de type Maximum a posteriori (MAP) fondé sur une fonctionnelle à minimiser permet une décision multi-classes pour gérer l'appartenance de chaque pixel.

Dans cette même ligne d'approche *hybride* mélangeant les caractéristiques des deux techniques nous pouvons citer les travaux de Cavallaro [CE04, Cav02] exploitant aussi les idées de prédiction d'objet et des régions détectées. Dans ces travaux, les régions ne sont plus des ensembles connexes résultants de la phase de détection d'objets mais des ensembles détectés et classifiés selon une caractéristique décrivant la région (couleur, texture, etc.). Les objets sont identifiés selon une approche région, à partir des mesures de distances fondées sur l'espace des caractéristiques.

Ce chapitre est organisé comme suit : la section 2.2 présente le schéma complet du processus de suivi d'objets en mouvement. La section 2.3 nous montre les descripteurs que nous avons sélectionnés pour alimenter l'étape d'identification. La section 2.4 est dédiée à l'estimation de mouvement. Deux approches y sont présentées, l'une exploite les équations de contrainte du mouvement, l'autre est fondée sur la mise en correspondance de points caractéristiques. La prédiction du modèle de mouvement permettant de se projeter dans l'image suivante est effectuée dans la section 2.4.3 via un filtre de Kalman. La section 2.5 décrit le processus de mise en correspondance qui va établir l'association régions-descripteurs possédant un lien spatio-temporel. Ces couples vont entrer en compétition dans le processus EM présenté dans la section 2.5.2 afin de fournir une segmentation finale pour chaque objet. La section 2.6 est dédiée au processus de mise à jour des descripteurs ainsi qu'au contrôle du nombre d'objets.

2.2 Procédé de suivi proposé

Comme nous l'avons énoncé dans l'introduction, le processus de suivi enchaîne les étapes de pré-segmentation de l'image et de mise en correspondance. La pré-segmentation extrait un masque binaire, noté M^k, représentant les pixels regroupés à l'aide d'un critère de segmentation (mouvement, chrominance, etc). Soit $R^k = \left\{R_i^k\right\} i = 1, \ldots, n$, l'ensemble des n régions de M^k pour l'image k. Comme l'indique la Fig. 2.2, ces régions se présentent à l'entrée de la chaîne de traitement dédiée à la poursuite des objets. Compte tenue de cette première segmentation, une recherche de correspondance des m objets sur la base de leur observation à l'image précédente peut être effectuée. Les objets sont représentés individuellement par un ensemble de plusieurs descripteurs. L'ensemble est noté DO_j^{k-1} avec j l'indice dédié à l'objet. Soit $DO^{k-1} = \left\{DO_j^{k-1}\right\} j = 1, \ldots, m$ les m ensembles de descripteurs mis à jour à l'image précédente. Parmi ces descripteurs, on peut trouver celui qui permet la modélisation spatiale de l'objet : ***la zone***. L'objectif du processus de suivi est de retrouver les m zones (une zone par objet) associées à chacun des objets. Soit $Z^k = \left\{Z_j^k\right\} j = 1, \ldots, m$, cet ensemble. Une première étape de mise en correspondance (section 2.5) met en liaison les régions et les objets ayant une cohérence spatio-temporelle. Cette cohérence est évaluée numériquement en utilisant les descripteurs. Spatialement, une première évaluation est obtenue à partir d'une prédiction de la zone notée $\tilde{Z}^k$, appelée ***Zone prédite***. Soit alors

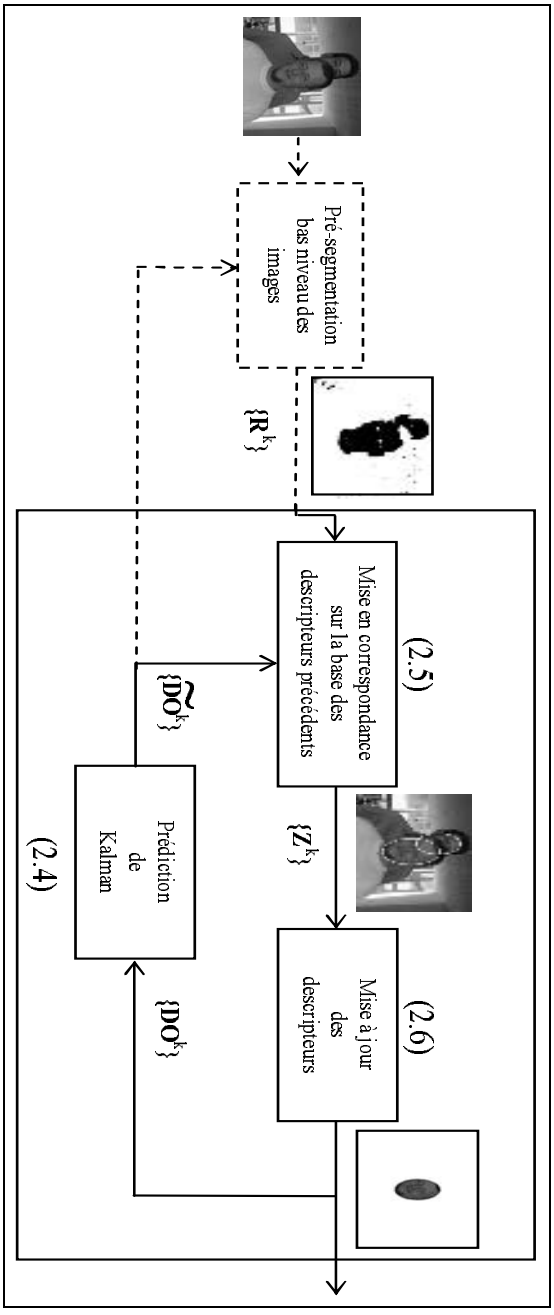


FIG. 2.2 – Schema complet du processus de suivi.

$$\tilde{Z}_j^k(p) = Z_j^{k-1}(p - \tilde{\Theta}_j^k) \tag{2.1}$$

la zone prédite, où p définit la position du pixel dans l'image et $\tilde{\Theta}_j^k$ définit la prédiction du modèle de mouvement noté Θ_j^k. La Zone prédite est utilisée pour prédire la zone spatiale dans l'image où l'objet est censé se positionner. Chaque couple région-descripteur mis en correspondance est construit sur la base d'une procédure de type Expectation-Maximization (EM) (section 2.5.2). Cette dernière procédure est une étape primordiale car c'est elle qui gère les cas ambigus tels que les cas de sur-segmentation ou d'occultation. Ce processus débouche sur une segmentation finale définissant une seule zone Z_j^k pour chaque objet. La Fig. 2.3 nous montre la différence entre une région R_i, un ensemble de descripteurs DO_j et une zone Z_j. La Fig. 2.4 nous montre la génération de la zone prédite à partir de la zone : la projection spatio-temporelle est effectuée selon le modèle du mouvement. Une fois que tous les objets ont été repérés nous passons à l'étape de mise à jour de tous les descripteurs. La prédiction est le point clef de la méthode de mise en correspondance qui permet de fermer la boucle du processus de suivi. La convergence, la rapidité et la qualité du suivi dépendent de la prédiction.

Avant de détailler une par une les étapes du processus de suivi, nous allons tout d'abord présenter les caractéristiques pouvant être sélectionnées pour constituer l'ensemble "descripteur". Il est à noter que le contenu de cet ensemble doit être adapté à l'application que l'on traite.

2.3 Sur la base de descripteurs

Chacun des objets est caractérisé par un ensemble de descripteurs regroupant des attributs spatiaux, un modèle de mouvement et une étiquette qui définit l'activité de l'objet. Cet ensemble constitue *"le modèle"* qui identifie l'objet le long de la séquence. Il permet de lever les ambiguïtés inévitables dans les cas d'occultation ou de sur-segmentation.

Comme la définition de ces descripteurs dépend de l'application envisagée, nous présentons dans la suite de ce paragraphe un ensemble de descripteurs génériques, nécessaires pour la bonne marche du processus de suivi.

Le premier d'entre eux, la *zone*, caractérise l'objet en termes spatiaux. La zone spécifie le contour, la taille et le centre de gravité ainsi que toute l'information attachée aux pixels contenus dans l'objet. La zone est la caractéristique principale du processus de mise en correspondance (section 2.5). Elle permet d'établir la relation spatiale entre la/les régions et les descripteurs. Le centre de gravité, noté (x_g, y_g), sert de point de référence

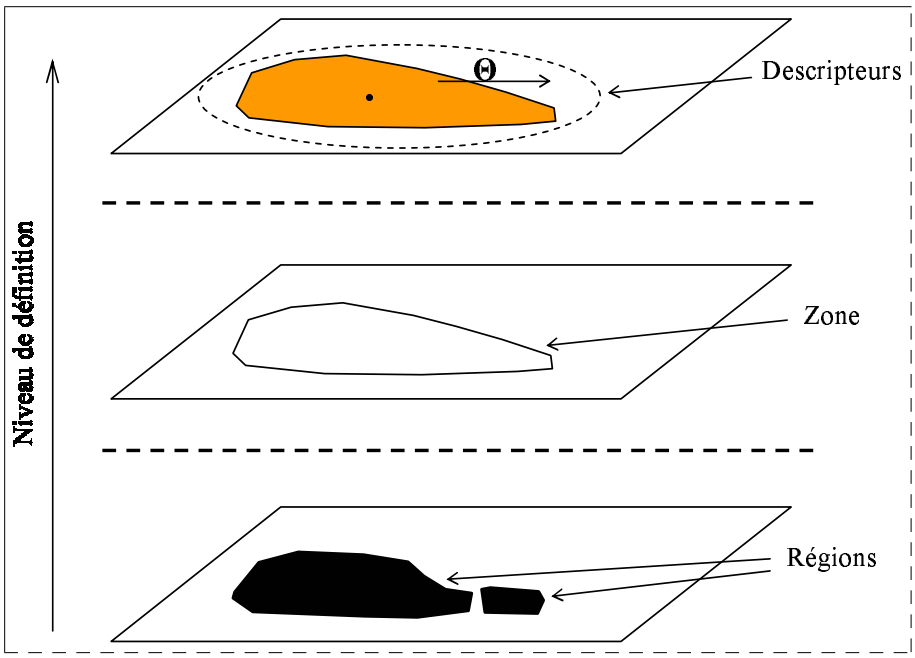


FIG. 2.3 – Différence entre une région, un ensemble de descripteurs (modèle) et la zone qui nous permet d'identifier l'objet.

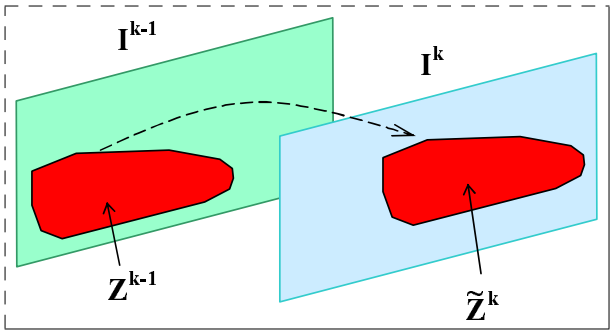


FIG. 2.4 – Zone prédite à partir de la projection spatio-temporelle de la zone selon le modèle de mouvement.

pour l'algorithme d'estimation de mouvement (section 2.4). Il nous donne la trajectoire de déplacement de l'objet dans l'image.

Nous utiliserons ici un modèle affine à six paramètres [OB98], défini comme suit :

$$\begin{cases} d_x = a_1 + a_2(x - x_g) + a_3(y - y_g) \\ d_y = a_4 + a_5(x - x_g) + a_6(y - y_g) \end{cases} \tag{2.2}$$

Le mouvement nous permet d'établir la liaison spatio-temporelle entre les régions R^k et l'ensemble des descripteurs calculés à l'image précédente, DO^{k-1}. Ce lien temporel est choisi à partir de la prédiction des attributs spatiaux de chacun des objets. Comme le présente la Fig. 2.3 et comme nous l'avons indiqué dans la section précédente, nous appelons *Zone prédite* la prédiction des attributs spatiaux des objets. Soit $\Theta = [a_1\ a_2\ a_3\ a_4\ a_5\ a_6]^T$ le vecteur contenant tous les paramètres du modèle de mouvement.

L'étiquette a pour but de décrire la situation de l'objet dans la séquence vidéo. Trois états sont définis. Ces états permettent de marquer un objet comme *entrant*, *sortant* ou *vivant* dans la scène :

1. *Objet entrant* : c'est l'état initial de tous les objets encore connectés au bord de l'image qui n'ont pas été vivants.
2. *Objet vivant* : c'est l'état des objets pleinement définis dans la scène. C'est un état intermédiaire entre entrant et sortant.
3. *Objet sortant* : ce sont des objets qui rentrent dans une zone de bord de l'image après être passés par l'état vivant.

La Fig. 2.5 nous montre le cadre définissant la limite pour qu'un objet prenne les états entrant, vivant ou sortant.

FIG. 2.5 – Cadre symbolisant sur l'image la limite pour classifier les objets comme entrant, vivant ou sortant de la scène.

Une seconde étiquette va spécifier le niveau d'interaction inter-objets. Nous avons trois cas à traiter :

1. Cas normal : c'est le cas de l'objet isolé.

2. Cas de sur-segmentation : l'objet est divisé en plusieurs régions lors de la phase de pré-segmentation.
3. Cas d'occultation : plusieurs objets sont superposés spatialement.

2.4 Estimation du modèle de mouvement

La mise en correspondance dépend de l'estimation du modèle de mouvement associé à l'objet. La reconnaissance de l'objet est assurée si la mise à jour du modèle de mouvement est pertinente. Afin de proposer un schéma efficace d'estimation du mouvement, nous avons étudié deux méthodes d'estimation du mouvement. La première est une méthode multi-résolution et incrémentale fondée sur l'équation classique de contrainte. La seconde est une méthode d'estimation du mouvement exploitant une approche d'appariement de points caractéristiques.

2.4.1 Méthode multi-résolution et incrémentale

La première approche que nous avons testée, est fondée sur le travail d'Odobez et Bouthemy [OB95]. Cette méthode propose l'estimation de mouvement fondée sur une technique de type Moindres Carrés (LMS en anglais) emploient l'équation de contrainte de mouvement [HS81] formulée sur la base d'un modèle affine à six paramètres (2.2). Les auteurs proposent une approche multi-résolution et incrémentale pour l'estimation de grands déplacements. Nous avons repris leur schéma en introduisant une approche incrémentale dans le temps permettant une réduction du temps de calcul tout en conservant la pertinence de la solution.

L'annexe A présente un résumé de la méthode complète développée par Odobez et Bouthemy. Nous détaillons dans la suite la modification introduite dans leur schéma, appelée processus incrémental dans le temps.

L'approche incrémentale et multi-résolution présentée par Odobez et Bouthemy nécessite une inversion matricielle à chaque étape. Cette approche ne peut pas être utilisée à chaque mise à jour du modèle de mouvement de l'objet car elle est trop coûteuse du point de vue calculatoire. Pour mettre à jour le modèle nous avons décidé d'implanter un algorithme s'adaptant aux variations du modèle au cours d'évolution de l'objet dans la scène. Ce processus est appelé un *"processus incrémental dans le temps"* puisqu'il intègre la même formulation que le processus incremental. Cette fois ci, à chaque instant k nous déclenchons une nouvelle itération du processus incrémental à partir du modèle prédit.

A chaque instant k, la variation de l'estimation de mouvement $\Delta\hat{d}_k$ est estimée avec la nouvelle image I^k conformément à la dernière estimation, $\hat{d}_{k-1}$. Nous avons :

$$\hat{d}_k = \hat{d}_{k-1} + \Delta\hat{d}_k \tag{2.3}$$

Cette méthode permet une adaptation aux changements de modèle tout en réduisant le coût calculatoire par rapport à la méthode multi-résolution et incrémentale originale.

2.4.2 Estimation du mouvement par appariement de points caractéristiques

De nombreux problèmes peuvent être à l'origine d'une mauvaise estimation sur l'équation de contrainte de mouvement. La sensibilité de la méthode dans les zones caractérisées par des gradients nuls, au d'ouverture, ainsi qu'une sensibilité au bruit due à l'utilisation des gradients sont autant de points délicats commentés dans la littérature [BFBB]. La Fig. 2.6 nous montre ces problèmes de divergence.

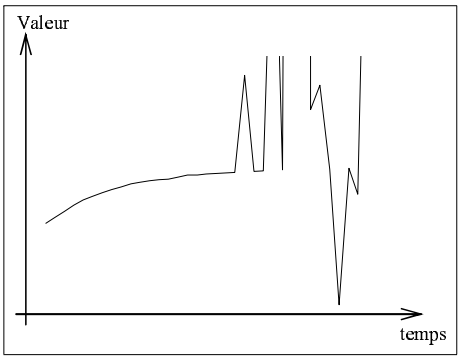


FIG. 2.6 – Divergence d'un paramètre du modèle de mouvement lors de l'éloignement de l'objet.

Afin de proposer une solution alternative, une nouvelle méthode d'estimation du modèle de mouvement est proposée. La méthode doit permettre une estimation plus robuste aux changements de taille de l'objet tout en diminuant le coût calculatoire de la méthode.

En se référant à la "littérature" dédiée à l'estimation du mouvement, une alternative à l'équation de contrainte du mouvement est possible en considérant le principe suivant.

La perception du mouvement d'un objet est d'autant plus performante que notre système visuel peut mettre en correspondance des zones texturées. Plus ces zones texturées persistent au cours du temps et plus l'interprétation du mouvement en terme de modèle est stable. Afin d'exploiter ces propriétés liées à la perception dans une approche algorithmique, certains auteurs comme Kanade [TK91] localisent les zones texturées à l'aide d'une analyse locale des gradients centrés sur un point. Ce point, appelé point caractéristique ou "feature point" (en anglais) est l'élément que l'on va chercher à propager dans les images suivantes afin d'estimer le mouvement.

Deux phases sont nécessaires pour la mise à jour d'une méthode d'estimation de mouvement utilisant les points caractéristiques : l'extraction et l'appariement de ces points. Ces deux phases dépendent de l'application. Ainsi, le chapitre 4 nous propose deux techniques de mise à jour du modèle de mouvement pour deux applications différentes : le suivi des véhicules pour la gestion du trafic routier et le suivi de visages.

Quelle que soit la technique utilisée de mise à jour du modèle de mouvement et quel que soit le modèle de mouvement utilisé, le processus de mise en correspondances des régions et des zones s'appuie sur une prédiction de la région spatiale occupée par l'objet. Dans la section suivante, nous développons la phase de prédiction à l'aide d'un filtre de Kalman dédié au modèle de mouvement.

2.4.3 Prédiction du modèle de mouvement

La mise en correspondance s'appuie sur une comparaison entre la prédiction de la position de l'objet (zone prédite) et celle observée (région). Cette prédiction est établie en prolongeant l'estimation du modèle de mouvement. Si la cadence d'acquisition est suffisamment élevée, nous observons que le mouvement de l'objet définit une trajectoire dans la scène. Cette trajectoire nous amène à considérer une variation du modèle de mouvement au cours du temps. Dans de nombreuses applications, cette trajectoire est assez "lisse" pour lui conférer une hypothèse de prédictibilité. Il est alors intéressant de proposer une modélisation régissant la loi d'évolution des paramètres. A titre d'exemple, la Fig. 2.7 nous montre la variation du paramètre de translation selon l'axe y du centre de gravité d'un véhicule se déplaçant. La convergence de la courbe correspond à l'éloignement de l'objet de la caméra interprété par la projection du mouvement 3D. Nous exploiterons l'espace d'observation dans les chapitres dédiés aux applications car il est préférable de s'adapter au cas étudié afin d'assurer une meilleure

prédiction. Par contre, pour toutes les applications, la loi d'évolution va s'appuyer sur la modélisation auto-régressive. En outre, la méthode de prédiction utilisée doit avoir la capacité de s'adapter à des ruptures de trajectoire. Dans ce contexte, le filtre de Kalman s'impose comme la solution adéquate.

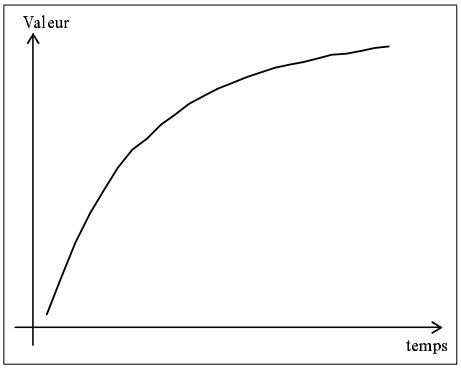


FIG. 2.7 – Variation du paramètre de translation du centre de gravité d'un objet en mouvement.

Les équations générales du filtre sont :

$$\begin{aligned} \bar{x}(k+1) &= \Phi(k+1,k) \cdot \bar{x}(k) + G(k) \cdot \bar{u}(k) \\ \bar{y}(k) &= H(k) \cdot \bar{x}(k) + \bar{v}(k) \end{aligned} \tag{2.4}$$

où $\bar{x}(k)$ représente le vecteur d'état du filtre, Φ la matrice de transition, G la matrice de commande, $\bar{u}(k)$ le vecteur du processus générateur, $\bar{y}(k)$ le vecteur de mesure, H la matrice d'observation et finalement $\bar{v}(k)$ le bruit de mesure [1]. Le processus générateur et le bruit de mesure sont considérés comme étant deux bruits blancs, de moyenne nulle et indépendants :

$$\begin{aligned} E\left\{ \begin{array}{cc} \bar{u}(k) & \bar{u}^T(l) \end{array} \right\} &= Q(k) \cdot \delta(k,l) \\ E\left\{ \begin{array}{cc} \bar{v}(k) & \bar{v}^T(l) \end{array} \right\} &= R(k) \cdot \delta(k,l) \end{aligned} \tag{2.5}$$

où $Q(k)$ et $R(k)$ sont les matrices de covariance du processus générateur et du bruit de mesure, et $\delta(k,l)$ représente le symbole de Kronecker.

Comme nous le verrons dans le chapitre dédié aux applications, la prédiction est réalisée sur la base des paramètres de mouvement (éq. 2.2) ou des paramètres d'une ellipse pour l'application sur les visages. Le but du filtre de Kalman est de prédire chacun des paramètres du modèle de mouvement. Soit b_i les paramètres considérés. La prédiction de ces paramètres est

[1] Les variables barrées sont des vecteurs et les variables en majuscules des matrices.

construite à partir d'un modèle auto-régressif (AR) caractérisant l'évolution des paramètres b_i du modèle de mouvement. Sa mise en équation donne :

$$b_i(k+1) = \sum_{j=0}^{p-1} \alpha_{ij} \cdot b_i(k-j) + v_i(k) \tag{2.6}$$

où p représente l'ordre du modèle et α_{ij} sont les coefficients du modèle AR de chaque composante b_i. A titre d'exemple, la Fig. 2.7 nous montre que les variations du paramètre à modéliser sont très corrélées, ce qui permet d'adapter le modèle AR. On constate grâce à la Fig. 2.8 que l'information spectrale est concentrée dans les basses fréquences, ce qui caractérise des variations lentes des paramètres b_i au cours du temps. Ces variations très corrélées nous amènent à un choix d'un ordre du modèle p réduit.

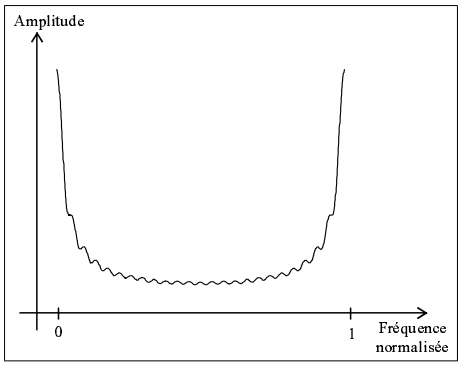


FIG. 2.8 – Analyse spectrale de la courbe des variations du paramètre de mouvement a_4 du modèle affine extrait sur la trajectoire d'un véhicule. Le choix de l'ordre p est en accord avec le contenu fréquentiel de la courbe.

Pour compléter la présentation du filtre de Kalman, il nous reste à identifier tous les termes de l'équation 2.4. Si on compare ces équations avec (2.6), nous constatons que dans le formalisme de Kalman le vecteur de mesures $\hat{y}(k)$ est représenté par les paramètres $b_i(k+1)$. Les coefficients α_{ij} forment le vecteur d'état $\hat{x}(k)$ et le vecteur des derniers paramètres estimés $b_i(k-j)$ constitue la matrice d'observation H. Pour la matrice de transition Φ, nous utilisons la matrice identité en admettant que la correction de l'état précédant alimente le nouveau vecteur d'état à l'itération suivante. Le processus générateur et le bruit de mesure sont à estimer dans l'application. Dans le chapitre 4 nous développerons deux exemples illustrant la mise en oeuvre de l'utilisation du filtre de Kalman.

2.5 Processus de mise en correspondance

2.5.1 Appariement des régions et des zones prédites

Disposant d'une prédiction des paramètres du modèle de mouvement pour chacun des objets, nous pouvons nous intéresser à la mise en correspondance.

La mise en correspondance établit les relations spatio-temporelles entre les régions issues de la détection précédente et les zones dans le but d'une mise à jour de ces mêmes zones et donc des descripteurs.

Pour mener à bien cette tâche, nous disposons des éléments suivants :

1. L'ensemble $R^k = \left\{R_i^k\right\} i = 1, \ldots, n$
2. L'ensemble $\tilde{DO}^k = \left\{\tilde{DO}_j^k\right\} j = 1, \ldots, m$ regroupant les sous-ensembles de descripteurs prédits dont la zone $\tilde{Z}_j^k$

Si on considère qu'aucun objet ne peut disparaître ou apparaître dans chaque nouvelle image [2], nous devons retrouver les m objets observés précédemment parmi les n régions en mouvement détectées dans la nouvelle image k. Pour retrouver les m objets, une redistribution des régions doit être réalisée. Une certitude est que nous ne pouvons pas nous appuyer sur les quantités m et n pour développer une stratégie de structuration de l'algorithme de suivi. Différents états "dégénérés" peuvent en effet apparaître rendant non significative la relation entre ces deux valeurs. Ces états "dégénérés" sont les suivants :

- Présence de régions dues au bruit : certains effets spéculaires, le mouvement de certains éléments non rigides de la scène (arbres, etc) peuvent être à l'origine de la génération de régions connexes parasites.
- Sur-segmentation des objets : la pré-segmentation proposée par la méthode de détection peut engendrer plusieurs régions pour un même objet.
- Occultation des objets : suivant la position de la caméra, les objets peuvent être partiellement ou totalement occultés. Une seule région représente alors plus d'un objet.

Ces configurations nous montrent donc que les régions obtenues lors de la phase de pré-segmentation ne peuvent pas à elles seules assurer le suivi. Ces régions ne traduisent pas directement la notion d'objet et il n'y a pas une relation univoque région-objet. Ce constat justifie la structure intermédiaire que nous avons créée : la zone.

[2] Voir section 2.7 pour établir les conditions d'initialisation et de disparition d'un objet.

Face à ces différents cas ambigus, plusieurs niveaux de traitement peuvent être envisagés permettant de solutionner la mise à jour des zones. Le premier niveau se caractérise par une élimination des régions inférieures à un seuil afin d'éliminer certaines régions parasites. Le niveau suivant va consister à mettre en correspondance les zones prédites et les régions observées afin de "scanner" l'ensemble des candidats sous l'hypothèse qu'il ne peut y avoir d'apparition ou de disparition d'objets dans la scène. Un objet doit suivre sa logique d'existence à savoir commuter séquentiellement entre les trois états définis dans le descripteur d'objet (section 2.3) : entrant - existant - sortant. Ce niveau a l'intérêt de dresser la liste exhaustive des régions et des zones que nous devons re-manipuler car correspondant à un cas de sur-segmentation ou d'occultation. Riche de ces informations, il reste le dernier niveau, le plus coûteux en calcul, fondé sur une extraction optimale d'un mélange de distributions modélisant l'erreur d'adéquation entre l'observation et les descripteurs prédits. C'est cependant lui qui permet de lever les dernières ambiguïtés en utilisant une approche de type EM.

La reconstruction des zones passe par une classification des pixels des régions. Pour alimenter cette étape, nous devons créer un certain nombre de partitions des différents ensembles à notre disposition. Le point de départ du processus de génération est la formation exhaustive des couples actifs [Région—Zone]. Chaque couple est défini comme suit :

Définition : *Couple actif.*

$$C_{ij}^k = \left[R_i^k | \tilde{Z}_j^k\right] \iff R_i^k \Cap \tilde{Z}_j^k \neq \varnothing$$

où $\Cap$ représente l'opérateur d'intersection spatiale.

A la fin de cette recherche, toutes les possibilités d'appariement ont été testées. Au cours du processus, trois possibilités peuvent apparaître :

1. Une région, R_i^k correspond à une seule zone prédite $\tilde{Z}_j^k$: dans ce cas la région représente sans ambiguïté l'objet.
2. Une région R_i^k correspond à plusieurs zones prédites : c'est le cas de **l'occultation** (voir Fig. 2.9). L'approche EM va devoir établir une nouvelle segmentation de la région afin d'identifier l'appartenance.
3. Une zone prédite, correspond à plusieurs régions : c'est le cas de **sur-segmentation** (voir Fig. 2.10). L'approche EM a pour rôle de fusionner toutes les régions pour ne former qu'une zone.

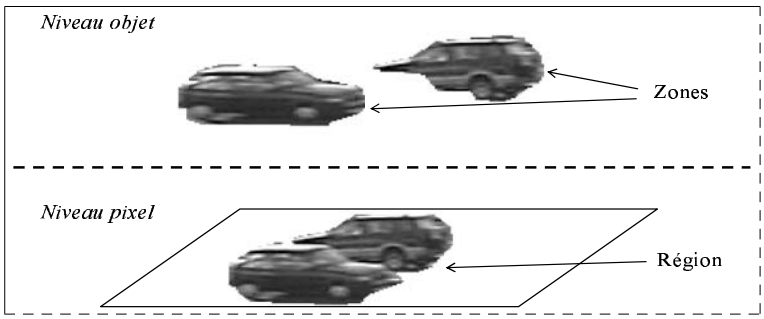


FIG. 2.9 – Représentation d'un cas d'occultation.

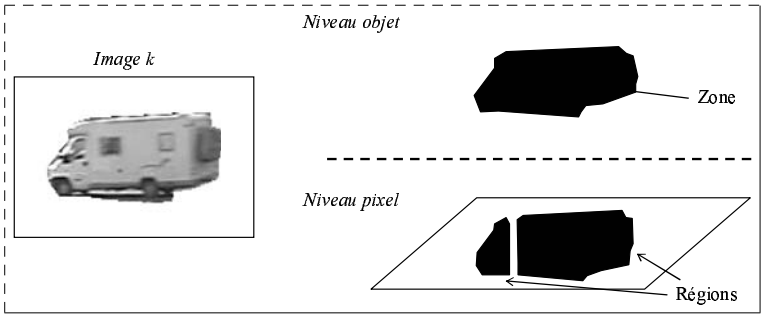


FIG. 2.10 – Représentation d'un cas de sur-segmentation.

A partir de tous ces couples actifs, nous allons créer des regroupements de nature symbolique différente. La partition de plus haut niveau concerne les objets qui entrent en occultation ou en sur-segmentation. Plus bas, nous avons les partitions construites à partir des régions et des zones.

Définition : *Partitions.*

$\mathcal{R}_l^k \equiv$ Partition l de l'ensemble des régions ayant en correspondance un groupe de zones connectées.

$\tilde{\mathcal{Z}}_l^k \equiv$ Partition l de l'ensemble des zones prédites ayant en correspondance une région appartenant à $\mathcal{R}_l^k$.

$\mathcal{P}_l^k \equiv$ Partition l de l'ensemble des descripteurs prédits $\tilde{DO}_i^k$ établie à partir de la partition des zones prédites, $\tilde{\mathcal{Z}}_l$.

Pour arriver à définir ces trois partitions, nous devons créer un nouvel opérateur, ♦, oeuvrant sur les partitions $\tilde{\mathcal{Z}}_l^k$ et $\mathcal{R}_l^k$, nous permettant de retrouver les couplages par famille.

Définition : *Opérateur de restauration des régions couplées.*

$$\blacklozenge\left[\tilde{\mathcal{Z}}_l^k\right] = \left\{R_i^k \mid \exists\ \tilde{Z}_j^k \in \tilde{\mathcal{Z}}_l^k\ /\ \exists\ C_{ij}^k\ \forall\ i \in [1,\ldots,n]\right\}$$

Définition : *Opérateur de restauration des zones prédites couplées.*

$$\blacklozenge\left[\mathcal{R}_l^k\right] = \left\{\tilde{Z}_j^k \mid \exists\ R_i^k \in \mathcal{R}_l^k\ /\ \exists\ C_{ij}^k\ \forall\ j \in [1,\ldots,m]\right\}$$

Nous obtenons la génération des partitions, $\mathcal{R}_l^k$ et $\tilde{\mathcal{Z}}_l^k$ à partir de cet opérateur à l'aide du schéma récursif suivant :

$$\begin{array}{l}
l = 0 \\
for\ j = 0,\ldots,m-1 \\
\quad if\left(\tilde{Z}_j^k \notin \left\{\tilde{\mathcal{Z}}_l^k\ \forall\ l\right\}\right) \\
\quad\quad \tilde{\mathcal{Z}}_l^k[0] = \varnothing,\ \tilde{\mathcal{Z}}_l^k[1] = \tilde{Z}_j^k,\ q = 1 \\
\quad\quad while\left(\tilde{\mathcal{Z}}_l^k[q] \neq \tilde{\mathcal{Z}}_l^k[q-1]\right) \\
\quad\quad\quad \mathcal{R}_l^k[q] = \blacklozenge\left[\tilde{\mathcal{Z}}_l^k[q]\right] \\
\quad\quad\quad \tilde{\mathcal{Z}}_l^k[q] = \blacklozenge\left[\mathcal{R}_l^k[q]\right] \\
\quad\quad\quad q = q+1 \\
\quad\quad endwhile \\
\quad\quad l = l+1 \\
\quad endif \\
endfor \\
L = l
\end{array} \tag{2.7}$$

où q fait référence au numéro de l'itération et L va définir le nombre de partitions définies à chaque image. Soient alors, m' le nombre de zones prédites contenues dans la partition $\tilde{\mathcal{Z}}_l^k$, et n' le nombre de régions contenues dans la partition $\mathcal{R}_l^k$:

$$\begin{aligned} m' &= Card(\tilde{\mathcal{Z}}_l^k) \\ n' &= Card(\mathcal{R}_l^k) \end{aligned} \tag{2.8}$$

La partition $\mathcal{P}_l^k$ est définie à partir des zones contenues dans $\tilde{\mathcal{Z}}_l^k$:

$$\mathcal{P}_l^k = \left\{\tilde{DO}_j^k\ \forall\ j\ /\ \tilde{Z}_j^k \in \tilde{\mathcal{Z}}_l^k\right\} \tag{2.9}$$

La Fig. 2.11 nous montre les différents sous-ensembles décrits précédemment.

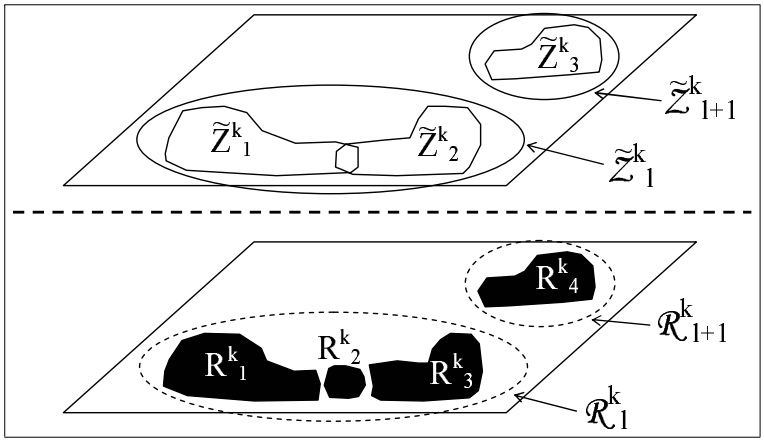


FIG. 2.11 – Représentation des différents sous-espaces créés.

La Fig. 2.12 nous montre un schéma représentatif de la création des partitions et des sous-ensembles à partir des appariements produits par le processus de mise en correspondance.

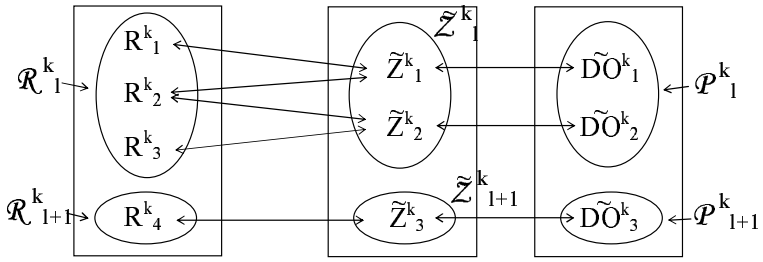


FIG. 2.12 – Représentation des différentes partitions et sous-espaces de façon schématique à partir des appariements donnés par le processus de mise en correspondance.

2.5.2 Construction des zones : approche EM

Modélisation stochastique du problème d'appariement

Dans le but de lever les ambiguïtés engendrées par les cas de sur-segmentation et d'occultation, une procédure de classification fondée sur une approche inférentielle est considérée. Initialisée par les couples [région—zone] en attente d'une mise en correspondance, l'identification doit déboucher sur la génération de l'ensemble des zones Z^k_j. Pour cela, nous devons ré-affecter

chacun des pixels sélectionnés à la fois par la détection courante, symbolisée par les R_j^k, et par la prédiction, symbolisée par le $\tilde{Z}_j^k$. Pour chaque partition, nous créons une groupement de pixels noté Ω_l tel que :

$$\Omega_l = \left\{ I^k(p) \,/\, I^k(p) \subseteq \mathcal{R}_l^k \right\} \tag{2.10}$$

où $I^k(p)$ désigne le pixel courant à la position $p = (x, y)$. A partir de cet ensemble Ω_l et ayant pour objectif de reconstruire les différents objets de la scène, nous émettons l'hypothèse que la loi statistique régissant l'appartenance d'un pixel $I^k(p)$ au groupement d'objets représenté par $\mathcal{P}_l^k$ est un mélange fini de distributions de formes analytiques connues. Ces différentes distributions expriment le lien conditionnel qui existe entre le pixel et la prédiction des descripteurs de chaque objet, $\tilde{DO}_j^k$. Formellement, le problème d'identification du mélange de distributions s'énonce ainsi :

$$\begin{gathered} f\left(I^k(p)\,|\Psi\right) = \sum_l \pi_l \cdot f\left(I^k(p)\,|\tilde{DO}_l^k, \Psi\right) \\ 0 \leq \pi_l \leq 1 \; et \; \sum_l \pi_l = 1 \end{gathered} \tag{2.11}$$

où Ψ est le vecteur regroupant les grandeurs caractérisant les distributions conditionnelles et marginales d'un objet par rapport à l'ensemble. La quantité π_l représente la probabilité d'avoir l'objet l et le terme $f\left(I^k(p)\,|\tilde{DO}_l^k, \Psi\right)$ la probabilité d'associer le pixel $I^k(p)$ à l'objet l conditionnelement à l'ensemble des prédictions des descripteurs de ce dernier. Ce type de formulation a déjà été étudié dans la "littérature". Concernant la segmentation fondée sur l'attribut mouvement, Shawney et al. ont proposé dans leurs travaux [SA96] d'exploiter un mélange fondé sur l'erreur de compensation liée au mouvement. Notre approche s'inspire de ce point de vue mais en le généralisant. Nous pouvons en effet imaginer de nombreuses formes d'expression pour la probabilité conditionnelle sachant que les descripteurs utilisés peuvent être de nature très différentes. En outre, nous avons l'opportunité dans notre schéma de séparer le mélange en exploitant les informations fournies par la boucle de rétroaction, c'est à dire, la prédiction. La généralisation et la boucle font que nous pouvons espérer une meilleure stabilité de notre procédé. Il est à noter, en outre, que nous avons réduit le nombre de modèles entrant dans le mélange, ce qui n'est pas le cas pour l'algorithme de Shawney qui a une approche globale sans partitionnement. L'intérêt de notre approche est de s'appuyer sur différentes grandeurs permettant de conditionner la méthode pour une plus grande stabilité. Nous avons en effet le contrôle sur :

- L'expansion des zones d'intérêt avec Ω_l.
- Le nombre d'objets en compétition avec $\mathcal{P}_l^k$
- l'expression des distributions conditionnelles ou marginales construites à partir de l'ensemble des descripteurs prédits permettant d'injecter des erreurs de compensation à partir des modèles de mouvement $\tilde{\Theta}_l$, des écarts colorimétriques ou fondées sur des mesures de corrélation par exemple.

Nous avons montré que l'approche proposée permet d'obtenir un niveau adéquat de description de la problématique. Si maintenant, nous nous intéressons à la méthode de résolution plusieurs solutions s'offrent à nous.

Une des méthodes itératives classiques de résolution d'un mélange est l'approche EM. Il s'agit d'enchaîner séquentiellement deux étapes :

- L'étape E (Expectation) : cette étape a pour objectif d'assigner les pixels aux différents objects contenus dans la partition $\mathcal{P}_l^k$.
- L'étape M (Maximization) : en supposant connue l'assignation, une phase d'estimation est réalisée sur la base d'un ensemble de paramètres constitué des grandeurs caractérisant les distributions conditionnelles et marginales.

La construction du vecteur Ψ dépend de l'application considérée car son contenu est conditionné par les spécificités de l'application en termes de besoin de description des objets et de leur environnement. Dans le chapitre 4, nous étudierons plusieurs propositions permettant de répondre aux problématiques de la surveillance autoroutière et du suivi de visage.

Une fois défini l'ensemble de l'espace paramétrique, nous proposons un critère global pour chaque groupement Ω_l correspondant à l'ensemble $\mathcal{P}_l^k$. Ce critère formulé sur la base du maximum a posteriori est le suivant :

$$\zeta\left(\Omega_l\right) = -\log\left(\prod_{p\in\Omega_l} f\left(I^k\left(p\right)|\Psi\right)\right) \tag{2.12}$$

Comme le processus d'affectation et d'estimation est itéré, différentes conditions d'arrêt peuvent être envisagées en observant la convergence des paramètres, l'évolution du critère global ou la stabilité des affectations.

En ce qui concerne l'affectation, nous utilisons les probabilités a posteriori d'avoir l'objet j conditionnellement au pixel $I^k(p)$. Nous avons :

$$\omega_{pj} = \frac{\pi_j \cdot prob\left(I^k\left(p\right)|\tilde{DO}_j^k\right)}{\sum\limits_{\forall l} \pi_l \cdot prob\left(I^k\left(p\right)|\tilde{DO}_l^k\right)} \tag{2.13}$$

L'ensemble de ces probabilités ω_{pj} forment m' cartes d'affectation :

$$M_l^k(p) = j \text{ si } \omega_{pj} \geq \omega_{pl} \; \forall \; l \neq j \tag{2.14}$$

Génération finale des zones

La construction des zones exploite les cartes d'affectation $M_l^k(p)$, les zones prédites $\tilde{Z}_j^k$ et les régions R_i^k issues de l'étape de pré-segmentation. Nous utilisons toutes ces informations et non pas l'une d'entre elles pour limiter les dérives dues à une prédiction moins précise. A partir d'un partitionnement des régions R_i^k contrôlé par les cartes d'affectation $M_l^k(p)$ nous allons assurer la mise à jour des zones Z_j^k. La Fig. 2.13 nous montre un exemple illustrant ce partitionnement pour une scène autoroutière. A partir d'un test de superposition spatiale entre les zones prédites et la/les régions, nous générons des sous-régions notées $\bar{R}_j^k$ et $\bar{R}_j^{*k}$. Les $\bar{R}_j^{*k}$ sont des portions de R_j^k qui sont associées à un seul objet et les $\bar{R}_j^k$ représentent celles appartenant à plusieurs objets (zones d'ambiguïté générées par l'occultation, voir Fig. 2.14). Une nouvelle partition, $\bar{\mathcal{R}}_l^k$, est créée afin de regrouper l'ensemble des sous-régions faisant partie de la partition $\mathcal{P}_l^k$. Pour construire cette nouvelle partition, nous définissons différents opérateurs.

Définition : *Intersection spatiale entre une zone prédite et une partition des régions* $\mathcal{R}_l^k$.

$$\bigstar[\tilde{Z}_j^k, \mathcal{R}_l^k] = \tilde{Z}_j^k \Cap \mathcal{R}_l^k$$

Cet opérateur nous permet de définir l'intersection spatiale entre une zone prédite $\tilde{Z}_j^k$ et la partition des régions $\mathcal{R}_l^k$.

Définition : *Intersection spatiale entre une zone prédite et le reste de zones prédites inclues dans sa partition* $\tilde{\mathcal{Z}}_l^k$.

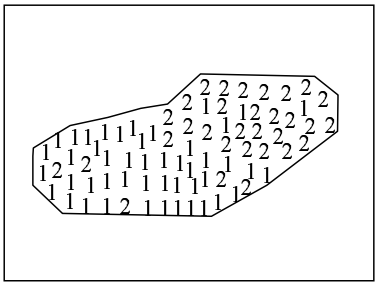
$$\clubsuit[\tilde{Z}_j^k] = \tilde{Z}_j^k \Cap \tilde{\mathcal{Z}}_{l-j}^k$$

où $\tilde{\mathcal{Z}}_{l-j}^k = \left\{ \tilde{Z}_p^k \in \tilde{\mathcal{Z}}_l^k \; / \; p \neq j \right\}$ représente le contenu de la partition $\tilde{\mathcal{Z}}_l^k$ sans la zone $\tilde{Z}_j^k$. Cet opérateur nous permet de définir l'intersection spatiale entre une zone prédite $\tilde{Z}_j^k$ et le reste des zones de la même partition.

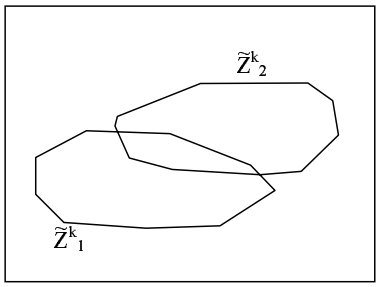
On notera la propriété suivante :

$$\bigcup \left(\cup \bar{R}_j^k, \cup \bar{R}_j^{*k} \right) = \cup R_j^k \tag{2.15}$$

(a) Région à segmenter.



(b) Cartes d'affectation établies par EM.



(c) Zones prédites.

FIG. 2.13 – Représentation des régions, cartes d'affectation et zones prédites dans un cas d'occultation dans une application routière.

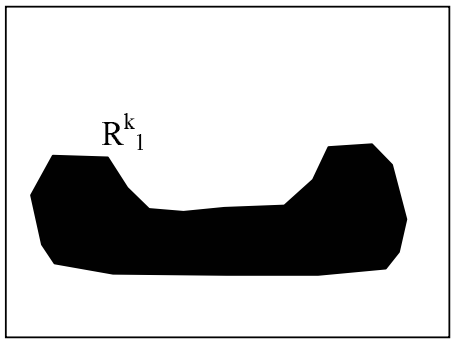
L'obtention de la partition $\bar{\mathcal{R}}_l^k$ est obtenue à partir de la procédure recursive suivante :

$$\begin{array}{l} \forall l \\ \bar{\mathcal{R}}_l^k[0] = \varnothing \\ for\ j = 1, \ldots, m' \\ \quad \bar{R}_j^k = \bigstar \left[\clubsuit[\tilde{Z}_j^k], \mathcal{R}_l^k \right] \ominus \bar{\mathcal{R}}_l^k[j-1] \\ \quad \bar{R}_j^{*k} = \bigstar \left[\left(\tilde{Z}_j^k \ominus \bar{R}_j^k \right), \mathcal{R}_l^k \right] \\ \quad \bar{\mathcal{R}}_l^k[j] = \bar{\mathcal{R}}_l^k[j-1] \bigcup \bar{R}_j^k \bigcup \bar{R}_j^{*k} \\ endfor \end{array} \tag{2.16}$$

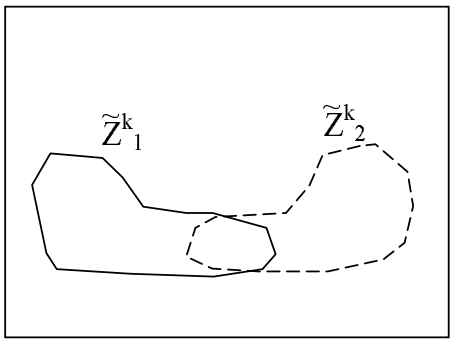
où $\ominus$ représente l'opérateur de substraction spatiale.

La Fig. 2.14 présente un exemple montrant la génération de trois sous-régions.

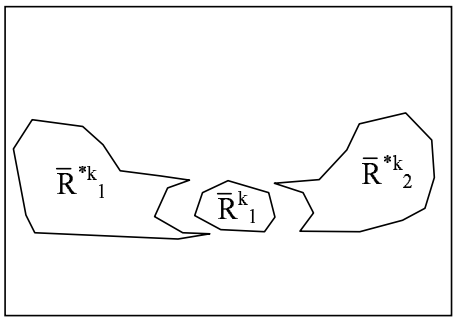
Pour l'affectation, nous cumulons les cartes d'affectation pour chaque sous-région. La Fig. 2.15 nous montre un exemple d'assignation.



(a) Région contenant les objets en occultation.



(b) Zones prédites des objets.



(c) Sous régions crées à partir de la région et des zones prédites.

FIG. 2.14 – Représentation graphique de la construction des sous-régions à partir d'un exemple constitué de deux zones prédites et d'une seule région.

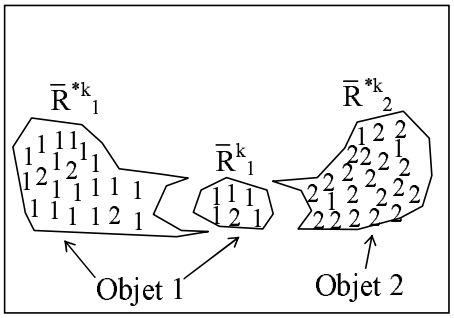


FIG. 2.15 – Exemple d'assignation des sous-régions aux objets identifiés par cumul des cartes.

Une fois que toutes les sous-régions ont été identifiées et assignées à un objet, nous pouvons réaliser la mise à jour des zones – forme univoque de l'objet. Pour cela, nous calculons l'enveloppe convexe associée à l'ensemble des sous-régions affectée à un objet. La Fig. 2.16 présente un exemple de construction.

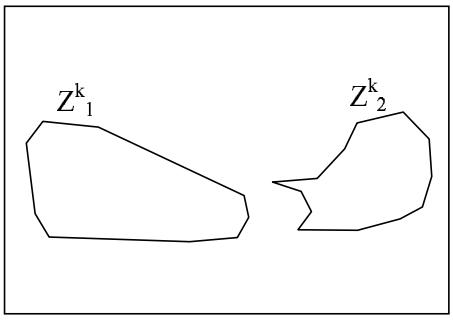


FIG. 2.16 – Création d'une zone à partir de la fusion de deux sous-régions différentes.

2.6 Mise à jour des descripteurs

La mise à jour de l'ensemble des descripteurs est la dernière opération à réaliser pour boucler le processus de poursuite.

Cette phase de mise à jour est réalisée objet par objet, et a pour but de réactualiser les descripteurs pour prendre en compte les changements de forme, de taille, d'état de l'objet et surtout pour s'adapter aux variations du modèle de mouvement.

Pour mettre à jour les descripteurs, nous utilisons les anciens descripteurs, DO_j^{k-1}, ainsi que les nouvelles zones extraites, Z_j^k, en considérant l'état de l'objet. L'état, comme nous l'avons présenté dans la section 2.2, est le descripteur qui rend compte de la situation de l'objet dans la scène. Il nous informe sur les configurations de sur-segmentation, d'occultation ou correspondant au cas normal. La mise à jour sera différente selon les cas :

- Pour le cas dit normal et pour le cas de sur-segmentation la zone qui va identifier l'objet correspond directement à l'objet et il n'y a donc aucun problème pour faire la mise à jour de tous les descripteurs.
- Pour le cas d'occultation, il faut d'abord savoir si l'objet est l'occulté ou l'occultant. Si l'objet est au premier plan alors la zone représente de façon complète l'objet. La mise à jour est directe. Si au contraire,

l'objet est occulté la zone ne représente plus complètement l'objet. Nous bloquons la mise à jour.

Les étiquettes qui définissent l'état de l'objet par rapport à sa situation dans l'image (entrant, vivant ou sortant) sont traitées après que tous les autres descripteurs aient été mis à jour.

2.7 Gestion dynamique des objets

Dans cette section, nous étudions la mise à jour du nombre d'objets. Nous envisageons les cas de création et de destruction d'objets conformément à sa situation dans l'image.

2.7.1 Création d'un objet

Dans la section 2.2 nous avons établi les conditions qualifiant l'entrée et la sortie d'un objet de la scène. Pour la naissance d'un objet, la procédure est mise en place lorsque l'on trouve une région non-affectée.

Son modèle de mouvement est alors initialisé en évaluant la distance maximale entre l'origine de la région et la limite de l'image la plus proche, comme on peut le voir sur la Fig. 2.17. Après sa sortie du cadre, l'état du pixel passe à l'état *entrant* dans la scène et il entre alors dans la boucle complète.

FIG. 2.17 – Création d'un nouveau modèle. Le modèle de mouvement est initialisé en évaluant la distance maximale au bord de l'image.

2.7.2 Destruction d'un objet

Pour pouvoir détruire un modèle il faut que son état passe à l'état *sortant* ou bien que la taille de l'objet soit inférieure à un seuil prédéfini.

2.8 Conclusion

Dans ce chapitre, nous avons proposé une architecture algorithmique générique dédiée à la poursuite d'objets en vidéo-surveillance. Comme nous l'indique la Fig. 2.2, le schéma fonctionnel du processus de suivi présente une boucle de rétroaction. Cette boucle a pour objectif de stabiliser le suivi grâce à un retour de connaissance. Ce *"feedback"* permet de palier aux imperfections des différentes étapes de segmentation et d'estimation que nous rencontrons tout au long de la boucle de traitement. Deux points conditionnent la pertinence de ce schéma de suivi. Le premier point est l'étape de pré-segmentation. Pour un problème donné, la phase de segmentation bas-niveau doit exploiter des attributs ou des caractéristiques suffisamment discriminantes pour la classe d'objets que l'on veut suivre (mouvement, couleur, forme, ...). Le second point est la construction du modèle de l'objet. Les descripteurs qui le composent doivent être suffisamment représentatifs pour assurer la reconnaissance individuelle de chaque objet.

Dans ce chapitre nous avons proposé différents descripteurs qui nous semblent indispensables quelle que soit l'application. Deux sont particulièrement importants.

1. Représentation univoque de l'objet, la zone est la signature spatiale de l'objet et c'est grâce à sa connaissance que l'on peut réaliser les tests d'appariement.
2. Le second est le modèle de mouvement qui permet une prédiction de la zone.

Dans les chapitres suivants de ce mémoire, nous allons nous attacher à illustrer la pertinence du schéma proposé. Le chapitre 3 est notamment dédié à l'étape de pré-segmentation. Nous proposons en effet une nouvelle méthode dédiée à la détection de mouvement. Cette problématique est évidemment très importante en vidéo-surveillance ce qui explique que nous lui consacrons un chapitre entier.

Chapitre 3

Détection de mouvement

3.1 Introduction

La détection de mouvement consiste à segmenter une image en régions regroupant les pixels dont le contenu est associé à un objet se déplaçant dans la scène.

Dans ce contexte on distingue deux problématiques : dans la première, on considère des images obtenues à l'aide d'un capteur fixe pour lesquelles il existe un fond stable et une avant scène constituée d'objets en mouvement.

La deuxième problématique concerne les situations où le capteur est déplacé durant la prise de vue, ce qui conduit à une image où toutes les régions sont en évolution. Le déplacement du capteur introduit alors un mouvement dominant que certains auteurs essaient de compenser afin de se ramener au premier cas.

Le travail présenté dans ce document se situe dans le cadre statique (caméra fixe ou après compensation du mouvement dominant). Dans ce cas, la plupart des techniques décrites dans la "littérature" sont fondées sur la détection des changements temporels des pixels de l'image. Parmi celles-ci, la plus ancienne, mais aussi la plus simple et la plus fréquemment utilisée, s'appuie sur la différence temporelle d'intensité lumineuse entre deux images consécutives, calculée pixel par pixel. Les pixels appartenant à une région en mouvement sont retenus si cette différence excède un certain seuil. En deçà de ce seuil le pixel est considéré comme appartenant à une région statique. Cette règle n'est pas exempte d'erreurs car des intensités identiques n'impliquent pas que l'objet observé soit statique. Par ailleurs, une différence temporelle non nulle n'implique pas nécessairement un objet en mouvement du fait du bruit présent le long de l'axe temporel. Le choix du seuil optimal

conduit au meilleur compromis permettant de minimiser simultanément les erreurs de première et deuxième espèce décrites précédemment.

Dans une approche simple, le seuil est préfixé a priori, tandis que dans les approches plus élaborées le seuil est adaptatif. Cette adaptation peut résulter d'une estimation du bruit affectant les images comme dans [LL02] ou d'une évaluation de l'entropie de la différence inter-image ou par des tests de vraisemblance utilisant le voisinage du pixel étudié. Cette dernière approche, plus complexe, est aussi la plus robuste et tire profit de la cohérence spatiale des pixels inclus dans le voisinage.

Une autre famille d'approches construit une image de référence correspondant au fond (background), dont certaines parties sont occultées par les objets en mouvement présents dans l'avant-scène (foreground). Les objets en mouvement sont alors détectés en réalisant la différence pixel à pixel entre l'image observée et l'image de référence et en comparant cette différence à un seuil. Cette méthode est confrontée aux mêmes problèmes de determination du seuil optimal auxquels se surajoutent la sensibilité aux vibrations du capteur et la difficulté d'élaborer l'image de référence surtout si la scène est affectée par des variations de luminosité. On contourne généralement ce dernier problème par une mise à jour adéquate de l'image de référence. Cette mise à jour peut être réalisée par différentes techniques.

Dans [Wen83] les auteurs présentent un procédé très élémentaire où l'utilisateur choisit l'image qui sera ultérieurement utilisée comme image de référence. Un tel procédé n'est évidemment pas adapté aux situations affectées par des changements de luminosité. Dans [WD84] la mise à jour est effectuée automatiquement lorsqu'aucun objet en mouvement n'a été détecté, ce qui suppose qu'une telle situation se rencontre avec une fréquence suffisante.

A l'opposé de ces méthodes fondées sur une prise de décision globale, les auteurs dans [DHA88b] et [KB90] proposent des techniques de décision locale permettant de faire évoluer le niveau de référence de chaque pixel indépendamment les uns des autres. Pour ce type d'approches, une famille de méthodes propose une construction adaptative d'une image de référence. Appelée approche ARI pour Adaptive Reference Image, la technique est implantée de façon à mettre à jour les pixels de l'image de référence de manière indépendante. Ainsi, dans [SG99], un mélange de distributions Gaussiennes est utilisé pour la modélisation temporelle de l'intensité de chaque pixel.

Lorsque l'hypothèse d'un faible taux moyen d'occultation est valide on peut estimer l'image de référence à l'aide d'un filtre temporel passe-bas. Il peut être implanté dans une forme récursive (filtre AR) selon l'équation suivante :

$$B^{k+1}=\alpha B^k+(1-\alpha)\, I^k \tag{3.1}$$

où B^k représente la $k^{ème}$ image de référence et I^k l'image actuelle et α est le facteur d'oubli qui détermine la dynamique d'évolution de l'image de référence. Ce dernier peut être préfixé comme dans [KB90] ou variable comme dans [VMBP96]. Dans cette technique les auteurs utilisent des images de gradients au lieu d'images d'intensité de luminosité. Cette version améliore la robustesse de la construction de l'image de référence aux variations de luminosité globales car les gradients sont moins sensibles à ce type de variations que les images d'intensité. Par ailleurs on obtient une image de détection qui ne prend en compte que les contours des objets, qui dégradent les performances du détecteur. De plus, les contours peuvent varier énormément d'une image à l'autre d'où leur utilisation peu fréquente dans les processus de détection de mouvement. Quel que ce soit le cas, à savoir, avec un α constant ou variable, la procédure de mise à jour de la référence est simple mais peu robuste. Dans les cas où l'illumination globale reste constante (comme c'est le cas des scènes d'intérieur avec lumière artificielle), l'image de référence est bien adaptée à certains phénomènes qui peuvent survenir (comme l'arrêt d'objets en plein milieu de la scène). Par contre, si la luminosité globale varie alors la référence est mal adaptée ce qui entraîne une mauvaise qualité de la détection de mouvement. Une modification est introduite dans l'équation (3.1) pour séparer les mises à jour des pixels qui font partie du background de celles du foreground [KWM94]. Nous reviendrons sur ce type de technique plus en détail dans le paragraphe 3.2.2.

Dans ce chapitre nous proposons un algorithme de détection de mouvement fondé sur une approche multi composantes. Il utilise de façon concomitante l'intensité lumineuse de l'image actuelle et de l'image de référence comme dans les méthodes précédentes, mais aussi la chrominance et les images de module de gradient. Le résultat de cette coopération de méthodes permet de détecter les objets en mouvement en les distinguant de l'ombre portée de ces mêmes objets, tout en rendant plus robuste et plus précis le résultat obtenu. Il s'agit d'une méthode locale visant à construire récursivement une image de référence (ARI) plus stable face aux changements de luminosité locale et aux croisements d'objets (occultation mutuelle). La décision de mise à jour du niveau de référence d'un pixel (état du pixel) est prise en fonction de l'analyse de l'évolution temporelle de l'intensité observée. Ce principe conduit à une segmentation de la courbe temporelle permettant de distinguer 2 états du pixel considéré : *occultation* ou

non occultation. Pour ce faire, une machine à états finis appropriée est mise en oeuvre.

Ce chapitre est organisé comme suit : la section 3.2 décrit le processus d'extraction de l'image de référence ou processus ARI. L'extraction de la carte de mouvement est décrite dans la section 3.3. Deux approches sont présentées dans cette section, l'approche classique et l'approche multi-composantes. L'approche multi-composantes permet la détection et la suppression de l'ombre accolée aux objets en mouvement permettant une caractérisation plus robuste de l'objet, indépendante des conditions de luminosité. Dans la section 3.4 on montre les différents résultats obtenus, comparés avec d'autres méthodes décrites dans la "littérature".

3.2 Image de Référence Adaptative (ARI)

Un nouveau processus d'Image de Référence Adaptative (ARI) est introduit dans les trois prochaines sous-sections. Ce processus établit une mise à jour de l'image de référence par rapport aux conditions de luminosité globales du fond de l'image (background). La Fig. 3.1 montre le diagramme complet du processus ARI. De façon à éviter des mises à jour inadéquates de l'image de référence, cette méthode propose une mise à jour différente pour les pixels appartenant au foreground et au background de l'image. Un détecteur de passage d'objets est développé dans la section 3.2.1 permettant de faire évoluer l'état du pixel. Une machine d'état (section 3.2.3) contrôle cette évolution et établit la mise à jour de la référence (section 3.2.2).

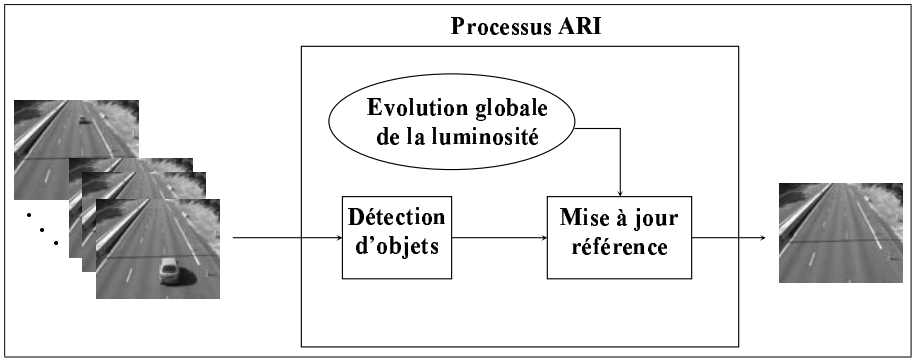
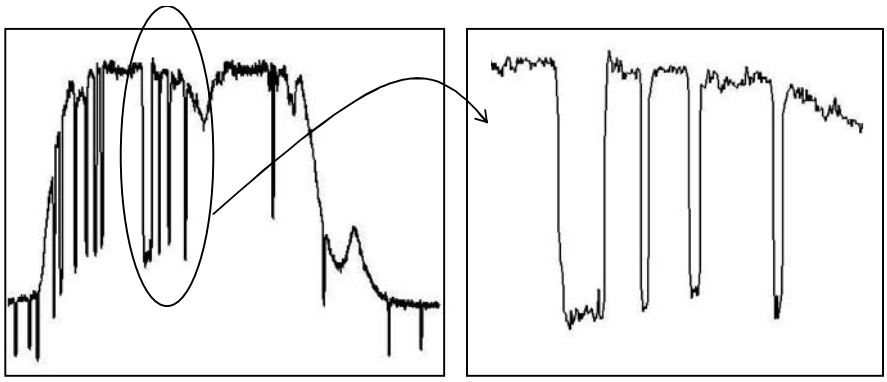


FIG. 3.1 – Diagramme complet du processus ARI.

3.2.1 Détecteur de passages d'objets

La détection d'un passage d'objet est fondamentale pour une mise à jour correcte de l'image de référence. Dans le but de développer ce type de détecteur, analysons l'évolution temporelle d'un pixel caractérisée par sa courbe d'intensité de luminosité. Il est à noter que la mesure luminance peut être remplacée par une mesure de chrominance. Soit $I^k(p)$ la fonction d'intensité de luminosité d'un pixel p à l'image k, où $p = (x, y)$. Cette fonction, dont un exemple est donné par la Fig. 3.2, nous montre les variations de la composante de luminance (ou chrominance selon le cas).

(a) Sélection d'un pixel p dans l'image.



(b) Détail de la courbe d'intensité du pixel p.

FIG. 3.2 – Evolution temporelle de la luminance d'un pixel p.

Dans la Fig. 3.3 on distingue l'évolution de l'éclairement (courbe pointillée) des périodes d'occultation qui se traduisent par une variation rapide du niveau (ici la luminance des objets est inférieure à celle du fond

pour le pixel considéré). Notons que l'évolution de l'éclairement du fond peut résulter de phénomènes naturels tels que le passage de nuages, ou l'évolution de l'ombre portée d'un objet statique. Malgré tout, l'évolution de l'éclairement correspond à une dynamique beaucoup plus lente que l'évolution de la luminance résultant d'une occultation. Cette propriété est utilisée pour distinguer les deux types d'évolution.

Dans la suite nous appellerons enveloppe la courbe représentant l'évolution de l'éclairement (courbe en pointillée de la Fig. 3.3).

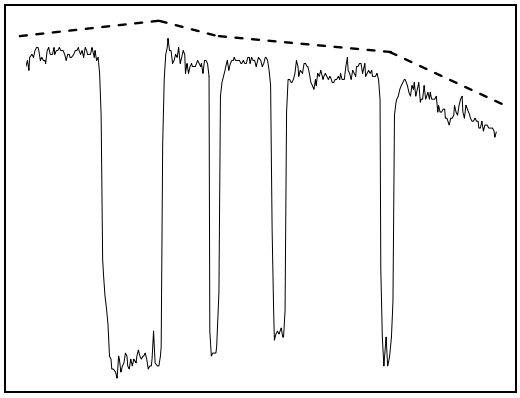


FIG. 3.3 – Evolution de l'éclairement (en pointillé) et occultations dues aux passages d'objets.

L'intensité de référence du pixel considéré est donc déterminée par l'enveloppe décrite ci-dessus. Deux problèmes sont donc à résoudre. Le premier consiste à détecter les occultations et le deuxième à estimer l'enveloppe.

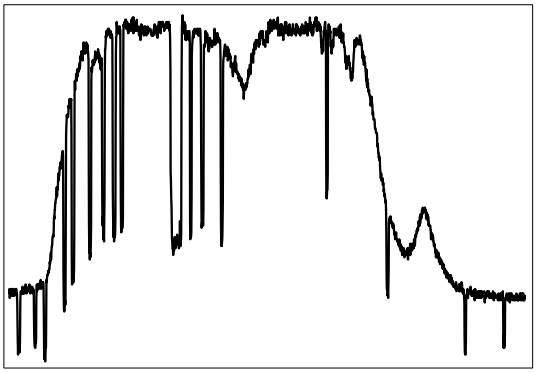
Un opérateur sensible aux transitions d'état du pixel occulté-non occulté et non occulté-occulté est fondé sur la variationnelle calculée sur un intervalle compact Ω de durée prédéterminée.

$$VL\,[I] = \int_{\Omega} \left| \frac{\partial I}{\partial t} \right| dt \tag{3.2}$$

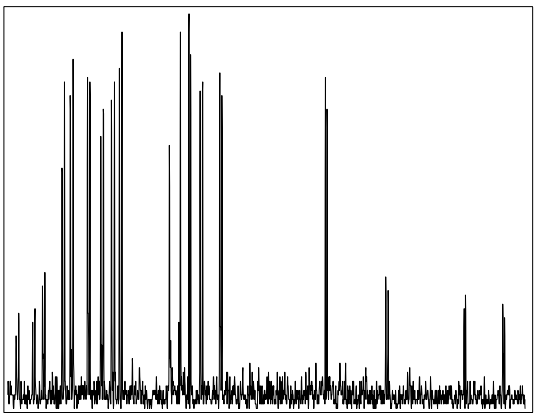
Le choix de la norme L_1 retenue ici est justifié par le modèle statistique sous-jacent qui présente une distribution appelée *"déviée de la normale"*. Dans ce contexte, on montre dans [Hub81] que la norme L_1 est plus robuste que la norme L_2.

La Fig. 3.4 montre le résultat de l'application de cet opérateur. On observe sur cette figure que les instants de transition sont nettement perceptibles et que le bruit résiduel est d'un niveau relativement faible. Notons que l'intervalle Ω doit être plus petit que la plus petite durée d'occultation afin

de séparer le début de la fin de l'occlusion. Notons enfin que si l'éclairement varie, le bruit résiduel ainsi que les pics transitionnels varient dans le même sens comme on le constate sur la Fig. 3.4.



(a) Evolution temporelle de l'intensité de luminosité d'un pixel.



(b) Résultat de l'opérateur de variationnelle appliqué à la fonction d'intensité de luminosité.

FIG. 3.4 – Fonction d'intensité de luminosité et son opérateur de variationnelle associé.

Pour prendre en compte ce phénomène nous avons développé un opérateur de détection de pics transitionnels utilisant un seuillage adaptatif.

La présence du pic transitionnel est représentée par la variable booléenne déterminée par :

$$OC_p^k = \begin{cases} 1 & si\ VL_p^k \geq T_p^k \\ 0 & ailleurs \end{cases} \tag{3.3}$$

où T^k représente le seuil adaptatif. Pour définir ce seuil on utilise le moment d'ordre deux, noté m, de l'opérateur de variationnelle VL_p^k. Le seuil T^k est mis à jour à partir de la racine carrée du moment m, tel que :

$$T_p^k = c \cdot \sqrt{m_p^k} \tag{3.4}$$

où c est une constante prédéfinie par l'utilisateur et m_p^k une estimation du moment d'ordre 2 de VL_p^k :

$$m_p^{k+1} = \beta_p \cdot m_p^k + (1-\beta_p) \cdot \left(VL_p^k\right)^2 \tag{3.5}$$

avec

$$\beta_p = 1 - \alpha \cdot \left(1 - OC_p^k\right) \tag{3.6}$$

où le paramètre α est la mémoire du système récursif que nous avons déjà introduit dans l'équation (3.1). L'opérateur OC_p^k est introduite dans (3.6) de façon à ne pas prendre en compte les passages d'objet. Ainsi la valeur du seuil est cohérente avec le niveau de bruit présent dans l'image. Un seuil bas est rajouté dans le calcul du moment m pour éviter qu'un pixel avec une valeur saturée (constante) puisse faire tomber malencontreusement la valeur du seuil T à des valeurs proches de zéro :

$$m_p^{k+1} = \begin{cases} m_p^{k+1} \ si \ m_p^{k+1} > \tau_l \\ \\ m_p^k \quad sinon \end{cases} \tag{3.7}$$

Si on superpose la courbe de la variationnelle à celle du seuil T on peut voir que le seuil varie uniformément par rapport au contenu du bruit dans l'image (Fig. 3.5). Le seuil T est toujours situé au dessus du bruit et en dessous des pics qui vont nous permettre d'effectuer une détection robuste des passages d'objets.

3.2.2 Mise à jour de la référence

L'image de référence doit être adaptée aux conditions globales de luminosité de la scène. Dans le cas de séquences d'extérieur, les variations de luminosité globales sont produites par des phénomènes naturels comme les passages nuageux, le lever ou le coucher du soleil, etc. Ces phénomènes vont provoquer une forte variation dans la courbe d'intensité de luminosité, comme on peut le voir sur la Fig. 3.2(b). Comme on l'a déjà dit plus haut, la mise à jour de l'image de référence doit se faire en conformité avec l'enveloppe de la courbe de luminosité.

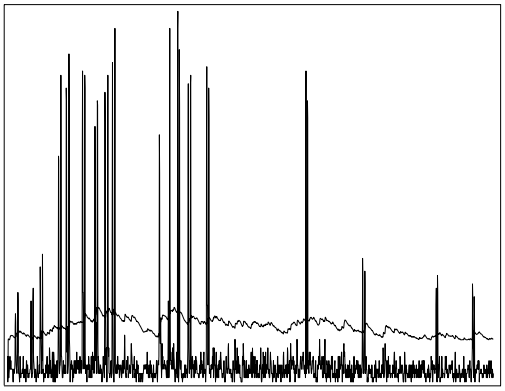


FIG. 3.5 – Seuil superposé sur la courbe de la variationnelle.

De nombreuses méthodes sont présentées dans la "littérature", notamment celle décrite dans l'équation (3.1) au paragraphe 3.1, qui met en oeuvre un filtre passe-bas pour séparer les fréquences basses dont l'origine traduit l'évolution lente de l'éclairage des fréquences élevées résultant des occultations brèves.

La fréquence de coupure du filtre passe-bas est contrôlée par le paramètre d'oubli α qui intervient dans l'implantation récursive.

Lorsque le taux d'occultation devient important la sortie du filtre n'est plus en conformité avec l'enveloppe recherchée et l'image de référence devient inadéquate.

Pour pallier cet inconvénient, Koller et al. [KWM94] proposent une approche qui permet d'effectuer la mise à jour de l'image de référence B avec une dynamique rapide si le pixel est considéré *non occulté* et une dynamique lente si le pixel est considéré *occulté*. Ces deux dynamiques sont respectivement contrôlées par les facteurs d'oubli α_1 et α_2 dans l'équation récursive :

$$B^{k+1} = B^k + \left(\alpha_1 \cdot \left(1 - M^k\right) + \alpha_2 \cdot M^k\right) \cdot \left(I^k - B^k\right) \tag{3.8}$$

où M^k est la carte des occultations (rappelons que l'occultation est synonyme de mouvement car le fond est occulté par un objet en mouvement).

Bien que procurant un net progrès, cette méthode reste imparfaite en particulier si les durées d'occultation sont importantes. La Fig. 3.6 montre en pointillé l'évolution de l'intensité d'un pixel de l'image de référence B qui reste assez éloignée de l'enveloppe souhaitée. Pour résoudre ce problème nous proposons dans la suite de ce chapitre une approche robuste pour la mise à jour de l'image de référence B, s'approchant de l'enveloppe recherchée

afin de limiter les fausses détections de mouvement (fantômes) ou l'absence de détection.

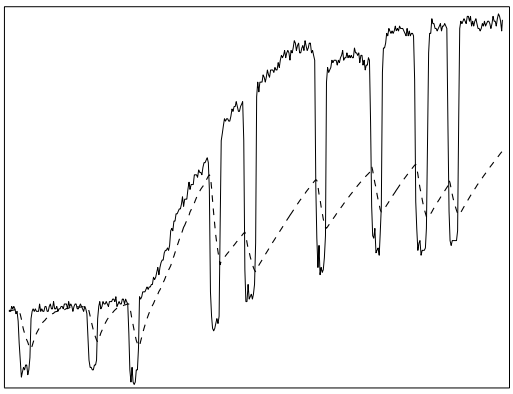


FIG. 3.6 – Estimation de la référence (en pointillé) à partir de la méthode de Koller et al.

Pour ce faire nous allons utiliser l'opérateur OC de détection de début et de fin d'occlusion pour élaborer le masque d'occultations M qui nous permettra de mettre en oeuvre les stratégies appropriées à la mise à jour de B en fonction de l'état du pixel (comme le montre la Fig. 3.7).

Mise à jour sans occultation

S'il n'y a pas de passage d'objet on se trouve dans une situation triviale où l'enveloppe se confond avec l'image courante (Fig. 3.7) :

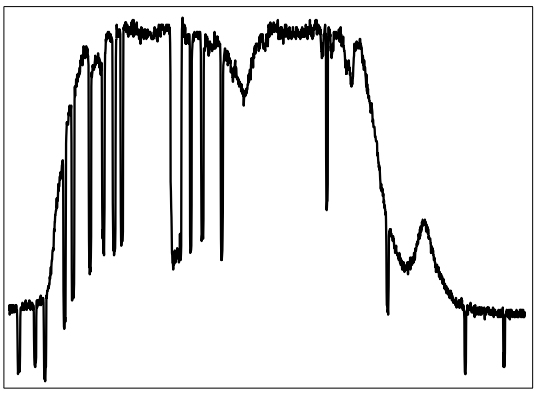
$$M^k = 0 \Rightarrow B^k = I^k \tag{3.9}$$

Mise à jour en cas d'occultation

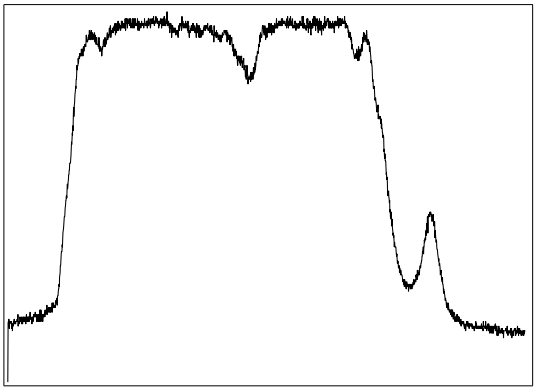
Ce cas est plus complexe et doit tenir compte de deux éventualités.

La première est à considérer lorsque l'éclairage ne varie pas durant l'occultation. Dans ce cas, l'image de référence B ne doit pas évoluer (ligne pointillée de la Fig. 3.8).

La deuxième se produit lorsqu'une évolution rapide de l'éclairement se produit durant le passage d'un objet. Bien que ce cas ne soit pas fréquent nous proposons un algorithme qui permet d'estimer l'évolution de l'image de référence durant l'occultation de manière à ce que l'enveloppe interpolée retrouve le niveau souhaité à la fin de l'occultation (ligne pointillée de la Fig. 3.9).



(a) Intensité de luminosité



(b) Enveloppe de la courbe

FIG. 3.7 – Courbes d'intensité de luminosité et son enveloppe.

Notons qu'il s'agit d'une interpolation en temps réel car la valeur d'arrivée du segment interpolée n'est pas connue. Il s'agit plus précisément d'une prédiction construite sur un extrapolateur d'ordre 1.

Cet algorithme est fondé sur une évaluation récursive de la pente P qui traduit une approximation linéaire locale de l'évolution de B selon la formule recursive :

$$P^k = \lambda P^{k-1} + (1-\lambda) \cdot \left(B^k - B^{k-1} \right) \tag{3.10}$$

L'équation de mise à jour s'écrit alors :

$$B^k = B^{k-1} + P^{k-1} \tag{3.11}$$

Ces deux équations se condensent en une seule :

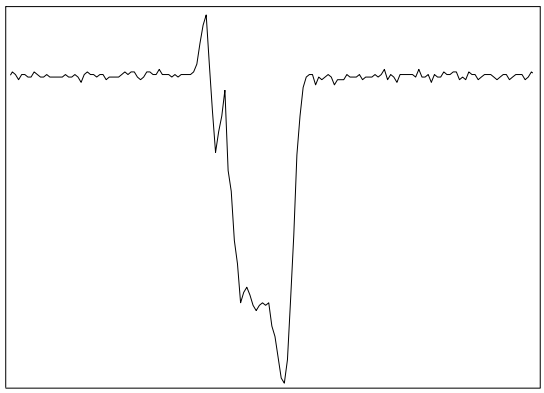


FIG. 3.8 – Enveloppe constante de la courbe d'intensité de luminosité pendant le passage d'un objet.

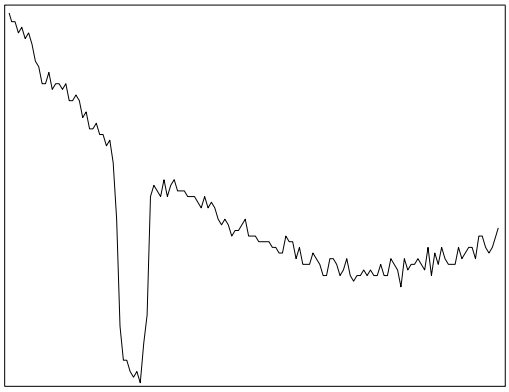


FIG. 3.9 – Changement de luminosité globale pendant le passage d'un objet.

$$B^{k+1} = (1 - M^k) \cdot I^{k+1} + M^k \cdot \left(B^k + P^k \right) \tag{3.12}$$

où en désignant par $\tilde{D}^{k+1}$ la différence $\tilde{D}^{k+1} = I^{k+1} - B^k$

$$B^{k+1} = B^k + \left(M^k \cdot P^k + (1 - M^k) \cdot \tilde{D}^{k+1} \right) \tag{3.13}$$

Des exemples d'application de cet algorithme sont montrés dans le cas d'éclairage constant (Fig. 3.10) et dans le cas d'éclairage variable (Fig. 3.11).

Notons que sur la Fig. 3.10 un pic négatif apparaît à la fin de l'occultation. Ce phénomène provient d'une détection de fin d'occultation trop

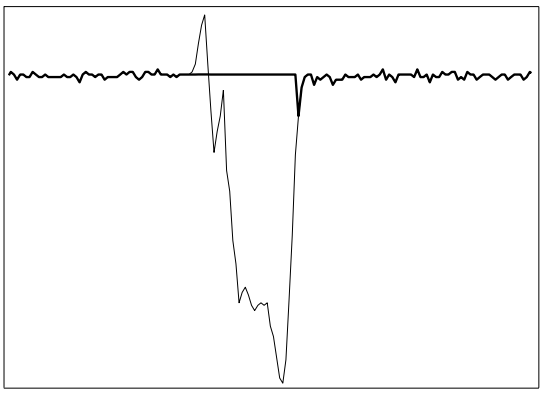


FIG. 3.10 – Mise à jour de la référence sans changement de luminosité globale.

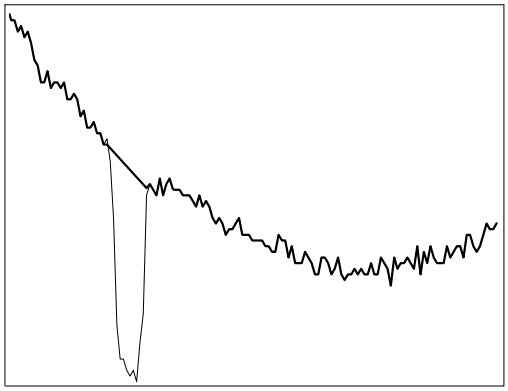


FIG. 3.11 – Mise à jour de la référence avec changement de luminosité globale.

précoce : la valeur de M^k s'est annulée avant la fin de l'occultation effective.

3.2.3 Analyse de l'état du pixel

Comme nous l'avons déjà décrit précédemment, chaque pixel de l'image se trouve dans un état d'occultation ou non. Le suivi de l'évolution de l'état sera effectué à l'aide d'une machine d'état contrôlée par le signal OC.

Dans cette section nous allons établir les lois d'évolution de l'état d'un pixel. Cet état du pixel va définir le comportement du pixel (i.e. la mise à jour de la référence, les passages d'objets, etc). Une machine d'états contrôle l'évolution de l'état du pixel à travers l'opérateur de détection de passages

d'objets OC^k obtenu dans (3.3). L'évolution de cet opérateur détermine les passages d'un état à l'autre à travers la machine d'état.

Deux états différents vont distinguer les situations d'un pixel :

1. ***État de non occultation*** : aussi appelé état de repos. La difficulté essentielle du procédé réside dans la détermination fiable de l'état d'un pixel selon qu'il est occulté ou non.

   Le cas dit *"élémentaire"* se produit lorsque les objets en mouvement sont bien texturés. Dans ce cas l'opérateur OC est égal à 1 durant l'occultation et à zéro dans le cas contraire.

   Lorsque l'objet n'est pas texturé, l'opérateur OC est nul durant l'occultation et ne passe à 1 qu'en début et fin d'occultation. Le cas le plus "problématique" survient lorsque l'objet présente une luminosité voisine du fond. Ainsi le début et la fin d'occultation seront difficilement perceptibles.

   L'état du trafic peut conduire aussi à des situations délicates. Par exemple en cas de trafic intense, des pixels peuvent être occultés pratiquement en permanence et la mise à jour de l'image de référence ne s'opère plus. De même en cas d'embouteillage et arrêt du trafic, OC ne passe plus à 1 : il y a risque de confusion entre les *pixels objets* et les *pixels fond*.

   La possibilité d'erreur d'étiquetage de l'état d'un pixel doit donc être prise en compte. Le processus de détermination de cet état doit permettre le recouvrement des erreurs.

   Pour ce faire il sera parfois nécessaire de faire appel à une couche de traitement de niveau supérieur dans laquelle s'opère le suivi des objets. Cette couche sera examinée au chapitre 4.

2. ***État d'occultation*** : dans cet état on va trouver les pixels pour lesquels une détection de passage d'objet a été reconnue (et bien entendu, on n'a pas trouvé la fin de passage de l'objet). Dans ce cas, la valeur d'intensité du pixel correspond à l'objet et non au fond de l'image.

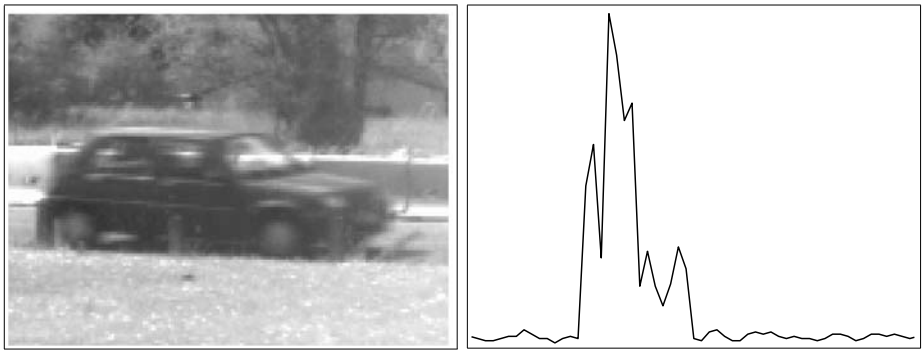
Avant de détailler l'évolution des différents états d'un pixel nous allons citer la liste des cas critiques qui peuvent survenir lors de la détection d'objets en mouvement. On va séparer les différents cas par rapport à deux axes majeurs : l'un concernant la texture de l'objet et l'autre la détection ou non de la fin de passage d'un objet. Ainsi on trouve :

Par rapport à la texture : Dans le cas d'un objet texturé, la variationnelle va être souvent supérieure au seuil (ce qui se traduit par l'activation de l'opérateur $OC = 1$) jusqu'au moment de la fin de passage de

l'objet (Fig. 3.12). Dans ce cas il suffit de déterminer quand l'opérateur OC vaut zéro. Par contre si l'objet est peu texturé il peut arriver que l'opérateur OC soit nul pendant un laps de temps significatif, comme le montre la Fig. 3.13. Dans ce cas, il faut avoir une deuxième condition pour reconnaître la fin du passage d'objet.

Par rapport à la fin de passage de l'objet : L'état d'occultation du pixel cesse lorsqu'on détecte la transition de fin d'occultation. Par exemple, si l'objet s'arrête, un tel signal ne se produit pas. De nombreux autres cas de confusion peuvent conduire à une décision erronée quant à l'état d'un pixel.

La résolution du problème posé par le recouvrement de ces erreurs doit être explorée en faisant appel à des techniques de haut-niveau.



(a) Objet texturé. (b) Courbe de la variationnelle sur l'objet.

FIG. 3.12 – Image réelle et courbe de variationnelle pour un objet texturé.

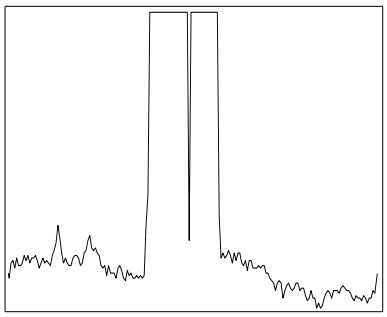
État 1 : État de non occultation

Cet état est aussi appelé état de repos. L'intensité du pixel correspond à celle du fond. Dans ce cas l'image de référence se déduit de façon élémentaire de l'image observée :

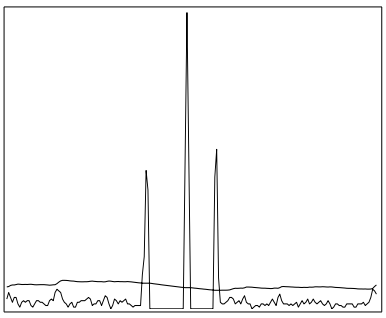
$$B^k = I^k \tag{3.14}$$

A l'instant initial de la séquence d'images les pixels sont considérés non occultés, ce qui n'est pas forcement exact. Dans ce cas il faudra donc corriger cette erreur. Ceci peut être généralement résolu en appliquant l'heuristique suivante : *les durées d'occultation sont plus faibles que les durées de non occultation.*

(a) Détail du pixel sur une zone très peu texturée de l'objet.



(b) Courbe d'intensité de luminosité.



(c) Courbe de la variationnelle.

FIG. 3.13 – Détail d'un pixel sur une zone très peu texturée de l'objet et sa courbe de variationnelle associée.

Dans le cas particulier d'un taux d'occultation élevé, cette propriété n'est pas toujours vérifiée. Il en est de même si l'engorgement de la circulation conduit à l'arrêt des véhicules. Pallier cette situation nécessite alors de la prise en compte de solutions algorithmiques de plus haut niveau.

État 2 : État d'occultation

Un pixel est affecté à cet état lorsque nous détectons le passage d'un objet sur lui. Dans cet état la mise à jour de la référence va suivre l'équation (3.11) à partir de laquelle on essaye de suivre les possibles variations de l'enveloppe de la courbe d'intensité de luminosité dues aux changements de luminosité globale. Le pixel va rester dans cet état jusqu'au moment où l'objet termine l'occultation du pixel.

Le seul cas où l'état du pixel ne comporte pas de confusion possible correspond à $OC = 1$. Dans les autres cas on va faire appel à un ensemble

TAB. 3.1 – Conditions pour la fin de passage d'un objet

Cas	Conditions
1er cas	Pas de passage d'objet & $\left\|D_p^k\right\| < \tau_l$
2ème cas	Pas de passage d'objet pendant N_1 images & $\left\|D_p^k\right\| < \tau_h$

de conditions supplémentaires pour limiter les risques d'erreur.

Pour ce faire on va s'appuyer sur l'image de référence en considérant le cas où l'actualisation de B^k s'est bien réalisée et le cas où elle n'est qu'approximative.

1er cas. Image de référence exacte.

Dans ce cas on va dire qu'un pixel est non occulté si les conditions suivantes sont vérifiées :

$$OC_p^k = 0 \ \& \ \left|I_p^k - B_p^k\right| < \tau_l \tag{3.15}$$

où τ_l est un seuil prédéfini par l'utilisateur. Ceci est le cas typique qui suppose cependant que la luminance de l'objet occultant est sensiblement différente de celle du fond.

2ème cas. Image de référence approximative.

Lorsque la durée d'occultation est importante, la mise à jour de B^k ne s'opère qu'à travers une estimation par extrapolation du 1er ordre. En fin d'occultation des écarts entre I^k et B^k peuvent être significatifs (voir Fig. 3.14), même en cas de non occultation. Dans ce cas on va accepter un écart $\tau_h > \tau_l$, mais on va exiger que OC^k reste nul pendant une séquence de N_1 images consécutives :

$$\left\{OC_p^{k-i} = 0\right\}_{i=0,1,..,N_1-1} \ \& \ \left|I_p^k - B_p^k\right| < \tau_h \tag{3.16}$$

Il se peut qu'aucune de ces conditions ne soit jamais vérifiée. C'est le cas par exemple si l'objet en mouvement s'est immobilisé en maintenant l'occultation, ce qui peut se produire en cas d'embouteillage ou de stationnement du véhicule. Dans ce dernier cas, il semble justifié que le véhicule en stationnement soit intégré dans l'image de référence. On voit que de telles situations ne peuvent être résolues par des approches de bas niveau.

La table 3.1 résume ces deux conditions.

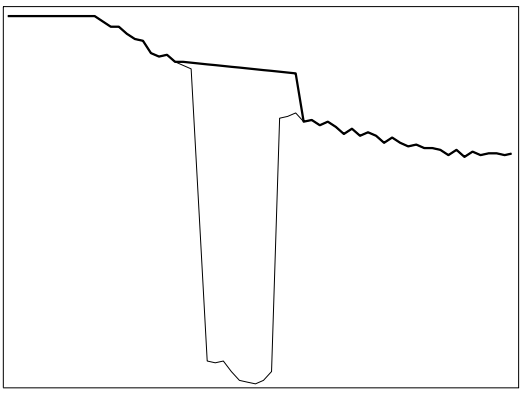


FIG. 3.14 – Estimation biaisée de la pente de la courbe qui amène une mauvaise estimation de B (en gras) à la fin de l'occultation.

Une fois que la mise à jour de la référence est assurée par rapport à l'évolution de l'état du pixel, on doit passer à l'étape détection de mouvement. Cette étape est analysée dans la section suivante et nous montre une nouvelle méthode fondée sur la collaboration de différentes composantes de l'image dans le but d'une meilleure robustesse de détection.

3.3 Détection des ombres portées

3.3.1 La problématique

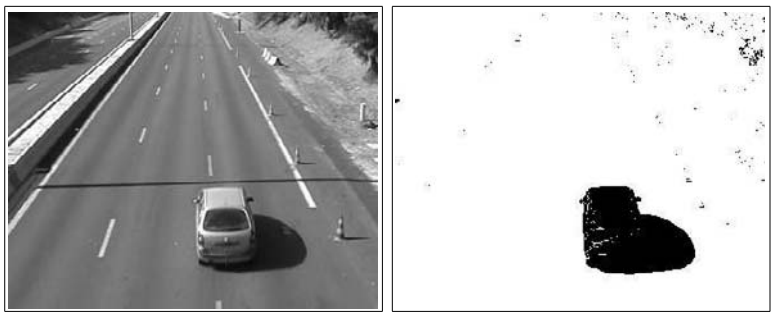
Une grande difficulté du traitement des images issues de scènes naturelles provient de l'existence de l'ombre portée qui est toujours associée à l'objet lorsqu'on utilise les approches de détection classiques[1] comme dans [KWM94], [VMBP96] ou [SG99] (Fig. 3.15).

Ce cas se présente surtout lorsque l'environnement est bien ensoleillé et que le soleil se rapproche de l'horizon.

Les ombres associées aux objets en mouvement se déplacent à la même vitesse que l'objet. Si l'ombre provient d'un objet fixe, elle affecte le fond de l'image (Fig. 3.16) et parfois l'ombre d'un objet en déplacement se projette sur un autre objet en déplacement dans le voisinage (Fig. 3.17).

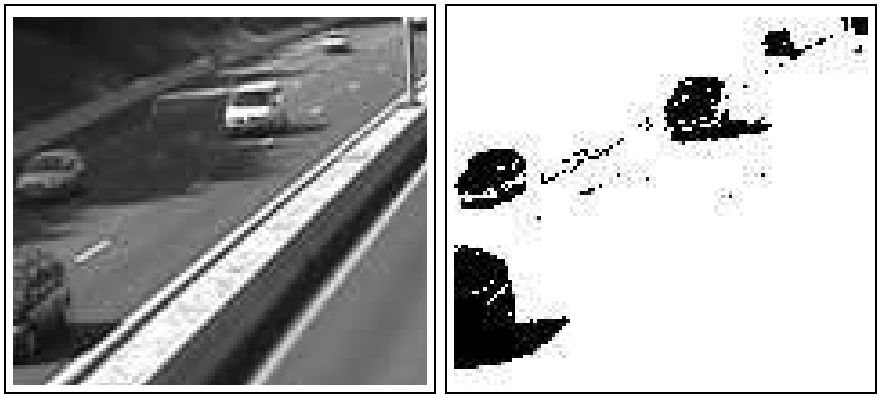
Si l'on n'est pas en mesure de distinguer l'objet de son ombre il s'ensuit des difficultés d'étiquetage de pixels qui rendent très difficile l'interprétation de l'image.

[1] Les approches classiques de détection de mouvement comparent l'image courante avec l'image de référence en terme de luminance.



(a) Image originale (b) Image de détection

FIG. 3.15 – Détection de mouvement selon l'approche classique.



(a) Image originale (b) Image de détection

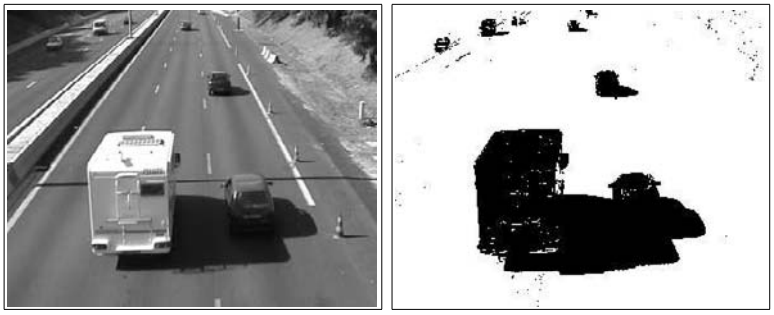
FIG. 3.16 – L'ombre fait varier la taille et l'apparence des objets selon les conditions d'illumination.

C'est pour résoudre ce problème que nous présentons dans ce qui suit un nouvel algorithme de détection des ombres fondé sur une analyse multi-composantes.

3.3.2 Approche Multi-Composantes pour la détection de l'ombre

Il est très difficile d'effectuer une détection correcte des ombres en utilisant seulement la luminance. Nous allons donc enrichir la donnée initiale en lui adjoignant la chrominance et la norme du gradient, aussi bien pour l'image en cours I que pour l'image de référence B.

Dans ce qui suit, nous désignons par I^Y la composante de luminance



(a) Image originale (b) Image de détection

FIG. 3.17 – L'ombre d'un objet peut compliquer la détection d'un objet voisin.

de l'image courante I et par B^Y la luminance de l'image de référence B. De même, les chrominances seront respectivement désignées par I^{C_1}, I^{C_2} et B^{C_1}, B^{C_2}. Enfin I^G et B^G désigneront les images respectives des modules du gradient calculés respectivement sur l'image luminance I^Y et B^Y.

La Fig. 3.18 compare les images RGB et les images YC_1C_2 de la même donnée. Les ensembles $\Omega_I = \left[I^Y, I^{C_1}, I^{C_2}, I^G\right]$ et $\Omega_B = \left[B^Y, B^{C_1}, B^{C_2}, B^G\right]$ constituent les données multi-composantes que l'on considère pour ce traitement.

Nous proposons un indicateur de disparité par composante $Z \in \Omega$ tel que :

$$MD^{\mathrm{Z}} = \begin{cases} 1 \text{ si } |I^{\mathrm{Z}} - B^{\mathrm{Z}}| > \tau_{\mathrm{Z}} \\ 0 \quad \text{ailleurs} \end{cases} \tag{3.17}$$

où Z est remplacé par la luminance, Y, la chrominance, C, et la norme des gradients, G (voir Fig. 3.19).

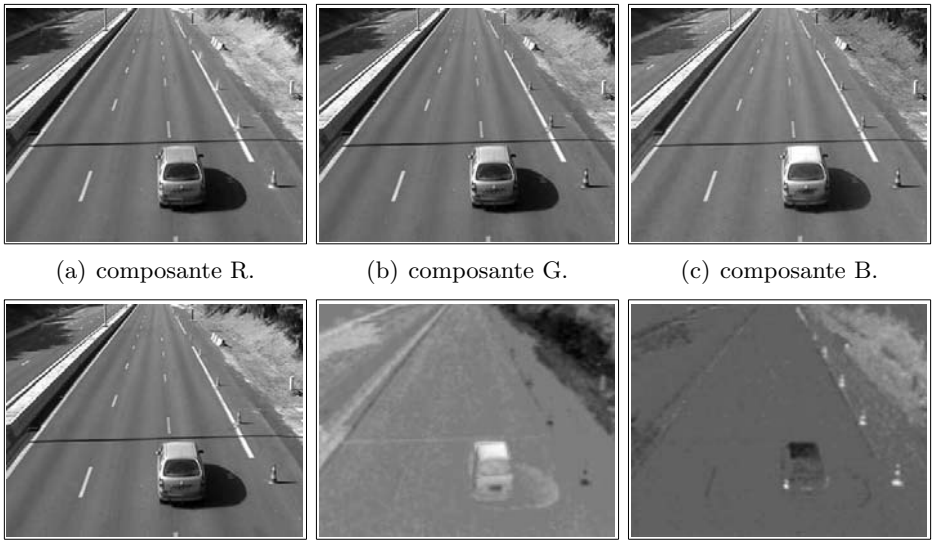
Avant de combiner différents opérateurs agissant sur chacune des composantes de Ω_I et Ω_B, nous allons étudier comment chacune des composantes est affectée par l'ombre.

Action de l'ombre sur les composantes

Luminance

Concernant la luminance Y, il est évident que la présence d'ombre conduit à une chute importante de celle-ci.

On caractérise cette chute par le rapport R_Y défini par :



(a) composante R. (b) composante G. (c) composante B.

(d) composante Y. (e) composante C_1. (f) composante C_2.

FIG. 3.18 – Décomposition de l'image en composantes RGB, luminance et chrominances.

$$R_Y = \frac{I^Y}{B^Y} \tag{3.18}$$

La valeur de R_Y fournit un opérateur de détection des ombres S_L :

$$S^L = \begin{cases} 1 \ si \ \delta_l < R_Y < \delta_h \\ 0 \quad ailleurs \end{cases} \tag{3.19}$$

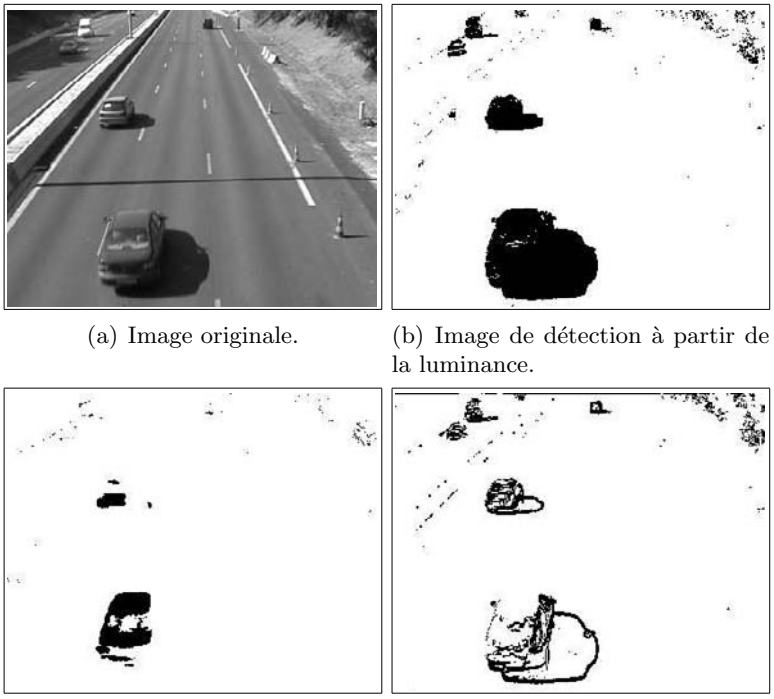
où δ_l et δ_h sont deux seuils préfixés avec $\delta_h > \delta_l$.

La Fig. 3.20 montre le résultat obtenu sur l'image test et révèle les limites de l'opérateur :

- Il existe des régions ombrées non détectées (ce qui signifie que δ_l est trop grand).
- Il existe des zones dans les objets en mouvement qui ont été étiquetées comme appartenant à l'ombre (ce qui signifie que δ_l est trop petit).

A partir de cet exemple on peut conclure que la luminance seule ne permet pas une détection robuste des ombres.

Chrominance



(a) Image originale.

(b) Image de détection à partir de la luminance.

(c) Image de détection à partir de la chrominance.

(d) Image de détection à partir de la norme du gradient.

FIG. 3.19 – Détection de mouvement multi-composante.

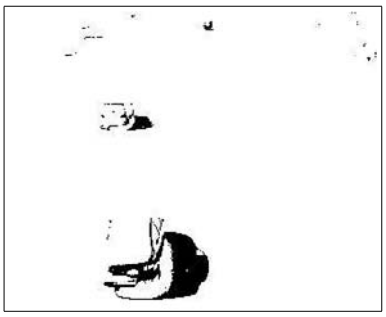


FIG. 3.20 – Caractérisation de la zone ombrée d'un objet.

L'ombre en général ne procure pas de changement de chrominance par rapport à celle de l'image de référence.

Un changement de chrominance résulte plutôt d'une occultation. Cette remarque nous permet de ré-étiqueter certains pixels classés par erreur dans

la catégorie ombre.

La Fig. 3.19(c) nous montre le résultat obtenu.

Gradient

La chute de la luminance provoque la chute des gradients à l'intérieur de la région ombrée et un très fort gradient sur la frontière de la régions ombrée, comme on peut le voir sur la Fig. 3.19(d).

Détection des ombres

Les différentes composantes de l'image vont dont être utilisées pour permettre la détection des objets en mouvement débarrassés de leur ombre.

La combinaison booléenne utilisée est la suivante :

$$M = \left(MD^L \cup MD^C \cup MD^G\right) \cap \bar{S^L} \tag{3.20}$$

où $M = 1$ si le pixel correspond à un objet en mouvement.

L'écart de chrominance global MD^C est extrait des écarts de chacune des composantes selon la relation :

$$MD^C = MD^{C_1} \cup MD^{C_2}$$

Sur l'image résultat (Fig. 3.21) on constate que la région ombrée a été bien supprimée sauf sur sa frontière où le gradient est important et induit donc une détection d'objet.

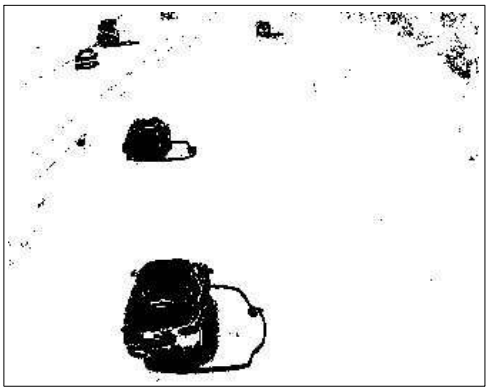


FIG. 3.21 – Masque de détection de mouvement sans ombre, M^k, extraite à partir de la combinaison des différentes composantes de l'image et de la caractérisation de l'ombre.

L'heuristique mise en oeuvre pour ré-étiqueter les frontières de la zone ombrée consiste à remarquer que ces limites ont une épaisseur relativement fine contrairement aux objets en mouvement. Ces limites se déplacent à la vitesse de l'objet. La durée de présence de cette limite a donc une faible durée dans la séquence d'images. La Fig. 3.22 montre l'évolution de M^k pour un pixel affecté par la limite de la zone ombrée.

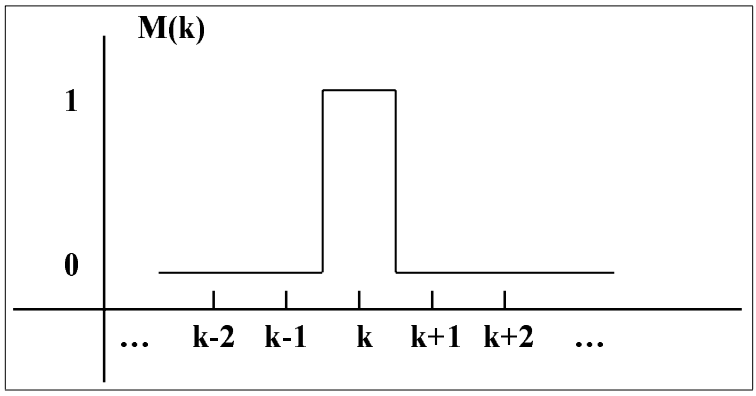


FIG. 3.22 – Gabarit temporel d'un pixel de contour d'une zone ombrée associée à un objet en mouvement.

Ce processus temporel ne permet pas toutefois d'exclure les limites de zones ombrées parallèles au déplacement car dans ce cas un pixel reste actif durant un laps de temps plus important. Cependant, ce type de région sera rejeté par une gestion de plus haut niveau.

La Fig. 3.23 montre le résultat de l'opérateur composite que nous venons de présenter.

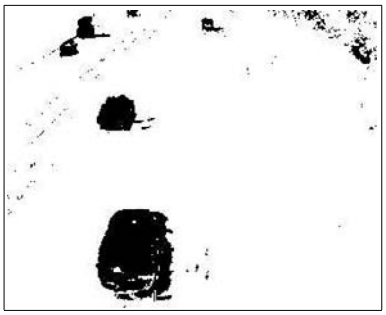


FIG. 3.23 – Masque final de détection d'objets en mouvement sans ombre.

La Fig. 3.24 présente la structure complète des processus mis en oeuvre

pour la mise à jour des images de référence (ARI) et de détection des objets en mouvement débarrassés de leur ombres.

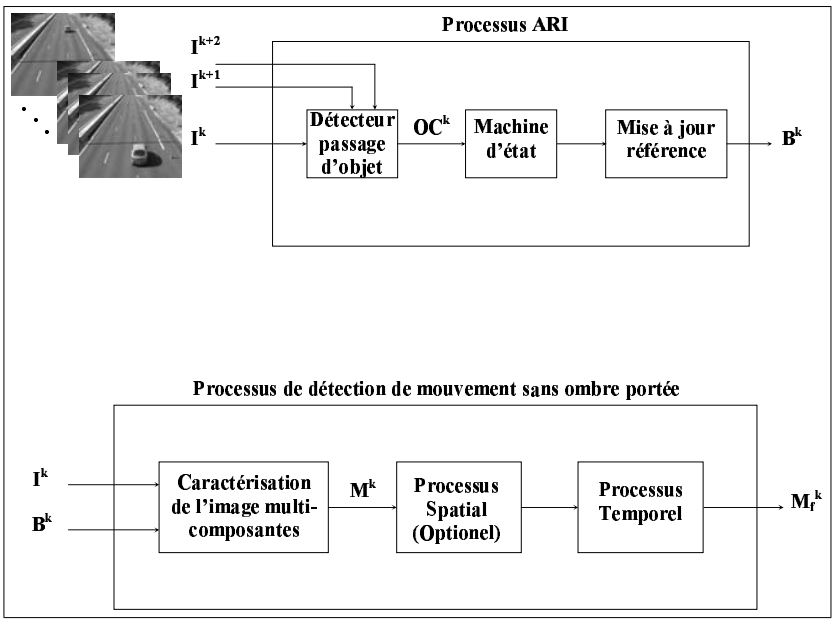


FIG. 3.24 – Diagramme complet du processus ARI et de la détection d'objets en mouvement sans ombre.

3.4 Résultats comparatifs

Si on compare nos résultats à ceux des méthodes proposées dans la "littérature", on peut observer que notre méthode est plus robuste face aux changements de luminosité globales, ce qui la rend plus adaptée aux séquences d'extérieur. Les Fig. 3.25 et 3.26 illustrent des résultats comparatifs de notre méthode avec deux méthodes présentées dans la "littérature" : l'approche de Koller [KWM94] et celle de Stauffer [SG99]. Trois conclusions peuvent être dégagées :

1. Notre approche montre que les fantômes résiduels sont en nombre réduit.
2. On obtient des frontières d'objet plus nettes.
3. Les régions en faible mouvement, telles que du feuillage agité par le vent, sont détectées de façon similaire dans les 3 méthodes, avec cependant une légère amélioration par rapport à la méthode de Koller.

Le cas de scènes sensibles à la présence d'ombres est présenté en Fig. 3.25. On constate que l'approche multi-composantes supprime les régions ombrées et améliore la définition de l'objet.

3.5 Conclusions

Ce chapitre a été l'occasion d'aborder la détection de mouvement, étape clef du procédé globale de suivi. Comme nous l'avons vu dans le premier chapitre, l'utilisation d'une image de référence présente de nombreux avantages tels que l'absence de zones incertaines (recouvrement, chevauchement, ...), une bonne restitution des contours de l'objet et la possibilité d'introduire un schéma récursif permettant de filtrer le bruit. Dans ce contexte, nous avons proposé une nouvelle méthode utilisant de façon concomitante l'intensité lumineuse de l'image courante et de l'image de référence. Fondée sur une analyse de la variationnelle de l'intensité lumineuse, nous avons pu réaliser une mise à jour de l'image de référence par rapport aux conditions de luminosité globales du fond de l'image. Nous avons noté que ce procédé était d'autant plus pertinent qu'une interaction avec un procédé plus haut niveau existé. Certaines informations délivrées par l'étape de reconnaissance sont en effet nécessaires afin de lever les ambiguïtés générées par la présence d'occultation et de la méconnaissance de fin de passage de l'objet. Ces informations sont délivrées par les descripteurs prédits ce qui permet de rendre plus robuste la détection.

Nous avons tous les éléments concernant notre architecture et les formats de grandeurs manipulées par les différents modules. Nous pouvons maintenant étudier la mise en application de ces techniques. Dans le chapitre suivant, nous allons focaliser notre étude sur deux applications : la première est dédiée au trafic routier et la seconde au suivi de visages.

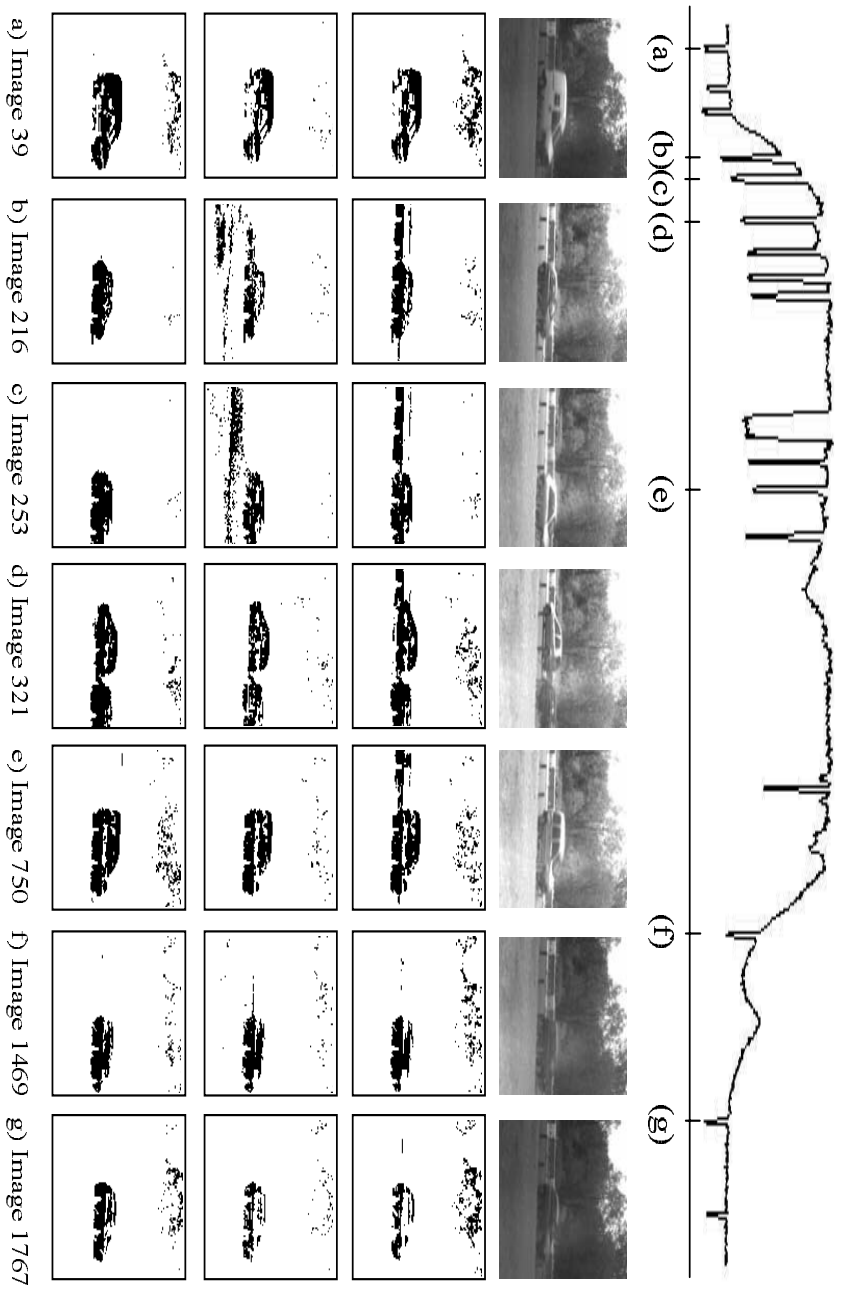


FIG. 3.25 – Résultats comparatifs pour les méthodes de Koller et al. (deuxième ligne), Grimson et Stauffer (troisième ligne) et notre approche (dernière ligne). La première ligne nous montre le courbe d'intensité de luminosité d'un pixel p situé au milieu de l'image. Les lettres nous indiquent les instants auxquels on a repéré les images.

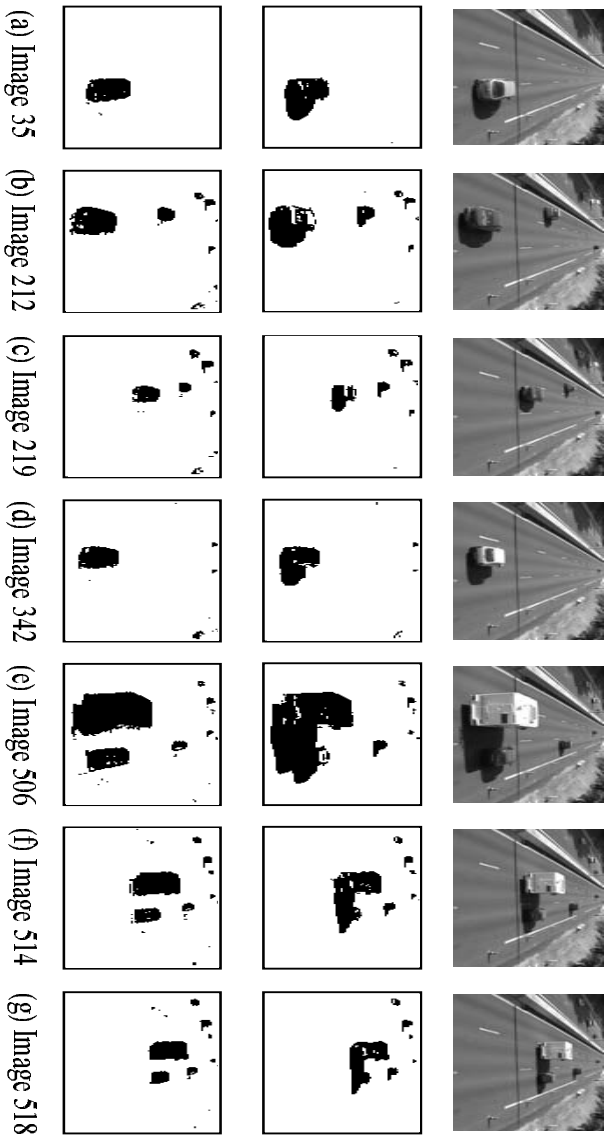


FIG. 3.26 – Résultats comparatifs entre l'approche classique (comparaison de l'image courante et de référence en terme de luminance) et notre approche multi-composantes avec suppression d'ombre.

Chapitre 4

Applications

4.1 Application 1 : suivi de véhicules pour la gestion du trafic routier

4.1.1 Introduction

Toutes les grandes agglomérations s'équipent aujourd'hui d'infrastructures dédiées à la gestion du trafic routier. Cependant, le déploiement des systèmes d'information dédiés à la gestion du trafic reste (malheureusement) contraint par une forte implication de l'opérateur humain. Dans un futur proche, afin de rendre plus performante et moins contraignante la supervision du trafic, il est primordial de disposer de systèmes capables d'interpréter automatiquement l'état du trafic, de prévenir en cas d'incident ou d'accident. Pour atteindre cette qualité de service, les sous-systèmes de traitement de l'information doivent extraire automatiquement le contenu pertinent pour ne faire remonter vers l'opérateur qu'une information parcellaire mais prioritaire (accidents) ou synthétique mais globale (saturation, pollution).

Parmi toutes les sources d'information disponibles pour alimenter ce système de gestion, la télésurveillance possède beaucoup d'atouts. La caméra est en effet le capteur délivrant le plus d'informations aussi bien sur le plan quantitatif que qualitatif. D'autres sources comme les technologies radar ou les boucles magnétiques sont intéressantes mais relativement pauvres sur le plan de la couverture spatiale de mesure et sur la richesse du signal délivré. Cependant, la pertinence de ce média est sans nulle doute aussi son point faible car l'analyse d'un contenu visuel pour des scènes extérieures même pour une caméra statique n'est pas aujourd'hui une problématique totalement résolue.

Dans ce mémoire, il ne s'agit pas de répondre à la problématique complète du suivi de véhicules quelles que soient les conditions météorologiques (brouillard, pluie, neige). Nous désirons surtout évaluer la pertinence de notre approche. La robustesse de la réponse aux conditions météorologiques est à rechercher dans une adaptation au cas par cas des différentes fonctionnalités qui implementent ce procédé. Nous répondons pour l'instant à un contexte d'exposition diurne avec une météo relativement clémente en acceptant les passages nuageux. Conformément au schéma 2.2 du chapitre 2 nous devons choisir une méthode de synthèse des entrées perceptives (détection), adapter le jeu de descripteurs, conditionner la méthode de prédiction et finaliser la forme paramétrique associée au modèle probabiliste pour la procédure EM.

4.1.2 Détection de véhicules

Dans l'étape de détection, nous exploitions la méthode proposée dans le chapitre 3, qui s'appuie sur la perception du mouvement pour focaliser le procédé de suivi. Dans cette application on peut noter qu'il est intéressant de raccorder la détection de mouvements à la boucle de rétroaction après estimation du mouvement de chaque véhicule. En effet, nous pouvons conditionner le masque de mouvement M^k qui contrôle la mise à jour de l'image de référence. L'intérêt est de pouvoir bloquer la mise à jour lorsqu'un véhicule a un déplacement nul, il ne faut pas qu'il soit intégré dans l'image de référence.

4.1.3 Descripteurs dédiés au suivi de véhicules

Dans le chapitre 2, nous avions présenté un jeu de descripteurs par défaut. Nous avions introduit *"la zone"* comme descripteur spatial de l'objet. La zone permet la manipulation du contour, de la taille, du centre de gravité de l'objet et de la distribution en amplitude de l'ensemble des pixels contenus dans l'objet.

Dans cette première application de suivi des véhicules, cette zone sera définie à partir du contour convexe associé à la détection de l'objet (voir Fig. 4.1).

En ce qui concerne le modèle de mouvement, nous conservons un modèle affine à six paramètres :

$$\begin{cases} d_x = a_1 + a_2(x - x_g) + a_3(y - y_g) \\ d_y = a_4 + a_5(x - x_g) + a_6(y - y_g) \end{cases} \tag{4.1}$$

Soit $\Theta = [a_1\ a_2\ a_3\ a_4\ a_5\ a_6]^T$ le vecteur regroupant tous les paramètres

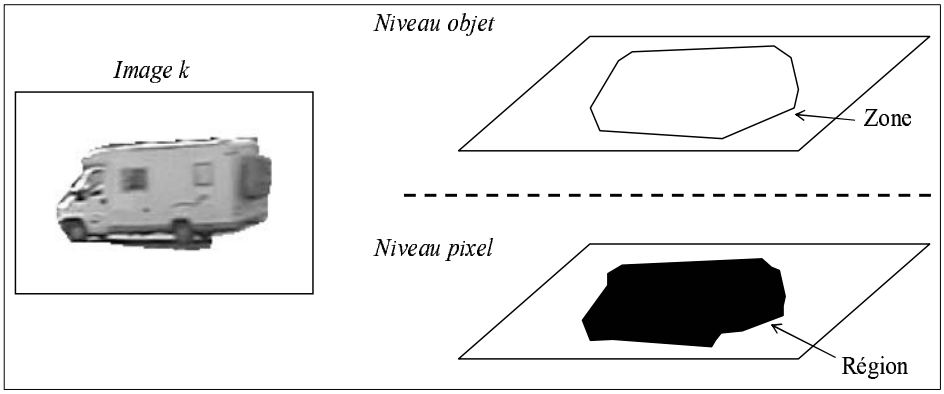


FIG. 4.1 – Définition de la zone représentant l'objet à partir de son contour convexe.

du modèle de mouvement. Ce modèle nous permet d'effectuer le suivi d'un objet évoluant selon un mouvement de translation et de rotation avec facteur d'échelle.

4.1.4 Mise à jour du modèle à partir des points caractéristiques

Dans la suite de ce paragraphe, nous développons la méthode d'estimation du mouvement utilisant les points caractéristiques. Deux phases sont nécessaires pour la mise en oeuvre de cette méthode : la phase d'extraction des points et la phase d'appariement.

Extraction des points caractéristiques

Le point caractéristique est le représentant d'une zone spatiale de forte activité en intensité. Cette activité est quantifiée à l'aide des gradients. L'amplitude n'est pas la seule information utilisée. Afin de ne pas tomber "dans le piège" de la texture mono-orientée qui engendre une indétermination quant à l'estimation du mouvement, nous recherchons les zones présentant des points de convergence de structures multi-orientées. Pour de raisons de simplicité, cette recherche est ramenée à la recherche de la plus forte courbure sur la base d'une matrice de covariance des gradients. Ce principe a déjà été développé dans la "littérature" : deux références majeures, Harris [HS98] et Kanade [TK91] proposent deux schémas de décision fondés sur la distribution des valeurs propres de la matrice de covariance des gradients.

La forme générique de la matrice de covariance est donnée par la relation suivante :

$$M = G(\tilde{\sigma}) \otimes M_{\nabla} \tag{4.2}$$

Le terme $G(\tilde{\sigma})$ représente un masque gaussien de convolution qui permet la gestion du voisinage. La matrice M_{∇} est construite à partir des gradients spatiaux. Ces gradients correspondent aux dérivées partielles d'un simple canal de luminance ou de façon plus complète des canaux de couleurs. Nous avons par exemple :

$$M_{\nabla} = \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \tag{4.3}$$

pour les images en niveau de gris, où I_x, et I_y sont les gradients spatiaux respectivement dans la direction x et y et pour les images en couleur :

$$M_{\nabla} = \begin{bmatrix} R_x^2 + G_x^2 + B_x^2 & R_x R_y + G_x G_y + B_x B_y \\ R_x R_y + G_x G_y + B_x B_y & R_y^2 + G_y^2 + B_y^2 \end{bmatrix} \tag{4.4}$$

où R_x, R_y sont les gradients spatiaux pour le canal R (rouge) respectivement dans la direction x et y.

A la suite de l'extraction des valeurs propres de M, nous obtenons les conditions d'extraction des points caractéristiques en utilisant les méthodes proposées par les deux auteurs présentés dans la table 4.1.

Tab. 4.1 – Conditions d'extraction des points caractéristiques

Kanade	$max(\sigma_1, \sigma_2)$
Harris	$\prod_{i=1}^{2} \sigma_i - k \cdot \left(\sum_{i=1}^{2} \sigma_i \right)^2$

La condition proposée par Harris utilise un paramètre k que l'on peut ajuster pour modifier la sensibilité [1]. La détection va consister à rechercher les maxima locaux de l'opérateur. Les Fig. 4.2 et 4.3 montrent des résultats obtenus avec les deux opérateurs pour des images de luminance ou en couleur. Les résultats sont très similaires. Par contre, il est à noter que le résultat de la détection, sans traitement approprié se caractérise par une concentration de points dans les zones fortement texturées. Afin d'assurer une bonne couverture spatiale de l'objet qui permet une estimation

[1] Les auteurs conseillent la valuer $k = 0.04$ comme valeur optimale.

correcte des paramètres de mouvement, nous devons rajouter une gestion hiérarchique avec zones d'exclusion lors de l'extraction des maxima locaux. La Fig. 4.4 nous montre un résultat en considérant une zone d'exclusion fondée sur un compromis entre nombre de points désirés et la taille de l'objet.

(a) Détecteur de Harris en Noir et Blanc. (b) Détecteur de Kanade en Noir et Balnc.

FIG. 4.2 – Extraction des points caractéristiques pour une image en noir et blanc avec les opérateurs de Harris et Kanade.

(a) Détecteur de Harris en couleur. (b) Détecteur de Kanade en couleur.

FIG. 4.3 – Extraction des points caractéristiques pour une image en couleur avec les opérateurs de Harris et Kanade.

En réalisant l'extraction sur deux images consécutives, il nous reste à établir l'appariement pour estimer le modèle du mouvement de l'objet.

(a) Détecteur de Harris en couleur. (b) Détecteur de Kanade en couleur.

FIG. 4.4 – Extraction des points caractéristiques pour une image en couleur avec les opérateurs de Harris et Kanade avec répartition des points autour de l'objet.

Identification des points : Block Matching

Une fois que nous disposons des points caractéristiques positionnés sur l'objet aux instants t et $t + T_e$, nous devons établir l'appariement entre ces deux familles de points. Un exemple est présenté à la Fig. 4.5.

(a) Image à l'instant t. (b) Image à l'instant $t + 1$.

FIG. 4.5 – Identification d'un point caractéristique sur un objet en mouvement. La mise en correspondance nous permet d'estimer le modèle de mouvement de l'objet .

L'identification des couples de points passe par une phase de recherche et de comparaison. Pour cette phase, nous avons choisi un algorithme de

type correspondance des blocs dit *"Block Matching"*.

L'idée fondamentale du *"block matching"* est résumée dans la Fig. 4.6. La mise en correspondance d'un certain pixel p en mouvement entre l'instant k et $k+1$ est établie à partir d'une comparaison s'appuyant sur un bloc de taille $N_1 \times N_2$ centré sur le pixel p. Un mode de parcours est établi afin de rechercher où se trouve le correspondant du pixel p sur la base du voisinage. La recherche du bloc ressemblant s'arrête par minimisation d'un critère de similarité. La zone de recherche est souvent limitée à une région de taille $(N_1 + 2M_1) \times (N_2 + 2M_2)$.

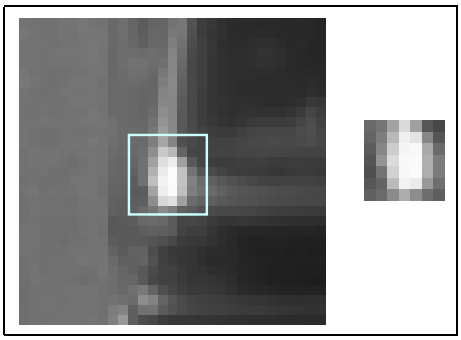


FIG. 4.6 – Présentation de la procédure de "block matching" : Bloc à chercher (à droite) de taille $N \times N$ et fenêtre de recherche (à gauche) de taille $(N+M) \times (N+M)$.

Les différents algorithmes de block matching peuvent se décliner selon les trois options suivantes :

1. Le critère de ressemblance ou matching criteria en anglais (mean square error, mean absolute difference, etc.).
2. La stratégie de recherche (Full search, three step search, etc.).
3. La gestion de la taille du bloc (hiérarchique, adaptative).

1.- Les critères de ressemblance

L'identification des blocs est effectuée par minimisation d'un critère entre le bloc de référence (à l'image t) et le bloc recherché (à l'image $t + T_e$). Ce critère peut prendre différentes formes. Nous trouvons dans la "littérature" la moyenne des erreurs quadratiques (MSE en anglais), la moyenne des erreurs absolues (MAD en anglais), le coefficient d'inter corrélation (CCF) ou sa version normalisée (NCCF). La mise en équation de ces critères est la suivante :

$$
\begin{aligned}
CCF(d_1,d_2) &= \sum_{(x,y)\in \mathrm{B}} \left(I^k(x,y) \cdot I^{k+1}(x+d_1, y+d_2) \right) \\
NCCF(d_1,d_2) &= \sum_{(x,y)\in \mathrm{B}} \left(\left(I^k(x,y) - \bar{I}^k \right) \cdot \left(I^{k+1}(x+d_1, y+d_2) - \bar{I}^{k+1} \right) \right) \\
MSE(d_1,d_2) &= \frac{1}{N_1 N_2} \sum_{(x,y)\in \mathrm{B}} \left[I^k(x,y) - I^{k+1}(x+d_1, y+d_2) \right]^2 \\
MAD(d_1,d_2) &= \frac{1}{N_1 N_2} \sum_{(x,y)\in \mathrm{B}} \left| I^k(x,y) - I^{k+1}(x+d_1, y+d_2) \right| \qquad (4.5)
\end{aligned}
$$

où I^k et I^{k+1} sont les niveaux d'intensité des images aux instants consécutifs k et $k+1$. Les deux premiers critères sont à maximiser, tandis que les deux autres sont à minimiser.

2.-Les stratégies de recherche

Pour établir quel est le meilleur bloc qui correspond au bloc original, il faut comparer les résultats des critères de ressemblance sur la région de recherche. Comme nous l'avons vu plus haut, la région de recherche de taille $(N_1 + 2M_1) \times (N_2 + 2M_2)$ est la région dans laquelle on va devoir identifier notre bloc. Pour chercher ce bloc, il nous faut parcourir la région selon une certaine stratégie. La recherche exhaustive ou aveugle va parcourir toute la région de recherche, en déplaçant le bloc recherché pour toutes les positions possibles de la région (voir Fig. 4.7). Bien évidemment, cette stratégie est très simple mais elle est coûteuse au niveau calculatoire. Par contre, si la taille de la région de recherche est suffisamment grande pour être sûr que le bloc à chercher y est contenu, la recherche exhaustive est optimale. Les stratégies sous optimales, présentées par la suite, sont beaucoup moins coûteuses comme le montre le tableau 4.2.

TAB. 4.2 – Comparaison des différentes méthodes de recherche du bloc pour une fenêtre de recherche de $(N+M)^2$, où N est la taille du bloc à chercher.

Type de recherche	Max. positions à tester	$w = 4$	$w = 16$
Full search	$(2M+1)^2$	81	1089
Three step	$1 + 8 \cdot log_2(M)$	17	33

Nous trouvons dans le tableau la recherche *Three step search* (recherche en trois étapes) proposé par Koga et al. [KIH+81]. Cette méthode va tester 8

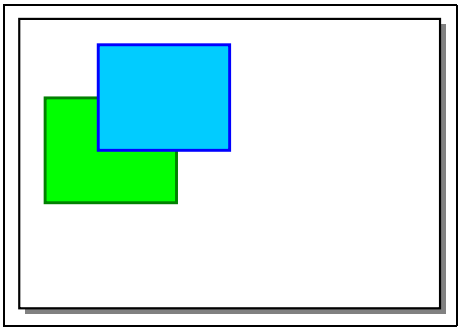


FIG. 4.7 – Méthode de recherche exhaustive.

positions autour du centre avec un rayon donné (voir Fig. 4.8). La meilleure option sert de point de départ pour une nouvelle recherche, mais cette fois-ci, le pas d'exploration est réduit. Une troisième itération met un point final à la recherche du bloc. Le pas initial couramment utilisé est de 3 pixels, décrémenté de 1 pixel à chaque itération, ce qui donne 3 étapes, d'où le nom de la méthode.

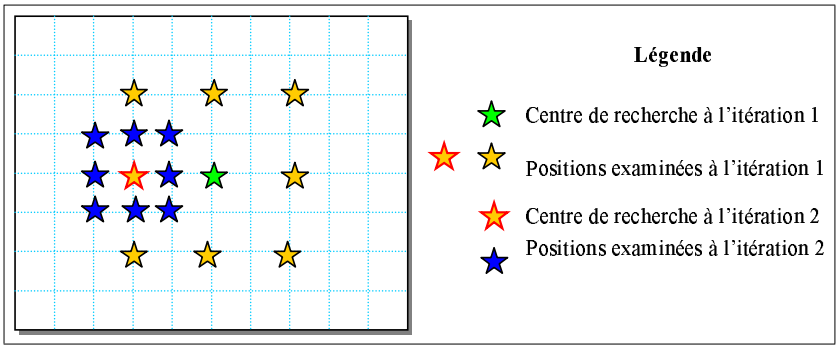


FIG. 4.8 – Méthode de recherche *Three step search*.

D'autres méthodes, plus compliquées peuvent aussi être examinées, comme la méthode de *recherche en diamant* [TSL+00]. Les auteurs proposent d'examiner 9 positions disposées en diamant autour d'un point central. A nouveau, le point optimal devient le centre d'une nouvelle recherche en diamant. On continue la recherche tant que le point optimal ne se situe pas au centre du diamant. A ce point, on fait une toute dernière recherche mais avec un diamant de dimension inférieure (voir Fig. 4.9).

Dans l'objectif de "bien" identifier tous les points de l'objet en mouvement, nous avons décidé d'utiliser la technique de recherche exhaustive avec

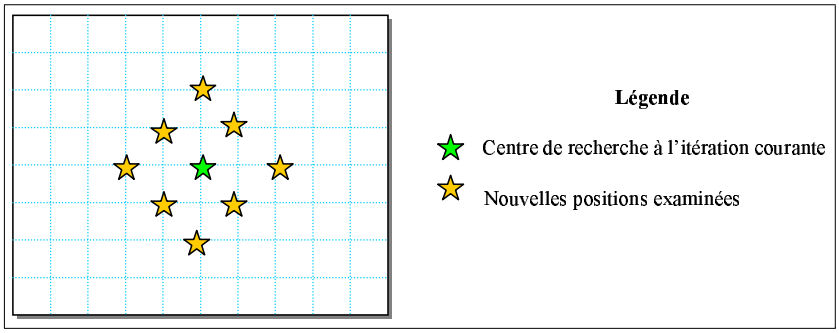


FIG. 4.9 – Méthode de recherche en diamant.

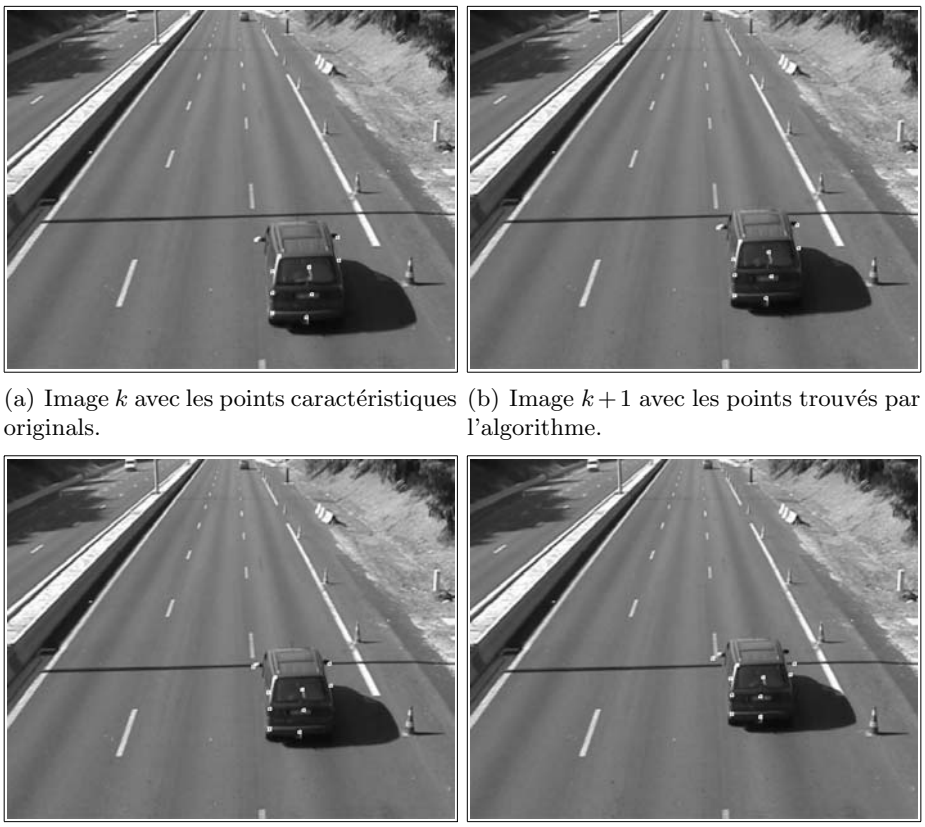
la mesure de coefficient de corrélation normalisé (NCCF). Ainsi et même si on perd plus de temps dans la phase de recherche, on sera sûr de bien récupérer le point caractéristique à l'image suivante. Les résultats d'identification des différents points peuvent être visualisés dans la Fig. 4.10.

4.1.5 Prédiction du modèle de mouvement

La prédiction du modèle de mouvement permet l'adaptation du déplacement de l'objet aux variations. Dans le cas particulier de l'application dédiée au suivi des véhicules, la prédiction de mouvement permet aussi la réduction de la taille de la région de recherche. Cette réduction s'explique par le fait que la prédiction nous donne un a priori sur la position du bloc à chercher.

Pour démontrer que le modèle AR s'ajuste parfaitement à la modélisation du déplacement d'un objet en mouvement, notamment dans cette application, pour le déplacement d'une voiture dans la scène, on va comparer les résultats du déplacement réel d'une voiture avec celui obtenu par prédiction en utilisant le modèle. Pour effectuer cette comparaison, il suffit de sélectionner la valeur des coefficients α_{ij} pour chaque paramètre du modèle de mouvement. Une fois qu'on les a définis, on traite l'évolution de tous les paramètres et on les compare avec les paramètres réels. On peut constater dans la Fig. 4.11 que le résultat est équivalent au cas réel. Grâce à cette simulation, on constate que la modélisation des paramètres de mouvement s'adapte aux besoins du déplacement réel des voitures. La Fig. 4.12 nous montre la reproduction du déplacement effectué par un objet en utilisant la modélisation AR. La reproduction s'ajuste parfaitement au déplacement réel effectué par la voiture.

La Fig. 4.13 nous montre l'évolution du paramètre a_i ainsi que l'évolution



(a) Image k avec les points caractéristiques originals.

(b) Image $k+1$ avec les points trouvés par l'algorithme.

(c) Image $k+2$ avec les points trouvés par l'algorithme.

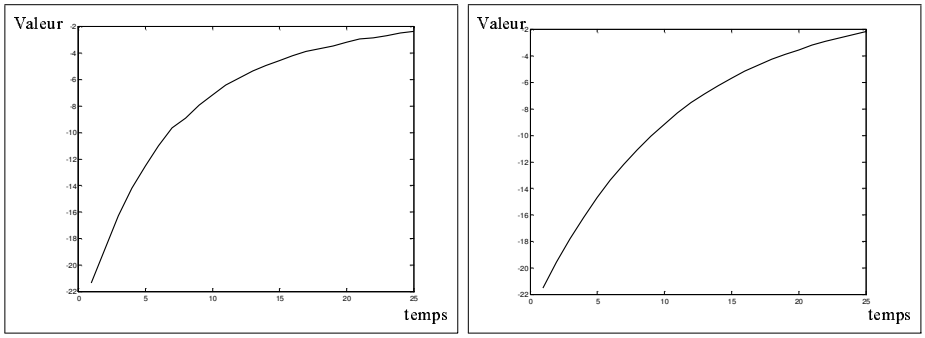
(d) Image $k+3$ avec les points trouvés par l'algorithme.

FIG. 4.10 – Identification des différents points caractéristiques sur un objet en mouvement.

des coefficients α_{ij} associés. On peut constater que ces coefficients évoluent au cours du temps pour s'ajuster aux variations de la courbe d'évolution du paramètre a_i.

La Fig. 4.14 montre que le choix d'un ordre p plus élevé ne rajoute aucune information supplémentaire puisque les coefficients α_{ij} sont quasiment nuls à partir de $j = 2, \cdots, p-1$ dans l'exemple proposé.

Nous pouvons remarquer que dans les séquences correspondant à l'observation de voies rapides, les paramètres a_i significatifs sont surtout ceux des translations et du zoom. Les Fig. 4.15 et 4.16 montrent différents résultats de prédiction.



(a) Exemple d'évolution réel des paramètres de mouvement.

(b) Exemple d'évolution estimée des paramètres à travers d'un modèle AR.

FIG. 4.11 – Comparaison de l'évolution des paramètres de mouvement .

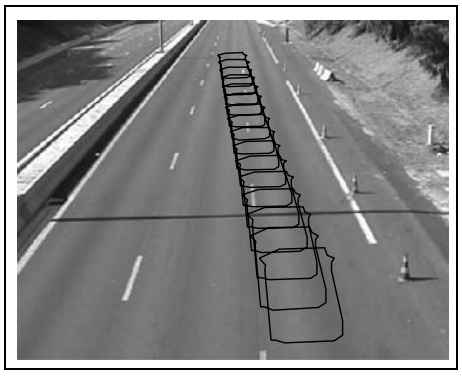
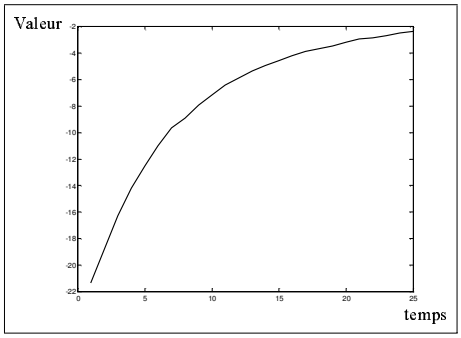


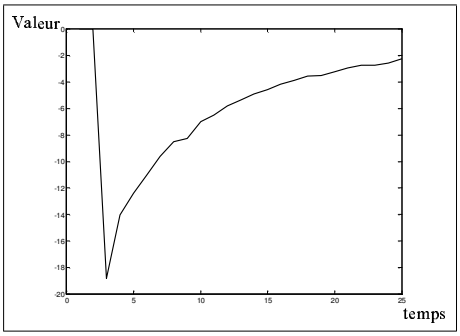
FIG. 4.12 – Reproduction du déplacement d'une voiture virtuelle à travers l'estimation du modèle de mouvement avec la modélisation AR.

4.1.6 La mise en correspondance

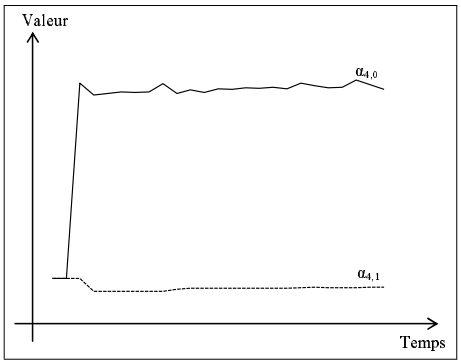
Le procédé de mise en correspondance permet l'établissement des liaisons spatio-temporelles entre la prédiction des zones et les régions issues de la détection de mouvement. Le processus de mise en correspondance nous donne ainsi les appariements [régions—zones prédites]. Les Fig. 4.17, 4.18 et 4.19 nous montrent les appariements des régions et des zones prédites dans le cas non ambigu (une région et une zone prédite), le cas de sur-segmentation et le cas d'occultation respectivement. Le traitement et la gestion de ces appariements sont réalisés par le processus EM, capable de gérer les cas ambigus et de retrouver les nouvelles zones qui définissent spatialement l'objet.



(a) Evolution temporelle réel du paramètre de mouvement a_4.



(b) Evolution temporelle de la prédiction du paramètre a_4.



(c) Evolution temporelle des coefficients α_{ij}

FIG. 4.13 – Exemple d'évolution des coefficients α_{ij} dans le cas du paramètre a_4 avec un ordre $p = 2$.

4.1.7 L'identification : Algorithme EM

Comme on l'a déjà introduit dans le chapitre 2, le but du procédé d'identification est de lever les ambiguïtés engendrées par les cas de sur-segmentation et d'occultation.

L'algorithme EM est initialisé par les couples [régions—zones prédites] extraits du processus de mise en correspondance. Comme résultat final, on obtient les nouvelles zones associées à chaque objet.

Dans cette application de suivi de véhicules, deux simplifications peuvent être faites :

· Les variances de chaque couche, σ_j, sont fixées a priori par l'utilisateur. Elles sont toutes mises à la même valeur afin d'en simplifier la gestion

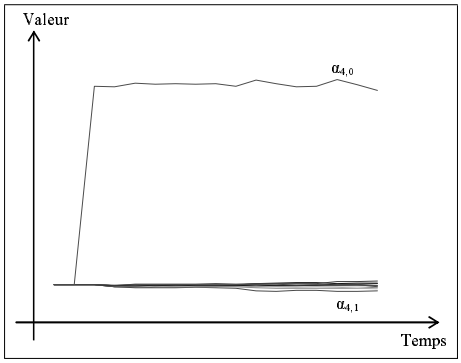


FIG. 4.14 – Prédiction des paramètres avec un ordre p élevé. On constate comme les coefficients α_{4j} sont tous très proches de la valeur zéro à partir de $j = 2$.

en supposant que le bruit est le même sur toute l'image.

· La probabilité d'existence d'un objet par rapport à l'ensemble est supposée être équiprobable. Aucun modèle n'est a priori prioritaire ou prépondérant par rapport à un autre. Nous posons : $\pi_j = 1/m'$.

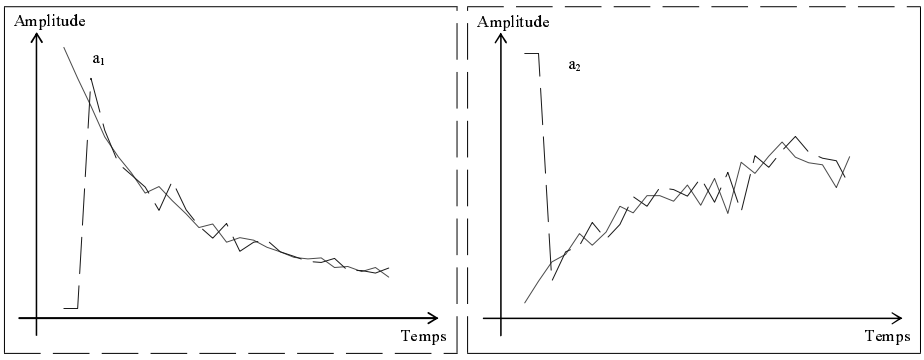
Si on injecte toutes ces simplifications dans l'équation (2.13) du chapitre 2, alors on obtient la nouvelle expression :

$$\omega_{pj} = \frac{\pi_j \cdot prob\left(I^k(p) \left| \tilde{DO}_j^k \right.\right)}{\sum_{l=1}^{m'} \pi_l \cdot prob\left(I^k(p) \left| \tilde{DO}_l^k \right.\right)} = \frac{\exp\left(-\frac{r_j^2(p)}{2\sigma_j^2}\right)}{\sum_{l=1}^{m'} \exp\left(-\frac{r_l^2(p)}{2\sigma_l^2}\right)} \tag{4.6}$$

Nous présentons dans les Fig. 4.20 et 4.21, les résultats obtenus par le processus d'identification dans différents cas de sur-segmentation et d'occultation.

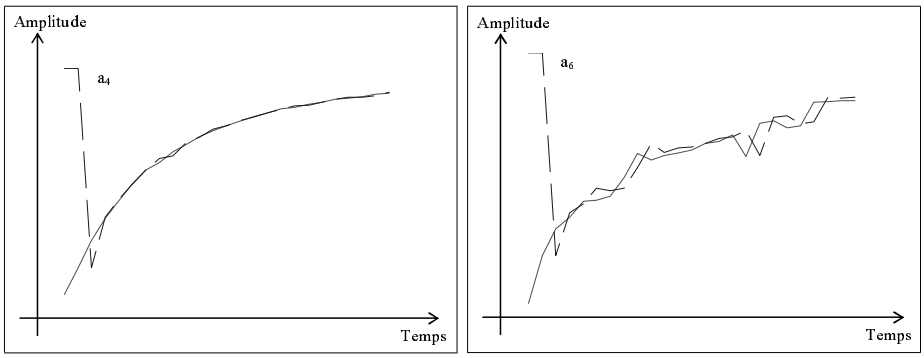
4.1.8 Conclusion

Dans cette application dédiée au suivi de véhicules, nous avons présenté l'ensemble des outils constituant la chaîne de traitement. Ainsi, l'étape de détection joue pleinement son rôle de pré-segmentation. Grâce à la prédiction, la phase de reconnaissance prend les "bonnes" décisions en utilisant efficacement les cartes d'appartenance. Nous avons montré, en effet, de nombreux exemples réels validant les résultats en cas d'occultation ou de



(a) Evolution de la courbe réelle (ligne continue) et de prédiction (pointillé) pour le paramètre a_1.

(b) Evolution de la courbe réelle (ligne continue) et de prédiction (pointillé) pour le paramètre a_2.



(c) Evolution de la courbe réelle (ligne continue) et de prédiction (pointillé) pour le paramètre a_4.

(d) Evolution de la courbe réelle (ligne continue) et de prédiction (pointillé) pour le paramètre a_6.

FIG. 4.15 – Courbes réelles et de prédiction pour des paramètres a_i.

conditions d'éclairage changeantes. L'étude de cette application a été l'occasion de proposer un schéma d'estimation du mouvement utilisant les points caractéristiques, la prédiction par le filtre de Kalman et le Block Matching. Grâce à la prédiction, nous obtenons une aide précieuse quant à la mise en correspondance des blocs.

Cette illustration de notre procédé de suivi nous a permis de valider l'interaction entre la détection et la reconnaissance. Nous avons montré comment en cas d'occultation ou d'arrêt d'un véhicule, l'utilisation des descripteurs prédits évite d'intégrer les pixels ambigus dans la référence.

Afin de valider le caractère générique de notre approche, nous allons dans la suite de ce chapitre étudier une nouvelle application, à savoir, le

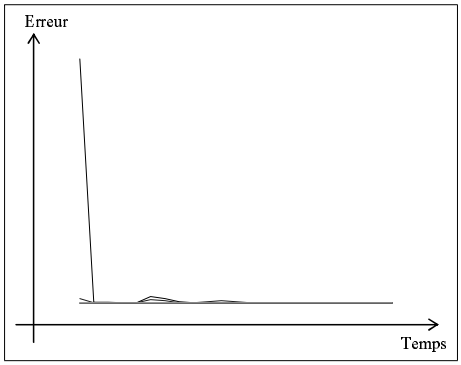


FIG. 4.16 – Erreurs de prédiction pour chaque composante a_i de mouvement.

(a) Image réelle. (b) Image de détection de mouvement. (c) Appariement des régions et zones prédites.

FIG. 4.17 – Résultat du processus de mise en correspondance pour un cas non ambigu : [1 région—1 zone prédite].

suivi de visage. Pour les véhicules, la génération des régions perceptives était associée à la caractéristique du mouvement et les mouvements étaient fortement contraints par les évolutions possibles dues aux infrastructures routières. Par la seconde application, nous avons voulu enrichir notre expertise en nous intéressant à une application utilisant une entrée perceptive complémentaire et une possible trajectoire plus difficile à prédire.

4.2 Application 2 : suivi de visages

4.2.1 Introduction

La détection et le suivi de personnes sont utilisés dans de nombreuses applications. La vidéo-surveillance, l'interprétation des gestes humains, l'iden-

(a) Image réelle. (b) Image de détection de mouvement. (c) Appariement des régions et zones prédites.

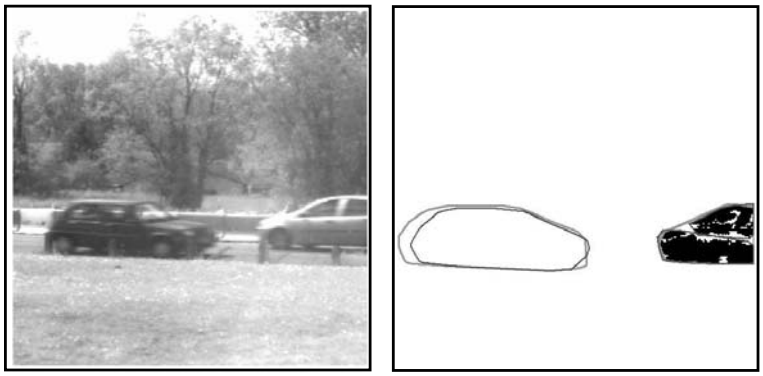
FIG. 4.18 – Résultat du processus de mise en correspondance pour un cas de sur-segmentation : [2 régions—1 zone prédite].

(a) Image réelle. (b) Image de détection de mouvement. (c) Appariement des régions et zones prédites.

FIG. 4.19 – Résultat du processus de mise en correspondance pour un cas d'occultation : [1 région—2 zones prédites].

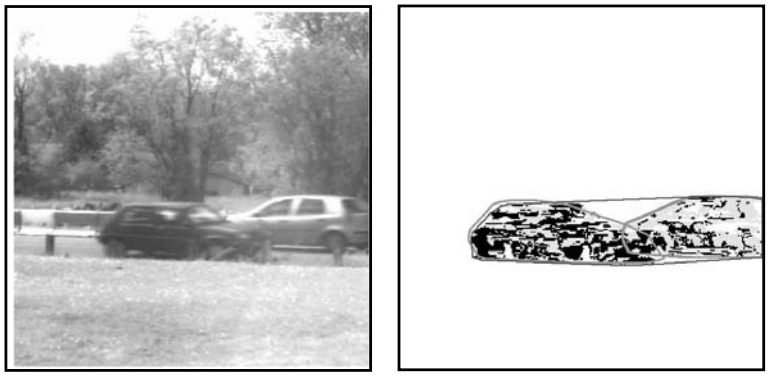
tification de personnes ou les systèmes d'aide au positionnement de personnes handicapées sont quelques exemples d'applications qui nécessitent la détection et le suivi du mouvement humain.

Dans cette section nous introduisons une méthode de suivi humain à partir du visage. L'analyse du mouvement du corps humain est une "problématique" difficile. De nombreuses parties du corps participent de façon indépendante au déplacement de la personne. Les bras et les jambes ont par exemple des mouvements quasi-périodiques. La tête se meut de façon autonome autour de la trajectoire principale imposée par le reste du corps (voir Fig. 4.22). Cette complexité nous amène à une situation que les algorithmes conventionnels d'estimation de mouvement n'arrivent pas à résoudre



(a) Image réelle. (b) Identification des zones.

FIG. 4.20 – Résultat du processus d'identification dans un cas de sur-segmentation.



(a) Image réelle. (b) Identification des zones.

FIG. 4.21 – Résultat du processus d'identification dans un cas d'occultation.

de façon simple et rapide. Dans notre étude, nous nous limitons au suivi du visage. Cette problématique trouve en effet des applications concrètes en télé-conférence, en visiophonie mais aussi dans des développements moins standards comme la surveillance du comportement d'un conducteur aux commandes d'un véhicule [SSdIL03].

Nous proposons dans cette section une modification de la méthode présentée dans le chapitre 2 de manière à l'adapter au suivi de visage. Deux modifications majeures vont être introduites pour réaliser cette adaptation, l'une concerne la méthode de pré-segmentation et l'autre le descrip-

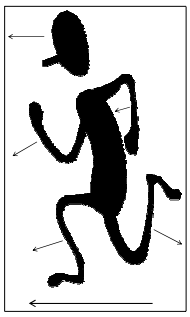


FIG. 4.22 – Déplacement d'une personne et mouvement des différentes parties du corps.

teur définissant la zone. Tout d'abord, nous avons besoin d'une méthode de segmentation bas-niveau focalisée sur la détection de visage. Pour ce faire, nous avons sélectionné une méthode de détection de peau fondée sur les attributs de chrominance [TDA98]. Un apprentissage statistique à partir d'une base de données est réalisé. Cet apprentissage a pour objectif de réaliser une estimation des statistiques d'ordre deux de l'intensité dans l'espace de chrominance. Ces valeurs sont alors injectées dans une mesure de dissemblance conditionnant la classification.

Afin d'obtenir une méthode de suivi stable, nous avons décidé d'attribuer au visage un descripteur de forme plus constraint qu'un simple contour. Nous modélisons le visage par un contour elliptique. Ce modèle géométrique a l'intérêt d'offrir un espace paramétrique bien conditionné pour envisager une prédiction à l'aide du filtre de Kalman. Nous allons, en effet, nous appuyer sur la modélisation dynamique de la loi d'évolution temporelle des coefficients de l'ellipse pour réaliser notre prédiction.

L'ellipse joue un rôle stabilisateur dans le processus de l'estimation du mouvement. Une fois la tête de la personne identifiée et son ellipse construite, nous pouvons déclencher la phase d'estimation du mouvement. Nous allons pour cela modifier la méthode d'appariement de points caractéristiques en utilisant l'ellipse.

4.2.2 Segmentation de la peau

Comme nous l'avons introduit précédemment, la première phase du suivi est la segmentation de la peau. Notre objectif est bien de retrouver le visage d'une personne qui se déplace tout au long d'une séquence vidéo. Pour concevoir la méthode de segmentation, nous nous appuyons sur les caractéristiques

de chrominance de la peau humaine fournies par le capteur vidéo.

Segmentation de la peau sur une image

Une grande variété de méthodes de détection du visage a été proposée ces dernières années [TC, STEK96]. De l'utilisation de méthodes fondées sur des **Eigenfaces** à celles qu'utilisent des attributs colorimétriques, les techniques présentent différents niveaux de complexité. Dans notre cas, nous voulons disposer d'une méthode de faible complexité nous permettant d'illustrer la pertinence de notre procédé de poursuite. Dans ce mémoire, nous avons décidé d'utiliser une approche exploitant la couleur comme base discriminante. Il est à noter que cette méthode peut être très largement enrichie en la couplant avec d'autre techniques utilisant des informations biométriques complémentaires comme les lèvres ou les yeux par exemple.

La couleur de la peau

La détermination de la couleur de la peau est une tâche à laquelle beaucoup de chercheurs se sont intéressés [Kim98, DN95] . La difficulté majeure est de se doter d'un modèle dédié à la couleur de peau suffisamment représentatif pour assurer une classification robuste. La classification consiste en l'extraction des pixels dont la couleur s'apparente à celle de la peau. Des travaux récents ont montré que la variabilité de la couleur de la peau tenait plus à une différence d'intensité que de la chromaticité [TDA98]. Il est connu, aussi, que la couleur de la peau contient une composante dominante de rouge dû au sang. Nous avons donc sélectionné l'espace des chrominances rouge et bleu permettant une certaine insensibilité à la diversité de types de peau, comme les types asiatiques, noirs et caucasiens. En outre, des études ont montré [AaD00, TDA98] que la couleur de peau analysée à partir d'echantillons positionnés dans l'espace normalisée C_bC_r a la forme d'un ellipsoïde. D'autres espaces de couleur peuvent être utilisés, mais nous ne développerons pas plus avant cette étude.

Classification supervisée

Sur la base de l'espace C_bC_r, nous utilisons une large banque d'images de peau afin de disposer d'un ensemble d'apprentissage. A partir de cet échantillon, nous modélisons la portion de l'espace de couleur associée à la peau, ce qui constitue notre connaissance a priori pour la classification dédiée à la peau. Différentes études proposent des modèles plus ou moins complexes partant de la simple distribution gaussienne en passant par le

mélange gaussien jusqu'à la méthode des noyaux [Vez02, PT03]. L'étape de décision s'appuie alors sur cette modélisation et sur une fonction discriminante pour réaliser la classification.

Pour notre application, nous avons sélectionné une modélisation *"elliptique"* avec une distance de Mahalanobis comme métrique de discrimination.

Identification des paramètres de la distribution associés à la couleur de peau

L'ensemble d'apprentissage est constitué par la concaténation de pixels "peau" sélectionnés par "contourage" à partir d'un nombre important d'images de visages. Soit n le nombre de pixels contenus dans l'ensemble d'apprentissage. Cet ensemble est ensuite réduit à l'espace des chrominances rouges et bleues normalisées.

Soient C_1 et C_2 les vecteurs d'échantillons de ce nouvel espace de couleurs. A partir de ces deux vecteurs, nous estimons les statistiques du premier et du second ordre qui permettent de caractériser la distribution des couleurs. Soit $M = (m_1, m_2)$ le vecteur contenant les moyennes de chacune des composantes définies par :

$$m_i = \frac{1}{n} \sum_{k=1}^{n} C_i [k] \tag{4.7}$$

Soit V la matrice des covariances associée à chacune des composantes :

$$V = \begin{bmatrix} \sigma_{C_1C_1} & \sigma_{C_1C_2} \\ \sigma_{C_1C_2} & \sigma_{C_2C_2} \end{bmatrix} \tag{4.8}$$

avec

$$\sigma_{C_iC_j} = \frac{1}{n^2} \sum_{k} \sum_{l} [C_i [k] - m_i] \cdot [C_j [l] - m_j]$$

Ces différentes quantités vont constituer notre modèle. Nous allons maintenant présenter la méthode de classification qui exploite ces paramètres.

Etape de discrimination

La classification consiste à prendre chaque pixel entrant comme un candidat et à le tester conformément aux paramètres statistiques estimés et à l'intégrer ou à le rejeter dans la classe peau à partir d'une métrique appropriée. Etant donné que nous avons un ellipsoïde comme représentant de forme, une métrique adaptée est la distance de Mahalanobis qui prend en

compte les paramètres d'échelle entre les axes principaux de l'ellipsoïde. La distance de Mahalanobis définit une valeur d_M qui mesure la distance d'une donnée $C = (c_1, c_2)$ au centre de gravité du modèle de peau, M. Cette distance est donnée par la relation :

$$d_M = (C - M) \cdot V \cdot (C - M)^T \tag{4.9}$$

La Fig. 4.23 nous montre une image contenant un visage ainsi que la carte associée des distances de Mahalanobis.

(a) Image réelle. (b) Carte de distance de Mahalanobis.

FIG. 4.23 – Image contenant un visage et carte de distance de Mahalanobis associée.

La segmentation finale de l'image est ensuite extraite à partir de cette carte des distances grâce à la fonction de probabilité estimée comme suit :

$$P_{peau} = e^{-d_M^2} \tag{4.10}$$

Un simple seuillage sur la fonction de probabilité, P_{peau} nous permet d'extraire la carte binaire de segmentation de peau de l'image. Soit S_p la carte de segmentation de la peau extraite comme suit :

$$S_p = \begin{cases} 1 \; si \; P_{peau} > seuil \\ 0 \quad\quad sinon \end{cases} \tag{4.11}$$

La Fig. 4.24 montre l'image des probabilités et la carte de segmentation associée.

4.2.3 Descripteur du visage

La méthode de suivi que nous proposons s'appuie sur une extraction bas niveau et haut niveau des objets. Pour la segmentation bas niveau, nous

(a) Image réelle. (b) Image des probabilités.

(c) Carte de segmentation de la peau.

FIG. 4.24 – Extraction de la carte de segmentation de la peau à partir de l'image des probabilités.

venons de présenter la méthode choisie qui nous permet d'obtenir des régions représentant des groupements connexes de pixels associés aux visages. Il est intéressant maintenant de voir dans ce cadre du suivi de visages quels descripteurs haut-niveau nous pouvons définir.

Dans le chapitre 2 nous avions fondé notre réflexion générale sur un descripteur de contour. Pour les visages, il est intéressant de noter qu'une forme générique peut-être introduite [TKO99]. En effet, la forme elliptique représente une bonne approximation de la forme d'un visage. Cette simplification va permettre d'introduire un filtrage de forme alimentant les étapes d'estimation de mouvements et d'identification.

Création du descripteur

Pour chaque visage, nous créons une ellipse donnée par :

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_g \\ y_g \end{bmatrix} + \underbrace{\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}}_{R_T} \cdot \begin{bmatrix} a \cdot \cos\alpha \\ b \cdot \sin\alpha \end{bmatrix} \tag{4.12}$$

avec $0 < b \leq a$, $\alpha \in [0, 2\pi]$ et R_T la *matrice de rotation*.

Cinq paramètres caractérisent donc l'ellipse :

1. Le centre de gravité de l'ellipse : (x_g, y_g).
2. Les valeurs du grand axe et du petit axe : (a, b).
3. L'angle de rotation de l'ellipse Θ.

Nous devons pour chaque image estimer ces différents paramètres pour ensuite évaluer les positions prédites afin de mettre en correspondance les ellipses.

Extraction des axes principaux

L'estimation des paramètres de l'ellipse est réalisée en effectuant une Analyse en Composantes Principales (ACP) du nuage de points issu de la phase de détection. L'annexe B présente un résumé des points clés de la méthode ACP. Soit $N(\Omega)$ le nuage de points et soit X la matrice de coordonnées centrées de chaque point. La matrice X est de dimension $(n, 2)$. L'ACP permet de retrouver les axes de plus grandes variations associés à $N(\Omega)$. Ces axes sont les vecteurs propres de la matrice de covariance associée à la matrice X. En outre, l'obtention des valeurs propres permet de retrouver les valeurs de a et de b. Nous avons :

$$R = \frac{1}{n} X X^T = U D U^T \tag{4.13}$$

avec

$$R = \begin{bmatrix} C_{x,x} & C_{x,y} \\ C_{y,x} & C_{y,y} \end{bmatrix}$$

où

$$\begin{aligned} C_{x,x} &= \frac{1}{n} \sum_i (x_i - x_g)^2 \\ C_{y,y} &= \frac{1}{n} \sum_i (y_i - y_g)^2 \\ C_{x,y} &= C_{y,x} = \frac{1}{n} \sum_i (x_i - x_g) \cdot (y_i - y_g) \end{aligned} \tag{4.14}$$

La matrice U contient les vecteurs propres orthonormés avec v_1 le vecteur propre associé à la plus grande valeur propre et D la matrice diagonale contenant les deux valeurs propres. La Fig. 4.25 montre un exemple du lien graphique entre le nuage et ces différentes grandeurs.

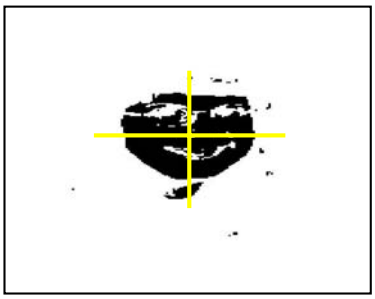


FIG. 4.25 – Exemple d'extraction des axes principaux de la segmentation.

Au final, nous avons les relations suivantes :

$$a = \sqrt{d_1} \tag{4.15}$$

$$b = \sqrt{d_2} \tag{4.16}$$

$$v_1 = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \tag{4.17}$$

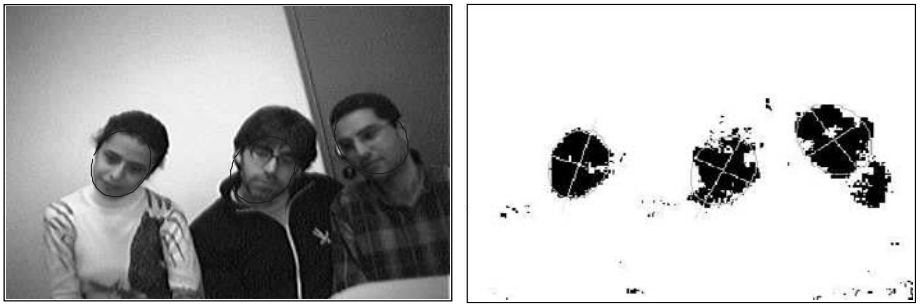
Les Fig. 4.26 et 4.27 nous montrent un résultat de modélisation du visage à partir du descripteur "ellipse".

4.2.4 Prédiction du descripteur de forme

Nous avions vu dans le chapitre 2 que la méthode proposée s'appuyait sur la prédiction pour assurer le suivi. Pour les visages, il nous reste à voir comment exploiter la contrainte de forme dans les phases d'estimation du mouvement et de la prédiction. La technique d'estimation de mouvement que nous avons développée pour l'application autoroutière était fondée sur un ensemble de points caractéristiques. Dans l'application visage, les mouvements 3D inhérents à l'évolution d'un visage rendent plus difficile la prédiction car le visage se déplace avec six degrés de liberté. L'extraction et la mise en correspondance de points caractéristiques sur le visage n'est pas une tâche aisée. Cependant ; l'ellipse constitue une solution alternative intéressante. L'ensemble des points qui constituent son contour forme un ensemble tout

(a) Image de segmentation. (b) Image réelle et ellipse superposée.

FIG. 4.26 – Modélisation par ellipses des résultats de la segmentation du visage humain.



(a) Image de segmentation. (b) Image réelle et ellipses superposées.

FIG. 4.27 – Deuxième exemple de la modélisation par ellipses des résultats de la segmentation du visage humain.

désigné pour la mise en correspondance. Il est plus simple de poursuivre et donc d'estimer le mouvement propre à l'ellipse. C'est cette solution que nous avons choisie. En considérant les points de l'ellipse, deux solutions s'offrent à nous pour gérer la prédiction : soit directement en modélisant l'évolution de paramètres de mouvement soit en considérant l'évolution des paramètres propres à l'ellipse.

Prédiction à partir du modèle affine à six paramètres

Le modèle affine à six paramètres décrit par l'équation 4.18 nous permet de définir parfaitement les déplacements d'un objet à partir des six

paramètres a_i :

$$\begin{cases} d_x = a_1 + a_2(x - x_g) + a_3(y - y_g) \\ d_y = a_4 + a_5(x - x_g) + a_6(y - y_g) \end{cases} \tag{4.18}$$

Nous avions aussi vu que pour la prédiction du modèle de mouvement il fallait trouver un modèle qui nous permette de modéliser les variations des paramètres. . Comme nous avions pu le constater, à travers un modèle AR simple comme celui de l'équation (4.19), nous arrivions à modéliser le déplacement effectué par un objet rigide en mouvement. La Fig. 4.28 nous montre comment ce même modèle s'adapte également aux déplacements du visage humain.

$$a_i(k+1) = \sum_{j=0}^{p-1} \alpha_{ij} \cdot a_i(k-j) + v_i(k) \tag{4.19}$$

où chaque a_i est un paramètre dans le modèle affine à six paramètres.

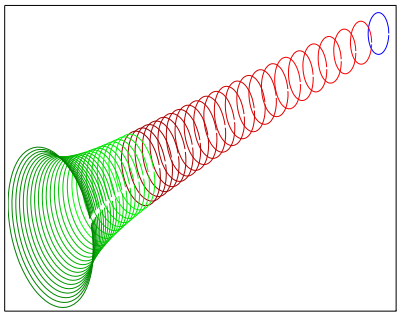
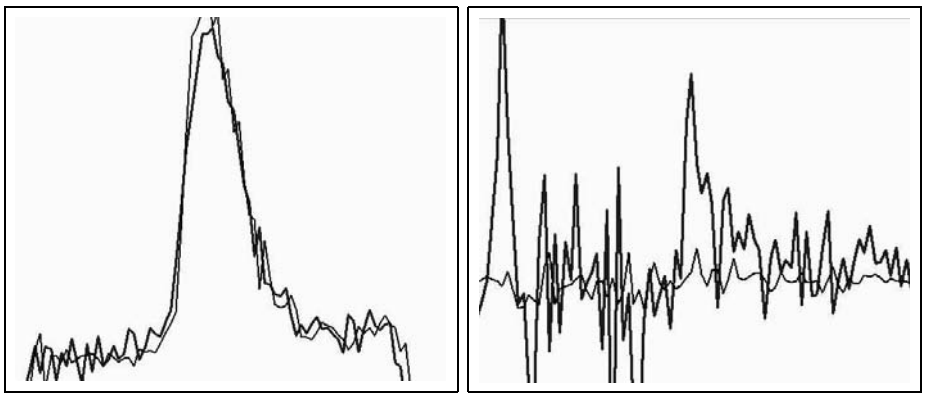


FIG. 4.28 – Evolution des positions estimées à travers un modèle AR d'un visage se rapprochant vers la caméra.

Si maintenant nous regardons les courbes réelles d'estimation et celles de prédiction des paramètres a_i on peut constater que les paramètres sont très instables à cause des petits mouvements de la tête qui font varier de façon assez importante leurs amplitudes. Ces fluctuations rendent peu précise la prédiction (Fig. 4.29).

Prédiction à partir du modèle de l'ellipse

La seconde solution est d'utiliser une loi d'évolution pour les paramètres de l'ellipse. Cette prise en compte directe nous permet d'éviter une phase d'estimation des paramètres de mouvement qui sont déjà donnés par les



(a) Evolution réelle (gras) et prédiction du paramètre a_1.

(b) Evolution réelle (gras) et prédiction du paramètre a_2.

FIG. 4.29 – Prédiction de la position du descripteur du visage humain à travers des paramètres de mouvement.

paramètres de l'ellipse. Nous avons en effet estimé par ACP les paramètres suivants :

1. Deux paramètres de translation, x_g^k, y_g^k : ces paramètres sont extraits à partir du centre de gravité de l'ellipse, et vont nous donner les amplitudes de translation du déplacement entre deux instants consécutifs : $t_x^k = x_g^k - x_g^{k-1}$ et $t_y^k = y_g^k - y_g^{k-1}$.
2. Deux paramètres, a, b, pour définir l'effet de zoom : ces deux paramètres sont extraits du grand et petit axe de l'ellipse et sont directement liés au zoom introduit par l'éloignement ou le rapprochement de l'ellipse.
3. Un paramètre pour définir la rotation, Θ.

Nous rassemblons tous ces paramètres[2] sous la forme d'un vecteur, que nous notons, t :

$$t = \begin{bmatrix} x_g \\ y_g \\ a \\ b \\ \Theta \end{bmatrix} \tag{4.20}$$

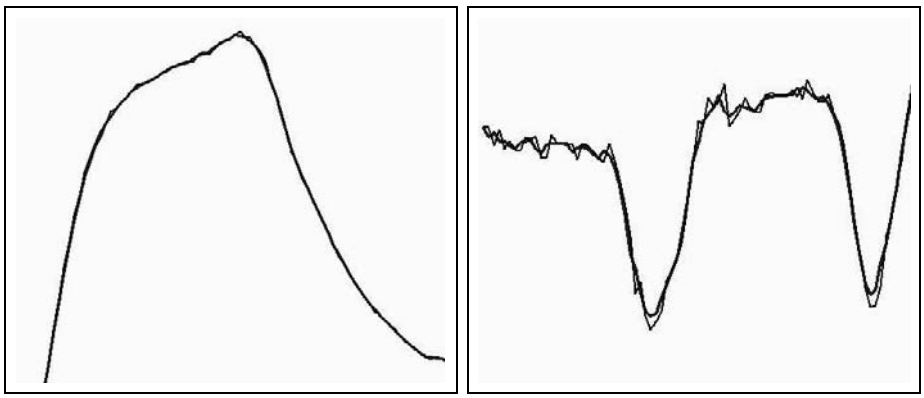
[2]Noter que si bien on accepte cinq paramètres, pour de raisons de simplicité au lieu de Θ on utilisera $\cos(\Theta)$ et $\sin(\Theta)$.

De la même manière que précédemment, où nous avions un modèle AR pour chaque paramètre du modèle affine, nous allons maintenant modéliser les paramètres de l'ellipse avec le même type de modèle AR sont :

$$t_i(k+1) = \sum_{j=0}^{p-1} \beta_{ij} \cdot t_i(k-j) + v_i(k) \tag{4.21}$$

où chaque t_i est un paramètre de l'ellipse.

Dans la Fig. 4.30 on représente les courbes d'évolution des paramètres réels et leur prédiction issue du filtre de Kalman utilisant le modèle AR.



(a) Evolution réelle (gras) et prédiction du paramètre t_1.

(b) Evolution réelle (gras) et prédiction du paramètre t_3.

FIG. 4.30 – Prédiction de la position du descripteur du visage humain à travers des paramètres de l'ellipse.

Nous pouvons ainsi constater que l'évolution des paramètres est plus lisse que dans le cas du modèle affine. Nous obtenons ainsi une prédiction plus ajustée aux valeurs réelles, les erreurs de prédiction sont minimisées et le visage est ainsi mieux suivi.

4.2.5 Conclusion

Nous avons présenté dans cette section une application permettant le suivi humain à partir du suivi des visages. Le fait de suivre les visages comme référence est fondé sur les différents mouvements indépendants présents sur le corps humain. Ainsi, les bras ou les jambes suivent des mouvements périodiques très différents. Suite à cette différence, les algorithmes classiques

d'estimation de mouvement ne sont pas capables d'extraire le modèle de mouvement adapté à l'ensemble du corps.

L'application que nous avons proposée est fondée sur une première phase de pré-segmentation extraite à partir de l'information de chrominance de la peau humaine. Une étude caractérisant les différents espaces de couleur a été aussi développé. A partir de ces informations, un descripteur d'objet précise la position et l'évolution du visage. Le descripteur choisi (une ellipse) nous permet la reconnaissance du visage grâce à la prédiction de sa position. Cette prédiction est établie, tel qu'on l'avait déjà avancé, à partir d'un filtre de Kalman sur le modèle de mouvement du descripteur. Deux modèles de mouvement ont été comparés. Le premier, le modèle affine à six paramètres classiques, n'assure pas une bonne prédictibilité à cause de l'instabilité des paramètres, notamment les paramètres de la rotation et de l'effet de zoom. Le second, le modèle de l'ellipse, étant plus stable, assure une bonne prédiction et permet une bonne qualité de la phase d'identification.

Il faut aussi noter que dans le but d'améliorer la phase d'identification à partir de l'approche EM, il reste à valider de nouveaux résidus plus discriminants. Des premiers résultats ont déjà été extraits à partir d'un résidu non local fondé sur le coefficient de corrélation entre deux zones prédites. Ces résultats nous encouragent pour continuer à travailler dans ce sens et à donner une suite au travail développé dans cette thèse.

Chapitre 5

Conclusion

Le travail présenté dans ce manuscrit s'inscrit dans le cadre de la conception d'une architecture de traitements permettant de réaliser le suivi d'entités mobiles à partir d'un flux continu d'images. Le suivi étant une tâche commune à de très nombreuses applications, nous nous sommes attachés à rechercher un formalisme facilitant l'implémentation algorithmique. En conséquence, nous avons proposé une architecture mettant en jeu divers modules algorithmiques génériques. Ces modules, au nombre de quatre, sont dédiés respectivement à la détection, à la reconnaissance, à l'identification et à la prédiction. Outre la mise en évidence du caractère générique des modules, nous nous sommes efforcés de proposer des prototypes standardisant le plus possible les données intra module et facilitant les interactions inter modules. A partir de ce canevas, l'établissement d'une solution consiste à particulariser les données internes et à sélectionner les techniques algorithmiques dédiées à chacun des modules selon le contexte applicatif et les objectifs désirés.

5.1 Bilan

S'inspirant du système visuel humain, nous avons tenté, tout au long de ce manuscrit, de répondre au problème de la poursuite d'objets en vision artificielle. Notre contribution s'articule autour des points suivants :

- Nous avons introduit la notion d'entrées perceptives, entrées qui vont directement alimenter le procédé de suivi. Obtenues dans l'étape de détection, elles sont le résultat d'une réduction de l'information visuelle brute. A l'instar des cartes de saillance générées par le système visuel, elles mettent en évidence les caractéristiques de texture, de

forme et de mouvement associées aux objets cibles. La sélection d'un nombre limité de caractéristiques alimentant les traitements haut niveau permet de réduire de façon drastique la complexité calculatoire du procédé global. Dans ce contexte, nous avons proposé une nouvelle méthode de détection du mouvement fondée sur une construction adaptative d'une image de référence. La méthode développée s'adapte à des séquences extérieures même dans le cas de fortes variations d'éclairage et gère en outre les ombres portées.

- Nous avons montré l'intérêt d'utiliser une description abstraite de l'objet afin d'obtenir un procédé robuste. Construite sous la forme d'un ensemble de descripteurs pour chaque objet, la représentation abstraite utilisée regroupe des descripteurs génériques (état, mouvement) et d'autres descripteurs plus spécifiques dictés par le contexte applicatif. C'est le cas du descripteur de forme. Celui-ci peut, en effet, n'exprimer qu'un caractère apparent de l'objet comme le contour extérieur obtenu par exploitation directe des entrées perceptives, sans aucune régularisation. Ce niveau de représentation suffit dans le cas d'objets rigides dont le mouvement est relativement contraint, voie routière par exemple. Cependant, certaines applications sont plus exigeantes. Dans le cas de mouvements plus libres, le descripteur de forme doit alors être associé à l'objet en régularisant son expansion spatiale à l'aide d'un modèle géométrique par exemple. L'utilisation d'une ellipse pour décrire le contour d'un visage illustre ce type de cas. Nous avons montré que l'utilisation d'une régularisation par modèle, même si elle rajoute une étape d'estimation, augmente la robustesse de la solution proposée.
- Nous avons mis en évidence l'intérêt de créer une interaction entre les procédés verticaux dédiés respectivement à la détection et à la reconnaissance. La phase de reconnaissance remet en cause les résultats de la détection en descendant jusqu'au pixel pour éditer des cartes d'appartenance afin de gérer les cas d'occultation. De son côté, la détection utilise les descripteurs pour contrôler ses paramètres décisionnels. Les descripteurs et les entrées perceptives sont alors directement combinés. Dans le contexte de la reconnaissance, nous avons proposé une technique dérivée du critère de Bayes pour la génération des cartes d'appartenance. Cette approche permet une gestion multi-objets. Des résultats fiables sont obtenus notamment en présence d'occultations.
- Nous avons mis en pratique les résultats ci-dessus dans le cadre de la mise en oeuvre de deux applications : le suivi de véhicules pour le contrôle du trafic routier et la poursuite de visages. De nombreux

résultats sur la base de séquences réelles illustrent le bien fondé de l'architecture générique que nous avons proposée.

5.2 Limites et perspectives

Comme pour tout travail de recherche, il est fondamental d'évaluer les limites des solutions proposées. Ces limites ouvrent autant de perspectives qui doivent être situées dans le contexte général de la vision artificielle.

Aspects techniques : Dans cette thèse, nous avons proposé un certain nombre de méthodes permettant de répondre au problème de suivi d'objets. Comme nous l'avions noté dans l'introduction, la majeure partie des applications concernées par ce type de "problématique" nécessite une gestion temps réel du suivi. Nous avons eu le soucis de réduire au maximum la complexité des méthodes proposées mais nous n'avons pas présenté dans ce manuscrit une étude approfondie, pour un contexte applicatif donné, de la complexité calculatoire dans le but de valider une mise en oeuvre quasi temps réel.

Aspects méthodologiques : Concernant les quatre phases du procédé, certaines limitations existent quant aux solutions proposées. Pour la phase d'estimation, les limites et les perspectives sont à évaluer en fonction d'une "littérature" très riche qui concerne le traitement du signal au sens large. Cette phase se décline en effet sur la base des nombreux algorithmes exploités en théorie de l'estimation (itératifs, robustes, multi-échelles ...) qui n'ont pas de spécificité particulière au regard de la problématique de suivi. Les trois autres phases par contre sont plus spécifiques à la tâche que nous étudions. Il est donc intéressant de faire certaines remarques les concernant.

Dans le cadre de la détection, certaines voies restent à explorer. Premièrement, le nombre et le choix des entrées perceptives sont des éléments déterminants quant à la pertinence du procédé. A ce titre, nous devons accorder à ce module une attention toute particulière. Une étude des travaux concernant l'obtention des cartes de saillance modélisant le processus attentionnel en vision est à réaliser. En outre, concernant plus spécifiquement le suivi de visages, l'utilisation de techniques prenant en compte des données biométrique est à développer.

Concernant la phase de reconnaissance, les techniques proposées peuvent être modifiées dans le but de s'adapter à des applications différentes. Des extensions peuvent être étudiées concernant la génération des descripteurs tels que ceux exploitant des contours ou des modèles géométriques plus complexes. Tout dépend de l'application étudiée et de l'objectif recherché. La

complexité des descripteurs doit être modulée en fonction de l'objectif selon que l'application nécessite juste une reconnaissance temporellement de l'objet ou l'identification de celui-ci en tant qu'élément d'une classe parmi d'autres (voiture, camion, piéton ...). De nombreux travaux réalisés dans le contexte de l'indexation peuvent sûrement être pris en compte pour atteindre cet objectif.

Si nous revenons maintenant sur l'approche proposée dans ce manuscrit, la méthode est fondée sur la règle de Bayes en exploitant une analyse locale de l'image afin de générer des cartes d'appartenance. Dans le but d'améliorer la pertinence de la mise en correspondance, il serait intéressant d'injecter un test s'appuyant sur une analyse plus globale de l'objet fondée sur la corrélation ou sur une métrique inter histogramme comme le suggèrent les travaux de Comaniciu [CR00]. Une double analyse locale/globale viendrait sans doute renforcer la pertinence de la décision.

Enfin, pour certaines applications et malgré un formalisme élégant, une efficacité certaine et une complexité réduite, le schéma de prédiction que nous avons développé peut présenter une limitation dans sa capacité à prédire de façon "correcte" un mouvement 3D associé à un objet non-rigide. C'est notamment le cas du corps humain pour lequel une représentation descriptive plus fine est nécessaire pour assurer la prédictibilité.

Annexe A

Méthode LMS multi-résolution et incrémentale

A.1 Estimation incrémentale

Soit I la fonction d'intensité de l'image. Soit $S(p,t)$ la fonction des dérivées temporelles de la position p de l'image. Nous avons :

$$S(p,t) = \vec{d} \cdot \bar{\nabla} I(p,t) + I_t(p,t) \tag{A.1}$$

où $\vec{d}$ est le vecteur de déplacement avec $\bar{d} = Q \cdot A$ et :

$$Q = \begin{pmatrix} 1 & x & y & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & x & y \end{pmatrix}$$

et

$$A = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{pmatrix}^T$$

avec $\bar{\nabla} I$ est le vecteur des gradients spatiaux et I_t le gradient temporel. Sous l'hypothèse de conservation de la luminosité établie par Horn et Schunck [HS81], nous trouvons que (A.1) s'écrit comme suit :

$$S(p,t) = 0$$

Si on suppose maintenant que la luminosité globale peut varier linéairement d'une image à l'autre l'équation (A.1) devient :

$$S(p,t) = -\xi$$

Par rapport à cette notation, le résidu de compensation s'écrit :

$$r = I\left(p + \bar{d}, t+1\right) - I\left(p, t\right) + \xi \tag{A.2}$$

avec $p = (x, y)$.

Si nous développons cette dernière équation en fonction des dérivées partielles, nous obtenons :

$$r = \bar{\nabla}_x d_x + \bar{\nabla}_y d_y + \bar{\nabla}_t + \xi \tag{A.3}$$

ou de façon matricielle :

$$r = \mathrm{X} \cdot \theta - \Upsilon \tag{A.4}$$

où $\mathrm{X} = \begin{pmatrix} \bar{\nabla}_x & \bar{\nabla}_x \cdot x & \bar{\nabla}_x \cdot y & \bar{\nabla}_y & \bar{\nabla}_y \cdot x & \bar{\nabla}_y \cdot y & 1 \end{pmatrix}$, $\Upsilon = -\nabla_t$ et $\theta = (A, \xi) = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & \xi \end{pmatrix}^T$

L'approche incrémentale définit la mise à jour comme suit :

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \Delta d_k \tag{A.5}$$

Si nous décomposons les deux termes de l'équation (A.5) par rapport à la parametrization, nous avons :

$$\begin{aligned} \hat{A}_{k+1} &= \hat{A}_k + \Delta A_k \\ \hat{\xi}_{k+1} &= \hat{\xi}_k + \Delta \xi_k \end{aligned} \tag{A.6}$$

En réalisant un développement au première ordre du résidu autour du point $p + \hat{d}_k$, nous obtenons :

$$r' = \mathrm{X}' \Delta d_k - \Upsilon' \tag{A.7}$$

où

$$\mathrm{X}' = \begin{pmatrix} \bar{\nabla}_x(p+\hat{d}) & \bar{\nabla}_x(p+\hat{d}) \cdot \left(x + \hat{d}_x\right) & \bar{\nabla}_x(p+\hat{d}) \cdot \left(y + \hat{d}_y\right) \cdots \\ \bar{\nabla}_y(p+\hat{d}) & \bar{\nabla}_y(p+\hat{d}) \cdot \left(x + \hat{d}_x\right) & \bar{\nabla}_y(p+\hat{d}) \cdot \left(y + \hat{d}_y\right) \quad 1 \end{pmatrix}$$

et $\Upsilon' = I(p, t) - I(p + \hat{d}, t+1) - \hat{\xi}$

La Fig. A.1 présente géométriquement l'approche incrémentale sur un signal à une dimension.

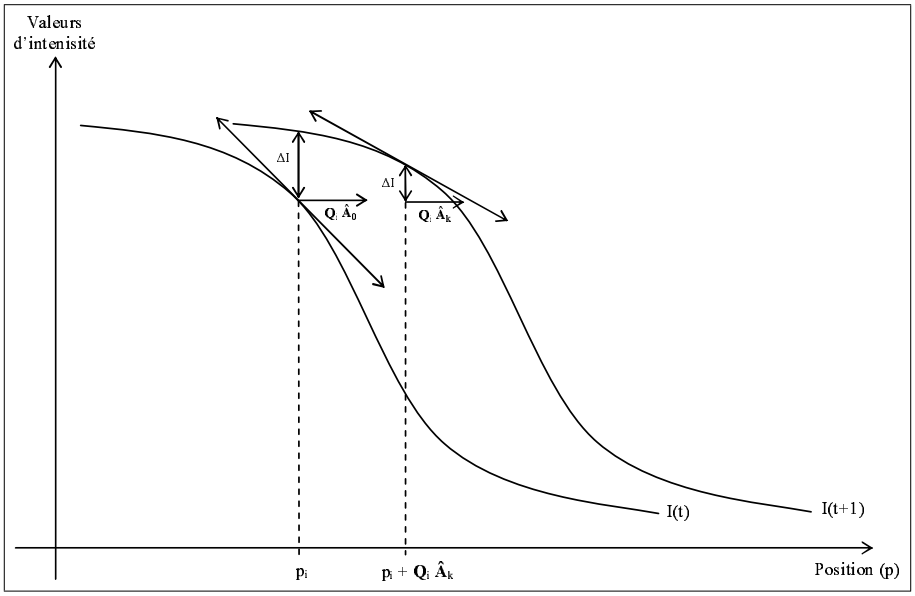


FIG. A.1 – Variations temporelles des grandeurs contribuant à la méthode incrémentale le long d'un profil 1D.

A.2 Estimation multi-résolution (coarse-to-fine)

Une approche multi-résolution consiste à estimer le modèle de mouvement entre deux images à différents niveaux de résolution des images. L'approche *"coarse-to-fine"* commence par le niveau de moindre résolution pour finir au niveau de résolution maximale. L'implantation de ce type d'approche nécessite la construction d'une pyramide Gaussienne de l niveaux. La Fig. A.2 nous schématise le résultat de la génération d'une pyramide dans les cas de $l = 3$ niveaux. Soit I_i l'image trouvée au niveau i de la pyramide. Cette image est sous-échantillonnée par un facteur deux par rapport à l'image du niveau inférieur I_{i+1}. Dans le dernier niveau (au niveau le plus bas) $i = l-1$, nous trouvons l'image de résolution complète.

L'algorithme *"coarse-to-fine"* commence au niveau le plus haut $i = 0$ avec l'image ayant la plus petite résolution. Pour chaque niveau, nous déployons le processus incremental comme nous l'avons décrit dans la section précédente. Le processus incremental s'arrête lorsque la différence entre deux itérations consécutives est inférieure à une certaine valeur (convergence des paramètres), ou bien lorsque un nombre d'itérations donnée est atteint. A partir de ce résultat nous descendons d'un niveau dans la pyramide et

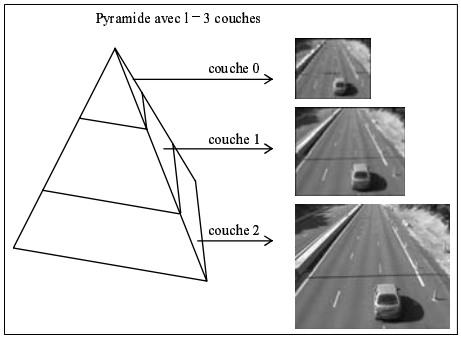


FIG. A.2 – Pyramide gaussienne avec $l = 3$ couches et les images pour les différentes résolutions.

nous relançons le processus d'estimation. Pour le passage d'un niveau i au niveau inférieur $i + 1$, le déplacement estimé $\hat{d}_i$ est à un facteur deux du déplacement du niveau inférieur :

$$\hat{d}_{i+1} = 2 * \hat{d}_i \tag{A.8}$$

Ce déplacement $\hat{d}_{i+1}$ est utilisé comme initialisation du processus incremental du niveau inférieur. Ce processus est répété jusqu'à que nous ayons atteint le niveau le plus bas. Le déplacement estimé représente alors le déplacement réel.

Annexe B

Analyse par Composantes Principales. Méthode ACP

L'Analyse par Composantes Principales, ou simplement méthode ACP fait parti des méthodes factorielles. Ces méthodes cherchent à représenter et traiter les données multidimensionnelles, c'est à dire, des données qui tient en compte plusieurs variables. La méthode ACP considère que les données peuvent être constituées sur un ensemble des variables non corrélées linéairement. Ces variables sont des combinaisons linéaires des variables en étude. En représentant les données sur un nuage des points, la méthode va chercher les directions principales de variation du nuage. Cette méthode avait déjà été introduite par K. Pearson en 1900 dans le cas de deux variables [Pea01]. H. Hotelling [Hot33] a été la première personne à étendre la méthode pour un nombre quelconque des variables.

Pour expliquer la méthode ACP on va s'appuyer dans le papier de Gérard Govaert [Gov03] qui introduit les méthodes d'analyse de données, notamment, la méthode ACP.

B.1 Axes principaux d'inertie

Soit X la matrice des données de dimension (n, p) qui contient n individus pour lesquels on connaît p variables différentes. Soit $N(\Omega)$ le nuage des points de $\mathbb{R}^p$ à partir duquel on veut obtenir les axes principaux de façon à représenter le nuage à partir d'un espace de faible dimension. Dans notre cas, les espaces de représentation seront les espaces affines, comme une droite, un plan, etc. La formulation de l'ACP nous dise : soit E_k un sous-espace affine de dimension k avec $k < p$ tel que l'inertie du nuage $N(\Omega)$ par rapport

à E_k, I_{E_k}, soit minimum :

$$I_{E_k} = \frac{1}{n} \cdot \sum_i d^2 (x_i, E_k) \tag{B.1}$$

où x_i représente chaque point contenu dans le nuage des points. On sait aussi que l'inertie totale du nuage, I peut s'exprimer comme la somme de $I_{E_k} + I_{E_k^\perp}$, où $I_{E_k^\perp}$ est l'inertie expliquée par E_k, où le terme d'inertie expliquée par E_k vient à définir l'inertie des points projetés orthogonalement sur E_k. On peut donc réformuler notre problème pour trouver le sous-espace E_k de dimension k tel que l'inertie $I_{E_k^\perp}$ expliqué par E_k soit maximum.

B.2 Solution au problème

Soit S la matrice des variances du nuage des points, étant symétrique. De cette façon on trouve que toutes les valeurs propres de S sont positives ou nulles et les vecteurs propres forment une base orthonormée. Soient $u_1, \ldots, u_p$ les vecteurs propres normés ordonnés suivant les valeurs propres en ordre décroissante. On peut voir comme la solution au problème à l'étape k vient à définir :

$$E_k = E_{k-1} \oplus \Delta u_k \tag{B.2}$$

avec $E_1 = \Delta u_1$. On peut aussi montrer comme l'inertie expliqué par chaque axe va verifier : $I_{\Delta u_k^\perp} = \lambda_k$.

B.3 Inerties expliquées

En effet on va obtenir que :

$$I_{E_k^\perp} = \lambda_1 + \ldots + \lambda_k \tag{B.3}$$

Pour prouver cette proposition, on sait que les vecteurs propres sont orthogonaux puisque la matrice S est symétrique. L'espace E_k étant décomposé en une somme directe de sous-espaces orthogonaux Δu_j (eq. B.2), alors :

$$I_{E_k^\perp} = \sum_{j=1}^{k} I_{\Delta u_j^\perp} \tag{B.4}$$

puisque $I_{\Delta u_j^{\perp}} = \lambda_j$ le résultat est donc démontré. En plus, si r c'est le rang de la matrice X (avec $r \leq min(p,n)$), en prenant $k = p$ on retrouve que $I = trace(S)$ où S est la matrice diagonale des valeur propres de X et aussi :

$$\lambda_1, \ldots, \lambda_r > 0 et \lambda_{r+1}, \ldots, \lambda_p = 0$$

et donc, $I_{E_r^{\perp}} = I$. Ce qui vient à dire que le nuage est dans les sous-espace vectoriel E_r engendré par les r premiers axes factories.

B.4 Calcul des composantes principales

Soit c^j les coordonnées de la projection de tous les points du nuage sur chaque axe factoriel :

$$c^j = \begin{pmatrix} c_1^j \\ \vdots \\ c_n^j \end{pmatrix} \tag{B.5}$$

On appelle c^j la j^{eme} composante principale. Soit C la matrice qu'on obtient à partir des vecteurs c^j mis en colonne. On peut constater que les composantes principales vont vérifier :

$$\mathbf{c}^j = X \cdot \mathbf{u}_j$$

où de façon matricielle :

$$C = X \cdot U$$

Pour démontrer cette proposition il suffit de projeter les x_i sur les vecteurs de la base orthonormée :

$$c_i^j = < \mathbf{x}_i, \mathbf{u}_j > = \mathbf{x}_i' \mathbf{u}_j \mathbf{c}^j = X \cdot \mathbf{u}_j C = X \cdot U$$

où U est la matrice des vecteurs propres normalisés.

Bibliographie

[AaD00] M. Al-aqrabawi and F. Du. Human skin detection using color segmentation. Term project ECPE 5554, March 2000.

[AB85] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Optical Soc. Am.*, A2 :284–299, 1985.

[AET98] Y. Altunbasak, P. Ehran Eren, and A. Murat Tekalp. Region-based parametric motion segmentation using color information. *Graphical Models and Image Processing*, 60(1) :13–23, 1998.

[AKM93] T. Aach, A. Kaup, and R. Mester. Statistical model-based change detection in moving video. *Signal Processing*, 31 :165–180, 1993.

[AMYT00] S. Araki, T. Matsuoka, N. Yokoya, and H. Takemura. Real-time tracking of multiple moving object contours in a moving camera image sequence. *Trans. Inf. and Syst.*, E.83-D(7) :1583–1591, July 2000.

[Ana89] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *Int. J. Computer Vision*, 2 :283–310, 1989.

[AS96] S. Ayer and H.S. Sawhney. Compact representation of videos through dominant and multiple motion estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(8) :814–830, 1996.

[ASB94] S. Ayer, P. Schroeter, and J. Bigün. Segmentation of moving objects by robust motion parameter estimation over multiple frames. In *Proc. of 3rd European Conf. on Computer Vision*, volume 801, pages 316–327, 1994.

[AT91] J. Aloimonos and D. Tsakiris. On the visual mathematics of tracking. *Image and Vision Computing*, 9 :235–251, 1991.

[AT97] Y. Altunbasak and A.M. Tekalp. Occlusion-adaptive, content-based mesh design and forward tracking. In *IEEE Trans. on Image Processing*, volume 6, pages 1270–1280, 1997.

[BA93] M.J. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Proc. of the 4th IEEE Int. Conf. on Computer Vision*, pages 231–236, 1993.

[BBDM94] B. Bascle, P. Bouthemy, R. Deriche, and F. Meyer. Tracking complex primitives in an image sequence. In *Proc. of the 12th IAPR Internation Conf. on Pattern Recognition*, pages 426–431, 1994.

[BD95] B. Bascle and R. Deriche. Region tracking through image sequences. In *Proc. of IEEE Int. Conf. on Computer Vision*, pages 302–307, 1995.

[BF93] P. Bouthemy and E. François. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *International Journal of Computer Vision*, 10(2) :157–182, 1993.

[BFBB] J.L. Barron, D.J. Fleet, S.S. Beauchemin, and T.A. Burkitt. Performance of optical flow techniques. *CVPR*, 92 :236–242.

[BGG96] V. Bruce, P.R. Green, and M.A. Georgeson. *Visual Perception.* Psychology Press, 1996.

[BI98] A. Blake and M. Isard. *Active Contours.* Springer ed., 1998.

[Bla94] M.J. Black. Recursive non-linear estimation of discontinuous flow fields. In *Proc. of the 3rd European Conference on Computer Vision*, pages 138–145, 1994.

[BPT88] M. Bertero, A. Poggio, and V. Torre. Ill-posed problems in early vision. In *Proc. of the IEEE*, volume 76, pages 869–889, 1988.

[Bré97] F. Brémond. *Environnement de résolution de problemes pour l'interprétation de séquences d'images.* PhD thesis, Inria Sophia (France), Oct. 1997.

[Bra74] O. Braddick. A short-range process in apparent motion. *Vision Research*, 14 :519–527, 1974.

[Cav02] A. Cavallaro. *From visual information to knowledge : semantic video object segmentation, tracking and description.* PhD thesis, Swiss Federal Institute of Technology, January 2002.

[CB99] G. Csurka and P. Bouthemy. Direct identification of moving objects and background from 2d motion models. In *Proc. of 7th IEEE Int. Conf. on Computer Vision*, pages 566–571, 1999.

[CE04] A. Cavallaro and T. Ebrahimi. Interaction between high-level and low-level image analysis for semantic video object extraction. *Applied Signal Processing, Special Issue on Object-based and semantic image and video analysis*, 6 :786–797, June 2004.

[CH96] I.J. Cox and S.L. Hingorani. An efficient implementation of reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(2) :138–150, 1996.

[CK97] L. Cohen and R. Kimmel. Global minimum for active contour models : a minimal path approach. *Int. Journal of Computer Vision*, 24(1) :57–78, 1997.

[CR00] D. Comaniciu and V. Ramesh. Robust detection and tracking of human faces with an active camera. In *Third IEEE Internation Workshop on Visual Surveillance*, pages 11–18, Dublin, 2000.

[DF90] R. Deriche and O. Faugeras. Tracking line segments. In *Proc. European Conference on Computer Vision*, pages 259–268, 1990.

[DHA88a] G.W. Donohoe, D.R. Hush, and N. Ahmed. Change detection for target detection and classification in video sequences. In *Internation Conference on Acoustic Speech and Signal Processing*, pages 1084–1087, 1988.

[DHA88b] G.W. Donohoe, D.R. Hush, and N. Ahmed. Change detection for target detection and classification in video sequences. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1988.

[Die91] N. Diehl. Object-oriented motion estimation and segmentation in image sequences. *Signal Processing : Image Communication*, 3 :23–56, 1991.

[DLC97] C. Dumontier, F. Luthon, and J.P. Charras. Real time implementation of an mrf-based motion detection algorithm on a dsp board. *IEEE Trans. on Image Processing*, 8(10) :1341–1347, 1997.

[DN95] Y. Dai and Y. Nakamo. Extraction of facial images from complex background using color information and sgld matrices. In *Workshop on Automatic Face and Gesture Recognition*, pages 26–28, June 1995.

[Dre78] L. Dreschler. *Using affinity for extracting images of moving objects from TV-frame sequences*. PhD thesis, Institut fur Informatik, 1978.

[DRMFT04] A. Delorme, G. Rousselet, M. Mace, and M. Fabre-Thorpe. Interaction of bottom-up and top-down processing in the fast visual analysis of natural scenes. *Cognitive Brain Research*, 19(2) :103–113, 2004.

[FET00] Y. Fu, A.T. Erdem, and A.M. Tekalp. Tracking visible boundary of objects using occlusion adaptive motion snake. *IEEE Trans. on Image Processing*, 9(12) :2051–2060, December 2000.

[FT79] C. Fennema and W. Thompson. Velocity determination in scenes containing several moving objects. *CGIP*, 9 :301–315, 1979.

[GB00] M. Gelgon and P. Bouthemy. A region-level motion-based graph representation and labeling for tracking a spatial image region. *Pattern Recognition*, 33(4) :725–745, 2000.

[GBR03] F. Galland, N. Bertaux, and Ph. Réfrégier. Minimum description length synthetic aperture radar image segmentation. *IEEE Image Processing*, 12(9) :995–1006, September 2003.

[GMP96] S. Gil, R. Milanese, and T. Pun. Combining multiple motion estimates for vehicle tracking. In *4th European Conf. on Computer Vision*, pages 307–320, 1996.

[Gov03] G. Govaert. Analyse de données et data mining, March 2003.

[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*, pages 152–157, 1994.

[HNR84] Y.Z. Hsu, H.H. Nagel, and G. Rekers. New likelihood test methodes for change detection in image sequences. *Computer Vision, Graphics and Image Processing*, 26 :73–106, 1984.

[Hot33] H. Hotelling. Analysis of a complex statistical variable into principal component. *Edu. Psy*, 24 :417–441 and 498–520, 1933.

[HS81] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17 :185–203, 1981.

[HS98] C. Harris and M. Stephens. A combined corner and edge detector. In *4th Alvey Vision Conference*, pages 189–192, Manchester, August 1998.

[Hub81] P.J. Huber. *Robust statistics*. Wiley, New York, 1981.

[HW87] N. Hoose and L.G. Willumsen. Automatically extracting traffic data from video-tape using the clip4 parallel image processor. *Pattern Recognition Letters*, 6(3) :199–213, 1987.

[IA98] M. Irani and P. Anandan. A unified approach to moving object detection in 2d and 3d scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(6) :577–589, 1998.

[IB96] M. Isard and A. Blake. Contour tracking for stochastic propagation of conditional density. In *4th European Conf. on Computer Vision*, 1996.

[IBB+03] D. Izquierdo, J. Becerra, Y. Berthoumieu, M. Donias, and Ph. Marchegay. Segmentation multi - descripteurs de scènes autoroutières. In *Compression et Représentation des Signaux Audiovisuels, CORESA*, Lyon, France, 2003.

[IBM02a] D. Izquierdo, Y. Berthoumieu, and Ph. Marchegay. High and low level object description for video tracking process. In *European Signal Processing Conference, EUSIPCO*, Toulouse, France, Sept. 2002.

[IBM02b] D. Izquierdo, Y. Berthoumieu, and Ph. Marchegay. Region level segmentation based on a derivative approach for video tracking process. In *IEEE Internation Conference on Image Processing, ICIP*, Rochester, USA, September 2002.

[IBM02c] D. Izquierdo, Y. Berthoumieu, and Ph. Marchegay. Système automatique de vidéo surveillance de scènes autoroutières. In *Conférence Internationale Francophone d'Automatique, CIFA*, Nantes, France, 2002.

[IBM03a] D. Izquierdo, Y. Berthoumieu, and Ph. Marchegay. Automatic surveillance system for real traffic video sequences. In *Intelligent Transport Systems, ITS*, Madrid, Spain, 2003.

[IBM03b] D. Izquierdo, Y. Berthoumieu, and Ph. Marchegay. Spatio-temporal segmentation for real rigid object tracking. In *Workshop COST 276*, Bordeaux, France, 2003.

[IRP92] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In *Proc. of 2nd European Conference on Computer Vision*, volume 588, pages 282–287, 1992.

[Jai81] R. Jain. Dynamic scene analysis using pixel based processes. *IEEE Trans. on Computers*, 14(8) :12–18, 1981.

[Jai84] R. Jain. Difference and accumulative difference pictures in dynamic scene analysis. *Image and Vision Computing*, 2(2) :99–108, 1984.

[JMA79] R. Jain, W.N. Martin, and J.K. Aggarwal. Segmentation through the detection of changes due to motion. *Computer Graphics and Image Processing*, 11 :13–34, 1979.

[JMN77] R. Jain, D. Militzer, and H.H. Nagel. Separating non-stationary form stationary scene components in a sequence of real world tv-images. In *Proc. of the 5th International Joint Conference on Artificial Intelligence*, pages 612–618, 1977.

[JN79] R. Jain and H.H. Nagel. On the analysis of accumulative difference pictures form image sequence of real world scenes. *IEEE Trans. on Pattern Analyasis and Machine Intelligence*, 1(2) :206–214, 1979.

[Joh73] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychologics*, 14(2) :201–211, 1973.

[KB90] K. Karmann and A. Brandt. *Time-Varying Image Processing and Moving Object Recognition, 2*, chapter Moving Object Recognition Using an Adaptive Background Memory, in V. Elsevier, Amsterdam, 1990.

[KBG90] K.P. Karmann, A. Von Brandt, and R. Gerl. Moving object segmentation based on adaptative reference images. *Signal Processing*, 5 :951–954, 1990.

[KH95] C. Kervrann and F. Heitz. A markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(1) :69–73, 1995.

[KIH+81] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro. Motion-compensated interframe coding for video conferencing. In *National Telecommunications Conference*, volume 4, pages G5.3.1–G5.3.5, New Orleans, December 1981.

[Kim98] S.-H. Kim. Object oriented face detection using range and color information. In *Automatic Face and Gesture Recognition*, pages 14–16, April 1998.

[Kol] Julian Kolodko. Real time dynamic obstacle reasoning. Ph.D. Cadidature Report.

[KSUS] T. Kurita, H. Shimai, S. Umeyama, and T. Shigehara. Estimation of background from image sequence with moving objects.

[KT94] J. Konrad and P. 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, 1994.

[KWM94] D. Koller, J. Weber, and J. Malik. Robust multiple car tracking with occlusion reasoning. In *European Conference on Computer Vision*, volume I, pages 189–196, 1994.

[LFP98] A. Lipton, H. Fujiyoshi, and R. Patil. Moving target classification and tracking from real-time video. In *IEEE Image Understanding Workshop*, pages 129–136, 1998.

[LG83] R. Lenz and A. Gerhard. Image sequence coding using scene analysis and spatio-temporal interpolation. In *Image Sequence Processing and Dynamic Scene Analysis*, NATO ASI series, pages 264–276. T.S. Huang ed., 1983.

[LL02] F. Luthon and M. Liévin. Entropy power for thresholding technique in image processing. In *EUSIPCO*, Toulouse, September 2002.

[LT90] W. Long and Y.H. Tang. Stationary background generation : an alternative to the difference ot two images. *Pattern recognition*, 23(12) :1351–1359, 1990.

[Mag02] D. Magee. Tracking multiple vehicles using foreground, background and motion models. In *ECCV Workshop on Statistical Methods in Video Processing*, 2002.

[Mak96] A. Makarov. Comparison of background extraction based intrusion detection algorithms. In *IEEE International Conference on Image Processing*, pages 521–524, 1996.

[Mar82] D. Marr. *Vision.* Freeman, San Francisco, 1982.

[MB94] F. Meyer and P. Bouthemy. Region-based tracking using affine motion models in long image sequences. *CVGIP : Image Understanding*, 60(2) :119–140, 1994.

[MBD95] M. Maurizot, P. Bouthemy, and B. Delyon. Determination of angular points in 2d deformable flow fields. In *Proc. of 2nd IEEE Int. Conf. of Image Processing*, pages 746–747, 1995.

[MDK96] F. Moscheni, F. Dufaux, and M. Kunt. Object tracking based on temporal and spatial information. In *Proc. of 3rd IEEE Int. Conf. on Image Processing*, 1996.

[Mec89] A. Mecocci. Moving object recognition and classification in external environments. *Signal Processing*, 18 :183–194, 1989.

[MF90] V. Markanday and B.E. Flinchbaugh. Multispectral constraints for optical flow computation. In *Proc. 3rd Int. Conf. on Computer Vision*, pages 38–41, 1990.

[MM97] F. Marques and C. Molina. Object tracking for content-based functionalities. In *SPIE Visual Communication and Image Processing*, volume 3024, pages 190–198, 1997.

[MP98] E. Mémim and P. Pérez. Dense estimation and object-based segmentation of the optical flow with robust techniques. *IEEE Trans. on Image Processing*, 7(5) :703–719, 1998.

[MWA87] A. Mitiche, Y.F. Wang, and J.K. Aggarwal. Experiments in computing optical flow with the gradient-based, multiconstraint method. *Pattern Recognition*, 20(2) :173–179, 1987.

[Nag76] H.H. Nagel. Experiences with yakimovsky's algorithm for boundary and object detection in real world images. In *Proc. of 3rd Internation Conference on Pattern Recognition*, pages 753–758, 1976.

[Nag78] H.H. Nagel. Formation of an object concept by analysis of systematic time variations in the optically perceptible environment. *Computer Graphics and Image Processing*, 7 :149–194, 1978.

[Nag83] H.H. Nagel. Displacement vectors derived form second-order intensity variations in image sequences. *Computer vision, Graphics and Image Processing*, 21 :85–117, 1983.

[Nag89] H.-H. Nagel. On a constraint equation for the estimation of displacement rates in images sequences. *IEEE Trans. PAMI*, 11 :13–30, 1989.

[Nak85] K. Nakayama. Biological image motion processing : a review. *Vision Research*, 25(5), 1985.

[OB94] J.M. Odobez and P. Bouthemy. Detection of multiple moving objects using multiscale mrf with camera motion compensation. In *IEEE Int. Conference on Image Processing*, pages 257–261, Texas, November 1994.

[OB95] J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4) :348–365, December 1995.

[OB97] J.M. Odobez and P. Bouthemy. *Video Data Compression for Multimedia Computing*, chapter 8, pages 295–311. H.H. Li, S. Sun and H. Derin ed., Kluwer, 1997.

[OB98] J.M. Odobez and P. Bouthemy. Direct incremental model-based image motion segmentation for video analysis. *Signal Processing*, 6(2) :143–155, 1998.

[Odo94] J.M. Odobez. *Estimation, détection et segmentation du mouvement : une approche robuste et markovienne.* PhD thesis, Université de Rennes 1, 1994.

[Pat98] S. Pateaux. *Segmentation spatio-temporelle et codage orienté-régions de séquences vidéo basés sur le formalisme MDL.* PhD thesis, Université de Rennes 1, 1998.

[PD98] N. Paragios and R. Deriche. A pde-based level-set approach for detection and tracking of moving objects. In *Proc. of IEEE International Conf. on Computer Vision*, pages 1139–1145, 1998.

[Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space, 1901.

[Pla85] T. Plassard. *Extraction de paramètres caractéristiques dans une séquence d'images de particules passant de l'état de brouillard à l'état d'agglomérats. Application à la génération automatique d'alarme en milieu industriel.* PhD thesis, Université de Rennes 1, 1985.

[PT03] V. Popovici and J.P. Thiran. Face detection using an svm trained in eigenfaces space. Technical report, Signal Processing Institute, Swiss Federal Institute of Technology Lausanne, 2003.

[RWBM96] D. Reynard, A. Wildenberg, A. Blake, and J. Marchant. Learning dynamics of complex motions from image sequences. In *4th European Conf. on Computer Vision*, pages 357–358, 1996.

[SA96] H. Sawhney and S. Ayer. Compact representation of videos through dominant and multiple motion estimation. *Pattern Analysis and Machine Intelligence*, 18(8), 1996.

[SG99] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR99*, volume 2, pages 246–252, 1999.

[SSdIL03] P. Smith, M. Shah, and N. de Itoria Lobo. Determining driver visual attention with one camera, December 2003.

[STEK96] E. Saber, A.M. Tekalp, R. Eschbach, and K. Knox. Automatic image annotation using adaptive color classification. *Graphical Models and Image Processing*, 58(2) :115–126, 1996.

[Tan82] I.S. Tang. Extraction of moving objects in textured dynamic scenes. In *IEEE Conference Publication*, 1982.

[TC] K. Takaya and K.Y. Choi. Detection of facial components in a video sequence by independent component analysis.

[TDA98] J.C. Terrillon, M. David, and S. Akamatsu. Automatic detection of human faces in natural scene images by use of skin color model and of invariants moments. In *Third Internation Conference on Automatic Face and Gesture Recognition*, pages 112–117, Nara (Japan), 1998.

[TEST96] C. Toklu, A.T. Erdem, M.I. Sezan, and A.M. Tekalp. Tracking motion and intensity variations using hierarchical 2d mesh modeling for synthetic object transfiguration. *Graphical Models and Image Processing*, 58(6) :553–573, 1996.

[TK91] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.

[TKO99] J. Tang, S. Kawato, and J. Ohya. Face detection from complex background. In *International Workshop on Very Low Bitrate Video Coding*, pages 29–30, Kyoto Reasearch Park, October 1999.

[TM93] P.H.S. Torr and D.W. Murray. Statistical dectection of independent movement from a moving camera. *Image and Vision Computing*, 11(4) :79–88, 1993.

[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, 1984.

[TP90] W.B. Thompson and T.C. Pong. Detecting moving objects. *Internation Journal of Computer Vision*, 4(1) :39–57, 1990.

[TSL+00] A. Tourapis, G. Shen, M. Liou, O. Au, and I. Ahmad. A new predictive diamond search algorithm for block based motion estimation. In *Visual Comm. and Image Proc.*, Perth, June 2000.

[Van97] P. Vannorenberghe. *Détection de mouvement par analyse de séquences d'images monoculaires. Application à l'estimation de flux de piétons en milieu urbain.* PhD thesis, Université du Littoral, 1997.

[Vez02] V. Vezhnevets. Face and facial feature tracking for natural human-computer interface. Technical report, Department of Applied Mathematics and Computer Science pf Moscow State University, 2002.

[VMBP96] P. Vannoorenberghe, C. Motamed, J.M. Blosseville, and J.G. Postaire. A motion detection for non-rigid objects. application to pedestrians monitoring in an urban environment. In *IEEE International Conference on IMACS, Computationnal Engineering in Systems Applications*, pages 438–442, Lille, France, July 1996.

[WA94] J.Y.A. Wang and E.H. Adelson. Representing moving images with layers. *IEEE Trans. on Image Processing*, 3(5) :625–638, 1994.

[WADP97] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder : real-time tracking of the human body. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7) :780–785, July 1997.

[WBPDB96] L. Wu, J. Benois-Pineau, Ph. Delagnes, and D. Barba. Spatiptemporal segmentation of image sequences for object-oriented low bit rate image coding. *Signal Processing : Image Communication*, 8 :513–543, 1996.

[WD84] R.C. Waterfall and K.W. Dickinson. Image processing applied to traffic. 2 : Practical experince. *Traffic Engineering and Control*, pages 60–67, 1984.

[Wen83] O.S. Wenstop. Motion detection for image information. In *Scandinavian Conference on Image Analysis*, pages 381–386, Tromso, Norway, July 1983.

[WG87] J. Wiklund and G.H. Granlund. Image sequence analysis for object tracking. In *Proc. of 5th Scandinavian Conference on Image Analysis*, pages 641–648, 1987.

[XG94] W. Xiong and C. Graffigne. A hierarchical method for detection of moving objects. In *Proc. of 1st IEEE Int. Conf. on Image Processing*, pages 795–799, 1994.

[Yak76] Y. Yakimovsky. Boundary and object detection in real world images. *Journal of the ACM*, 23(4) :599–618, 1976.

[YAT81] M. Yachida, M. Asada, and S. Tsuji. Automatic analysis of moving images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 3(1) :262–279, 1981.

[ZB95] H. Zheng and D. Blostein. Motion-based object segmentation and estimation using the mdl principle. *IEEE Trans. on Image Processing*, 4(9) :1223–1235, 1995.

[ZF92] Z. Zhang and O. Faugeras. Three-dimensional motion computation and object segmentation in a long sequence of stereo frames. *Int. Journal of Computer Vision*, 7(3), 1992.

Résumé

Dans cette thèse, nous nous intéressons à l'étude des procédés numériques en vue du développement d'un système automatique de *suivi* multi - objets sur la base d'un flux continu d'images. Le procédé complet proposé décompose la tâche de suivi en une phase de localisation complétée d'une phase de reconnaissance de chacun des objets tout au long de la séquence vidéo. Les outils développés permettent l'analyse des séquences prises à l'aide d'une caméra statique en extérieure. Afin d'assurer la tâche de localisation/reconnaissance au sens de la mise en correspondance nous avons développé les quatre étapes suivantes :

- la *détection* qui est la mise en œuvre de méthodes de segmentation générant des entrées perceptives permettant d'initialiser le procédé de suivi. Généralement, la primitive caractérisant au mieux les objets est le mouvement,
- la *reconnaissance* qui a pour objectif de comparer les résultats obtenus lors de la phase de détection (bas niveau) à une description haut niveau de l'objet (modèle),
- l'*estimation* qui permet une mise à jour des descripteurs de l'objet, ensemble de caractéristiques définissant le modèle décrivant les objets,
- Et la *prédiction* qui prolonge l'évolution de l'objet, notamment, en terme de position.

Deux applications, la première consacrée au suivi de véhicules, et la seconde au suivi de visages, vont permettre d'évaluer les performances des méthodes proposées pour chacune des quatre étapes afin de valider le procédé complet.

Mots-clés
Suivi spatio - temporelle, détection de mouvement, mise en correspondance, reconnaissance d'objets, estimation de mouvement, points caractéristiques.

Abstract

In this Ph.D., we are interested on the digital techniques focusing an automatic tracking system for multiple objects. The tracking process is decomposed on a localisation phase and a recognition phase for every object over the image sequence. The tools let image processing over an image sequence with static camera on outdoor scene. Assuring the localisation/recognition tasks we have developed four different steps:

- *detection*: implementation of a segmentation process generating the input necessary for the tracking initialisation. The most typical primitive characterising objects is the movement,
- *recognition*: comparison of the low level input (the detection phase results) with a high level object representation (model),
- *estimation*: update of the object model,
- *prediction*: knowledge of the object evolution, in terms of position.

Two different applications, a vehicle tracking application and a face tracking application, let show multiple results validating the four proposed steps.

Key-words
Spatio-temporal tracking, movement detection, matching process, object recognition, movement estimation, feature points.