Laboratoire d'InfoRmaique
en Image et Systèmes d'information
UMR 5205 CNRS/
INSA de Lyon/Université Claude Bernard Lyon 1/
université Lumière Lyon 2/Ecole Centrale de Lyon

Christine LEIGNEL
Jean-Michel JOLION

20-01-2008

Projet ANR CAnADA

« Comportements Anormaux Analyse Détection Alerte »

Étude bibliographique suivi de personnes et trajectoire dans un réseau de caméras

Les machines un jour pourront résoudre tous les problèmes, mais jamais aucune d'entre elles ne pourra en poser un.
Albert Einstein

Table des matières

Introduction

Dans le cadre d'une absence d'offre technologique en matière de détection en temps réel à partir de la vidéo des comportements anormaux de personnes dans un lieu public, tel un lieu de vente, des industriels comme YOUG'S et Thales expriment ce besoin.

Le but du projet CAnADA « Comportements Anormaux : Analyse, Détection, Alerte » est de proposer une approche pour la détection en temps réel de comportements inhabituels pouvant mettre en péril la sécurité des personnes et des biens dans des lieux publics, comme les centres commerciaux, les magasins, les métros. Les informations détectées seront transmises à une application capable de rendre en temps réel une alarme et de ramener la situation à un niveau normal via un affichage par exemple. Dans ce cadre, un réseau de caméras est mis en place, comportant à certains endroits de la scène des zones aveugles dont il faudra tenir compte (un individu peut se cacher dans une telle zone afin de définir une stratégie de vol, hors des caméras). Les traitements mis en place consistent à extraire les trajectoires des personnes, ainsi que leurs activités, en tenant compte du contexte de la scène (disposition des caméras et des objets de la scène), et en traitant les cas d'occultations (une personne cachant une autre personne tout ou partie, ou bien un objet cachant un membre d'une personne). Les zones du visage des personnes suivies doivent être masquées, car il ne faut pas avoir accès à l'identité des personnes, le partenariat CNIL nous guidant dans cet aspect.

Plusieurs partenaires scientifiques, juridiques et industriels sont regroupés dans ce consortium, couvrant ainsi des compétences complémentaires :
-Le LIRIS (Laboratoire d'InfoRmatique en Images et Systèmes d'information), INSA de Lyon, et la société FOXTREAM, tous deux spécialiste dans l'analyse des objets en mouvement, la gestion des occultations, la reconnaissance de visages, et l'indexation des données vidéo;
-Le LIFL (Laboratoire d'Informatique Fondamentale de Lille) - TÉLÉCOM LILLE 1, pour la fouille de données, et l'analyse des situations à un niveau sémantique;
-ARMINES-EMD (Centre Commun Ecole des Mines de Douai) pour le suivi de trajectoires multiples en temps-réel, et analyse « bas-niveau » des séquences de mouvements;
-URECA (UFR de Psychologie, Université de Lille 3) pour l'interprétation des comportements individuels et collectifs;
-IREENAT (Institut de Recherches sur l'Evolution de l'Environnement Normatif des Activités Transnationales Université de Lille 2) pour l'analyse des problèmes juridiques;
-Les partenaires industriels, YOUG'S et Thales sont une interface avec les industriels potentiellement intéressés par le projet.

Depuis quelques années, la vision par ordinateur témoigne d'un intérêt croissant, d'une part du fait des technologies meilleur marché et de plus en plus compétentes, et d'autre part des besoins en sécurité et télésurveillance qui ont vu le jour depuis les évènements du 11 septembre 2001.

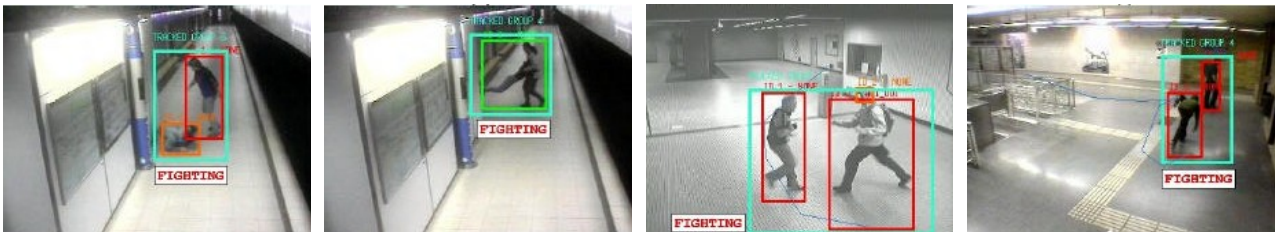
Traditionnellement, le flux vidéo était traité par un opérateur humain, remplacé progressivement par un traitement automatique sur les données enregistrées contenant des évènements anormaux. Actuellement, l'objectif est de détecter ces évènements en temps réel et de façon automatique.

L'analyse de la vidéo de façon automatique est centrée sur la détection des situations anormales dans diverses activités, en surveillance du trafic routier, pour la détection de congestion, la détection d'accidents, et dans la sécurité des personnes comme la délinquance, la détection des colis suspects. Pour la détection de colis dangereux, seule une détection de mouvement est nécessaire. En revanche, s'il s'agit de reconnaître des comportements tels qu'une agression, il faudra une interprétation haut niveau de la scène. Les systèmes de vidéo surveillance peuvent être totalement automatisés et servir au déclenchement d'actions externes selon les observations, comme la régulation du trafic routier, ou le déclenchement d'alarmes. Les systèmes bas niveau de détection sont de moins en moins usités au profit de systèmes de reconnaissance de situations anormales ou

dangereuses, en analysant les comportements de la foule ou entre des personnes, et faisant appel à la coopération de divers modules bas niveau.

A titre d'exemple, en surveillance d'activité humaine, [Chleq N., Thonnat M.] ont réalisé un système d'aide à la décision d'opérateur de vidéo surveillance, déclenchant une alarme dans une situation à risque. [Nagel H.-H] a réalisé le même genre de système mais en surveillance routière. [Choi S., Seo Y., Kim H., Hong K.] analysent les scènes sportives d'une équipe de football. [Pentland A.] réalise un système de compréhension du langage des sourds et muets grâce à l'analyse de gestes. Les applications sont diverses et la demande est de plus en plus fournie.

Pour la détection de comportements dangereux dans les métros, [Cupillard F., Avanzi A., Bremond F., Thonnat M.] propose une approche avec plusieurs caméras pour reconnaître des personnes isolées, des groupes de personnes ou la foule (cf. figure 1). Cet exemple peut être utilisé comme introduction aux étapes clefs de l'analyse de séquences en vue de détection de comportements. Trois composantes définissent ce système :
1. détection de mouvement et suivi image par image;
2. combinaison de plusieurs caméras;
3. suivi long terme de une ou plusieurs personnes.
Pour chacun des acteurs suiveur, le module de reconnaissance possède trois niveaux de reconnaissance: états, évènements et scénarios.



(a) Personne couchée au sol | (b) Variation dans la largeur du groupe | (c) Séparation de personnes dans un groupe | (d) Variation dans la trajectoire du groupe

Figure 1. Chaque image représente une configuration de bagarre, nécessitant une reconnaissance par le système automatique [Cupillard F., Avanzi A., Bremond F., Thonnat M.].

La plupart des applications nécessitent une focalisation sur le mouvement de la personne humaine. [Johansson G.] a démontré dans les années 1970 qu'on peut reconnaître des personnes familières par leur démarche grâce aux lumières fixées sur leurs articulations mais cela n'est pas possible dans un système non contraint et de plus ce système ne tient pas compte de l'apparence de la personne. Au vu du changement d'apparence, il faudrait de multiples représentations de la personne. La reconnaissance de visages changeant moins que la reconnaissance d'apparence, elle pourrait être combinée avec la reconnaissance de la marche pour identifier une personne. Contrairement à l'approche des MLD (« Moving Light Display ») mettant en évidence les articulations, l'approche choisie par [Lee L., Grimson W.E.L.] est celle de l'apparence de la marche pour la reconnaissance de personnes par leur démarche. La marche humaine est une primitive d'identification de personne déterminant son poids, la longueur de ses membres, et sa posture habituelle : elle peut être utilisée comme une mesure biométrique pour reconnaître des personnes connues et classer des sujets inconnus. Dans des situations pour lesquelles l'information de visage ou de regard n'est pas valable, la marche est une information disponible à basse résolution.

D'autres applications comme la téléconférence, l'indexation vidéo, la réalité virtuelle nécessitent un suivi robuste dans un environnement réel en temps réel. Tous ces domaines requièrent l'identification des parties du corps humain et l'estimation de la pose et des paramètres de mouvement. L'estimation de la pose d'une

personne dans une image fixe ou son suivi dans une séquence vidéo consiste à déterminer les coordonnées, dans le plan 2D ou dans l'espace 3D, des membres du corps dans chacune des images. Les méthodes utilisées pour l'estimation de la pose ou pour le suivi se classent en fonction du nombre de caméras (mono/stéréo), du modèle du corps, du nombre de personnes, de la nature stochastique ou déterministe des méthodes d'estimation. Les principales difficultés dans l'analyse du mouvement du corps humain proviennent de la nature 3D non rigide du mouvement, des changements de luminosité, des occultations entre membres du corps ou avec un objet de la scène, des changements de fond et de la nature parfois ample des vêtements. La plupart des approches existantes introduisent des simplifications soit par une approche basée modèle soit par des hypothèses sur les divers types de mouvements. Bon nombre de travaux sont basés sur des modèles non déformables pouvant approcher le corps humain , comme les cylindres généralisés, mais ils ne peuvent s'adapter aux différentes tailles du corps humain. Une alternative à cette limitation serait de segmenter l'image et de mettre en correspondance un modèle déformable avec les membres du corps humain issus de la segmentation *a priori* et définissant différentes dimensions anthropométriques. En deux dimensions, aucune technique n'existe pour acquérir de façon automatique un modèle 2D du corps humain. L'estimation de la pose et du mouvement en 3D n'est pas résolue à cause de la difficulté dans l'intégration de multiples points de vue et du traitement des occultations entre les différentes membres du corps.

Les techniques utilisées pour l'identification de personnes et la reconnaissance des activités sont classées en différentes catégories en fonction de la précision de l'analyse et de la résolution vidéo requise. D'un côté, les techniques caractérisées par une grande résolution vidéo et une faible quantité de texture dans la scène ont pour objectif de reconnaître une personne grâce à une grande base de données, sur la démarche par exemple. L'autre extrême est caractérisée par une basse résolution vidéo et des scènes très bruitées, desquelles il est alors souvent impossible d'obtenir des objets discriminants. Dans ce cas, les personnes sont détectées par leur présence, déterminée par le mouvement. Entre les deux extrêmes précédentes à base de « template » ou à base de « blob », la représentation par « blob » peut être raffinée à l'aide de modèles articulés hiérarchiques ([Niu W., Jiao L., Han D., Wang Y.-F.], [Black M.J., Jepson A.D.], [Collins R., et al.a], [Haritaoglu I., Harwood D., Davis L.S. 00]), permettant aux divers membres (tête, mains, bras, buste, torse) une identification individuelle spécifiant les activités de façon plus précise.

Le traitement de séquence d'images pour le suivi de personnes se divise en trois niveaux hiérarchisés (cf. figure 2), qui se distinguent soit par une approche ascendante soit par une approche descendante. Au bas niveau la détection, au niveau intermédiaire le suivi (**appariement temporel**) et au haut niveau la reconnaissance des actions ou des personnes. Le suivi est une étape intermédiaire entre la détection et la reconnaissance en vue de la description sémantique de la scène.

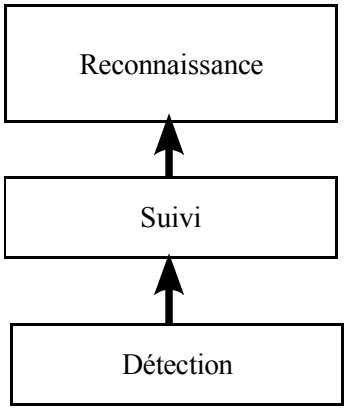


Figure 2 : Les trois niveaux hiérarchisés du traitement d'une séquence d'images

Le **chapitre 1** est consacré à l'estimation du mouvment et au suivi de trajctoires. Aprés la définition du mouvement réel, apparent et estimé dans la section 1, deux sections sont dédiées à l'estimation de mouvment, soit par des modèles paramétriques (section 2), soit par des modèles non paramétriques (section 3) avec une approche de mise en correspondance ou Markovienne. La section 4 reprend les différentes méthodes de suivi de trajectoires, de façon déterministe ou bien probabiliste. Nous concluons ce chapitre par deux exemples d'applications probabilistes à base de graphe.

Le **chapitre 2** expose les différentes méthodes de suivi, soit dans la séquence entière, soit image par image avec les approches basées modèle du corps en 2D ou en 3D (section 1). Dans la section 2, les approches d'extraction de caractéristiques sont passées en revue. La section 3 présente quelques travaux dont ceux de Nicolas Thome qui a effectué sa thèse (2003-2007) dans la société FOXTREAM à Lyon.

Le **chapitre 3** met en exergue les différents systèmes de vidéo surveillance à l'heure actuelle, tant chez les académiques que chez les industriels (sections 1 à 6, 9 et 10). Dans la section 7, les travaux effectués par Ionel Pop chez Nagel dans le cadre du suivi de piétons dans un réseau routier sont développés. Le suivi des trajectoirs des tâches de couleur est explicité à la section 8. Ces travaux ont été réalisés par Rémi Megret au cours de sa thèse (2000-2003) au sein de notre laboratoire LIRIS INSA de Lyon.

Nous terminons par le **chapitre 4**, plus adapté à notre problématique de suivi multi caméras. Une première section effectue une introduction sur le suivi dans un réseau de caméras. La section 2 propose de suivre, dans un cadre mono caméra ou stéréo, le haut du corps d'une personne à l'aide de filtres à particules à travers un réseau bayésien. Chaque noeud du réseau est alors attaché à un membre du corps humain. De façon analogue, un réseau bayésien peut être modélisé pour suivre une personne entière dans un réseau de caméras, chaque noeud du réseau étant lié à une caméra (et non plus à un membre du corps). La section 3 regroupe les travaux de recherche de Jean-Charles Noyer en vue de l'obtention de la H.D.R à l'Université du Littoral Côte d'Opale, Laboratoire d'Analyse des Systèmes du Littoral. Il s'agit de fusionner des informations issues de plusieurs capteurs pour l'estimation de la structure d'un objet et la détection de son mouvement. Une partie de ce travail de recherche, présentée dans la dernière sous-sectione est consacrée à la fusion multi capteurs par filtrage « particulaire » pour la reconstruction 3D, l'estimation du mouvement 3D et le suivi d'objets 3D.

Enfin, la **conclusion** propose une approche multi caméras, dans un réseau bayésien, chaque caméra représentée par un noeud du réseau. Les messages sont envoyés d'une caméra à l'autre par propogation de croyance, symbolisant la croyance qu'une personne vue dans une caméra puisse se trouver un instant plus tard dans le champ de l'autre caméra, en fonction de la configuration des caméras, et de l'analyse de la scène.

Chapitre 1 – Estimation de mouvement

1 Mouvement réel, mouvement apparent, mouvement estimé

Le mouvement dans une séquence d'images en 2D est perceptible grâce aux variations des intensités lumineuses.
Le mouvement réel et le mouvement observé sont souvent différents dans une image. Les images représentent la projection du monde réel 3D. Le mouvement observé à partir des changements de la distribution spatiale d'intensité lumineuse entre plusieurs images de la séquence, dit mouvement apparent, est la projection du mouvement réel 3D dans le plan de l'image 2D. On parle aussi de « **flot optique** » [Horn B.K.P, Schunk B.G.] ou de « **champ de vitesses** » pour désigner le champ des vitesses apparentes. Le champ de « **mouvement apparent** » s'appelle aussi « **mouvement projeté** » [Aggarwall J.K., Nandhakumar N.] du fait qu'il représente la projection du mouvement 3D dans le plan image. L'objectif de l'estimation de mouvement est d'estimer le champ de mouvement 2D ou 3D à partir d'une séquence d'images 2D ou 3D évoluant au cours du temps. Il y a donc un mouvement réel, un mouvement observé dit apparent et un mouvement estimé.
Le mouvement « apparent » ne correspond pas toujours au mouvement réel projeté dans le plan de l'image. En effet, les vitesses apparentes des points situés sur une sphère uniforme en rotation sont nulles, ce qui n'est pas le cas des projections des vraies vitesses de ces points. La présence de variations lumineuses (ombres, flash, etc.) non dues au mouvement entraîne un mouvement apparent différent du mouvement réel. Le mouvement apparent est donc une combinaison de deux sources de mouvement : Le mouvement propre des objets en 3D dans la scène, et le mouvement de la caméra.
Le vecteur déplacement estimé correspondant au déplacement d'un point dans le plan image, est défini par le champ de mouvement apparent, c'est-à-dire par les variations locales d'intensité lumineuse. Le vecteur vitesse estimé correspond à la variation temporelle du déplacement par unité de temps. Dans une séquence d'images, il n'est possible que d'estimer le champ de mouvement (déplacement ou vitesse) apparent et non le champ de vitesse réel. Le champ de déplacement est le champ de vecteurs déplacement estimé, il en va de même pour le champ de vitesse et le champ de vecteurs vitesse estimé.
Pour estimer le mouvement à partir du champ de mouvement apparent, il faut faire l'hypothèse que l'intensité lumineuse reste constante au cours du mouvement [Horn B.K.P, Schunk B.G.]. Cette hypothèse de conservation de l'intensité lumineuse en chaque point de la trajectoire du mouvement s'exprime par l'équation des différences entre les images déplacées (« **DFD** » **Displaced Frame Difference**) entre deux instants successifs : $DFD=I(x+d_x, y+d_y, t+\Delta t)-I(x, y, t)=0$, avec I(x, y, t) l'intensité au point (x, y) à l'instant t.
En estimation de mouvement, on utilise l'estimation arrière ou inverse. L'estimation avec compensation avant ou directe du mouvement est utilisée dans la compression prédictive des séquences d'images.

Un champ de mouvement estimé (le « **flot optique** ») est caractérisé soit par le champ des vecteurs vitesse soit par le champ des vecteurs déplacement ou de correspondance, ce sont deux approches similaires si l'intervalle de temps entre deux images est court et constant. C'est pourquoi on s'intéresse à l'estimation du champ de vecteurs déplacement. **L'estimation de mouvement est un problème mal-posé car il n'a pas toujours de solution dans le cas d'une occultation, et s'il en a une, elle n'est pas toujours unique à cause du problème d'ouverture.**

1.1 L'occultation

L'occultation est le recouvrement ou le non-recouvrement d'une surface, à cause de la translation ou rotation d'un objet dans le champ. Une surface recouverte, en général le fond d'une image, correspond à une zone recouverte par un objet à un instant donné, donc les pixels d'une région qui sera recouverte à l'image suivante n'auront pas de correspondant dans l'image suivante (cf. figure 3). A l'inverse, la région du fond découverte par l'objet en mouvement aura ses pixels qui n'auront pas de correspondant dans l'image précédente (cf. figure 3).

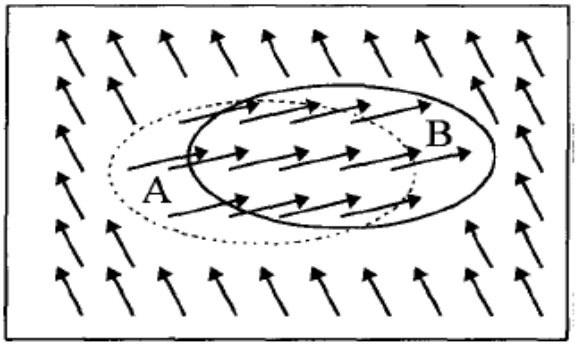


Figure 3 : Champ de vecteurs obtenu par mise en correspondance avec recherche « en avant ». Aucun vecteur ne « pointe » vers la zone découverte (A) et aucun vecteur ne « sort » de la zone recouverte (B) [Orkisz M., Clarysse P.].

1.2 Le problème de l'ouverture

Le problème de l'ouverture indique que seule la composante normale au déplacement est mesurable, c'est-à-dire seule celle orthogonale au contour local de l'image, orientée dans la direction du gradient spatial de l'intensité, au point considéré. Supposons un objet dont l'un des coins est en mouvement dans la direction verticale haut. Il n'est possible de déterminer si l'objet est en mouvement dans la direction verticale supérieure ou dans la direction normale au bord de l'objet (cf. figure 4).

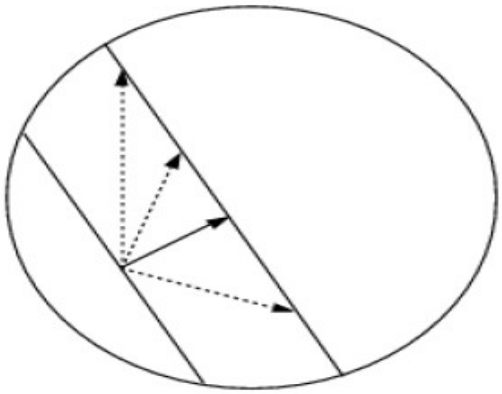


Figure 4 : Le problème de l'ouverture : seule la composante du mouvement parallèle au gradient d'intensité, orthogonale au contour est déterminée [Ricquebourg Y. 97].

Ce problème est dû à l'hypothèse de conservation de l'intensité lumineuse ou de la luminance [Horn B.K.P, Schunk B.G.] $I(x+d_x, y+d_y, t+\Delta t)=I(x, y, t)$. Le développement en série de Taylor conduit à l'équation :

$\frac{(\partial I)}{(\partial x)}(x, y, t)\cdot u(x, y, t)+\frac{(\partial I)}{(\partial y)}(x, y, t)\cdot v(x, y, t)+\frac{(\partial I)}{(\partial t)}(x, y, t)=0$ soit l'équation de contrainte du mouvement apparent ECMA, appelée aussi équation du flux optique (EFO) $\nabla I^t . w+\frac{(\partial I)}{(\partial t)}=0$ avec ∇I le gradient spatial et $w=(u, v)$, $u=\frac{d_x}{d_t}$, $v=\frac{d_y}{d_t}$. Donc l'équation ECMA peut s'écrire

$\frac{(\partial I)}{(\partial x)} \cdot u + \frac{(\partial I)}{(\partial y)} \cdot v + \frac{(\partial I)}{(\partial t)} = 0$. Elle relie les gradients spatio-temporels au vecteur vitesse apparente.
Nous avons 2 inconnues u et v pour une seule équation, donc un **problème mal posé**.

Seule la projection du vecteur vitesse dans la direction du gradient spatial de l'intensité est déterminée. Cette projection est localement perpendiculaire aux frontières photométriques, c'est la « composante normale » du vecteur vitesse. Pour la composante « tangentielle », il faut régulariser l'estimation, c'est-à-dire introduire une contrainte supplémentaire afin de réduire l'espace des solutions. Il s'agit de la « contrainte de continuité » dite « de lissage » du champ de vitesses, exprimant que les points voisins sont animés de mouvements « très semblables ». La régularisation conduit alors à la minimisation d'une fonctionnelle comportant un « terme d'attache aux données » et un « terme de lissage », pondérés par un coefficient de régularisation $\alpha > 0$.

Pour résoudre ce problème d'indétermination, [Horn B.K.P, Schunk B.G.] ont proposé une méthode consistant à minimiser une énergie de la forme [Jehan-Besson S.] :

$$\min_{\mathbf{w}} E(\mathbf{w})$$
$$E(\mathbf{w}) = \int_{\Omega} \underbrace{\left(\nabla I^T \mathbf{w} + \frac{\partial I}{\partial t} \right)^2}_{\text{Terme d'observation}} dxdy$$

$E(w)$ est le **terme d'attache aux données**, traduisant la première hypothèse de conservation de l'intensité. Il s'agit donc de minimiser une fonctionnelle par rapport aux vecteurs mouvements $\min_{w} E(w)$, et seule la composante normale peut être extraite ici.

La seconde énergie à minimiser correspond à un **terme de lissage**, une contrainte supplémentaire de lissage qui suppose que tous les points voisins ont un mouvement semblable [Jehan-Besson S.] :

$$\min_{\mathbf{w}} E(\mathbf{w})$$
$$E(\mathbf{w}) = \int_{\Omega} \left(\nabla I^T \mathbf{w} + \frac{\partial I}{\partial t} \right)^2 + \underbrace{\alpha \left(|\nabla u|^2 + |\nabla v|^2 \right)}_{\text{Contrainte sur la régularité du flot optique}} dxdy$$

La contrainte de régularité du flot optique représente les gradients horizontaux et verticaux de la vitesse apparente.
C'est une énergie de connaissance *a priori* sur le champ de déplacement, ou terme de régularisation, qui contraint le problème pour le rendre bien-posé. Il s'agit de la seconde hypothèse : la cohérence spatiale des vecteurs mouvement.

A cause de la sous-détermination du problème de l'estimation de mouvement (problème de l'ouverture), nous avons vu qu'il faut introduire des contraintes supplémentaires sur le champ de mouvement [Stiller C., Konrad J.] mais des modèles paramétriques ou non paramétriques de ce champ peuvent aussi lever l'indétermination.

2 Modèles paramétriques pour l'estimation de mouvement

L'approche avec la contrainte de lissage fait l'hypothèse que les points lissant le champ appartiennent à un même objet. Ceci n'est pas le cas en présence de **discontinuité du mouvement**, c'est-à-dire dans le cas d'objets

différents ayant des mouvements indépendants, ou dans le cas d'un même objet articulé présentant des parties avec des mouvements différents (un humain par exemple).
On fait l'hypothèse que les objets en mouvement correspondent à des régions homogènes du point de vue photométrique, donc que des discontinuités de mouvement correspondent avec des frontières photométriques.
L'estimation du mouvement de tous les contours, donc aussi des discontinuités de mouvement est décrite dans ([Bouthemy P. 87], [Bouthemy P. 88], [Bouthemy P. 89]) qui exploite la dimension temporelle d'une séquence d'images, en modélisant un contour en mouvement par une portion de surface dans l'espace-temps. La détection des contours et l'estimation de leurs mouvements se font simultanément, par un test du rapport de vraisemblance de ces deux hypothèses.
La **segmentation au sens du mouvement** en régions homogènes recherche des zones de continuités ou discontinuités comme contours des régions délimitées. La vitesse apparente étant une variable non observable (cachée), la segmentation du champ estimé s'appuie sur l'information de composante normale du vecteur vitesse, ainsi que sur un modèle paramétrique du mouvement. Le modèle paramétrique le plus souvent utilisé est le modèle linéaire 2D affine [Bouthemy P. 87] :

$$v=\begin{bmatrix}u\\v\end{bmatrix}=\begin{bmatrix}p\\q\end{bmatrix}+\begin{bmatrix}a & b\\c & d\end{bmatrix}.\begin{bmatrix}x-x_0\\y-y_0\end{bmatrix}=T+M\cdot(x-x_0) \quad ,$$

avec x_0 le point de référence du mouvement, par exemple le centre de gravité de la région considérée, le vecteur T la translation parallèle au plan de l'image, la matrice M une combinaison de rotation, homothétie et déformation. Le plus simple des modèles est un modèle constant : tous les points d'une région effectuent un même mouvement de translation, décrit par le vecteur T , la matrice M **étant alors nulle** [Bouthemy P. 87].
Le modèle de mouvement le plus simple est le mouvement constant ou modèle de translation, tous les pixels du bloc effectuant le même déplacement d . Un bloc B_r de pixels de l'image de référence, de dimensions $B_x \times B_y$, centré sur le pixel (r) de coordonnées (x, y) dans l'image de référence à l'instant t , est mis en correspondance avec le bloc B_c centré en pixel c , dans l'image cible à l'instant $t+1$, soit $B_r(x, y, t)=B_c(x+d_x, y+d_y, t+1)$, avec $d=(d_x, d_y)$ le vecteur déplacement.
Les méthodes paramétriques pour estimer le mouvement, tel que le modèle de translation du mouvement, sont simples, mais seulement applicable pour les mouvements rigides. Elles sont inappropriées aux mouvements complexes.

3 Modèles non paramétriques pour l'estimation de mouvement

Les méthodes non paramétriques peuvent être utilisées pour régulariser les mouvements complexes (régler le problème de l'ouverture). Parmi celles-ci, nous pouvons distinguer :

1. Les méthodes de mise en correspondance : On suppose l'image divisée en régions, chacune correspondant à un mouvement particulier et donc à un objet. Les mises en correspondance sont soit dans le plan image, soit dans le plan transformé (la plus connue est la corrélation de phase);
2. Les méthodes statistiques, parmi lesquelles les méthodes Bayésiennes ou Markoviennes. Pour estimer le champ de déplacement, ces méthodes utilisent des contraintes probabilistes de lissage sous la forme d'un champ aléatoire, éventuellement de Gibbs, mais elles nécessitent beaucoup de calculs;
3. Les méthodes différentielles : Elles sont basées sur les gradients spatiaux et temporaux d'intensité lumineuse;
4. Les méthodes récursives : Elles sont basées sur la correction d'une prédiction ou d'une estimée du vecteur déplacement.

3.1 Algorithmes de mise en correspondance de blocs

L'estimation de mouvement dans une séquence d'images a pour rôle d'associer à chacun des pixels dans une image à l'instant t les pixels correspondants au même objet dans l'image suivante à l'instant t+1. Les vecteurs d'estimation de mouvement sont calculés entre l'image à l'instant t et les images précédentes t-1.
La mise en correspondance de bloc appelée « **block matching** » exploite les redondances temporelles entre les images consécutives. Supposons que nous voulons estimer le mouvement de divers objets contenus dans une séquence d'images. Pour simplifier, on considère le mouvement entre deux images successives, l'image courante et l'image précédente appelée « image de référence ». L'image courante est divisée en blocs de taille égale 8x8 ou 16x16 pixels (cf. figure 5).

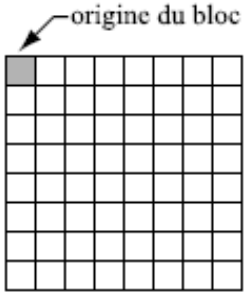


Figure 5 : Illustration d'un bloc de taille 8x8 pixels [Garcia V.].

On suppose le mouvement uniforme dans chacun des blocs et pour chaque bloc d'une zone déterminée appelée « fenêtre de recherche » dans l'image précédente, on calcule un critère de comparaison entre les deux blocs.

3.1.1 Critères de comparaison entre deux blocs

Ce critère de comparaison est une mesure de la ressemblance entre les valeurs des pixels contenus dans chaque bloc. La plupart des articles traitant de l'estimation de mouvement considèrent seulement la luminance du bloc du fait que l'oeil humain est plus sensible à l'intensité lumineuse qu'à la chrominance.
On considère des blocs carrés de même dimension, et soit le bloc de référence B_r et le bloc courant B_c . Les blocs sont codés en YUV (luminance Y, chrominance U, chrominance V). La luminance du pixel (i, j) du bloc de référence est donnée par $B_r(i, j, 1)$ et les chrominances par $B_r(i, j, 2)$ et $B_r(i, j, 3)$

avec $i, j \in [1, m]$. En considérant le bloc centré en (x, y) et l'image de référence notée I_r au format YUV, nous avons $B_r(i, j, 1) = I_r(x+i, y+j, 1)$.

La moyenne du bloc de référence en ne considérant que la luminance est :

$\bar{B}_r(1) = \frac{(\sum_{i=1}^{w} \sum_{j=1}^{w} B_r(i, j, 1))}{w^2}$, soit en allégeant les notations $\bar{B}_r(1) = \frac{(\sum_i \sum_j B_r(i, j, 1))}{w^2}$.

3.1.1.1 La somme des différences au carré SSD

C'est la somme des différences au carré « Square Sum Difference » entre les pixels correspondants des deux blocs

$SSD(B_c, B_r) = \sum_i \sum_j [B_c(i, j, 1) - B_r(i, j, 1)]^2$, i et j parcourent les lignes et les colonnes des blocs.

Ce critère très simple ne prend pas en compte la couleur, ce qui n'est pas le cas de cet autre critère adapté à la couleur : $SSDColor(B_c, B_r) = \sum_i \sum_j \sum_{c=1}^{3} [B_c(i, j, c) - B_r(i, j, c)]^2$.

3.1.1.2 La valeur absolue AV

Ce critère « Absolute Value » très similaire au précédent SSD, ne considère pas le carré de la différence mais la valeur absolue de la différence et a pour particularité de considérer toutes les différences de la même façon, tandis que dans le SSD, les grandes erreurs sont plus pénalisées.

$Av(B_c, B_r) = \sum_i \sum_j |B_c(i, j, 1) - B_r(i, j, 1)|$,

et en simplifiant : $AvColor(B_c, B_r) = \sum_i \sum_j \sum_{c=1}^{3} |B_c(i, j, c) - B_r(i, j, c)|$.

3.1.2 Prédictions avant-forward et arrière-backward

L'algorithme du block matching estime le mouvement des blocs entre deux images aux instants t et t-1. Le mouvement calculé va permettre de prédire les blocs de l'image à l'instant t+1 grâce aux blocs de l'image à l'instant t et aux vecteurs de mouvements. Deux types de prédictions sont possibles : la prédiction « avant » dite « forward » (cf. figure 6) et la prédiction « arrière » dite « backward » (cf. figure 7).

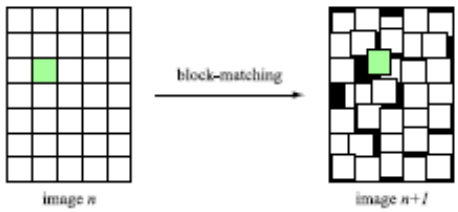


Figure 6 : « block matching » avec prédiction avant [Garcia V.].

La prédiction « avant » divise l'image t en blocs et cherche la position de chacun des blocs dans l'image suivante t+1. La prédiction « arrière » divise l'image t+1 en blocs et cherche leur position dans l'image t.
L'image prédite par prédiction « avant » présente des « trous », du fait que tous les blocs n'ayant pas le même mouvement, certains se recouvrent, et donc certains blocs dans l'image t+1 ne sont pas prédits. Ce problème n'existe pas dans la prédiction « arrière ».

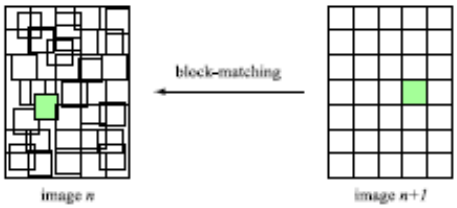


Figure 7 : « block matching » avec prédiction arrière [Garcia V.].

Le bloc pour lequel le critère de comparaison est le plus petit, c'est-à-dire le bloc le plus semblable, est déterminé pour chaque bloc de l'image de référence. Un **vecteur de déplacement** est ainsi associé à chaque bloc. Différents algorithmes de mise en correspondance de blocs sont présentés dans la suite.

3.1.3 Algorithme de recherche « Full Search »

L'algorithme de recherche exhaustive « Full Search » (cf. figure 8) parcourt de manière exhaustive l'ensemble des pixels de la fenêtre de recherche, et le bloc retourné est celui qui minimise le critère de comparaison.
L'algorithme de recherche exhaustive étant trop lourd en calculs, divers algorithmes de recherche rapides avec une stratégie de recherche ont été développés.
L'idée principale des algorithmes de recherche stratégique est que « le critère de comparaison de blocs diminue de façon monotone vers le minimum global de la fenêtre », donc il n'est plus nécessaire de parcourir tout le bloc. Il suffit de parcourir la fenêtre en se rapprochant pas à pas vers le minimum global.

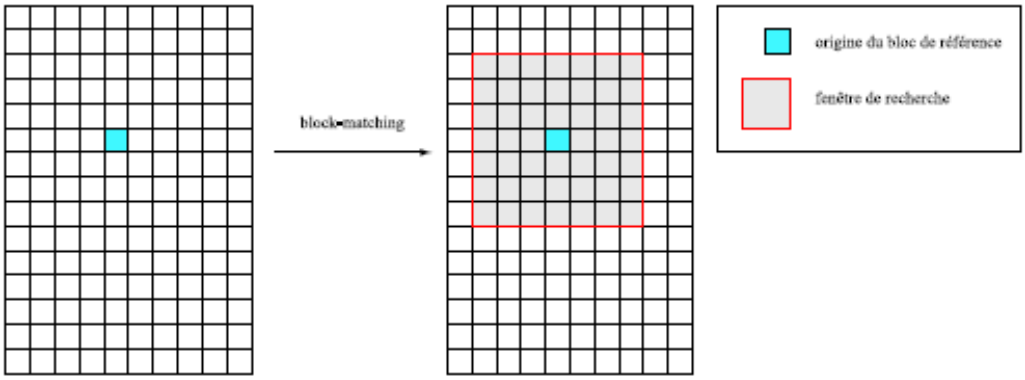


Figure 8 : Algorithme « Full Search » [Garcia V.].

3.1.4 Algorithme de recherche « Three Step Search Algorithm »

Nous présentons l'algorithme de recherche multi résolution à trois étapes « Three Step Search Algorithm » [Koga T., Linuma K., Hirano A., Lijima Y., Ishiguro T.], dit « recherche en n-pas » ou « recherche en log-n », le premier de cette catégorie d'algorithmes. Le pixel noté « 0 » représente le pixel courant. A la première itération, les 9 pixels comprenant le pixel « 0 » et les pixels notés « 1 » sont soumis à un critère de ressemblance. Si le critère optimal correspond au pixel « 0 », il n'y pas de déplacement estimé. Sinon, à l'itération suivante, le pas est égal à la moitié du déplacement maximal admis d_{max} (7 pixels dans la figure 9 ci-dessous, fenêtre de recherche de [-7, 7]x[-7, 7]) arrondi à l'entier supérieur [$d_{max}/2$], soit 4 pixels à la première itération et 2 pixels à la seconde itération pour une fenêtre [-7, 7]x[-7, 7]. Dans la figure 9, le pixel noté « 1 » en haut à droite (entouré) est le premier qui minimise le critère de distance. La flèche indique la

direction et le sens de la recherche pour le pas suivant. Une fenêtre de recherche autour de lui, avec un pas de 2 pixels (pixels notés « 2 ») est construite pour tester de nouveau les 8 pixels notés « 2 » avec le nouveau pixel central noté « 1 ». C'est le pixel de la ligne du haut au centre (entouré) qui l'emporte et devient le nouveau pixel central. L'amplitude du pas, à chaque itération, décroît selon une loi logarithmique, c'est-à-dire un pas de 1 pour la troisième itération, et les pixels notés « 3 » sont comparés avec le pixel central noté « 2 ». C'est le pixel en haut à droite noté « 3 » (entouré) qui est celui qui minimise le critère. Le nombre total de points de comparaison est (9+8+8)=25, et en général pour des fenêtres de recherche plus large, avec la même stratégie, le nombre n de points de comparaison nécessaire est $n=1+8.[\ln(d_{max}+1)]$.

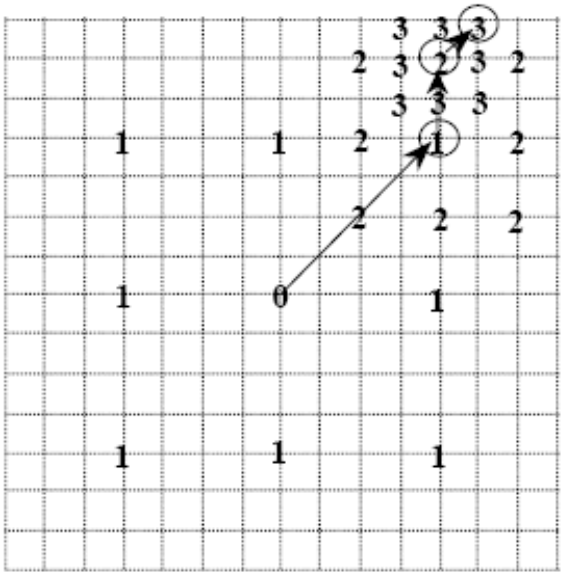


Figure 9 : Le principe de recherche en trois points [Grava C.].

3.1.5 Algorithme de recherche « *Diamond Search Algorithm* »

L'algorithme de recherche sur un grille en diamant « Diamond Search Algorithm » (DS) [Zhu S., Ma K.] a deux méthodes de recherche, présentées sur les figures 10 et 11, dérivées du modèle du diamant.

Le modèle « Large Diamond Search Pattern » (LDSP) est composé de neufs points, dont huit situés sur le bord du diamant à une distance de deux pixels, et le neuvième au centre, formant ainsi un diamant. Le modèle « Small Diamond Search Pattern » (SDSP) a cinq points dont quatre sur le bord situés à une distance de un, et le cinquième au centre du diamant.

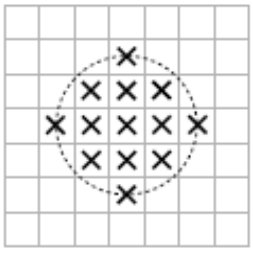
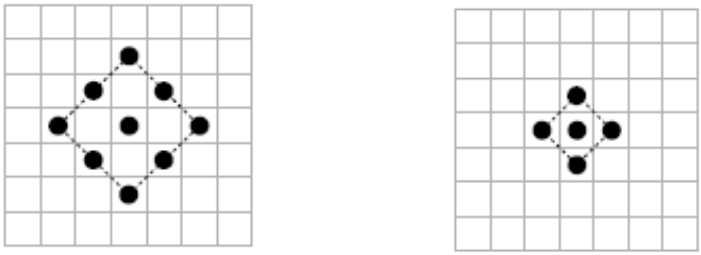


Figure 10 : Modèle de recherche : Disque de rayon deux pixels [Garcia V.].



(a) « Large diamond search » pattern (b) « Small diamond search » pattern

Figure 11 : Modèles de recherche dérivés de la figure 9 et utilisés dans l'algorithme « Diamond Search » [Garcia V.].

La figure 12 montre les étapes de l'algorithme de block-matching par « Diamond Search ».

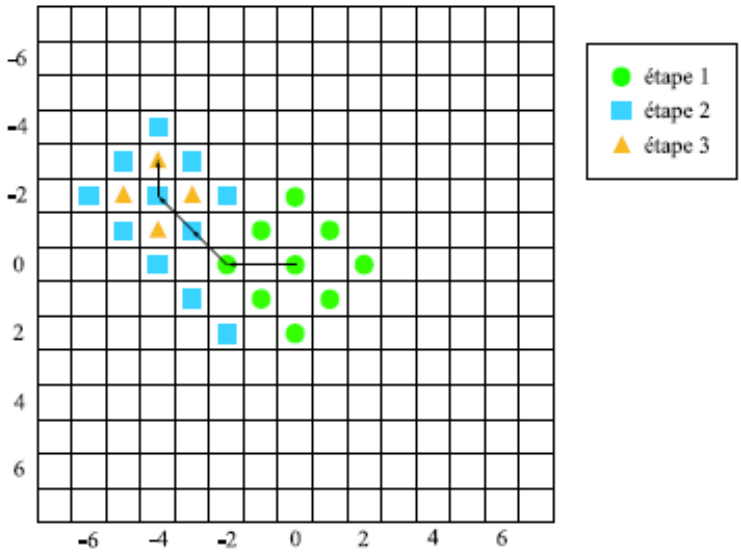


Figure 12 : Algorithme « Diamond Search » [Garcia V.].

Le diamant large est centré à l'origine du bloc courant et les neuf blocs sont testés selon le critère qui recherche le minimum local. Si le bloc qui minimise ce critère est le bloc central, on passe directement à la dernière étape sinon on continue à la seconde étape. Celle-ci consiste à se repositionner sur le bloc précédent correspondant au minimum local. Ce bloc devient le centre d'un nouveau LDSP et les neuf blocs du modèle diamant sont évalués. Si le nouveau bloc correspondant au minimum local est le bloc central, on continue avec la troisième étape, sinon on réitère la seconde étape. A la troisième étape, le modèle est le plus petit SDSP et de nouveau les blocs alentours sont testés. La solution finale correspond au bloc minimisant l'erreur.
Cet algorithme de « Diamond Search » permet d'optimiser la recherche par rapport au « Three Step Search » car la recherche est plus rapide (moins de pixels visités). D'autre part, les résultats sont meilleurs en qualité d'estimation qu'avec les algorithmes utilisant des modèles carrés car les points sont à une distance de 2 pour la norme L^1 alors qu'ils sont à une distance de 4 dans la première étape des modèles carrés pour la norme L^∞. Avec la norme L^2, les points du modèle « Diamond Search » sont à une distance de 2 ou de $\sqrt{2} \approx 1.4142$ dans les directions diagonales (cf. figure 13). Certaines directions sont donc privilégiées et le voisinage n'est donc pas homogène. L'algorithme « Heaxgon-Based Pattern » (HEXBS) [Zhu X.L.S., Chau L.] propose une solution homogène quant au voisinage.

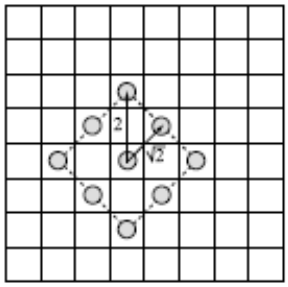


Figure 13 : Non homogénéité du voisinage dans le « Diamond Search » [Garcia V.].

3.1.6 Algorithme de recherche « *Hexagon-Based Search Algorithm* »

L'algorithme de recherche sur une grille hexagonale « Hexagon-Based Search Algorithm » et présenté à la figure 14.

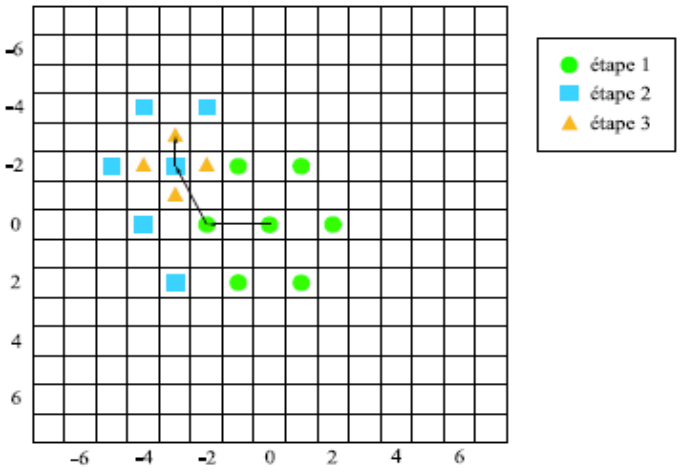


Figure 14 : Algorithme « Hexagon Based Search » [Garcia V.].

Il utilise comme le « Diamond Search » deux modèles de recherche (cf. figure 15), un modèle large (« Large Hexagonal Search Pattern » LHSP) et un modèle plus petit (« Small Hexagonal Search Pattern » SHSP).

(a) « Large hexagonal search pattern » (b) « Small hexagonal search pattern »

Figure 15 : Modèles de recherche pour l'algoritme « Hexagon-Based Search » [Garcia V.].

Cependant, le SHSP est identique au SDSP, et le LHSP contient sept points alors que le LDSP en contient

neuf. Il y a donc moins de points testés dans l'algorithme hexagonal, un avantage supplémentaire par rapport à l'algorithme diamant.
L'hexagone est une meilleure approximation du cercle que le carré pour la norme L^2, les pixels étant situés sur le LHSP à une distance de 2 ou de $\sqrt{5} \approx 2.2361$. L'algorithme HEXBS est le même que celui du DS, hormis l'utilisation des modèles de recherche LHSP et SHSP.

L'algorithme de recherche 2D-logarithme [Jain R.] effectue une recherche en croix à chaque itération. Le pas initial est de $[d_{max}/4]$. Il est divisé par deux si le pixel optimal se trouve au centre ou bien au bord de la fenêtre de recherche, sinon il ne change pas. Lorsque le pas vaut 1, les 8 points voisins du pixel central est testés. Sur la figure 16, nous avons deux cas : en haut il faut n=5+3+3+8=19 points de calcul, en bas et à droite il faut n=5+3+2+3+2+8=23 points de calcul.

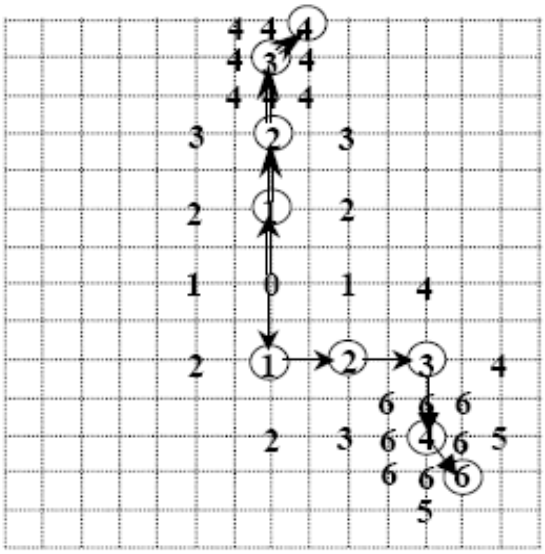


Figure 16 : Le principe de recherche « 2D-logarithmique » [Grava C.].

L'algorithme de recherche orthogonale (cf. figure 17) [Puri A., Hang H.M., Schilling D.L.] compare des paires de pixels horizontaux et verticaux avec une décroissance logarithmique du pas. La dimension initiale du pas est $[d_{max}/2]$.

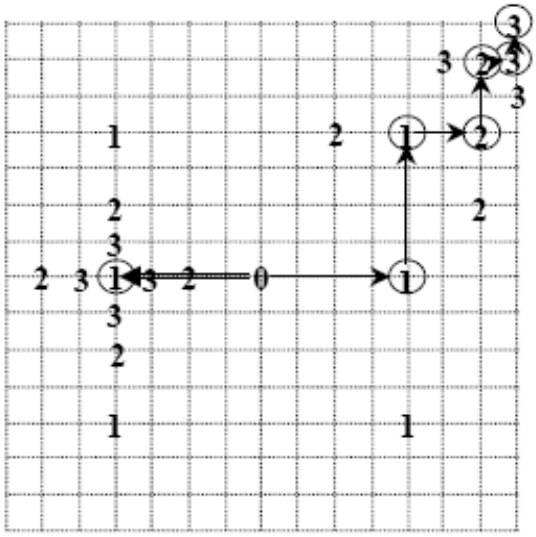


Figure 17 : Principe de recherche « orthogonale » [Grava C.].

Une itération correspond à deux étapes. La première étape calcule le critère de ressemblance en trois pixels horizontaux « 0 » et « 1 ». Le pixel minimisant le critère devient alors le centre de la deuxième étape dans la direction verticale, avec le même pas que dans la direction horizontale. A l'itération suivante, on renouvelle la même stratégie dans les directions horizontales et verticales mais avec le pas réduit de moitié. L'algorithme s'arrête quand le pas est égal à un. Dans l'exemple présenté à la figure 17, en haut à droite, la recherche orthogonale a besoin de n=3+2+2+2+2+2=13 pixels de calcul pour estimer le critère, et dans le cas général, il faut $n=1+4.[\log_2(d_{max}+1)]$ points de calcul.

3.2 L'estimation de mouvement par une approche Markovienne

Nous proposons d'étudier les méthodes markoviennes pour l'estimation de mouvement. Un champ aléatoire de Markov est noté **MRF** pour « **Markov Random Field** ».

L'estimation de mouvement est, comme nous venons de le voir, un **problème mal posé**, à moins d'y introduire des **contraintes**, ce qui **existe déjà dans la théorie des champs de Markov**, comme la continuité du mouvement à l'intérieur des objets ou des discontinuités du mouvement aux frontières de ces objets.

On appelle site s_i chaque pixel d'une image et l'image est composée d'un ensemble de sites $S=\{s_1, s_2, ..., s_{L\times L}\}$. A chaque site est associé un descripteur qui peut être son niveau de gris, une étiquette, etc. Les interactions locales entre les sites nécessitent l'introduction des relations spatiales entre les divers sites. S est donc muni d'un système de voisinage tel que : $\nu_s=\{t\}$ tels que $s\notin\nu_s$ ou bien $t\in\nu_s\rightarrow s\in\nu_t$. A partir d'un système de voisinage, un système de cliques est défini. Une clique est soit un singleton de S , soit un ensemble de sites tous voisins les uns des autres. Selon le voisinage choisi, le système de cliques sera différent, comme illustré sur la figure 18 ci-dessous.

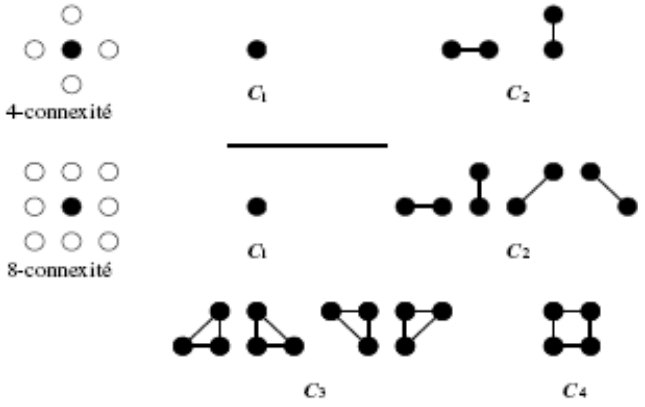


Figure 18 : Les cliques associées à deux systèmes de voisinage en dimension 2 [Tupin F., Sigelle M].

Les interactions locales entre descripteurs de sites voisins peuvent s'exprimer par un potentiel de clique. On associe à une clique c le potentiel U_c dont la valeur dépend des descripteurs des pixels de la clique.

L'énergie globale de l'image est alors la somme pondérée de toutes les cliques : $U=\sum_{c\in C} U_c$, et **l'énergie locale en un site** est la sommes des potentiels de toutes les cliques auxquelles le site appartient : $U_s=\sum_{c\in C/i} {}_{s\in c} U_c$.

L'image doit alors être modélisée de façon probabiliste, comme une réalisation d'un champ aléatoire. Pour tout site s de l'image, on peut lui associer une variable aléatoire X_s à valeurs dans E l'ensemble des descripteurs de l'image. Le niveau de gris x_s en s n'est qu'une réalisation de la variable aléatoire X_s . On définit le champ aléatoire $X=(X_s, X_t, ...)$ prenant ses valeurs dans $\Omega=E^{|S|}$. L'image est dans ce cas une réalisation x du champ. La probabilité globale de x , $P(X=x)$ donne la vraisemblance de

l'image, et les probabilités conditionnelles locales d'une valeur en un site donnent le lien statistique entre un niveau de gris (par exemple comme descripteur) et le reste de l'image. L'hypothèse markovienne permet d'évaluer ces quantités, puisque dans cette hypothèse, « **X est un champ de Markov si et seulement si la probabilité conditionnelle locale en un site n'est fonction que de la configuration du voisinage du site considéré** ». Donc pour tout x_s la valeur d'un descripteur prise au site s et $x^s=(x_t)_{t\neq s}$ la configuration de l'image sauf au site s : $P(X_s=x_s/x^s)=P(X_s=x_s/x_t,t\in\nu_s)$. Le niveau de gris en un site ne dépend donc que des niveaux de gris des pixels voisins de ce site, ce qui est le cas de la plupart des images naturelles constituées de zones homogènes ou texturées.

En pratique, il est difficile de déterminer les probabilités conditionnelles $P(X_s=x_s/x_t,t\in\nu_s)$ déterminées par les caractéristiques locales d'un champ de Markov, et donnant la probabilité de réalisation d'une distribution $P(X_s=x_s/x^s)$. Mais le théorème de **Hammersley-Clifford** [Geman S., Geman D.] va permettre l'accès aux probabilités conditionnelles locales $P(X_s=x_s/x^s)=P(X_s=x_s/x_t,t\in\nu_s)$, grâce à l'équivalence entre champs de Markov et champs de Gibbs.

Une mesure de Gibbs de fonction d'énergie $U:\Omega\rightarrow\Re$ est la probabilité définie sur Ω par $P(X=x)=\frac{1}{Z}\cdot\exp(-U(x))$ avec $U(x)=\sum_{c\in C}U_c(x)$, C le système de cliques associé au système de voisinage U , $Z=\sum_{x\in\Omega}\exp(-U(x))$ est une constante de normalisation appelée fonction de partition de Gibbs, et $U_c(x)=U_c(x_t,t\in c)$.

Un champ aléatoire X est un champ de Gibbs de potentiel associé au système de voisinage de ν si la probabilité $P(X=x)$ est une mesure de Gibbs associée au système de voisinage ν , et donc nous avons : $P(X=x)=\frac{1}{Z}\cdot\exp(-U(x))=\frac{1}{Z}\cdot\exp(-\sum_{c\in C}U_c(x))$, les $U_c(x)$ sont les énergies locales qui sont à relier aux probabilités conditionnelles locales. Plus une configuration d'un champ de Gibbs a une énergie faible, plus elle est probable. Le théorème de **Hammersley-Clifford** indique que : « **X est un champ de Markov relativement à ν et $P(X=x)>0\quad\forall x\in\Omega$ si et seulement si X est un champ de Gibbs de potentiel associé à ν ». On établit ainsi l'équivalence entre champ de Markov caractérisé par ses propriétés locales $P(X_s=x_s/x^s)=P(X_s=x_s/x_t,t\in\nu_s)$ et champ de Gibbs caractérisé par sa propriété globale $P(X=x)=\frac{1}{Z}\cdot\exp(-U(x))$, la distribution de Gibbs.**

3.2.1 Estimation stochastique du mouvement avec le MAP

Dans l'estimation stochastique du mouvement, les images et les champs de déplacement sont modélisés par des champs aléatoires de Markov (MRF). L'estimation du MAP du mouvement entre les images I_{t-1} et I_t entre les instants t-1 et t consiste à trouver la meilleure estimation du vecteur déplacement $\hat{d}$ qui maximise la probabilité $P(d/I_t,I_{t-1})$.

La formule de Bayes nous donne $P(d/I_t,I_{t-1})\overset{Bayes}{=}\frac{(p(I_t/d,I_{t-1})\cdot p(d/I_{t-1}))}{(p(I_t/I_{t-1}))}$,

$$\hat{d}\overset{MAP}{=}argmax_d\,p(d/I_t,I_{t-1})\overset{Bayes}{=}argmax_d\,p(I_t/d,I_{t-1})\cdot p(d/I_{t-1})$$

donc $$\frac{1}{Z}\cdot argmax_d\exp(\{-[U(I_t/d,I_{t-1})+U(d/I_{t-1})]\})$$

$$argmin_d[U(I_t/d,I_{t-1})+U(d/I_{t-1})]$$

Il s'agit donc de maximiser la probabilité à posteriori, ce qui revient à minimiser l'énergie à posteriori (appelée aussi « critère du MAP »). $U(I_t/d,I_{t-1})$ est l'énergie d'attache aux données, $U(d/I_{t-1})$ est l'énergie *a priori* ou terme de régularisation.

3.2.2 Algorithmes de minimisation du critère du MAP

Pour minimiser le critère du MAP, il existe divers algorithmes de minimisation (cf. Annexe 1) :
-Les algorithmes **stochastiques**, de type recuit simulé (recuit avec dynamique de « Metropolis », échantillonneur de «Gibbs avec recuit »), les algorithmes génétiques, les algorithmes déterministes (les modes conditionnels itérés « ICM » « Iterated Conditional Modes », la non-convexité graduelle « GNC » « Graduated Non-Convexity », le recuit en champ moyen « MFA » « Mean Field Annealing »).
-Les algorithmes **déterministes** sont plus rapides que ceux stochastiques mais peuvent être piégés dans un **minimum local** de l'énergie du critère du MAP au lieu d'un minimum global assuré pour l'algorithme stochastique.

Les algorithmes effectuant simultanément l'estimation de mouvement et la segmentation de l'image fonctionnent avec une approche bayésienne. Une méthode basée sur le **test de vraisemblance** [Bouthémy'87] est fondée sur le schéma division/fusion. L'image est découpée en blocs carrés de taille 16x16 pixels. Pour chaque bloc, on calcule le **rapport de vraisemblance de deux hypothèses** :
-H_0 le bloc est homogène au sens du mouvement avec un modèle de mouvement défini par une translation;
-H_1 le bloc est composé de deux parties animées chacune d'un mouvement de translation différent.
A chaque hypothèse est associée une fonction de vraisemblance. On recherche ensuite quelle est l'hypothèse qui minimise le rapport logarithmique des fonctions de vraisemblance. Si c'est H_1, le bloc est divisé et la même procédure est appliquée à chacune des ses parties. Si c'est H_0, les sous parties sont fusionnées.
Dans un cadre bayésien, des distributions de probabilité semblables aux fonctions de vraisemblance sont utilisées. Soit un ensemble ou champ O de variables aléatoires appelées les observations. On cherche à estimer un ensemble de variables aléatoires que sont les **étiquettes** ou primitives E (numéro de régions, vecteurs de paramètres, vecteurs vitesse, etc). **Les observations sont les dérivées spatio-temporelles de la fonction intensité lumineuse**. En supposant que e et o sont des réalisations particulières des variables aléatoires E et O , on cherche les primitives maximisant la probabilité globale à posteriori $p(E=e/O=o)$. Il s'agit d'un **estimateur au sens du maximum à posteriori (MAP)** et d'après le théorème de Bayes :

$$p(E=e/O=o)=\frac{(p(O=o/E=e)\cdot p(E=e))}{(p(O=o))} \quad .$$

Le dénominateur étant une constante, maximiser $p(E=e/O=o)$ revient à maximiser $p(O=o/E=e)\cdot p(E=e)=p(O=o,E=e)$.
Le premier terme $p(O=o/E=e)$, la probabilité à posteriori des observations, relie les primitives aux observations, et prend une forme gaussienne. C'est un **terme d'attache aux données**.
Le second terme $p(E=e)$ est une probabilité *a priori* qui a un **rôle régularisant** et peut être décrite par le formalisme Markovien. L'image en deux dimensions est composée de pixels, les sites s . $s\in S$, S est le support du champ à estimer, sur lequel on définit un voisinage B . Le champ E est de Markov pour le voisinage B si toutes les réalisations ont une probabilité de se réaliser non nulle, et pour chaque site la loi de probabilité de son étiquette sachant les étiquettes de tous les autres sites est la même que la loi de probabilité de son étiquette sachant les étiquettes uniquement des sites voisins. D'après le théorème de **Hammersley-Clifford**, les probabilités associées à un champ de Markov suivent la **distribution de Gibbs** :
$p(E=e)=\frac{(\exp(-U(e)))}{Z}$, Z est la fonction de partition, $U(e)=\sum_{c\in C} V_c(e)$, $V_c(e)$ sont les fonctions potentiels locales. Chaque potentiel est associé à une clique $c\in C$, une clique est un sous-ensemble de S composé d'un seul site ou de sites tous mutuellement voisins. Maximiser la probabilité $p(E=e)$ revient à minimiser l'énergie $U(e)$. Il faut donc des potentiels bas aux configurations que l'on veut privilégier, et des potentiels hauts aux configurations que l'on veut décourager. En ce qui concerne la segmentation en régions homogènes, on privilégie les cliques composées de sites avec des étiquettes identiques.

Le problème d'estimation bayésienne au sens du MAP est défini en termes de minimisation d'une énergie de Gibbs avec un terme d'attache aux données et un terme régularisant [Odobez J.M., Bouthemy P. 94]. L'énergie globale est minimisée grâce à un **algorithme déterministe de type ICM** [Odobez J.M., Bouthemy P.

94], qui évite la lenteur de convergence des algorithmes stochastiques tel que le recuit simulé mais conduisent parfois à un minimum local. Une **approche multi-résolution** proposée par [Odobez J.M., Bouthemy P. 94] permet d'éviter ce piège.

3.2.2.1 Approche multi résolution

Quand les déplacements sont trop importants, les méthodes de l'ECMA sont inadaptées et les procédés de relaxation itératifs trop lents. Une approche **multi résolution** permet de palier à ces problèmes. [Meyer F., Bouthemy P. 94] proposent une approche multi résolution : des pyramides d'images d'une même scène sont construites à des résolutions successives (cf. figure 19). Les mouvements les plus importants sont estimés aux résolutions les plus basses. Au fur et à mesure qu'on monte dans les résolutions, le mouvement est de plus en plus fin et la solution est précisée par une reformulation de l'ECMA.

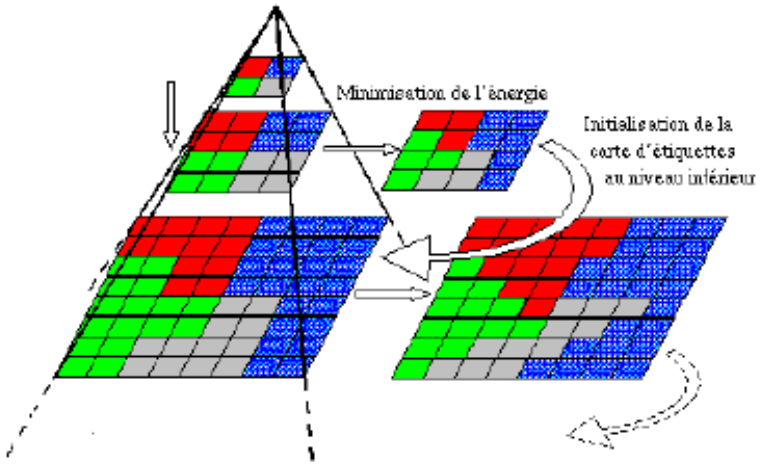


Figure 19 : Principe de résolution multi grille du problème de minimisation. La configuration optimale se trouve dans une de ces images emboîtées, où la contrainte d'homogénéité par bloc des étiquettes est progressivement levée. La résolution est effectuée au niveau le plus haut de la pyramide (espace de recherche très restreint) et progressivement raffinée [Gelgon M.].

Les approches multi résolution proposent des solutions meilleures en un temps de convergence plus court, et en tenant compte des déplacements importants. Toutefois les recouvrements entre les projections successives des objets mobiles doivent être importants. Dans les autres cas, des alternatives à l'estimation instantanée du vecteur vitesse apparente ont été proposées comme les méthodes de « block matching », mise en correspondance de blocs.
Dans le contexte de la **vidéo surveillance**, **les techniques différentielles ECMA sont inadaptées** car elles nécessitent de faibles amplitudes de mouvement, y compris en multi résolution où les grandes amplitudes sont envisagées mais les recouvrements doivent être importants en deux instants successifs. **Les techniques de « block matching » sont, elles, coûteuses en temps de calcul et comportent bon nombre d'ambiguïtés d'appariement.**

3.2.2.2 Approche à base du mouvement des contours

Une approche à base de calcul du **mouvement 2D des contours** a ainsi vu le jour pour pallier aux inconvénients précédents [Ricquebourg Y. 97], **adaptée aux structures articulées que sont les personnes.** Dans l'espace (x, y, t), un contour génère au cours de son déplacement, une **surface spatio-temporelle** donnant

la composante du vecteur vitesse apparente normale au contour [Bouthemy P. 89]. La mesure de déplacement des contours a été placée dans un cadre statistique markovien [Ricquebourg Y., Bouthemy P.].
Pour estimer le déplacement normal des contours de façon robuste, des informations contextuelles locales doivent être prises en compte. Ainsi un modèle markovien d'interaction locale sur une carte des contours est défini. Les points de contours sont préalablement extraits via un détecteur spatial classique ([Deriche R.], [Shen J., Castan S.]). Une étape de polygonalisation transforme les structures en segments. Ils sont chaînés, formant les chaînes des contours qui constituent les primitives de base de l'algorithme, le « support topologique », comportant des sites, un voisinage et des cliques. Les cliques sont les composantes des vitesses orthogonales aux contours, ce sont les points appartenant aux chaînes de contours détectés. Le voisinage est défini d'ordre 1, il est constitué du prédécesseur et du successeur d'un site le long de la chaîne de contour, et l'ensemble des cliques associées au système de voisinage. On note E le champ des étiquettes à estimer, les vitesses orthogonales au contour en chaque site défini, ainsi que le champ O des observations, les intensités lumineuses. Pour estimer le champ E des étiquettes par des champs de Markov, une approche bayésienne fondée sur le critère du MAP, maximum à posteriori, maximise la probabilité jointe $p(E=e,O=o)$. L'équivalence entre champs de Markov et distribution de Gibbs, expliquée dans le **théorème de Hammersley et Clifford** [Geman S., Geman D.] indique que la densité de probabilité : $p(E=e,O=o)=\frac{(\exp(-U(e,o)))}{Z}$ avec $Z=\sum_{e\in\Omega}\exp(-U(e,o))$, et Ω est l'ensemble des réalisations possibles du champ des étiquettes.

$U(e,o)$ est une fonctionnelle appelée « fonction d'énergie ». C'est la somme pondérée de deux termes : $U(e,o)=U_1(e)+U_2(e,o)$, avec $U_1(e)$ un terme de « régularisation » du champ E correspondant à un *a priori* sur les propriétés de la répartition des étiquettes de la solution recherchée, et $U_2(e,o)$ un terme « d'adéquation étiquettes-observations » dit « d'attache aux données », correspondant à la vraisemblance de la solution par rapport aux observations.

$U_1(e)$ et $U_2(e,o)$ sont des termes d'énergie qui peuvent s'exprimer sous la forme d'une somme de potentiels locaux sur les cliques du voisinage, grâce à la modélisation markovienne des champs E et O. Afin de définir la vraisemblance $U_2(e,o)$ reliant observations et étiquettes (les déplacements normaux en chaque site) et avec la modélisation choisie d'un élément de contour en mouvement par une portion de surface spatio-temporelle, [Bouthemy P. 89] propose de considérer la séquence d'images à traiter comme un volume 3D de l'espace (x, y, t) composé de deux dimensions spatiales (x, y) correspondant à chaque plan image et d'une dimension temporelle t. Dans un espace, un élément de contour en 2D génère une portion de surface (2D +t) élémentaire, caractérisée par un vecteur de paramètres. L'estimation du mouvement des contours revient alors à déterminer ces portions de surfaces.
Une fois la structuration des primitives en niveaux sites, segments, chaînes de contours effectuée, un algorithme de relaxation déterministe minimise la fonction globale d'énergie, donnant ainsi une solution à l'estimation du champ E des déplacements normaux aux contours.

4 Suivi de trajectoires

4.1 Les techniques de suivi

Les techniques de suivi sont issues des « radaristes » pour le suivi de cibles, à base de théorie de l'estimation. L'objectif est de combiner les informations issues de divers capteurs et d'obtenir un état estimé le plus proche possible de l'état observé. Le filtre de Kalman estime l'état d'un système dynamique linéaire à l'aide d'un modèle d'observation probabiliste [Kalman R.E.], mais il se limite au modèle de bruit gaussien. Le « Filtrage Particulaire » ou algorithme de « Condensation », plus connu en vision par ordinateur, estime les états dans les cas non linéaires ou non gaussiens. Son principe est de calculer des lois de probabilité des états par une somme finie pondérée des lois de Dirac avec des poids évoluant en fonction des observations.

Le suivi de personnes est fondamental dans les systèmes de vidéo surveillance puisqu'il est la base de l'analyse de comportements, la reconnaissance des activités et la détection des évènements d'intérêts. Toutefois, le suivi

est conditionné par la qualité de la reconnaissance. Le fondement des problèmes de suivi consiste à suivre et associer correctement les individus. C'est un **problème d'association de données (« Probabilistic Data Association », « PDA »)** qui a été identifié dans la littérature radar et sonar [Fortmann T.E., Bar-Shalom Y., Scheffe M.], avant que le suivi vidéo ne soit d'actualité.
Dans le domaine du suivi vidéo, le problème d'association de données a été résolu par des primitives telles que le mouvement, l'apparence et la forme qui doivent avoir des modèles cohérents au cours du temps ([Wren C.R., et al.], [Haritaoglu I., Harwood D., Davis L.S. 00], [Lipton A.J., Fujiyoshi H., Patil R.S.], [Zhao T., Nevatia R.], [McKenna S., Raja Y., Gong S.], [Fuentes L.M., Velastin S.A.]).

Dans les **problèmes d'association de données**, les algorithmes de suivi de points sont constitués des méthodes **directes** fondées sur la prédiction de mouvement, et des méthodes **multi-hypothèses**.
-**Dans les méthodes directes, on choisit le point le plus similaire dans la zone de recherche localement**, à chaque instant, s**ans prendre en compte une certaine durée.** La mise en correspondance directe [Meyer F., Bouthemy P. 92] cherche à trouver les correspondants dans **l'image suivante à partir de l'image actuelle, tout en minimisant l'erreur d'appariement;**
-Le s**uivi multi hypothèses, contrairement aux méthodes directes, est centré sur le problème d'association de données points/trajectoires**. Le filtre **MHT (Multi-Hypothesis Filter** ou Multi-Hypothesis Tracking) existe en suivi de cibles ([D.B. Reid.], [Cox I.J.]). Ce filtre modélise l'initialisation, le maintien et la terminaison de pistes. Il génère des arbres d'hypothèses dont les branches forment des associations candidates. La dimension temporelle est prise en compte, en plus de l'association des trajectoires avec les primitives points. Les trajectoires et les hypothèses sont sélectionnées sur la durée, permettant ainsi de lever les ambiguïtés dues aux occultations, de traiter la concurrence entre les associations, et de régler l'initialisation des trajectoires grâce à plusieurs hypothèses possibles. Un certain nombre d'hypothèses d'association sont envisagées, chacune assortie d'une probabilité, et la meilleure est sélectionnée. L'arbre des hypothèses croît de façon exponentielle, et des stratégies d'élagage permettent de ne conserver que les combinaisons d'associations les plus probables, par exemple avec l'algorithme des k-meilleures hypothèses de [Cox I.J, Hingorani S.L.]. Il arrive cependant que des solutions sont supprimées trop tôt, avant qu'elles n'aient eu le temps de montrer leur efficacité. [Cox I.J, Hingorani S.L.] a utilisé le MHT pour le suivi de coins.

D'autres méthodes statistiques de suivi temporel multi pistes de plusieurs primitives simultanément existent :
-L'algorithme des **plus proches voisins NN (Nearest Neighbors)** met en correspondance les observations en fonction d'une distance;
-Le filtre **Probabilistic Data Association Filter (PDAF**) estime le maintien de pistes [Bar-Shalom Y., Li X.];
-Le filtre **Joint Probabilistic Data Association Filter (JPDAF**) est une amélioration du précédent. **Le JPDAF traite les associations image par image, et non sur la durée**, comme le font les méthodes MHT et PMHT que nous détaillons plus loin;
-Le **filtre de Kalman Distribué DKF** [Rao B.S.Y., Durrant-Whyte H.F., Sheen J.A.] combine divers filtres de Kalman en parallèle pour augmenter la robustesse;
-Un **filtre hybride, mélange de MHT et JPDAF** utilise le MHT pour l'initialisation et la terminaison de pistes, et le JPDAF pour le maintien des pistes;
-L'approche par **optimisation combinatoire**.

Nous présentons quelques unes de ces approches les plus usitées, le MHT, le PMHT, le JPDAF, et l'approche combinatoire.

4.1.1 Le MHT « Multiple Hypothesis Tracking »

Le MHT « Multiple Hypothesis Tracking » est la méthode la plus utilisée parmi celles de suivi multi hypothèses. Les données sont associées de façon probabiliste, et le MHT génère des hypothèses explicitement. La première fois que l'approche des hypothèses multiples a été utilisée fut par [Housewright R.B., Singer R.A., Sea R.G.] , mais elle ne tenait compte que d'une unique cible et ne se souciait guère de l'initialisation de piste. [D.B. Reid.] a remédié à ces deux problèmes, plusieurs pistes sont initialisées et suivies dans un environnement encombré. Une arborescence d'hypothèses est mise en place, avec des probabilités calculées

pour chaque hypothèse, que celle-ci provienne d'une cible déjà connue, d'une nouvelle cible ou qu'elle soit une fausse alarme. Les états des pistes sont estimés pour chaque hypothèse en utilisant un filtre de Kalman. Les probabilités des hypothèses sont calculées récursivement au fur et à mesure que de nouvelles mesures arrivent. Les hypothèses peu vraisemblables sont éliminées afin de ne conserver qu'un petit nombre d'hypothèses et les hypothèses correspondant à une même cible sont combinées.
[Cox I.J, Hingorani S.L.] applique cet algorithme à des points d'intérêts issus d'une séquence vidéo. Il prédit la position des points par filtrage de Kalman. Les points candidats les plus similaires sont triés selon un critère de corrélation sur la luminance du voisinage des points. Plusieurs candidats sont générés pour une trajectoire, il s'agit soit d'une association avec chacun des points candidats, soit de la disparition d'un point, soit d'une fausse alarme. Des hypothèses globales sont générées, avec l'ensemble des associations entre les trajectoires et les points. La procédure répétée sur plusieurs images forme un arbre d'hypothèses. La décision concernant les candidats est différée jusqu'à la confirmation d'une hypothèse au cours du temps, donné par le calcul d'une probabilité liée à l'ensemble des trajectoires correspondant à une hypothèse globale. La croissance de l'arbre des hypothèses étant exponentielle, il faut l'élaguer.

4.1.2 Le PMHT « Probabilistic Multiple Hypothesis Tracking »

Le PMHT, « Probabilistic Multiple Hypothesis Tracking » est une approche probabiliste, dans laquelle l'affectation trajectoires/mesures est définie par des variables aléatoires, et les hypothèses dans leur ensemble au cours du temps [Gauvrit H., Le Cadre J.P.] et non individuellement, instant après instant, évitant l'énumération explicite des hypothèses. L'initialisation des vecteurs d'état de position et de vitesse doit être correcte, sinon les résultats manquent de robustesse, notamment pour les trajectoires au niveau d'un croisement. [Gauvrit H., Le Cadre J.P.] utilise cet algorithme pour les données sonar unidimensionnelles. La méthode a été appliquée aux séquences vidéo par [Gelgon M.] avec une segmentation en régions homogènes par le mouvement. Le suivi est assuré par mise en correspondance de régions entre images successives, et l'algorithme multi hypothèses permet alors de relier les bouts de trajectoires obtenues de chaque côté des zones d'occultations.

4.1.3 Le JPDAF « Joint Probabilistic Data Association Filter »

Le PDAF est une extension du filtre de Kalman pour le suivi dans le contexte de mesures multiples. Dans le filtre de Kalman, l'innovation est construite à partir d'une combinaison de mesures, pondérées par la confiance qu'on leur accorde. Plusieurs PDAF utilisés simultanément pour le suivi multi pistes posent le problème des filtres qui suivent la même piste. Le JPDAF ajoute un principe d'exclusion aux associations du PDAF et remédie à ce problème.

4.1.4 Le JPDAF « Joint Probabilistic Data Association Filter »

Cet algorithme ([Blackman S.S.], [Bar-Shalom Y., Fortmann T.E.]) prend en compte toutes les observations dans le voisinage de la position de la cible prédite pour mettre à jour l'estimée de la position en utilisant une probabilité à posteriori. Plusieurs hypothèses concurrentes sur l'origine des données sont formées, mais les décisions finales sont prises séquentiellement. Les hypothèses sont combinées. Le JPDAF prend en compte les incertitudes liées aux mesures peu fiables, en particulier en cas d'occultations partielles, le suivi agissant sur les parties non occultées. L'incertitude sur les pistes est modélisée par les matrices de covariance associées aux trajectoires. L'avantage de cette méthode est sa récursivité, qui n'a pas besoin de stocker les observations passées ni les multiples hypothèses candidates. En revanche, il n'y a pas de mécanisme d'initialisation de piste. La méthode «**Joint Probabilistic Data Association Filter»** est une méthode d'entretien de trajectoires déjà initialisées et **ne permet pas de détecter l'apparition de nouveaux points**. La solution serait alors d'utiliser l'algorithme **MHT pour l'initialisation des trajectoires et l'algorithme JPDAF pour la maintenance** de celles-ci. Le choix du MHT ou du JPDA est fonction de la densité de fausses alarmes. Pour un grand nombre de fausses alarmes, le MHT n'est pas envisageable. Pour un petit nombre, le MHT est plus pertinent.

Les algorithmes MHT et JPDA énumèrent de façon exhaustive les hypothèses d'association entre les mesures et les pistes. Leur nombre croissant exponentiellement avec le nombre de cibles, des techniques ont été

élaborées afin de se limiter aux hypothèses les plus vraisemblables. [Streit R.L., Luginbuhl T.E. 94] ont proposé une nouvelle approche modélisant les associations des mesures aux pistes comme des variables aléatoires à estimer. Aucune énumération n'est nécessaire mais le vecteur d'association est considéré comme une donnée du problème qu'il s'agit d'estimer.

4.1.5 L'approche par optimisation combinatoire

Comme dans l'approche probabiliste du PMHT, d**ans l'approche par optimisation combinatoire**, il s'agit d'éviter l'énumération explicite des hypothèses. [Gauvrit H., Le Cadre J.P.] considère la séquence dans son ensemble, et par une méthode d'optimisation, minimise un coût défini à partir des probabilités des trajectoires, sous contrainte d'unicité des correspondances. Cette méthode a le défaut de l'explosion combinatoire que [Gauvrit H., Le Cadre J.P.] résout en combinant l'algorithme combinatoire qui permet d'obtenir une solution approchée, et qui sert d'initialisation à l'algorithme probabiliste.

4.1.6 L'appariement temporel

4.1.6.1 La mise en correspondance

La première famille est la mise en correspondance directe. Il faut **détecter les régions** correspondants aux objets en mouvement dans la scène, soit par une soustraction du fond [Stauffer C., Grimson W.E.L.b], soit par une différence d'images soit par une combinaison des deux [Collins R., et al.b]. Il s'agit d'apparier les détections obtenues dans les images précédentes avec celles des images courantes, en minimisant l'erreur commise par cet appariement.

4.1.6.2 Le filtre de Kalman

La seconde famille **associe les régions détectées dans des images consécutives et calcule de façon récursive la trajectoire** des objets à suivre. Des méthodes ont été développées pour résoudre ce problème, de l'estimation par filtre de Kalman aux arbres multi hypothèses, des méthodes d'inférence avec degrés de confiance (**JPDAF** [Bar-Shalom Y., Fortmann T.E.]) et **filtres à particules** [Isard M., Blake A., 98]. Les méthodes bas niveau ne résolvent pas les problèmes d'occultation et les trajectoires estimées sont perdues dans ce cas. Un réseau bayésien est alors utilisé pour relier différentes trajectoires appartenant au même objet en leur assignant un label commun. Le suivi de l'objet s'effectue de deux façons. Les opérations bas niveau permettent de détecter les régions en mouvement et d'associer des régions par paires dans les images consécutives. Les opérations bas niveau produisent un ensemble de trajectoires, chacune décrivant l'évolution d'un objet ou d'un groupe d'objets dans le flux vidéo. Pour extraire la trajectoire complète de chacun des objets, il est nécessaire de relier plusieurs segments de la trajectoire. Il faut alors attribuer un label en assignant un label probabiliste à chaque fois. Les interactions entre les différents labels peuvent être modélisés par un réseau bayésien. Les noeuds du réseau bayésien sont les labels et les liens représentent les dépendances causales modélisées par les tables de probabilités conditionnelles. La meilleure configuration d'étiquetage peut être obtenue par inférence probabiliste grâce à un arbre de jonction [Jensen F.b].

Cette famille s'oppose à la précédente dans l'utilisation des mesures courantes. Le traitement a lieu en deux étapes : l'image précédente sert à prédire l'image courante puis la prédiction est comparée avec les mesures réellement obtenues. Le filtre de Kalman tel que utilisé par ([Baumberg A., Hogg D.], [Choi S., Seo Y., Kim H., Hong K.] [Ricquebourg Y. 97]) est un filtre récursif prédisant l'état courant du système à partir des mesures précédentes si l'évolution dynamique du système est considéré comme linéaire. Le vecteur d'état X(t) est défini en fonction de X(t-1) et un terme correspondant à l'évolution dynamique du système. Ce filtre a deux intérêts : obtenir une prédiction de l'état courant indépendamment des mesures obtenues et connaître la fiabilité du modèle de mouvement grâce à ces mesures. Les deux inconvénients sont la nécessité de la modélisation de l'évolution dynamique des objets de la scène, et sa sensibilité aux valeurs initiales à cause de son caractère itératif.

4.2 Les techniques de suivi de trajectoires

Les techniques de suivi de trajectoires sont aussi issues du domaine radar. En radar, il s'agit d'établir des pistes à partir des mesures obtenues en les associant aux mesures précédentes afin de mettre à jour les pistes, en terminer certaines qui seraient sorties de la zone de surveillance, en initialiser d'autres, etc. **C'est un problème identique à l'établissement des trajectoires long terme.** Selon la façon d'associer les observations aux pistes, nous distinguons trois familles :
-les approches déterministes;
-les approches probabilistes non bayésiennes;
-les approches bayésiennes.
Le suivi long terme aborde le problème de disparition, réapparition et occultation des objets suivis, levé par la prédiction déterminant l'évolution des traces des objets.

4.2.1 Approches déterministes

Pour résoudre **l'association modèles/observations**, la méthode de base est le « **filtre du plus proche voisin** » qui utilise l'observation la plus proche de l'observation prédite. Les résultats obtenus sont pauvres dans le cas d'environnement bruité. En effet, ils ne tiennent pas compte du fait que la mesure utilisée dans le filtre peut provenir d'une autre source que la cible d'intérêt. De plus, il n'y a pas de modèle d'évolution dynamique du système pour parer à cela. Les associations ne sont jamais remises en cause ni affectées d'une incertitude. Ce n'est pas le cas des approches probabilistes contenant un modèle d'évolution dynamique et des tests d'hypothèses ainsi que le filtrage linéaire optimal des trajectoires ([Blackman S.S.], [Bar-Shalom Y., Fortmann T.E.]).

4.2.2 Approches probabilistes non bayésiennes basées sur des fonctions de vraisemblance

Une méthode d'estimation au sens du maximum de vraisemblance où la vraisemblance dépend de l'erreur résiduelle entre les prédictions et les mesures, permet de décider la construction d'une **trajectoire**. L'algorithme le plus connu de cette famille est celui du « track split » [Buechler G., Smith P.] qui sépare la piste tant que plus d'une détection est observée dans le voisinage de la mesure prédite. Une fonction de vraisemblance est calculée sur chaque trajectoire et les trajectoires dont la vraisemblance est en dessous d'un seuil sont éliminées. Les résultats sont convenables pour l'initialisation des pistes et leur mise à jour, mais les temps de calcul peuvent devenir prohibitif dans des environnements complexes.
Le défaut majeur de ces algorithmes est que les décisions sont binaires, les trajectoires sont acceptées ou rejetées. L'estimation de l'état résultant et les covariances ne prennent pas en compte l'incertitude des décisions, caractéristique des approches non bayésiennes.

4.2.3 Approches probabilistes

Les **associations observations/trajectoires** sont modélisées comme un événement aléatoire auquel est associé une probabilité. Ces techniques estiment les probabilités a posteriori selon la règle de Bayes :

$$p(H_t|D)=\frac{(p(D|H_t)\times p(H_t))}{(p(D))}$$

- H_t est l'hypothèse de la source à l'origine des données;
- D est l'ensemble des données reçues le plus récemment;
- $p(H_t)$ est la probabilité *a priori* que l'hypothèse H_t soit correcte;
- $p(H_t|D)$ est la probabilité à posteriori de H_t ;
- $p(D)$ est la probabilité de recevoir l'ensemble des données D ;
- $p(D|H_t)$ est la probabilité conditionnelle de recevoir D étant donné H_t .

4.2.3.1 Les travaux sur la segmentation par le mouvement

[Gelgon M.] a **étendu la segmentation par carte d'étiquettes au suivi de deux partitions spatiale et mouvement** au cours d'une séquence d'images. Les deux segmentations spatiale et mouvement sont définies comme des étiquetages statistiques permettant d'exploiter la **cohérence temporelle des cartes de segmentation.**
L'analyse du mouvement dans les images est une tâche difficile du fait que le mouvement apparent dans l'image est une variable cachée, dont les discontinuités spatiales sont *a priori* inconnues. Le champ de mouvement 2D est la projection du mouvement 3D dans l'image et il ne peut être mesuré que dans les zones de l'image où il provoque des variations spatio-temporelles de l'intensité. L'équation de contrainte du mouvement apparent, liant le vecteur vitesse aux mesures de gradients spatio-temporels de l'intensité, est la base des techniques d'estimation de mouvement. Elle ne détermine cependant que la composante normale du vecteur de vitesse, c'est le classique problème de l'ouverture [Mitiche A., Bouthemy P.]. Une contrainte supplémentaire est alors introduite, favorisant la similarité des vecteurs de mouvement de pixels voisins [Horn B.K.P, Schunk B.G.] ou se basant sur des modèles de mouvement paramétriques. Cette contrainte additionnelle suppose la continuité spatiale du champ de vitesse apparent, mais le problème se pose aux discontinuités du mouvement apparent. Ces discontinuités sont localisées aux contours des projections des éléments en mouvement. Ainsi, pour bien estimer le mouvement, il faut connaître les régions homogènes en mouvement. A l'inverse, segmenter les régions au sens du mouvement nécessite la connaissance des mouvement, donc l'estimation du mouvement et la segmentation des régions en mouvement sont étroitement liées. Il faut donc déterminer un partitionnement en régions de l'image dont le mouvement est homogène, et le mouvement de ces régions est estimé sous forme dense (champ de vitesse 2D) ou sous forme d'un modèle paramétrique.

Deux types de méthodes de segmentation par le mouvement existent :
-les méthodes dites « **indirectes** » qui segmentent un champ de mouvement préalablement estimé;
-les méthodes dites « **directes** » qui segmentent directement à partir de l'image.

Certaines méthodes, dites **séquentielles**, extraient les divers régions en mouvement de l'image, de façon successive, en partant de l'image considérée dans sa globalité. Le modèle de mouvement du fond de la scène est estimé et les régions ne suivant pas ce modèle sont repérées. Dans ces régions, un autre modèle de mouvement est estimé, et de nouveau les régions ayant un autre mouvement sont détectées et ainsi de suite, jusqu'à un critère d'arrêt. [Bouthemy P., François E.] alterne les phases d'estimation et de segmentation dans un cadre d'étiquetage statistique markovien où l'affectation d'un pixel peut changer au cours de la phase de segmentation.

On peut aussi traiter **de façon conjointe l'estimation et la segmentation du mouvement** avec une approche markovienne de la carte de segmentation, visant à introduire une information contextuelle. En effet, le cadre statistique markovien permet de modéliser le problème de segmentation par une fonction d'énergie comportant un terme d'attache aux données (les gradients spatio-temporels de l'intensité) et une contrainte contextuelle ([Bouthemy P., François E.], [Odobez J.M., Bouthemy P. 98]). La minimisation de cette fonction d'énergie correspond à la carte de segmentation la plus probable au sens du maximum à posteriori (MAP). La modélisation de la fonction d'énergie s'effectue par une méthode d'optimisation de cette fonction, qui est non convexe. Les méthodes stochastiques assurent la convergence vers un minimum global bien que lentement, tandis que les méthodes déterministes convergent vers un minimum local plus rapidement. Des stratégies multi-échelles leur sont souvent associées [Odobez J.M., Bouthemy P. 98], afin d'atteindre un minimum plus vite et meilleur qu'avec une seule échelle, car l'espace de recherche étant plus restreint, la fonction d'énergie est plus convexe. La segmentation par le mouvement apparent peut être également modélisé par une méthode à base de **mélange de lois**. On considère le mouvement apparent dans l'image comme un mélange de différents modèles et il s'agit d'estimer les paramètres de ces lois (estimation des modèles de mouvement) et les affectations des données aux différentes lois (phase de segmentation). L'algorithme EM est couramment utilisé pour l'estimation et la segmentation conjointement, estimant alternativement les modèles, connaissant les affectations des pixels aux modèles, puis mettant à jour ces affectations d'après les nouveaux modèles.

Enfin une troisième classe de méthodes de segmentation par le mouvement est le **regroupement de primitives élémentaires.** Ces méthodes ont pour but de former des ensembles homogènes au sens du mouvement en regroupant des primitives tells que des contour ou des régions. Le regroupement en régions homogènes élémentaires conduit à une **partition de l'image au sens du mouvement**. Pour cela il faut un critère d'homogénéité des régions élémentaires [Bouthemy P., Santillana Rivero J.].

Les problèmes d'estimation et de segmentation à partir de plusieurs images sont plus faciles à résoudre qu'avec seulement deux images, car l'information augmente avec le temps, tandis que le bruit se moyenne, et enfin les ambiguïtés sur l'explication des gradients spatio-temporels de l'intensité par le champ de mouvement 2D sont levées. Des **contraintes de lien temporel** entre les cartes de segmentation peuvent être ajoutées, soit en initialisant la segmentation par une prédiction issue d'une ou plusieurs segmentations à des instants passés ([Bouthemy P., François E.], [Odobez J.M., Bouthemy P. 98]) ou bien en incluant une contrainte dans la segmentation, favorisant la stabilité temporelle [Odobez J.M., Bouthemy P. 98].

4.2.3.2 Méthode proposée par [M. Gelgon]

La méthode proposée par [Gelgon M.] formule la **recherche de régions homogènes en deux étapes : segmentation au sens d'un critère « statistique » et formation de groupes de régions élémentaires cohérents au sens du mouvement.** A ces fins, les **champs markoviens** sont utilisés pour l'étape de segmentation. On suppose qu'il existe une distribution de probabilité de cette segmentation, qu'il faut maximiser, et la segmentation est considérée comme un problème d'étiquetage statistique contextuel. Chaque pixel de l'image se voit affecté d'une étiquette indiquant la région à laquelle il appartient. Soit une grille de sites S correspondant à la grille des pixels, E la champ aléatoire des étiquettes, O le champ des observations sur la grille des sites.
La segmentation recherche le champ des étiquettes $\hat{e}$ le plus probable par le critère du Maximum à Posteriori $\hat{e} \overset{MAP}{=} argmax_{e\in\Omega}\, p(E=e/O=o) \overset{MAP}{=} argmax_{e\in\Omega}\, p(O=o/E=e).\, p(E=e)$ d'après la règle de Bayes, Ω est l'ensemble des configurations d'étiquettes possibles.
$p(O=o/E=e)$ est la vraisemblance conditionnelle exprimant le lien entre les étiquettes, $p(E=e)$ est la probabilité *a priori* du champ des étiquettes. On suppose que E le champ aléatoire des étiquettes est markovien. D'après le théorème de **Hammersley-Clifford** [Geman S., Geman D.], nous avons vu précédemment (§ 3.2) que la distribution de probabilité jointe associée à un champ de Markov est donnée par une distribution de Gibbs : $p(E=e)=\frac{(\exp(-U_2(e)))}{Z}$.
Donc $\hat{e} \overset{MAP}{=} argmax_{e\in\Omega}\, p(O=o/E=e).\, p(E=e)$ est équivalent à $\hat{e}=argmin_{e\in\Omega} U_1(e,o)+U_2(e)$ avec $U_1(e,o)=-\ln(p(o/e))$. La recherche de la carte d'étiquettes optimales est équivalent à minimiser une fonctionnelle d'énergie $U(e,o)=U_1(e,o)+U_2(e)$.
Des champs markoviens peuvent être définis non sur une grille de pixels mais sur un graphe de primitives préalablement extraites de l'images ; ils modélisent les interactions entre ces primitives. Deux champs markoviens sont utilisés chez [Gelgon M.], un au niveau des pixels et un au niveau régions, chaque niveau avec sa phase d'étiquetage.
Une technique de segmentation non supervisée est proposée [Gelgon M.], on ne connaît ni les caractéristiques des régions recherchées, ni le nombre de régions. Pour la première étape de segmentation au niveau des pixels nommé par [Gelgon M.] « segmentation spatiale », les principes décrits par [Bouthemy P., François E.] sont repris et regroupent les critères d'intensité, de couleur ou de texture.
L'énergie est minimisée en parcourant les sites de S selon l'algorithme ICM (« Iterated Conditional Modes »). De façon itérative, l'étiquette permettant la plus grande baisse d'énergie localement (conditionnellement aux autres étiquettes) est recherchée. Pour chaque site visité, soit un ensemble d'étiquettes candidates comprenant l'étiquette courante et les étiquettes des sites voisins. L'étiquette pour laquelle la variation d'énergie locale, par l'affectation de chacune de ces étiquettes, est la plus forte, est affectée au site visité. Chaque fois qu'une étiquette est modifiée, les statistiques des régions concernées sont mises à jour. Chaque site se voit attribué une étiquette de « **stabilité** » [Bouthemy P., Lalande]. Tout site visité et étiqueté

devient stable, si son étiquette a changé, ses voisins deviennent instables. Les statistiques des régions sont remises à jour après chaque changement d'étiquette. Dans [Bouthemy P., François E.], la partition initiale choisie est toute l'image. La première minimisation est une détection de zones non conformes à la caractéristique dominante dans l'image, le mouvement dominant pour [Bouthemy P., François E.].
Le suivi d'une primitive à un instant donné t vers un instant t+1 est réalisé grâce à la prédiction de la position de cette primitive à t+1, à l'aide de sa position à t et d'un modèle dynamique de mouvement. Les primitives sont extraites à t+1, parfois avec l'aide de la prédiction, ou bien la prédiction intervient seulement dans le choix d'association entre mesures et trajectoires en construction. Dans ce dernier cas, des ambiguïtés peuvent apparaître. Des modèles de suivi existent ainsi que des **techniques d'association temporelle de mesures aux trajectoires** présentées au § 4.1.1 et § 4.1.3.
Parmi les techniques basées sur les contours pour le suivi de primitives, citons le suivi par **contour actif** [Blake A., Isard M. 98], le contour correspondant en effet à un contraste d'intensité. Le problème des techniques par contour actif est son initialisation qui doit être proche du vrai contour. [Blake A., Isard M. 98] ont réalisé des travaux sur le suivi de contours dans des conditions difficiles (fond texturé) par des techniques de « **Condensation** » [Isard M., Blake A., 98].

Dans les méthodes de suivi d'une partition représentée par une carte d'étiquettes sur les pixels, les méthodes markoviennes d'étiquetage statistique peuvent mettre à jour de façon incrémentale les régions et les suivre ([Bouthemy P., François E.], [Odobez J.M., Bouthemy P. 98]). La correspondance temporelle région par région est assurée par l'étiquetage.
Le suivi par partitions par propagation d'étiquettes a lieu par une **partition spatiale** de la première image et propagation de celle-ci. Cette partition correspond à des groupes de régions dont le mouvement 2D est cohérent. La connaissance des groupes de régions rend le suivi de la partition plus efficace que le suivi des pixels seuls. **Dans une application de surveillance, il est utile de disposer des cartes de segmentation spatiale de la séquence.** Si les régions spatiales suivies correspondent à une segmentation de l'image en objets identifiés, la connaissance des propriétés de la région spatiale dans laquelle les objets doivent se trouver peut aider à les localiser. Le suivi peut être réalisé par la prédiction et la mise à jour des configurations d'étiquettes définissant les partitions. Les régions sont initialement regroupées au sens du mouvement, c'est la segmentation spatiale au sens du mouvement, on suppose que des pixels proches spatialement ont le même mouvement. Sur chacune des régions, un modèle de mouvement est estimé entre les instants t et t+1 . A partir de la carte des étiquettes $\hat{e}(t)$ (segmentation spatiale) estimée à l'instant t, et des modèles de mouvement estimés sur les régions au sens du mouvement, on construit une carte d'étiquettes $\tilde{e}(t+1)$ qui est une prédiction de la carte de segmentation spatiale à l'instant t+1. L'étiquette spatiale de chacun des sites pixels est projetée au sens du mouvement estimé en ce site, avec un modèle affine de mouvement ([Bouthemy P., François E.], [Odobez J.M., Bouthemy P. 98]) vers un site dans l'image à t+1. La prédiction ainsi établie est la base de la configuration initiale d'étiquettes pour la minimisation de l'énergie $U(e(t+1), o(t+1))$ relative à la segmentation spatiale à l'instant t+1 . La partition spatiale est mise à jour à t+1 par la minimisation de cette énergie.
Cette technique de segmentation spatio-temporelle associée à une phase de suivi « court-terme » fournit des étiquettes aux régions extraites. **L'alternance prédiction et mise à jour établit un lien temporel entre les régions se correspondant dans des images successives.** Cependant, cette méthode est à « mémoire courte ». Dans un cas d'occultation, ou si l'élément suivi est immobile pendant un temps court, le lien temporel peut ne plus exister, ce qui a pour conséquence un changement d'étiquette dans la séquence de cartes de segmentation. Une « piste partielle » est un ensemble de régions extraites liées dans le temps par l'identité de leur étiquette de mouvement. Il s'agit alors maintenant d'identifier des associations entre pistes partielles, formant une piste unique cohérente. Il faut également estimer les **trajectoires** des objets en mouvement dans la scène, en tenant compte des associations entre les pistes partielles. La trajectoire de la région comprend la silhouette de celle-ci ainsi que le modèle d'évolution temporelle, silhouette et région doivent être estimés.
La méthode proposée fonctionne en « batch », elle s'applique une fois l'ensemble des cartes de segmentation obtenues et non au fur et à mesure de l'extraction des mesures. Le traitement immédiat des cartes de segmentation est nécessaire quand il s'agit d'extraire les pistes pour une action immédiate concernant les cibles suivies, ce que permet le MHT ou le JPDAF. L'analyse différée est intéressante s'il n'y a pas d'action

immédiate, et cela permet de tenir compte de toutes les mesures disponibles.

Le domaine de l'extraction multi-pistes concerne le domaine de surveillance radar et sonar. [Cox I.J, Hingorani S.L.] a appliqué ces techniques au suivi dans des séquences vidéo. Le PMHT (probabilistic Multi-Hypothesis Tracking) a été proposé par [Streit R.L., Luginbuhl T.E. 93]. [Gelgon M.] propose d'adapter le PMHT au problème d'estimation et d'association de trajectoires dans un contexte de vision par ordinateur, en reprenant les travaux de [Gauvrit H., Le Cadre J.P.]. [Gelgon M.] propose une phase d'initialisation des pistes, l'introduction d'un modèle géométrique et l'identification du nombre de pistes.
On désigne par mesure une région extraite à un instant donné, avec ses informations. Les trajectoires doivent être estimées à partir d'associations entre mesures et modèles, ces associations sont soit binaires comme le MHT, soit probabiliste comme le JPDAF. Le MHT répond à ce problème en énumérant les associations possibles, en évaluant la pertinence des trajectoires construites pour chacune des hypothèses d'association, et en retenant la piste la plus vraisemblable. **La méthode proposée par** [Gelgon M.] **propose une alternative évitant cette énumération des hypothèses et sa combinatoire.**
L'idée du PMHT est d'affecter toutes les mesures à toutes les pistes avec une certaine probabilité, plutôt que d'affecter de manière unique les mesures aux pistes. On suppose que la source peut être à l'origine de plusieurs mesures, ce qui signifie que les variables d'affectation sont indépendantes. Cette hypothèse du PMHT rend possible la décomposition de la probabilité jointe sur l'ensemble des mesures d'une image, permettant d'éviter l'énumération des hypothèses d'association. [Gelgon M.] prend en compte les liens temporels court terme dans l'estimation des trajectoires par la technique du PMHT.

4.3 Exemple de deux applications probabilistes à base de graphe

4.3.1 Travaux de [Rota N.]

L'interprétation de séquences d'images a suscité de nombreux travaux déjà vus ([Chleq N., Thonnat M.], [Nagel H.-H], [Choi S., Seo Y., Kim H., Hong K.], [Pentland A.]). L'objectif de [Rota N.] est de détecter, **reconnaître et suivre plusieurs personnes** dans une **station de métro avec une seule caméra**. Chaque individu parcourt la scène au cours du temps, et le système de suivi doit suivre sa piste. Une piste est un ensemble de points correspondant aux positions des objets au cours du temps. Les problèmes délicats concernant les pistes sont **l'initialisation, la terminaison, le mélange de pistes et l'éclatement des pistes** (cf. figure 20).

Figure 20 : initialisation, la terminaison, le mélange de pistes et l'éclatement des pistes [Rota N.].

L'initialisation d'une piste correspond au premier élément caractéristique d'un objet entrant dans la scène. La terminaison d'une piste est le dernier élément de la suite des positions de la trajectoire de l'objet, correspondant

à la disparition de l'objet de la scène. Mais dans le cas d'une occultation, la disparition de l'objet ne correspond pas à une disparition réelle. La fusion de pistes correspond par exemple au cas d'un groupe de personnes qui se rejoignent, les pistes définies par leurs trajectoires se fusionnent. A l'inverse, l'éclatement des pistes correspond au cas d'un groupe de personnes qui se séparent.
La première étape de tout système de traitement pour le suivi de personnes est de détecter le mouvement des régions mobiles dans l'image. La reconnaissance correspond à la classification des objets en voiture, personnes, etc.

Le projet **ESPRIT PASSWORDS** est le point de départ de ce travail réalisé en 1996, dont la contribution principale est **l'apport d'information externe concernant les humains et l'apport de connaissances externes sur la scène 3D.** Une fois l'extraction des régions en mouvement par combinaison d'images avec une image de référence, et la reconnaissance de personnes via un modèle de personnes à trois paramètres (vitesse, hauteur, largeur), l'appariement temporel s'effectue via un **graphe tempore**l comprenant les objets de la scène et les filiations entre les détections au cours du temps. Un critère de recouvrement spatial entre deux régions de deux images successives permet de définir si les deux régions appartiennent au même objet suivi. La filiation entre les détections successives met en évidence la correspondance d'un objet d'une image à la suivante. **L'appariement par graphe temporel est toutefois compromis dans des scènes très peuplées, mettant en évidence les problèmes de mélange et d'éclatement de piste.**

Pour rendre le système robuste (cf. figure 21), de l'information contextuelle a été introduite, via un **modèle d'humain 3D**, et le « **contexte statique** », le décor de la scène, que ce soit les objets de décor statique ou sur les personnes à suivre.

Figure 21 : résultat du suivi avec et sans contexte [Rota N.].

En effet, sans information contextuelle, les occultations ne sont pas gérées et le graphe temporel est inexploitable, alors qu'avec l'information de contexte, le graphe temporel résout les problèmes d'occultations. Le graphe temporel pourrait être rendu plus robuste grâce au contrôle du nombre de personnes dans la scène sur une échelle de temps appropriée par exemple.

4.3.2 Travaux de [Han M., Xu W., Gong Y.]

Pour suivi une personne, [Han M., Xu W., Gong Y.] proposent un **suivi multi hypothèses** qui intègre le processus de détection dans le processus de suivi, et la **trajectoire globale** est recherchée dans les **multiples hypothèses**. Des hypothèses sont détectées et générées. Un modèle d'observation autorise le suivi de multiples trajectoires. **Le suivi de trajectoires multi hypothèses** utilise un HMM qui maximise la probabilité jointe entre la séquence des états et la séquence d'observations. Chaque objet suivi est représenté par son index et son état à l'instant t par un vecteur comprenant sa localisation, sa vitesse, son apparence et son échelle. La probabilité jointe d'une séquence d'états donné X et d'une séquence d'observation Z est supposée sous

l'hypothèse markovienne. L'espace des trajectoires possibles est très grand et le problème est résolu dans les algorithmes de suivi multi-hypothèses (MHT) pour de petites cibles ([Cox I.J, Hingorani S.L.], [D.B. Reid.]), en trouvant toutes les combinaisons possibles des observations courantes et des trajectoires existantes à l'intérieur de groupes de points.
Le suivi de trajectoire peut traiter les difficultés temporelles causées par des fonds texturés, des interactions multi-objets et des occultations.

L'image sert au calcul de la vraisemblance et donne une mesure de comment une configuration, incluant le nombre d'objets et leurs états, **explique** les pixels d'avant-plan. L'image permet de restreindre la détection des objets à la recherche uniquement dans les zones de l'avant-plan afin de réduire les calculs. Elle fournit l'information d'apparence des objets qui va aider au suivi. Le détecteur renvoie une boîte englobante dont la taille correspond à l'échelle donnée par le meilleur score de détection à chaque localisation. L'apparence de l'objet à cette localisation est représentée par l'histogramme coloré calculé dans la boîte englobante. Divers systèmes de détection et suivi de multiples personnes sont basés sur la silhouette [Haritaoglu I., Harwood D., Davis L.S. 99], les modèles de forme [Zhao T., Nevatia R., Lv F.].

Un modèle d'observation est composé de l'image originale, d'un masque de non fond et d'une carte du score de la détection d'objets généré par un détecteur d'objets. L'image fournit l'apparence des objets pour les relier au cours du temps. Le modèle du fond et le détecteur d'objet sont utilisés pour rendre une décision basée sur l'image tandis que le suivi de trajectoire rend une décision globale sur le nombre et la configuration des objets. Le masque d'avant-plan, généré par un modèle du fond de mélange de gaussiennes [Stauffer C., Grimson W.E.L.b], permet à la vraisemblance de considérer la présence de divers objets. La carte de détection, consistant en un score de détection des objets basés pixel, fournit des indices pour localiser les objets. Une méthode de détection des objets permet de générer la carte de détection : un réseau de neurones pour la détection des piétons [Le Cun Y., Bottou L., Bengio Y., Haffner P.] recherche à chaque pixel à différentes échelles un score de détection. Le score de détection correspond au meilleur score parmi toutes les échelles. Un algorithme d'inférence conduit par la détection est proposé. Il utilise un module de détection pour générer des hypothèses d'objets et exploiter les informations image afin de suivre les identités des objets et résoudre les interactions multi-objets.

Le module de suivi accumule les résultats de la détection dans une **structure de graphe** et maintient de multiples hypothèses des trajectoires des objets. Le module de suivi comporte trois étapes :
-la génération d'hypothèses;
-le calcul de la vraisemblance;
-le management d'hypothèses.
Les noeuds du graphe représentent les résultats de la détection des objets. Les liens du graphe sont établis en fonction de la similarité entre deux noeuds correspondants à deux objets identiques détectés. Le suivi de plusieurs objets rend une décision globale sur les trajectoires des objets en sélectionnant l'hypothèse qui s'est accumulée au cours du temps la plus probable.
Les **comportements anormaux** pour la vidéo surveillance sont ainsi détectés.
Plus précisément, les divers modules se décomposent comme suit.
-Dans le module de **génération d'hypothèses**, une structure par graphe est maintenue dans l'algorithme de suivi d'objets multiples pour chacune des trajectoires. Les noeuds du graphe représentent les détections. Chaque noeud est composé de la probabilité de détection des objets, sa taille et son échelle, sa localisation et son apparence. Un histogramme par boîte englobante représente l'apparence de l'objet. La force de chaque lien du graphe est calculée en fonction de la proximité, similarité en taille et en apparence entre deux noeuds (objets détectés). Le graphe est continûment étendu à travers le temps pendant le suivi. A chaque image, les résultats de la détection d'objets étant donné, la génération d'hypothèses calcule les connections entre les noeuds du graphe maintenu et les noeuds dans l'image courante. La génération d'hypothèses évite les occultations par séparation et regroupement de noeuds, car si un objet réapparaît après une occultation, le noeud précédent se sépare en deux traces d'objets. Dans l'autre sens, si un objet est en occultation, le noeud correspondant est regroupé avec le noeud occulté (cf. figure 22). Ce module traite également les données manquantes, et les fausses détections.

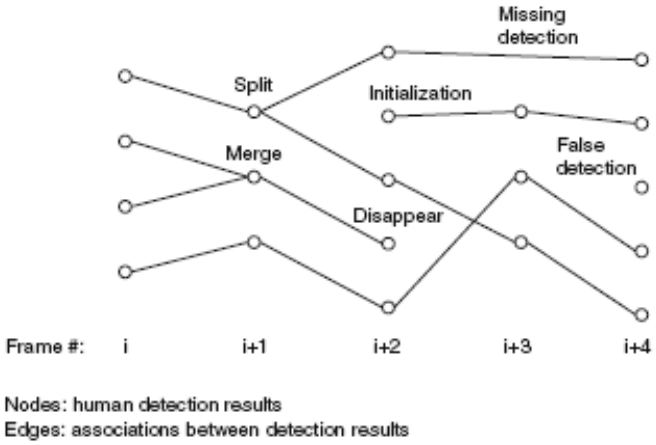
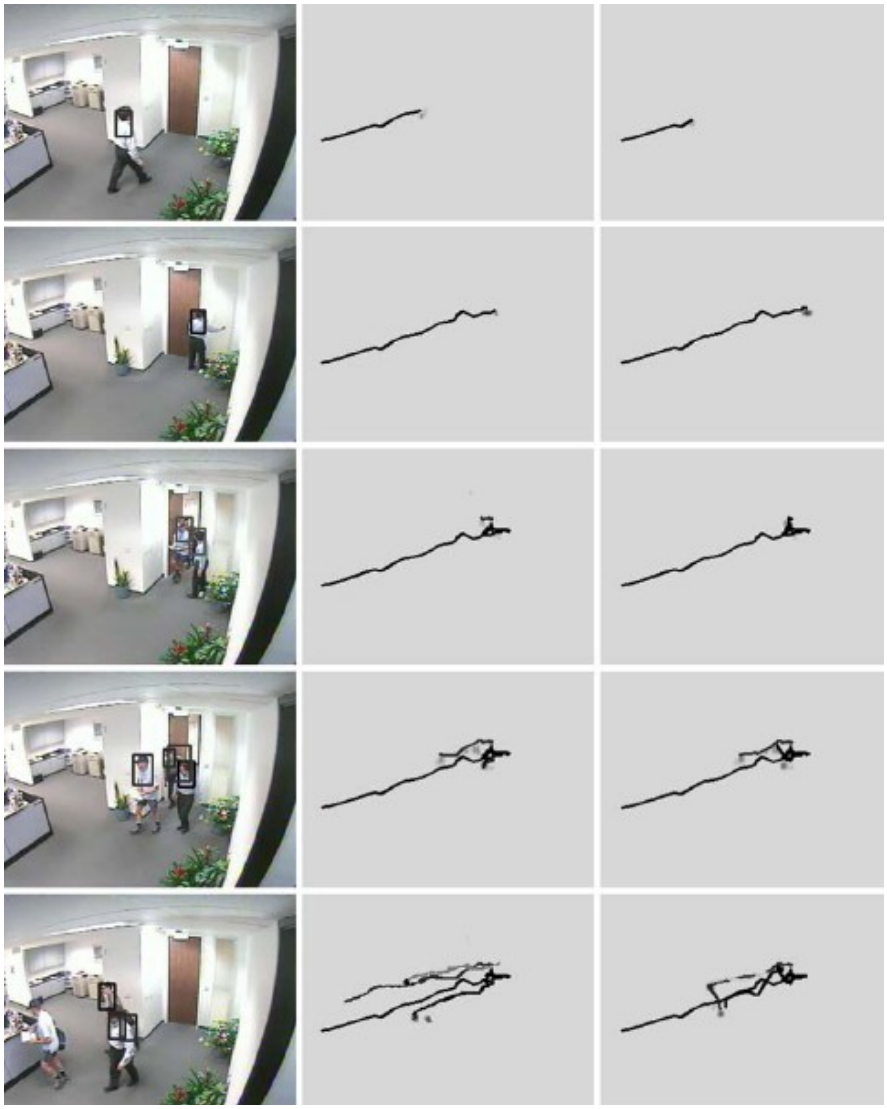


Figure 22 : structure de graphe d'une trajectoire multi objet [Han M., Xu W., Gong Y.].

-La **vraisemblance** ou probabilité de chacune des hypothèses générées à l'étape de génération d'hypothèses est calculée selon la probabilité de détection, et l'analyse de la trajectoire. Le graphe de structure permet d'inclure les objets détectés les plus récents et de générer de multiples hypothèses sur les trajectoires. Une image de vraisemblance est calculée afin de fournir une probabilité à chacune des hypothèses. Les probabilités calculées sur toute la séquence d'images correspondent à la vraisemblance des hypothèses. La vraisemblance des hypothèses est calculée à chaque instant, elle fournit une description globale des résultats de la détection. Les hypothèses avec les vraisemblances les plus fortes sont composées des meilleures détections d'objet. Les vraisemblances des hypothèses sont accumulées à travers la séquence d'images.
La probabilité des observations sachant un état caché, décrit comment un état (caché) du système ressemble aux observations. Une **fonction de vraisemblance basée objet** est calculée comme le score de mise en correspondance entre la représentation de l'objet avec l'image au lieu où se trouve l'objet. Une telle fonction de vraisemblance n'explique pas à elle seule toute l'image. Mais d'un autre côté, une **fonction de vraisemblance basée image** explique chaque pixel dans l'image grâce aux états objets. L'avantage d'une vraisemblance basée image est que si le suiveur est dans une mauvaise localisation, comme un fond texturé, la vraisemblance est faible car la vraie cible ne peut pas être expliquée avec les autres objets. Une fonction de vraisemblance composée d'un terme de vraisemblance basé objet pour l'image originale, et d'un terme de vraisemblance basé image pour le masque d'avant plan et la carte de détection, est proposée. En combinant les trois termes de vraisemblance, celle de l'image sachant la séquence d'états, celle de la carte de détection et celle du masque d'avant-plan, l'algorithme de suivi d'objets multi trajectoire (plusieurs objets) préfère les pistes qui sont des connections d'objets détectés avec de grands scores de détection, des apparences similaires au cours du temps, et explique bien les régions d'avant-plan. Les indices visuels forts rendent la configuration de la séquence, ayant la meilleure probabilité des états observés joints, survivante à des détections manquantes ou fausses grâce à la vue globale de la séquence d'images, des occultations et des fonds texturés (cf. figure 23).
-Le module de **management des hypothèses** range les hypothèses en fonction de leur vraisemblance. Afin d'éviter une explosion combinatoire du nombre d'hypothèses, la structure de graphes manage de multiples hypothèses et effectue un élagage pour avoir des performances raisonnables. Les détections successives sont vérifiées, par des prédictions de la localisation des objets dans les images successives. Cette vérification donne un meilleur score de probabilité aux objets détectés qui ont vérifié la prédiction. Un nombre limité d'hypothèses est ainsi maintenu dans la structure de graphe.
Le module de suivi fournit une prédiction au module de détection d'objets pour améliorer les performances de détection locale.



(a) Image originale. (b) résultats de suivi avec la meilleure probabilité jointe état-observation (trois traces incluses). (c) résultats de suivi avec la moins bonne probabilité jointe état-observation (seulement deux traces incluses).

Figure 23 : résultats de suivi de trajectoire de personnes avec des détections manquantes et des fausses détections [Han M., Xu W., Gong Y.].

Chapitre 2 – Suivi

1 Analyse du mouvement

Les approches pour l'analyse du mouvement ([Abrantes A., Marques J., Lemos J.], [Bar-Shalom Y., Fortmann T.E.], [Cohen I., Medioni G.], [Collins R., et al.b], [Cox I.J, Hingorani S.L.], [Isard M., Blake A., 98], [Isard M., Mac Cormick J.P.], [Jensen F.b], [Stauffer C., Grimson W.E.L.b], [Haritaoglu I., Harwood D., Davis L.S. 00]) sur de longues séquences peuvent se diviser en deux catégories, soit la séquence entière est utilisée et ensuite le mouvement dans cet espace (espace spatio-temporel, espace des phases, etc...) est analysé, soit le suivi a lieu image après image et le résultat obtenu à l'image précédente est inclus dans l'analyse du mouvement courant. Ce chapitre se focalise principalement sur l'analyse du mouvement humain même si la plupart des techniques sont utilisables dans d'autres contextes.

1.1 Analyse du mouvement dans la séquence

1.1.1 Espace spatio-temporel

Dans **l'espace spatio-temporel XT**, [Adelson E. H., Noyogi S. A.] effectuent la reconnaissance d'un piéton dans un volume XYT. Comme il est difficile de segmenter la silhouette d'une personne dans une seule image, les informations de niveau de gris pouvant être peu caractéristiques, l'information temporelle génère une bande inclinée du corps en mouvement segmenté par les contours actifs. La tête d'une personne marchant parallèlement à la caméra et en translation génère dans le plan spatio-temporel une bande inclinée. Une fois le contour de la personne extrait, une technique simple de reconnaissance est basée sur la distance euclidienne entre le contour extrait et la trace caractéristique dans le plan XT d'un piéton.

1.1.2 Espace des phases

Dans **l'espace des phases**, [Aaron Bobick, Lee Campbell] reconnaissent le mouvement humain dans des domaines sportifs où des catégories de mouvement sont bien définies (athlétisme, danse, etc). Cet espace est le produit de l'espace ordinaire (x, y, z) par l'espace des vitesses. Un point matériel est repéré dans cet espace par les coordonnées (x, y, z) de son vecteur position r ainsi que par celles de son vecteur vitesse v, notées (v_x, v_y, v_z). L'idée fondamentale est qu'il est possible de reconnaître un mouvement simplement à l'aide des contraintes du mouvement (par exemple les bras sont attachés aux épaules). En cherchant les contraintes produites par un mouvement et qui ne sont valables que pour ce mouvement, il est envisageable de trouver un modèle caractéristique pour chaque mouvement. Cet espace a l'avantage d'être invariant par rapport aux changements de vitesse. Le modèle de mouvement est appris pour reconnaître les neuf mouvements fondamentaux du ballet classique.

1.1.3 Espace des échelles

Dans **l'espace des échelles**, [Rangarajan K., Allen W., Shah M.] reconnaissent la différence entre deux objets de même forme mais de mouvement différent, ou entre deux objets de même mouvement mais de forme différente. L'idée de base est de considérer que si un objet a un mouvement prédéfini, les trajectoires de plusieurs points sur un objet peuvent servir pour identifier de façon unique l'objet. L'entrée est un ensemble de trajectoires 2D provenant d'un objet suivi à travers une séquence d'images. La structure et les trajectoires 3D de chaque objet sont stockées dans le modèle. Une mise en correspondance est effectuée entre les projections 2D des trajectoires 3D du modèle et les trajectoires 2D afin de déterminer s'ils représentent le même objet. Les trajectoires 2D sont converties en 2 signaux 1D basés sur la vitesse et la direction. Les signaux sont ensuite représentés par des images échelle-espace pour simplifier la mise en correspondance et parce que cette représentation est invariante par rotation et par translation.

1.1.4 Intégration temporelle

Par **intégration temporelle**, [Polona R., Nelson R. 94b] proposent une technique non-paramétrique de détection de périodicité afin de distinguer la marche, la course, le saut à pied joint, la balançoire, etc. Une analyse fréquentielle de l'intensité lumineuse le long des trajectoires associées au mouvement d'ensemble est

effectuée. La fréquence fondamentale donne la période du mouvement relatif. Le volume XYT est partitionné en cellules régulières, dans chacune d'elles un vecteur d'attributs est calculé, et comparé aux attributs des mouvements modèles, afin d'en établir une mesure de similarité indiquant le plus proche.

1.2 Analyse du mouvement image par image

Le suivi peut s'effectuer sur une ou plusieurs personnes [Zhao H.-X., Huang Y.-S.], avec une ou plusieurs caméras [Khan S., Shah M.], pour la détection et l'analyse des mouvements de la foule [Beymer D.], dans le cas de situations temps réel et de traitement simultané de divers flux vidéos [Ruiz-del-Solar J., Shats A., Verschae R.], avec des occultations entre objets ou la rencontre entre plusieurs personnes, et aussi avec des caméras mobiles non calibrées.

Les pionniers dans le domaine du suivi de personnes [Siebel N.T.] sont [O'Rourke J., Badler N.] et [Hogg D.]. Les régions issues du suivi sont classées en individus, groupes de personnes et d'autres classes d'objets. La sortie du suivi sert à construire un **graphe** de suivi facilitant le suivi d'individus sur une longue période même s'ils rejoignent ou bien quittent des groupes. Dans le domaine de la surveillance, il existe bon nombre d'algorithmes de suivi ([Baumberg A.M.], [Bremond], [Cai, Q., Mitiche, A., Aggarwal, J.K**],** [Gavrila D.M., Davis L.S.], [Haritaoglu I., Harwood D., Davis L.S. 00], [Johnson N.], [Khan S., Javed O., Rasheed Z., Shah M.], [Lipton A.J., Fujiyoshi H., Patil R.S.], [Sidenbladh H., Black M.J., Fleet D.J.], [Wren C.R., et al.]). Les systèmes de surveillance comme ceux proposés par [Hongeng S., Bremond F., Nevatia R.], [Pentland A., Liu A.] et [Xiang T., Gong S., Parkinson D.] présentent quelques difficultés pour analyser le comportement dans des scènes de bagarre ou vandalisme vues de plusieurs caméras, et dans des environnements texturés telles que les scènes de métro.

L'approche suivie par [Rohr K.] pour analyser une démarche est de reconnaître une personne dans une image et de suivre les mouvements de ses membres dans chaque image. Le corps humain est modélisé par un ensemble de cylindres articulés. Les régions correspondants à des objets en mouvement sont extraites grâce à la détection des changements temporels de l'intensité. La position 3D de la personne est déterminée par la projection des contours du modèle (des lignes droites) sur les contours dans l'image approchés par des lignes droites. L'approche est étendue à une séquence d'images en estimant les paramètres dynamiques du modèle. L'estimation de ces paramètres dans les images successives est faite en utilisant un filtre de Kalman, autorisant le suivi sur une séquence d'image.

1.3 Les méthodes de suivi image par image

La première étape de tout système de traitement de séquence d'images pour le suivi de personnes consiste à détecter le mouvement des régions mobiles dans l'image.

Nous pouvons classer les méthodes de suivi en six catégories :
-Catégorie 1: les méthodes, parfois sans modèle, basées **région ou suivi de « blobs »**, basés sur la couleur, la texture, les primitives ponctuelles, les contours ([Bremond], [Cai, Q., Mitiche, A., Aggarwal, J.K**],** [Khan S., Javed O., Rasheed Z., Shah M.], [Lipton A.J., Fujiyoshi H., Patil R.S.], [Wren C.R., et al.]);

-Catégorie 2 : les méthodes utilisant un **modèle d'apparence 2D** du corps humain ([Baumberg A.M.], [Haritaoglu I., Harwood D., Davis L.S. 00], [Johnson N.] avec son modèle d'apparence temporelle), les **approches 2D avec modèle explicite de la forme**, et les **approches 2D sans modèle explicite de la forme.**

-Catégorie 3 : les méthodes avec un **modèle articulé en 3D** du corps humain ([Gavrila D.M., Davis L.S.], [Sidenbladh H., Black M.J., Fleet D.J.]);

-Catégorie 4 : Les méthodes par **soustraction du fond** procède par soustraction de l'image courante avec une image du fond ([Haritaoglu I., Harwood D., Davis L.S. 98], [Wren C.R., et al.]). Le système peut être plus robuste dans des environnements texturés en combinant la couleur, la texture, et le mouvement pour segmenter

l'avant-plan.
Voyons en détail les travaux de [Ali M.A., Indupalli S., Boufama B.] qui utilisent une méthode par soustraction du fond. [Ali M.A., Indupalli S., Boufama B.] font de la détection de personnes en mouvement et du suivi dans un environnement complexe avec un fond inconnu, pour la vidéo surveillance. Une méthode de mise en correspondance des primitives des blobs par corrélation dans une séquence d'intérieur est proposée. Le fond est modélisé par une méthode statistique et remis à jour continuellement. La segmentation des objets d'avant-plan est effectuée par un algorithme de soustraction du fond et un algorithme de clustering K-means. L'espace HSV (invariance en luminance) est utilisé pour minimiser l'effet des ombres. Pour le suivi, la plupart des travaux font de la prédiction sur les primitives et comparent les valeurs prédites et estimées pour remettre à jour le modèle, via un filtre de Kalman. [Ali M.A., Indupalli S., Boufama B.] présentent une méthode par corrélation de « Pearson » : après avoir détecté les blobs, les primitives sont extraites et comparées avec les primitives de blobs dans les images précédentes via la corrélation de « Pearson ». Les occultations ont été résolues par des boites englobantes autour des blobs et de l'information de mouvement;

-Catégorie 5 : La différence temporelle (deux ou trois images) [Anderson C., Burt P., Van Der Wal G.].
Les méthodes à base de différence des catégories 4 (soustraction du fond) et 5 (différence temporelle) calculent une carte binaire de mouvement, et les pixels de mouvement sont regroupés en « blobs », régions de pixels connexes ([Haritaoglu I., Harwood D., Davis L.S. 00], [Jabri S., Duric Z., Wechsler H., Rosenfeld A.], [Zhao T., Nevatia R., Lv F.]). Les mouvements et les interactions entre les personnes sont obtenus par le suivi des « blobs ».
La différence temporelle est bien adaptée aux environnements dynamiques mais souffre du « **problème d'ouverture »** dû aux couleurs homogènes d'objets en mouvements et effectue une mauvaise extraction des primitives. La soustraction du fond permet d'extraire les objets en mouvements mais le fond doit être bien modélisé, et cette méthode est très sensible aux changements de lumière ou aux mouvements des objets dans le fond. Le flot optique est une technique très robuste, même en présence de mouvement de caméra, mais est très chère en coût de calcul et donc peu usitée pour les applications temps réel. Seule la soustraction du fond requiert une modélisation du fond (des gaussiennes ou mélanges de gaussiennes), et est plus rapide que les autres méthodes;

-Catégorie 6 : Une autre approche complémentaire aux catégories 5 et 6, est **l'approche différentielle à base d'estimation du champ de vitesse** en tous points de l'image, aussi dite par **détection de mouvement**. Elle consiste à connaître les vecteurs vitesses dans la scène, en faisant l'hypothèse d'invariance entre t et t+dt, c'est-à-dire que la fonction d'intensité lumineuse en un point (x, y, z) est identique en (x+dx, y+dy, t+dt). On définit une fonction d'erreur appelée DFD « Deplaced Frames Difference » : $DFD(x, y, t) = f(x, y, t) - f(x+dx, y+dy, t+dt)$ et $DFD(x, y, t) = 0 \; \forall x \; \forall y \; \forall t$, équation connue sous le nom « Equation de Contrainte du Mouvement Apparent (ECMA) ». On cherche à minimiser la DFD pour tout (x, y) de l'image à l'instant t. Cette famille comprend la méthode par « **flot optique** ». L'estimation du mouvement par « **flot optique** » [Barron J.L., Fleet D.J., Beauchemin S.S.] en fonction des variations spatio-temporelles de la fonction d'intensité lumineuse est une façon d'appréhender le mouvement dans une scène. La détection de mouvement met en évidence des régions mobiles dans l'image courante. La soustraction du fond est bien adaptée pour les environnements intérieurs dans lesquels la lumière est stable et les mouvements d'arrière plan peu nombreux, tandis que la détection de mouvement par « **flot optique** » correspond aux environnements texturés avec des mouvements dans le fond.
Enfin, il existe dans cette même catégorie, les approches basées sur la **corrélation à base de similitudes spatio-temporelles pour l'estimation du mouvement d'ensemble.**

Nous présentons dans la suite les méthodes basées primitives, les approches basées modèle du corps en 2D ou 3D, les méthodes avec modèle d'apparence en 2D, les approches en 2D avec modèle explicite de la forme, les approches en 2D sans modèle explicite de la forme, et les méthodes avec modèle articulé en 3D.

1.3.1 Les méthodes basées primitives

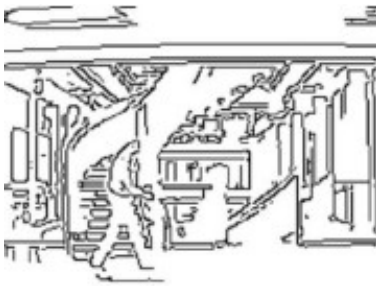
Dans les **méthodes de suivi de caractéristiques**, l'objectif est de détecter des descripteurs liés à des points

particuliers, et décrivant l'objet par un ensemble d'attributs géométriques (points, segments, courbes paramétriques, arêtes, contours), ou des régions de l'image. Ces méthodes ont l'avantage d'une bonne robustesse aux occultations car des associations qui n'ont pas pu se faire sur certains points de l'objet cachés dans l'image, ne mettent pas en échec le suivi sur l'ensemble des points.
[Du L., Sullivan G., Baker K.] et [Koller D., Daniilidis K., Nagel H.-H] détectent des véhicules par extraction des angles du toit et du capot. Ces primitives peuvent aussi être la taille, la position, la vitesse, le rapport des deux axes de l'ellipse. Elles sont obtenues par une extraction de contours ([Deriche R.], [Shen J., Castan S.]) et analyse de la norme du gradient de l'image. Cette approche a l'avantage d'avoir un contenu sémantique (points précis sur les objets, comme le toit d'une voiture). Le suivi de **primitives ponctuelles** est un procédé de mise en correspondance d'une image à la suivante. Les primitives des « blobs » sont extraites pour une mise en correspondance dans la séquence, par la distance euclidienne ou l'approche basée corrélation. La **trajectoire** peut alors être évaluée par le regroupement de ces primitives tout au long de la séquence. Cette approche n'identifie que quelques points sur l'objet suivi et non l'objet en entier, ce que font les approches contour et région.

Le suivi par **contours** appelé « contours actifs » suit le bord de l'objet, il suffit pour cela qu'il ait assez de contrastes au niveau de ses contours ou bien à cause de son mouvement. Des boites englobantes représentent le contour externe des objets remis à jour dynamiquement dans les images successives. Cette approche est sensible à l'initialisation et limitée en terme de précision de suivi.
Dans les approches de **suivi par l'apparence basé contour**, les contours actifs appelés « snakes » estiment la frontière de l'objet à chaque instant mais ils sont très sensibles à l'initialisation du contour. Une autre approche pour le suivi de contours consiste à l'approcher par un ensemble de points et le suivi est rendu possible par l'utilisation de Modèles de Markov Cachés ou HMM. [Chen Y., Rui Y., Huang T.S.] représentent le contour par une ellipse, chacun des points représente un état du HMM. L'objet d'intérêt est suivi au cours du temps grâce à son contour, soit par mise en correspondance du contour de l'objet soit en suivant le contour. La mise en correspondance du contour, dans une approche descendante, consiste à la minimisation d'une distance entre les positions du contour entre deux instants successifs.

Les approches descendantes recherchent directement le corps humain ([Dimitrijevic M., Lepetit V., Fua P.] [Mori G., Malik J.]) à partir de la mise en correspondance entre l'image et le « template ». Un mélange d'arbre représente le corps pour gérer les occultations [Ioffe S., Forsyth D.A., 03], ou des gabarits spatio-temporels pour détecter la marche d'une personne [Dimitrijevic M., Lepetit V., Fua P.] (cf. figure 24).

(a) Image originale.



(b) Contours d'après l'algorithme de Canny.

(c) Gabarit utilisé pour la détection de piétons.

Figure 24 : Template matching sur les contours ([Dimitrijevic M., Lepetit V., Fua P.], [Noriega P. a]).

L'approche par **région** se caractérise par l'extraction dans l'image courante de régions dénommées « blobs », ensemble de pixels connexes et regroupés en fonction d'un critère déterminé, par exemple les pixels dont la valeur est différente avec ceux de l'image précédente, et le suivi des régions homogènes au cours de la séquence. Cette méthode est basée sur la variation du mouvement dans les régions de l'image. Elle ne résout pas les occultations entre objets. L'hypothèse est faite qu'à l'intérieur d'une région, l'apparence est invariante et

le mouvement est homogène, par exemple le suivi de « blobs » par filtrage de Kalman [Crowley J.L., Demazeau Y.]. [Chleq N., Thonnat M.] et [Baumberg A., Hogg D.] utilisent la différence absolue entre l'image courante I_t et une image de référence I_O : $I_{résultat}=|I_O-I_t|$. L'inconvénient de cette méthode est la mise à jour de l'image de référence I_O. L'autre méthode très usitée est la différence d'images successives [Jain R., Martin W., Aggarwal J.] : $I_{résultat}=Max(|I_t-I_{t-1}|, |I_{t-1}-I_t|)$, lorsqu'on ne dispose pas d'image de référence, mais elle ne prend pas en compte les mouvements des régions uniformément colorées. Seules les régions texturées sont détectées [Ricquebourg Y. 93].
Dans les approches de **suivi par l'apparence basé régions**, les méthodes à base de densité de probabilité modélisent la répartition colorimétrique ou texturale sous forme par exemple de mélange de gaussienne. [Fieguth P., Terzopoulos D.] calculent la couleur moyenne de la boite englobante de l'objet suivi. [Perez P., Hue C., Vermaak J., Gangnet M.] font du suivi basé apparence par le calcul des similarités entre des histogrammes dans l'espace HSV, obtenue avec le coefficient de Bhattacharya. Le suivi a lieu avec l'algorithme de « condensation » [Isard M., Blake A., 98], estimant la densité du vecteur d'état de l'objet suivi. Dans le contexte de la vidéo surveillance, un modèle d'apparence proposé par [Haritaoglu I., Harwood D., Davis L.S. 00] W^4 est appris en même temps que des personnes sont suivies. Le fond est soustrait, mettant en évidence les régions en mouvement qui sont mises en correspondance au cours du temps. Le modèle d'apparence est constitué d'un prototype de forme représentant la probabilité qu'un pixel appartienne à la personne, et un prototype de texture contenant les informations d'intensité lumineuse et de texture. Au fur et à mesure du déroulement de la séquence, le modèle d'apparence est formé dynamiquement intégrant l'aspect temporel du suivi. Il existe des modèles d'apparence qui varient avec le point de vue [Black M.J., Jepson A.D.].

Le système W^4 [Haritaoglu I., Harwood D., Davis L.S. 00], avec une seule caméra en niveau de gris ou en infra rouge, pour la détection, le suivi et la surveillance (2000) temps réel analyse ce que les personnes font (what), où (where), quand (when) et qui (who) le fait. Le système suit la tête, le torse, les bras et les jambes d'une personne debout en temps réel. L'objectif est de suivre des blobs de l'avant-plan par une approche basée primitives, avec des images basse résolution, nécessitant un détecteur de mouvement précis et très robuste. La détection des personnes en mouvement a lieu par soustraction de l'image courante avec un modèle du fond gaussien bimodal. Les histogrammes des blobs d'avant-plan sont projetés le long des axes principaux. Une mesure de similarité les compare avec des histogrammes appris, déterminant s'il s'agit de simple personne ou d'un groupe. Les membres du corps sont suivis par un modèle de mouvement pour la position et par une mise en correspondance des silhouettes des « blobs » avec des « prototypes » ou « template » de texture temporelle, et des techniques de corrélation prédisent les occultations des membres du corps dans l'image suivante. W^4 suit correctement et étiquette les membres du corps. L'avantage de W^4 est sa généricité, il peut détecter et suivre différentes postures, et cela en temps réel. En revanche, comme le suivi est basé sur des blobs, ils doivent être bien détectés. Du coup le suivi se perd en cas d'ombres, de bruit et de changement d'illumination. Dans le dernier cas le modèle du fond est recalculé.

[Landabaso J.L., Xu L.Q., Pardas. M.] font du suivi de personnes, groupe de personnes ou de voitures pour la vidéo surveillance avec une seule caméra. Les pixels de non fond sont détectés par soustraction avec un modèle du fond adaptatif, composé d'un mélange de gaussiennes, et appris de façon statistique via l'intensité, la couleur, les contours, et les textures. Une analyse de la connectivité des pixels permet de les regrouper en « blobs ». Les blobs sont suivis via des « template » temporelles comprenant des primitives caractéristiques : vitesse, taille, ratio, l'orientation des axes principaux de l'ellipse, la couleur dominante. Le « template » de chaque objet suivi donne lieu à un ensemble de filtres de Kalman qui vont prédire les valeurs des paramètres caractéristiques à l'image suivante. Les objets en mouvements qui se rejoignent et se séparent ne sont pas traités.

[Lee M.W., Cohen I.] associent un détecteur de visage pour la recherche du visage, des contours actifs pour détecter les épaules, les blobs de teinte chair pour repérer la teinte chair, et l'axe médian des jambes afin de détecter la tête, épaules, jambes, main (cf. figure 25).

(a) détection du visage | (b) détection des épaules d'aprés un contour actif | (c) blobs de teinte chair | (d) axe médian des jambes

Figure 25 : Association de détecteurs ([Lee M.W., Cohen I.], [Noriega P. a]).

1.3.2 Les approches basées modèle du corps en 2D ou 3D

Pour détecter et identifier les différents membres du corps, il faut avoir un modèle géométrique en 2D (sans ou avec modèle explicite de la forme) ou en 3D. Le suivi avec **modèle explicite** compare les données issues de l'image avec un modèle de l'objet ou de la personne à suivre. Cette méthode requiert le développement d'un modèle 2D ou 3D de la personne, selon l'application.
-Si le modèle du corps est simple, il sera facile à implémenter, rapide, mais sujet aux occultations, et peu précis par rapport aux variations de posture, d'angle, et d'apparence (fonction du point de vue et des occultations). Pour les applications où la capture de la pose n'est pas nécessairement exacte comme le suivi de personnes pour la télésurveillance, l'approche 2D est appropriée. Il en est de même des applications avec une seule personne impliquant des contraintes sur le mouvement et un point de vue simple (estimation de la posture de la main en reconnaissance de la langue des signes face à la caméra, reconnaissance de la marche latéralement à la caméra);
-Les approches 3D correspondent aux applications de suivi de mouvements complexes et non contraints (interactions entre personnes comme se serrer la main, danser ou se battre). La pose du corps humain représentée par des angles 3D est indépendante du point de vue car moins sensibles aux variations dans la taille des personnes. Les approches 3D sont plus exactes et résolvent les occultations et collisions, en revanche elles ne sont pas adaptées au temps réel.

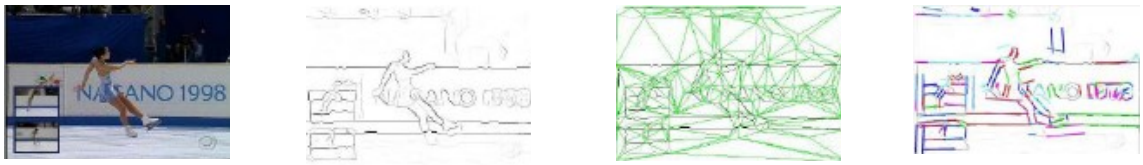
Les méthodes basées modèle sont robustes aux occultations car elles possèdent une connaissance *a priori* d'un modèle de la forme, contrairement aux autres méthodes sans modèle, mais elles demandent un coût de calcul important. L'information structurelle du modèle de la forme sert à mettre en correspondance les données image avec le modèle, soit par une approche ascendante associant des hypothèses images, soit par une approche descendante où on cherche le modèle ayant le maximum de corrélation avec les données image. Mais elles demandent un coût de calcul important.

1.3.2.1 Méthodes de mise en correspondance image/modèle

Dans une **approche descendante**, un modèle et des informations *a priori* au plus haut niveau de la hiérarchie doivent expliquer les observations au bas niveau par mise en correspondance entre un modèle du corps et l'image, et dans le cas de mouvements cycliques tel que la marche, des poses clés servent à la reconnaissance. Le modèle géométrique est donc utilisé de façon directe.

En revanche, dans une **approche ascendante**, les membres du corps sont recherchés à partir des caractéristiques bas niveaux extraits dans chaque image, sans modèle *a priori*. Par la suite, le modèle voit ses paramètres modifier pour tenter de correspondre au mieux aux caractéristiques image. Il permet alors d'identifier les membres candidats. [Ren X., Berg A.C., Malik J.] (cf. figures 26 et 27), dans une stratégie

ascendante similaire à celle de [Mori G., Ren X., Efros A.A., Malik J.] (cf. figure 28), détectent des contours, qui sont décomposés en segments et une triangulation de Delaunay s'appuyant sur ces segments est mise en place.



(a) Image originale (b) Contours (c) Triangulation de Delaunay (d) Affichage des membres candidats

Figure 26 : Détection de membres candidats ([Ren X., Berg A.C., Malik J.], [Noriega P. a]).

(a) Image originale (b) Sélection des membres candidats d'après les critères géométriques sur les segments extraits de l'image, sélection finale des membres (c) Résultat de la pose en 2D d'après des critères anthropomorphiques

Figure 27 : Reconnaissance des membres par une approche montante ([Ren X., Berg A.C., Malik J.], [Noriega P. a]).

(a) Image originale (b) Contours (c) Segmentation par coupe normalisée

Figure 28 : Segmentation des membres ([Mori G., Ren X., Efros A.A., Malik J.], [Noriega P. a]).

Des ensembles d'arêtes parallèles produisent des membres hypothèses. Seuls sont conservés ceux qui vérifient les contraintes du modèle, permettant d'étiqueter les membres. [Haritaoglu I., Harwood D., Davis L.S. 99] détectent la silhouette par suppression du fond et recherchent le déplacement qui maximise la corrélation de deux silhouettes entre deux instants différents dans une stratégie ascendante.

1.3.2.2 Les méthodes avec modèle d'apparence en 2D

L'apparence est un autre indice pour mettre en **correspondance** les objets au cours du temps dans une séquence d'images. L'apparence peut être une caractéristique de couleur, de forme ou de texture. Ces approches de suivi par la texture se dénomment « **Suivi Visuel** » ou « **Visual Tracking** » en anglais. Parmi ces **approches de suivi basées apparence**, nous pouvons considérer les approches région et les approches contours. Dans les approches région, l'apparence d'un objet peut être définie soit par un prototype soit par un modèle probabiliste. Dans le cas du prototype, il s'agit d'une méthode descendante, le maximum de corrélation entre le prototype de l'objet à l'instant t et l'objet dans l'image à t+1 est recherché.

En ce qui concerne le **suivi visuel** d'objets, [Isard M., Mac Cormick J.P.] proposent un suivi bayésien multi « blob » et un filtre à particule pour l'inférence. [Hue C., Le Cadre J.P., Perez P.] décrivent une extension du filtre à particule classique où le vecteur stochastique est estimé par un échantillonneur de Gibbs. Ces algorithmes de suivi sont basés sur le (Maximum à Posteriori) **MAP** marginal. Un algorithme de suivi doit fournir la meilleure séquence des états observés. Mais quand les états dynamiques et les vraisemblances sont des distributions gaussiennes, l'état observé joint est aussi une distribution gaussienne. Dans ce cas la solution MAP de la distribution jointe et la distribution marginale sont identiques, et il n'est pas utile de travailler avec le suivi de trajectoire. Mais pour des distributions générales, le suivi marginal MAP n'est pas une bonne approximation du suivi de trajectoire et le problème est plus accru pour le suivi multi objets.
Parmi les méthodes avec modèle d'apparence en 2D, le système de l'université de LEEDS « People Tracker » de Adam Baumberg (cf. figure 29), sous la supervision de David Hogg [Baumberg A.M.], est basé sur un modèle 2D des contours externes de la personne (modèle d'apparence 2D du corps humain). L'algorithme de suivi, avec une seule caméra, est un modèle de la forme active qui se cale sur les contours d'un piéton, généré via une étape d'entraînement et les contours extraits du modèle sont analysés par analyse en composantes principales. Il fonctionne bien tant que la personne est visible et peu en occultation. Cette méthode 2D est assez rapide pour une utilisation en temps réel.

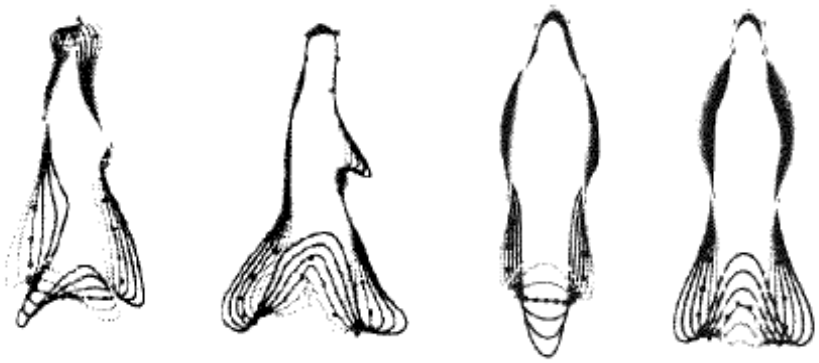


Figure 29 : Le modèle d'apparence d'objet non-rigide de [Hogg D.]

La détection de personnes se fait en plusieurs étapes, du bas niveau au haut niveau. Initialement, le mouvement est initialement détecté par soustraction des pixels avec le fond de l'image et seuillage. Les pixels du fond sont remis à jour dans l'image courante si leur valeur varie sur plusieurs images consécutives. Les zones de l'image correspondant à des personnes couramment suivies sont masquées du détecteur de mouvement, mais pas de la remise à jour du fond. L'objectif est de détecter des mouvements seulement pour les nouveaux objets tandis que les anciens objets continuent d'être suivis. Dans une seconde étape, la forme est initialisée par une fonction B-Spline. Enfin, le modèle est automatiquement généré dans une phase d'entraînement via un ensemble de vidéos contenant des piétons. Une analyse en composante principale sur la silhouette des données d'entraînement met en exergue divers modèles. Un détecteur de contours extrait les contours et la distance de Mahalanobis évalue la proximité avec le modèle. A chaque nouvelle personne détectée dans l'image, les paramètres du modèle PCA ainsi que la position de la personne dans l'image sont

stockés et réutilisés pour la partie suivie. Le suivi des personnes dans les différentes images a lieu via un modèle de mouvement du second ordre, le filtre de Kalman modélisant la vitesse et l'accélération de la personne suivie et prédisant la position dans l'image courante. La position initiale estimée ainsi que les paramètres de la forme courante constituent un point de départ pour la détermination de la position et du contour dans l'image courante.

L'avantage de « People Tracker » est sa robustesse et sa rapidité, mais les personnes qui marchent, celles qui sont assises, les groupes, et les personnes seules ayant un faible contraste avec le fond, ne sont pas détectées. Le suiveur s'initialise mal lorsque les personnes entrent par deux ou plus dans la scène, et s'il perd des personnes suivies, une fois qu'elles seront suivies de nouveau, il n'y aura pas reconnaissance possible de la personne.

1.3.2.3 Approche en 2D avec modèle explicite de la forme

Les approches 2D avec modèle explicite de la forme ont une connaissance *a priori* du corps humain en 2D.
Le modèle peut être une **figure en bâtons** (« fils de fer » [Karaulova I.A., Hall P.M., Marshall A.D.]), entourées de rubans ou « blobs ». La silhouette du corps est détectée par soustraction du fond, en supposant le fond stationnaire et la caméra fixe. Les régions homogènes sont identifiées par la couleur ou la texture. Le modèle 2D contient les contraintes d'articulations entre les régions correspondants aux membres du corps humain. Le mouvement est détecté et le fond est séparé des objets en mouvement à chaque nouvelle image. La teinte chair permet de détecter le visage et les mains. Le filtre de Kalman [Rigoll G., Eickeler S.] permet d'estimer les paramètres du modèle tout au long de la séquence. Les modèles de Markov [Rigoll G., Eickeler S.] et le filtrage particulaire [Chen Y., Rui Y., 2004] sont des techniques de modélisation statistique des paramètres du modèle.

Dans les approches **descendantes** (haut/bas), le modèle de la pose du corps va permettre d'estimer la vraisemblance des hypothèses. Les modèles de contours 2D comme celui proposé par [Ju S., Black M., Yacoob Y.] modélisent le corps humain par un modèle « cardboard » (cf. figure 30).

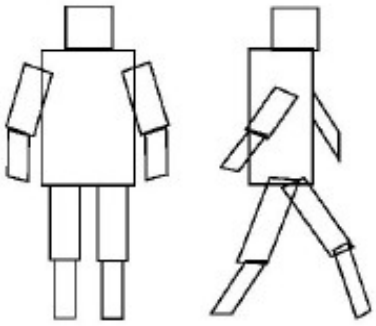
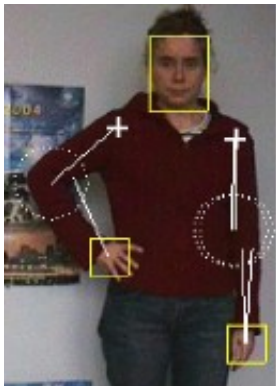


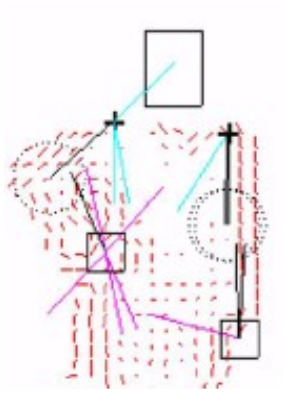
Figure 30 : Modèle « cardboard » : les membres de la personne sont représentés par des « patchs » plans [Ju S., Black M., Yacoob Y.].

Les membres du corps sont modélisés par des « patchs » planaires (rectangulaires « cardboard » pour [Ju S., Black M., Yacoob Y.]) ou blobs, reliés entre eux. Les patchs « cardboard » sont composés d'un rectangle pour chacun des membres. Leurs projections dans l'image permettent d'évaluer la vraisemblance des hypothèses en deux dimensions. Un modèle simplifié est celui du « cardboard » proposé par [Cham T.J., Rehg J.M.], composé de rectangles connectés entre eux. Chaque « patch » est suivi au cours du temps avec un modèle de mouvement, Le modèle est utilisé de façon explicite et l'information *a priori* est propagée de manière descendante dans les couches hiérarchiques. [Felzenszwalb P.F., Huttenlocher D.P. 00] ainsi que [Forsyth D.A., Fleck M.M.] ont proposé un modèle articulé 2D dans une approche descendante conduisant à identifier un certain nombre de membres candidats pour chacun d'eux, mais la mesure de similarité pour détecter les

membres est basée apparence avec comme hypothèse forte que les membres portent un vêtement de teinte chair. [Ronfard R., Schmid C., Triggs B.] proposent de remplacer cette hypothèse par un apprentissage à base d'une Machine à Vecteur de Support (SVM).
Dans [Leignel C., Viallet J.E.] (cf. figure 31), les membres sont détectés sous la forme de segments, conduisant ainsi au squelette 2D, mais l'approche est ascendante.



(a) Image originale et les membres supérieurs complets.



(b) Image des gradients « robustes » - Détection des mains par la teinte chair (carrés), gradients robustes en tirets fins déterminant la direction globale (segments) des bras et avant-bras, à partir des épaules (croix) et des mains, et qui se croisent au niveau du coude (cercle).

Figure 31 : Détection des membres supérieurs complets candidats [Leignel C., Viallet J.E.].

1.3.2.4 Approche en 2D sans modèle explicite de la forme

Les approches 2D sans modèle explicite de la forme décrivent le mouvement humain par des caractéristiques 2D bas niveau issues des régions d'intérêt. Les modèles du corps issus de ces primitives bas niveau sont statistiques. Les caractéristiques extraites de l'image sont dans ce cas mises en relation avec la pose de la personne suivie. La « **structure from motion** » permet de retrouver les coordonnées 3D d'une personne suivie au cours du temps grâce aux points 2D en mouvement dans une série d'images prises sous des angles différents. Le codage des contours d'une silhouette extraite de l'image d'après un descripteur de forme de type « **shape context** » (cf. figure 32) [Agarwal A., Triggs B.] permet de comparer l'image avec une base apprise.

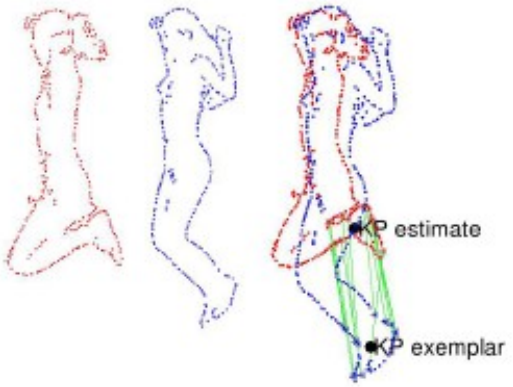


Figure 32 : « **shape context** » - localisation des articulations. Les points échantillonnés le long de la silhouette exemple (à gauche) et de test (au centre) sont mis en correspondance ([Mori G., Malik J.], [Noriega P. a]).

Une approche implicite modélisant la pose humaine consiste à comparer une base de poses apprises avec chaque nouvelle image. Beaucoup d'applications ont vu le jour comme l'estimation de la pose de la main dans la reconnaissance de la langue des signes ou le dialogue basé sur le geste. Une soustraction du fond suivie de la détection de la couleur de la peau permet d'extraire la forme de la main et son mouvement.

1.3.2.5 Les méthodes avec modèle articulé en 3D

Les modèles 3D représentent la structure articulée en trois dimensions, levant les ambiguïtés des modèles 2D dépendant de la pose. Les membres sont modélisés par des cylindres [Hogg D.] ou des cônes [Sminchisescu C., Triggs B. 01], tandis que [D.M. Gavrila and L.S. Davis] font le choix de super quadriques (cf. figure 33).

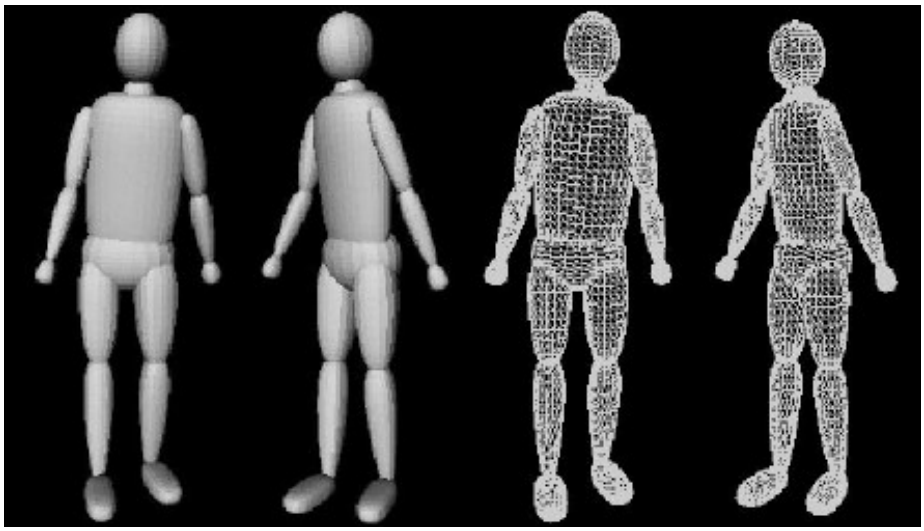


Figure 33 : Modélisation des membres par des primitives ellipsoïdales [Noriega P. a].

Parfois des gaussiennes 3D ou « métasphères » modélisent chaque muscle du corps ([Plänkers R., Fua P. 01] [Plänkers R., Fua P. 03]) (cf. figure 34).

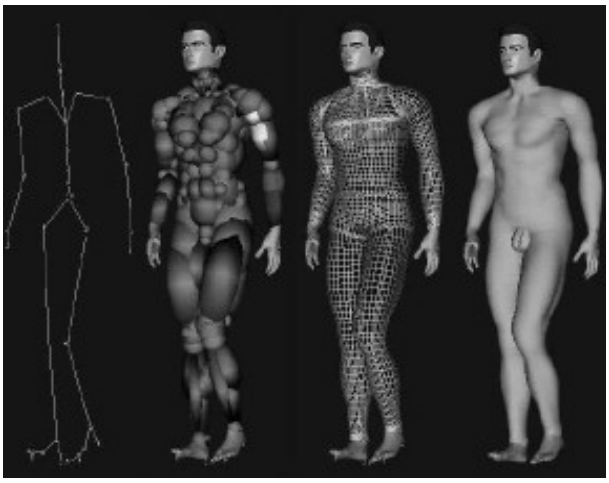


Figure 34 : Modélisation des tissus musculaires avec des primitives gaussiennes ([Plänkers R., Fua P. 01], [Noriega P. a]).

Parmi les méthodes utilisant un modèle articulé en 3D du corps humain pour le suivi de personnes, le modèle de [Gavrila D.M., Davis L.S.] de l'université de Maryland sous la direction de L.S. Davis [Gavrila D.M.,

Davis L.S.] proposent 22 degrés de libertés comprenant des cylindres et des ellipsoïdes, et décrit par les angles d'articulations. Les mesures sont prises avec deux caméras calibrées; dans chaque vue orthogonale, la segmentation d'une forme en 2D est issue du résultat du détecteur de contours.
Ce modèle donne de bons résultats mais il n'est **pas adapté pour des applications de surveillance** du fait qu'il faut **deux caméras stéréo orthogonales** et les occultations ne sont pas prises en compte. De plus, le détecteur de contours a nécessité que les personnes portent des vêtements colorés afin de différencier les différentes parties du corps. Enfin, le temps réel n'est pas envisageable.
Le modèle de Sidenbladh « 3D people tracker » [Sidenbladh H., Black M.J., Fleet D.J.], sous la direction de Michael Black à l'université de Brown, est également un modèle complexe 3D articulé, avec une seule caméra. Le modèle du corps est représenté par un ensemble de cylindres articulés contenant 25 degrés de libertés. Ce modèle de suivi est composé de deux parties : un modèle probabiliste est estimé en utilisant les données d'activités typiques obtenues par un ensemble de données de mouvement 3D. Les mouvements répétitifs, tels que la marche, sont décomposés en une séquence de modèles temporels, les « cycles de mouvement ». La construction de ces modèles temporels provient de la segmentation des données d'entraînement. Puis ce modèle probabiliste est injecté comme « prior » d'une distribution bayésienne d'un filtre à particule (propagation). Le suivi est correct mais la grande complexité du modèle rend l'algorithme très coûteux en temps de calcul, ce qui le rend inapte aux problématiques de vidéo surveillance.

Pour détecter et suivre une personne dans une scène, le modèle est d'autant meilleur qu'il est détaillé, surtout dans les situations difficiles. Le modèle complexe 3D, tel que celui de Sidenbladh [Sidenbladh H., Black M.J., Fleet D.J.], est trop lent pour une utilisation en temps réel. Certains systèmes nécessitent de plus une calibration et un système de caméras orthogonales, comme celui de [Gavrila D.M., Davis L.S.]. C'est la raison pour laquelle, en général, **les systèmes de surveillance visuelle sont basés région ou bien avec un modèle d'apparence 2D.**

Le modèle est ensuite affiné par rapport aux caractéristiques extraites de l'image par un méthode soit **déterministe**, soit **stochastique**, soit par **apprentissage**, soit **à base de règles**, soit **descendante**.

1.4 Les approches pour affiner le modèle

1.4.1 Les approches déterministes

Les approches **déterministes** cherchent le modèle le plus proche des caractéristiques extraites de l'image grâce à l'optimisation d'une fonction de coût, quelquefois sous la forme d'une probabilité [Demirdjian D., Taycher L., Shakhnarovich G., Grauman K., Darrell T.]. Dans cette approche, les produits d'exponentiels de « twists » permettent de suivre le corps humain [Bregler C., Malik J.] avec une seule caméra.

1.4.2 Les approches stochastiques

Les approches **stochastiques** sont nécessaires lorsqu'il existe des imperfections dans modèle et des sources d'incertitudes dans les observations liées au bruit des caméras. Dans ce cas, une fonction de probabilité modélise plus correctement le modèle. Des hypothèses sont générées et vérifiées à partir des observations de l'image, grâce à des techniques comme les HMM (Hidden Markov Model) [Lan X, Huttenlocher D.P.], les MCMC (Monte Carlo Markov Chain) [Lee M.W., Cohen I.], le filtre à particules (et sa variante « condensation » [Andrew Blake and Michael Isard.]) ou le filtre à grille [Taycher L., Demirdjian D., Darrell T., Shakhnarovich G.].
L'approche probabiliste dans tout problème de vision par ordinateur consiste à trouver le maximum à posteriori (MAP) de la densité de probabilité des N paramètres du modèles $x=(v_1, v_2, v_3)$ d'après les observations y sur l'image : $x=(v_1, v_2, v_3)=arg_{max}P(x/y)$, et d'après l'écriture bayésienne : $P(x/y) \propto P(y/x).P(x)$.
La probabilité P(y/x) est appelée la vraisemblance et notée L(x, y), plus facile à calculer si on considère un modèle x et les observations image qu'il génère y, P(x) est la probabilité *a priori* sur le modèle qui peut être calculée par apprentissage par exemple ([Gao J., Shi J.], [Lan X, Huttenlocher D.P.]). Le MAP a pour objectif de propager les hypothèses pertinentes au cours du temps et le suivi multi hypothèses permet de retrouver la

bonne solution même après une erreur de suivi.

[Lee M.W., Cohen I.] estiment la densité à posteriori par un échantillonnage de type « **Metropolis Hasting** ». La **densité de proposition** est remplacée par une fonction conditionnée par les observations. C'est une technique d'échantillonnage par MCMC conduite par les données de l'image (data driven Monte Carlo chain) et permettant de faire converger l'algorithme vers un optimum global plus efficacement. Il est alors nécessaire d'extraire des cartes de probabilités pour chacun des membres, les « proposal maps » sont des hypothèses pondérées par leur confiance, provenant des indices extraits de l'image, et modélisées sous la forme de gaussiennes 2D sur l'image.
La densité à posteriori n'étant pas possible à exprimer analytiquement, il faut en trouver une approximation, conduite par un échantillonnage d'importance séquentiel suivant une distribution de proposition dans le cadre du filtre à particules [Isard M., Blake A., 98]. Un ré échantillonnage est nécessaire pour empêcher la dégénérescence des échantillons vers une solution unique. Le nombre de particules, les échantillons, doit être assez important pour représenter toutes les hypothèses possibles mais l'espace de recherche dans le cas du suivi étant de grande dimension, le nombre de particules nécessaire est souvent trop grand. Afin d'éviter de rechercher un optimum dans un espace trop petit, le « street light effect » [Demirdjian D., Taycher L., Shakhnarovich G., Grauman K., Darrell T.], il faut améliorer le ré échantillonnage. La fonction de vraisemblance possédant des maxima allongés sous la forme de vallées conduit à ré échantillonner avec une covariance qui suit ces vallées par la technique de « covariance scaled sampling » [Sminchisescu C., Triggs B. 01]. Une seconde amélioration a été donnée par [Sminchisescu C., Triggs B. 03a], des échantillons sont générés vers les poses 2D qui présentent une ambiguïté avec la projection du modèle dans l'image. En exploitant la propriété multi hypothèses du filtre à particules, le procédé de saut cinématique [Sminchisescu C., Triggs B. 03a] permet de raccrocher le suivi après que la pose 2D ne soit plus ambiguë grâce aux contraintes temporelles (cf. figure 35).

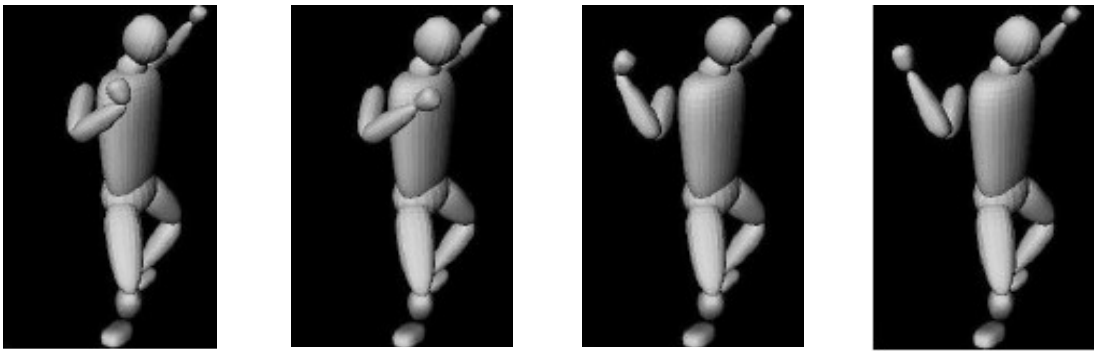


Figure 35 : Ambiguïtés 3D-2D : un membre avec deux articulations vu en 2D peut générer quatre positions 3D qui ont la même projection dans l'image. Cet exemple est celui du bras complet muni des articulations du coude et du poignet ([Sminchisescu C., Triggs B. 03a], [Noriega P. a]).

Le formalisme de Bayes associé à un réseau bayésien [Gao J., Shi J.] permet de calculer la probabilité à posteriori de la posture. Un réseau bayésien représente les dépendances des probabilités par des **liens** entre les **noeuds** correspondant aux états paramétrés des membres du corps humain. La probabilité jointe est dans ce cas le produit des probabilités indépendantes entre les membres non adjacents. Ces probabilités sont issues d'un apprentissage préalable et la vraisemblance d'un membre est fonction du nombre de pixels appartenant au membre détecté en mouvement. La **propagation de croyances discrètes** peut se faire avec un algorithme de **filtre à particules en interactions** [Bernier O., Cheung-Mon-Chang P.]. Les **potentiels d'interaction** entre les membres adjacents sont calculés pour chaque paire de particules appartenant à ces membres. La propagation de croyances peut se faire dans un espace **continu**, **les potentiels et les messages sont approximés par des mélanges de gaussiennes** [Sigal L., Bhatia S., Roth S., Black M.J., Isard M.], mais la mise à jour des gaussiennes créant une **explosion combinatoire du nombre de gaussiennes du fait des mélanges, un échantillonneur de Gibbs similaire à l'algorithme PAMPAS** [Isard M.] permet de garder le même nombre de

gaussiennes tout au long du processus.

1.4.3 Approches à base de règles

Ces approches utilisent plusieurs caractéristiques bas niveau qui, reliées par des règles de haut niveau, permettent de suivre le corps d'une personne par ses membres. Par exemple [Leignel C., Viallet J.E.], un système d'intelligence artificielle composé de trois niveaux d'intelligence hiérarchique, chacun lié à un tableau noir hiérarchisé. Au plus bas niveau de la hiérarchie, les spécialistes effectuent les traitements bas niveau, pour la détection des membres. Les éléments bas niveau issus de ces traitements sont regroupés dans le tableau noir bas niveau : épaule, bras, avant-bras, main, buste, tête. Au niveau intermédiaire, les tâches regroupent les éléments bas niveaux pour la constitution de membres supérieurs complets : une main + un avant-bras + un bras + une épaule, et les membres supérieurs complets sont stockés dans le tableau noir intermédiaire. Au plus haut niveau de la hiérarchie, la stratégie regroupe les membres supérieurs complets avec le buste et la tête, formant ainsi le haut du corps complet et le résultat est stocké dans le tableau noir haut niveau lié à la stratégie. Les tableaux noirs de niveaux supérieurs activent les tâches et les spécialistes de niveau inférieur en fonction des hypothèses qu'ils contiennent.

1.5 Suivi lors des occultations

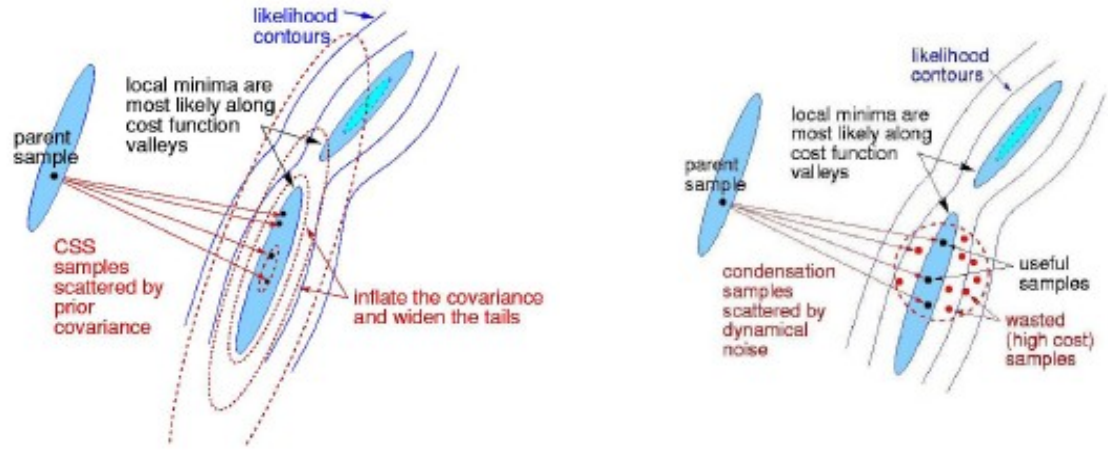
La nécessité de détecter et suivre des personnes en mouvement, y compris en cas d'occultations, est requise dans beaucoup d'applications de surveillance (W^4, Pfinder). Pfinder s'occupe de suivi de personnes dans des scènes complexes mais est restreint à une seule personne sans occultations. [Niu W., Jiao L., Han D., Wang Y.-F.] font du suivi multi-personnes en présence d'occultations avec un filtre de Kalman dans un environnement extérieur. Le système proposé par [Rerkrai K., Fillbrandt H.] suit des personnes sous des occultations partielles mais limité à une seule personne-via un filtre de Kalman. L'avant plan est segmenté par une méthode de soustraction du fond. Une silhouette moyenne est trouvée par régression linéaire, et de la connaissance *a priori* est appliquée à la régression linéaire pour définir la présence d'occultations. Le modèle de caméra est utilisé pour calculer la position et hauteur de la personne suivie. Une carte de profondeur des objets dans le fond permet de détecter les occultations de la scène. Une personne est en occultation si sa profondeur est supérieure que celle d'un objet. Tout ce qui est en occultation et appartenant à la personne est mis au même plan que la personne dans l'image de profondeur. Ainsi la forme de la personne (silhouette) est localisée pendant les occultations. Les ombres sont également éliminées, du faut qu'une ombre crée une petite réduction d'intensité sans changer la couleur de l'image.

Dans le cas des occultations, les méthodes déterministes, en mettant à jour les paramètres d'un modèle de façon unique, risquent de manquer le suivi et de décrocher pour toujours de l'objet suivi. C'est le cas du filtre de Kalman ([Kalman R.E.], [Kalman R.E.], [Welch G., Bishop G.]) qui sait gérer des distributions normales mais pas des densités de probabilités avec plusieurs modes. Il s'agit d'un estimateur récursif avec un modèle de mouvement linéaire. L'hypothèse est faite que le bruit de la dynamique du processus et le bruit de mesure suivent des lois normales, ce qui est limitant dans le cadre du suivi d'objets. Dans bon nombre de situations courantes dans les séquences vidéo (occultations), la distribution du bruit ne suit pas une loi normale. ([Isard M., Blake A., 96], [Isard M., Blake A., 98]) ont par la suite proposé d'approcher la densité de probabilité multi modale, c'est l'algorithme de « condensation » (« Conditional Density Propagation »), connue aussi sous le nom de « filtrage particulaire », utile lorsque les densités de probabilités sont non-Gaussiennes et multi-modales. Ce filtre approche les distributions de manière non paramétrique et peut suivre plusieurs hypothèses simultanément.

Trois grandes étapes définissent cet algorithme de « filtrage particulaire » (cf. Annexe 2) : **propagation, pondération, ré échantillonnage**. [Thome N.] a adapté cet algorithme de la façon suivante. Un vecteur d'état X correspond au centre de la boîte englobante de la silhouette de la personne et la densité de probabilité des paramètres est estimée à partir de la position initiale avant occultation.
-L'étape de **propagation** comporte deux termes, un correspond à la vitesse estimée de l'objet, l'autre au bruit de propagation aléatoire des particules;

-L'étape de mesure image consiste à attribuer un **poids** à chaque particule, évaluée par une mesure de corrélation entre l'apparence de la particule et le modèle de l'objet, par calcul d'une distance euclidienne. Le résultat est le poids et l'état moyen du filtre est calculé comme la moyenne des positions des différentes particules pondérées par leur poids;
-L'étape de **ré échantillonnage** des particules est effectué en associant une probabilité de tirer une particule par rapport à son poids.
L'état moyen du filtre est la somme pondérée des particules par leur poids, c'est l'estimation de la position de la personne. Le filtre à particules est capable d'estimer des distributions multi modales et est robuste aux occultations. La remise à jour du modèle d'apparence permet d'être invariant à la taille des personnes dans l'image et de supporter les cas d'occultations. Le modèle W^4 [Haritaoglu I., Harwood D., Davis L.S. 00] ne remet pas à jour le modèle d'apparence.
Dans les cas d'occultations, le suivi de l'objet se fera correctement dés lors que l'occultation a disparu et que l'objet est de nouveau visible, grâce à la propagation de l'information de façon diffuse pendant qu'il y a occultation et que la mise en correspondance est difficile.

Bien que l'algorithme de condensation sache gérer les problèmes d'occultations, la diffusion des particules avec un bruit dynamique isotrope est en échec dans le cas d'un problème de grande dimension. C'est la raison pour laquelle [Sminchisescu C., Triggs B. 03b] proposent d'adapter l'algorithme de condensation en grande dimension pour trouver la solution optimale par un échantillonnage optimal efficace « Covariance Scaled Sampling » (cf. figure 36). Le principe est de propager les particules de façon à coller au mieux à la fonction de coût. Le ré échantillonnage des particules intervient en utilisant un modèle de densité de probabilité à queue longue (permettant une meilleure diffusion), dont la covariance résulte du calcul précédent, d'où la dénomination « Covariance Scaled Sampling ».



(a) Le ré échantillonnage se fait d'après une ellipsoïde qui suit les vallées à fort vraisemblance. Avec l'algorithme « Condensation classique », ce ré échantillonnage a lieu selon un mouvement brownien.

(b) Avec l'algorithme « Covariance scaled sampling », de nouveaux maxima sont découverts plus probablement.

Figure 36 : « Covariance scaled sampling ». A gauche ([Sminchisescu C., Triggs B. 03b], [Noriega P. a]).

La figure 37 présente quelques résultats de suivi monoculaire en 3D.

Figure 37 : Suivi monoculaire en 3D. Les améliorations sur le ré échantillonnage du filtre à particules offrent un environnement peu contraint mais pas encore assez robuste ([Sminchisescu C., Triggs B. 03a], [Sminchisescu C., Triggs B. 03b], [Noriega P. a]).

1.6 La reconnaissance

La reconnaissance des objets concerne les modèles d'apparence d'objets et les modèles d'objets réels.

1.6.1 Les modèles d'apparence d'objets

Les **modèles d'apparence d'objets** sont utilisés lorsqu'il est difficile de classifier l'objet à cause du bruit dans l'image ou de la complexité des objets. [Baumberg A., Hogg D.] définissent un modèle d'apparence d'objet non rigide complexe pour l'analyse du mouvement humain. Un modèle 2D de la forme est extrait grâce aux points de contours de la projection de la personne dans le plan image, en se basant sur un modèle de distribution appelés PDM « Point Distribution Model » issu des travaux de ([Cootes T.F., Taylor C.J.], [Cootes T.S, Taylor C.J., Cooper D.H., Graham J.]). Dans ce modèle, une différence dx est mesurée entre un modèle quelconque et le modèle moyen $dx=x-x_{moyen}$, x est un vecteur de points, x_{moyen} est le vecteur moyen. Les vecteurs différence donnent les modes de variation principaux (les « eigenshape ») du modèle. Les résultats sont obtenus sur un ensemble d'apprentissage réduit à des personnes marchant latéralement. L'avantage de ces méthodes basées apparence est de venir juste après la détection composée d'indices pour l'apparence. En revanche, il n'y a pas d'aspect volumique ni de caractéristiques intrinsèques de l'objet.

1.6.2 Les modèles d'objets réels

Les modèles d'objets réels utilisent des caractéristiques de l'objet. Les objets rigides, tels que les véhicules, peuvent être décrits sur des considérations volumiques. [Koller D., Daniilidis K., Nagel H.-H] décrivent un véhicule par un polyèdre 3D, et les contours obtenus à l'étape de détection sont mis en correspondance avec le modèle par minimisation d'une distance de Mahalanobis, pour chaque arête du polyèdre. Les modèles non rigides sont ceux pour lesquels il n'y a pas pas d'invariance de la forme au cours du temps, aux

transformations affine prés (rotation, translation, homothétie). C'est le cas des humains. Pour une personne donnée, les caractéristiques qui la décrivent varient au cours du temps à cause des variations liées aux déplacements (oscillations des bras et des jambes). De plus, le modèle est différent pour deux personnes distinctes. Dans le cas des objets non rigides tels que le corps humain, [Akita K.] utilise un modèle composé d'un squelette de six segments (deux bras, deux jambes, torse et tête). [Chen Z., Lee H.] définit un modèle de squelette de dix-sept segments, et [Rohr K.] un modèle volumétrique avec quatorze cylindres elliptiques.

1.7 L'interprétation sémantique de la scène

L'interprétation sémantique d'une scène se décompose en méthodes discriminatives et génératives
Les approches **discriminatives** partent des caractéristiques extraites des images et proposent une classification directe du mouvement, soit par apprentissage par classification supervisée, soit par classification non supervisée. Ces méthodes modélisent le mouvement de façon implicite, celui-ci est dans le processus d'apprentissage. Dans les méthodes **génératives**, il est possible de fabriquer un ensemble d'instances de mouvements à partir de la même classe du fait qu'il y a une description explicite du mouvement.
Parmi les **méthodes discriminatives, nous différencions les méthodes d'apprentissage de type « template matching »**, proposant une stratégie d'apprentissage pour discriminer les types de mouvements **et les approches de classification non supervisée**, pour la **détection d'évènements rares**.

2 Les différentes approches d'extraction des caractéristiques

Les caractéristiques de l'image à extraire en vue de la détection bas niveau des indices, pour le suivi du corps, sont la couleur, les contours, la texture, le mouvement, la profondeur. Par la suite, les positions des membres sont estimées, soit par une approche ascendante soit par une approche descendante.

2.1 Extraction de la caractéristique couleur

Pour la **caractéristique couleur**, PFINDER est un système temps réel pour le suivi de personnes et l'interprétation de ses comportements, à l'aide d'un modèle statistique multi classes 2D de la couleur (les zones de teinte chair [Leignel C., Viallet J.E.]) et de la forme représentant la tête et les mains. La soustraction du fond permet de découper la silhouette d'une personne dans l'image [Leignel C., Viallet J.E.]. Mais l'information de couleur seule ne suffit pas, elle n'est pas robuste aux variations de luminosité. La chrominance normalisée permet de s'abstenir de la luminosité. Les ombres portées sont des régions qui ne correspondent pas au fond, mais qu'il est nécessaire de supprimer. Les pixels correspondant à ces ombres ont une luminosité plus faible que l'image de référence mais la même chrominance. Dans un espace colorimétrique invariant en luminance et donc aussi à l'influence des ombres (Lab, Luv), on peut séparer les ombres des objets, ce que fait PFINDER. L'objectif de PFINDER est de modéliser des personnes en mouvement devant une caméra, afin de les insérer dans un environnement virtuel. Une personne est modélisée par un ensemble de « blobs » détectés en temps réel et qui suivent le corps humain, mettant à jour le modèle du corps. Les vecteurs de primitives associés à chacun des pixels sont composés des coordonnées spatiales et spectrales des composantes de l'image. Les pixels sont ensuite regroupés en régions 2D connexes de propriétés colorimétriques et spatiales similaires, les « blobs ». La tête et les mains sont ainsi modélisées par des « blobs », représentés par leur moyenne et matrice de covariance, et par un modèle gaussien pour leurs statistiques spatiales. Le fond statique est modélisé par une autre gaussienne. Les deux modèles, pour le fond et pour l'avant-plan, sont remis à jour régulièrement. PFINDER a été adopté pour les interfaces non contraintes et le codage bas débit.

2.2 Extraction de la caractéristique contour

D'un autre côté, **les contours**, bien que plus robustes que l'information colorée aux variations d'éclairage, le sont moins au bruit dans l'image. [Cristian Sminchisescu and Bill Triggs] estiment la vraisemblance de leur modèle par rapport aux observations issues de la détection de contours par l'algorithme de Sobel et le calcul de flot optique. [Mori G., Ren X., Efros A.A., Malik J.] segmentent l'image par un détecteur de Canny, différentes zones dans l'image correspondent aux membres de la personne. Dans [Leignel C., Viallet J.E.], un algorithme de Shen Castan détecte les contours dans l'image et une transformée de Hough rassemble les

contours en droites, correspondant aux directions générales des membres candidats.

2.3 Extraction de la caractéristique mouvement

Dans le cas du suivi, **l'estimation du mouvement** dans une image permet de prédire la position des membres à l'image suivante, en supposant la variation du mouvement négligeable sur une courte période de temps. Une des techniques utilisée est par calcul du **flot optique** entre les images afin d'estimer le mouvement dans la scène.

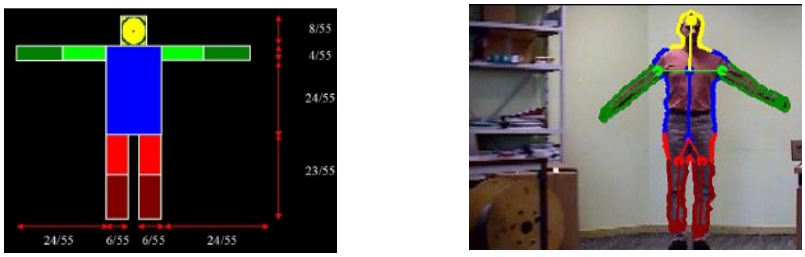
2.4 Extraction de la caractéristique profondeur

En plus des informations de **couleur, contour et mouvement, l'information de profondeur** permet de lever l'ambiguïté dans le cas des occultations des membres. L'information 3D est obtenue par l'écart de position (la disparité) entre deux pixels qui « regardent » le même point de la scène dans deux images issues de deux caméras calibrées. L'estimation de la vraisemblance du modèle est mesurée en 3D et non en 2D. Cependant, cette méthode nécessite une calibration au préalable des caméras afin de déterminer leurs paramètres intrinsèques et les paramètres extrinsèques des positions mutuelles entre les caméras. La calibration n'est pas nécessaire avec une seule caméra, ce qui constitue un avantage dans une application de vidéo surveillance.

3 Quelques exemples

3.1 Présentation des travaux de [Thome N.]

[Thome N.] détecte les personnes dans des séquences d'images **monoculaires, en temps réel et avec des solutions non invasives** (environnement non contraint). Une **segmentation au sens du mouvement** par différence de l'image courante avec un modèle du fond est effectuée, suivie d'une mise en correspondance dynamique de régions, correspondant à un **problème d'associations de données**. L'approche originale de suivi de personnes est un **modèle d'apparence articulé** (cf. figure 38), invariant aux transformations affines, permettant un suivi correct même dans les cas d'occultations.

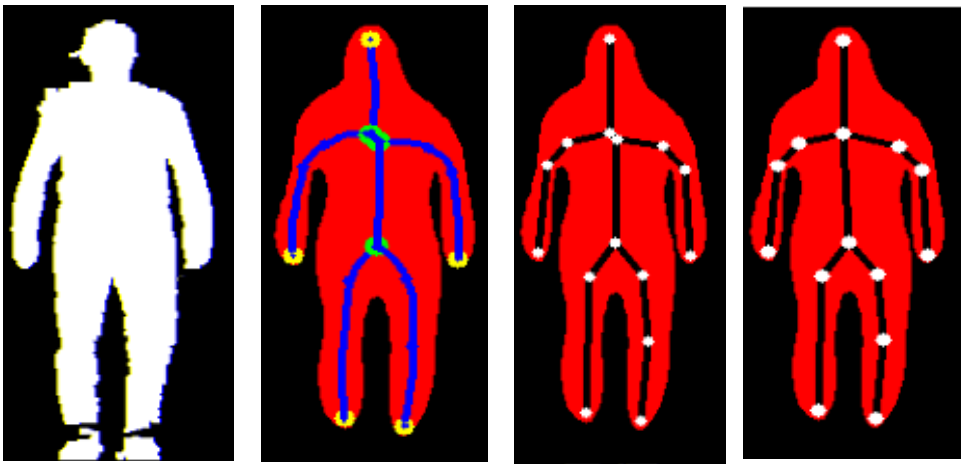


(a) Géométrie du modèle d'apparence (b) Étiquetage final

Figure 38 : Géométrie du modèle d'apparence et étiquetage final [Thome N.].

L'utilisation d'un modèle articulé capturant les informations structurelles pour chacun des membres est nécessaire, surtout lorsque les caractéristiques globales de couleur, représentées par des densités de probabilités, sont insuffisantes. La **silhouette** met en évidence les membres de la personne par analyse de la forme.

La forme du **squelette** est représentée sous la forme d'un **graphe** et une technique de **mise en correspondance de graphes** identifie chaque membre (cf. figure 39).



(a) Silhouette initiale | (b) Calcul du squelette sur la silhouette lissée | (c) ensemble de segments après polygonisation | (d) Suppression des petites arêtes

Figure 39 : Extraction des segments de la silhouette [Thome N.].

3.1.1 Modèle d'apparence articulé

Les composantes connexes ou « blobs » sont extraites à l'étape de segmentation de mouvement et mis en correspondance au cours du temps. Le **suivi basé région** relie les régions entre les instants successifs en **associant les données**. Un **modèle d'apparence articulé** est composé des caractéristiques de forme, couleur et texture, utiles pour traiter les situations difficiles (occultations). Le modèle d'apparence est généré et mis à jour à chaque instant, et les membres étiquetés. Il s'agit de mettre en **correspondance** un **graphe** image issu du squelette (topologie des segments extraits du squelette) de la silhouette extraite de la personne suivie avec un graphe modèle. Les pixels mobiles de la scène sont extraits, par une soustraction entre l'image courante et un modèle du fond statique, conduisant à une carte binaire de mouvement. L'image du fond dite « image de référence » modélise les parties statiques de la scène analysée. Il existe différentes approches pour modéliser et mettre à jour cette image de référence. [Thome N.] choisit un mélange de gaussienne. La méthode avait été au préalable introduite par [Stauffer C., Grimson W.E.L.b]. Une mise à jour dynamique du modèle du fond permet de gérer les variations d'illuminations de la scène, les ajouts ou suppression des objets du fond. L'approche proposée par [Thome N.] repose sur un nombre variable de gaussiennes, déterminé automatiquement et évoluant au cours du temps. La méthode est similaire à celle de [Shimada A., Arita D., Taniguchi R.I.]. Dans les scènes extérieures, le fond peut changer à tout moment, de part le mouvement ou les ombres par exemple dans certaines parties de l'image et dans ces cas, le fond nécessite une modélisation multi gaussienne. Enfin, une analyse en composantes connexes regroupe les pixels qui ont un mouvement suffisant, en objets avec des propriétés sémantiques.

3.1.2 Mise en correspondance de blobs

Les objets extraits sous la forme de « blobs » à l'instant t sont mis en correspondance avec ceux détectés à l'instant t+1. Une prédiction du vecteur d'état X, associé à l'objet suivi sous forme paramétrique, est mise à jour dynamiquement en fonction des observations dans l'image à chaque image de la séquence. Autrement dit la position de chaque objet est prédite en fonction de sa position à l'instant précédent t-1, et d'un modèle de mouvement, simple à vitesse constante ou complexe et paramétrique, permettant de restreindre l'espace de recherche. Une matrice de similarité détermine les liens entre les objets entre deux instants successifs.

L'objectif du suivi est de minimiser une fonction de coût entre le vecteur d'état estimé et le vecteur d'état extrait de l'image observée. Ce problème d'association de données peut être résolu par un algorithme de type « Plus Proche Voisin » mais il risque d'échouer dans les cas d'objets multiples proches. Pour contourner cette difficulté, deux autres grandes techniques de suivi multi-hypothèses existent dont le principe est d'associer des objets non plus seulement sur deux images mais sur plusieurs images consécutives : les associations probabilistes de données (Joint Probabilistic Data Association Filter) [Bar-Shalom Y., Fortmann T.E.] et le Suivi Multi Hypothèses (Multiple Hypothesis tracking) ([D.B. Reid.], [Cox I.J, Hingorani S.L.], [Cox I.J.], [Cox I.J, Hingorani S.L.]). La **cohérence temporelle** permet de savoir si l'objet suivi est réellement un objet d'intérêt ou du bruit. Pour cela il devra être suivi pendant un certain nombre d'images. Comme dans les approches « **multi hypothèses »**, la décision sur les mises en correspondances dépend des associations des données sur un intervalle de temps.

3.1.3 Étiquetage des membres

Après l'étape de mise en correspondance des « blobs », il s'agit d'étiqueter les membres afin de former un **modèle d'apparence articulé**. L'étiquetage est possible par une mise en correspondance de graphes à partir d'un modèle 3D du squelette humain, indépendante de la pose de la personne, du point de vue, de la géométrie ou de l'apparence des membres. Pour représenter l'apparence, il existe deux approches, soit par des descriptions statistiques soit par l'utilisation de « templates ». Parmi les approches statistiques, citons [Wren C.R., et al.] qui détectent et suivent les membres d'une personne dans une vidéo, en environnement intérieur, par une modélisation gaussienne multi dimensionnelle pour la position et la couleur de chaque « blob ». La modélisation par « template » ou gabarit revient à mémoriser « l'imagette » de l'objet d'intérêt. Mais la mise à jour du modèle d'apparence au cours du temps demande un stock considérable « d'imagettes ». Dans W^4 [Haritaoglu I., Harwood D., Davis L.S. 00], le modèle d'apparence, une boite englobante, est mis à jour pour chaque personne suivie et est utilisé en cas d'occultations. [Thome N.] met à jour un modèle d'apparence articulé pour chaque personne suivie, en étiquetant les membres à partir de la silhouette segmentée, des segments candidats ayant été identifiés pour les différents membres. Dans ce modèle, les occultations partielles n'empêchent pas de mettre à jour le modèle car il n'y a que les modèles d'apparence des membres détectés qui sont modifiés.
Pour suivre les personnes dans les cas d'occultations, un algorithme de type « C**ondensation »** estime la position de la personne suivie à partir des caractéristiques d'apparence, et donne une approximation robuste de la densité même dans les cas multi modaux.

3.2 Approche avec une caméra à champ large

Dans les **applications de surveillance**, il est parfois utile d'avoir la position de l'objet suivi en trois dimensions. Les caméras vidéo stéréo à champ large fournissent l'information de position à de grandes distances, ce qui n'est pas possible avec les caméras vidéo stéréo standard [Hampapur A., Brown L., Connell J., Ekin A., Haas N., Lu M., Merkl H., Pankanti S.]. Pour établir la correspondance entre les deux images, les apparences des objets sont mis en correspondance, par leurs histogrammes colorés, par la distance de Bhattacharya entre toutes les paires possibles.
La première étape est de **détecter** des objets d'intérêts et de les suivre dans chaque champ de vue de la caméra via un **modèle d'apparence**. **Les objets en 2D suivis sont combinés via la stéréo champ large** pour former des objets 3D suivis. La tête est détectée en 2D puis les centroïdes de la tête dans les deux vues sont combinés pour détecter et suivre la position de la tête en 3D par triangulation. Chaque caméra se voit affecter une orientation et un zoom en fonction de l'objet suivi, le **système sélectionne de façon automatique la caméra qui va se suivre la tête**. Le système recherche donc le visage et une fois détecté, la caméra se centre sur celui-ci et le **zoom** est augmenté. **L'orientation** est également contrôlée en fonction du déplacement relatif du centre du visage par rapport au centre de l'image.
La **classification** d'objets est appliquée à tous les objets suivis, générant trois types de label, les véhicules, **les groupes de personnes et les personnes seules**, en fonction des primitives de forme comme la compacité, les paramètres de l'ellipse englobante et les primitives de mouvement (vitesse et direction). A partir d'un petit ensemble d'entraînement, les objets sont classés par un classificateur des plus proches voisins et d'une

information de cohérence temporelle. Le suivi de plusieurs objets recherche des **trajectoires** en **combinant à la fois l'apparence des objets et les caractéristiques du mouvement**. **Les modèles d'apparence sont des « templates » basés image** et un nouveau modèle d'apparence est crée lorsqu'un nouvel objet entre dans la scène.
Une alerte temps réel est déclenchée selon des critères prédéfinis, comme la détection de mouvement dans certaines zones, la détection d'objets abandonnés, etc. Un index des vidéos est stocké, il contient la trajectoire des objets dans la scène, la taille des objets, le type des objets, l'apparence des objets, et le fond dynamique.

3.3 Approche avec suivi de visage

[Ruiz-del-Solar J., Shats A., Verschae R.] est un système de **suivi robuste de personnes** dans un environnement réel et temps réel. Trois composantes sont mises en oeuvre : l'analyse de mouvement, l'analyse de la couleur, et l'analyse de visage.
Les systèmes de contrôle d'accès sécurisés sont devenus de plus en plus importants et sont permis par le suivi de visages. [Ruiz-del-Solar J., Shats A., Verschae R.] proposent un système de suivi temps réel basé sur la **détection du visage via la couleur de la peau et des règles heuristiques**. Les **régions de teinte chair** sont obtenues via une **table de teinte chair**. Les autres objets de la scène avec les mêmes caractéristiques colorimétriques qu'un visage dans le fond de la scène sont immobiles. Un visage étant en général en mouvement dans une scène, l'information de mouvement permet d'éliminer les faux candidats.

3.4 Approche par modèle de Markov caché pour la détection des évènements rares

Une personne peut être différenciée d'une autre personne par la reconnaissance de sa démarche. [Kale A. et al.] élaborent une distance de la silhouette de la personne (obtenue par détection des régions en mouvement) obtenue à chaque image, avec un ensemble de poses de silhouettes représentatives du mouvement. Cette mesure du vecteur d'observation d'un Modèle de Markov Caché, permet de modéliser le mouvement de la marche pour une personne et de le distinguer d'une autre personne.

3.4.1 Définition du modèle de Markov Caché

Un HMM (« Hidden Markov Models » ou « modèle de Markov caché ») est un modèle statistique dans lequel les états sont reliés à un vecteur d'observation, exprimé par une probabilité d'observation. La matrice de transition modélise la probabilité de passage d'un état à un autre, c'est-à-dire la vraisemblance avec laquelle les états sont susceptibles de se suivre. [Lan X, Huttenlocher D.P.] suivent un sujet animé de la marche grâce à un modèle de Markov caché, comprenant les postures clés de la marche observées depuis huit angles de vue, et décalées de 45° (cf. figure 40).

Les observations associent chaque image avec un état du modèle de Markov d'après la distance de Chanfrein entre le modèle « cardboard » et la silhouette. [Yamato J., Ohya J., Ishii K.] analysent le mouvement détecté dans les images, grâce aux caractéristiques des blobs de mouvement, de couleur et de texture. Ces observations images sont associées aux états du HMM. [Nair V., Clark J.] détectent des activités inhabituelles grâce à un HMM. Le vecteur d'observation pour le HMM est constitué des paramètres extraits sur les régions en mouvement. Des mouvements *a priori* sont définis : « entrer », « sortir d'une pièce » et « roder ». La vraisemblance de ces mouvements *a priori* avec les images de la séquence conduit à la reconnaissance des mouvements inhabituels d'une personne dans un couloir. Si la séquence ne peut pas être expliquée par aucun de ces trois modèles, alors un événement inhabituel est détecté.

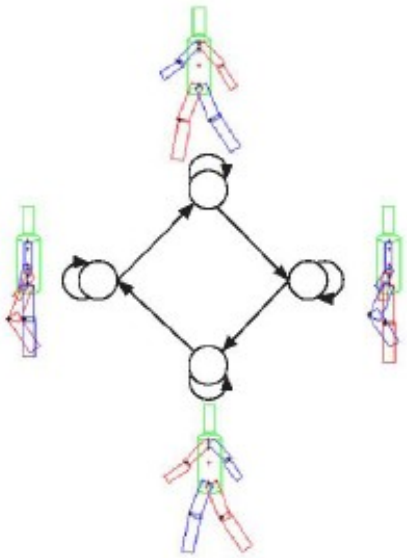


Figure 40 : Modèle de Markov comportant les positions clés de la marche pour la vue de côté ([Lan X, Huttenlocher D.P.], [Noriega P. a]).

3.4.2 Définition du réseau bayésien

Les Réseaux Bayésiens Dynamiques (« Dynamic Bayesian Networks », « DBN »), appelés aussi « modèles graphiques dynamiques », sont une généralisation des Modèles de Markov Cachés [Jensen F.a]. Ils consistent en une représentation graphique orientée où chaque état est influencé par un nombre quelconque de variables (une seule dans les HMM), et avec une extension temporelle indéfinie (une seule composée du passé immédiat dans le cas des HMM). Les Réseaux Bayésiens Dynamiques sont appliqués à l'analyse du comportement humain dans la vidéo ([Brand M., Kettnaker V.], [Buxton H., Gong S.]).

La plupart des méthodes de suivi détectent les régions en mouvement, soit par une soustraction du fond soit par une différence d'images, et suivent les régions en mouvement via leur trajectoire, soit par un filtre de Kalman, soit par un arbre multi hypothèses [Cox I.J, Hingorani S.L.], soit avec une méthode d'inférence à base de degrés de confiance comme l'algorithme JPDAF [Bar-Shalom Y., Fortmann T.E.] ou le filtre à particule [Isard M., Blake A., 98]. Une approche différente est utilisée avec les réseaux bayésiens (BN) [Abrantes A., Marques J., Lemos J.]. Le réseau bayésien est défini par le graphe (ensemble de dépendances causales) et le modèle probabiliste associé à chacun des noeuds. Ils sont proposés pour modéliser les interactions entre les trajectoires des objets dans des applications de suivi. Ils permettent de lever les ambiguïtés sur les conflits entre les superpositions des diverses régions actives (les groupes d'objets) ou au sujet des occultations. **Dans les cas d'occultations, les trajectoires sont rompues avec la plupart des méthodes, tandis qu'avec un réseau bayésien, les différents segments appartenant au même objet sont reliées en leur assignant un label commun. Le réseau bayésien est construit automatiquement pendant la phase de suivi et il tente de modéliser les interactions causales entre les trajectoires des objets en mouvements.** [Jorge P.M., Marques J.S., Abrantes A.J] estiment l'architecture du réseau bayésien à l'aide de méthodes d'apprentissage supervisé par un réseau de neurones, un perceptron multi couches. Le suivi d'objets a lieu en deux étapes. La première étape détecte les régions actives et associe des régions par paires dans des images consécutives. Un ensemble de segments de chaque trajectoire est extrait, chacun correspondant à l'évolution d'un objet ou d'un groupe d'objets dans la vidéo. La trajectoire entière de chaque objet est extraite en reliant les différents segments trajectoires, ce qui revient à une opération d'étiquetage des segments de trajectoire, en assignant une probabilité à chaque segment. **Les interactions entre les segments sont modélisés dans le réseau bayésien.** Les noeuds sont les étiquettes et les liens les dépendances causales

modélisées par des probabilités conditionnelles entre les noeuds. Le perceptron multi couches va classifier chaque lien comme pertinent ou non. Il est entraîné pour remplacer les règles heuristiques pas toujours adaptées pour traiter les cas d'occultations de plusieurs personnes. Cela permet une réduction significative de la complexité du réseau tout en traitant les cas d'occultation non traités précédemment.

Un **réseau bayésien** est proposé par [Jorge P.M., Marques J.S., Abrantes A.J] pour le suivi d'objets afin de **modéliser les interactions entre les trajectoires détectées** et d'obtenir une **identification des objets en présence d'occultations**. Les réseaux bayésiens sont composés de règles simples et ne sont pas robustes dans certains cas. [Jorge P.M., Marques J.S., Abrantes A.J] proposent une nouvelle méthode pour estimer l'architecture des réseaux bayésiens à partir des séquences vidéo grâce à des techniques d'apprentissage supervisés.

[Oliver N., Horvitz E., Garg A.] ont modélisé les interactions entre les personnes avec des HMM couplés pour la détection et la classification des interactions. [Han M., Xu W., Gong Y.] ont travaillé également avec des HMM pour des applications de reconnaissance des interactions des véhicules d'aéroport. Ces approches utilisent des modèles probabilistes couplés représentant les relations entre les trajectoires individuelles segmentées. Une alternative serait d'utiliser un modèle probabiliste joint couvrant l'ensemble de la scène. [Galata A., Johnson N., Hogg D.] déterminent les trajectoires en encodant les relations spatiales et temporelles entre les objets en mouvement et en interaction (automobiles) via des modèles appris à partir des observations.

3.4.3 Cas des comportements inhabituels/anormaux

Pour [Junejo I.N., Shah O., Shah M.], l'objectif est d'apprendre les routes ou les chemins les plus communément pris par les objets et de détecter les **comportements inhabituels**, une personne marchant dans une zone non piéton etc. Un chemin est défini par une ligne de parcours, et une trajectoire comme un chemin avec un objet en mouvement. Les applications sont en vidéo surveillance, comme dans les aéroports où il s'agit de détecter la présence d'intrus dans des zones surveillées. [Grimson W.EL., Stauffer C., Romano R., Lee L.] ont utilisé un **système distribué de caméras** pour couvrir la scène entière. Les pistes sont clustérisées à l'aide de primitives spatiales basées sur une quantification vectorielle. Les comportements inhabituels sont alors détectés par mise en correspondance des trajectoires avec les clusters. [Junejo I.N., Shah O., Shah M.] proposent une nouvelle approche de détection de chemin avec des primitives multiples. Le système est entraîné sur des séquences prises depuis **une seule caméra**, mais le système peut être étendu à plusieurs caméras. La trajectoire de l'objet est défini par une succession de points, de longueur variable. Les trajectoires similaires obtenues lors de l'entraînement sont clustérisées. **Un noeud du graphe représente une trajectoire. Chacun des noeud est connecté aux autres noeuds, rendant le graphe complet.** Le poids d'un lien entre deux noeuds est la distance de Hausdorff mesurée entre deux trajectoires. L'avantage de cette mesure est qu'elle compare deux ensembles de cardinalité différente, donc deux trajectoires de longueurs différentes.

Le « NEC Laboratories America » a été développé pour le système de vidéo surveillance SmartCatch, pour plusieurs aéroports aux états unis [Gong Y.]. Ce système est capable de détecter **des comportements anormaux portant atteinte à la sécurité des aéroports, les apparitions et disparitions d'objets, ainsi que les interactions avec les autres objets de la scène**.
Les méthodes traditionnelles de suivi d'objet traitent la détection comme un processus à part initialisant le suivi. Une fois l'objet détecté, son suivi est assuré uniquement par le module de suivi. L'inconvénient de cette approche est que des erreurs de suivi peuvent arriver lors de changements d'apparences et d'illuminations de la scène. De plus des erreurs apparaissent aussi à cause des occultations entre objets. Divers méthodes ont tenté de suivre des objets de façon plus robuste, comme le **filtre à particule**. Mais ces méthodes ne conservent qu'**une seule hypothèse par objet suivi**, celle ayant la probabilité à posteriori la plus grande, basée sur l'observation courante et précédente. Des **méthodes multi-hypothèses sont plus robustes aux occultations**, aux fonds texturés et à la confusion entre plusieurs objets, car le résultat du suivi correspond à l'état de la séquence qui maximise la probabilité jointe de l'observation. Il existe dans cette gamme d'algorithme, le **MHT « Multiple Hypothesis Tracking », le JPDAF « Joint Probabilistic Data Association Filter »**, mais l'objet suivi doit être simple.

Le suivi multi objets doit résoudre deux problèmes, le problème **d'estimation** comme un problème de suivi traditionnel, et le problème **d'association** de données spécialement dans le cas d'interactions multi objets.

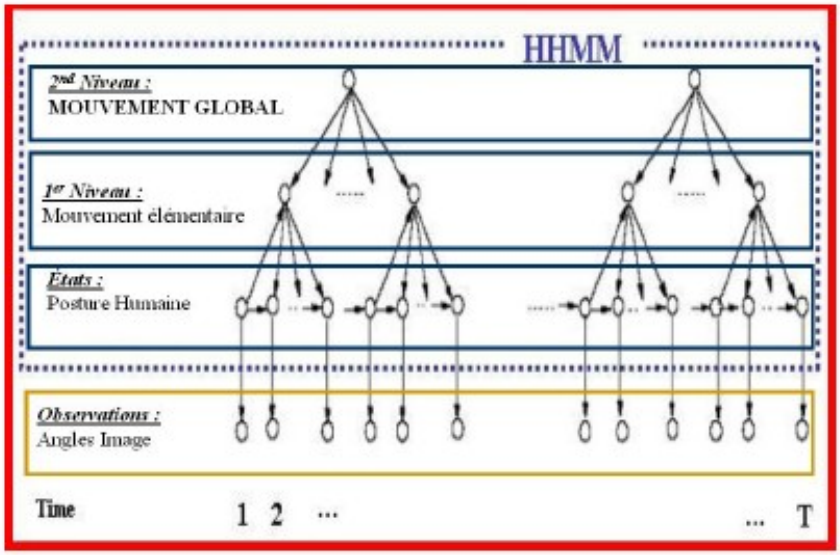
Bon nombre d'algorithmes de suivi résolvent le problème d'estimation par un maximum à posteriori MAP [Bar-Shalom Y., Li X.], l'hypothèse courante étant celle ayant la probabilité à posteriori maximale basée sur les observations courantes et précédentes. **La formulation MAP peut être simplifiée si on suppose un problème markovien HMM** [Rabiner L.R.]. **Cette approche échoue à cause des fonds texturés, des occultations et des ambiguïtés multi objets**. Un **autre type d'algorithme de suivi** estime la distribution de séquence d'observations d'états jointe. Le résultat du suivi correspond à une séquence d'états qui maximise la probabilité jointe entre les états de la séquence et les observations de la séquence. Les états de la séquence indiquent les trajectoires des divers objets suivis, c'est du **suivi de trajectoire.**
Un travail bien connu dans le suivi de trajectoire est le **suivi multi hypothèse** (MHT) développé par Reid [D.B. Reid.], décomposé en une estimation des états et des composantes d'association de données. Le filtre **JPDAF** d'association de données de probabilités jointes [Fortmann T.E., Bar-Shalom Y., Scheffe M.] détecte les états estimés en évaluant les probabilités d'association des suivis mesurés.

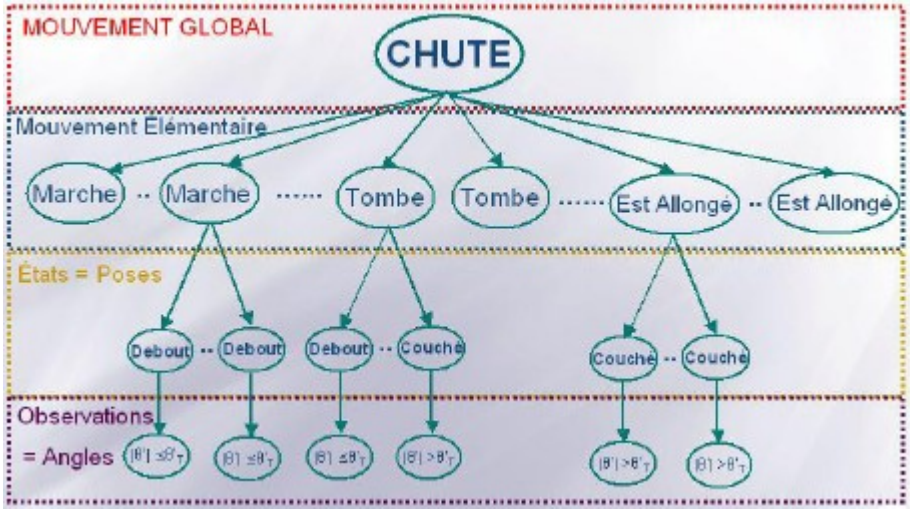
3.4.4 Cas de la détection de chute

Les évènements « rares » sont des évènements anormaux, difficiles à décrire par les **méthodes basées modèle**, mais faciles à reconnaître. [Zhong et al.] proposent de reconnaître les évènements anormaux dans une cantine d'hôpital en comparant chaque évènement à l'ensemble de ceux présents dans la séquence vidéo afin de constituer une base d'évènements similaires. Tous les évènements sont comparés à cette base afin de déterminer s'il s'agit d'un événement inhabituel. La détection d'un évènement rare est donc très dépendante du contexte.

[Nait-Charif H., McKenna S.] extraient la trajectoire des personnes suivies de façon automatique grâce à une caméra omni-directionnelle. Un filtre à particules suit les paramètres d'une ellipse modélisant la personne. Le mouvement est expliqué dans un langage naturel compréhensible par les humains. L'évènement rare (la chute de la personne âgée) correspond à un événement fort différent des évènements appris.

La contribution de [Thome N.] pour la détection de chute consiste à interpréter la séquences de postures avec un Modèle de Markov Caché Hiérarchique (« Hierarchical Hidden Markov Model »). La pose de la personne, debout ou couché, est déterminée dans une image. La carte binaire de mouvement est extraite. Le suivi de la personne est effectué grâce aux informations de mouvement, forme et apparence. Le rectangle minimal de la région extraite suivie permet d'extraire ses axes principaux, et l'angle entre le grand axe et la direction verticale comme caractéristique de la verticalité de la personne dans la séquence. Le détecteur de verticalité différencie une personne debout d'une personne couchée. La séquence est ensuite analysée, à partir de la suite des postures, avec un HHMM pour reconnaître un mouvement anormal comme la chute. Un modèle HHMM à deux niveaux analyse chaque mouvement courant (cf. figure 41) :
-Le premier niveau de mouvement correspond aux mouvements dits « élémentaires ». Pour la détection de chute, ce niveau comprend trois instances : « marche », « tombe », « couché ».
-Le second niveau correspond à des mouvements globaux, des séquences de mouvement primitif sur des durées plus longues.





(b) Exemple de la modélisation d'un mouvement de chute

Figure 41 : Architecture du Modèle de Markov Caché Hiérarchique [Thome N.]

Le HHMM permet de travailler à des échelles de temps différentes, allant des mouvements brusques de la chute avec des mouvements sur un intervalle de temps plus long, grâce aux contraintes haut niveau données par les modèles de mouvement. L'architecture hiérarchique est bien adaptée à l'interprétation sémantique de la scène pour la détection de mouvements brusques et permet de filtrer les fausses alarmes issues d'erreur bas niveau.

3.5 Représentation symbolique

Opposé à ces méthodes probabilistes, il existe des représentations symboliques. [Ivanov Y., Bobick A.F.] décrivent la séquence temporelle des évènements par une représentation grammaticale et développent une technique probabiliste. La principale application concerne les interactions voiture/personne. [Intille S.S., Bobick A.F., 95] exploitent une représentation probabiliste et symbolique pour la **reconnaissance des 22 joueurs de football américain**. L'avantage de la représentation symbolique est la capacité d'encoder plus facilement une connaissance *a priori* du domaine, surtout dans les situations où la quantité d'observations est limitée (comportements inhabituels). L'inconvénient par rapport aux méthodes purement probabilistes est le risque d'échec pour représenter les interactions.

Chapitre 3 – Systèmes de vidéo surveillance

1 Les différents systèmes de vidéo surveillance existants

Nous présentons dans cette section les systèmes de vidéo surveillance en développement ou finalisé, chez les chercheurs mais aussi les industriels. Parmi les projets de recherche, nous présentons de façon détaillée les projets VSAM, ADVISOR, BEHAVE, CASSIOPEE, VIGITEC, CAVIARE, et PASSWORDS. Trois d'entre eux sont développés à l'INRIA : ADVISOR, CASSIOPEE, CAVIARE.

1.1 Le projet VSAM

Le projet **VSAM** (Video Surveillance and Activity Monitoring) [Collins R., et al.a] a été développé par le Robotics Institute de l'Université de Mellon Carnegie of Southern California (CMU) et l'Institut Sarnoff fondé par DARPA. Ce projet a eu lieu entre octobre 97 et janvier 2000. V.S.A.M avait pour objectif de développer des algorithmes de détection et suivi automatique de plusieurs personnes et véhicules dans un environnement urbain et complexe avec un réseau de caméras distribuées, pour la surveillance automatique de séquences vidéo prises à partir de drônes, avions automatiques volant à haute altitude. Les zones de surveillance sont connues et concernent le franchissement de ponts, les points de contrôle routier et le suivi de convois militaires.

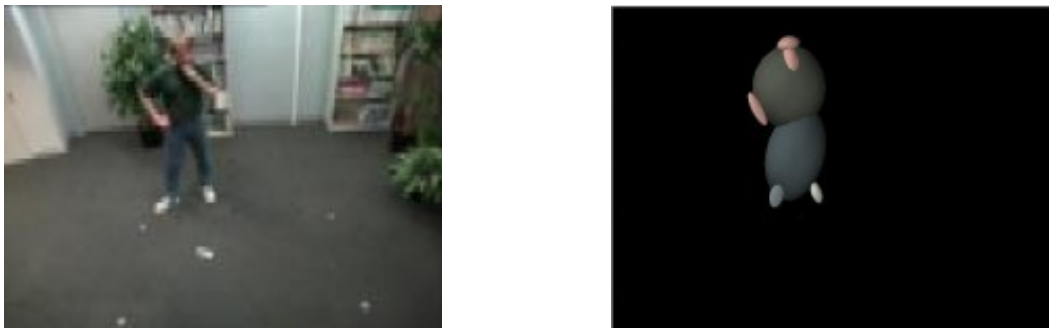
Différentes caméras permettent de suivre une cible en transmettant des évènements symboliques à un contrôleur, qui a un résumé des activités détectées dans une zone de couverture assez large. Une seule personne ne pouvant contrôler en même temps des dizaines de caméras, un des objectifs de ce projet était qu'une seule personne assure le contrôle d'une grande zone, à l'aide de capteurs multiples. Le suivi de plusieurs personnes, voitures et leurs interactions dans un environnement complexe urbain est une tâche difficile. L'approche de VSAM est de fournir une interface graphique et interactive qui place des agents dynamiques de façon automatique, et représentant les personnes et les voitures dans une vue synthétique de l'environnement. La visualisation des évènements de la scène n'est pas fixée à une résolution initiale et à un angle de vue d'une seule caméra. L'interface est une carte de la zone avec tous les objets et les vidéos superposés à la carte. Des caméras thermiques ont été ajoutées en fin de projet.

Dans VSAM, l'approche est similaire à celle de [Grimson E., Viola P.]. La soustraction du fond est robuste grâce à un modèle statistique dynamique du fond. Les objets individuels doivent être suivis au cours du temps. Pour cela, les blobs générés par la détection de mouvement sont mis en correspondance dans les images successives de la vidéo.
Beaucoup de systèmes de suivi sont basés sur le filtre de Kalman qui ne tolère pas des hypothèses multiples à cause de la nature des densités gaussiennes unimodales. [Isard M., Blake A., 96] proposent une approche stochastique (algorithme de « Condensation ») qui gère des hypothèses alternatives. Dans VSAM, une approche plus simple à base de fonction de coût de mise en correspondance image par image est proposée. Les blobs sont mémorisés avec les informations de trajectoire des centroïdes des objets (position et vitesse), « l'apparence » des blobs sous la forme d'un « template » image, la taille des blobs en pixels et l'histogramme de couleur des blobs. La position et la vitesse de chaque « blob » sont déterminées à partir de l'instant précédent et servent à la prédiction de la position dans l'image courante. Une fonction de coût est déterminée entre un objet identifié et un « blob » candidat en mouvement, pour la mise en correspondance de chaque « blob » image après image.

L'objectif final de VSAM est sa capacité à identifier des classes d'objets (humain, groupes et voitures) et déterminer les actions individuelles (dans les catégories « courir » ou « marcher » pour le mouvement des personnes), grâce à un réseau de neurones permettant une classification de l'objet. La compréhension des activités humaines est un problème encore ouvert dans le domaine de la vidéo surveillance automatique.

Depuis les années 1997, la détection et l'analyse des mouvements humains en temps réel à partir d'images vidéo est devenue possible grâce à l'algorithme PFINDER (cf. figure 42) de [Wren C.R., et al.] et W^4 [Haritaoglu I., Harwood D., Davis L.S. 98]. Le corps humain est détecté par ses membres (mains, pieds, tête) qui sont suivis et mis en correspondance avec un modèle *a priori*, tel le modèle « cardboard » [Ju S., Black

M., Yacoob Y.].



(a) Image vidéo d'entrée (b) Représentation 2-D des statistiques des blobs

Figure 42 : PFINDER

Dans les scènes d'extérieur, en général une unique caméra est insuffisante pour suivre un objet pendant longtemps. Les objets peuvent se trouver en occultation par les éléments extérieurs : les arbres et les bâtiments. Une solution prometteuse est d'utiliser un **réseau de caméras** pour un suivi d'objets de façon coopérative, et de façon coordonnée d'une caméra à l'autre. [Matsuyama T.] a présenté une telle approche dans des environnements intérieur où quatre caméras suivent un objet en mouvement sur le sol.

Dans VSAM, les objets sont géolocalisés afin de déterminer où chaque caméra doit regarder. L'orientation et le zoom des caméras les plus proches sont contrôlés pour amener l'objet dans son champ de vue, et les objets d'intérêts en mouvement sont recherchés.

1.2 Le projet ADVISOR-INRIA

ADVISOR (§ 2.2) « Annoted Digital Video for Surveillance and Optimised Retrieval » [Siebel N., Maybank S. et al.] est un projet européen (IST-1999-11287) de l'équipe ORION de l'INRIA Sophia-Antipolis, en vidéo surveillance multi-caméras, impliquant trois partenaires académiques (Univ. Kingston, Univ. Reading, KCL Londres) et trois partenaires industriels (THALES, BULL, VIGITEC), entre janvier 2000 et mars 2003. Ce système a pour objectif de sécuriser les transports publics par la détection automatique de **situations anormales** temps réel pouvant conduire à des accidents, de la violence ou des actes de vandalisme (cf. figure 43).

(a) Suivi d'un groupe de personnes (b) Analyse du mouvement de la foule

Figure 43 : Analyse d'images dans ADVISOR [Siebel N., Maybank S. et al.]

Sans système automatisé, dans un système de sécurité contenant une centaines de caméras vidéos dans une station de métro, les opérateurs humains ne pouvant visualiser que quelques caméras à chaque instant, certains

accidents détectés tardivement engendraient vandalisme et violence. Le système ADVISOR permet d'assister la surveillance humaine, en sélectionnant les écrans de surveillance. Une analyse temps réel des vidéos génère des alarmes dans les cas de comportements dangereux détectés. Les séquences vidéos d'intérêt sont archivées pour les problèmes de surveillance avec une application à la surveillance des métros. ADVISOR est le premier système intégrant à la fois le suivi de personnes, le contrôle des foules, et l'analyse des comportements. Ce système a été testé dans les stations de métro de Barcelone, Londres et Bruxelles.

1.3 Le projet BEHAVE

BEHAVE est un projet anglais de Robert Fischer à « l'Engineering and Physical Science Research Council » (EPSRC), pour la détection des comportements anormaux et/ou criminel [Andrade E., Blunsden S., Fisher R.].
Le projet est composé de deux volets. Le **volet 1** s'intéresse à la compréhension des interactions subtiles entre les personnes, à partir des méthodes de reconnaissance des comportements habituels dans un petit groupe d'individus, visant à différencier une salutation d'une bagarre, à l'aide de modèles de Markov cachés dynamiques pour le suivi des individus. Parmi les patterns de flux utiles dans le suivi court terme, une classification statistique permet de différencier les patterns normaux (les supporters quittant un stade de foot ont des patterns de mouvement standard) et anormaux (la densité de la foule peut rendre impossible le suivi individuel, donc l'identification des bagarreurs, et l'interruption du flux est alors détectée). Le but est de différencier les comportements normaux et anormaux par des modèles probabilistes du flux, issus du suivi court terme. Le **volet 2** intervient quand le suivi d'individus court terme n'est plus possible du fait du nombre de personnes en interaction croissant, car les individus ne peuvent être suivis que pendant quelques images et dans des images contenant peu de personnes. Le volet 2 permet l'analyse de la foule (pour la compréhension des scènes dynamiques, [Remagnino P., Shihab A., Jones G.] et [Buxton H., Gong S.]). Dans ce cas, une interprétation symbolique des comportements devient impossible, il faut analyser le flux de façon statistique.

Les résultats sont bons en ce qui concerne les interactions discrètes à l'aide d'un modèle probabiliste et d'une représentation symbolique. En revanche, il est plus difficile de modéliser les interactions subtiles parmi d'autres interactions peu différentes, comme les comportements de bagarre. En ce qui concerne la reconnaissance de comportements dans la foule tels les évènements sportifs, la question est toujours d'actualité pour la détection d'évènements criminels et la prévention.

1.4 Le projet AVITRACK

AVITRACK (§ 2.1) [Fusier F., Valentin V., Bremond F, Thonnat M.] est un projet IST européen en collaboration avec Silogic S.A. Toulouse (France), University of Reading (UK), CCI Aéroport Toulouse Blagnac (France), Fedespace (France), Tekever LDA, Lisbon (Portugal), ARC Seibersdorf research GMBH, Wien (Austria), Technische Universitaet, Wien, (Austria), IKT (Norway) et Euro Inter Toulouse (France). Ce projet a débuté en février 2004 et s'est terminé en Février 2006. Son objectif principal était l'identification des activités autour des avions stationnés à leur parking, allant des événements simples impliquant un objet mobile comme l'arrivée ou le départ des véhicules au sol aux scénarios complexes comme faire le plein ou le chargement de bagage.

1.5 Le projet CASSIOPEE-INRIA

Le projet **CASSIOPEE** [Avanzi A., Bremond F., Tornieri C., Thonnat M.], du groupe Crédit Agricole avec Eurotelis (Securitas) et Ciel (4 ans 2002-2006) pour la conception de sites de vidéo surveillance bancaire, a pour objectif de détecter des comportements définis comme « à risque » à partir d'une acquisition vidéo continue et d'une connaissance a priori de l'agence. L'objectif n'est pas de détecter à coup sûr les comportements mais plutôt de détecter des comportements potentiellement intéressants afin que leur prise en charge soit réalisée par la station centrale de vidéo surveillance pour lever le doute. Le projet regroupe les compétences d'une banque, d'un intégrateur de systèmes d'acquisition vidéo, d'un opérateur de télésurveillance et de l'INRIA.

1.6 Le projet VIGITEC

VIGITEC, appelé également « Videa », est un projet qui a commencé en novembre 2003 et s'est terminé en novembre 2005 [Velastin S.]. Le but de ce projet est de transférer une partie de la technologie de vidéo surveillance de l'équipe d'ORION dans des produits industriels pour l'identification de comportements humains spécifiques, tels que le contrôle d'accès de bâtiment et les violences urbaines.

1.7 Le projet CAVIARE-INRIA

CAVIARE « Context Aware Vision Picture-based Active Recognition » [Jorge P.M., Marques J.S., Abrantes A.J] est un projet européen (IST 2001) de l'INRIA Grenoble, qui étudie des techniques d'analyse d'images pour améliorer les performances des systèmes de surveillance dans les environnements urbains et les centres commerciaux.

1.8 Le projet PASSWORDS

Le projet **PASSWORDS** est un projet européen **ESPRIT** avec VIGITEC, SEPA, DIBE, AUCHAN (3 ans, 1994 à 1997) dont l'objectif est d'assurer la transition entre les ingénieurs experts ([Chleq N., Thonnat M.], [Bogaert M., Chleq N., Cornez P., Regazzoni C., Teschioni A., Thonnat M.]).

1.9 Les projets dans l'industrie

Des systèmes de surveillance existent pour la **détection de colis abandonnés dans les terminaux d'aéroports**, les personnes **dangereusement proches des rails dans les stations de métro** (le métro londoniens contient 6000 caméras à lui seul), les voitures conduisant à **contresens dans les tunnels et sur les routes** [Siemens].

Un système de vidéo surveillance a été développé au département « **Real-time Vision and Modeling Department** » au « **Siemens Corporate Research** » (SCR) à Princeton, New Jersey. [Zhu Y., Comaniciu D., Pellkofer M., Koehler T.] a développé une technique appelée « Robust Information Fusion » qui est une méthode statistique visant à pondérer les données issues des différentes sources. Un autre domaine de recherche du SCR est l'apprentissage statistique afin d'améliorer la robustesse des systèmes de vidéo surveillance. En effet, le modèle statistique explique les variations observées dans les données. Des applications en **reconnaissance de trafic autoroutier** ont vu le jour et le SCR s'est équipé de caméras additionnelles telles les caméras radar, infrarouge, et ultrasons afin de communiquer avec les automobilistes trop proches les uns des autres et éviter les collisions. Des technologies de vision basées modèle sont en développement, afin de suivre un modèle en 3D, déterminer sa position et son orientation, et sa structure 3D à partir du mouvement.

Dans les **aéroports**, « **the Sistore CX EDS (Enhance Detection Solution) monitoring system** » de **Siemens Building Technologies** (SBT) à Karlsruhe, Allemagne, est un système de détection automatique de mouvement, et de suivi automatique des objets. Le senseur vidéo peut apprendre les situations « normales » en mémorisant les états les plus fréquents pendant un intervalle de temps, afin de reconnaître une situation « **anormale** ». EDS peut aussi extraire des primitives, telles que la taille et la vitesse, pour distinguer une personne d'un animal ou d'un véhicule par exemple. Un autre avantage de connaître ce qu'est un fond normal est de pouvoir détecter de façon automatique les actions de sabotage. Si une personne malveillante tourne la caméra la faisant pointer dans une autre direction, modifiant ainsi le fond, la caméra ne reconnaît pas son environnement usuel et provoque une alarme. Siemens a installé un système digital Sistore complet pour les **jeux asiatiques de 2006**, à Doha au Qatar. Plus de 1300 caméras détectent, évaluent et suivent des mouvements suspects dans la cité sportive. Le système de surveillance a également été mis en place en Allemagne. La **police fédérale Berlinoise** protège certains quartiers avec ce système de vidéo surveillance.

Siemens participe également à un projet américain pour la **sécurité des vols aériens** : le **projet SAFEE** « Security od Aircraft in the Future European Environment » enregistre les évènements à bord d'un avion et les

compare avec des images enregistrées. Si le système détecte des mouvements ou des conversations suspectes, il déclenche une alarme et envoie un message crypté immédiatement. Si des terroristes pénètrent dans le cockpit et tente de détourner l'avion, SAFEE compare la position de l'avion avec les limites de zones enregistrées, et remet l'avion dans sa trace originale automatiquement.

Le bureau « **Fraunhofer Allianz Vision** » office à Erlangen en Allemagne existe depuis une dizaine d'années pour créer de la synergie entre les différents instituts Fraunhofer. Les applications sont les systèmes d'**assistance à la conduite pour la conduite d'engins, les systèmes de détection automatique pour l'industrie alimentaire**, la santé par exemple pour vérifier la qualité de l'air et la température ambiante, enfin pour la **reconnaissance biométrique du visage dans le but d'identifier une personne**. Des systèmes de surveillance permettent d'identifier des **voleurs à la sauvette dans une foule** grâce à leur mouvement. Cependant, il faut encore développer des systèmes prenant en compte les variations d'apparence des personnes car les systèmes travaillent par comparaison d'images avec des images de référence.

2 Présentation détaillée de quelques systèmes de vidéo surveillance

2.1 AVITRACK

[Fusier F., Valentin V., Bremond F, Thonnat M.] proposent un système de compréhension vidéo temps réel pour la reconnaissance des activités sur des séquences vidéo en trois étapes : suivi, maintenance de la cohérence et compréhension (cf. figure 44).

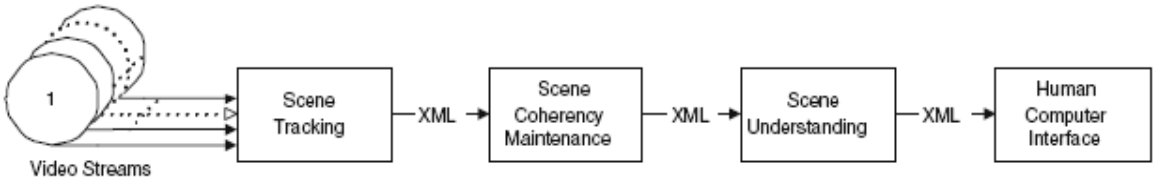


Figure 44 : système de compréhension vidéo temps réel en trois parties, suivi, maintenance de la cohérence et compréhension [Fusier F., Valentin V., Bremond F, Thonnat M.].

Le but est de faire du suivi robuste en vue de la reconnaissance d'évènements extérieur dans des conditions d'applications temps réel avec un **réseau de caméras**, et de reconnaître de façon automatique des évènements complexes avec plusieurs acteurs en interaction. Le système a été évalué en monitoring des **activités d'aéroport** sous des conditions normales d'utilisation, démontrant sa robustesse. **Les erreurs de suivi sont corrigées par la maintenance de la cohérence.** Le système fonctionne en **vision cognitive** mais aussi dans les activités d'aéroports, avec des perspectives dans les **stations de trains**.
Trois étapes sont prévues : le suivi de la scène, la maintenance de la cohérence, et la compréhension de la scène. Le **suivi de la scène** consiste à détecter des objets basés sur leur mouvement, les suivre au cours du temps et les classifier (personne, voiture, avion), et par fusion de données calculer la position 3D des objets mobiles dans un système de coordonnées globales. La **maintenance de la cohérence** a pour objectif de calculer une représentation cohérente de la scène 3D avec une remise à jour de son évolution au cours du temps. La **compréhension de la scène** reconnaît en temps réel des évènements vidéo. L'objectif est un suivi robuste capable de reconnaître des évènements quelles que soient les conditions, sur une aire d'aéroport pour le projet européen **AVITRACK**.

2.1.1 La détection de mouvement

La détection de mouvement segmente l'image en régions connectées des pixels d'avant-plan représentant les objets en mouvement. Un algorithme de segmentation par soustraction du fond est basé sur une distribution

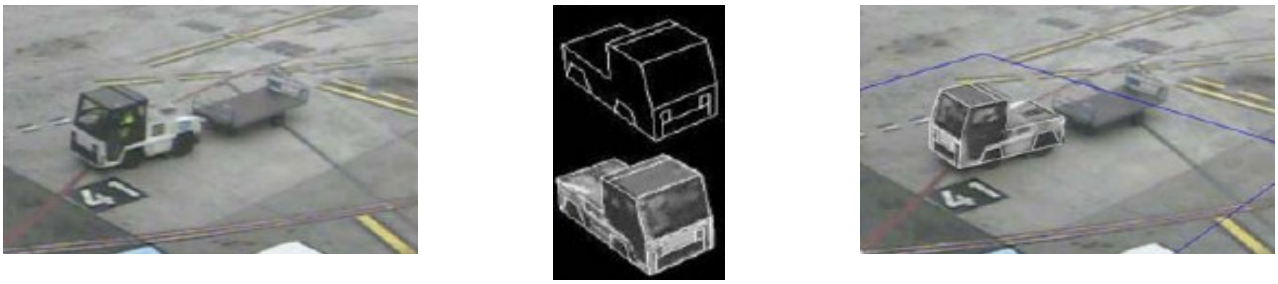
gaussienne (couleur moyenne et variance) en couleur RGB normalisée pour modéliser le fond [Wren C.R., et al.]. Cela a été étendu en incluant une détection des composantes des ombres et des hautes lumières [Horprasert T., Harwood D., Davis L.] rendant la détection de mouvement robuste aux changements d'illumination. Ces résultats permettent de suivre des objets dans différentes images.

2.1.2 Suivi d'objet

Le module de suivi de scène de AVITRACK comprend deux étapes, suivi d'objet 2D avec une seule caméra, et suivi d'objet dans un monde 3D. Le suivi d'objet par caméra détecte des objets en mouvement, puis les suit et les classe par une reconnaissance d'objets hiérarchiques. Les objets suivis par les huit caméras sont envoyés à un serveur central où les observations multiples sont fusionnées.
Le suivi d'objets temps réel peut être décrit comme un problème de correspondance et implique de trouver des objets en correspondance d'une image à l'autre. L'algorithme de suivi de Kanade-Lucas-Tomasi (KLT) pour le suivi de primitives [Shi J., Tomasi C.] est utilisé dans AVITRACK. Mais cet algorithme considère que les primitives sont indépendantes et les suit de façon individuelles. Pour passer à un niveau de suivi d'objets, le KLT est incorporé dans un processus de suivi haut niveau regroupant des primitives en objets, maintenant une association entre eux et prenant en compte les interactions complexes entre les objets.

2.1.3 Reconnaissance d'objets

Les objets sur le tarmac tels que les personnes, les voitures au sol (cf. figure 45), les avions, sont classés par un classifieur obtenu par mélange de gaussiennes entraînées sur des descripteurs tels que la largeur 3D, la hauteur, la dispersion et le ratio, inspirés des travaux de Collins [Collins R., et al.b].



(a) Image montrant un véhicule de transport sur une zone d'aéroport.

(b) Le modèle basé contour et apparence 3D du véhicule de transport, pour la classification.

(c) Le modèle d'apparence adapté au véhicule.

Figure 45 : Classification d'un véhicule de transport en zone aéroportuaire [Fusier F., Valentin V., Bremond F, Thonnat M.].

2.1.4 Fusion de données

La méthode de fusion de données est basée sur une approche par filtre de Kalman et plus proches voisins [Bar-Shalom Y., Li X.] avec un modèle à vitesse constante. L'étape d'association de données associe des pistes prédites avec des mesures observées dans chaque caméra. Dans l'algorithme des plus proches voisins, la meilleure mise en correspondance est définie pour être la seule observation par caméra. Pour de multiples pistes suivies, et différentes caméras, l'algorithme des plus proches voisins associe le plus proche voisin par caméra pour chaque piste suivie. Le filtre de Kalman est remis à jour pour chaque piste avec les mesures fusionnées.

2.1.5 Maintenance de la cohérence dans des scènes 3D dynamiques

L'interprétation haut niveau des scènes 3D est issue de la coopération du suivi de scène et de la compréhension

de la scène. Le but de la maintenance de la cohérence est d'analyser la dynamique des objets mobiles afin d'améliorer la robustesse du suivi de scène, par exemple en gérant les **occultations** ou mauvaises détections sur plusieurs images et les **changements d'objets mobiles** (**disparition**, **apparition**). La maintenance de la cohérence est décrite dans une tâche de **suivi long term**e et dans une tâche de **suivi global**. Le **suivi long terme** utilise une fenêtre temporelle pour augmenter les performances du suivi des objets mobiles, grâce à une analyse de graphe temporelle, dés que le suiveur image par image rencontre des difficultés comme en cas d'occultations. Bien que le suivi image par image et long terme soient efficaces, ils montrent quelques limites impliquant l'utilité d'un suivi global. Ces limites sont par exemple la perte des objets suivis à cause des occultations ou l'intégration des objets mobiles dans le fond après une longue période, des sur ou sous détections à cause des ombres ou du manque de contraste, et un mélange de l'identité des objets suivis quand plusieurs objets se regroupent. Pour éviter ces problèmes, un module haut niveau appelé le **suiveur global** a en charge d'augmenter les compétences du suivi long terme afin de fournir des données cohérentes à la compréhension de la scène. Le suiveur global utilise la connaissance *a priori* de l'environnement observé et une analyse 3D spatio-temporelle, avec un ensemble de règles (si alors).
Dans le suivi long terme, c'est la cohérence temporelle de chaque objet mobile qui est vérifiée, tandis que dans le suivi global c'est la cohérence spatio-temporelle de tous les objets mobiles.
Le suiveur global a connu des applications ayant du succès comme le contrôle d'accès dans les bâtiments, les agences bancaires, le monitoring des activités des aéroports (cf. figure 46), témoignant de sa généricité. Son utilité est démontrée dans le cas de la perte des objets suivis. Par exemple, lorsqu'un véhicule est resté très longtemps au même endroit, la remise à jour de l'image de référence, l'image du fond, tend à intégrer le véhicule dans le fond, résultant en une mauvaise détection et une perte de ce véhicule. Cela peut être le cas d'un véhicule Tanker stationné sous les ailes de l'avion pendant le remplissage de gasoil de l'avion.

(a) Avant le suiveur global, un véhicule de chargement est détecté comme plusieurs objets mobiles (sur détection).

(b) Aprés le suiveur global, le véhicule de chargement est correctement détecté comme un seul objet mobile.

Figure 46 : Apport du suiveur global dans un aéroport [Fusier F., Valentin V., Bremond F, Thonnat M.].

Le suiveur global et long terme ont amélioré les performances du système et autorisé le suivi sur une large gamme de personnes, voitures et avions qui interagissent ensemble sur le tarmac. La **compréhension de la scène** est ainsi capable de reconnaître des activités dans des situations plus complexes.

2.1.6 Compréhension de la scène

Le but de la compréhension de scène est de fournir une **interprétation haut niveau des trajectoires des objets mobiles suivis** en termes de **comportements humains**, activités des véhicules, ou de leurs interactions. Deux catégories d'approches ont été utilisées pour reconnaître des évènements vidéo, soit avec un réseau de neurones probabiliste, soit avec un réseau symbolique correspondant aux évènements à reconnaître. Pour la communauté de **vision par ordinateur**, c'est l'approche par réseau de neurone qui est préférée. Les noeuds du

réseau correspondent aux évènements vidéo reconnus à un instant donné grâce à une probabilité calculée [Hongeng S., Bernard F., Nevatia R.] et cela fonctionne bien pour les évènements courts mais pas pour les évènements complexes impliquant plusieurs personnes. Pour la communauté **intelligence artificielle**, un événement vidéo est reconnu par un **réseau symbolique** dont les noeuds correspondent à une reconnaissance booléenne des évènements vidéo [Pinhanez C., Bobick. A.]. Une approche traditionnelle est basée sur une représentation déclarative des évènements vidéo définis comme un ensemble de contraintes logiques et spatio temporelles. [Chleq N., Thonnat M.] ont propagé des contraintes temporelles pour la vidéo surveillance. Cette méthode reconnaît un scénario par prédiction des évènements vidéo attendus, afin d'être reconnus à l'instant suivant. [Vu T., Bremond F., Thonnat M.] ont étendu cette dernière méthode dans le cas de la reconnaissance des activités complexes impliquant plusieurs objets physiques de différents types (individus, véhicules, avions) dans un champ large observé par une caméra en réseau et pendant longtemps.

La méthode proposée reconnaît des évènements vidéo à l'aide d'un raisonnement spatio-temporel prenant avantage de la connaissance *a priori* de l'environnement observé et des modèles d'évènements vidéo. Une représentation par formalisme aide les experts à décrire les évènements vidéo d'intérêts qui arrivent dans la scène observée. Une connaissance du contexte de la scène observée est une information *a priori* que le système doit connaître pour interpréter les activités. La connaissance contextuelle est statique et dynamique. La connaissance statique correspond à l'information des objets statiques et de la scène 3D vide (description géométrique et sémantique des zones spécifiques). La connaissance dynamique contextuelle concerne les zones d'intérêt des véhicules qui peuvent interagir avec les autres voitures ou personnes. Cette connaissance est nécessaire si on veut reconnaître des activités impliquant des véhicules en stationnement.

2.1.7 La reconnaissance d'évènements vidéo

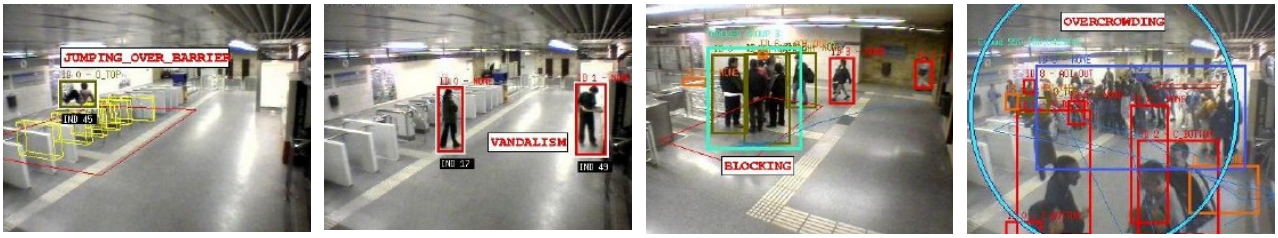
La reconnaissance des activités vidéo automatique est une tâche difficile pour la **Vision Cognitive** du fait qu'elle s'intéresse à la reconnaissance des activités complexes impliquant plusieurs objets physiques de différents types. Un algorithme temps réel de reconnaissance d'évènements vidéo est décrit dans [Vu T., Bremond F., Thonnat M.]. Le processus de reconnaissance d'évènements du projet **AVITRACK** se sert de la cohérence du suivi d'objets, de la connaissance *a priori* de la scène statique et dynamique et des modèles d'évènements prédéfinis.

2.1.8 Compréhension vidéo pour le monitoring des activités aéroportuaires

La compréhension vidéo a été validée dans une application d'activité d'aéroport dans le projet européen AVITRACK. Le système a démontré ses capacités dans la compréhension de la scène dans les environnements d'aéroports afin de reconnaître de façon automatique les activités autour des parkings sur un tarmac, donc reconnaître en temps réel les interactions entre personnes et véhicules. Le langage de description des évènements a démontré son efficacité dans bon nombre d'applications comme le **monitoring des stations de métro** [Cupillard F., Avanzi A., Bremond F., Thonnat M.], des agences bancaires [Georis B., Maziere M., Bremond F., Thonnat M.], de l'intérieur des trains, parking et tarmac d'aéroport.

2.2 ADVISOR

[Cupillard F., Avanzi A., Bremond F., Thonnat M.] se proposent de reconnaître des personnes isolées, des groupes de personnes, ou bien des comportements de foule dans le contexte de la surveillance des scènes de métro utilisant plusieurs caméras (cf. figure 47). Les scènes décrites sont celles de bagarre ou de vandalisme dans des environnements texturés, les stations de métro. Ce travail s'inscrit dans le projet **ADVISOR**.
Le système de vidéo interprétation est composé d'un module de vision et d'un module de reconnaissance de comportement. Le module de vision comprend :
-un détecteur de mouvement;
-un suivi image par image générant un graphe des objets en mouvement dans chaque caméra calibrée;
-un graphe global pour toute la scène observée par l'ensemble des caméras.



(a) « Fraude » reconnue par un automate.

(b) « Vandalisme » reconnu par un réseau de contraintes temporelles.

(c) «Blocage» reconnu par un automate.

(d) « Foule » reconnu par un arbre ET/OU.

Figure 47 : Illustration de 4 comportements reconnus par le système d'interprétation [Cupillard F., Avanzi A., Bremond F., Thonnat M.].

Le détecteur de mouvement détecte les régions en mouvement dans la scène et les classifie dans une liste d'objets mobiles avec des labels correspondant à leur type (« une personne »). La détection des personnes en mouvement dans la scène a lieu grâce à la différence entre l'image courante et un modèle du fond contenant l'apparence de la scène sans personne, remis à jour périodiquement pour s'adapter aux changements de lumière et aux mouvements de la caméra. Une image binaire de mouvement (contenant certains pixels classés en mouvement) et une image du fond sont fournies au système d'analyse de niveau supérieur afin d'en extraire une description des objets de la scène de façon plus abstraite. Un pixel de l'image courante est classé en « mouvement » s'il n'est pas expliqué par le modèle. Le modèle de fond le plus simple est un fond statique sans personne dans l'image. Dans le cas des changements de luminosité ou de mouvements du fond (les arbres ou le vent), des modèles statistiques plus élaborés sont nécessaires, comme les mélanges de gaussiennes du MIT ([Stauffer C., Grimson W.E.L.b]) ou le modèle non-paramétrique de l'université du Maryland [Elgammal A.M., Harwood D., Davis L.S]. Une personne en mouvement peut être détectée par un modèle décrivant l'apparence de la personne et ajustable aux mesures dans l'image grâce aux paramètres du modèle. Mais le problème peut être complexe du fait de la complexité du corps articulé et des changements dans l'apparence selon les points de vue, ainsi que des problèmes d'occultations. Les modèles simples sont plus rapides lorsqu'il s'agit de travailler en temps réel mais ils rencontrent des difficultés en présence d'occultations. Un modèle plus complexe est favorable dans ce cas mais le temps réel n'est plus toujours atteint.

Une liste d'objets mobiles est obtenue à chaque image, chacun décrit par les paramètres 3D (centre de gravité, position, etc.) et par une classe sémantique. Le rôle du suivi image par image est de relier d'une image à l'autre la liste des objets mobiles calculés par le détecteur de mouvement. La sortie de ce module est un **graphe des objets mobiles** (cf. figure 48), **qui fournit toutes les trajectoires possibles qu'un objet mobile** peut avoir. Le lien entre un nouvel et un ancien objet mobile est calculé en fonction de trois critères: la similitude entre les classes sémantiques, leur distance 2D dans l'image et 3D dans le monde réel.

Les graphes individuels sont combinés en un graphe global afin de prendre avantage des caméras calibrées qui voient la même scène mais de divers points de vue, et utilisées pour le suivi long terme. Un modèle 3D de la scène pour chaque caméra et une connaissance *a priori* contextuelle de la scène observée sont utilisés. Le modèle de la scène comprend les positions 3D et dimensions des objets statiques de la scène (une machine à vendre les tickets par exemple) et les zones d'intérêt. Les attributs sémantiques sont associés aux objets ou aux zones d'intérêt pour être utilisés dans la reconnaissance du comportement.

ADVISOR détecte et suit des individus autant que des groupes d'individus. L'analyse de la foule s'effectue via l'analyse de mouvement à partir du modèle du fond. La surpopulation ou la congestion de zones pré définies (sorties ou Escalator) sont détectées, ainsi que la stationnarité d'objets et de personnes, et le flux de personnes à contre sens.
Le module d'analyse de comportement peut détecter un certain nombre de comportements comme la violence entre personnes, le vandalisme contre les équipements comme les machines vendeurs de billets, l'évasion de

personnes escaladant les barrières au lieu d'utiliser un ticket.

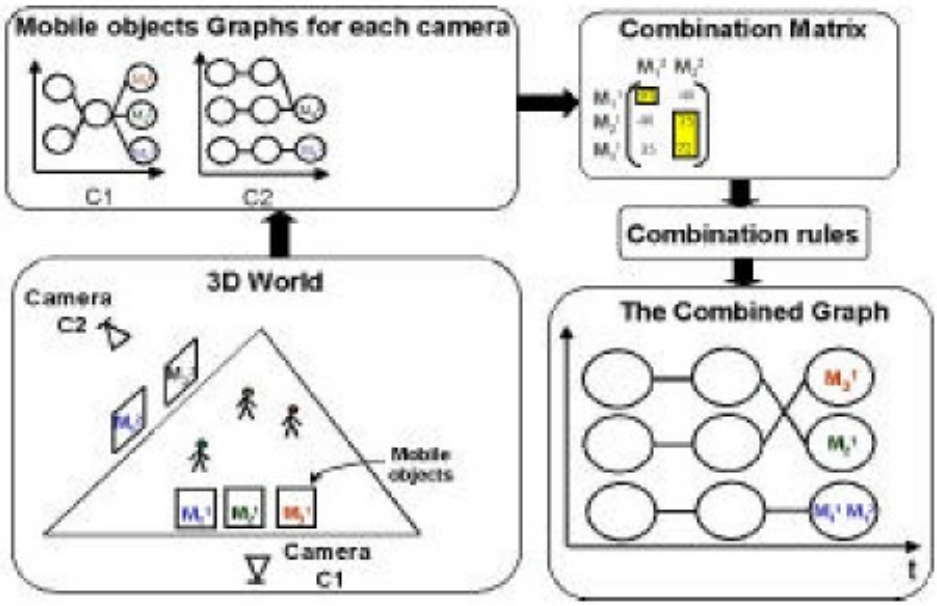


Figure 48 : Illustration de la combinaison de multiples caméras [Cupillard F., Avanzi A., Bremond F., Thonnat M.]. Trois personnes évoluent dans la scène. La caméra C1 détecte les trois objets mobiles tandis que la caméra C2 détecte seulement deux objets mobiles. La matrice de combinaison détermine une forte correspondance entre l'objet mobile M_1^1 de C1 et l'objet mobile M_1^2 de C2. Ces deux objets mobiles sont fusionnés dans le graphe combiné. La matrice de correspondance détermine également une correspondance ambiguë entre les deux objets mobiles M_2^1 et M_3^1 de C1 et l'objet mobile M_2^2 de C2. Les deux objets mobiles M_2^1 et M_3^1 détectés par C1 sont sélectionnés dans le graphe combiné.

2.3 La vidéo surveillance avec une architecture à base de connaissances

Une **architecture à base de connaissances** est proposée pour la surveillance vidéo dans [Georis B., Bremond F., Thonnat M.] (cf. figure 49). Une base de connaissance est composée de trois types de connaissances : La connaissance du domaine, la connaissance de l'environnement de la scène, et la connaissance des traitements vidéo. Chaque type de connaissance est fourni par des experts. Le rôle de la composante du contrôle (raisonnement) est d'exploiter toutes les connaissances *a priori* ou apprises contenues dans la base de connaissance afin de produire un plan. Le contrôle est conduit par les données pour guider le processus de prise de décision. Il contient des règles sous la forme condition/action.
Le formalisme utilisé pour permettre aux experts d'exprimer leurs connaissances directement utilisées par le système, est dédié à la représentation des connaissances pour les programmes de supervision [Thonnat M., Moisan S., Crubezy M.], incluant des règles de production.
La première tâche est la détection et la classification des objets d'intérêt présents dans la scène. Une fois l'image acquise, il faut générer une image du fond (image de référence) permettant de détecter des régions en mouvement par soustraction de l'image courante à l'image de référence. Le résultat seuillé donne lieu à des « blobs » ou régions en mouvement, associés avec un ensemble de primitives comme la densité ou la position. Dans l'étape de classification, les petits « blobs » correspondant au même objet physique sont regroupés pour corriger les erreurs de segmentation, et une séparation est opérée sur les larges « blobs » correspondant à plusieurs objets physiques. Un ensemble de primitives 3D comme la position 3D, la largeur et la hauteur sont calculés pour chacun des « blobs ». En comparant cet ensemble de primitives 2D et 3D avec des modèles

prédéfinis, ces « blobs » sont classés en diverses classes prédéfinies (personne, groupe, voiture, etc). Ces « blobs » avec leur label sont appelés les « objets physiques d'intérêt ». Une fois la détection obtenue, la liste des « objets physiques d'intérêt » est suivie dans une analyse spatio temporelle, composée des étapes de **suivi image par image**, **fusion** et **suivi long terme**.

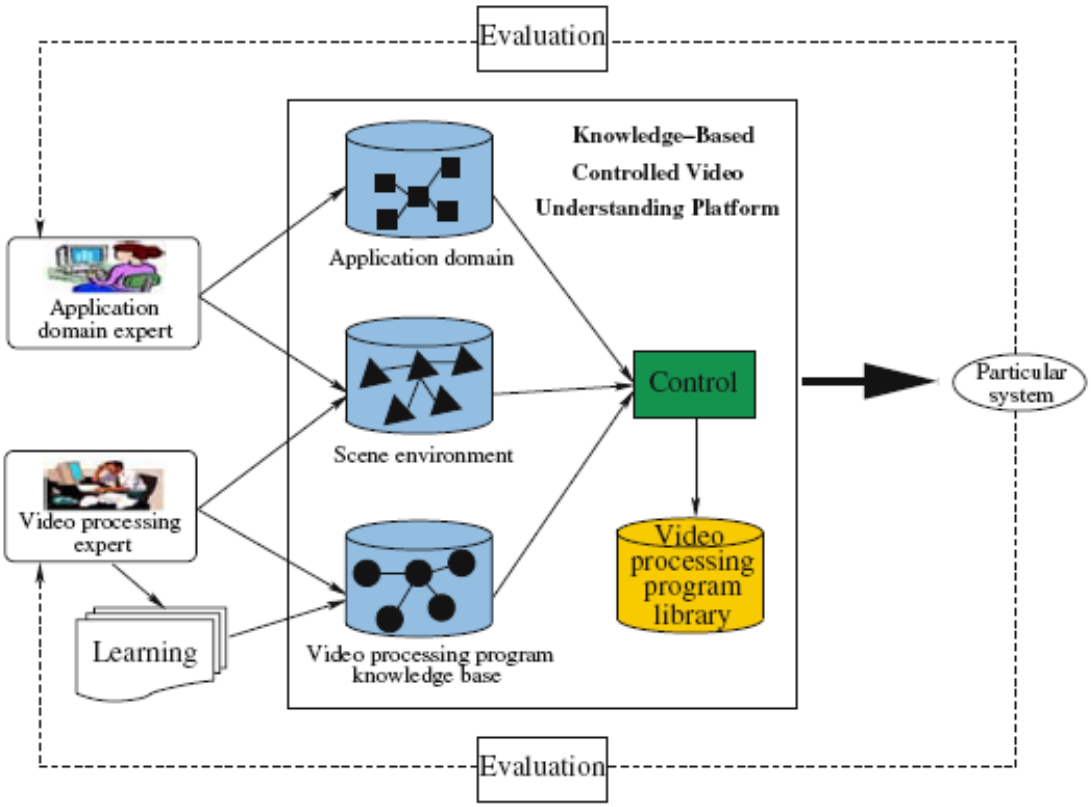


Figure 49 : Un système de surveillance vidéo avec un architecture à base de connaissances, composée de la base de connaissances (bleu), du contrôle (vert) et de la librairie de programmes (jaune) [Georis B., Bremond F., Thonnat M.].

2.3.1 Suivi image par image

Le **suivi image par image** a pour objectif de relier les objets physiques d'intérêt d'une image à la suivante [Georis B., Bremond F., Thonnat M., Macq B.]. Un graphe contenant les objets physiques détectés remis à jour à chaque instant et un ensemble de liens entre les objets détectés à l'instant t et ceux détectés à l'instant t-1 est crée. Un objet physique associé à des liens temporels vers les précédents est appelé un « objet physique d'intérêt suivi ». Ce **graphe** fournit toutes les **trajectoires** possibles pour un objet.

2.3.2 Fusion des suivis

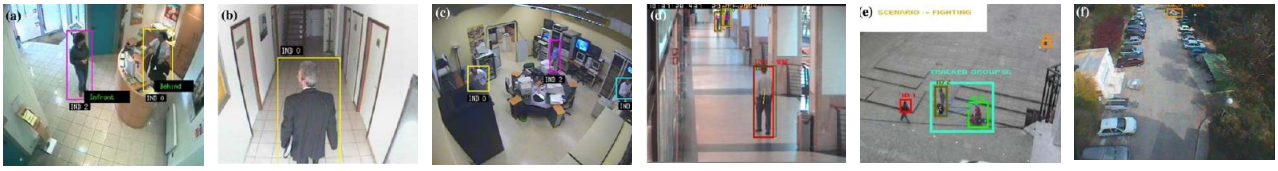
La fusion des suivis consiste à n'obtenir qu'un seul graphe à partir des graphes des objets physiques d'intérêts en provenance des différentes caméras avec des champs de vue se recouvrant. Des matrices établissent la correspondance entre les différentes vues du même objet. Le graphe des objets physiques d'intérêt fusionnés obtenu, contient tous les liens temporels des objets originaux fusionnés et leurs primitives 3D sont les moyennes pondérées des primitives 3D originales. Les pondérations sont calculées en fonction des distances des objets originaux aux caméras correspondantes. De cette façon les primitives 3D résultantes sont en général plus exactes que les originales.

2.3.3 Suivi long terme

Sur les graphes fusionnés des objets, le **suivi long terme** est mis en place. Un ensemble de chemins est calculé dans le graphe, représentant les **trajectoires possibles pour les objets suivis**. Les objets physiques d'intérêt sont suivis avec un délai pour comparer l'évolution des différents chemins. A chacune des images, le meilleur chemin dans le graphe est sélectionné.

2.3.4 Reconnaissance d'évènements

Une fois les objets physiques d'intérêts suivis, la **reconnaissance d'évènements** est mise en place. Selon le type d'évènements à reconnaître (cf. figure 50), il existe différentes méthodes, soit par les **réseaux bayésiens** [Moenne-Locoz N., Bremond F., Thonnat M.] dans le cas d'évènements rapides avec de l'incertitude, soit par les **arbres AND/OR** pour les évènements avec une grande variété d'invariants comme les bagarres. Enfin, pour des évènements impliquant de multiples « objets physiques d'intérêt » et des relations temporelles complexes, la technique est un **réseau de contraintes** dans lequel les noeuds correspondent à des sous évènements et les arêtes à des contraintes temporelles [Vu T., Bremond F., Thonnat M.]. La sortie est une liste des évènements reconnus.



(a) surveillance des banques | (b) comptage de personnes dans un hall d'immeuble | (c) détection de vandalisme dans un bureau | (d) contrôle de magasin | (e) détection de violence sur une place publique | (f) surveillance d'un parking de véhicule

Figure 50 : Cette figure montre 6 applications traitées avec la base de connaissance du système [Georis B., Bremond F., Thonnat M.].

Dans cette architecture, le contrôle appelle les tâches de traitements vidéo, et les programmes venant de n'importe quelle librairie de traitement vidéo peuvent être intégrés, permettant d'étendre la capacité du système sans modifier le contrôle.

2.4 Un réseau synergétique à deux niveaux pour les interactions multi personnes

La reconnaissance des activités des personnes est un processus compliqué, surtout dans les environnements non contraints à cause du bruit et des ambiguïtés des activités des personnes. La structure spatio temporelle des activités des personnes est analysée à différents niveaux de détail [Park S., Trivedi M.M. 07].
-Au plus **haut niveau**, l'activité d'une personne est analysée en termes de suivi de boites englobantes en mouvement appelé l'**analyse de niveau suivi**. Les systèmes de surveillance basés sur le suivi sont utiles pour beaucoup de situations incluant le temps puisque la région en mouvement peut être extraite.
-Au **niveau de détail**, l'activité d'une personne est analysée en termes de la coordination des membres individuels du corps appelé l'**analyse du niveau du corps**. Dans les situations de **surveillance en intérieur**, l'analyse du niveau du corps a été très étudiée. Dans des **situations extérieures**, les performances du système dépendent des processus de vision bas niveau comme la robustesse de la modélisation du fond, une segmentation efficace, etc.

Des systèmes de surveillance basés sur le « **suivi** » existent avec des véhicules et des personnes en mouvement, dans le cas d'un parking ([Oliver N.M., Rosario B., Pentland A.P.], [Remagnino P., Shihab A., Jones G.],

[Velastin S., Boghossian B., Lo B., Sun J., Vicencio-Silva M.]) ou sur « **l'analyse du corps** » [Haritaoglu I., Harwood D., Davis L.S. 00]. Dans certaines applications de suivi, une représentation du corps humain sous forme de boite englobante ou d'une ellipse peut être suffisante pour suivre une personne [Oliver N.M., Rosario B., Pentland A.P.]. D'autres recherches se sont concentrées sur une description plus détaillée du corps humain comme les régions en mouvement ou « blob » ([Remagnino P., Shihab A., Jones G.], [Velastin S., Boghossian B., Lo B., Sun J., Vicencio-Silva M.]).

[Park S., Trivedi M.M. 07] présentent un réseau synergetique à deux niveaux (cf. figure 51) pour les interactions multi personnes et les activités dans des environnements extérieurs, incluant les variations lumineuses, les changements de temps, les ombres en mouvement, les perspectives de caméra et les variations de lieux. Un mécanisme de **bascule adaptative au contexte** est proposé pour passer de l'« analyse » du corps à celle du « suivi » de celui-ci. Le concept de l'espace spatio-temporel pour modéliser les aspects de l'écologie humaine dans les interactions interpersonnels est aussi défini.

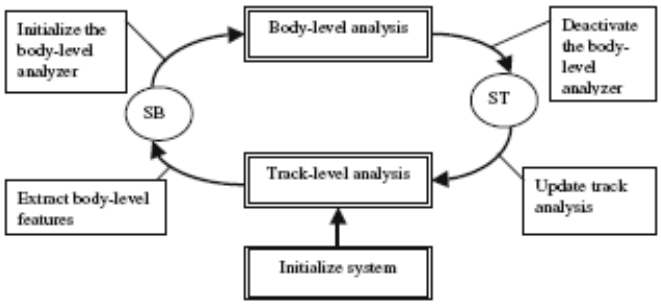


Figure 51 : Les deux étapes de l'analyse d'activités : bascule entre le niveau du corps « body level » (SB) en cas d'information sur le corps, et le niveau du suivi « track level » (ST) quand l'information du corps est insuffisante .

2.4.1 Le niveau « suivi »

Le niveau « suivi » est robuste aux fluctuations environnementales, tandis que le niveau « analyse », basé sur l'apparence, échoue sous des faibles illuminations et des apparences de personnes similaires.
Le niveau « suivi » représente les activités du corps humain en termes des mouvements de son corps. Le niveau « analyse du corps » effectue une représentation plus détaillée des activités en termes d'angles des articulations du squelette ou de positions des membres. La sensibilité du niveau « corps » est affectée par plusieurs sources d'incertitudes comme les occultations, la perspective de la caméra. L'analyse au niveau « suivi » est plus robuste à ces sources d'incertitudes. Le réseau proposé est un intermédiaire efficace entre le « suivi » robuste si l'apparence du corps est dégradée et la moins robuste « analyse » pour l'interprétation des activités.

2.4.2 Le niveau « analyse du corps »

« L'analyse du corps » se focalise sur des activités plus détaillées de personnes individuelles. [Haritaoglu I., Harwood D., Davis L.S. 00] analysent les contours de la silhouette pour détecter des membres du corps comme la tête, les mains, le torse et les jambes. La posture du corps est estimée à partir de la configuration des membres.

Une autre différence est à faire entre les environnements intérieurs et extérieurs.
-Les **environnements intérieurs** ont des conditions de lumière et des fonds plus stables mais les personnes peuvent se trouver en occultation plus facilement par d'autres objets de la scène. Les systèmes de **surveillance intérieur** présentent un faible champ de vue et peuvent fournir des images **haute résolution**. [Park S.,

Aggarwal J.K. 04a] ont analysé les interactions entre personnes dans un environnement intérieur grâce une représentation détaillée du corps humain par de multiples « blobs ».
-Les systèmes de **surveillance extérieure** présentent des variations environnementales comme les changements de conditions climatiques entre le matin et le soir, et les changements de fond. Les systèmes de **surveillance extérieure** doivent être robustes à ces variations. ont un système robuste pour la surveillance basée sur le suivi et la protection privée dans les environnements extérieurs. La plupart des systèmes de surveillance en extérieur font du « **suivi »** et non de « **l'analyse du corps** », à cause du grand champ de vue de la scène impliquant une **faible résolution**. Certains travaux comme ceux de [Zhao T., Nevatia R.] utilisant à la fois **le suivi et les modèles du corps bas niveau** pour analyser les comportements de piétons comme la marche, la course et la position debout.

Un développement récent en surveillance vidéo est l'utilisation de système distribués couvrant différentes zones de la scène avec des champs de vue différents. [Remagnino P., Shihab A., Jones G.] ont présenté un module multi-agent basé surveillance avec une intelligence décentralisée ([Valera M., Velastin S.]). La coopération des différents niveaux de détails (« **analyse au niveau suivi** » et au « **niveau du corps** ») à partir des différentes caméras hétérogènes et l'intégration des multiples niveaux d'analyse des activités humaines est l'objectifs des travaux de .

2.4.3 L'analyse des activités humaines en deux étapes

L'avant plan est segmenté par une modélisation du fond ([Chalidabhongse T., Kim K., Harwood D., Davis L.], [Hall D., Crowley, J. et al.]), suivi par un « **graphe relationnel d'attributs** » pour segmenter et suivre les membres du corps humain.

2.4.4 Représentation multi niveau des mouvements du corps humain

Le suivi avec le « **graphe relationnel d'attribus** » des différents corps humain utilise des boites englobantes et des ellipses gaussiennes 2D pour suivre les personnes dans l'avant-plan. Le mouvement du corps est représenté à différents niveaux: boites englobantes, ellipse 2D, et membres du corps segmentés. La boite englobante et l'ellipse représentent le mouvement global du corps au niveau « suivi » (translation du corps), tandis que les membres du corps segmentés représentent les membres individuels en mouvement au niveau « membres du corps ». Le « **suivi** » sait résoudre les occultations mais l'analyse n'est pas détaillée, tandis que le niveau « **analyse du corps** » fournit une information plus riche sur le corps humain mais échoue en cas d'occultation, d'où l'idée de développer une analyse à deux niveaux synergiques et un mécanisme adaptatif pour passer du niveau suivi au niveau analyse du corps.

2.4.5 La modélisation des activités au niveau des activités du corps humain

Bon nombre d'activités humaines et d'interactions se font tandis que les personnes sont dans la même position. L'analyse du « **suivi** » ne peut pas traiter les primitives d'activités humaines détaillées faites par des personnes stationnaires, comme « **se serrer la main** », « **danser** », etc. Les activités au niveau du « **corps** » des personnes est formulée en termes d'une estimation stochastique des postures et des gestes en utilisant les modèles de Markov caché. Un geste humain est représenté par une séquence de mots de code, et reconnus par des HMM. **Les approches basées HMM pour la reconnaissance des activités** ([Huang K.S., Trivedi M.M.], [Oliver N., Horvitz E., Garg A.]) **avec des ensembles de primitives différentes existent déjà.**
utilisent un ensemble indépendant de HMM pour représenter les gestes du haut du corps, les translations du corps, avec les hypothèses que les gestes individuels ont une évolution indépendante d'un membre à l'autre.

2.4.6 La modélisation des interactions

Il est nécessaire d'associer des primitives visuelles avec des concepts et des symboles pour construire les évènements sémantiques des activités d'une personne. Dans cette approche, la représentation des activités multi-personnes est basée sur une hiérarchie d'évènements [Park S., Aggarwal J.K. 04b]. Une interaction

humaine est une combinaison des actions d'une simple personne et l'action d'une seule personne est composée des gestes de différents membres du corps, tels le mouvement du torse et le mouvement des bras et des jambes. Chaque geste d'un membre du corps humain est un événement élémentaire du mouvement. Il est composé d'une séquence de postures instantanées à chaque image. Un simple événement peut être au niveau du « **suivi** » ou au niveau des activités du « **corps** » humain, en fonction de l'application.

2.5 Suivi de trajectoires à l'aide d'un SVM

De multiples caméras sont déployées dans un parking pour la vidéo surveillance de multiples personnes, en temps réel, gérant des occultations temporaires [Niu W., Jiao L., Han D., Wang Y.-F.]. Il n'y a pas de modèle 3D ni d'analyse des interactions ni de détections de comportements douteux.
La segmentation et le suivi de plusieurs personnes en temps réel est un problème délicat mais important en vidéo surveillance. Dans des applications proches pour lesquelles une analyse à **haute résolution** est nécessaire, il existe plusieurs techniques en reconnaissance de visage, de geste et de la marche. Dans le cadre de la vidéo surveillance, la reconnaissance de la marche est un des problèmes les plus intéressants, classée entre les **méthodes basées modèle et les méthodes sans modèle**. Une méthode basé modèle est donné dans [Lee L., Grimson W.E.L.] où sept ellipses sont utilisées pour représenter différentes parties de la silhouette d'une personne. L'hypothèse est faite que la personne est vue par la caméra perpendiculairement à sa démarche. La silhouette de la personne est segmentée par rapport au fond grâce à un algorithme de soustraction du fond adaptatif [Stauffer C., Grimson W.E.L.a]. Afin de rendre la représentation insensible aux changements de vêtements et à la distance entre la camera et le marcheur, la couleur de l'avant plan n'est pas utilisée, uniquement une silhouette binaire normalisée en échelle. Cette représentation est basée sur des primitives (les moments extraits de chaque silhouette de la marche). Afin d'obtenir de la robustesse au bruit et un modèle simple, plus fin que la silhouette grossière, la silhouette est découpée en 7 zones et dans chaque zone une ellipse est définie ainsi que des paramètres moyenne, orientation, soit un total de 28 paramètres calculés dans deux vues orthogonales. Ces paramètres constituent les moments extraits de chaque silhouette de la marche. Ces primitives sont ensuite utilisées pour reconnaître des individus par leur démarche apparente et pour prédire le genre de démarche inconnue. Cette approche fonctionne bien sous des conditions de lumière variées. Un exemple d'une méthode sans modèle serait avec des « blobs ».

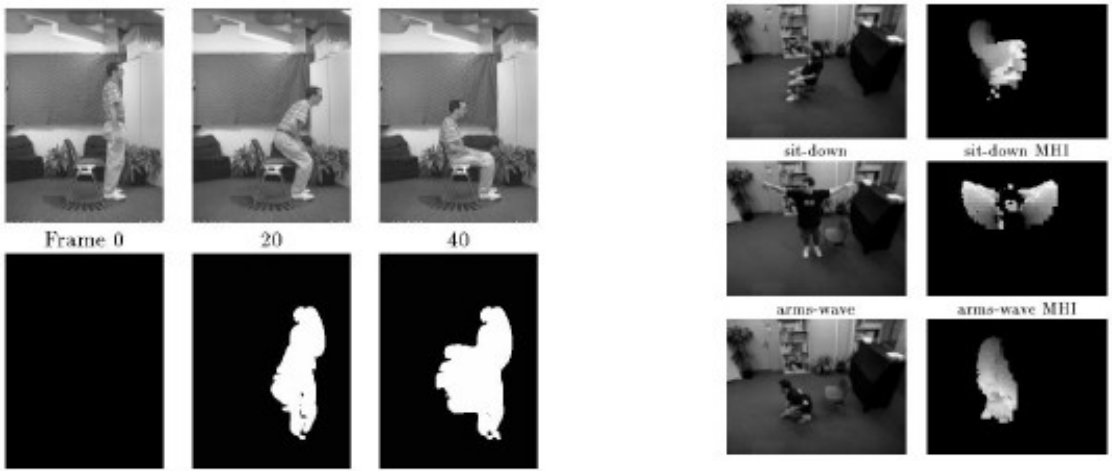
2.5.1 Moyenne résolution

A **moyenne résolution**, le but est de reconnaître des activités génériques comme le mouvement des bras et des jambes au lieu d'essayer de faire correspondre une action à un acteur particulier.
Un « template » ou gabarit a pour objectif la formation d'une image statique du mouvement étudié, afin d'en extraire des caractéristiques en vue d'une classification et d'une reconnaissance. Parmi les méthodes de « template matching », citons les travaux de ([Bobick A.F., Wilson A.D.], [Davis J.W., Bobick A.F.]) qui reconnaissent différents mouvements à l'aide de « template » spatio-temporels, une image binaire appelée « **Motion History Image »** (MHI) (cf. figure 52) et une image en niveaux de gris, « **Motion Energy Image »** (MEI) (cf. figure 52) .

Une carte de mouvement est extraite à chaque instant par suppression du fond. Le MHI mémorise chaque pixel détecté en mouvement pour une durée infinie, tandis que le MEI donne un poids plus important aux mouvements plus récents. [Chomat O., Crowley J.L.] génèrent des « templates » grâce à un banc de filtres spatio-temporels, mis en place via une analyse spatio-temporels en composantes principales (ACP). Mais les « templates » sont dépendants du point de vue.

Par exemple à **moyenne résolution**, les images « mouvement-histoire » (MHI) enregistrent la segmentation et l'information de mouvement temporel. Le MHI est une simple image composée d'une superposée d'images d'une séquence d'objets en mouvement segmentés et avec un poids évoluant au cours du temps (couverture spatiale et temporelle d'une activité). Les pixels de l'avant-plan les plus récents se voient assigner une couleur claire tandis que les pixels de l'avant-plan appartenant au passé sont progressivement assombris. Les **MHI n'utilisent pas de structure pour modéliser le corps humain**. Un vecteur de sept moments est calculé pour

chaque MHI. Les activités sont reconnues en trouvant la meilleure correspondance entre le vecteur de moments des MHI et les « patterns » entraînés. Les images contenant l'historique du mouvement peuvent être utilisées ou les CHMM pour suivre la tête et les deux mains. Ce sont des HMM couplés représentant les interactions entre personnes. Un modèle bâton 2D représentant le torse, les bras et les jambes, ou bien un modèle 3D, peuvent également être utilisés.



(a) MHI pour une personne s'asseyant (b) MEI pour des mouvements variés

Figure 52 : « Templates » spatio-temporels de ([Davis J.W., Bobick A.F.], [Bobick A.F., Wilson A.D.]).

2.5.2 Basse résolution

A **basse résolution**, l'objectif du suivi de personnes est de détecter la présence, et d'identifier les mouvements et les interactions de plusieurs personnes via le suivi de « blobs ». A basse résolution, des systèmes existent avec ou sans modèle du corps. Avec modèle, le corps peut être décrit par des ellipses. Ceux, sans modèle, ont des primitives et l'algorithme PCA permet de réduire la dimension de l'espace des primitives. La classification est faite basée sur les k plus proches voisins, ou avec un HMM. Le système **VSAM** suit le corps humain comme un « blob » entier, sans même savoir ce qu'il suit. Une soustraction du fond adaptative combinée à une différence entre trois images permet de détecter les objets en mouvement et le filtre de Kalman permet de suivre les objets en mouvement au cours du temps. Un réseau de neurones est entraîné pour reconnaître quatre classes: une personne seule, un groupe de personnes, des voitures, et la texture. Le système **VSAM** est très efficace dans le suivi de personnes et de voitures, et dans la discrimination entre différents types de voitures mais se limite à la reconnaissance de la marche en ce qui concerne la reconnaissance des activités. Le système **W^4** est dédié aux tâches de surveillance extérieure et pour les situations d'éclairage nocturne ou de faible lumière, sur des images monoculaires en niveau de gris. Un modèle de mouvement de second ordre incluant la vitesse et l'accélération est utilisé pour suivre le mouvement global du corps et le mouvement des divers membres. Ce système permet d'analyser et de reconnaître les activités humaines mais le modèle « cardboard » utilisé pour prédire la posture et la position du corps est restreint à la pose debout. [Paragios N., Deriche R.] ont une approche par contours actifs géodésique mais le suivi est limité à l'information de contour, le suivi de régions pourrait être ajouté pour augmenter la robustesse. Le temps de calcul de cette méthode est très long, une approche multi-échelle lui est préféré pour le temps réel.

2.5.3 Analyse des trajectoires de véhicules

Un système de vidéo surveillance pour le suivi de véhicule a été mis en place par l'**analyse des trajectoires de véhicules**. Différentes caméras sont déployées dans un parking afin d'augmenter la couverture. La trajectoire d'une voiture est issue des résultats de suivi de divers caméras. L'algorithme de reconnaissance prend la trajectoire en entrée et est composé de deux étapes : transformer les données numériques de la trajectoire en

une description **sémantique** comme « tourner », « s'arrêter », etc, et utiliser des **SVM** et **HMM** pour reconnaître les « patterns » de mouvement. [Niu W., Jiao L., Han D., Wang Y.-F.] reprennent ces travaux pour les appliquer au **suivi de personnes**.

2.5.4 Analyse des trajectoires de personnes

Quand les personnes sont éloignées, on peut les décrire par un « blob ». L'objectif est de les suivre même en présence d'occultations, de bruit, de courtes périodes d'absence et d'apparition des silhouettes, de longues périodes d'inactivités face à une des caméras. Pour la détection et le suivi, une différence d'image est effectuée suivie d'une corrélation. Les comportements ont été étudiés dans un parking, et les comportements suspicieux sont distingués des comportements normaux. La représentation de la personne par un « blob » permet de décrire la trajectoire du « blob » par sa vitesse et sa position. La formulation est basée sur un algorithme de condensation, utilisé dés lors que le bruit n'est pas gaussien et les états ne sont pas unimodaux, donc le filtre de Kalman est inadapté. C'est plus un estimateur bayésien qui permet de suivre plusieurs hypothèses qu'un estimateur du maximum de vraisemblance (cas Kalman). Chaque « blob » est représenté par un état comprenant la position, vitesse et accélération estimés de la région en mouvement. La position du « blob » est propagée au cours du temps en utilisant la vitesse et accélération enregistrés dans le vecteur d'état. La prédiction est validée par rapport à l'image observée (signatures au cours du temps de la couleur et texture). Un SVM avec une fonction à noyau gaussien est utilisée pour l'entraînement et la classification. Les taux de reconnaissance sont meilleurs qu'avec un HMM ou CHMM. De plus la complexité de l'approche statistique est plus faible que celle d'une approche structurelle qui en plus doit apprendre des modèles de Markov. C'est pourquoi **une approche statistique est plus apte à distinguer des comportements anormaux.**
Dans un premier temps les objets sont détectés automatiquement (extraits par soustraction avec un modèle de fond adaptatif). Le système de vidéo surveillance peut détecter des objets en mouvement et les classifier en catégories sémantiques comme les voitures et les personnes. Le filtre de Kalman est utilisé pour suivre chacune des personnes. Plusieurs personnes peuvent entrer et sortir de la scène. Si des silhouettes sont regroupées par erreur, le suivi est maintenu et corrigé une fois que le groupe de personnes se sépare. La détection d'objet débute par une soustraction du fond de façon adaptative [Collins R., et al.a] en faisant l'hypothèse d'un fond statique. Un filtre de second ordre de Kalman (états incluant position, vitesse et accélération) est utilisé pour modéliser le mouvement de chacune des personnes de la scène. Un SVM reconnaît les primitives de mouvement. Ce système peut suivre plusieurs personnes en temps réel. Des occultations temporaires sont acceptées si la même personne revient dans le champ de vue après une période de temps courte.

2.6 Suivi de trajectoires à l'aide d'une gestion haut niveau

De plus en plus de systèmes de vidéo surveillance disposent de caméras en réseaux [Regazzoni C.S, Sacchi C., Gera G.], permettant le temps réel du fait de la **distribution des traitements**. Chaque caméra donne lieu à un traitement local bas niveau et ce sont uniquement les informations haut niveau qui circulent d'une caméra à l'autre. Dans une stratégie de « **perception active** », les capteurs et les traitements sont activés à leur tour en fonction du contexte. La gestion des incertitudes liées aux décisions haut niveau est donnée par les notions de possibilité/nécessité. La possibilité d'une hypothèse correspond à un degré de compatibilité et la mesure de nécessité à une notion de certitude de cette hypothèse. Enfin, **l'information étant parfois incomplète** au moment d'une prise de décision, il est nécessaire de pouvoir **revenir sur une décision** en tenant compte d'une nouvelle information, c'est le rôle de la « **fusion temporelle** ». L'approche « perception active », associée à la gestion de l'incertain et à la « fusion temporelle » correspond à la « **gestion haut niveau** » du système.

2.6.1 Applications

L'application consiste à réaliser un système de surveillance pour la **Détection Automatique d'Incidents (DAI) en environnement autoroutier**. Les scénarios dangereux sont les arrêts sur la voie d'urgence, les accidents, les bouchons, les véhicules à contresens et les véhicules lents. Dans ce genre d'environnement, il faudrait un nombre beaucoup trop conséquent de caméras pour couvrir tout le territoire à surveiller. **Un système de suivi multi caméras permet de réduire le nombre de caméras tout en assurant la sécurité de la**

zone.
La première étape est la détection du mouvement, sujette à beaucoup d'imperfections. C'est la raison pour laquelle des descripteurs robustes et complémentaires ont été choisis, des modèles polyédriques. Les véhicules sont donc modélisés par un parallélépipède. Quelques travaux existent déjà [Koller D., Daniilidis K., Nagel H.-H]. Une distance de Mahalanobis est mesurée entre le modèle et l'objet détecté, intégrant les imprécisions de la mesure.
Une classification floue est élaborée afin de tenir compte des imprécisions des classes et aussi des mesures issues des observations. Les classes sont construites par un expert à partir des données contextuelles. Afin de tester le comportement des associations sur des situations complexes, un **simulateur de trafic** a été développé, il génère des données multi capteurs.

2.6.2 Détection de mouvement avec une image de référence

[Motamed C.] a travaillé sur la **détection du mouvement avec une seule caméra fixe.** Ce module a pour objectif la détection des zones en mouvement dans l'image. Un masque des objets mobiles est obtenu par détection de mouvement, afin de ne s'intéresser qu'aux zones d'intérêt de l'image. La méthode consiste à soustraire à l'image courante une image de référence sans objet en mouvement. L'**approche Markovienne** a été utilisée pour la détection de mouvement [Perez P.], considérée alors comme un problème statistique d'étiquetage en plusieurs classes de régions. Mais les temps de calcul sont trop importants pour la vidéo surveillance.
Cependant, la mise à jour de cette image de référence est une tâche difficile. Il s'agit d'une image de la scène récente sans aucun objet mobile. Une fois l'image de référence acquise à la première image et réactualisée, la détection de mouvement a lieu via une méthode de **gradient directionnel**, peu sensible aux variations de luminosité par rapport à la différence d'image simple.

2.6.3 Phase de mise en correspondance et gestion d'un système distribué de suivi

Un système de suivi de plusieurs objets simultanément doit maintenir les pistes des objets suivis, sans les perdre ni les confondre. C'est la phase de « **mise en correspondance** » ou « **d'association** ». En présence de plusieurs caméras, la mise en correspondance se situe non seulement au niveau des objets mais également entre chacune des vues issues de chaque caméra pour un objet donné.
Les objets sont suivis à partir de la détection de leur mouvement (masque de détection de l'objet). Le système de suivi doit pouvoir gérer les occultations d'un objet par un autre, les **créations de pistes des nouveaux objets**, le **maintien des pistes des objets suivis**, et la **terminaison de pistes** lorsque l'objet quitte la scène. [Motamed C.] propose une stratégie de « **gestion haut niveau** » actif, dans le sens où il s'adapte en fonction des situations présentes dans la scène. C'est un **système d'interprétation « haut niveau »** car un certain nombre de situations sont reconnues par celui-ci et intégrées de façon **dynamique**. Cette « gestion haut niveau » est innovante par rapport à des systèmes existants [Kettnaker V., Zabih R.].

L'approche proposée par [Motamed C.] est de type **coopérative et qualitative**. Ces techniques sont issues de l'**intelligence artificielle** et usitées en **vision par ordinateur** [Kholer Ch., Ottlik A., Nagel H.-H, Nebel B.]. [Kholer Ch., Ottlik A., Nagel H.-H, Nebel B.] interprète des scènes de trafic urbain avec l'équipe de Nagel. Un raisonnement haut niveau et des informations qualitatives spatiales et temporelles permettent d'éviter les occultations entre véhicules et donc certaines situations incertaines, augmentant ainsi la robustesse du suivi de véhicules. L'approche de [Motamed C.] est similaire à celle de Nagel mais le **modèle de comportement des objets de la scène** est plus générique. La modélisation du comportement des objets de la scène a pour but de guider le suivi, lui même ayant pour objectif l'obtention des pistes des objets mobiles.
La « gestion haut niveau » du suivi est centralisée par un superviseur, qui a accès à une base de données « BDO » (Base de Données Objet) et aux résultats fournis par les modules bas niveaux spécialistes. La « BDO » contient les pistes des objets suivis ou en cours de suivi, et un modèle visuel de l'objet, comprenant les dimensions de l'image et l'histogramme de la couleur de l'objet. L'objectif central du système de suivi est le maintien des pistes, en prenant en compte les occultations. Deux types de traitement sont appliqués, selon qu'il

s'agisse de suivre des objets isolés ou un groupe d'objets (les objets dont les régions ont été fusionnées forment un groupe). Le superviseur active, selon les cas, un des modules de maintien des pistes (par objet ou par groupe) pour chaque objet suivi et gère les objets et les groupes par deux autres modules à l'origine de la création, maintien et suppression de pistes des objets ou des groupes. Deux indicateurs, consistance et 'identité attribuent une qualité au suivi. Le module de maintien des pistes des objets isolés utilise une approche par « plus proche voisin NN » (Nearest Neighbor). La ressemblance est estimée par la distance entre descripteurs visuels utilisés (ceux de l'objet et ceux d'une observation), dans ce cas l'histogramme de couleur de l'objet. Le module de maintien des pistes en présence de groupes estime la position de chaque objet au sein d'un groupe, grâce à l'algorithme Mean-Shift, approche statistique non paramétrique pour la recherche de régions candidates à partir de l'histogramme de couleur, utilisé dans le suivi de régions en temps réel. La distance entre les histogrammes colorés des régions candidates et le modèle de l'objet est donnée par la distance de Bhattacharya. Mais l'algorithme Mean-Shift est basé sur l'apparence et ne donne pas de bons résultats dans le cas des occultations. Un autre traitement, basé sur la notion de groupe, rassemble des objets isolés en présence d'occultation, ceci par une approche haut niveau, limitant ainsi les inconvénients des approches apparence.
Cette approche a été validée pour la reconnaissance de comportements de piétons à partir de leurs trajectoires. Pour chaque comportement, les zones de la scène et les durées des enchaînements entre les zones sont modélisées par des variables floues. La gestion de l'incertitude de la reconnaissance a utilisé la théorie des possibilités, pour gérer les imprécisions spatio temporelles des trajectoires observées et choisir le comportement le plus plausible à partir du couple Nécessité/Possibilité.

Une extension multi-caméras a vu le jour dans le cadre du **projet régional « Gymnase Intelligent »** avec Sportica 2000 (1999-2002) à Gravelines et le département Sciences et Techniques des Activités Physiques et Sportiuves (STAPS) de l'Université du Littoral, pour l'analyse de scènes de **basket-ball**. Différentes caméras ayant des points de vue complémentaires couvrent tout le terrain de jeu et l'objectif de ce système d'interprétation est d'analyser les déplacements des joueurs. **L'objectif du suivi multi caméras est de lever les ambiguïtés lors des phases d'occultation, celles-ci n'apparaissant pas simultanément sur les différentes caméras. Chacune d'elles est sélectionnée de façon active.**
L'interprétation multi caméras possède deux niveau de suivi hiérarchiques, un suivi bas niveau local associé à chaque caméra, et un suivi haut niveau global qui regroupe les résultats issus des différentes caméras au niveau local.
[Motamed C.] a donc proposé une architecture globale pour le **suivi d'objets à partir de leurs régions (apparence)** pour la vidéo surveillance. La gestion « haut niveau » permet de dépasser les ambiguïtés telles que les occultations, gérées par la notion de groupe. La configuration multi caméras a pour objectif de sélectionner de façon active une des caméras en fonction des situations d'occultation. C'est le système d'interprétation « haut niveau » qui a en charge le choix de la caméra à chaque instant.

Ce système est constitué de caméras éloignées, leurs champs d'observation ne se recouvrent pas et ne couvrent pas obligatoirement la scène en entier. L'objectif d'un tel système est la **ré identification** des objets entre les caméras. Ce problème est courant dans tous les sites de surveillance, gares, aéroports, etc. Ces applications nécessitant le temps réel, une architecture de type « Vision Coopératuve Distribuée » [Motamed C., Wallart O.] composé d'un groupe de capteurs intelligents et communicants.
La ré identification est un problème de mise en correspondance entre une observation et un objet préalablement observé. Un superviseur central récupère les décisions locales de suivi afin de coordonner les caméras dans une surveillance globale.
La mise en correspondance est effectuée non pas de façon binaire mais floue. Les associations observation/objet attendu sont parfois ambiguës. Cette ambiguïté est traduite par un degré de confiance, issu de la théorie des possibilités (le degré de possibilité d'une hypothèse correspond à un degré de confiance). Une distribution de possibilité est partagée entre tous les experts pour chaque observation et pour la caméra pointant l'observation. Chaque expert donne son avis sous forme de nécessité d'une association (observation/objet) candidate (une hypothèse) après avoir intégré les avis des autres experts. Si le degré de nécessité est important, l'expert valide l'association, sinon il y a ambiguïté. Les experts haut niveau mettent en place une **fusion temporelle** dans ce cas. Les **arbres de type MHT** (Multiple Hypothesis Testing) issus de la

poursuite radar [D.B. Reid.] sont utiles dans le cas d'un trop grand nombre de données à fusionner. Les hypothèses du MHT forment une configuration d'association entre les objets, et le MHT propage récursivement les hypothèses (les branches de l'arbre). Le MHT est ainsi développé dynamiquement en fonction des observations. La fusion temporelle met en exergue les hypothèses les plus fortes : à chaque nouvelle observation, la qualité des hypothèses est estimée par les degrés de nécessité de chaque hypothèse de l'arbre. La mesure de nécessité représente l'écart relatif entre la possibilité d'une hypothèse vis-à-vis des autres hypothèses concurrentes générées par l'arbre.
La plupart des travaux sur le suivi multi caméras en vision par ordinateur présentent des configurations avec des capteurs ayant des champs en communs [Chang T.H., Gong S., Ong E.J.] et l'objectif est de ramener toutes les informations dans un même repère et de les fusionner, grâce à la redondance et la complémentarité des informations.
[Kettnaker V., Zabih R.] a utilisé des caméras distantes, [Kettnaker V., Zabih R.] pour la surveillance des bâtiments, à travers une approche bayésienne. L'approche de [Motamed C.] est aussi une **stratégie basée sur l'approche bayésienne, intégrant les connaissances *a priori* liées à l'apparence des objets et à leur comportement dynamique**, aidant aux décisions de reconnaissance. **La présence de zones aveugles importantes fournit des informations partielles et complémentaires, mais pas temporellement redondantes sur un même sujet.** Dans les scènes de transport routier, le suivi est une tâche difficile du fait de nombreuses contraintes, des informations incomplètes et incertaines. L'architecture dispose alors d'un **raisonnement distribué** et **temporel**, et gérant **l'incertitude** des décisions. Diverses informations **contextuelles** sont exploitées, comme celle de la configuration de la scène avec la disposition des caméras et les zones aveugles, les informations de classes des objets mobiles et dynamiques (comportements cinématiques). Lors de la mise en correspondance, une gestion intelligente a été mise en place, permettant au système une focalisation géographique et temporelle. La **focalisation géographique** indique qu'un objet ne peut emprunter qu'un nombre restreint de chemins possibles et donc seules certaines caméras sont sélectionnées, celles correspondants aux lieux des objets attendus possibles. La **focalisation temporelle** indique une fenêtre temporelle devant une caméra donnée, ainsi le nombre d'objets prévus à chaque instant pour apparaître devant une caméra est géré de façon dynamique. L'architecture choisie est de type **multi agents basée sur les « sociétés d'experts »** [Matsuyama T.]. La communication entre experts est réalisée par envoi de messages. La tâche de suivi d'objets par leur ré identification est réalisée de façon coopérative et distribuée. Pour chaque objet, un ensemble de chemins possibles est envisagé et le système vérifie les hypothèses. Il peut alors à ce moment là décider d'associer les objets.
La **création de nouveaux objets** est prévue afin d'initialiser les pistes. Les nouveaux objets sont ceux trop différents des objets attendus. La **terminaison de pistes** correspond aux objets attendus mais non détectés et à ceux qui terminent vraiment leur piste. La **phase de maintien des pistes** prend avantage de la **fusion temporelle** dont le rôle est de combiner des informations au cours du temps et d'améliorer ou de rendre une décision. Cette approche est initiée par le contexte dynamique avec des informations évolutives au cours du temps de façon incrémentale. Ainsi les informations peu crédibles perdent en véracité en attendant des informations complémentaires.

2.7 Suivi de piétons dans un réseau routier

Un piéton est plus facile à modéliser qu'un humain, car seuls certains mouvements sont possibles dans la rue: en position debout, un piéton marche ou est stationnaire. Mais beaucoup de cas d'occultations sont à envisager, ainsi que les variations d'illuminations, et la présence des ombres.
[Papageorgiou C., Oren M., Poggio T.] détectent des piétons avec une méthode à base de **SVM** (Support Vector Machines). [Chen H.T., Lin H.H., Liu T.L.] font du suivi par **mise en correspondance dynamique de graphes**. L'avant-plan est extrait du fond, et les pixels regroupés en « blobs ». Chaque agent a un profil enregistré en mémoire, et remis à jour de façon dynamique. Les objets détectés dans chacune des images sont mis en correspondance avec les profils enregistrés. S'il n'y a pas de correspondant, un nouveau profil est crée. Chaque profil est doté d'un âge. Si un profil n'a pas de correspondant, son âge augmente. Le profil est supprimé s'il devient trop vieux. En revanche, son âge est réinitialisé s'il est mis en correspondance avec un objet détecté. Les relations entre les profils et les objets sont modélisées par un graphe complet en deux parties (bipartite) avec d'un côté, les profils, et de l'autre, les objets détectés. Le coût du graphe est la somme des

fonctions de similarité entre profils et objets, appliquées aux paires de profils et correspondants. [Viola P., Jones M., Snow D.] détectent des objets par extraction de primitives dans une fenêtre glissante et comparaison avec des primitives référence. Ils détectent par la suite des piétons dans une séquence vidéo avec une cascade de **classifieurs**. [Heisele B., Wöhler C.] utilisent la **périodicité de la marche** pour détecter des piétons, dans une séquence d'images acquise avec une caméra en mouvement. Dans chaque image, un nombre défini de clusters coloré segmente l'image. Les clusters sont mis en correspondance d'une image à la suivante, avec un algorithme « k-means ». La taille du cluster varie périodiquement pendant la marche, mais le pied du piéton appartient au même cluster. L'analyse de la périodicité des variations de taille des clusters permet d'identifier les clusters contenant le pied du piéton. Un **réseau de neurones** TDNN (« Time Delay Neural Network ») est utilisé pour augmenter les performances en reconnaissance.

2.7.1 Comportements multi agents

Pour décrire des interactions entre plusieurs personnes, deux types de méthodes ont été examinées [Pop I.]. **La première est basée sur les modèles de Markov Cachés (HMM), et la seconde sur les réseaux bayésiens (propagation de croyance)**.

-L'avantage du HMM est la présence de la notion temporelle, mais son inconvénient majeur est que les valeurs sont numériques et le vecteur d'observation de taille fixe;
-Tandis que **le réseau de croyance intègre des notions conceptuelles**, mais pas le temps.

[Oliver N.M., Rosario B., Pentland A.P.] présentent un modèle dérivé des HMM pour l'analyse les interactions entre deux agents. Les **HMM chaînés** (« Chained HMM », « CHMM ») sont composés de deux chaînes de HMM, avec leurs observations et leurs états. Un CHMM est différent d'un HMM classique : les probabilités de transition sont estimées différemment. Dans un CHMM, l'état futur d'une chaîne de CHMM dépend non seulement de l'état courant de la chaîne, mais aussi de l'état courant des autres chaînes. Les primitives utilisées sont la distance entre agents ainsi que leur direction relative, leur vitesse, leur orientation, et leur position.

[Intille S.S., Bobick A.F., 01] analysent les **interactions entre les joueurs (les « agents ») pendant un match de football Américain et identifient les scénarios d'attaque et de défense**. Les informations sur les agents de position les uns par rapport aux autres, ou par rapport à un lieu, sont analysées, grâce à leur trajectoire et leur position. Les informations sont soumises à un réseau bayésien afin d'estimer le rôle d'un agent, « frapper une balle », « course entre deux joueurs », etc. Un **graphe** sur la séquence entière montre l'évolution de la probabilité des objectifs de chacun des agents. Les graphes sont comparés par analyse temporelle, et une relation temporelle est définie entre les objectifs des agents pour chacun d'eux.
Par la suite, un expert met en place des scénarios dans un réseau bayésien multi agents. Dans ce réseau, toutes les informations concernant les objectifs des agents sont prises en compte, ainsi que leurs relations temporelles, permettant ainsi d'estimer une probabilité pour chaque scénario possible.

2.7.2 Description du scénario

Il s'agit d'une intersection entre une route à 4 voies et une route à deux voies à l'université de Karlsruhe (cf. figure 53). Beaucoup de piétons traversent ces routes. Il y a un pont qui traverse la quatre voie où des piétons peuvent traverser. Les images capturées du pont ont assez de détails pour identifier et suivre des piétons. La scène est filmée par plusieurs caméras calibrées (paramètres intrinsèques et extrinsèques). Les enregistrements sont ainsi corrigés, et **un modèle de la scène est nécessaire pour prédire des éventuelles occultations des agents (voiture et personne) par des éléments statiques de la scène**.

Le comportement des véhicules est simple comparé au comportement humain. Une voiture peut aller tout droit ou tourner, il s'agira donc de déterminer la route sur laquelle elle se trouve ainsi que le moment auquel elle tourne. Le comportement des piétons est plus complexe car les piétons peuvent changer de direction à tout moment. Pour traverser une route, il faut déterminer les localisations de début et de fin de parcours.
Si dans un premier temps, le comportement du piéton est analysé comme un agent isolé, dans un deuxième

temps, les interactions entre les agents sont analysées sous la forme de relations cause/effet. Des « patterns » d'interaction sont construits et les actions des différents agents sont comparés à ces « patterns », par exemple un piéton attendant qu'une voiture passe. Les interactions entre les véhicules, entre véhicules et piétons et entre piétons sont détectés.

Les **applications** sont diverses, citons l'automatisation des feux tricolores, la surveillance des parkings générant un texte décrivant le comportement des piétons et des véhicules, plus concis que l'enregistrement de la séquence entière.

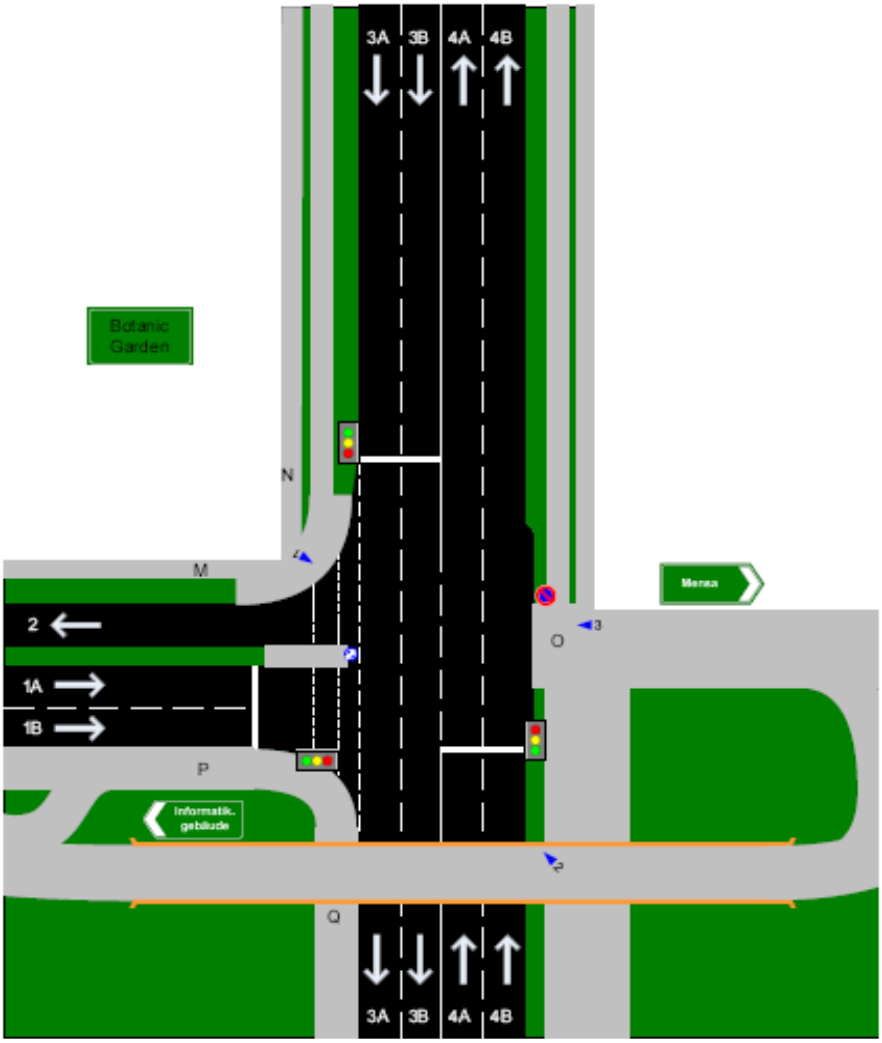


Figure 53 : Un modèle simple de l'intersection étudiée. Les flêches bleues indiquent la position des caméras. Les flêches blanches sur le bitume témoignent du sens de circulation des véhicules [Pop I.].

Les « SGT » [Arens M., Nagel H.-H.] sont des structures de graphes modélisant les comportements des agents. Son unité de base est un schéma de situation représentant l'état et l'action d'un agent à un instant donné. Les situations sont regroupées en graphe. Les situations sont connectées par des liens de prédiction.
Dans un **modèle conduit par la localisation**, on considère toutes les localisations possibles d'un piéton et on recherche sa trajectoire. Tous les chemins possibles sont structurés dans un SGT.
Dans un **modèle conduit par le comportement**, les comportements des piétons sont considérés. La différence avec le modèle conduit par la localisation réside dans l'absence de la localisation dans les schémas de situations.
Une liste de prédicats et de règles d'inférence [Gerber R., Nagel H.-H.] interprètent la scène.

L'architecture adoptée fournit de bons résultats pour décrire le comportements des véhicules. Le propos de cette étude [Pop I.] est d'adapter les programmes pour la **reconnaissance et le suivi de piétons** et pour générer du texte décrivant le comportement des piétons et leurs interactions avec les véhicules.

2.7.3 Model-Based Tracking in Image Sequences Motris

Motris (« Model-Based Tracking in Image Sequences ») constitue une des applications utilisées pour obtenir en langage naturel la description des comportements des piétons et de leurs interactions. Il s'agit d'un réseau pour le suivi 2D/3D. Motris estime la position des acteurs dans la scène et fournit une liste de prédicats exprimant la localisation de chacun des agents dans chaque image. Motris connaît deux types de suivi, en 2D ou en 3D. En 2D, aucune information dans la scène n'est utilisée et le suivi a lieu par l'analyse de l'image, sans connaissance explicite de l'environnement 3D. Les objets suivis sont modélisés par des ellipses. En 3D, le suivi nécessite les informations de calibration de la caméra et un modèle 3D de l'agent suivi.

2.7.4 Suivi des voitures et des piétons

Le but du prototype est de configurer le système pour suivre des voitures et des piétons et pour décrire leur comportement. Les interactions ne sont pas prises en compte dans un premier temps.
Pour les véhicules, leur suivi est basé sur un modèle de l'objet. Les problèmes d'occultations ne sont pas résolus car le système n'a aucune connaissance *a priori* de la scène. Si un agent se trouve en occultation, le suivi a lieu grâce à la prédiction, sinon le suivi s'effectue par le calcul du flot optique. Les problèmes d'occultations sont résolus avec un modèle 3D des véhicules qui s'adapte au véhicule grâce à divers images de différentes positions des agents (véhicule dans ce cas).

Pour les piétons, deux méthodes similaires de suivi en 2D ont été testées. Dans l'algorithme original, un agent est initialisé avec des composantes connectées. La position et la taille des agents est mise en correspondance de façon itérative avec les composantes connectées associées. Dans la seconde méthode, qui est une version améliorée de la première, il n'y a pas de suivi par composantes connectées. C'est la taille et la position des agents qui sont mises en correspondance avec un ensemble de pixels d'avant-plan via l'algorithme EM. Dans cette nouvelle approche, le suivi a bien lieu y compris en cas d'occultation partielle d'un agent.
Bien que les résultats en suivi 3D sont meilleurs que ceux du suivi 2D, il subsiste encore des problèmes, comme la détermination des objets en mouvements propre à des piétons. L'algorithme d'analyse des irrégularités de la trajectoire est utilisé dans ce but. Pendant le suivi, la taille et la position de l'ellipse est adaptée à la forme de l'agent suivi, du fait du changement de forme du piéton au cours de la marche. La variation de la taille du « blob » entourant le piéton est fonction de la fréquence de la marche et détermine les variations dans la trajectoire des piétons. La fréquence des variations est utilisée comme mécanisme de classification des piétons.

2.7.5 Lien entre la localisation et les actions des piétons

Dans le cas du croisement de routes étudié ici, il existe une relation forte entre la localisation et les actions des piétons. Le modèle conduit par la localisation et le modèle conduit par le comportement sont équivalent. Le modèle conduit par la localisation analyse les informations des piétons par les schémas de situations spécialisés, tandis que le modèle conduit par le comportement utilise des prédicats. Les coordonnées sont ensuite transformées du 2D vers le 3D grâce aux informations de calibration de la caméra relative à la scène. La vitesse et la direction de l'agent sont calculés via la vitesse de l'ellipse dans l'image. Afin que le système puisse reconnaître un piéton, un modèle 3D est associé, ce modèle est similaire à celui des véhicules, mais les dimensions sont ajustées de façon à ce qu'il corresponde à la taille des piétons. Ce système est génératif à tout type d'objet (vélos, etc.).

2.7.6 Modélisation des intéractions entre les agents

Pour modéliser les interactions entres les agents dans une scène, **trois approches** sont proposées, soit en étendant la sémantique du SGT, soit en ajoutant un module d'interaction basé sur le ODHMM, soit en utilisant

les résultats du SGT pour alimenter un réseau bayésien.

-Dans la **première approche**, il n'y a plus d'agent actifs, toutes les informations sont valables à l'instant courant au sujet de tous les agents, et elles sont toutes utilisées pour évaluer la prochaine situation;
-Dans **l'approche par ODHMM**, les interactions entre agents sont analysées par ODHMM, une extension es HMM. Mais les ODHMM n'utilisent pas des informations conceptuelles dérivées des SGT, ils nécessitent des informations numériques comme les HMM. Il est donc difficile de détecter des comportements de haut niveau, surtout si une connaissance de la scène est requise. L'ODHMM proposé aurait trois noeuds, l'un pour deux personnes s'approchant l'une de l'autre, l'autre pour deux personnes parlant, et le troisième pour les deux personnes marchant ensemble. Les primitives sont basées sur la distance entre les deux agents et leur orientation relative;
-La **troisième approche** proposée ressemble aux travaux de [Intille S.S., Bobick A.F., 01]. La différence essentielle est que [Intille S.S., Bobick A.F., 01] utilise des SGT pour détecter la probabilité des buts « simples » des agents. L'information au sujet des buts est rétro propagée vers un **réseau bayésien multi agent**, qui va estimer la probabilité de chaque but « interaction ». Ce réseau bayésien est généré dynamiquement, basé sur l'information des buts des agents et leur rôle dans le temps. **Un scénario est ainsi décrit comme un piéton qui attend qu'une voiture ne passe pour traverser.**

2.7.7 Modèle proposé

C'est la première architecture qui est retenue, c'est une structure multi agent transversal. Il existe aussi une version centralisée, dans laquelle toutes les informations sont stockées dans une base de connaissance.

Le modèle de comportement entre deux piétons peut être plus précis si on dispose de l'information de direction du regard. Le principal avantage est qu'il est possible de différencier entre les piétons qui marchent ensemble par hasard et ceux qui échangent des informations et marchent ensemble intentionnellement. Le modèle d'interaction entre véhicules est plus complexe, l'information étant contenue uniquement dans leurs trajectoires. Le suivi d'un groupe de personnes est plus délicat, un groupe étant considéré par le système comme un agent unique. Une façon de résoudre ce problème serait de construire un modèle 3D des piétons et d'utiliser le suiveur 3D, bien que cette solution ralentirait considérablement le suivi. Un « détecteur d'acteur vivant » permettrait d'oublier les acteurs morts, ne se préoccupant que des acteurs dans les zones d'entrée et d'accès de la scène. Finalement, le suivi dans Motris a lieu en 2D, fournissant de bons résultats dans le cas de non occultations par d'autres agents en mouvements.

Deux architectures ont été définies, une centralisée et une distribuée, mais aucune des ces deux architectures ne résout les cas d'interactions nombreuses. C'est le propos du **SGT** qui doit en limiter le nombre, par le choix des conditions d'instantiations pour chaque SGT.
Le premier modèle multi agent SGT a modélisé les interactions entre les véhicules et les piétons au croisement étudié [Pop I.]. Les résultats de l'analyse de comportements sont converties en texte, qui représentent une adaptation plus conviviale pour délivrer les résultats. Bien que le texte généré contienne des informations au sujet des interactions entre les agents, il ne s'agit pas d'une réelle description des interactions.

[Pop I.] propose d'ajouter une méthode de classification des piétons, de façon individuelle, même si le piéton appartient à un groupe. Une composante basée comportement doit être rajoutée au suiveur, ainsi qu'un générateur de texte décrivant les interactions entre les agents. Par la suite une architecture distribuée permettrait d'augmenter la vitesse des traitements. Un exemple de résultat est donné dans la figure 54.

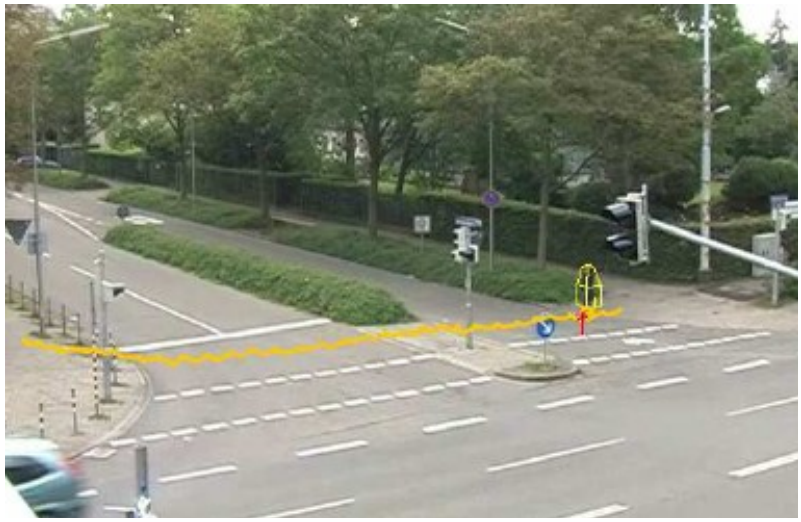


Figure 54 : La trajectoire d'un piéton traversant l'intersection [Pop I.] .

2.8 Le suivi des trajectoires des tâches de couleur

La méthode de [Megret R.] est basée sur le **suivi des trajectoires des tâches de couleur**, extraites sous la forme de « blobs » gaussiens et laplaciens. Les pixels à l'intérieur d'une tâche de couleur sont regroupés du fait qu'ils présentent des **caractéristiques colorimétriques, texturales et spatiales similaires**.

PFINDER [Wren C.R., et al.] présente une approche paramétrique puisque les tâches sont modélisées dans l'espace des caractéristiques par des densités gaussiennes paramétrisées par leur centre et leurs matrices de covariances à déterminer.
[Comaniciu D., Meer P.] présentent une approche non paramétrique de la distribution spatiale et colorimétrique des tâches de couleur, via l'estimation empirique des modes des distributions (cf. figure 55). Chacun des modes correspond à une classe de points regroupant un ensemble de pixels proches spatialement et spectralement. La distribution spatiale et spectrale est obtenue par l'estimateur de la fenêtre de Parzen. Cette méthode de classification, par recherche de modes basée sur le « **Mean-Shift** », permet d'associer chaque point à un mode sans estimer explicitement la fonction de densité.

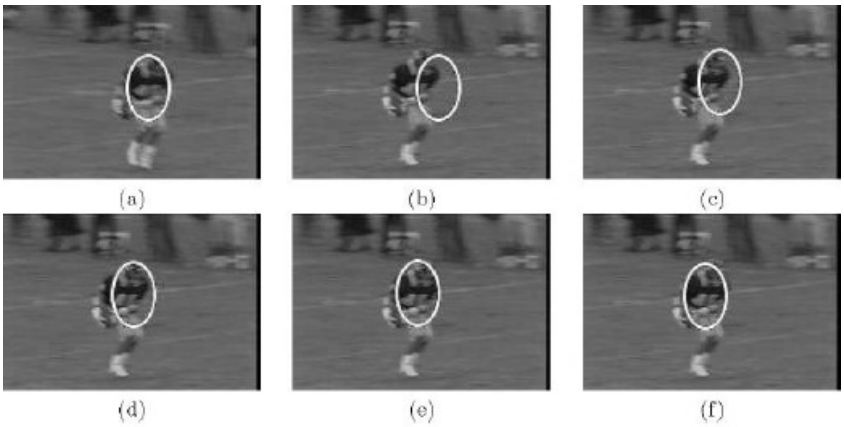


Figure 55 : Résultats de suivi par l'algorithme de Mean-Shift [Comaniciu D., Meer P.] .

Une **analyse multi-échelle** est utile pour extraire des structures présentes à divers échelles sous une forme hiérarchique, et les liens entre les échelles.

Le **suivi temporel cherche à effectuer une mise en correspondance entre les primitives** spatiales, points, segments et régions. En ce qui concerne les points, l'estimation de trajectoire associe une trajectoire à chaque point. [Allmen M., Dyer C. R.] calculent des courbes de flot spatio-temporel (« spatio-temporal flow curves ») par intégration du flot de mouvement local dans le temps. Un champ de mouvement local est estimé pour chaque paire d'images et chaque point est associé à une trajectoire. Le suivi des « points d'intérêt » uniquement est une méthode plus couramment usitée, permettant de s'abstenir d'accumuler des erreurs sur tous les points de l'image. Les points d'intérêts sont détectés à chaque image et mis en correspondance de façon temporelle. Pour limiter les correspondance, l'hypothèse est faite d'une invariance temporelle des caractéristiques et d'une continuité temporelle du mouvement. Dans le cas du suivi de personnes, il n'y a pas d'objet d'intérêt, il faut considérer l'ensemble des primitives de l'image par l'invariance des caractéristiques visuelles et la régularité du mouvement. Les points de Harris [Harris C., Stephens M.] sont des points d'intérêt qui ont été utilisés au sein du laboratoire LIRIS INSA Lyon, mais ils ont montré leur faiblesse au niveau de la non homogénéité de la répartition des points puisque la majeure partie des points se trouvent dans les zones de forte variance. Les « blobs » de leur côté, présentent une distribution plus régulière et détectent des zones contrastées comme les yeux.

Pour suivre les « blobs », [Megret R.] a utilisé la méthode de suivi des points d'intérêt multi hypothèses présentée dans [Cox I.J, Hingorani S.L.]. Des **arbres d'hypothèses d'appariement** sont construit entre les images qui se suivent. Les hypothèses qui présentent un conflit sont éliminées, permettant ainsi d'élaguer l'arbre. Les **initialisations** et **terminaisons de trajectoires, et les disparitions accidentelles temporaires de primitives sont traitées dans cet arbre d'hypothèses**. La trajectoire est considérée comme terminée lorsque les primitives ont disparu suffisamment longtemps, et considérée comme nouvelle si les primitives réapparaissent, évitant ainsi des erreurs d'appariement entre des primitives issues d'objets différents.
Les trajectoires doivent être regroupées afin de décrire le mouvement d'ensemble d'un objet (cf. figure 56).

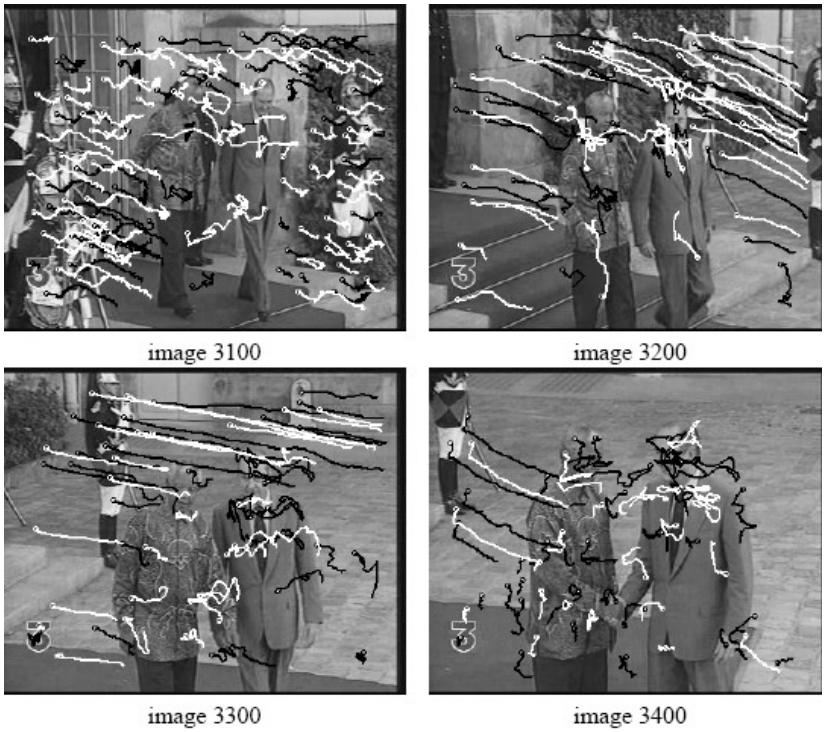


Figure 56 : Suivi sur une longue séquence « Mandela » et représentation des trajectoires [Megret R.].

Le regroupement de trajectoires a pour objectif soit de caractériser une trajectoire, soit de segmenter le mouvement. En **caractérisation de mouvement**, les trajectoires ainsi obtenues correspondent à des types de

mouvement, caractéristique de situations ou comportements. En **segmentation par le mouvement**, il s'agit de regrouper un ensemble de trajectoires de mouvement similaire, afin de segmenter un objet en mouvement par rapport au reste de la scène fixe. Dans la seconde catégorie, segmentation par le mouvement, [Megret R.] s'intéresse à la segmentation par classification.

2.8.1 Segmentation du bloc spatio-temporel

L'ensemble des pixels d'une séquence, constituant la base de l'analyse de celle-ci, est regroupé en un bloc spatio-temporel. Les approches de segmentation du bloc se scindent en deux catégories : celles *a priori*té spatiale réalisent la segmentation dans l'image et recherchent par la suite la cohérence temporelle, tandis que celles du domaine à la fois spatiale et temporel intègrent les liens temporels en même temps que la segmentation. Dans les approches de segmentation spatiale, les liens temporels entre images sont estimés via une segmentation spatiale existante. Il existe deux familles de segmentation spatiale : par le mouvement ou par les caractéristiques statistiques telles que la couleur ou la texture. Parmi les techniques de segmentation spatiale par le mouvement, coexistent celles basées sur la similarité de mouvement, et les autres sur l'estimation de modèles. Les premières, par la similarité de mouvement, font appel au mouvement estimé localement et aux caractéristiques de mouvement associées à chaque pixel ou région. Par estimation de modèle, les paramètres du mouvement sont estimés sur des groupes d'éléments. Ces deux méthodes sont basées sur un modèle de mouvement, implicite pour la première (critère de régularité spatiale du mouvement), et explicite pour la seconde, et paramétrique.
[Gelgon M.] calcule les paramètres de mouvement sur chaque région individuelle. Pour chaque paire de région voisine, la différence moyenne est évaluée entre les prédictions du champ de mouvement issues des paramètres de mouvement respectifs. Ces différences sont utiles pour la segmentation probabiliste par champ de Markov, car elles avantagent ou au contraire désavantagent l'étiquetage identique des régions voisines.

2.8.2 Cohérence temporelle

La segmentation d'une image, au sens du mouvement, de la couleur ou de la texture est spatiale. Elle doit se doter de liens temporels entre les images pour être une structure spatio-temporelle. Ces liens sont la cohérence temporelle des segmentations successives.

2.8.3 Mise en correspondance

Chaque région est reliée à la région de l'image suivante de meilleure similarité.

2.8.4 Hiérarchies de segmentation

Une fois les régions segmentées et mise en correspondance par des liens 1-1, association de chacune des régions à maximum une région correspondante temporellement, [Gomila C.] utilise des hiérarchies de segmentation couleur.

2.8.5 Extension de l'horizon temporel

La cohérence temporelle (par exemple l'invariance de l'apparence) peut être plus contrainte avec un horizon temporel plus large.

2.8.6 Segmentation dans le domaine joint spatio-temporel

Les approches pour l'extraction de structures spatio-temporelles de la vidéo, sont constitueés de deux sortes : par similarité ou spatio-temporelle. Les **méthodes par similarité** cherchent des classes cohérentes dans le bloc spatio-temporel, tandis que les **méthodes spatio-temporelles** ont un modèle global pour tout le bloc.

2.8.7 Segmentation de graphes

Les méthodes à base de graphe détectent des similarités entre les pixels du bloc spatio-temporel. Chaque noeud

est associé à un pixel du bloc, chaque arête est pondérée par la similarité entre les noeuds. Les arêtes connectent les pixels spatialement et temporellement, d'où la dénomination de segmentation spatiale et temporelle jointe.

Les méthodes par graphe sont basés sur les relations binaires entre noeuds, localement, tandis que la méthode par modélisation paramétrique traite le problème globalement.

2.8.8 Modélisation paramétrique du bloc vidéo

On étend sur une période temporelle, la représentation d'une image par un mélange de gaussiennes. La modélisation de la scène étendue au bloc vidéo, associe une classe par région de couleur homogène. Elle se déplace dans le temps à vitesse constante. Nous obtenons une segmentation spatio temporelle avec un modèle de gaussiennes. Si la tâche de couleur se déplace devant la face, nous voyons apparaître un cylindre généralisé, dont la forme (la génératrice) parcourt le temps en suivant un axe spatio-temporel (la directrice).
Un **tube de couleur spatio-temporel** est défini par un ensemble de pixels de couleur voisine, et situés dans le bloc vidéo autour d'une directrice droite (cf. figure 57). Il est semblable à la tâche de couleur, ensemble de pixels de couleur similaire au voisinage spatial d'un point central, mais avec une translation à vitesse constante, régissant un déplacement de la tâche au cours du temps.

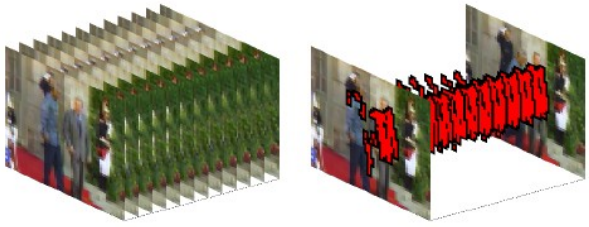


Figure 57 : Exemple de tube détecté dans un bloc vidéo. A gauche, un bloc vidéo vu sous la forme d'un empilement d'images. A droite, mise en évidence d'un tube particulier dans ce bloc [Megret R.].

2.8.9 Classification

Le vecteur de caractéristiques est associé à un vecteur de sept caractéristiques : trois pour la couleur, deux pour la direction et deux de position spatio-temporelle. La classification dans l'espace des caractéristiques est non paramétrique et hiérarchique. Lorsque deux centres de deux « clusters » différents sont très proches, ils sont regroupés hiérarchiquement : un nouveau noeud est crée, dont les fils sont les noeuds associés aux centres proches. Cette méthode s'applique à l'ensemble à l'ensemble des pixels du bloc vidéo. Les regroupements hiérarchiques sont spatio-temporels.

2.8.10 Comparaison avec les autres méthodes

2.8.10.1 Segmentation de graphes

Par rapport à la segmentation de graphes, dans [Megret R.] les pixels lointains temporellement peuvent être regroupés dans le même tube spatio-temporel de pixels, pour peu qu'ils aient des caractéristiques communes.
Dans la segmentation de graphes, des pixels éloignés spatio-temporellement ont des liens moins forts, avantageant ainsi la localité spatiale des regroupements, y compris dans le cas de pixels de caractéristiques distincts.

2.8.10.2 Mélange de gaussiennes

La méthode de [Megret R.] nécessite l'estimation du flot optique, mais n'a pas besoin d'initialisation. Le

nombre de classes n'est pas fixé a priori, on peut obtenir plusieurs classifications de niveaux de détails différents.

2.8.10.3 Réseau spatio-temporel de primitives

Pour avoir une représentation plus complète de la séquence que le tube spatio temporel, il est nécessaire de prendre en compte les relations entre les primitives, ce qui conduit à une structure spatio temporelle.

2.8.10.4 Structures spatio-temporelles par regroupement

[Megret R.] définit la structure spatio-temporelle basée par le regroupement. Celui-ci peut avoir lieu **spatialement par similarité des caractéristiques visuelles statiques (couleur, texture)**, **spatialement par cohérence du mouvement**, ou par **continuité temporelle**.

Les **structures spatio-temporelles** sont un regroupement récursif de pixels du bloc vidéo. Les relations spatio-temporelles à la base des regroupements de structures sont décomposées en **un aspect temporel, la projection temporelle**, et **un aspect spatial, la relation synchrone**.

Deux niveaux de détails déterminent les regroupements :
-Le premier niveau de détail se décompose en similarité de couleur ou de texture, similarité de mouvement, et continuité temporelle;
-Le second niveau de détail possède trois critères pour les regroupements : la similarité de couleur/texture, la similarité de mouvement, la proximité spatiale.

L'approche de [Megret R.] de la **segmentation de trajectoires par le mouvement** est basée sur des relations synchrones entre les trajectoires, permettant de comparer aussi bien des couleurs dans une image, que de mesurer la similarité du mouvement entre deux structures sur un intervalle temporel étendu.

2.9 Suivi basé sur l'apparence avec un réseau de caméras disjointes

[Madden C., Dahai Cheng E., Piccardi M.] ont proposé une méthode de suivi de personnes à travers un réseau de **caméras de surveillance disjointes** basé sur la **représentation de l'apparence invariante à la luminosité** ([Huang T., Russell S.J.],[J. Orwell, P. Remagnino, G.A. Jones], [Chang T.H., Gong S.], [Javed O., Rasheed Z., Shafique K., Shah M.], [Javed O., Shafique K., Shah M.], [Piccardi M., Cheng E.D.]).

Si les caméras ou les individus sont disjoints en temps et en espace, la **cohérence** n'est pas maintenue. Dans la plupart des systèmes, les vues sont **disjointes** car pour un opérateur humain, il n'est pas nécessaire de voir continuement une personne pour la suivre. Si des systèmes automatiques peuvent suivre des personnes à travers des vues disjointes, alors la vidéo surveillance devient possible.
[Madden C., Dahai Cheng E., Piccardi M.] mettent en correspondance des individus à partir de caméras **disjointes** dans des scénarios de vidéo surveillance typique. **L'approche simplifiée consiste à segmenter et suivre chaque personne dans une seule caméra et l'information pertinente (masque de l'objet et valeurs des pixels dans chaque image) est stockée dans un enregistrement (une trace). Le but est alors de trouver des correspondances entre les traces.**

Le suivi de personnes pendant qu'elles bougent, au travers de vues disjointes, est un problème difficile puisque **leur apparence varie significativement d'une vue à l'autre** à cause des variations des conditions lumineuses. Les changements dans l'apparence sont dûs aux variations d'illuminations et à la géométrie déformable des personnes. ([Javed O., Rasheed Z., Shafique K., Shah M.], [Javed O., Shafique K., Shah M.]) proposent un algorithme pour **compenser les diverses conditions d'illuminations** en estimant la fonction de transfert d'intensité entre chaque paire de caméra pendant une **phase d'entraînement initiale**.

[Madden C., Dahai Cheng E., Piccardi M.] proposent une **approche basée sur la représentation de l'apparence** et une transformation de l'intensité adaptative aux données, pouvant tolérer des variations

d'illuminations arrivant dans des scènes typiques de surveillance.
Une représentation par l'apparence invariante à l'illumination, basée sur un algorithme de « clustering » couleur « k-means » capable de traiter les faibles changements de pose d'une personne en mouvement, donne lieu à un ensemble de couleurs clustérisées, le « Major Colour Spectrum Histogram Representation » (MCSHR) décrivant les principales couleurs de l'objet. Une représentation de l'apparence ainsi définie sera utilisée pour chacun des objets segmentés dans l'image. Une **mesure de similarité compare les représentations par apparence entre deux individus** afin de quantifier la **similarité globale** entre deux MCSHR. Pour augmenter la validité de la mise en correspondance, **l'intégration des décisions de mise en correspondance est calculée tout au long du suivi**, le long des traces individuelles.
[Yu Y., Harwood D.] présentent un **modèle d'apparence** établissant la **correspondance** de personnes entre les images successives. Dans les **métro**, un ensemble de caméras observent les activités humaines décalées en espace et temps. Il faut donc établir des **correspondances** entre les observations des personnes qui **disparaissent** et **apparaissent** selon les caméras. **En supposant que l'apparence d'une personne ne change pas d'un point de vue à l'autre, les primitives d'apparence peuvent donc être utilisées pour mettre en correspondance les images. Cette mise en correspondance doit s'accommoder des variations d'illuminations, de postures et de changements de vues** et le critère de mise en correspondance doit refléter les réelles différences entre les observations (cf. figure 58).

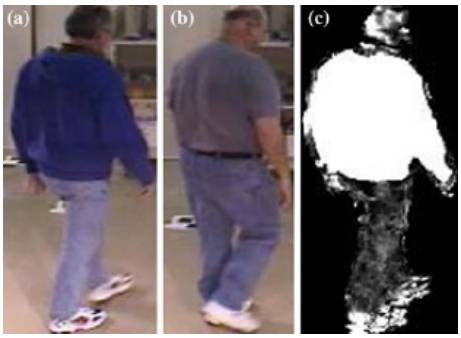


Figure 58 : L'image de log-vraisemblance (c) reflète la différence dans l'apparence locale entre l'image test (a) et l'image modèle (b). Un pixel brillant indique un ratio de la log-vraisemblance élevé [Yu Y., Harwood D.].

Un modèle d'apparence est construit par l'estimation d'un noyau de densité, basé sur les primitives statistiques spatiales et colorées. Pour présenter l'information d'une séquence vidéo dans laquelle la posture et la vue des personnes peuvent changer, des **images clés** sont choisies dans la séquence et une mesure de similarité entre séquences est calculée par la distance entre images clés. Les résultats montrent une invariance en illumination, insensible à la posture de la marche, ce qui est important pour un modèle d'apparence discriminante.

Le plus connu des **modèles d'apparence** est **l'histogramme de couleur** ([Comaniciu D., Ramesh V., Meer P.], [Fieguth P., Terzopoulos D.]) mais bien que robuste aux déformations non rigides, il ne contient aucune information géométrique, donc il ne peut discriminer des apparences possédant la même distribution colorée mais différentes dans la structure des couleurs. Par exemple, une personne portant une chemise bleue et un pantalon marron ne sera pas différenciée d'une autre personne portant une chemise marron et un pantalon bleu. Pour intégrer l'information de structure, [Elgammal A., Duraiswami R., Davis L.S.] propose un espace contenant à la fois les valeurs des primitives et la position spatiale des primitives. Mais cette approche pouvant différencier les structures est sensible à la pose. Par exemple, une personne qui marche avec le pied gauche au sol et le pied droit en l'air sera différente d'une personne qui marche avec le pied droit au sol et le pied gauche en l'air. Donc les primitives d'apparence invariantes à la pose sont préférées. Si une transformée géométrique [Li J.,Chellappa R.] est appliquée aux différents membres du corps humain, alors un modèle d'apparence invariant à la posture est obtenue. [Shan Y., Sawhney H., Pope A.] utilisent des histogrammes de la forme pour construire et mettre en correspondance des véhicules des séquences images.

D'autres **modèles d'apparence** ont été proposés pour la reconnaissance de visages et la mise en correspondance de véhicules. [Shan Y., Sawhney H., Kumar R.] proposent d'aligner les **contours des véhicules** et d'utiliser les primitives d'alignement pour **mettre en correspondance les véhicules**. Il serait intéressant d'**appliquer cela à l'apparence humaine puisque les plis des vêtements donnent lieu à des contours**.

[Yu Y., Harwood D.] proposent une soustraction de fond [Elgammal A., Duraiswami R., Harwood D., Davis L.S.] segmentant les personnes en mouvement, et des opérations morphologiques de fermeture et d'analyse en composantes connectées permettent d'obtenir la **silhouette** de la personne. En supposant que **l'apparence** actuelle d'une personne change peu entre les observations, les primitives d'apparence idéales doivent différencier différentes apparences et tolérer des changements comme le mouvement ou l'illumination. Une fois les traces des personnes générées, la mise en correspondance de modèles d'apparence est basée sur une trace. La similarité entre la distribution d'une image test et la distribution d'un modèle d'apparence est évaluée par la distance de Kullback-Leiber.
Quand une personne marche, ses mains peuvent, en bougeant, cacher le torse. De plus une personne peut faire demi tour et de nouvelles primitives apparaissent. Une solution consiste à utiliser toutes les images de la séquence et une mise en correspondance image par image sur la séquence entière. Cependant, ceci demande beaucoup de stockage et ne prend pas avantage de la redondance entre les images. Il est alors judicieux de sélectionner des **images clés** contenant toutes les informations de la séquence.

2.10 Panoramic Appearance Maps

[Gandhi T., Trivedi M.M.] présentent le concept de « Panoramic Appearance Maps » (PAM) pour la réidentification des personnes dans un **réseau multi caméra.** Leur groupe de recherche au « Computer Vision and Robotics Research Laboratory » (CVRR) à l'université de California San Diego effectue des recherches au sujet des « distributed interactive video array » (DIVA) systèmes. La soustraction de fond permet de détecter des personnes dans chaque caméra. La correspondance est établie entre les personnes détectées. Une nouvelle approche d'analyse des gestes avec des « **Shape Context** » 3D, un histogramme multi couche cylindrique basé sur la voxelisation du corps humain, est décrite dans [Huang K.S., Trivedi M.M.]. Le concept de PAM complémentaire des 3D « **Shape Context** », puisque ce dernier est basé sur l'information volumétrique, tandis que le premier est basé sur l'information d'apparence de surface.
Chaque personne est suivie dans plusieurs caméras et la position du sol est déterminée par triangulation. En utilisant la géométrie de la caméra et la localisation de la personne, une **carte panoramique** centrée sur la personne est crée. L'axe horizontal de la carte représente l'angle azimuth par rapport au système de coordonnées du monde, et l'axe vertical représente la hauteur de l'objet au dessus du sol. Le PAM combine les informations issues de toutes les caméras (la distribution colorée à différentes hauteurs au dessus du niveau du sol et à différentes azimuth autour de la personne), formant ainsi une seule signature du corps de la personne, qui sera utiliser pour la ré identification. L'information colorée est utilisée comme primitive d'apparence pour la comparaison. Chaque pixel de la carte possède une information colorée des caméras qui l'observent. Mais d'autres informations d'apparence comme la texture peut aussi être intégré. La carte générée de deux événements différents peut être comparée pour trouver les mises en correspondance potentielles. Une mesure de la distance euclidienne pondérée est proposée pour la comparaison des cartes entre les divers suivis, sélectionnant la meilleure mise en correspondance. L'intégration temporel améliore la mise en correspondance. Les personnes sont correctement ré identifiées en comparant leurs cartes d'apparence.

Chapitre 4 – Suivi dans un réseau de caméras

1 Introduction sur le suivi dans un réseau de caméras

La **surveillance visuelle** est devenue un domaine de recherche active dans les années récentes ([Hu W., Tan T., Wang L., Maybank S.], [Valera M., Velastin S.]). Le système **PRISMATICA** [Velastin S., Boghossian B., Lo B., Sun J., Vicencio-Silva M.] développé par la fondation EU traite de la sécurité dans les transports publics. [Remagnino P., Shihab A., Jones G.] ont introduit le concept des « **agents intelligents »**, modules autonomes regroupant des informations de plusieurs caméras et construisant le modèle de la scène de façon incrémentale. Des caméras multiples avec des champs de vue se recouvrant offrent une couverture de la scène plus importante, fournissant une information 3D plus riche et autorisant des occultations, des estimées exactes de la position du sol et des hauteurs des personnes, et l'observation des primitives de plusieurs perspectives. **D'un autre côté, les caméras ayant des vues qui ne se recouvrent pas peuvent fournir une couverture d'une grande zone sans perdre en résolution**. Un des problèmes de ces applications est de **ré identifier les objets** qui sont sortis du champ d'une caméra et entrent de nouveau, soit dans le champ de la même caméra, soit dans le champ d'une autre caméra. Ce problème est souvent difficile puisqu'un objet peut avoir un certain nombre de correspondants et il n'est pas toujours possible de différencier les correspondants. Dans ce cas, il est préférable d'identifier tous les correspondants possibles avec des primitives au niveau bas comme la couleur, la texture, et les transitions temporelles entre les caméras afin d'élaguer la recherche.
Les travaux initiaux sur la ré identification furent initiés par les applications de trafic routier où les véhicules sont rigides, ont des couleurs uniformes et sont situés sur des chemins bien définis. [Trivedi M.M., Gandhi T., Huang K.S.] décrivent les **mises en correspondance** de véhicules avec des primitives de **couleur et taille.** Le suivi de personnes et la ré identification sont souvent plus complexes car les personnes sont articulées, se meuvent de façon arbitraire et souvent sont vêtues avec des couleurs différentes. [Kettnaker V., Zabih R.] proposent d'utiliser la similarité des vues de personnes, et la plausibilité de transition d'une caméra à la suivante dans un **réseau bayésien**. [Javed O., Rasheed Z., Shafique K., Shah M.] utilisent de multiples primitives basées sur l'espace-temps (localisations des entrées/sorties, vitesse, temps de voyage) et **l'apparence (histogramme coloré)** dans un **réseau probabiliste** pour identifier les meilleures mises en correspondances. [A. Mittal, L. Davis] proposent un système de suivi de personnes multi-camera appelé « **M2-tracker** ». Ils développent un algorithme **stéréo basé région** qui trouve la position 3D grâce à la connaissance des régions appartenant à l'objet dans les deux vues. [Chang T.H., Gong S.] développent un **réseau bayésien** pour **fusionner** les informations de différentes caméras pour le suivi de personnes. Ils maintiennent les identités des objets pendant les occultations temporaires grâce à la **forme et l'apparence** des modèles de personnes. [Wu T., Matsuyama T.], grâce à de multiples caméras, obtiennent une reconstruction de la forme basée sur les voxels en temps réel. Voyons à présent l'analogie entre le suivi d'une personne dans un réseau de caméras et le suivi des membres d'une personne avec une seule ou deux caméras. A cette fin, nous présentons le suivi du haut du corps à travers un réseau bayésien.

2 Suivi du haut du corps avec des filtres à particules à travers un réseau bayésien

L'estimation de la posture d'un modèle articulé et son suivi est un problème complexe, que ce soit en mono, stéréo ou plusieurs caméras, et aussi à cause de l'aspect temps réel.

Les algorithmes pour le suivi du corps doivent traiter avec un espace de haute dimension dans lequel la probabilité jointe est hautement multidimensionnelle.

Dans ce contexte, les méthodes peuvent être classées entre :
-les approches **déterministes** ([Bregler C., Malik J.], [Plänkers R., Fua P. 03]);
- les approches **stochastiques** [[Demirdjian D., Taycher L., Shakhnarovich G., Grauman K., Darrell T.], [Jiang Gao and Jianbo Shi], [Sminchisescu C., Triggs B. 01]), la dernière étant plus robuste.

Les méthodes **déterministes** peuvent suivre en temps réel avec des caméras vidéos [Demirdjian D., Ko T., Darrell T.]**,** mais échouent dans le contexte monoculaire à cause des optimums locaux conduisant à des ambiguïtés pour des mouvements rapides ou des occultations [Demirdjian D., Taycher L., Shakhnarovich G., Grauman K., Darrell T.]. La propagation de croyances fournit un cadre judicieux pour réduire la dimension de l'espace des hypothèses générées, rendant le filtre à particules approprié.
Les algorithmes **stochastiques** sont utiles en vision **monoculaire** pour résoudre les **ambiguïtés** résultant de l'inférence de la posture en 2D ou en 3D, en particulier quand un filtre à particules [Andrew Blake and Michael Isard.] à plusieurs hypothèses est utilisé. [Sigal L., Bhatia S., Roth S., Black M.J., Isard M.] proposent un modèle graphique pour estimer la posture du haut du corps à partir de plusieurs images via une **propagation de croyances** non paramétriques. Mais la grande dimension de l'espace des paramètres interdit le temps réel de toutes ces techniques, surtout celles statistiques.

2.1 Les modèles de graphes

Les techniques probabilistes ont rencontré beaucoup de succès en vision par ordinateur, tant au niveau des **modèles image basés sur les pixels** [Jojic N., Petrovic N., Frey B.J., Huang T.S.], **au suivi dans les espaces paramétriques haut niveau** [Isard M., Blake A., 98] mais lorsque la taille de l'espace augmente, il est nécessaire de décomposer le problème en un **modèle graphique structuré** [Jordan M.I., Sejnowski T.J., Poggio T.]. **Les composantes de base sont des noeuds d'un graphe où chacun des noeud est conditionnellement indépendant de tous sauf des voisins adjacents.**

Lorsque les noeuds sont des éléments image, les voisins peuvent être :
-proches spatialement dans l'image;
-des niveaux adjacents dans une représentation multi échelle;
-des instants proches dans une séquence.

Des objets complexes peuvent être décomposés en graphes où les noeuds sont des sous parties de l'objet et un lien indique les deux parties connectées. Cette représentation permet une **inférence** computationnelle **linéaire** et non exponentielle par rapport à la taille du graphe. L'inférence exacte sur des modèles de graphes est possible dans des circonstances précises. Dans les autres conditions, l'**échantillonneur de Gibbs** est utilisé pour générer des échantillons approximatifs issus de la distribution jointe [Geman S., Geman D.], mais **pour la plupart des applications en vision par ordinateur cette technique n'est pas possible**.
Des méthodes d'inférence approchée peuvent être utilisées pour les modèles gaussiens linéaires conditionnels. Deux méthodes récentes permettent **d'approcher** l'inférence sur des graphes plus généralistes:
-Les **propagations de croyance en boucles** (« Loopy Belief Propagation LBP » [Yedidia J.S., Freeman W.T., Weiss Y.]) applicables aux graphes avec cycles;
-Les **filtres à particules** [Doucet A., De Freitas N., Gordon N.] autorisant l'utilisation de distributions plus générales sur des variables aléatoires à valeur continues mais appliquées seulement sur des graphes avec une simple structure en chaîne linéaire. La restriction aux variables cachées gaussienne est très onéreuse, ce qui a rendu le filtre à particule très populaire.

A partir des idées des filtres à particules et de la propagation de croyance (« belief propagation » BP), [Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.] ont développé une propagation de croyances non paramétrique (« **Nonparametric Belief Propagation** » **NBP**) applicable aux graphes. Le NBP est appliqué à l'inférence de relations entre les composantes d'un modèle de visages composé de primitives le décrivant. Bien que l'inférence exacte dans un graphe discret général est trop complexe, l'inférence approximative comme la propagation de croyances en boucle BP produit de bons résultats dans beaucoup de cas. Pour les problèmes d'inférence temporelle, les filtres à particules [Isard M., Blake A., 96] ont montré leur efficacité et constituent une alternative à la discrétisation (pour approcher des modèles graphiques à valeurs continues). **Ils constituent la base de bon nombres d'algorithmes de suivi** [Sidenbladh H., Black M.].

Le filtre à particule approche les densités conditionnelles non paramétriques par une collection d'éléments représentatifs. Bien qu'il soit possible de mettre à jour ces approches de façon déterministe par une

linéarisation locale, la plupart des implémentations utilisent des **méthodes de Monte Carlo pour la remise à jour stochastique des échantillons pondérés.** Les filtres à particules sont très efficaces, ils sont spécialisés dans les problèmes temporels dont les correspondants en graphes sont les chaînes de Markov (cf. figure 59).

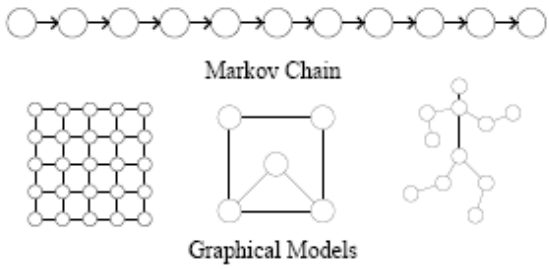


Figure 59 : Les filtres à particules font l'hypothèse que les variables vérifie l'hypothèse de Markov. L'algorithme NBP étend la technique du filtre à particule aux modèles graphiques structurés arbitraires [Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.].

L'algorithme « **Non Parametric Belief Propagation** » NBP est différent de l'algorithme BP pour deux raisons :
-En premier lieu, pour les graphes avec des cycles, on ne forme pas des arbres de jonction mais on itère la remise à jour locale de messages jusqu'à convergence dans un BP bouclé. Ceci permet de réduire la dimension de l'espace dans lequel les distributions sont inférées;
-Deuxièmement, un algorithme de remise à jour de messages, adapté aux graphes contenant des potentiels non gaussiens et continus, est fourni. Les messages produits peuvent être calculés en utilisant un **échantillonnage de Gibbs local**.

Les modèles graphiques associent chacun des noeuds à une variable aléatoire cachée non observée, et une observation locale du bruit. Pour les graphes acycliques ou en structure d'arbres, la distribution conditionnelle souhaitée peut être calculée directement par propagation de messages de façon locale, la propagation de croyances BP. Pour les modèles graphiques avec des variables cachées continues, une évaluation analytique de l'intégrale de remise à jour est souvent impossible et donc on représente les messages non paramétriques par une densité noyau estimée.

La remise à jour de BP se décompose en deux étapes. Les produits de messages combinent les informations des membres voisins avec la croyance locale, conduisant à une fonction résumant toutes les connaissances potentielles au sujet de la variable cachée, c'est **la fonction de vraisemblance**. Cette fonction de vraisemblance est combinée avec une **fonction potentielle**. **L'algorithme stochastique de propagation de croyance non paramétrique** approche ces deux étapes, produisant des représentations non paramétriques consistantes des messages.

En supposant que les fonctions potentielles sont des mélanges de gaussiennes pondérées, le produit de gaussiennes est lui même une gaussienne. On peut utiliser un échantillonneur de Gibbs [Geman S., Geman D.] pour dessiner de façon asymptotique des échantillons du produit de densités.

Dans les « **simulated annealing** », l'échantillonneur de Gibbs remet à jour la chaîne de Markov dont la dimension de l'état est proportionnelle à la dimension du graphe. Au contraire, **NBP utilise des échantillonneurs de Gibbs locaux**, impliquant chacun quelques noeuds. Dans certaines applications, le potentiel d'observation est spécifié par des fonctions analytiques. L'échantillonneur de Gibbs peut être adapté dans ce cas par une **fonction d'importance** [Doucet A., De Freitas N., Gordon N.].
L'algorithme NBP dans une seconde étape propage chacun des échantillons à partir du produit des messages en approchant la remise à jour des croyances. **PAMPAS** [Isard M.] a proposé une généralisation du filtre à

particules avec un noyau déterministe.
Les modèles graphiques gaussiens fournissent une des distributions continues pour lesquelles l'algorithme BP peut être implémenté de façon exacte. Pour cette raison, les **modèles gaussiens** sont utilisés pour tester l'**exactitude de l'approche non paramétrique** faite par le **NBP**.

[Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.] utilisent les NBP pour inférer des relations entre les coefficients PCA dans un modèle basé composantes du **visage**. Le modèle étend l'approche de ([Felzenszwalb P.F., Huttenlocher D.P. 03], [Moghaddam B., Pentland A.]) pour **estimer la localisation mais aussi l'apparence des membres cachés**. Les modèles d'apparence locaux basés sur les membres ont des indices communs avec les modèles articulés utilisés pour le suivi.

[Isard M.] décrivent l'algorithme **PAMPAS** (**Particle Message PASsing**) combinant le LBP « Loopy Belief Propagation » avec les idées du filtre à particule.

La propagation de croyances (« Belief Propagation » BP) est en vogue depuis quelques années pour calculer des **inférences** dans des **réseaux bayésiens** [Jordan M.I., Sejnowski T.J., Poggio T.] **et a été** récemment appliquée aux **graphes avec cycles** sous la dénomination de **« Loopy Belief Propagation »** [Yedidia J.S., Freeman W.T., Weiss Y.]. La méthode consiste à passer des messages entre les noeuds du graphe.
Lorsque le graphe est une chaîne, un filtre à particules [Doucet A., De Freitas N., Gordon N.] peut être utilisé. **Il représente les probabilités marginales dans une forme non paramétrique, l'ensemble des particules.** Le filtre à particule, largement utilisé en vision par ordinateur, fonctionne bien avec les modèles de vraisemblance des images. **L'algorithme PAMPAS modifie le BP pour se servir des ensemble de particules comme des messages et donc permettre une inférence approchée sur des modèles graphiques à valeur continues.**

La **propagation de croyance** peut être analysée comme des couples d'ensemble de variables cachées X et d'ensemble de variables observées Y. Le réseau de particules se propage dans le réseau par propagation de croyances. Le principe de n'importe quel algorithme de propagation de croyances avec des ensembles de particules est celui de l'**approximation de Monte-Carlo**. Une solution consiste à utiliser **l'échantillonnage d'importance** pour certaines particules.

Une propriété des modèles de vision par ordinateur est que la **fonction potentielle** peut s'écrire comme un mélange de gaussiennes et la vraisemblance est difficile à échantillonner. L'algorithme **PAMPAS** est spécialiste pour calculer la propagation de croyances avec ce type de modèle. [Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.] a développé un algorithme presque similaire pour le calcul de la propagation de croyances avec l'aide des ensembles de particules, qu'ils ont appelé **NPB**, « **Non parametric belief propagation** ». Afin de traiter l'explosion exponentielle des messages produits, ils introduisent un **échantillonneur de Gibbs**. Tandis que [Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.]] a démontré que **l'échantillonneur de Gibbs** est très efficace pour quelques applications, dans les scènes texturées l'algorithme peut générer des échantillons avec une faible masse de probabilité dans quelques régions. Un graphe représentant les objets liés a été construit et l'algorithme **PAMPAS** est appliqué combiné à l'échantillonneur de Gibbs [Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.] pour la propagation de messages et de croyances, afin de localiser les objets dans une scène bruitée.

L'algorithme **PAMPAS** peut être efficace pour la localisation des structures articulées dans des images, en particulier des personnes. Des algorithmes de détection de personnes existent déjà pour la localisation et le groupement des membres du corps [Mori G., Malik J.] mais ne correspondent pas à un réseau probabiliste. Récemment des chercheurs ont eu de bons résultats en cherchant des structures liées dans des images grâce à une représentation en modèles graphiques ([P.F. FELZENSZWALB & D.P. HUTTENLOCHER. 00], [Ioffe S., Forsyth D.A., 03]). Un des avantages de la représentation en modèle graphique est qu'elle s'étend naturellement au suivi, en augmentant la taille du graphe et les liens entre les noeuds à des intervalles de temps

adjacents.
Beaucoup d'approches pour la détection et le suivi de personnes sont basées sur les modèles articulés du corps, dans lesquels le corps est vu comme un arbre cinématique en deux dimensions comme le modèle « cardboard » [Ju S., Black M., Yacoob Y.] ou en trois dimensions [Sminchisescu C., Triggs B. 01], conduisant à un espace paramétrique de grande dimension.
Cet **espace de grande dimension peut être réduit par une représentation hiérarchique de la personne exploitant la structure en arbre du modèle** [MacCormick J., Isard M.]. Malgré tout, cette méthode a des inconvénients comme l'impossibilité d'incorporer des traitements bas niveau ou **l'initialisation** automatique.

2.2 Avec un modèle de membres « lâches »

[Sigal L., Isard M., Sigelman B.H., Black M.] proposent un modèle du corps avec ses membres « lâches », c'est-à-dire non connectés de façon rigide mais plutôt en attraction l'un vers l'autre. Le corps est représenté par un modèle graphique dans lequel chacun des noeuds du graphes correspond à un membre du corps (torse, bras, etc). Chacun des membres est paramétré par un vecteur définissant sa position et son orientation dans un espace 3D de coordonnées globales, et chaque membre est traité indépendamment. **Le corps entier est assemblé par inférence globale sur tout le modèle graphique.** Les contraintes spatiales entre les membres du corps (relations spatiales et angulaires entre les membres adjacents) sont traitées dans les liens du graphes. On fait l'hypothèse que les variables d'un noeud sont conditionnellement indépendantes des noeuds non immédiatement voisins, connaissant les valeurs des noeuds voisins. Ceci est mis à défaut dans le cas de l'auto occlusion d'un membre. Chacun des membres est modélisé par un cylindre avec 5 paramètres fixes (longueur du membre, etc), et 6 paramètres estimés contenant la position en 3D (3 paramètres) et l'orientation en 3D (3 paramètres) du membre dans le système de coordonnées globales. Chaque lien a une distribution de probabilité conditionnelle qui modélise les dépendances probabilistes entre les membres adjacents, et approchée par un mélange de gaussiennes. Chacun des noeuds du graphe possède une fonction de **vraisemblance** qui modélise la probabilité d'observation des images conditionnées sur la position et l'orientation des membres. A chacun des membres est associé une probabilité dans le sens descendant dans le graphe (de la cuisse vers le mollet par exemple) ou ascendant (de la cuisse vers le torse par exemple).
La détection de personnes ou son suivi exploite la propagation de croyances pour estimer la distribution de croyances sur les paramètres. **L'inférence probabiliste combine un modèle du corps avec un modèle de vraisemblance probabiliste.** Un algorithme estime les distributions de croyance pour chacun des membres du corps. L'algorithme adopté est celui de (**PAMPAS** [Isard M.], [Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.]), **une généralisation de l'algorithme à particules qui autorise des inférences sur n'importe quel graphe plutôt que uniquement sur une chaîne**. L'ensemble des particules propagées dans un filtre à particules standard est traité comme une approximation des messages utilisés dans l'algorithme de propagation de croyances, en remplaçant la distribution conditionnelle par un produit des messages arrivant, qui peuvent être approchés par **l'échantillonnage d'importance**.

Une personne peut être suivie en temps réel grâce à un filtre à particules propagé à travers un réseau bayésien. [Sigal L., Bhatia S., Roth S., Black M.J., Isard M.] présentent une méthode probabiliste pour détecter et suivre le corps d'une personne en 3D de façon automatique par un modèle de membres « lâches » avec des paramètres continus représentant la localisation et la posture de la personne (cf. figure 60). L'inférence sur ce modèle est menée par la **propagation des croyances sur un ensemble de particules**. La propagation de croyance permet d'éviter la distinction entre initialisation et suivi, et autorise à utiliser des détecteurs bas haut pour les membres du corps afin de stabiliser l'estimation du mouvement et de fournir une initialisation à chaque instant. De plus, les probabilités conditionnelles entre les membres dans l'espace et le temps sont apprises à partir de données d'entraînement, et une fonction **de Gibbs** apprise à partir des données d'entraînement modélise les dépendances conditionnelles entre les mesures sur les images.Cette approche pourrait être étendue aux images mono caméras ou avec des caméras en mouvement. Un des inconvénient de cette approche provient du fait de l'hypothèse d'indépendance des membres de même nature entre la droite et la gauche conditionnellement à la position du torse, omettant des postures quand un membre est caché par un autre. Le problème serait traité plus facilement avec un modèle cinématique du corps sous forme d'un arbre,

donc le modèle à membre lâches serait une étape intermédiaire entre les détecteurs bas niveau et le modèle cinématique complet.

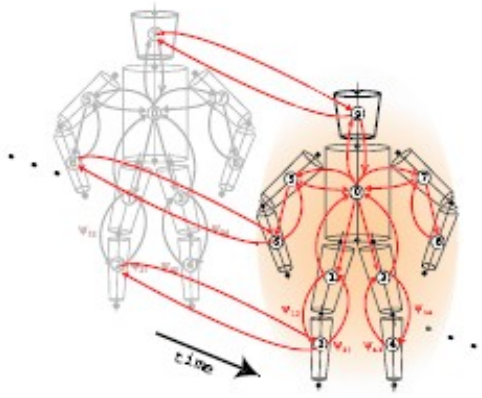


Figure 60 : Modèle graphique pour une personne. Les noeuds représentent les membres et les flêches représentent les dépendances conditionnelles entre les membres. Les dépendances temporelles sont montrées entre deux mages sur cette figure. Dans la réalité, chaque membre est connecté par un arc flêché au même membre dans les images précédente et suivante [Sigal L., Bhatia S., Roth S., Black M.J., Isard M.].

Les probabilités conditionnelles des poses des membres en 3D sont apprises à partir de données d'entraînement. Dans ce travail par rapport au précédent [Sigal L., Isard M., Sigelman B.H., Black M.], il ne s'agit pas uniquement de l'estimation de la pose mais également du suivi de celle ci.

La plupart des techniques courantes modélisent le corps humain par un arbre cinématique en deux dimensions [Ju S., Black M., Yacoob Y.], ou trois dimensions ([Bregler C., Malik J.], [Deutscher J., Blake A., Reid I.], [Sidenbladh H., Black M.J., Fleet D.J.], [Sminchisescu C., Triggs B. 01]), conduisant à un espace de grande dimension. La recherche de la solution dans un tel espace étant impossible, **les méthodes courantes reposent sur une initialisation manuelle du modèle du corps**. Quand de tels algorithmes perdent le suivi, la dimension de l'espace de recherche rend difficile de recouvrir au suivi. **L'utilisation d'un modèle de « membres lâches » et articulés et la propagation de croyances fournit un bon moyen pour incorporer l'information des divers détecteurs de membres.**

Le modèle du corps nécessite une spécification des relations probabilistes entre les articulations à un instant donné et au cours du temps. Le modèle nécessite aussi une image de la mesure de vraisemblance pour chacun des membres. En utilisant les données d'entraînement de membres connus dans l'image, on peut apprendre un nouveau modèle de la vraisemblance qui capture les statistiques des articulations. Les vraisemblances sont apprises grâce à un modèle de Gibbs [Zhu S., Wu Y., Mumford D.]. Quatre caméras calibrées participent au suivi dans un environnement intérieur. Les modèles du corps ne sont pas nouveaux pour le suivi articulés, par exemple ([Ioffe S., Forsyth D.A., 01], [Ioffe S., Forsyth D.A., 03], [Ioffe S., Forsyth D.A., 99], [Ramanan D., Forsyth D.]). ([Ioffe S., Forsyth D.A., 01], [Ioffe S., Forsyth D.A., 03]) détectent les membres du corps et les regroupent en des figures dans une approche « bas haut ».

Dans les travaux précédents [Sigal L., Isard M., Sigelman B.H., Black M.], les fonctions potentielles reliant les membres se construisent manuellement tandis que dans ce travail elles sont apprises à partir de données d'entraînement. Chacun des liens entre deux membres possède une fonction potentielle associée qui code la compatibilité entre les configurations des paires de membres et peut être vu comme la probabilité de la configuration d'un membre conditionnellement à la configuration d'un autre membre. La fonction potentielle est en générale non gaussienne et est approchée par un mélange de gaussiennes. L'image de vraisemblance (d'observation des mesures sur l'image conditionnellement à la pose d'un membre) est un modèle probabiliste qu'il faut combiner avec le modèle du corps.

De nombreux indices incluant les filtres de contours multi échelle [Sidenbladh H., Black M.] sont mis en place, et les dépendances conditionnelles sont modélisées entre les diverses réponses des filtres par apprentissage de la densité jointe en utilisant le modèle de Gibbs ([Roth S., Sigal L., Black M.], [Zhu S., Wu Y., Mumford D.]). L'inférence de la posture du corps est définie comme une estimation de la croyance dans le modèle graphique.

Afin de s'abstenir de l'espace des paramètres en six dimensions pour chacun des membres, les densités conditionnelles entre les membres qui sont non gaussiennes, et les vraisemblances non gaussiennes, on utilise une forme non paramétrique de la propagation de croyances ([Isard M.], [Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.]), généralisation de l'algorithme à particules ([Doucet A., De Freitas N., Gordon N.]) qui permet des inférences sur une graphes arbitraire plutôt que sur une simple chaîne.

Les messages envoyés dans la propagation de croyances standard sont ici approchées par un ensemble de particules, et la distribution conditionnelle utilisée dans l'algorithme à particules standard est remplacée par un produit des messages entrants, nécessaire pour la propagation de croyances. On utilise ici l'algorithme **PAMPAS** [Isard M.] plus adapté à la problématique et l'échantillonneur de Gibbs pour évaluer les produits de messages.

Les messages de la tête, des deux bras et des deux jambes sont envoyés vers le torse (cf. figure 61).

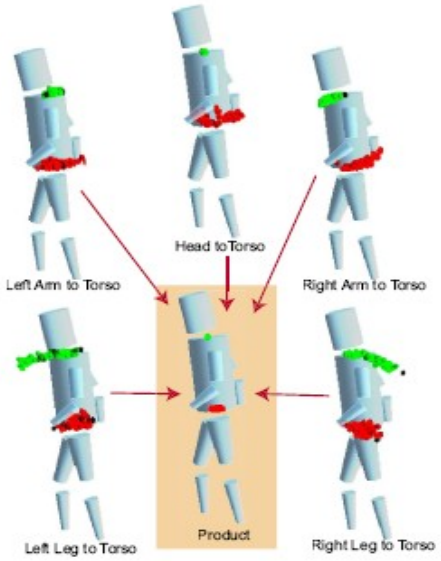


Figure 61 : Produit de messages : les messages de la tête, des deux bras et des deux jambes sont envoyés vers le torse [Sigal L., Bhatia S., Roth S., Black M.J., Isard M.].

Ces messages sont des distributions représentées par un ensemble d'échantillons avec des poids comme dans le filtre à particules.

Nous présentons un exemple de suivi avec un modèle de « membres lâches » (cf. figure 62).

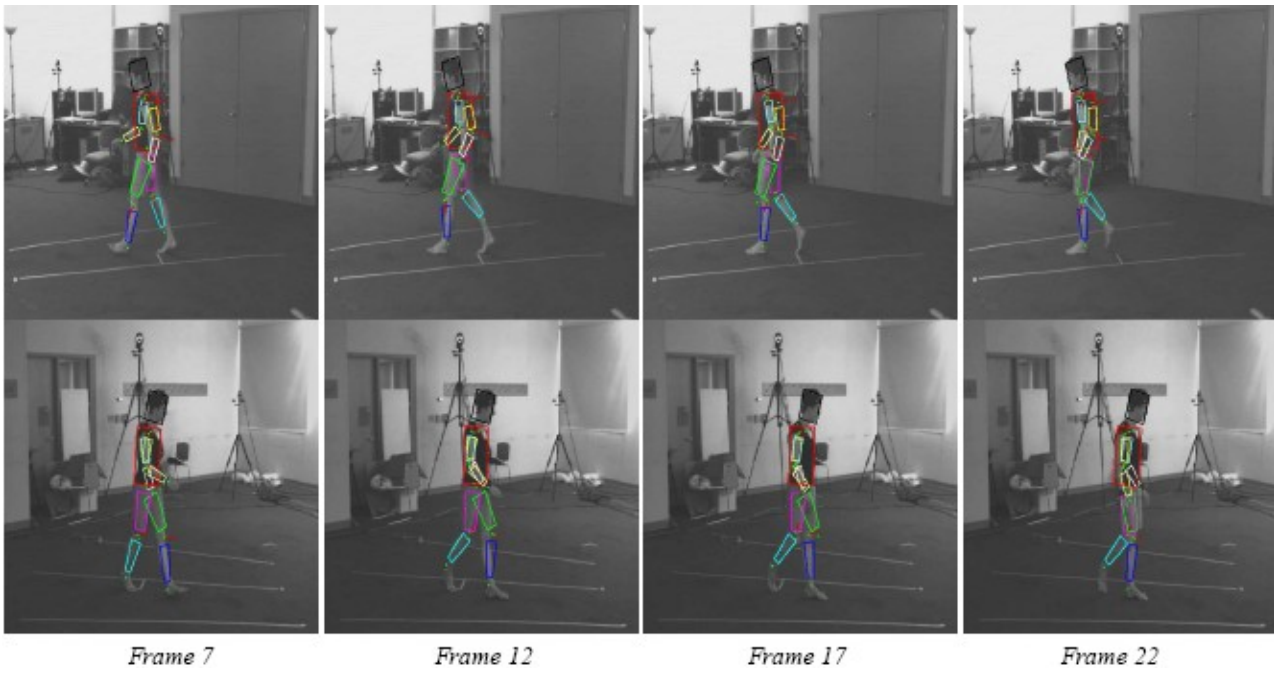


Figure 62 : Suivi avec modèle de « membres lâches » - quelques résultats sur une séquence [Sigal L., Bhatia S., Roth S., Black M.J., Isard M.].

2.3 Avec une seule caméra

[Noriega P. b] décrit une méthode **avec un modèle graphique articulé**, représentant la structure articulée du corps humain, pour le suivi du haut du corps dans un environnement non contraint (vêtements et lumière), dans des scènes couleur **monoculaires** (cf. figure 63).

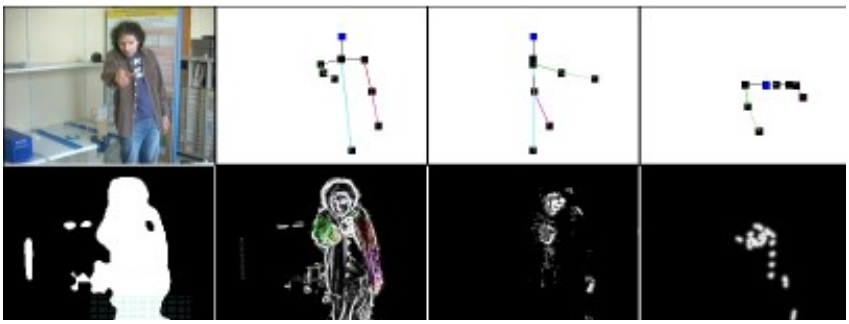


Figure 63 : Suivi du haut du corps. Dans la 1ère ligne, l'image originale, de face, de côté et du dessus, des positions obtenues des membres, avec une seule caméra. Dans la 2ème ligne, la soustraction du fond, les contours, la carte de couleurs du visage, et la carte de distance d'énergie du mouvement [Noriega P. b].

La propagation de croyances sur des **graphes factoriels** autorise le calcul des probabilités marginales des membres (cf. figure 64).

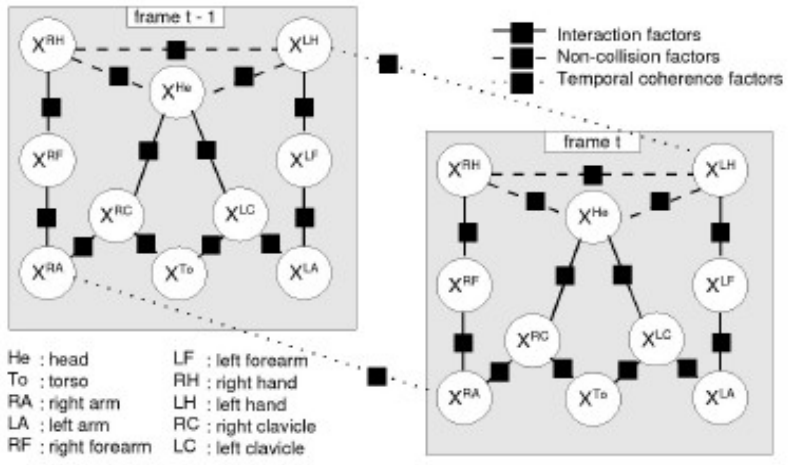
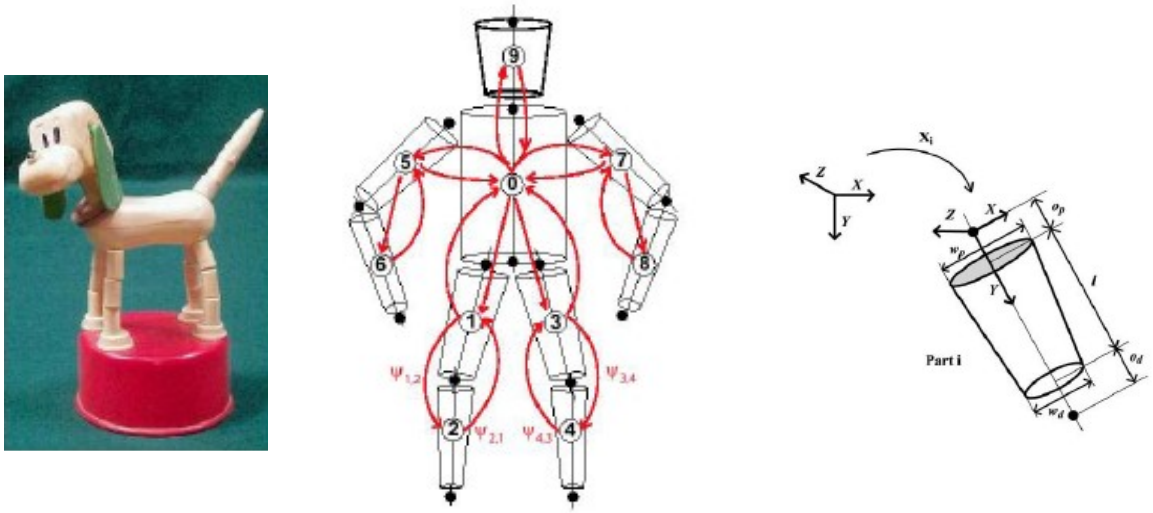


Figure 64 : Graphe factoriel. Les cercles correspondent aux noeuds qui sont des variables, les états des membres, et les carrés sombres aux noeuds factoriels (cohérence temporelle et interaction ou non-collision factorielle). Deux images consécutives sont représentées [Noriega P. b].

Le modèle du corps est formé de « membres lâches » [Sigal L., Isard M., Sigelman B.H., Black M.] incluant les contraintes articulatoires facilement intégrées dans des facteurs d'attraction. Pour résoudre les ambiguïtés liées au suivi monoculaire, les indices sont les contours robustes, les couleurs, et une carte d'énergie de mouvement (cf. figure 63).

Une façon d'éviter le problème de la haute dimension de l'espace de recherche des algorithmes stochastiques, consiste à utiliser un modèle du corps avec des « membres lâches » [Sigal L., Bhatia S., Roth S., Black M.J., Isard M.] (cf. figure 65), où la vraisemblance de chacun des membres est évaluée indépendamment. De cette façon, **un filtre à particule peut être associé avec chaque membre réduisant la dimension de l'espace de recherche au nombre de degrés de libertés du membre [Bernier O., Cheung-Mon-Chang P.]. L'influence entre les membres est prise en compte par la propagation de croyance des membres** à travers un **graphe factoriel** [Kschischang, Frey, Loeliger]. Une technique similaire est utilisée dans le cadre des scènes monoculaires [Gao J., Shi J.] avec pour indice l'énergie du mouvement (cf. figure 66). Le nombre d'indices est augmenté ici [Noriega P. b] pour accroître la robustesse du suivi. La tête et les mains sont suivies grâce à l'information colorée et aux niveaux de gris: soustraction du fond, énergie du mouvement, carte d'orientation des contours.
Une gaussienne sur la distance, entre deux points articulés, est utilisée pour calculer les **facteurs d'interactions** entre les membres articulés. Les **facteurs de comptabilité des images** sont calculés à partir de scores représentant la comptabilité entre un membre hypothèse et des indices extraits des images. Les bras ont tendance à bouger rapidement et sont sujets à des occultations partielles. Ainsi afin d'atteindre un niveau suffisant de robustesse, une fusion des indices de contour et d'énergie de mouvement est calculée. Le score de l'énergie du mouvement est calculée en considérant la distance gaussienne entre chacun des membres projetés et le pixel le plus proche où un mouvement a été détecté. La détection de mouvement est fournie par des différences d'images adjacentes.



(a) Analogie avec un jouet à poussoir pourvu d'articulations élastiques

(b) Modèle graphique

(c) 11 paramètres définissant un membre

Figure 65 : Modèle de tronc de cônes à membres indépendants dits « laches » ([Sigal L., Bhatia S., Roth S., Black M.J., Isard M.], [Noriega P. a]).

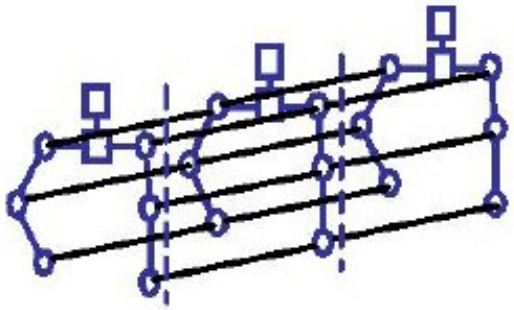


Figure 66 : Graphe intégrant une fenêtre temporelle sur trois images ([Gao J., Shi J.], [Noriega P. a]).

Le suivi bayésien récursif

Le haut du corps est modélisé par un graphe comprenant les membres représentés par des noeuds et des liens correspondant à des articulations, et des contraintes de non collision entre les membres. Un modèle de Markov peut être utilisé pour représenter la structure.

Le **graphe factoriel** complet inclut les états précédents pour prendre en compte la **cohérence temporelle**. Le **facteur de cohérence temporel** est une simple gaussienne, indépendante pour chacun des paramètres, centré sur la valeur de l'image précédente. Pour les mains qui peuvent bouger très rapidement, le facteur de cohérence temporelle est un mélange de deux gaussiennes similaires, une centrée sur les paramètres précédents et l'autre centrée sur la prédiction des paramètres courants en utilisant la vitesse précédente de la main. La probabilité jointe connaissant les observations, la probabilité marginale des états des membres est obtenue en utilisant la propagation de croyances sur un graphe factoriel [Bernier O., Cheung-Mon-Chang P.].

Les messages sont représentés par des ensembles d'échantillons pondérés. D'une image à la suivante, ils sont calculés via un algorithme de filtre à particules consistant en une étape de ré échantillonnage suivie par une

étape de prédiction basée sur la cohérence temporelle des facteurs [Andrew Blake and Michael Isard.]. L'algorithme de propagation de croyance en boucle (« loopy belief propagation ») est alors réduit, pour l'image courante, à un algorithme de propagation en boucle dans l'espace des états discrets, l'espace des états de chacun des membres restreint à ses échantillons. L'algorithme est équivalent à un ensemble de filtres à particules en interaction, où les échantillons pondérés sont réévalués à chaque image à travers une propagation de croyances prenant en compte les interactions entre les membres.

De plus, l'utilisation des filtres à particules en interaction avec la propagation de croyances [Bernier O., Cheung-Mon-Chang P.] simplifie l'algorithme en calculant de façon **récursive** une estimation dans un espace discret, au lieu d'utiliser un échantillonneur de Gibbs dans un espace continu [Sigal L., Bhatia S., Roth S., Black M.J., Isard M.]. Des contraintes sur les articulations sont construites dans les facteurs de compatibilité. Quelques exemples de suivi en monoculaire [Noriega P. b] sont présentés ci-dessous (cf. figure 67).

Figure 67 : Suivi 3D en monoculaire avec des poses difficiles incluant les occultations, les fonds texturés et les environnements non contraints (lumière et vêtements) [Noriega P. b].

2.4 Avec des caméras stéréo

[Bernier O.] présente un **modèle statistique** pour le **suivi 3D rapide du haut du corps articulé avec une caméra stéréo en temps réel**, similaire au **modèle de « membres lâches »** [Sigal L., Isard M., Sigelman B.H., Black M.] mais où la **cohérence inter images** est prise en compte, via la probabilité marginale de chacun des membres dans l'image précédente, comme information *a priori*. La **propagation de croyances** sert à estimer la probabilité marginale courante de chacun des membres. L'algorithme résultant correspond à un ensemble de particules, un pour chacun des membres, où le poids de chacun des échantillons est recalculé en **prenant en compte les interactions entre les membres**.

De façon analogue à [Sigal L., Bhatia S., Roth S., Black M.J., Isard M.], un modèle graphique représente le haut du corps, composé de M membres, chacun dans un état donné X (cf. figure 68).
Chacun des membres génère une observation, une image Y, et le modèle est composé de liens entre les membres représentant les articulations mais aussi des contraintes de non intersection. Chacun des états des membres est dépendant de son état à l'instant précédent. Les paramètres du modèle sont les probabilités conditionnelles P(Y/X), la prédiction de probabilité *a priori* des états des membres $P(X_t/X_{t-1})$ et le potentiel d'interaction pour chacun des liens entre les membres.
L'inférence dans un modèle graphique avec boucle peut se faire par la propagation de croyances avec boucle « Loopy Belief Propagation » pour les états discrets, ou la méthode de propagation de croyances non paramétrique [Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.] afin d'obtenir la distribution

marginale de chacun des membres sur l'image courante, qui à son tour peut être utilisée comme « prior » dans les images suivantes.

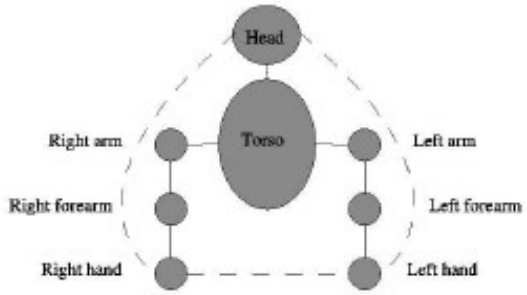


Figure 68 : Modèle graphique : les lignes représentent des les liens entre les membres, les tirets représentent les contraintes de non intersection [Bernier O.].

La probabilité marginale précédente est représentée par un ensemble d'échantillons pondérés. La croyance locale de chacun des membres est estimée par la méthode du filtre à particules standard : ré échantillonnage, prédiction, et nouvelle pondération par les probabilités des observations. La croyance est alors représentée par une somme pondérée des échantillons.

L'algorithme de propagation de croyance se réduit à un algorithme de propagation bouclé dans l'espace des états discrets. Dans cet espace chaque membre est représenté par ses échantillons. De plus la probabilité marginale est représentée par une somme pondérée des mêmes échantillons que ceux utilisés pour la croyance, l'estimation est donc récursive.

L'algorithme est équivalent à un ensemble de filtre à particules en interaction où les poids sont réévalués à chaque image à travers la propagation de croyances pour prendre en compte les liens entre les membres.

Initialement, un détecteur de visage à base de RN détecte le visage. L'information de profondeur est prise en compte. Les probabilités prédites sont des gaussiennes, indépendantes pour chacun des paramètres, centrées sur la valeur de l'image précédente. Les observations sont des points estimés 3D avec un facteur de confiance et une probabilité de couleur pour le visage. Les observations sont supposées indépendantes pour chacun des membres et chaque pixel. Pour chaque membre, la vraisemblance est proportionnelle à un score S.
-Pour la tête, le score S est une distance gaussienne à une sphère, multipliée par la probabilité colorée de la tête;
-Pour le torse, le score S est une distance gaussienne dont la forme est composée de deux cylindres plats;
-Pour les bras et les avant bras, le score est la distance à un « patch » rectangulaire, parallèle à l'image plane dans la direction du plus petit contour;
-Pour les potentiels d'interaction des liens, une gaussienne de la distance entre deux points liés est utilisée.

Le système suit correctement (cf. figure 69) même en présence d'auto occultations et cette méthode peut être généralisée à d'autres systèmes de suivi, monoculaire ou le corps entier. Les limitations de la méthode sont liées à l'impossibilité de l'étape de prédiction des échantillons de générer des échantillons dans des régions à forte vraisemblance. Pour résoudre ce problème, l'étape de prédiction devrait être conduite avec des « proposal maps » [Lee M.W., Cohen I.] pour chacun des membres, générant les échantillons pour les régions de grande vraisemblance, surtout pour les mains et les avant bras.

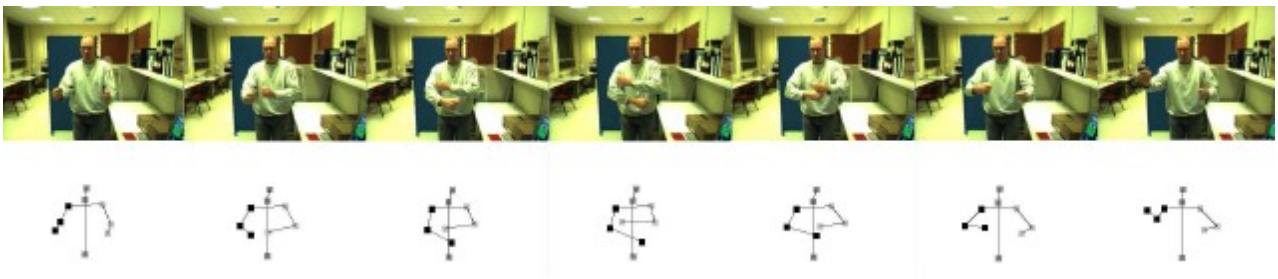


Figure 69 : Résultats de suivi sur une partie de la séquence [Bernier O.].

3 Fusion d'informations pour l'estimation de la structure d'un objet et la détection de son mouvement

Les travaux de [Noyer J-C.] s'intègrent dans la thématique de la fusion d'informations pour l'estimation et la détection, avec comme domaine applicatif la **vision par ordinateur**. Les outils développés concernent les **méthodes de Monte-Carlo séquentielles** telles que le **filtrage particulaire**, appliqué à la **perception multi capteurs.**

3.1 Fusion multicapteurs pour l'estimation de la structure et du mouvement 3D d'objets : une approche primitive

On cherche à estimer le mouvement 3D d'un objet et sa structure à partir d'un ensemble de capteurs parfois de nature physique différente. Les primitives points ou segments décrivent la forme à estimer. Une structure globale de filtre est mise en place grâce à la fusion multi capteurs, la « fusion centralisée » [Bar-Shalom Y., Li X.].
La solution du problème d'estimation repose sur la densité de probabilité conditionnelle que l'on peut décomposer en deux densités de probabilités élémentaires : une loi de transition et une loi d'observation. Dans le cas linéaire gaussien [Kalman R.E.], ces densités de probabilités restent gaussiennes au cours du temps, et le filtre de Kalman est la solution optimale. Dans le cas non-linaire (cas général), il n'y a pas de solution en dimension finie des équations du filtrage. Nombre d'auteurs ont proposé des solutions analytiques au problème d'estimation mais les solutions sont sous-optimales:
-1965 : filtrage de Kalman étendu;
-1991: filtrage particulaire;
-1996 : filtrage condensation; etc.
Le filtre de **Kalman étendu**, adapté au problème d'estimation multi capteurs, permet de suivre dans une séquence d'images un objet avec des informations issues de plusieurs capteurs non nécessairement synchrones, et d'estimer son mouvement. L'objet est décrit par des points caractéristiques ([Cox I.J, Hingorani S.L.], [Koller D., Daniilidis K., Nagel H.-H]), des contours [A. Blake and M. Isard], des régions. Ce filtre donne lieu à une **linéarisation des équations d'état**, lui permettant de se ramener à une solution localement linéaire gaussienne. Cependant, la linéarisation des équations d'états dans le filtre de Kalman étendu conduit à une perte en précision d'estimation de la structure et du mouvement, contrairement au cas linéaire gaussien.
Au début des années 90, le problème est résolu par l'utilisation d'un « **filtrage particulaire** », traitant les non-linéarités des modèles sans faire d'approximations, pour la résolution du problème d'estimation multi capteurs de la structure et du mouvement conjointement 3D d'objets, le suivi de la forme dans la séquence et la fusion des mesures issues des divers capteurs.

[Noyer J-C.] propose une **modélisation globale du problème d'estimation multi capteurs de la structure et du mouvement 3D**. Des équations d'état modélisent le problème en intégrant les mesures des capteurs hétérogènes.
L'approche de [Noyer J-C.] pour le suivi d'objets est à base de **primitives et de segments, convenant aux**

scènes d'intérieur contenant des objets modélisés sous forme polyédrique. Un objet en mouvement est caractérisé par son vecteur d'état composé des caractéristiques de la primitive (mouvement, structure) : les coordonnées 3D de chacune des extrémités du segment, et ses paramètres de mouvement.
Un système d'équations d'état modélise l'état avec une équation de dynamique décrivant l'évolution de l'état du système et une équation de mesure pour l'observation partielle que l'on en a. La résolution du problème multi capteurs ainsi posé dans le système d'équations, passe par **deux méthodes issues de la théorie de l'estimation dynamique : le filtrage de Kalman étendu, et le filtrage « particulaire ».**

[Noyer J-C.] résout le problème d'estimation multi capteurs des structure et mouvement 3D par filtrage de Kalman étendu. Il réalise **estimation et suivi**, mais également l'étape de **fusion des informations issues de chaque capteur**. Le filtre de Kalman comporte deux parties : une prédiction pour le calcul de la loi de transition entre deux états successifs, et la correction qui utilise l'observation pour donner une estimation de l'état. Dans le cas non linéaire, la propriété de gausienneté de la densité de probabilité n'existe plus. La résolution passe alors par la linéarisation des équations d'état autour de l'estimée et on obtient ainsi le filtre de Kalman étendu. La solution proposée par [Noyer J-C.] reprend la structure du type prédiction/correction dans un problème d'estimation multi capteurs multi cibles pour le suivi de cibles au cours du temps. Le filtre prédit les positions 3D des extrémités du segment primitive et ses paramètres de mouvement à l'instant t à partir des mesures disponibles à l'instant t-1.
Dans une étape de mise en correspondance, chaque primitive doit être suivie dans la séquence, et l'estimée doit être mise à jour (correction de l'estimation). La segmentation en points et segments générant trop de candidats, le suivi multi cibles multi capteurs peut avoir lieu avec des méthodes d'association probabilistes de données (PDAF) [Bar-Shalom Y., Li X.], le suivi multi-hypothèses (MHT) [D.B. Reid.], ou bien la minimisation d'une fonctionnelle de coût.
[Noyer J-C.] a choisi la distance de Mahalanobis (couramment utilisée en vision par ordinateur) entre deux observations (position, intensité et profondeur) pour chaque segment et pour chaque capteur. Les segments mis en correspondance participent à la correction de l'estimation de l'état à l'instant t. Un filtre de Kalman étendu traite séquentiellement les données provenant des divers capteurs, et fournit l'estimée de l'état, comprenant les paramètres de positions 3D du segment et ses paramètres de mouvement. Le schéma général est le suivant (cf. figure 70) : le premier capteur effectue une mise en correspondance entre l'état prédit et l'état observé, la différence sert à la correction utilisée par le capteur suivant, et ainsi de suite pour tous les capteurs, estimant en fin de chaîne la position et le mouvement 3D. Pour les capteurs ne réussissant pas à faire une mise en correspondance à cause de mesures manquantes, l'étape de « correction » n'est pas effectuée, ceci constitue l'intérêt de la structure séquentielle. Le traitement est temps réel.
Au début des années 90, le filtrage « particulaire » a permis de traiter le filtrage non linéaire sans aucune approximation. L'idée principale est de trouver une estimation directe de la densité de **probabilité conditionnelle** $P(X_t|Z_0{}^t)$), solution du problème de filtrage, par une approximation de type Monte-Carlo. On cherche à estimer l'évolution du processus X à partir de l'observation du processus Z.

X_t représente l'état du système, c'est-à-dire le processus à l'instant t, et $Z_0{}^t$ représente les mesures $Z_0{}^t=\{Z_0, Z_1, \ldots, Z_t\}$ jusqu'à l'instant t. La solution à ce problème d'estimation est donnée par convergence lorsque le nombre fini de particules du filtre augmente. L'ensemble de ces approches est regroupé sous le terme de « filtrage particulaire ». En vision par ordinateur, la dénomination fréquente est « algorithme de condensation » de [Andrew Blake and Michael Isard.].

La méthode « particulaire » repose sur la décomposition de la loi de probabilité conditionnelle $P(X_t|Z_0{}^t)$. La méthode du filtre à particules représente $P(X_t|Z_0{}^t)$ par un ensemble de mesures ponctuelles $\delta_{X^i}(X_t)$ d'amplitude $p_i{}^t$: $P(X_t|Z_0{}^t)=\sum_{i=1}^{N} p_t{}^i \delta_{X^i}(X_t)$.

$\delta_{X^i}(X_t)$ est la mesure de Dirac et ces mesures ponctuelles sont qualifiées de « particules » du fait du caractère dynamique de la représentation de l'espace de probabilité. $p_i{}^t$ représentent les poids associés.

On approche la densité de probabilité *a priori* par : $P(X_O,\dots,X_t)\approx\frac{1}{N}\cdot\sum_{i=1}^{N}\delta_{X_{0,\dots,}^{i},X_t^{i}}(X_{0,\dots,}^{i}X_t)$.

On obtient l'approximation suivante de la loi conditionnelle qui est la base du filtrage « particulaire » :

$P(X_t|Z_0{}^t)\approx\sum_{i=1}^{N}p_t{}^i\,\delta_{X^i}(X_t)$.

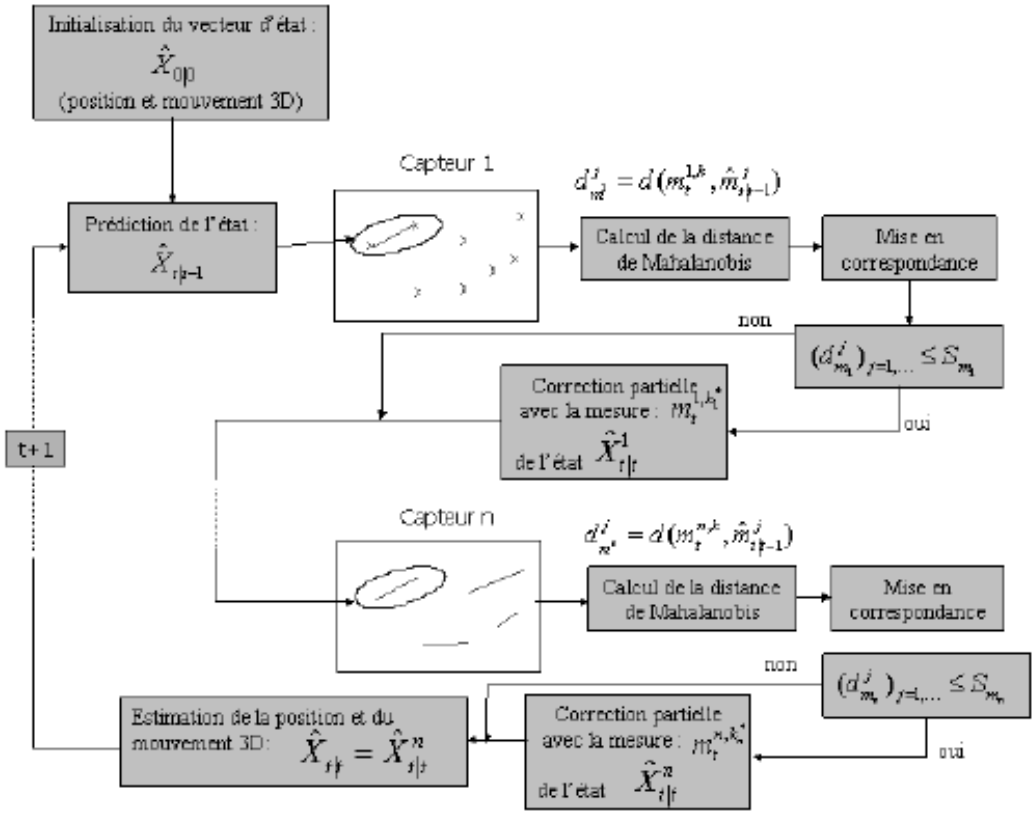


Figure 70 : Structure générale de la méthode proposée basée sur le filtrage de **Kalman étendu** (basé sur des capteurs synchrones) [Noyer J-C.].

Le noyau du filtre « particulaire » est composé de deux étapes :

-Une étape d'évolution des particules X_t^i ;

-Une étape de calcul des pondérations $p_t{}^i$ à partir des équations des capteurs.

Le déroulement du filtre « particulaire » est le suivant (cf. figure 71) :

1. Initialisation : Comme tout filtre récursif, on affecte les N particules X_0^i dans l'espace d'état en fonction de la probabilité *a priori* $P(X_0)$;
2. Évolution : Chaque particule X^i se voit affecter sa propre dynamique dictée par l'équation d'évolution ;
3. Pondération : Chaque particule évolue librement dans l'espace d'état et l'étape de pondération permet d'évaluer la probabilité associée à la région explorée. Un poids $p_t{}^i\in[0,1]$ associé à chaque particule est calculé selon la loi des capteurs ou de manière récursive $p_t{}^i=f(p_{t-1}{}^i)$;
4. Estimation : Chaque particule permet de dessiner une approximation de la loi conditionnelle $P(X_t|Z_0{}^t)$ et l'estimateur est alors construit par : $\hat{X}_{t|t}=\sum_{i=1}^{N}p_t{}^i\,X_t^i$

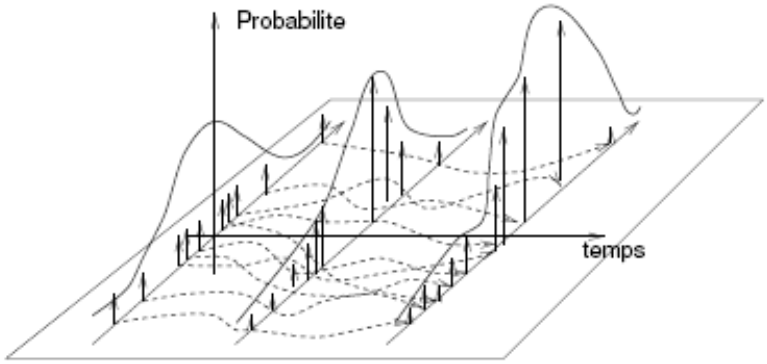


Figure 71 : Représentation schématique de l'évolution du réseau « particulaire » [Noyer J-C.].

Initialement les particules sont distribuées selon la loi $P(X_0)$ et sont équiprobables. Elles évoluent ensuite aléatoirement selon l'équation de la dynamique et la pondération permet de dessiner une discrétisation de la densité de probabilité conditionnelle. On voit aussi sur la figure 71 la nécessité d'introduire une étape supplémentaire de régulation de l'évolution des particules (« **redistribution** »), car un nombre croissant de particules ont un poids qui tend vers 0 lorsque t augmente.

3.1.1 Fonction de redistribution

Le filtre « particulaire » présente parfois certaines pondérations très faibles, rendant leur contribution très faible. Une solution consiste à réaffecter les particules de poids faible vers des régions plus probables de l'espace d'état, caractérisées par les particules de poids forts, donnant naissance à plus de particules tandis que celles de poids faibles disparaissent. C'est l'étape de « redistribution ». On utilise la fonction de répartition, solution du problème d'estimation, pour retirer chaque particule selon cette loi. Les particules (X_t^i, p_t^i) pour proposent ainsi un ré échantillonnage de cette loi de probabilité, puisqu'elles dessinent une discrétisation de cette densité de probabilité par des mesures ponctuelles de Dirac : $P(X_t|Z_0^t)=\sum_{i=1}^{N} p_t^i \delta(X_t^i)$ avec $Z_0^t=\{Z_0,...,Z_t\}$.

Les nouvelles particules donnent un nouvel échantillonnage de la densité de probabilité conditionnelle; leurs poids sont réinitialisés à 1/N.

3.1.2 Estimation multi capteurs de la structure et du mouvement 3D

Les travaux de la thèse de Christophe Boucher [Boucher C.] ont conduit à une formulation globale du problème de fusion multi capteurs pour l'estimation de la structure et du mouvement 3D. La solution proposée repose sur un filtre centralisé de type Kalman étendu qui fusionne les mesures de tous les capteurs, assure le suivi de la forme et l'estimation des paramètres caractéristiques (position et mouvement). **La solution « particulaire » a permis par la suite d'éviter les problèmes liés à la linéarisation des équations d'état dans un cadre multi capteurs.** La fusion est centralisée et la description de l'objet sous forme de points et segments polyédrique pour les scènes d'intérieur. La solution « particulaire » reprend les équations d'état précédentes et la structure de filtre sous forme prédiction/correction :

1. Initialisation : les N particules X_0^i sont initialisées selon l'information *a priori* sur les paramètres initiaux de mouvement;
2. Évolution : chaque particule est animée de sa propre dynamique et cette étape permet de prédire les positions 3D des extrémités de chaque segment ainsi que les paramètres de mouvement;
3. Mise en correspondance : il s'agit de suivre les particules au cours du temps. On va pour cela se servir

des mesures prédites au moyen de chaque particule X_t^i . Grâce à l'équation de mesure, on prédit la mesure sur chaque capteur. Cette mesure prédite sert de base au calcul de la distance de Mahalanobis autorisant la mise en correspondance. Cette distance est fonction des mesures effectuées sur les capteurs et de la covariance de l'erreur de prédiction de la mesure pour le capteur j . Cette covariance calculée dans le filtre de Kalman doit aussi être évaluée pour le filtre « particulaire » au moyen de N particules;

4. Pondération : les N particules ont exploré librement l'espace d'états et chacune d'elles doit être évaluée à partir de la mesure retenue sur chaque capteur (par l'étape de mise en correspondance) :
   $p_t^{\,i} = f(p_{t-1}^{\,i})$ avec f une fonction des mesures obtenues sur chaque capteur et permettant la pondération des particules;
5. Estimation de la structure et du mouvement 3D : l'estimation « particulaire » est donnée par $\hat{X}_{t|t} = \sum_{i=1}^{N} p_t^{\,i} X_t^i$;
6. Redistribution : la solution « particulaire » de base se doit d'être complétée par une étape de « réaffectation » des particules dans l'espace d'état. Certaines particules peuvent en effet avoir un poids $p_t^{\,i}$ qui tend vers 0 lorsque t augmente et ne contribuent pas à l'estimateur $\hat{X}_{t|t} = \sum_{i=1}^{N} p_t^{\,i} X_t^i$.

La méthode de résolution « particulaire » repose sur un filtre unique qui suit les primitives au cours du temps et estime la structure et le mouvement 3D. Le filtre assure aussi la fusion multi capteurs via un traitement centralisé (cf. figure 72).

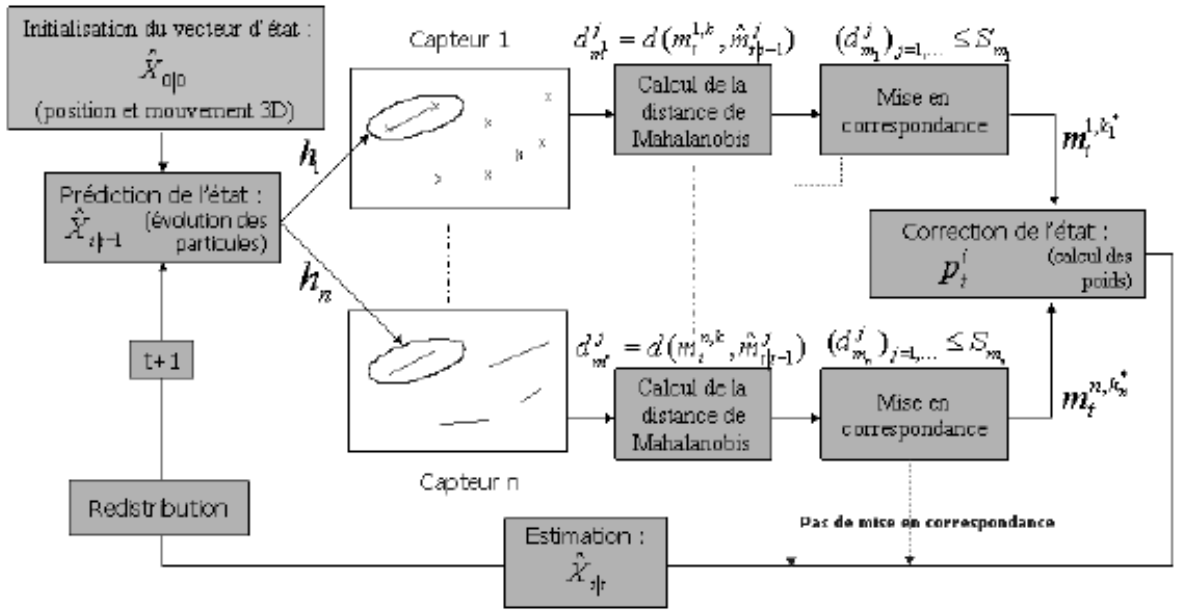


Figure 72 : Structure générale de la méthode proposée basée sur le **filtrage « particulaire »** [Noyer J-C.].

Le vecteur d'état est initialisé en position et mouvement. Les particules évoluent contribuant à prédire l'état. Chaque capteur évalue en parallèle la distance de Mahalanobis entre les mesures observées et les prédiction afin d'effectuer la mise en correspondance pour chacun d'eux, permettant de corriger l'état et d'estimer le nouvel état, ensuite redistribué et qui participera de nouveau à la prédiction. Cette solution est temps réel. **La**

solution initiale proposée par [Boucher C.] par filtrage de Kalman étendu présente une structure de fusion centralisée et un traitement séquentiel des mesures. L'inconvénient du filtre de Kalman étendu est qu'il procède à une linéarisation des équations d'état, ce qui ne garantit pas la convergence. D'où l'idée du filtrage « particulaire » qui peut traiter des modèles non linéaires et accéder au temps réel. Tout type de capteur peut être intégré.

3.2 Fusion multi capteurs pour l'estimation des positions et mouvement 3D et suivi 3D : une approche dense

Le filtrage de Kalman étendu a été remplacé par un filtrage non-linéaire, le filtrage « particulaire » qui a montré un bon suivi et une estimation correcte. Mais c'est une méthode basée primitives et ses résultats sont dépendants de l'extraction des primitives, donc de la segmentation (en profondeur et en intensité).
Parmi les approches issues de l'analyse de scènes, outre celles basées sur le suivi de primitives, nous avons les méthodes issues des données mesurées par les capteurs [Lanvin P.]. L'objet est alors décrit de manière **ponctuelle** et non plus par des primitives extraites. L'intérêt de ce genre d'approche réside dans l'absence de pré traitement des données, étape pouvant générer des erreurs dans la localisation des primitives. Une description a priori de la forme à poursuivre est utilisée. Le filtre doit alors assurer :
-La détection de la forme sur chaque mesure;
-Le suivi temporel;
-L'estimation des paramètres de position et de mouvement;
-La fusion des données issues des plusieurs capteurs.
L'approche adoptée par [P.Lanvin] repose sur un filtre « particulaire » adapté (à la forme). Ce genre d'approches existe dans de nombreuses applications de scènes naturelles (suivi de cibles, télésurveillance, etc.) où on a une information *a priori* sur la forme à suivre (avion, voiture, etc.). Contrairement aux approches type primitives qui font peu d'hypothèses sur l'objet, l'approche proposée exploite **l'information *a priori* disponible sur le type d'objet suivi**. Cela permet d'améliorer les performances en suivi, notamment lors de faibles rapports signal à bruit.

Trois problématiques dans le stage de DEA et la thèse de [Lanvin P.] ont été poursuivis :

1. Suivi d'objets et estimation des paramètres de position et de mouvement 2D à partir d'une séquence d'images monoculaire;
2. Suivi d'objets et estimation des paramètres de position et de mouvement 3D par une approche monoculaire : étendre la modélisation au cas des objets 3D pour le suivi;
3. Suivi d'objets et estimation des paramètres de position et de mouvement 3D dans un contexte multi capteurs, augmentant ainsi la robustesse des traitements.

L'ensemble des solutions s'intègre dans la théorie du filtrage non linéaire dont la solution est le filtre **« particulaire » centralisé adapté à la forme suivie**. Sa capacité à traiter des modèles non linéaires autorise une telle modélisation globale du problème d'estimation et de suivi d'objets. Le filtre peut réaliser le suivi d'un objet grâce à divers modèles génériques de formes et détecter à chaque instant l'hypothèse de forme la plus adaptée.

3.2.1 Suivi d'objets et estimation des paramètres de position et de mouvement 2D à partir d'une séquence d'images monoculaire

[Noyer J-C.] développe une méthode de suivi d'objets dans une séquence d'images d'intensité, par une description dense de la forme. Ce problème d'estimation est modélisé sous forme d'équations d'état caractérisant l'évolution de la forme (suivi bidimensionnel) dans la séquence et caractérisant l'observation associée (estimation des paramètres caractéristiques-position, mouvement).
Le problème d'estimation dynamique d'un processus aléatoire peut être modélisé par un système d'équations d'état qui décrit son évolution (équation de dynamique) et l'observation partielle que l'on en a (équation de mesure). L'objet est décrit par une forme générique dont les paramètres géométriques et cinématiques (position du centre de gravité, paramètres de mouvement) doivent être estimés. Le vecteur d'état est donc composé des coordonnées 2D du centre de gravité, du vecteur de translation, de la vitesse angulaire, de l'orientation

angulaire de l'objet et du facteur d'échelle de la forme.
L'objet évolue selon un mouvement rigide 2D (rotation et translation) ou affine dont les paramètres doivent être estimés. Il s'agit alors de trouver l'équation de dynamique de l'état.

[Noyer J-C.] cherche à réaliser le suivi d'objet dans une séquence d'images monoculaire. L'idée est de suivre directement l'objet sur l'image à niveaux de gris, sans détection, et sans réaliser de pré traitements usuels (par exemple extraire des primitives). La méthode qui consiste à extraire des primitives ne nécessite pas d'*a priori* sur la forme suivie, mais elle peut donner des erreurs de localisation de la primitive. Dans la thèse de [Lanvin P.], une connaissance *a priori* sur le type de forme suivie permet de décrire l'objet par un modèle à niveau de gris dont les paramètres de position et de mouvement doivent être estimés, rendant le suivi plus robuste.

3.2.2 Estimation des positions et mouvement 2D par filtrage « particulaire »

Une telle méthode évite de pré traiter l'image en assurant une prise en compte optimale des informations fournies par le capteur. La forme est décrite par le vecteur d'état X_t qui caractérise les paramètres géométriques et cinématiques (équation de la dynamique), Z_t est caractérisée par l'équation de mesure.
La méthode de résolution est celle de la structure générale du filtre « particulaire » :

1. Initialisation : les N particules sont initialisées selon l'information *a priori* $P(X_0)$;
2. Évolution : chaque particule X^i évolue dans l'espace d'état selon le modèle de dynamique;
3. Pondération : la probabilité associée à chaque particule est calculée à partir de la mesure Z_t disponible à l'instant t (image). Les poids de pondération peuvent être vus sous l'hypothèse de bruit de mesure gaussien de moyenne nulle et covariance R;
4. Estimation : l'estimée « particulaire » est donnée par : $$\hat{X}_{t|t} = \sum_{i=1}^{N} p_t{}^i X_t^i$$
5. Redistribution : on procède à une étape de redistribution des particules en ré échantillonnant l'espace d'état à partir de la fonction de répartition.

Contrairement à l'approche primitive, il n'y a pas ici d'étape de mise en correspondance, du fait du défaut de segmentation. L'approche proposée traite l'ensemble des mesures disponibles dans l'image complète sans pré traitement. La mise en correspondance est réalisée par le calcul des poids de manière transparente, par un calcul de corrélation entre l'image mesurée et la reconstruction de l'image associée à la particule. Le filtre réalise le suivi et l'estimation des paramètres de mouvement et de structure.

3.2.3 Extension au problème de détection

Les résultats précédents dans la première partie de la thèse de [Lanvin P.] à base de modèle *a priori* de la forme, sont améliorés par la détection parmi un ensemble de modèles de formes, du modèle le plus probable.

3.2.3.1 Modélisation

On cherche une méthode de détection et suivi d'objets dans une séquence d'images monoculaire. La solution proposée détecte le modèle au meilleur suivi par rapport à la précision de l'estimation des paramètres de position et mouvement. Ce modèle adapté est construit à l'aide d'un filtre unifié qui résout le problème d'estimation/détection/suivi afin de minimiser les pertes d'information à chaque étape du traitement.

3.2.3.2 Solution « particulaire » du problème d'estimation-détection et suivi 2D

Le filtre doit estimer l'état X caractéristique de la forme, c'est--dire estimer conjointement les paramètres de position et de mouvement ainsi que le modèle de forme.
Le problème d'estimation détection est résolu grâce à un filtre adapté à chaque mode. A chaque instant, les probabilités du mode k permet de détecter le modèle le plus probable.
La solution « particulaire » proposée repose donc sur l'utilisation de m filtres adaptés (à chaque mode)

fonctionnant en parallèle. Chaque filtre « particulaire » a la structure habituelle :

1. Initialisation : les N particules sont initialisées selon l'information disponible *a priori* pour chaque filtre;
2. Évolution : les particules évoluent dans chaque mode selon le flot du système au moyen de N réalisations indépendantes;
3. Pondération : cette étape permet d'évaluer la probabilité associée à chaque particule grâce à la règle de Bayes. Elle utilise pour cela l'ensemble des mesures image à l'instant t pour construire le poids de la particule dans le mode k;
4. Estimation : on construit l'estimation « particulaire » associée au mode k;
5. Détection : chaque filtre adapté au mode k estime conjointement la position et le mouvement et parallèlement, on cherche à détecter l'hypothèse la plus probable;
6. Redistribution : chaque particule est redistribuée pour chaque mode k selon la fonction de répartition de la loi $P(X_{k,t}|Z_0{}^t)$.

La solution du filtre « particulaire » permet d'estimer conjointement les paramètres de position et de mouvement selon chaque hypothèse, en détectant à chaque itération le modèle le plus probable. De plus le filtre traite directement les images à niveau de gris, ce qui évite la détection des primitives pouvant engendrer des erreurs de localisation de la forme.

[Noyer J-C.] a ainsi proposé une méthode de détection et suivi 2D de formes dans une séquence d'images monoculaire. Une formulation conjointe a conduit au filtrage « particulaire », adapté à la modélisation non linéaire, que le Kalman étendu n'aurait pas permis de traiter sans s'affranchir de la segmentation. Il n'y a en effet pas de pré traitement, pas de suivi de primitives entraînant des erreurs de localisation. Le problème d'estimation hybride posé est résolu par n filtres adaptés aux différents modèles de forme qui évoluent en parallèle. On calcule donc conjointement la probabilité associée à chaque mode et une estimation des caractéristiques géométriques et cinématiques associées. Les hypothèses de forme peuvent être des hypothèses de modèles dynamiques ou un mélange des deux, permettant de détecter à la fois le modèle d'évolution le plus probable mais aussi le modèle de forme adapté.

3.2.3.3 Suivi d'objets et estimation des positions et mouvement 3D par une approche monoculaire

Des résultats intéressants ont été obtenus par une description bidimensionnelle de l'objet pour l'estimation des paramètres de forme et de la détection. La modélisation retenue repose sur l'utilisation d'un **mouvement rigide** de la forme. Bien que valide dans bon nombre d'applications, elle ne convient pas aux modèles déformables. La modélisation étendue au cas 3D permet une meilleure prise en compte des déformations de la forme. On représente la forme à suivre par l'ensemble de ses paramètres géométriques et cinématiques, résumées dans X_t . Son mouvement est modélisé par un modèle rigide 3D (rotation/translation). Un modèle *a priori* de la forme, sous la forme d'objets 3D, est défini pour le suivi dans la séquence d'images monoculaires. La fonction de dynamique est non linéaire. L'équation de mesure conserve la même forme que dans le cas 2D.

3.2.3.4 Reconstruction 3D et estimation du mouvement 3D par filtrage « particulaire »

Une partie des travaux de thèse de [Lanvin P.] fut la mise en oeuvre d'une méthode de résolution « particulaire » pour le suivi de forme dans une séquence d'images monoculaire, mais aussi pour estimer les paramètres de position et de mouvement 3D. La modélisation globale du problème d'estimation et de suivi permet de proposer un filtre non linéaire qui résolve de manière conjointe ces problèmes. Le filtre « particulaire » est solution de ce problème d'estimation (du vecteur d'état X_t).

3.2.3.5 Extension au cas de la détection d'objets 3D

La description bidimensionnelle précédente ne permettait pas de prendre en compte très précisément les légères déformations de l'objet dans la séquence, la problématique s'étend au cas 3D. On introduit m hypothèses

H_k de forme 3D. M filtres « particulaires » sont adaptés à chaque hypothèse de la forme 3D. L'objet est modélisé en 3D et le filtre cherche à adapter la forme retenue à l'objet à suivre.

3.2.3.6 Estimation des positions et mouvement 3D dans un contexte multi capteurs

[Noyer J-C.] étudie, à la suite du cas multi capteurs, le problème d'estimation des positions et mouvement 3D à partir d'un système multi capteurs. On pourra ainsi lever l'ambiguïté au problème de reconstruction 3D. En effet, [Lanvin P.] a autorisé la reconstruction 3D avec un modèle *a priori* de la forme 3D. Ceci constitue une hypothèse légitime dans certains domaines applicatifs comme la **vidéo surveillance**, le transport. Ils peuvent intégrer un modèle de forme 3D *a priori* permettant d'assurer un suivi efficace de l'objet et sa reconstruction. Le déplacement de l'objet 3D est modélisé par un mouvement rigide 3D dont les paramètres (rotation, translation) doivent être estimés. Comme pour l'approche primitive, c'est une structure de fusion centralisée prenant en compte les mesures issues des capteurs en un noeud central de traitement, qui est retenue.

3.3 Fusion multi capteurs par filtrage « particulaire » pour la reconstruction 3D, l'estimation du mouvement 3D et le suivi d'objets 3D

La dernière partie de la thèse de [Lanvin P.] concerne l'utilisation de la fusion multi capteurs pour l'estimation des positions et mouvement 3D d'objets et le suivi. La solution repose sur un filtre « particulaire » **unique** qui fait l'estimation de ces paramètres et la fusion centralisée des informations des capteurs. La structure du filtre est composée des étapes suivantes :

1. Initialisation : les particules $(X_0^i)_{i=1,\ldots,N}$ représentent N réalisations aléatoires de la loi $P(X_0)$ et sont donc initialisés selon l'information disponible *a priori*;
2. Pondération : les particules sont pondérées en fonction des mesures obtenues sur chaque capteur, dans le cas d'un système de fusion centralisée;
3. Estimation : l'estimation multi capteurs des positions et mouvement 3D;
4. Redistribution.

L'originalité de l'approche réside dans l'utilisation d'une modélisation d'état pour décrire ce problème d'estimation. Les équations d'état permettent de modéliser l'évolution des paramètres caractéristiques de forme (position, mouvement, facteurs d'échelle, ..) et leur lien avec la mesure. La nature non linéaire de ces équations a conduit au filtrage « particulaire ». La solution repose sur un filtre unique qui réalise non seulement l'estimation des paramètres caractéristiques de l'objet (position, mouvement), mais également la fusion des informations issues de chaque capteur dans un schéma de **fusion centralisée**. Cette méthode est bien adaptée à la détection de formes car elle propose une formulation globale du problème d'estimation-détection en évitant toute décorrelation des traitements. Les mesures délivrées par plusieurs capteurs sont intégrées. Les capteurs ne sont pas nécessairement synchrones et peuvent être de nature physique différentes.

Conclusion

Au vu des travaux précédents exposés dans cette bibliographique, il nous semble judicieux de proposer une approche multi caméras, dans un réseau bayésien, chaque caméra représentée par un noeud du réseau. Les messages sont envoyés d'une caméra à l'autre par propogation de croyance, symbolisant la croyance qu'une personne vue dans une caméra puisse se trouver un instant plus tard dans le champ de l'autre caméra, en fonction de la configuration des caméras, et de l'analyse de scène.

Il faut pour cela un module « **bas niveau** » qui détecterait le mouvement, à base de soustraction de fond et de détection de mouvement, par flot optique par exemple. Un autre module procéderait à la ré identification des personnes lors de leur passage d'une caméra à l'autre. Il faut pouvoir identifier chaque personne et suivre chacune d'elles individuellement. Pour cela un modèle d'apparence, silhouette et couleur des vêtements et de la teinte chair, pourrait nous y aider. Une fois la ré identification possible, il faut pouvoir suivre les trajectoires de chaque personne à travers le réseau de caméras. A cette fin, il est nécessaire d'avoir un modèle de la scène avec le positionnement des caméras dans la scène et les unes par rapport aux autres, le positionnement des rayons et des allées, les zones « aveugles » (non visitées par les caméras) où peuvent se positionner des individus mal intentionnés, le contenu des rayons (attractif?), le parcours possible à l'intérieur du magasin. Il est également utile d'avoir une « **gestion haut niveau** », le superviseur, contenant une description sémantique de l'action de chaque individu suivi à chaque instant. Il faut en effet pouvoir **coupler une approche globale par suivi d'une caméra à l'autre (entre les zones du magasin) avec une analyse locale dans chaque caméra**. L'approche globale serait gérée par le superviseur, en tenant compte de la trajectoire totale de la personne suivie, laquelle trajectoire serait obtenue à partir des informations bas niveau issues d'une analyse locale dans chaque caméra. On pourrait envisager une architecture à tableau noir avec une base de données bas niveau. Ainsi, un score (probabilité) serait attribué à chaque action, déterminant ainsi le degré de « dangerosité ».

Un système de « perception active », où les caméras seraient activées (zoom, translation, rotation) individuellement, et à tour de rôle, en fonction de la description sémantique des personnages dans la scène n'est cpendant pas envisageable. En effet, il semble difficile de mobiliser une caméra sans risquer, par un zoom, de perdre une partie de l'observation de la scène.

Annexe 1 – Minimisation du critère du MAP

Algorithmes de minimisation du critère du MAP

Pour minimiser le critère du MAP, il existe divers algorithmes de minimisation :
-Les algorithmes **stochastiques**, de type recuit simulé (recuit avec dynamique de « Metropolis », échantillonneur de «Gibbs avec recuit »), les algorithmes génétiques, les algorithmes déterministes (les modes conditionnels itérés « ICM » « Iterated Conditional Modes », la non-convexité graduelle « GNC » « Graduated Non-Convexity », le recuit en champ moyen « MFA » « Mean Field Annealing »).
-Les algorithmes **déterministes** sont plus rapides que ceux stochastiques mais peuvent être piégés dans un **minimum local** de l'énergie du critère du MAP au lieu d'un minimum global assuré pour l'algorithme stochastique.

Les algorithmes stochastiques d'optimisation sont une analogie avec le **procédé de recuit en métallurgie et en verrerie**. Le matériau est porté à très haute température et refroidit très lentement, afin d'obtenir la meilleure cristallisation, c'est-à-dire l'état le plus ordonné possible. Le recuit simulé appartient à la famille des algorithmes de **relaxation stochastique de type Monte-Carlo**. A chaque pas de l'algorithme, la solution précédente est vue comme une perturbation aléatoire. Le recuit simulé permet d'éviter de converger vers un minimum local alors que les algorithmes déterministes itératifs basés sur la minimisation du gradient suivent la décroissance de la fonction à minimiser et peuvent converger sur un minimum local.
Dans l'algorithme du « recuit simulé », un paramètre appelé température est à l'origine de la probabilité d'accepter une croissance de la fonction à minimiser. Initialement, le système est porté à une très haute température et le nouvel état du système est évalué. La haute température permet d'accepter tous les états possibles du système. La température est progressivement diminuée selon une loi de refroidissement. Le nouvel état est calculé, et ainsi de suite, jusqu'à atteindre une température qui permette la convergence vers un état d'équilibre, **minimum global** de la fonction à minimiser.

1 Algorithme du recuit simulé

Le recuit simulé permet de trouver les configurations les plus probables correspondant aux états d'énergie minimale. Ces réalisations sont obtenues par l'algorithme de « recuit simulé avec dynamique de Metropolis » et « l'échantillonneur de Gibbs avec recuit », qui permettent d'échantillonner selon la loi de probabilité de Gibbs associée au champ de Markov. Ces deux algorithmes synthétisent donc les réalisations d'un champ de Markov. Étant donné un champ de Markov, on réalise le tirage d'une configuration image en suivant la loi de probabilité de Gibbs caractéristique de ce champ. Dans les années 50, [Metropolis N. et al.] a mis au point un algorithme de relaxation probabiliste issu de la physique statistique. Une suite d'images est construite qui sont des tirages selon la loi du champ de Markov après un grand nombre d'itérations. A l'initialisation, la température étant élevée, tous les états sont équiprobables. Lorsque la température diminue, la configuration la plus probable correspond au minimum global de l'énergie. Le matériau est un cristal parfait quand la température tend vers zéro.

2 Cas d'une image

Dans le cas d'une image, on considère que la grille des pixels représente les atomes du matériau et les niveaux de gris leurs états possibles. Une image est une configuration X , à laquelle est associée l'énergie $U(x)$ du système correspondant à la configuration x , une probabilité de réalisation $P(X=x)=\frac{1}{Z}\cdot\exp(\frac{-(U(x))}{T})$ avec $Z=\sum_x \exp(\frac{-(U(x))}{T})$ la fonction de partition du système, et $T>0$. Le terme de température provient de l'analogie avec la physique statistique. La variation de température est supposée assez lente pour que le système évolue vers un état le plus ordonné possible et qui corresponde à l'équilibre thermique à cette température, et à cette température le système se trouve dans l'état X dont la probabilité est donnée par $P(X=x)=\frac{1}{Z}\cdot\exp(\frac{-(U(x))}{T})$.
A chaque température, on effectue une petite perturbation au système jusqu'à ce qu'il se trouve dans son état

d'équilibre « thermique », par exemple, modifier légèrement la valeur d'un pixel en lui ajoutant une valeur aléatoire appelée grain. A chaque pas de l'algorithme, on génère aléatoirement une nouvelle perturbation candidate. Si cette solution produit une diminution de l'énergie $U(x)$ ($\nabla U(x)<0$), la solution est acceptée, sinon elle est acceptée selon la probabilité $P(X=x)$:

$P(X=x)=\{\exp(-\Delta T),\ si\ -\Delta U>0\ \ ou\ bien\ 1\ ,si-\Delta U<0\}$, T est la température et ΔU est la variation d'énergie suite à la perturbation.

A température grande, il y a plus de chances d'accepter la configuration qu'en cas de faible température car on va chercher à baisser la température. Cet algorithme nécessite beaucoup d'itérations pour converger du fait que les perturbations sont générées aléatoirement, mais la convergence vers un minima local est évitée grâce à l'acceptation des configurations d'énergie supérieure.

3 Algorithmes de Gibbs et Metropolis

L'échantillonneur de Gibbs est un algorithme proposé par [Geman S., Geman D.], il repose sur la construction itérative d'une suite d'images. Cependant, l'analogie avec le processus physique de recuit est moins évidente car on n'attend pas que le système se stabilise à chaque température. **La différence principale avec l'algorithme de Metropolis** réside dans la génération des perturbations. En effet, au lieu de générer des perturbations de manière aléatoire et de décider ensuite si elles sont acceptées ou non, les perturbations sont générées selon des fonctions de densité de probabilité conditionnelles locales, dérivant d'une distribution de Gibbs. A la convergence, les images générées sont des réalisations tirées selon la loi de Gibbs globale :

$$P(X_s=x_s/V_s)=\frac{(\exp(-U_s(x_s/V_s)))}{(\sum_{\xi\in E}\exp(-U_s(\xi/V_s)))}\ .$$

A l'itération n en partant de l'itération $n-1$, on choisit un site s selon une loi uniforme ou un balayage de l'image, la condition étant de balayer tous les sites un très grand nombre de fois. Selon la configuration des voisins V_s pour l'image $x^{(n-1)}$, on calcule la probabilité conditionnelle locale : $P(X_s=x_s/V_s)=\frac{(\exp(-U_s(x_s/V_s)))}{(\sum_{\xi\in E}\exp(-U_s(\xi/V_s)))}$. Enfin, on met à jour le site s par tirage aléatoire selon la loi $P(X_s=x_s/V_s)$. On considère que l'algorithme a convergé après un grand nombre d'itérations ou lorsque le nombre de changements est faible.
Cet algorithme construit une suite d'images $x^{(n)}$ qui sont les observations d'une suite $X^{(n)}$ de champs aléatoires formant une chaîne de Markov. Lorsque la séquence balaye chaque site une infinité de fois, on a le théorème suivant : $\forall x^{(0)}\ \ \forall x\in\Omega\ \lim_{n\to\infty}P(X^{(n)}=x/X^{(0)}=x^{(0)})=P(x)$, P est la mesure de Gibbs associée au champ de Markov. Après un grand nombre d'itérations, les images $x^{(n)}$ générées sont des réalisations de la loi globale $P(X)$ $\forall x^{(0)}$ la configuration initiale.
L'algorithme de Gibbs est connu sous le terme d'algorithme de « relaxation », car il met à jour de façon successive des sites et de façon probabiliste du fait du tirage aléatoire.

Par rapport à l'algorithme de Gibbs, l'algorithme de Metropolis tire au sort le nouveau descripteur (niveau de gris dans notre cas) au lieu de considérer la loi définie par tous les descripteurs. Les balayages des sites et le critère d'arrêt sont similaires entre les deux algorithmes. Cependant, l'algorithme de Metropolis est plus rapide à chaque étape que l'échantillonneur de Gibbs, mais la convergence peut être plus lente car l'algorithme de Metropolis a un taux d'acceptation inférieur à 1, alors que l'échantilonneur de Gibbs accepte toutes les transitions.

Une **distribution de Gibbs** est une probabilité $P(X=x)=\frac{1}{Z}\cdot\exp(\frac{-(U(x))}{T})$. Pour une température

infinie, on démontre que tous les états sont équiprobables (converge vers la probabilité uniforme). Pour une température qui tend vers 0, on démontre que la probabilité est uniformément distribuée sur les minima globaux de l'énergie, c'est-à-dire sur les configurations les plus probables. Ceci est la **base de l'algorithme de recuit simulé.**

4 Fonctionnement de l'algorithme du recuit simulé

Cet algorithme a pour objectif **non plus l'échantillonnage, mais la recherche de la configuration d'énergie minimale d'un champ de Gibbs**. C'est un algorithme de simulation itératif qui établit la solution progressivement. **L'algorithme de recuit simulé est le suivant** avec n le numéro de l'itération :

1. On choisit une température initiale $T^{(0)}$ assez grande;
2. On choisit une configuration initiale quelconque $x^{(0)}$;
3. A l'étape n , on simule une configuration $x^{(n)}$ pour la loi de Gibbs d'énergie $\frac{(U(x))}{T^{(n)}}$ à partir de la configuration $x^{(n-1)}$; la simulation a lieu soit par l'échantillonneur de Gibbs soit par l'algorithme de Metropolis. On balaie l'image à la température $T^{(n)}$. Puis on fait diminuer la température lentement.
4. On arrête si le changement est faible.

L'algorithme de recuit simulé, contrairement à l'échantillonneur de Gibbs et à l'algorithme de Metropolis qui en échantillonnant selon la loi de Gibbs peuvent donner toutes les configurations possibles, fournit des images uniques correspondant aux minima globaux de l'énergie. L'algorithme de recuit simulé atteint un **minimum global** car il permet des remontées en énergie. En faisant décroître la température assez lentement pour ne pas rester piégé dans un minimum local de l'énergie, les sauts d'énergie sont progressivement supprimés en se rapprochant de l'optimum global.
Les algorithmes stochastiques du type recuit convergent en probabilité vers un minimum global du critère du MAP, indépendamment de la configuration initiale. Si l'énergie du MAP est une somme de termes locaux, on utilise l'échantillonneur de Gibbs avec recuit, sinon on utilise le recuit simulé avec l'algorithme dynamique de Metropolis [Chadhury S., Subramanian S., Parthasaraty G.]. Cependant, les algorithmes de recuit ont un coût de calcul important.

5 Algorithme ICM Iterated Conditional Mode

L'algorithme de recuit simulé est très long en calculs, puisqu'il faut générer beaucoup de configurations en même temps que la température décroît. L'ICM (« Iterated Conditional Mode ») proposé par [J. Besag] est plus rapide mais il n'assure pas de convergence vers un minimum global. Cet algorithme est itératif, modifiant à chaque étape les valeurs x_s de l'ensemble des sites de l'image, mais la modification est maintenant déterministe. Cet algorithme ne permet d'atteindre qu'un minimum local de l'énergie, la transition d'une configuration à une autre n'étant possible que si l'énergie est inférieure. Les algorithmes déterministes tels que descente de gradient, gradient conjugué ou modes conditionnels itérés (ICM) risquant de rester piégés dans un minimum local, des algorithmes ont été développés afin de fournir des estimées de bonne qualité. Citons le Non-Convexité Graduelle (GNC « Graduated Non-Convexity ») et le recuit en champ moyen (MFA « Mean Field Annealing »).
L'algorithme ICM est appelé aussi « recuit gelé » ou « Metropolis gelé », ou « Gibbs gelé », car c'est un cas particulier de l'algorithme de Metropolis ou de l'échantillonneur de Gibbs : la probabilité d'accepter des perturbations qui augmentent l'énergie est toujours nulle. Cet algorithme est similaire à l'échantillonneur de Gibbs, mais on choisit pour chaque pixel la valeur maximisant la probabilité conditionnelle locale, au lieu de tirer une valeur aléatoire d'une distribution de probabilité conditionnelle.
On construit, à partir d'une configuration initiale $x(0)$ une suite d'images $x(n)$ convergeant vers une approximation du MAP $\hat{x}$ recherché. Une itération est une mise à jour d'un site, un tour correspond à la visite de tous les sites de l'image, et une étape est l'accomplissement d'un tour.
Le déroulement de l'étape n s'effectue en parcourant tous les sites et en chacun d'eux, on effectue deux

opérations :

1. On calcule les probabilités conditionnelles locales pour toutes les valeurs possibles de λ dans E du site : $P(X_s=\lambda/\hat{x}_r(k),\ r\in v_s)$;
2. On met à jour la valeur de λ qui maximise la probabilité conditionnelle locale $\hat{x}_s(k+1)=Argmax_\lambda P(X_s=\lambda/\hat{x}_r(k),\ r\in v_s)$.

On arrête quand le nombre de changements d'une étape à l'autre devient faible. L'énergie globale de la configuration $\hat{x}$ diminue à chaque itération. L'algorithme ICM converge plus rapidement que les algorithmes stochastiques de type recuit simulé, mais sa qualité dépend de l'initialisation car il converge vers un minimum local, étant donné qu'il n'accepte que les perturbations de variation d'énergie $\nabla U(x)$ négative . L'ICM ressemble à une descente de gradient (l'énergie baisse à chaque itération) ou à un recuit simulé gelé à température nulle (d'où sa dénomination de « recuit gelé »), et donc peut rester bloqué dans le minimum local le plus proche de l'initialisation. Tandis que le recuit simulé, grâce aux remontées en énergie qu'il se permet via le paramètre température, permet d'atteindre le minimum global. Il faut donc choisir une estimée initiale convenable. Mais si le nombre d'états possibles du système est petit, comme le déplacement maximal en estimation de mouvement ou le nombre d'étiquettes en segmentation, l'ICM converge très rapidement.

Nous allons expliquer des algorithmes de simulation permettant de générer des réalisations d'un champ de Markov quelconque, **pour des applications en segmentation**.

6 Cas de la segmentation

Le problème est modélisé dans un **cadre bayésien**. On suppose une image y et une réalisation d'un champ aléatoire Y . Le champ markovien est ici décrit sur un autre espace de configurations que Y car seules quelques étiquettes sont considérées, celles correspondant aux diverses classes recherchées. Le processus de passage de X , le champ des étiquettes ou des labels pour la segmentation (le champ des intensités pour la restauration), à Y ne décrit pas le processus d'acquisition mais l'apparence des classes dans l'image. Nous cherchons une réalisation x de l'image segmentée (ou restaurée dans le cadre de la restauration), modélisée par un champ de Markov X . Le champ X est la réalité terrain tandis que le champ Y est l'image bruitée. La segmentation (ou la restauration) permet de remonter à une réalisation de X à partir de l'observation de l'image bruitée y . Il s'agit alors d'un champ de Markov caché pour X , ou de données incomplètes puisque y n'est pas une réalisation de X . Grâce au **critère du maximum à posteriori**, on recherche la configuration $\hat{x}$ maximisant la probabilité suivante définie par la règle de Bayes :

$$P(X=x/Y=y)=\frac{(P(Y=y/X=x)\cdot P(X=x))}{(P(Y=y))} \quad ,$$

$P(Y=y/X=x)$ correspond à l'observation des données image (probabilité de réalisation d'une configuration connaissant son étiquetage, c'est-à-dire la classe de chaque pixel). On fait l'hypothèse courante d'indépendance conditionnelle des pixels sites les uns par rapport aux autres (bruit non corrélé par exemple), et que le niveau de gris y_s en un site s ne dépend que de l'étiquette x_s en ce site :

$$P(Y=y/X=x)=\prod_s P(Y_s=y_s/X_s=x_s) \quad .$$

Les valeurs des probabilités conditionnelles sont données par l'histogramme conditionnel des niveaux de gris pour une classe donnée. On fait de plus l'hypothèse sur le champ X de markoviennité : $P(X=x)=\frac{(\exp(-U(x)))}{Z}$.

On démontre que $P(X=x/Y=y)\propto\exp(-\upsilon(x/y))$, avec $P(X=x/Y=y)\propto\exp(-\upsilon(x/y))$, avec $\upsilon(x/y)=-\sum_{s\in S}\ln(p(y_s/x_s))+\sum_{c\in C}U_c(x)$. Donc la distribution à posteriori est une distribution de Gibbs et le champ des étiquettes X conditionnellement à y est aussi un champ de Markov (**théorème de Hammersley-Clifford**), et d'énergie de Gibbs $\upsilon(x/y)$. Le terme d'ordre 1 exprime la cohérence des données (le niveau de gris doit correspondre à la classe), et le terme d'ordre 2 la contrainte de régularisation. Il est ainsi possible de simuler des réalisations de ce champ à l'aide de l'échantillonneur de Gibbs ou de

l'algorithme de Metropolis. Il est nécessaire de déterminer les états d'énergie minimale correspondant au maximum de la probabilité d'un champ markovien. Autrement dit, la configuration x recherchée est celle qui maximise la probabilité à posteriori, c'est-à-dire la réalisation la plus probable du champ de Gibbs ou celle qui minimise l'énergie $\upsilon(x/y)$. **L'algorithme de recuit simulé permet de trouver ces configurations.**

Prenons un exemple, et faisons l'**hypothèse de la présence d'un mouvement dominant**, par exemple dans le cas du mouvement du fond. Dans ce cas, il faut segmenter les objets de l'avant-plan. Le mouvement dominant est alors recherché pour toute l'image courante I(t), c'est le modèle global A. L'image précédente I(t-1) est compensée : chaque point de I(t-1) est déplacé grâce aux paramètres du modèle trouvé A. L'erreur de compensation « **Deplaced Frame Difference** » est calculée pour chaque pixel. Les objets non conforme au mouvement du modèle dominant sont mal compensés. Ils seront détectés par seuillage de l'erreur de compensation, mais il existe aussi des méthodes markovienne plus complexes [Odobez J.M., Bouthemy P. 94].

Annexe 2 – Filtrage particulaire

1 Le filtre particulaire

Le filtre de Kalman n'est pas optimal dans les applications de « suivi visuel » (suivi d'un objet via les caractérstiques d'apparence) car les hypothèses de normalité du bruit de mouvement et d'observation ne sont pas toujours satisfaites.

Le suivi probabiliste est réalisé grâce à la couleur. Il s'agit d'estimer le vecteur d'état composé de la position et du facteur d'échelle de la boîte englobante de l'objet. Les situations qui engendrent le décrochage du filtre de Kalman sont les suivantes :

1. Lorsque le fond de la scène présente une apparence similaire à l'objet suivi, connu sous le nom de « clutter ». Sur la figure 73a, les cartons du fond de la scène présentent une apparence similaire à la peau du bébé, conduisant à un décrochage à partir de l'image 50;
2. Lorsqu'il y a une occultation de l'objet suivi. Sur la figure 73b, le suivi est complètement décroché à partir de l'image 103.

Dans le premier cas, l'hypothèse de normalité du vecteur d'état conditionnellement aux observations n'est pas vérifiée. Lorsque l'objet suivi se trouve dans une zone similaire d'un point de vue de l'apparence, plusieurs positions du vecteur d'état correspondent au modèle et en deviennent équiprobables.
Dans le second cas, lors d'une occultation, l'information d'apparence de la personne suivie n'étant plus visible, la répartition de l'état sera multi modale et l'hypothèse de normalité conditionnellement aux observations n'est plus vérifiée.

(a) Fond d'apparence similaire à 'objet d'intérêt et variation de point de vue

(b) occultation

Figure 73 : Limite du filtrage de Kalman dans le cas du suivi visuel ([Perez P., Hue C., Vermaak J., Gangnet M.], [Thome N.]).

Le filtre de Kalman propose une solution optimale lorsque l'hypothèse de normalité est vérifiée, mais dans le cas contraire, la recherche déterministe par filtre de Kalman risque de dériver vers une mauvaise solution, sans aucune chance de retrouver la personne suivie quelques images plus loin.
D'autre part, le filtre de Kalman suppose la recherche de l'observation dont la corrélation avec le modèle est la

plus grande. Ceci signifie qu'en cas de décrochage, il n'est plus possible de rattraper le suivi.
Une alternative au filtrage de Kalman a été proposée afin d'éviter ses limites dans le cas du suivi visuel. Au lieu d'utiliser une loi de paramètres *a priori* connus pour la densité de probabilité de l'état, on approche la distribution recherchée par simulation numérique. Les méthodes de Monte-Carlo permettent de trouver une solution, mais dans le cas du suivi visuel, le filtrage particulaire est un exemple très connu d'application des méthodes de Monte-Carlo, pour l'estimation du vecteur d'état d'un système Markovien non linéaire et non gaussien.
Le filtrage particulaire est une méthode d'exploration de l'espace d'état du problème par des « particules » dont la dynamique évolue aléatoirement. L'ensemble des particules est distribué selon la probabilité du processus à estimer, conditionnellement aux observations délivrées par les capteurs. Comme cette méthode ne nécessite pas une résolution explicite des équations, elle est applicable dans le cas de non linéarité ou non gaussieneté.

Le but du filtrage particulaire, en tant qu'estimateur bayésien, est d'estimer récursivement la densité de probabilité *a posteriori* $p(x_k / z_{1:k})$ du vecteur d'état x_k à l'instant k conditionnellement sur l'ensemble des mesures $z_{1:k} = z_1, \ldots, z_k$. L'idée est d'approcher la distribution de probablité de l'état X de la personne suivie par un ensemble de n échantillons $x^{(i)}$ associés à des poids $\pi^{(i)}$:

$X = \{(x^i, \pi^i), i = 1, \ldots, n\}$, et $\sum_{i=1}^{N} \pi^{(i)} = 1$. Chaque échantillon $x^{(i)}$ est appelé une particule, représentant une instance de l'état X dans l'espace dans lequel il est défini. Le poids $\pi^{(i)}$ correspond à la probabilité que $X = x^{(i)}$.
A chaque instant k, la densité $p(x_k / z_{1:k})$ est approchée grâce à la distribution ponctuelle $p(x_k / z_{1:k}) \approx \sum_{i=1}^{N} \pi_k^{(i)} \delta(x_k - x_k^{(i)})$, et $\sum_{i=1}^{N} \pi^{(i)} = 1$, exprimant la sélection d'une « particule » $x_k^{(i)}$ avec la probabilité ou « poids » $\sum_{i=1}^{N} \pi^{(i)}$, $i = 1, \ldots, N$.
Les particules $x_k^{(i)}$ évoluent de façon stochastique dans le temps et sont échantillonnées selon une **fonction d'importance** qui a pour but d'explorer de façon adaptative les zones « pertinentes » de l'espace d'état.
L'initialisation de l'**algorithme générique du filtrage particulaire** consiste à définir un ensemble de particules pondérées décrivant la distribution *a priori* $p(x_0)$, en affectant des poids identiques $\pi_0^{(i)} = \frac{1}{N}$ à des échantillons $x_1^{(0)}, \ldots, x_N^{(0)}$ indépendants identiquement distribués (i.i.d) selon $p(x_0)$. La détermination de l'ensemble des particules pondérées $\{x_k^{(i)}, w_k^{(i)}\}$ associée à la densité *a posteriori* $p(x_k / z_{1:k})$ se fait en deux étapes :

1. Les $x_k^{(i)}$ sont échantillonnés selon la fonction d'importance $q(x_k / x_{k-1}, z_k)$;
2. Les poids $\pi^{(i)}$ sont ensuite mis à jour de façon à assurer la cohérence de l'approximation
$p(x_k / z_{1:k}) \approx \sum_{i=1}^{N} \pi_k^{(i)} \delta(x_k - x_k^{(i)})$

L'ensemble des particules pondérées a donc pour objectif d'estimer la densité de probabilité de l'état (cf. figure 74). Les n particules évoluent en parallèle, et chaque particule progresse en fonction des mesures fournies par les capteurs à chaque instant, en simulant une « trajectoire » possible. Cette trajectoire représente le déroulement d'un processus qui a les mêmes équations que le processus à estimer. Chaque particule fournit l'information du vecteur d'état similaire au vecteur d'état du processus à estimer, et le poids représentatif de la probabilité que ce vecteur soit celui du processus à estimer.
Pour un grand nombre de particules, on démontre que l'ensemble des états des particules pondérés par leurs poids respectifs correspond à la loi de probabilité conditionnelle du vecteur d'état du processus.

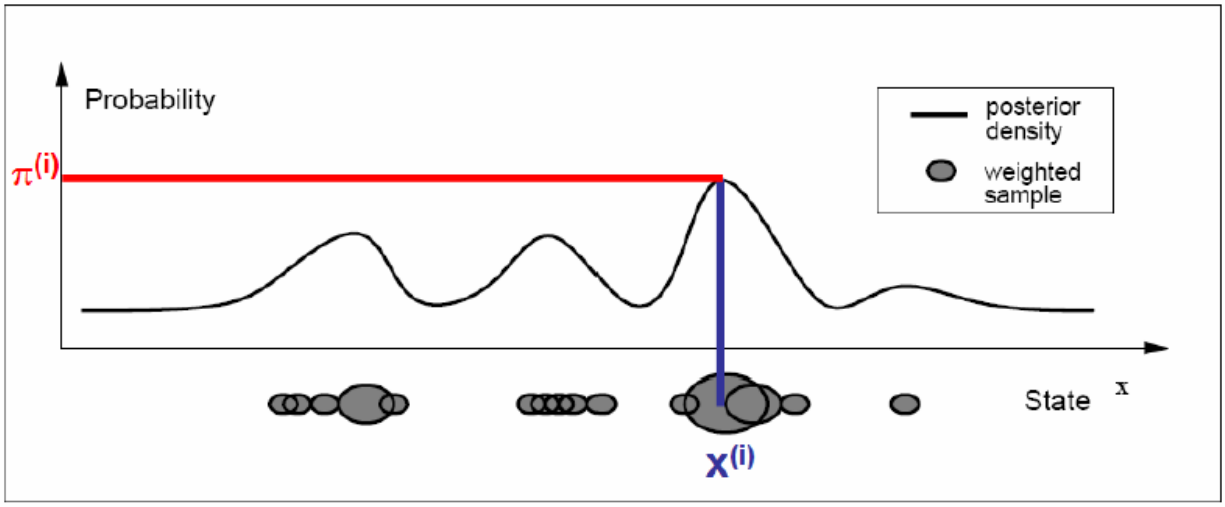


Figure 74 : Estimation de la distribution de l'état par simulation de n particules [Thome N.].

Les trajectoires des particules sont représentées par les courbes et leur poids respectifs par la hauteur des flêches (cf. figure 75).

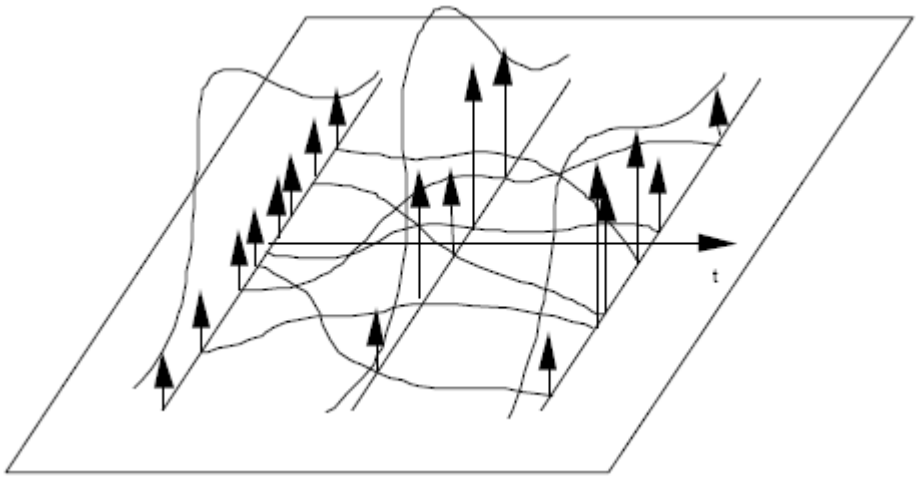


Figure 75 : Trajectoires des particules et leur poids respectifs réprésentés par la hauteur des flêches.

Dans le but d'augmenter la capacité d'exploration de l'estimateur, sans augmenter le nombre de particules, on **redistribue** périodiquement les particules selon leur probabilité.
La **redistribution** permet d'explorer au maximum le réseau de particules dans les régions de probabilité maximale,afin d'améliorer la précision de l'estimation. Nous voyons l'effet de la redistribution à la figure 76 : plusieurs particules sont nées au même endroit, tandis que d'autres ont disparu. Ceci vient du fait que les particules les plus « lourdes » sont favorisées en donnant naissance à plusieurs particules à la même position, alors que les particules aux régions les moins probables sont peu choisies, et disparaisssent de cette façon.

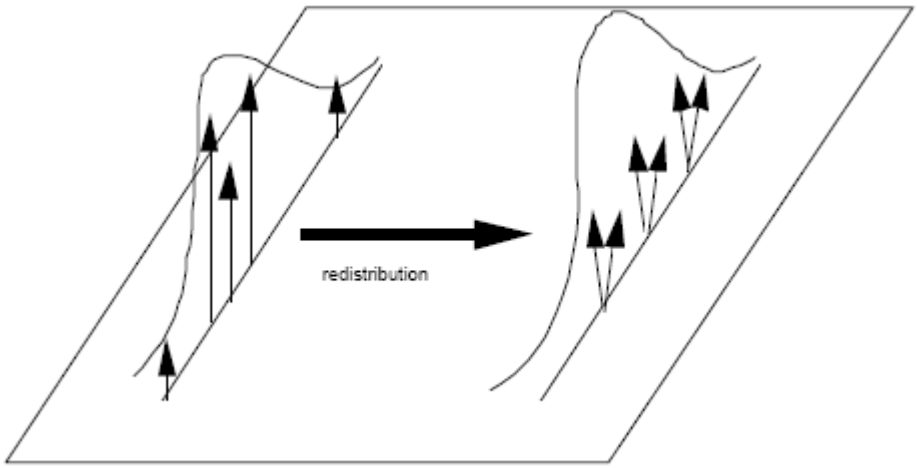


Figure 76 : Effet de la redistribution.

Pour estimer la densité de probablité grâce aux particuls pondérées, il faut trois étapes : propagation, pondération, et rééchantillonnage. En effet, il faut trouver le nombre de particules nécessaires, et la manière dont l'échantillonnage va être effectué. Ensuite, il faut une mesure associée aux observations image pour déterminer le poids de chaque particule. C'est ce que font les trois étapes de **propagation**, **pondération**, et **rééchantillonnage**.

1. **Etape de propagation** : les particules sont diffusées selon un modèle de mouvement dont les paramètres sont estimés. La propagation contient deux termes (cf. figure 77). Un terme correspond au mouvement déterministe et est une dérive (« drift ») en fonction du passé. Un second terme est aléatoire et correspond à un bruit dynamique dit « le mouvement brownien », avec des paramètres statistiques. Ce terme aléatoire parcourt l'espace d'état en cherchant de nouvelles solutions (« diffuse »).
2. **Etape de pondération** : une mesure de similarité avec les données images est effectuée. A l'initialisation du suivi, un modèle est généré, et une distance est calculée entre le modèle généré et la mesure image correspondant à la valeur $x^{(i)}$ de l'espace d'état de la particule en question. A chacune des étapes de pondération, le poids $\pi^{(i)}$ de chaque particule est calculé. Il est d'autant plus grand que la mesure image correspond bien au modèle. A la fin de l'étape de pondérarion, on calcule l'état moyen du système par la somme des différentes particules pondérées par leurs poids. Cet état moyen correspond à l'estimation du vecteur d'état renvoyé par le filtre particulaire : $\hat{x} = E[X] = \sum_{i=1}^{n} \pi^{(i)} \cdot x^{(i)}$.
3. Toute méthode de simulation séquentielle de type Monte Carlo présente un problème de dégénérescence : aprés quelques itérations, les poids non négligeables vont se concentrer sur une seule particule. Afin de limiter ce problème, une **étape de rééchantillonnage** (appelée aussi étape de redistribution) est introduite en fin de chaque cycle de l'algorithme de filtrage particulaire. Cette étape a pour but de tirer un ensemble de n particules, chacune avec la probabilité correspondant aux poids calculés lors de l'étape de pondération, donc selon la densité estimée de l'état. N nouvelles particules sont obtenues par rééchantillonnage avec remise dans l'ensemble $\{x_k^{(i)}\}$. Les particules associées à des poids $\pi_k^{(i)}$ élevés sont dupliqués, au détriment de celles, faiblement pondérées, qui disparaissent. Le tirage avec remise est appelé **« échantillonnage d'importance » ou « Sampling Importance Resempling »**, c'est l'**algorithme SIR**. Cette étape de redistribution peut être soit appliquée systématiquement, soit être déclenchée seulement lorsqu'un critère d'efficacité du filtre est en dessous d'un certain seuil. Cette étape ne doit pas être oubliée, sinon cela correspondrait à simuler le

jeu de n particules initial une fois pour toutes, sans remise, c'est-à-dire sans prendre en compte la pondération calculée grâce aux données image, alors que cette étape permet d'approcher la densité recherchée.

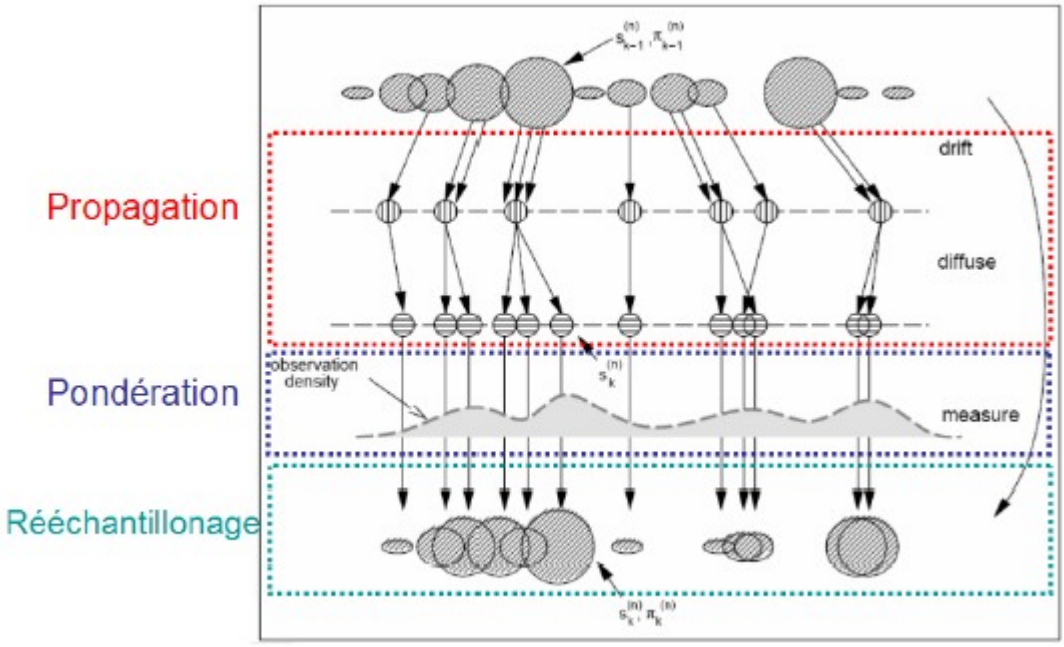
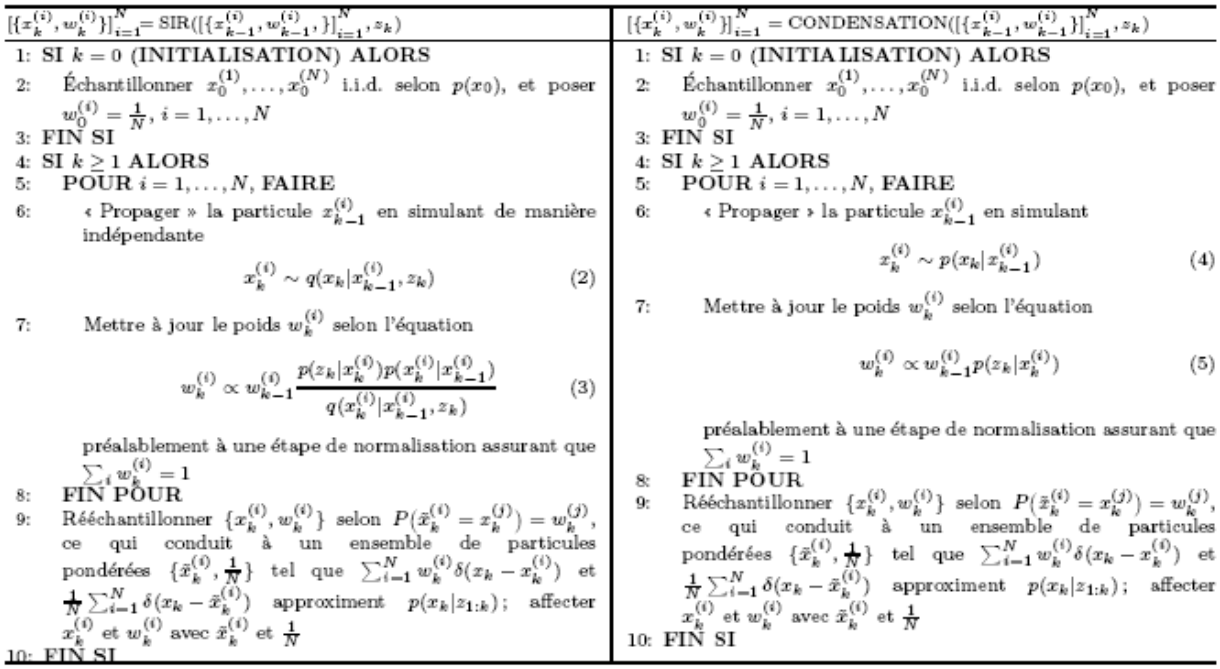


Figure 77 : Etapes de l'algorithme du filtrage particulaire [Thome N.].

2 L'algorithme de CONDENSATION

L'algorithme de CONDENSATION (CONditional DENSity propagATION for visual tracking) [Isard M., Blake A., 98] fut la première application de filtrage particulaire. Il peut être vu comme le cas particulier de l'algorithme SIR (cf. figure 78) où la fonction d'importance est relative à la dynamique du processus d'état.
Ceci donne à la CONDENSATION une structure prédiction/mise à jour comparable à celle du filtre de Kalman, puisque la densité ponctuelle $\sum_{i=1}^{N} \pi_k^{(i)} \delta(x_k - x_k^{(i)})$ approche la prédiction $p(x_k / z_{1:k})$. De plus, la mise à jour des poids rappelle la formule de Bayes correspond à l'étape de mise à jour de l'estimé de Kalman.

Dans le cas du suivi visuel, l'algorithme de CONDENSATION original définit les vraisemblances des particules à partir des primitives visuelles telles que les contours.



$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = \text{SIR}([\{x_{k-1}^{(i)}, w_{k-1}^{(i)},\}]_{i=1}^N, z_k)$

1: SI $k = 0$ (INITIALISATION) ALORS
2: Échantillonner $x_0^{(1)}, \ldots, x_0^{(N)}$ i.i.d. selon $p(x_0)$, et poser $w_0^{(i)} = \frac{1}{N}$, $i = 1, \ldots, N$
3: FIN SI
4: SI $k \geq 1$ ALORS
5: POUR $i = 1, \ldots, N$, FAIRE
6: « Propager » la particule $x_{k-1}^{(i)}$ en simulant de manière indépendante

$$x_k^{(i)} \sim q(x_k | x_{k-1}^{(i)}, z_k) \quad (2)$$

7: Mettre à jour le poids $w_k^{(i)}$ selon l'équation

$$w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(z_k | x_k^{(i)}) p(x_k^{(i)} | x_{k-1}^{(i)})}{q(x_k^{(i)} | x_{k-1}^{(i)}, z_k)} \quad (3)$$

préalablement à une étape de normalisation assurant que $\sum_i w_k^{(i)} = 1$
8: FIN POUR
9: Rééchantillonner $\{x_k^{(i)}, w_k^{(i)}\}$ selon $P(\tilde{x}_k^{(i)} = x_k^{(j)}) = w_k^{(j)}$, ce qui conduit à un ensemble de particules pondérées $\{\tilde{x}_k^{(i)}, \frac{1}{N}\}$ tel que $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$ et $\frac{1}{N} \sum_{i=1}^N \delta(x_k - \tilde{x}_k^{(i)})$ approximent $p(x_k | z_{1:k})$; affecter $x_k^{(i)}$ et $w_k^{(i)}$ avec $\tilde{x}_k^{(i)}$ et $\frac{1}{N}$
10: FIN SI

$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = \text{CONDENSATION}([\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N, z_k)$

1: SI $k = 0$ (INITIALISATION) ALORS
2: Échantillonner $x_0^{(1)}, \ldots, x_0^{(N)}$ i.i.d. selon $p(x_0)$, et poser $w_0^{(i)} = \frac{1}{N}$, $i = 1, \ldots, N$
3: FIN SI
4: SI $k \geq 1$ ALORS
5: POUR $i = 1, \ldots, N$, FAIRE
6: « Propager » la particule $x_{k-1}^{(i)}$ en simulant

$$x_k^{(i)} \sim p(x_k | x_{k-1}^{(i)}) \quad (4)$$

7: Mettre à jour le poids $w_k^{(i)}$ selon l'équation

$$w_k^{(i)} \propto w_{k-1}^{(i)} p(z_k | x_k^{(i)}) \quad (5)$$

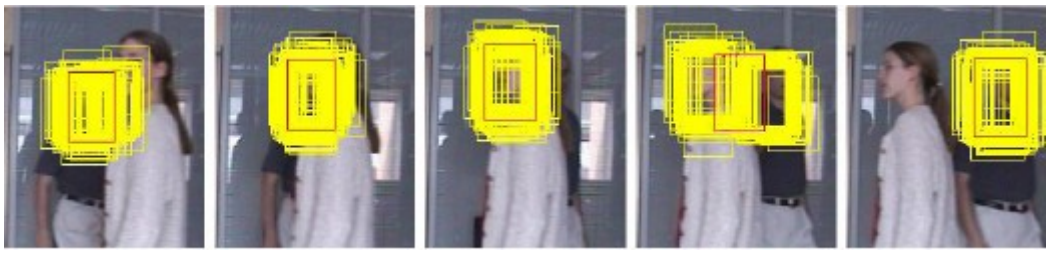
préalablement à une étape de normalisation assurant que $\sum_i w_k^{(i)} = 1$
8: FIN POUR
9: Rééchantillonner $\{x_k^{(i)}, w_k^{(i)}\}$ selon $P(\tilde{x}_k^{(i)} = x_k^{(j)}) = w_k^{(j)}$, ce qui conduit à un ensemble de particules pondérées $\{\tilde{x}_k^{(i)}, \frac{1}{N}\}$ tel que $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$ et $\frac{1}{N} \sum_{i=1}^N \delta(x_k - \tilde{x}_k^{(i)})$ approximent $p(x_k | z_{1:k})$; affecter $x_k^{(i)}$ et $w_k^{(i)}$ avec $\tilde{x}_k^{(i)}$ et $\frac{1}{N}$
10: FIN SI

Figure 78 : Algorithme générique de filtrage particulaire (SIR) et CONDENSATION [Brèthes L., Danès P., Lerasle F.].

3 Présentation des travaux de [Perez P., Hue C., Vermaak J., Gangnet M.]

Nous présentons les résultats de suivi avec le filtre à particules proposé par [Perez P., Hue C., Vermaak J., Gangnet M.] à la figure 79. Le vecteur d'état, cmme indiqué précédemment, est composé de la position et du facteur d'échelle de la boîte englobante de l'objet suivi, et la pondération des particules est obtenue par une mesure de distance entre histogrammes dans l'espace HSV. A la figure 79, les boîtes jaunes représentent les différentes particules $x^{(i)}$ générées pour estimer la densité de probabilité de l'état, et la boîte rouge correspond à l'état estimé par le filtre à particules. Cet état estimé correspond à la moyenne pondérée des particules. Nous pouvons constater que le suivi est correct dans les deux cas. En effet, le filtre à particules adapte la recherche dans l'espace d'état en fonction de la forme de la distribution qui est approchée. Dans les situations simples, quand l'objet suivi a une apparence différente du fond de la scène, un petit nombre d'échantillons a un poids important et la « diffusion » dans l'espace de recherche est faible. Il n'est en effet pas utile de chercher des régions candidates loin de l'objet suivi. La modélisation de la densité par une Gaussienne étant possible, le filtre de Kalman aurait pu convenir également. Cependant, dans le cas d'occultation par un fond où l'hypothèse de normalité n'est plus vérifiée, le filtre de Kalman fonctionne mal. Le filtre à particules, à l'inverse, s'adapte aux mesures images et le mode correspondant à l'état précédemment suivi devient de moins en moins marqué, ce qui a pour effet qu'un grand nmbre de particules se voit attribuer un poids non négligeable. Ceci apparaît sous la forme d'un nuage de particules beaucoup plus diffus (cf. figure 79), permettant de parcourir des zones de l'espace d'état plus lointaines, pour chercher des particules fortement discriminantes. Ainsi à la figure 79a, le bébé revient dans une pose où son apparence est similaire au modèle établi avant le suivi, et à la figure 79b à la fin de l'occultation, la personne qui se trouvait dans le fond revient dans le champ. Le filtre particulaire est plus souple que le filtre de Kalman, dans le sens où il permet le raccrochage aprés des situations d'occultations ou de camouflage.

(a) Fond d'apparence similaire à 'objet d'intérêt



(b) Suivi par filtrage particulaire dans le cas d'occultations

Figure 79 : Performances du filtrage particulaire dans le cas du suivi visuel ([Perez P., Hue C., Vermaak J., Gangnet M.], [Thome N.]).

4 Présentation des travaux de [Brèthes L., Danès P., Lerasle F.]

Enfin, citons les travaux de [Brèthes L., Danès P., Lerasle F.] traitant du suivi visuel de personnes à partir d'une caméra embarquée sur un robot mobile en environnement humain, *a priori* encombré et évolutif. Le but est alors de guider les visiteurs d'une exposition et d'interagir avec eux. Des mesures visuelles sur la couleur, la forme ou le mouvement sont décrites, ainsi que différentes stratégies de filtage prenant en compte plus ou moins ces mesures. Les mesures visuelles combinées définissent une fonction d'importance selon laquelle les particules sont échantillonnées, et si elles sont fusionnées à l'intérieur d'un modèle de mesure, alors celui-ci sert de base à la définition des poids.
Le repositionnement des particules par la fonction d'importance puis l'association d'informations hétérogènes dans le modèle de mesure augmente la robustesse et la précision du suivi.
La figure 80 montre un exemple de suivi incluant une occultation.

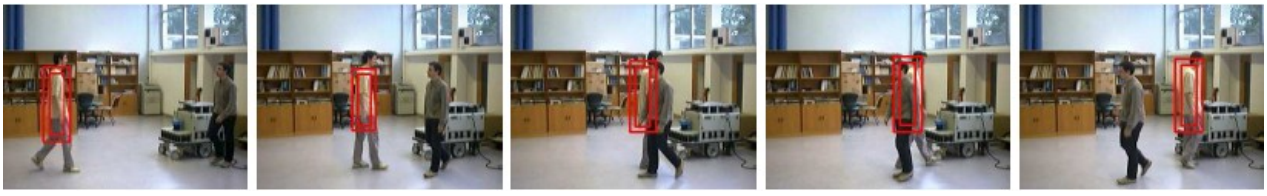


Figure 80 : Exemple de suivi incluant une occultation [Brèthes L., Danès P., Lerasle F.].

Références

[Abrantes A., Marques J., Lemos J.], « Long Term Tracking Using Bayesian Networks ». In IEEE International Conference on Image Processing.Vol. 3, pages 609-612, Rochester, Sept. 2002.
[Adelson E. H., Noyogi S. A.], « Analysing and recognizing walking figures in xyt ». In Proc. CVPR, Vol. 309, pages 469-474, Seattle, Wash., 1994.
[Andrade E., Blunsden S., Fisher R.], « Performance Analysis of Event Detection Models in Crowded Scenes ». In Proc. Workshop on « Towards Robust Visual Surveillance Techniques and Systems » at Visual Information Engineering 2006, Bangalore, India, pages 427-432, Sept 2006.
[Aggarwall J.K., Nandhakumar N.], « On the computation of motion from sequences of images - A review ». In Proceedings of IEEE, 1998, Vol. 76, N°8, pages 917-935.
[Agarwal A., Triggs B.], « Recovering 3D human pose from monocular images ». In IEEE Transactions on Pattern Analysis & Machine Intelligence, Vol. 28, N°1, January 2006.
[Akita K.], « Image sequence analysis of real world human motion ». In Pattern recognition, Vol. 17, N°1, pages 73-83, 1984.
[Ali M.A., Indupalli S., Boufama B.], « Tracking Multiple People for Video Surveillance », University of Windsor, Canada.
[Allmen M., Dyer C. R.], « Computing spatiotemporal relations for dynamic perceptual organization ». In Computer Vision, Graphics and Image Processing: Image Understanding, Vol. 3, N°58, pages 338–351, 1993.
[Anderson C., Burt P., Van Der Wal G.], « Change detection and tracking using pyramid transformation techniques ». In Proceedings of SPIE – Intelligent Robots and Computer Vision, Vol. 579, pages 72-78, 1985.
[Arens M., Nagel H.-H.], « Behavioral Knowledge Representation for the Understanding and Creation of Video Sequences ». In Proceedingsof the 26th German Conference on Artificial Intelligence(KI-2003), 15-18 September 2003, Hamburg, Germany. LNAI, Vol. 28, N°21, pages 149-163. Springer: Berlin Heidelberg New York/NY 2003.
[Avanzi A., Bremond F., Tornieri C., Thonnat M.], « Design and Assessemnt of an Intelligent Activity Monitoring Platform ». In EURASIP Journal on Applied Signal Processing, Special Issue on « Advances in Intelligent Vision Systems: Methods and Application », August 2005, Vol. 2005, N°14, pages 2359-2374.
[Barron J.L., Fleet D.J., Beauchemin S.S.], « Performance of Optical Flow techniques ». In Int. Journal. Comp. Vision, Vol. 12, N°1, pages 43-77, 1994.
[Baumberg A.M.], « Learning Deformable Models for Tracking Human Motion ». PhD thesis, School of Computer Studies, University of Leeds, Leeds, UK, 1995.
[Baumberg A., Hogg D.], « An adaptative eigenshape model ». In British Machine Vision Conference BMVC, Birmingham, 1995.
[Bar-Shalom Y., Fortmann T.E.], « Tracking and data association », Academic Press, 1988.
[Bar-Shalom Y., Li X.], « Multitarget-Multisensor Tracking: Principles and Techniques ». YBS Publishing, 1995.
[Bernier O.], « Real-Time 3D Articulated Pose Tracking using Particle Filters Interacting through Belief Propagation », ICPR, 2006.
[Bernier O., Cheung-Mon-Chang P.], « Real-time 3D articulated pose tracking using particle filtering and belief propagation on factor graphs ». In British Machine Vision Conference, Vol.01, pages 005-008, 2006.
[Besag J.], « Spatial interaction and the statistical analysis of lattice systems ». J. Royal Statist. Soc., 36 B: 192-236, 1974.
[Beymer D.], « Person counting using stereo ». In Workshop on Human Motion, December 2000.
[Blackman S.S.], « Multiple-target tracking with radar applications », Artech House, 1986.
[Blake A., Isard M. 98], « Active Contours ». In Springer Verlag, 1998.
[Black M.J., Jepson A.D.], « EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation », ECCV, Vol. II, pages 329–342, 1996.
[Bobick A., Campbell L.], « Recognition of human body motion using phase space constraints ». In Technical Report 309, M.I.T Media Laboratory, Perceptual Computing Section, 1994.
[Bobick A.F., Wilson A.D.], « A state-based technique for the summarization and recognition of gesture ». In Proceedings of 5th International Conference on Computer Vision, pages 382-388, Cambridge, 1995.

[Bogaert M., Chleq N., Cornez P., Regazzoni C., Teschioni A., Thonnat M.], « The passwords project ». In International conference on Image Processing (ICIP'96). Proceeding in IEEE ICIP. Vol 3, pages 675-678. Lausanne, Switzerland, September 1996.
[Boucher C.], « Contribution à la Fusion d'Informations Par Filtrage Non-Linéaire : Application à l'Estimation de la Structure et Du Mouvement 3D Dans un Contexte Multi-Capteurs ». In PhD thesis, Université du Littoral Côte d'Opale, Octobre 2000.
[Bouthemy P. 87], « Estimation et structuration d'indices spatio–temporels pour l'analyse du mouvement dans une séquence d'images », Traitement du Signal, Vol. 4, N°3, pages 239-257, 1987.
[Bouthemy P., Santillana Rivero J.], « A hierarchical likelihood approach for region segmentation according to motion-based criteria ». In Proc. of 1rst Int. Conf. on Computer Vision, pages 463-467, Londres, 1987.
[Bouthemy P. 88], «Modèles et méthodes pour l'analyse du mouvement dans une séquence d'images », 2nd Atelier Scientifique Traitement d'Images : du Pixel à l'Interprétation, Aussois, 1988, XXV–1 à XXV–19 .
[Bouthemy P. 89], « A Maximum Likelyhood Framework for Determining Moving Edges ». In IEEE Trans. PAMI, Vol. 11, N°5, pages 499-511, May 1989.
[Bouthemy P., François E.], « Motion segmentation and qualitative dynamic scene analysis from an image sequence ». In Int. Journal of Computer Vision, Vol. 10, N°2, pages 157-182, April 1993.
[Bouthemy P., Lalande], « Recovery of moving object masks in an image sequence using local spatio-temporal contextual information ». In Optical Engineering, Vol. 32, N°6, pages 1205-1212, 1993.
[Brand M., Kettnaker V.], « Discovery and segmentation of activities in video ». In IEEE Trans. Pattern Anal. Mach. Intell., Vol. 22, N°8, pages 844-851, 2000.
[Bregler C., Malik J.], « Tracking people with twists and exponential maps ». In CVPR, pages 8-15, 1998.
[Bremond], « Environnement de résolution de problèmes pour l'interprétation de séquences d'images ». PhD thesis, INRIA-Université de Nice Sophia-Antipolis.
[Brèthes L., Danès P., Lerasle F.], « Stratégies de filtrage particulaire pour le suivi visuel de personnes : description et évaluation ». In RFIA 2006, Tours France.
[Buechler G., Smith P.], « A branching algorithm for discriminating and tracking multiple objects ». In IEEE Trans. Automat. Contr., Ac-Vol.20, pages 101-104, February 1975.
[Buxton H., Gong S.], « Advanced Visual Surveillance using Bayesian Networks ». In International Conference on Computer Vision, Cambridge, Massachusetts, June 1995.
[Cai Q., Mitiche A., Aggarwal J.K.], « Tracking human motion in an indoor environment ». In Proceedings of the 2nd International Conference on Image Processing (ICIP'95), pages 215-218, 1995.
[Chadhury S., Subramanian S., Parthasaraty G.], « Heuristic search approach to shape matching in image sequences », in Proceedings of IEEE, Vol. 138, N°2, pages 97-105, 1991.
[Chalidabhongse T., Kim K., Harwood D., Davis L.], « A perturbation method for evaluating background subtraction algorithms ». In Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Nice, France, 2003.
[Cham T.J., Rehg J.M.], « A multiple hypothesis approach to figure tracking ». In Perceptual User Interfaces, pages 19-24, November 1998.
[Chang T.H., Gong S., Ong E.J.], « Tracking multiple people under occlusion using multiple cameras ». In Proceedings of the 11th British Machine Vision Conference, 2000.
[Chang T.H., Gong S.], « Tracking multiple people with a multicamera system ». In Proceedings of IEEE ICCV Workshop on Multi-Object Tracking, pages 19-26, Vancouver, 2001.
[Chen Y., Rui Y., 2004], « Real-time Speaker Tracking Using Particle Filter Sensor Fusion ». In Proceeding of the IEEE, Vol. 92, N°3, pages 485-494, March 2004.
[Chen Z., Lee H.], « Knowledge-guided visual perception of 3D gait from a single image sequence ». In IEEE Transactions on systems, man and cybernetic, Vol. 22, N°2, pages 336-342, 1992.
[Chen H.T., Lin H.H., Liu T.L.], « Multi-Object Tracking Using Dynamical Graph Matching ». In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, pages 210-217, 9-14 December 2001, Kauai Marriott, Hawaii.
[Chen Y., Rui Y., Huang T.S.], « JPDAF based HMM or real-time contour tracking ». In CVPR, Vol. 1, pages 543-550, 2001.
[Chleq N., Thonnat M.], « Realtime image sequence interpretation for videosurveillance ». In IEEE, editor, International Conference on Image Processing, Lausanne, Switzerland, pages 801-804, 1996.

[Choi S., Seo Y., Kim H., Hong K.], « Where are the ball and players? soccer game analysis with color-based tracking and image mosaik ». In ICIAP, 1997.
[Chomat O., Crowley J.L.], « Recognizing motion using local appearance ». In International Symposium on Intelligent Robotic Systems, University of Edinburgh, 1998.
[Cohen I., Medioni G.], « Detecting and Tracking Moving Objects for Video Surveillance ». In IEEE Proceedings of Computer Vision and Pattern Recognition, Fort Collins, Jun. 1999, pages 1-7.
[Collins R., et al.a], « A System for Video Surveillance and Monitoring ». CMU-RI-TR-00-12, Robtics Institute, CMU, May, 2000.
[Collins R., et al.b], « A System for Video Surveillance and Monitoring: VSAM Final Report ». In Technical report CMU-RI-TR-00-12, 2002.
[Comaniciu D., Meer P.], « Mean shift : A robust approach toward feature space analysis ». In IEEE Trans. Pattern Analysis Machine Intell., Vol. 24, N°5, pages 603-619, 2002.
[Comaniciu D., Ramesh V., Meer P.], « Kernel-based object tracking ». In PAMI, Vol. 25, N°5, pages 564-577, 2003.
[Cootes T.F., Taylor C.J.], « Active shape models - `Smart snakes' ». In British Machine Vision Conference, pages 276-285, september 1992.
[Cootes T.S, Taylor C.J., Cooper D.H., Graham J.], « Active shape models-Their training and application » , Computer Vision and Image Understanding, Vol. 61, N°1, pages 38-59, January 1995.
[MacCormick J., Isard M.], « Partitioned sampling, articulated objects, and interface-quality hand tracking ». In ECCV, Vol. 2, pages 3–19, 2000.
[Cox I.J.], « A review of statistical data association techniques for motion correspondence ». In Int. J. of Computer Vision, Vol. 10, N°1, 1993.
[Cox I.J, Hingorani S.L.], « An Efficient Implementation of Reid's Multiple Hypothesis Traking Algorithm and Its Evaluation for the Propose of Visaul Traking ». In IEEE Trans. Pattern Anal. Mach. Intell., Vol. 18, N °2, pages 138-150, February 1996.
[Crowley J.L., Demazeau Y.], « Principles and Techniques for Sensor Data Fusion », Signal Processing (EURASIP), Vol. 32, pages 5-27. [
[Cupillard F., Avanzi A., Bremond F., Thonnat M.], « Video understanding for metro surveillance ». In IEEE International Conference on Networking, Sensing and Control, March 2004.
[Le Cun Y., Bottou L., Bengio Y., Haffner P.], « Gradient-based learning applied to document recognition ». In Proc. IEEE, Vol. 86, N°11, pages 2278-2324, 1998.
[Davis J.W., Bobick A.F.], « The representation and recognition of human movement using temporal templates ». In Proceedings on the Computer Vision and Pattern Recognition, pages 928-934, 1997.
[Demirdjian D., Ko T., Darrell T.], « Constraining human body tracking ». In ICCV '03: Proceedings of the 9th IEEE International Conference on Computer Vision, page 1071, IEEE Computer Society, 2003.
[Demirdjian D., Taycher L., Shakhnarovich G., Grauman K., Darrell T.], « Avoiding the « streetlight effect »: Tracking by exploring likelihood modes ». In ICCV, pages 357-364, 2005.
[Deriche R.], « Using Canny's criteria to derive a recursively implemented optimal edge detector », International Journal of Computer Vision, Vol. 2, pages 167-187, 1987.
[Deutscher J., Blake A., Reid I.], « Articulated body motion capture by annealed particle filtering », CVPR, Vol. 2, pages 126-133, 2000.
[Dimitrijevic M., Lepetit V., Fua P.], « Human body pose recognition using spatio-temporal templates ». In ICCV, 2005.
[Djeraba C.], « State of Art in Body Tracking », Publication interne N°6, Laboratoire d'Informatique Findamentale de Lille, Université des Sciences et Technologies de Lille, 2005.
[Doucet A., De Freitas N., Gordon N.], « Sequential Monte Carlo methods in practice ». In Stats. for Eng. and Info, Sciences, Springer Verlag, 2001.
[Du L., Sullivan G., Baker K.], « Quantitative analysis of the view point consistency constraint in mode-based vision ». In International Conference of Computer Vision, pages 632-639, Berlin, 1993.
[Elgammal A., Duraiswami R., Davis L.S.], « Probabilistic tracking in joint feature-spatial spaces ». In Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, 2003.
[Elgammal A., Duraiswami R., Harwood D., Davis L.S.], « Background and foreground modeling using non-parametric kernel density estimation for visual surveillance ». Proc. IEEE Vol. 90, N°7, pages 1151-1163,

2002.
[Elgammal A.M., Harwood D., Davis L.S], « Non-parametric Model for Background Subtraction ». In Proceedings of the 6th European Conference on Computer Vision-Part, Vol. 2, pages 751-767, June 26-July 01, 2000.
[Felzenszwalb P.F., Huttenlocher D.P. 00], « Efficient Matching of Pictorial Structures ». In Proceedings on the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 66-75, Hilton Head Island, USA, 2000.
[Felzenszwalb P.F., Huttenlocher D.P. 03], « Pictorial structures for object recognition ». Submitted to IJCV, 2003.
[Fieguth P., Terzopoulos D.], « Color-based tracking of heads and other mobile objects at video frame rates ». In CVPR '97 : Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97), page 21, Washington, DC, USA, 1997. IEEE Computer Society.
[Forsyth D.A., Fleck M.M.], « Body plans ». In Proceedings on the IEEE Conference on Computer Vision and Pattern Recognition, pages 678-683, Puerto Rico, USA, 1997.
[Fortmann T.E., Bar-Shalom Y., Scheffe M.], « Sonar tracking of multiple targets using joint probabilistic data association ». In IEEE J. Oceanic Eng. OE-8, pages 173-184, 1983.
[Fuentes L.M., Velastin S.A.], « People tracking in surveillance applications ». In Proceedings of the IEEE Workshop on Performance Evaluation of Tracking and Surveillance (PETS2001), 2001.
[Fusier F., Valentin V., Bremond F, Thonnat M.], « Video understanding for complex activity recognition ». In Machine Vision and Applications (2007), Special Issue Paper, Vol. 18, pages 167-188. Springler-Verlag 2007.
[Galata A., Johnson N., Hogg D.], « Learning variable length markov models of behaviour ». In Journal of Computer Vision and Image Understanding, pages 398-413, 2001.
[Gandhi T., Trivedi M.M.], « Person tracking and reidentification: Introducing Panoramic Appearance Map (PAM) for feature representation ». In Machine Vision and Applications (2007), Special Issue Paper, Vol. 18, pages 207-220. Springler-Verlag 2007.
[Gao J., Shi J.], « Multiple frame motion inference using belief propagation ». In FGR, pages 875-882, 2004.
[Garcia V.], Rapport de DEA Image Vision, « Estimation de mouvement subpixélique par blocs adaptée à la couleur avec modèle de mouvement », Laboratoire I3S, Equipe CreATIVe, soutenu le 14 Septembre 2004.
[Gauvrit H., Le Cadre J.P.], « A formulation of multitarget tracking as an incomplete data problem ». In IEEE Trans. Aerosp. Electron. Systems. Vol. 33, N°4, pages 1242-1257, 1997.
[Gauvrit H.], « Extraction multi-pistes : approche probabiliste et approche combinatoire ». Thèse Université de Rennes 1 IRISA, décembre 1997.
[Gavrila D.M., Davis L.S.], « 3-D Model-Based Tracking of Humans in Actions: A Multi-View Approach ». In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 73-80, San Francisco, USA, June 1996.
[Gelgon M.], « Segmentation spatio-temporelle et suivi dans une séquence d'images : application à la structuration et à l'indexation de vidéo ». Thèse de doctorat, Université de Rennes 1, 1998.
[Geman S., Geman D.], « Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images ». In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, n°6, novembre 1984, pages 721-741.
[Gerber R., Nagel H.-H.], « Representation of « Occurrences » for Road Vehicle Traffic ». In Internal Report, 31 March 2006. Institut für Algorithmen und Kognitive Systeme, Fakultät für Informatik der Universität Karlsruhe(TH), 76128 Karlsruhe.
[Georis B., Bremond F., Thonnat M., Macq B.], « Use of an evaluation and diagnosis method to improve tracking performances ». In Hamza,M.(ed.) Proceedings of the 3rd IASTED International Conference on Visualization, Imaging and Image Processing (VIIP'03), pages 827-832. Acta Press, Benalmadera, Spain, 2003.
[Georis B., Bremond F., Thonnat M.], « Real-time control of video suveillance systems with program supervision techniques ». In Machine Vision and Applications, Special Issue Paper, Vol. 18, pages 189-205. Springler-Verlag 2007.
[Georis B., Maziere M., Bremond F., Thonnat M.], « A video interpretation platform applied to bank agency monitoring ». In Proceedings of IDSS'04-2nd Workshop on Intelligent Distributed Surveillance Systems London,UK, 2004.

[Gomila C.], « Mise en correspondance de partitions en vue du suivi d'objets ». Thèse de doctorat, École Nationale Supérieure des Mines de Paris, 2001.
[Gong Y.], « Integrated Object Detection and Tracking by Multiple Hypothesis Analysis ». In NEC J Adv Technol. Vol. 2, N°1, pages 13-18, 2005.
[Grava C.], « Compensation de mouvement par réseaux neuronaux cellulaires. Application en imagerie médicale », Thèse de doctorat, soutenue le 12 décembre 2003, INSA de Lyon.
[Grimson E., Viola P.], « A forest of sensors ». In Proceedings of DARP -VSAM workshop II, November 1997.
[Grimson W.EL., Stauffer C., Romano R., Lee L.], « Using Adaptive Tracking to Classify and Monitor Activities in a Site ». In CVPR archive Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society Washington, DC, USA, 1998.
[Hall D., Crowley, J. et al.], « Comparison of target detection algorithms using adaptive background models ». In IEEE VS-PETS. Beijing, China, 2005.
[Hampapur A., Brown L., Connell J., Ekin A., Haas N., Lu M., Merkl H., Pankanti S.], « Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking ». in Signal Processing Magazine, IEEE, Vol. 22, N°2, pages 38-51March 2005.
[Han M., Xu W., Gong Y.], « Multi-object trajectory tracking », Machine Vision and Applications, Special Issue Paper, Vol. 18, pages 221-232. Springler-Verlag 2007.
[Haritaoglu I., Harwood D., Davis L.S. 98], « Ghost : A human body part labeling system using silhouettes ». In Fourteenth International Conference on Pattern Recognition, Brisbane, Vol. 8, 1998.
[Haritaoglu I., Harwood D., Davis L.S. 99], « Hydra: multiple people detection and tracking using silhouettes ». In IEEE Workshop on Visual Surveillance, 1999.
[Haritaoglu I., Harwood D., Davis L.S. 00], « W^4 : Real-Time Surveillance of People and Their Activities ». In IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 22, N°8, pages 809-830, August 2000. Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04) 1051-4651/04 $ 20.00 IEEE.
[Harris C., Stephens M.], « A combined corner and edge detector ». In Alvey Vision Conference, pages 147-151, 1988.
[Heisele B., Wöhler C.], « Motion-Based Recognition of Pedestrians ». In Fourteenth International Conference on Pattern Recognition, Brisbane, Qld., Australia, 16-29August 1998, Vol. 2, pages 1325-1330.
[Hogg D.], « Model-based vision: A program to see a walking person ». In Image and Vision Computing, Vol. 1, pages 5-20, 1983.
[Hongeng S., Bernard F., Nevatia R.], « Representation and optimal recognition of human activities ». In IEEE Proceedings of Computer Vision and Pattern Recognition, South Carolina, USA, 2000.
[Hongeng S., Bremond F., Nevatia R.], « Bayesian framework for video surveillance application ». In Proc. of the 15th International Conference on Pattern Recognition, Barcelona, Spain, September 2000.
[Horn B.K.P, Schunk B.G.], « Determining optical flow ». In Artificial intelligence, Vol. 17, pages 185-204, 1981.
[Horprasert T., Harwood D., Davis L.], « A statistical approach for real-time robust background subtraction and shadow detection ». In IEEE ICCV'99 FRAME-RATE Workshop, Kerkyra, 1999.
[Housewright R.B., Singer R.A., Sea R.G.], « Derivation and evaluation of improved tracking filters for use in dense multitarget environments », In IEEE Transactions on Information Theory, Vol. 20, July 1974, pages 423-432.
[Hu W., Tan T., Wang L., Maybank S.], « A survey on visual surveillance of object motion and behaviors ». In IEEE Trans. Syst. ManCybern. Part C, Vol. 34, N°3, 334-352, 2005.
[Huang T., Russell S.J.], « Object identification in a Bayesian context ». In Proceedings of IJCAI1997, pages 1276-1283, 1997.
[Huang K.S., Trivedi M.M.], « 3D shape context based gesture analysis integrated with tracking using omni video array ». In Proceedings of the IEEE Workshop on Vision for Human-Computer Interaction (V4HCI). San Diego,USA, 2005.
[Hue C., Le Cadre J.P., Perez P.], « Tracking multiple objects with particle filtering ». In IEEE Trans. Aerosp Electron. Systems, Vol. 38, N°3, 791-812, 2002.
[Intille S.S., Bobick A.F., 95], « Closed world tracking ». In 5th International Conference on Computer Vision

ICCV, Cambridge.
[Intille S.S., Bobick A.F., 01], « Recognizing Planned, Multiperson Action ». In Computer Vision and Image Understanding, Vol. 81, N°3, pages 414-445, 2001.
[Ioffe S., Forsyth D.A., 99], « Finding people by sampling ». In ICCV, pages 1092-1097, 1999.
[Ioffe S., Forsyth D.A., 01], « Probabilistic methods for finding people ». In IJCV Vol. 43, N°1, pages 45-68, 2001.
[Ioffe S., Forsyth D.A., 03], « Human tracking with mixtures of trees ». In ICCV, pages 690-695, 2001.
[Isard M.], « Pampas : Real-valued graphical models for computer vision ». In CVPR, Vol. 1, pages 613-620, 2003.
[Isard M., Blake A., 96], « Contour tracking for stochastic propagation of conditional density ». In the 4th Proceedings of the European Conference on Computer Vision, pages 343-356, Cambridge UK, April 1996, LNCS 1065.
[Isard M., Blake A., 98], « Condensation-Conditional Density Propagation for Visual Tracking ». In IEEE Intl J. Computer Vision, Volume 29, N°1, pages 5-28, 1998.
[Isard M., Mac Cormick J.P.], « BraMBLe: A Bayesian Multiple-Blob Tracker ». In IEEE Proc. 8th Int. Conf. on Computer Vision, Vol. 2, Vancouver, July 2001, pages 34-41.
[Ivanov Y., Bobick A.F.], « Recognition of visual activities and interactions by stochastic parsing ». In PAMI, 2000.
[Jabri S., Duric Z., Wechsler H., Rosenfeld A.], « Detection and location of people in video images using adaptive fusion of color and edge information ». In Proc. 15th International Conference on Pattern Recognition, Barcelona, Spain, Vol. 4, N°9, pages 627-630, 2000.
[Jain R., Martin W., Aggarwal J.] (1979). « Segmentation throught the detection of changes due to motion ». In Computer Graphics and Image Processing, Vol. 2, pages 13-34.
[Jain R.], « Dynamic scene analysis using pixel based processes », In IEEE Transactions on Computers, Vol. 14, N°8, août 1981, pages 12-18.
[Javed O., Rasheed Z., Shafique K., Shah M.], « Tracking across multiple cameras with disjoint views ». In Proceedings of IEEE International Conference on Computer Vision, pages 1-6, 2003.
[Javed O., Shafique K., Shah M.], « Appearance modeling for tracking in multiple non-overlapping cameras ». In IEEE CS Conf. Comput.Vis.Pattern Recognit. Vol. 2, pages 26-33, 2005.
[Jehan-Besson S.], Présentation « Analyse vidéo. Introduction, formats, applications », GREYC-Images, ENSICAEN option Image/Multimédia & Telecom, Septembre 2004.
[Jensen F.a], « An introduction to bayesian networks ». Springer, pages 398-413, 1996.
[Jensen F.b], « Bayesian Networks and Decision Graphs », Springer, 2001.
[Johansson G.], « Visual perception of biological motion and a model for its analysis ». In Perception and Psychophysics. Vol. 14, N°2, pages 201-211, 1973.
[Johnson N.], PhD thesis, « Learning Object Behaviour Models ». School of Computer Studies, University of Leeds, Leeds, UK, September 1998. http://www.scs.leeds.ac.uk/neilj/ps/thesis.ps.gz.
[Jojic N., Petrovic N., Frey B.J., Huang T.S.], « Transformed hidden markov models: estimating mixture models of images and inferring spatial transformations in video sequences ». In CVPR, Vol. 2, pages 26-33, 2000.
[Jorge P.M., Marques J.S., Abrantes A.J], « Estimation of the Bayesian network architecture for object tracking in video sequences ». In Proceedings of the 17th International Conference on Pattern Recognition ICPR, Vol. 2, pages 732-735, Cambridge, August 2004.
[Jordan M.I., Sejnowski T.J., Poggio T.], « Graphical Models : Foundations of Neural Computation ». In MIT Press, 2001.
[Ju S., Black M., Yacoob Y.], « Cardboard people : A parameterized model of articulated image motion ». In FG '96 : Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition (FG '96), pages 38-44, Washington, DC, USA, 1996. IEEE Computer Society.
[Junejo I.N., Shah O., Shah M.], « Multi Feature Path Modeling for Video Surveillance ». In Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04), Vol. 2, pages 716-719, 2004.
[Kale A. et al.], « Identification of humans using gait ». In IEEE Transactions on Image Processing, 2004.
[Kalman R.E.], « A new approach to linear filtering and prediction problems ». In Transaction of the ACME Journal of basic ingineering, pages 343–356, 1960.

[Karaulova I.A., Hall P.M., Marshall A.D.], « A hierarchical model of dynamics for tracking people with a single video camera ». In British Machine Vision Conference, pages 352-361, 2000.
[McKenna S., Raja Y., Gong S.], « Tracking color objects using adaptive mixture models ». Image Vis. Comput. Vol. 17, pages 225–231, 1999.
[Kettnaker V., Zabih R.], « Bayesian multi-camera surveillance ». In IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, pages 253-259, 1999.
[Khan S., Shah M.], « Consistent labeling of tracked objects in multiple cameras with overlapping fields of view ». In IEEE Pattern Analysis and Machine Intelligence,Vol. 25, N° 10, October 2003, pages1355-1360.
[Khan S., Javed O., Rasheed Z., Shah M.], « Human tracking in multiple cameras ». In Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV 2001), Vancouver, Canada, July 9-12, pages 331-336, 2001.
[Kholer Ch., Ottlik A., Nagel H.-H, Nebel B.], « Qualitative Reasoning Feeding Back into Quantitative Model-Based Tracking », Technical Report N°204, Fakultat fur Informatik, Albert-Ludwigs-Universitat, http://cogvisys.iaks.uni-karlsruhe.de.2004.
[Koga T., Linuma K., Hirano A., Lijima Y., Ishiguro T.], « Motion compensated interframe coding for video conferencing », Proc. Nat. Telecommun. Conf., 1981.
[Koller D., Daniilidis K., Nagel H.-H], « Model-based object tracking in monocular image sequence of road trafic scenes ». In International Journal of Computer Vision, Vol. 3, N°10, pages 257-281, 1993.
[Kschischang, Frey, Loeliger], « Factor graphs and the sum-product algorithm ». In IEEETIT: IEEE Transactions on Information Theory, Vol. 47, 2001.
[Lanvin P.], « Suivi de formes par filtrage particulaire ». In Technical report, DEA Automatique et Informatique Industrielle, Université Lille 1, Juin 2001.
[Lan X, Huttenlocher D.P.], « A unified spatio-temporal articulated model for tracking ». In CVPR, Vol. 1, pages 722-729, 2004.
[Landabaso J.L., Xu L.Q., Pardas. M.], « Robust Tracking and Object Classification Towards Automated Video Surveillance ». In International Conference on Image Analysis and Recognition ICIAR 2004, Part II, pages 463-470, Porto, Portugal, September 29-October 1, 2004.
[Lee L., Grimson W.E.L.], « Gait analysis for recognition and classification », In 5th IEEE International Conference on Automatic Face and Gesture Recognition, May 2002.
[Lee M.W., Cohen I.], « Proposal maps driven MCMC for estimating human body pose in static images ». In CVPR, Vol.2, pages 334-341, 2004.
[Leignel C., Viallet J.E.], « A blackboard architecture for the detection and tracking of a person ». In RFIA, Toulouse, 2004.
[Li J.,Chellappa R.], « Appearance modeling under geometric context ». In the 10th IEEE International Conference on Computer Vision, 2005.
[Lipton A.J., Fujiyoshi H., Patil R.S.], « Moving target classification and tracking from real-time video ». In Proceedings of the DARPA Image Understanding Workshop(IUW'98), pages 129-136, Monterey, USA, 1998.
[Madden C., Dahai Cheng E., Piccardi M.], « Tracking people across disjoint camera vieuws by an illumination-tolerant appearance representation », Machine Vision and Applications, Special Issue Paper, Vol. 18, pages 233-247. Springler-Verlag 2007.
[Matsuyama T.], « Cooperative distributed vision ». In Proceedings of DARPA Image Understanding Workshop, Vol. 1, pages 365-384, November 1998.
[Maybeck P.S], « Stochastic models, estimation, and control ». Vol. 141 of Mathematics in Science and Engineering. Academic Press, 1979.
[Megret R.], « Structuration spatio-temporelle de séquences vidéo », thèse de doctorat, soutenue le 17 décembre 2003, Laboratoire d'InfoRmatique en Image et Systèmes d'Information LIRIS, INSA de Lyon.
[Metropolis N. et al.], « Equations of state calculations by fast computing machines », Journal of Chemical Physics, Vol. 21, pages 1087-1091, 1953.
[Meyer F., Bouthemy P. 92], « Region-based tracking in an image sequence ». In Proc. Second European Conference on Computer Vision, S. Margherita, Ligure, Italy, May 1992, G. Sandini (ed.), Lecture Notes in Computer Science 588, Springer-Verlag, Berlin, Heidelberg, New York, 1992, pages 476-484.
[Meyer F., Bouthemy P. 94], « A Region-based tracking using affine motion models in long image sequences », Computer Vision, Graphics and Image Processing. In Image Understanding, Vol 60, n°2, pages 119-140,

1994.
[Mitiche A., Bouthemy P.], « Computation of image motion: a synopsis of current problems and methods ». In Int. Journ. of Comp. Vis., Vol. 19, N°1, pages 29-55, 1996.
[A. Mittal, L. Davis], « M2 tracker: a multi-view approach to segmenting and tracking people in a cluttered scene ». I n Int.J. Comput.Vis. Vol. 51, N°3, pages 189-203, 2003.
[Moenne-Locoz N., Bremond F., Thonnat M.], « Recurrent bayesian network for the recognition of human behaviors from video ». In Crowley J., Piater J.,Vincze M., Paletta L. (eds.) Proceedings of the 3rd International Conference on ComputerVision Systems (ICVS'03). Lecture Notes in Computer Science, pages 68-77. Springer, Graz, 2003.
[Moghaddam B., Pentland A.], « Probabilistic visual learning for object representation ». In IEEE Trans. PAMI, Vol. 19, N°7, pages 696-710, July 1997.
[Mori G., Malik J.], « Estimating human body configurations using shape context matching ». In Proc. 7th European Conf. Computer Vision, pages 666–680, 2002.
[Mori G., Ren X., Efros A.A., Malik J.], « Recovering human body configurations : Combining segmentation and recognition ». In CVPR, Vol. 2, pages 326-333, 2004.
[Motamed C.], Habilitation à Diriger des Recherches H.D.R, « Contribution à la conception de systèmes d'interprétation de séquences d'images », Université du Littoral Côte d'Opale, 2006.
[Motamed C., Wallart O.], « Suivi d'objets dans une scène étendue par un système de vision distribué, application à la surveillance d'environnements autoroutiers », revue Traitement du signal, Vol. 20, N°1, pages 87-100, 2003.
[Nagel H.-H], « The representation of situations and their recognition from image sequences ». In RFIA, pages 1221-1229, Lyon-Villeurbanne, 1988.
[Nair V., Clark J.], « Automated visual surveillance using hidden markov models ». In ICVI, Vol. 5, pages 88-93, 2002.
[Nait-Charif H., McKenna S.], « Activity summarisation and fall detection in a supportive home environment ». In ICPR, pages 323–326, 2004.
[Niu W., Jiao L., Han D., Wang Y.-F.], « Real-Time Multi-person Tracking in Video Surveillance », ICICS-PCM 2003, 15-18 Decembre 2003.
[Noriega P. a], « Modèle du corps pour le suivi du haut du corps en monoculaire », thèse de doctorat, LORIA, Nancy 1, soutenue le 11 Octobre 2007.
[Noriega P. b], « Multicues 3D Monocular Upper Body Tracking Using Constrained Belief Propagation », 2007. In British Machine Vision Conf., Warwick, UK, September 10-13 2007.
[Noyer J-C.], « Fusion multicapteurs par filtrage non-linéaire : application à la détection et au suivi de formes en vision par ordinateur », H.D.R. Habilitation à Diriger des Recherches, soutenue le 19 décembre 2003, Université du Littoral Côte d'Opale, Laboratoire d'Analyse des Systèmes du Littoral.
[Odobez J.M., Bouthemy P. 94], « Detection of Multiple Moving Objects Using Multiscale MRF with Camera Motion Compensation », Int. Conf. Image Proc., Austin TX (USA), Vol. 2, pages 257-261, 1994.
[Odobez J.M., Bouthemy P. 98], « Direct incremental model-based image motion segmentation for video analysis ». In Signal Processing, Vol. 66, N°3, pages 143-156, May 1998.
[Oliver N., Horvitz E., Garg A.], « Layered representations for human activity recognition ». In Proceedings of the 4th IEEE Int. Conf. on Multimodal Interfaces, pages 3-8, 2002.
[Oliver N.M., Rosario B., Pentland A.P.], « A Bayesian computer vision system for modeling human interactions ». In IEEE Trans. Pattern Anal. Mach. Intell. Vol. 22, N°8, pages 831-843, 2000.
[Orkisz M., Clarysse P.], « Estimation du flot optique en présence de discontinuités : une revue ». In Traitement du Signal, Vol 13, N°5, Spécial 1996.
[J. Orwell, P. Remagnino, G.A. Jones], « Multi-camera colour tracking ». In Proceedings of the IEEE International Workshop on Visual Surveillance, June 26, Fort Collins, Co, pages 14-21, 1999.
[Papageorgiou C., Oren M., Poggio T.], « A General Framework for Object Detection ». In Proceedings of 6th International Conference on Computer Vision, Bombay, India, 4-7 January 1998, pages 555-562. IEEE Computer Society 1998.
[Paragios N., Deriche R.], « Geodesic active contours and level sets for the detection and tracking of moving objects ». In IEEE Trans. Pattern Anal. Mach. Intell., pages 266-280, 2000.
[Park S., Aggarwal J.K. 04a], « A hierarchical bayesian network for event recognition of human actions and

interactions ». In Multimedia Systems: Special Issue on Video Surveillance, pages 164-179, 2004.
[Park S., Aggarwal J.K. 04b], « Semantic-level understanding of human actions and interactions using event hierarchy ». In IEEE Workshop on Articulated and Nonrigid Motion. Washington, DC,USA, 2004.
[Park S., Trivedi M.M. 07], « Multi-person interaction and activity analysis: a synergetic track- and body-level analysis framework », Machine Vision and Applications, Special Issue Paper, Vol. 18, pages 151-166, Springler-Verlag 2007.
[Pentland A.], « Machine understanding human action ». In 7th International Forum on of Frontier of Telecommunication Technology, Tokyo, 1995.
[Pentland A., Liu A.], « Modeling and prediction of human behaviour ». In Neural Computation, pages 229-242, 1999.
[Perez P.], « Champs markoviens et analyse multi-résolution de l'image : application à l'analyse du mouvement », thèse de doctorat, Université de Rennes 1, IRISA, 1993.
[Perez P., Hue C., Vermaak J., Gangnet M.], « Color-based probabilistic tracking ». In Eur. Conf. on Computer Vision, ECCV'2002, LNCS 2350, pages 661-675, Copenhaguen, Denmark, June 2002.
[Piccardi M., Cheng E.D.], « Multi-frame moving objects track matching based on an incremental Major Color Spectrum histogram matching algorithm ». In IEEE International Workshop on Object Tracking and Classification in and Beyond theVisible Spectrum (OTCBVS'05), San Diego, CA, USA, June 20, 2005.
[Pinhanez C., Bobick. A.], « Human action detection using pnf propagation of temporal constraints ». In M.I.T. Media Laboratory Perceptual Section Report, Vol. 423, 1997.
[Polona R., Nelson R. 94a], « Low level recognition of human motion (or how to get your man without finding his body parts », 1994.
[Polona R., Nelson R. 97], « Detection and Recognition of Periodic, Nonrigid Motion », Int'l J. Computer Vision, Vol. 23, N°3, pages 261-282, 1997.
[Polona R., Nelson R. 94b], « recognizing activities ». In International Conference on Pattern Recognition, 1994.
[Pop I.], Rapport de DEA, sous la direction de H.-H. Nagel, « On the interaction between pedestrians and vehicles using trafic videos », Institut des systèmes cognitifs et algorithmiques de la faculté d'informatique, Université de Karlsruhe, Allemange.
[Puri A., Hang H.M., Schilling D.L.], « An efficient block matching algorithm for motion-compensated coding », Proc. of IEEE Int. Conf. Acoust, Speech and Signal Proc., pages 1063-1066, 1987.
[Rabiner L.R.], « A tutorial on hidden markov models and selected applications in speech recognition ». In Proc. IEEE 77, pages 257-286, 1989.
[Plänkers R., Fua P. 01], « Articulated soft objects for video-based body modeling ». In ICCV, pages 394-401, 2001.
[Plänkers R., Fua P. 03], « Articulated soft objects for multiview shape and motion capture ». In IEEE Trans. Pattern Anal. Mach. Intell., Vol. 25, N°9, pages 1182-1187, 2003.
[Ramanan D., Forsyth D.], « Finding and tracking people from the bottom up ». In CVPR, Vol. 2, pages 467-474, 2003.
[Rangarajan K., Allen W., Shah M.], « Matching Motion Trajectories Using Scale-Space ». In Pattern Recognition, Vol. 26, N°4, pages 595-610, 1993.
[Rao B.S.Y., Durrant-Whyte H.F., Sheen J.A.], « A fully decentralized multi-sensor system for tracking and surveillance », The International Journal of Robotics Research, Vol. 12, N°1, February, 1993.
[Regazzoni C.S, Sacchi C., Gera G.], « Intelligence distribution of a third generation people counting system transmitting information over an urban digital radio link ». In Proceedings of the 2nd Europena Workshop on Advanced Video-based Surveillance Systems, Kingston, UK, pages 53-69, August 2001.
[Reid D.B.], « An algorithm for tracking multiple targets ». In IEEE Trans. on Automatic Control, Vol. 24, N °6, pages 843-854, 1979.
[Remagnino P., Shihab A., Jones G.], « Distributed intelligence for multi-camera visual surveillance ». In Pattern Recognit. : Special Issue on Agent-Based ComputerVision, Vol. 37, N°4, pages 675-689, 2004.
[Rerkrai K., Fillbrandt H.], « Tracking Persons under Partial Scne Occlusion Using Linear regression ». In 8th International Student Conference on Electrical Engineering POSTER 2004, Prague, Faculty of Electrical Engineering, Czech Technical University, May 2004.
[Ricquebourg Y. 93], « Segmentation et suivi d'objets mobiles par modèles structurels adaptatifs ». Master's

thesis, Institut national des sciences appliquées de Rennes, 1993.
[Ricquebourg Y., Bouthemy P.], « A statistical regularization framework for estimating normal displacememts along contours with subpixel accuracy », Lectures Notes in Computer Science, Vol. 970, Vaclav Hlavac et Radim Sara ed., pages 73-81, 6th international conference on Computer Analysis of Images and Patterns, Prague, Czech Republic, septembre 1995.
[Ricquebourg Y. 97], « Analyse de mouvements articulés: mesure et suivi 2D; application à la télésurveillance ». Thèse de doctorat, Université de Rennes I, 1997.
[Rigoll G., Eickeler S.], « Real-time tracking of moving persons by exploiting spatiotemporal image slices ». In IEEE Pattern Analysis and Machine Intelligence, Vol. 22, N°8, pages 797-808, August 2000.
[Rohr K.], « Towards model-based recognition of human movements in image sequences ». In CVGIP: Image Understanding, Vol. 59, pages 94-1, January 1994.
[Rota N.], Rapport de DEA : « Système adaptatif pour le traitement de séquences d'images pour le suivi de personnes », Septembre 1998, sous la direction de Monique Thonnat, et Nicolas Chleq, Projet ORION, INRIA, Sophia-Antipolis.
[Ronfard R., Schmid C., Triggs B.], « Learning to Parse Pictures of People ». In Proceedings on the European Conference on Computer Vision, pages 700-714, Copenhagen, Denmark, 2002.
[Roth S., Sigal L., Black M.], « Gibbs likelihoods for Bayesian tracking ». In CVPR, 2004.
[O'Rourke J., Badler N.], « Model-based image analysis of human motion using constraint propagation ». In IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 2, pages 522-536, 1980.
[Ruiz-del-Solar J., Shats A., Verschae R.], « Real-time tracking of multiple persons », 12th International Conference on Image Analysis and Processing, pages 109-114, September 2003. IEEE Computer Society Washington, DC, USA.
[Shan Y., Sawhney H., Kumar R.], « Unsupervised learning of discriminative edge measures for vehicle matching between non-overlapping cameras ». In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[Shan Y., Sawhney H., Pope A.], « Measuring the similarity of two image sequences ». In Asia Conference on Computer Vision, 2004.
[Shen J., Castan S.], « An optimal linear operator for step edge detection » , Computer Vision, Graphics and Image Processing, Vol. 54, N°2, March 1992, pages 13-17.
[Shi J., Tomasi C.], « Good features to track ». In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pages 593-600, 1994.
[Shimada A., Arita D., Taniguchi R.I.], « Dynamic control of adaptive mixture-of-gaussians background model ». In AVSS, 2006.
[Sidenbladh H., Black M.J., Fleet D.J.], « Stochastic tracking of 3D human figures using 2D image motion ». In Vernon D., ed. : 6th European Conference on Computer Vision (ECCV 2000), Dublin, Ireland, Springer Verlag, pages 702-718, 2000.
[Sidenbladh H., Black M.], « Learning the statistics of people in images and video ». In IJCV, Vol. 54, N°13, pages 183-209, 2003.
[Siebel N.T.], PhD, « Design and Implementation of People Tracking Algorithms for Visual Surveillance Applications », March 2003.
[Siebel N., Maybank S. et al.], « The ADVISOR Visual Surveillance System ». In Proceedings of the ECCV 2004 Workshop, « Applications of Computer Vision » (ACV'04), Prague, Czech Republic, pages 103-111, May 2004, ISBN 80-01-02977-8.
[Siemens], The magazine for Research and Innovation | Fall 2006, Siemens.
[Sigal L., Isard M., Sigelman B.H., Black M.], « Attractive people: Assembling loose-limbed models using non-parametric belief propagation ». In NIPS, 2003.
[Sigal L., Bhatia S., Roth S., Black M.J., Isard M.], « Tracking loose-limbed people ». In CVPR, Vol. 1, pages 421-428, 2004.
[Sminchisescu C., Triggs B. 01], « Covariance scaled sampling for monocular 3D body tracking ». In Proceeding. of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA, Vol. 1, pages 447-454. In IEEE Computer Society Press, December 2001.
[Sminchisescu C., Triggs B. 03a], « Kinematic jump processes for monocular 3D human tracking ». In International Conference on Computer Vision and Pattern Recognition CVPR, Vol. 1, pages 69-76, June 2003.

[Sminchisescu C., Triggs B. 03b], « Estimating articulated human motion with covariance scaled sampling ». In International Journal of Robotics Research, Vol. 22, N°6, pages 371-391, June 2003. Special issue on Visual Analysis of Human Movement.
[Stauffer C., Grimson W.E.L.a], « Adaptive background mixture models for real-time tracking ». In CVPR, 1999.
[Stauffer C., Grimson W.E.L.b], « Learning Patterns of Activity Using Real-Time Tracking ». In IEEE Trans. on Patt. Anal. and Machine Intell., Vol. 22, N°8, pages 747-757, August 2000.
[Stiller C., Konrad J.], « On models, criteria and search strategies for motion estimation in images sequences », IEEE Signal Procesing Magazine, pages 1-41, 1998.
[Streit R.L., Luginbuhl T.E. 93], « A probabilistic multi-hypothesis tracking algorithm without enumeration and pruning », in Proc. of the 6th Joint Service Data Fusion Symposium, pages 1015-1024. Laurel, June 1993.
[Streit R.L., Luginbuhl T.E. 94], « Maximum likelihood for probabilistic multi-hypothesis tracking », SPIE International Symposium, Orlando, USA, April 1994.
[Sudderth E. B., Ihler A. T., Freeman W. T., Willsky A. S.], « Nonparametric belief propagation ». In Proc. Conf. Computer Vision and Pattern Recognition, Vol. 1, pages 605-612, June 2003.
[Taycher L., Demirdjian D., Darrell T., Shakhnarovich G.], « Conditional random people : Tracking humans with crfs and grid fillters ». In CVPR '06 : Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 222-229, Washington, DC, USA, 2006. IEEE Computer Society.
[Thome N.], PhD, « Représentations hiérarchiques et discriminantes pour la reconnaissance des formes, l'identification des personnes et l'analyse des mouvements dans les séquences d'images », INSA Lyon, Juin 2007.
[Thonnat M., Moisan S., Crubezy M.], « Experience in integrating image processing programs ». In H. Christensen(ed.) Proceedings of the 1st International Conference on Vision Systems, Lecture Notes in Computer Science, pages 200-215. Springer, Las Palmas, Gran Canaria, 1998.
[Trivedi M.M., Gandhi T., Huang K.S.], « Distributed interactive video arrays for event capture and enhanced situational awareness ». In IEEE Intell. Sys. Spec. Issue AI Homeland Security Vol. 20, N°5, pages 58-66, 2005.
[Tupin F., Sigelle M], Cours donné à l'ENST Paris, « Définition et simulation d'un champ de Markov », Octobre 2006.
[Valera M., Velastin S.], « Intelligent distributed surveillance systems: a review ». In IEEE Proc. Vis. Image Signal Process. Vol. 152, N°2, pages 192-204, 2005.
[Velastin S.], ADVISOR, « Annotated Digital Video for Surveillance and Optimised Retrieval », EU, IST Programme, IST-1999-11287 with Thales Research Ltd, Reading University, INRIA, Vigitec, Bull, S.A. Velastin, Value £1,533,042, Duration 36 months, 2001.
[Velastin S., Boghossian B., Lo B., Sun J., Vicencio-Silva M.], « Prismatica: toward ambient intelligence in public transport environments ». In IEEE Trans. Syst. Man Cybern. Part A35, Vol. 1, pages 164-182, 2005.
[Viola P., Jones M., Snow D.], « Detecting Pedestrians Using Patterns of Motion and Appearance ». In Proceedings of 9th IEEE International Conference on Computer Vision, pages 734-741, Nice, France, 13-16 October 2003.
[Vu T., Bremond F., Thonnat M.], « Automatic video interpretation: a novel algorithm for temporal scenario recognition ». In Proceedings of the 18th International Joint Conference on Artificial Intelligence, pages 1295-1300 Acapulco, Mexico, 2003.
[Welch G., Bishop G.], « An introduction to the kalman filter ». TR-95-041, Dept. of Computer Science, Univ. of North Carolina at Chapel Hill., 2004.
[Wu T., Matsuyama T.], « Real-time active 3D shape reconstruction for 3D video ». In Proceedings of 3rd International Symposium on Image and Signal Processing and Analysis, vol. 1, pages 186-191, 2003.
[Wren C.R., et al.], « PFINDER : Real-time tracking of the human body ». In IEEE Trans. Pattern Anal.Mach. Intell., Vol. 19, N°7, pages 780-785, 1997.
[Xiang T., Gong S., Parkinson D.], « On the Structure of Dynamic Bayesian Networks for Complex Scene Modelling ». In Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveilflance (VS-PETS). Nice (France), October 2003.
[Ren X., Berg A.C., Malik J.], « Recovering human body configurations using pairwise constraints between

parts ». In Proc. 10th Int'l. Conf. Computer Vision, Vol. 1, pages 824-831, 2005.
[Yamato J., Ohya J., Ishii K.], « Recognizing human action in time-sequential images using Hidden Markov Model ». In CVPR, pages 379-385, 1992.
[Yedidia J.S., Freeman W.T., Weiss Y.], « Understanding belief propagation and its generalizations ». In Technical Report TR2001-22, MERL, 2001.
[Yu Y., Harwood D.], « Human appearance modeling for matching across video sequences », Machine Vision and Applications (2007), Special Issue Paper, Vol. 18, pages 139-149. Springler-Verlag 2007.
[Zhao H.-X., Huang Y.-S.], « Real-time multiple-person tracking system », In International Conference on Pattern Recognition, August 2002.
[Zhao T., Nevatia R., Lv F.], « Segmentation and tracking of multiple humans in complex situations ». In IEEE PAMI, Vol. 9, 2004.
[Zhao T., Nevatia R.], « Tracking multiple humans in complex situations ». In IEEE Trans. Pattern Anal. Mach. Intell. Vol. 26, N°9, pages 1208-1221, 2004.
[Zhu X.L.S., Chau L.], « Hexagon-based search pattern for fast block motion estimation ». In IEEE Trans. On Circuits and Systems for Video Technology, Vol. 12, May 2002.
[Zhu Y., Comaniciu D., Pellkofer M., Koehler T.], « Reliable Detcection of Overtaking Vehicles Using Robust Information Fusion ». In IEEE Trans. Intelligent Transportation Systems, Vol. 7, N°4, pages 401-414, 2006.
[Zhu S., Ma K.], « A new diamond search algorithm for fast block-matching motion estimation ». In IEEE Trans. On Image Processing. Vol. 9, February 2000.
[Zhu S., Wu Y., Mumford D.], « FRAME: Filters, random field and maximum entropy: Towards a unified theory for texture modeling ». PAMI, Vol. 27, N°2, pages 1-20, 1998.