UNIVERSITE PIERRE ET MARIE CURIE - PARIS VI

Mémoire de recherche présenté en vue de l'obtention du :

DIPLOME D'ETUDES APPROFONDIES

Spécialité :

INTELLIGENCE ARTIFICIELLE

RECONNAISSANCE DES FORMES ET APPLICATIONS

par

Nathanaël ROTA

Sujet :

Système adaptatif pour le traitement de séquences d'images pour le suivi de personnes

Mai - Septembre 1998

Direction : **Monique THONNAT : INRIA (Projet ORION)**
Nicolas CHLEQ : INRIA (Projet ORION)
Mr. X : UPMC (lip6)

A l'accueil et au professionnalisme
de Mme M. THONNAT
Au sourire de Mlle M. MARCOS
A l'enthousiasme de Mlle M. CRUBEZY
A l'efficacité de Mlle A. CORTEL
A la vitalité de Mr P. ITEY
Au chapeau melon de Mr C. COINTE
Aux bottes de cuir de Mlle M. RIBIERE
et une mention particulière
A la patience, à la rigueur et
aux enseignements de Mr N. CHLEQ

Mots clés :

- Traitement de séquences d'images
- Classification des problèmes
- Gestion du contexte statique
- Base de connaissances

Résumé :

Le sujet de ce stage fait partie du domaine de l'interprétation de séquences d'images. On recueille à partir d'un signal de caméra une série d'images envoyée vers un logiciel capable de la traiter. Pour ce faire, les images de la caméra sont acquises à une cadence fixe puis numérisées. Ensuite, la séquence d'images est traitée image par image afin d'en extraire le contenu sémantique. Enfin, le résultat est traité dans l'optique d'obtenir une analyse et une interprétation de haut niveau. L'objectif de ce rapport est double. Le premier objectif est de fournir une grille d'explication des problèmes posés par le traitement de séquences d'images pour le suivi de personnes, à ceux souhaitant réaliser un système opérationnel. Ce but sera poursuivi, d'une part, par l'étude d'une classification des problèmes rencontrés, d'autre part, par les différentes réponses apportées par la littérature. Le second but de ce rapport est de montrer la nécessité d'introduire des connaissances dans les traitements. Nous verrons, une solution à base de connaissances incluant une gestion du contexte statique.

Table des matières

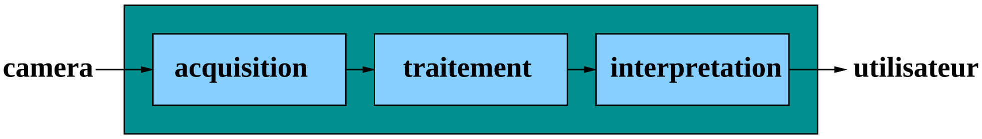

Fig. 1 – *L'interprètation de séquences d'images*

1 Introduction

1.1 Le domaine

Le sujet de ce stage fait partie du domaine de l'interprétation de séquences d'images. On recueille à partir d'un signal de caméra, une série d'images envoyée vers un logiciel capable de la traiter. Pour ce faire, les images de la caméra sont acquises à une cadence fixe puis numérisées. Ensuite, la séquence d'images est traitée image par image puis dans la globalité afin d'en extraire le contenu sémantique. Enfin, le contenu est traité dans l'optique d'obtenir une analyse et une interprétation de haut niveau (cf. figure 1).

- (1) Acquisition et numérisation d'images.
- (2) Traitement de la série d'images.
- (3) Interprétation et analyse.

L'intérêt du sujet sur le plan applicatif a été démontré à de nombreuses reprises. En surveillance d'activité humaine, [Chleq and Thonnat, 1996] ont réalisé un système d'aide à la décision d'opérateur de vidéo surveillance, déclenchant une alarme lorsqu'une situation à risque était détectée. Le même dessein était servi par [Nagell, 1988] en surveillance routière. En analyse de scène sportive [Choi et al., 1997] exploitent les résultats du suivi de personnes afin d'analyser le comportement d'une équipe de football. En analyse de gestes [Pentland, 1995], un niveau plus fin de description permet de réaliser des systèmes capables de comprendre le langage des sourds-muets.

1.2 Le stage

Un des hypothéses de l'interprétation (point (3)) est d'utiliser des résultats fiables afin de résoudre ses propres problèmes. Nous avons donc dans ce stage, réalisé un système de traitement de séquences d'images (point (2)) dont l'objectif est la fiabilité de la représentation statique et dynamique de la séquence traitée. Les détails de ces traitements seront étudiés dans la suite, mais dès lors il nous faut signaler que ce traitement se déroule en trois temps. En premier lieu, une étape de détection extrait de l'image courante les indicateurs du mouvements, témoignant de l'existence possible d'objets mobiles. Puis ces indicateurs sont classés, triés, organisés pour reconnaître les objets de la scène qui nous intéresse. Enfin, ces objets sont suivis d'une image à l'autre, afin de rendre compte de l'aspect dynamique de la séquence (cf. figure 2).

- (1) Détection de mouvement.

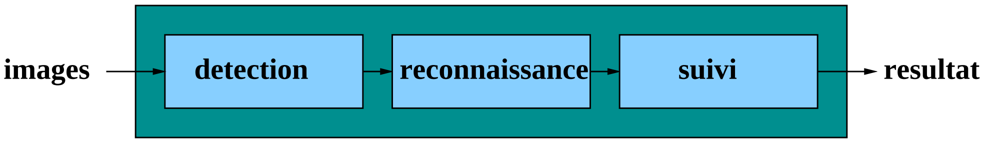

FIG. 2 – *Le traitement de séquences d'images*

- (2) Reconnaissance d'objets.
- (3) Appariement temporel.

La figure 3 illustre un exemple de traitement dans une station de métro. Le problème de représentation se reformule alors en la minimisation de l'erreur faite par le calcul de la représentation automatisée d'une scène par rapport à la réalité de cette scène. Malheureusement cette erreur peut difficilement être mesurée autrement que de façon intéractive et se révèle parfois tout à fait subjective. Malgré tout certains indicateurs objectifs peuvent être décrit pour mesurer l'erreur commise par le système de traitement de séqences d'images. On peut notamment citer des critères tels que le nombre de personnes dans une scène, leur taille, leur position au cours du temps. Ce sont sur ces critères que l'on peut alors définir la robustesse de tel ou tel système. Quoi qu'il en soit, il est clair que notre contrainte de robustesse sera remplie moyennant une bonne détection, une bonne reconnaissance et un bon suivi. Il existe quelques hypothèses qui vont nous permettre de réduire le champ d'étude : tout d'abord nous ne nous intéresserons, pour notre système, qu'à l'étude d'images issues d'une caméra monoculaire fixe dans une scène. Ce système a été réalisé en C à partir d'une architecture déjà existante et surtout d'une librairie d'algorithmes de vision développée dans l'équipe Orion de l'INRIA Sophia-Antipolis .

1.3 L'équipe

Le projet Orion est une équipe pluridisciplinaire, à la frontière des domaines de la vision par ordinateur, des systèmes à base de connaissances et du génie logiciel. L'objectif d'Orion est de concevoir et de développer des techniques et des logiciels pour d'une part, l'interprétation automatique d'images et d'autre part, la réutilisation et le pilotage automatique de programmes. Cet objectif est poursuivi en développant des langages d'expression des connaissances ainsi que des mécanismes d'apprentissage et de traitement de ces connaissances, adaptés à des classes de problèmes spécifiques. Orion se focalise sur l'étude des connaissances qui interviennent dans les deux types de problèmes étudiés : connaissances sur les objets et les scénarios à reconnaître pour l'interprétation automatique d'images, connaissances sur les programmes et leur utilisation pour le pilotage automatique de programmes. Orion étudie plus particulièrement les techniques de représentation des connaissances hybrides (à base de frames, de réseaux sémantiques et de règles de production) ainsi que les techniques de planification. Pour le raffinement de bases de connaissances, nous étudions des techniques d'apprentissage symbolique. Les applications qui intéressent Orion relèvent principalement du domaine de la vision par ordinateur.

FIG. 3 – *Exemple de détection de personnes dans une station de métro*

A terme, l'équipe souhaite élargir le champ d'application des méthodes et techniques développées par Orion à d'autres domaines (traitement du signal, calcul scientifique).

1.4 Le plan du rapport

L'objectif de ce rapport est double. Le premier objectif est de fournir une grille d'explication des problèmes posés par le traitement de séquences d'images pour le suivi de personne, à ceux souhaitant réaliser un système opérationnel. Ce but sera poursuivi, d'une part, dans la première section, sous forme d'une classification des problèmes rencontrés, d'autre part, dans la second section, par les différentes réponses apportées par la littérature. Le second but de ce rapport est de montrer la nécessité d'introduire des connaissances dans les traitements. Nous verrons, dans la troisième section une solution à base de connaissances explicites et déclaratives sur la scène et son contexte.

FIG. 4 – *Problème d'occultation statique partielle*

2 Analyse du problème

Cette section est consacrée à établir une classification des problèmes rencontrés lors de l'élaboration d'un système de traitement de séquence d'images, afin d'en faciliter la conception. Cette classification se décompose en trois types de difficultés. La première regroupe les problèmes d'occultations; la seconde vise à classifier les problèmes de suivi multipistes des objets dans l'image; la dernière est consacrée aux problèmes d'incertitudes.

2.1 Les problèmes d'occultations

L'utilisation d'images issues de caméra monoculaire est à l'origine de nombreux problèmes. Contrairement au cadre de la vision stéréoscopique, notre champ d'études utilise des informations d'entrée en deux dimensions : les images. Or la réalité de la scène est en 3D (voire 3D+t). une quantité d'information non négligeable est donc perdue lors de la projection dans le plan focal de la caméra. Cette projection est à l'origine de nombreux problèmes d'occultations. On choisira, pour catégoriser ces problèmes, les deux axes suivants : un axes qualitatif et un axe quantitatif.

2.1.1 L'aspect qualitatif

Selon l'axe qualitatif, les problèmes d'occultations nous amènent à se demander si un objet est occulté de façon statique ou dynamique. Ceci revient à se demander si un objet, dont on qualifie l'occultation, l'est par un objet immobile de la scène (ex. du mobilier) ou par un autre objet mobile. Les réponses à apporter, algorithmiquement parlant, ne sont en effet pas du tout les mêmes (cf.figures 4 et 5).

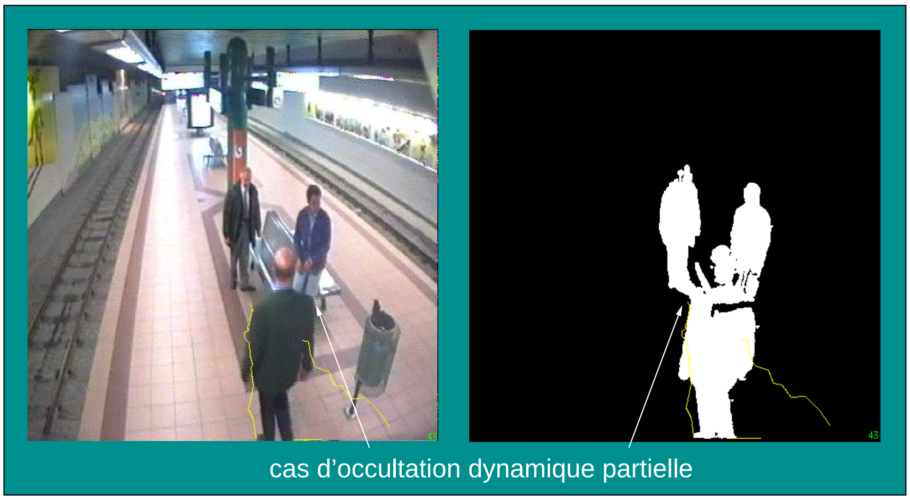

Fig. 5 – *Problème d'occultation dynamique partielle*

2.1.2 L'aspect quantitatif

Selon l'axe quantitatif, les problèmes d'occultation se posent en termes de quantité de surface occultée. On distingue alors deux cas: l'occultation partielle laissant une partie de l'objet considéré visible et l'occultation totale, l'objet ayant dans ce cas totalement disparu de l'image. La encore, les recours algorithmiques ne sont pas du tout de la même nature (cf.figures 4 et 5).

2.2 Les problèmes dynamiques

La dernière catégorie de difficultés est l'ensemble des problèmes liés au suivi. Le terme de piste est utilisé pour qualifier une suite continue de positions d'un objet au cours du temps. On note l'existence de deux couples de comportements problématiques de ces pistes : l'initialisation et la terminaison ainsi que le mélange et l'éclatement.

2.2.1 L'initialisation de piste

L'initialisation de piste est définie par le premier élément d'une suite de caractéristiques représentant un objet.(voir figure 6) En fait, le problème de légitimité de cette initialisation intervient très souvent dans le sens où il n'existe pas forcément d'adéquation entre l'arrivée d'un objet dans la scène et le commencement d'une piste. Un objet peut apparaître sans qu'une piste soit automatiquement ouverte. De même, une piste peut être ouverte sans qu'un nouvel objet soit apparu, dans le cas d'une piste correspondant à un objet précédemment occulté, par exemple.

2.2.2 La terminaison de piste

Le problème dual de celui posé par l'initialisation de piste est celui de la terminaison de piste. Une terminaison de piste est définie par l'existence d'un dernier élément de la suite de positions. On constate que ce dernier élément est

FIG. 6 – *Les problèmes dynamiques*

associé à la disparition d'un objet de la scène, mais la disparition d'un objet de l'ensemble des détections ne signifie pas forcément sa disparition réelle. La encore, ce phénomène peut être un problème d'occultation par exemple. De façon générale, les problèmes d'initialisation et de terminaison de pistes posent le problème des entrées-sorties d'objet dans la scène.

2.2.3 La fusion de pistes

Le problème de fusion de pistes se pose de façon différente des deux problèmes précédents. Une fusion de pistes fait intervenir au moins deux pistes et est qualifiée si il existe une filiation père-fils de deux objets fils de l'image I_{t-1} distincts vers un unique objet père de l'image I_t. Un problème de légitimité se pose encore car certains cas de mélange peuvent avoir un sens et une légitimité, alors que d'autres ne peuvent être qualifiés que d'accident dans les traitements. (voir figure 6) Dans le cas d'un groupe de personnes qui se rejoint en un point de la scène, on peut considérer comme légitime que les pistes définies par les trajectoires de chacune des personnes se mélangent. En revanche, dans le cas d'occultation dynamique (occultation d'un objet mobile par un autre objet mobile) la légitimité du mélange des pistes se pose plus difficilement.

2.2.4 L'éclatement de pistes

Le problème dual du mélange est le problème d'éclatement de piste. On définit un éclatement par une filiation père-fils d'un unique objet père de l'image I_{t-1} vers deux objets fils distincts de l'image I_t. Là encore, certains de ces cas paraissent avoir du sens tel que la séparation d'un groupe de personnes (voir figure 6), dont on aurait suivi une unique piste (à cause de la proximité des personnes dans le groupe). D'autres cas, en revanche, relèvent de l'accident de traitement (un objet réel de la scène bien détecté jusqu'à l'instant t, dont la détection à un instant t+1 donne plusieurs objets).

2.3 Les problèmes d'incertitude

Une catégorie de problèmes du domaine est la gestion de l'incertitude à toutes les étapes des traitements. En fait, il existe différents niveaux d'incertitude que le système doit savoir gérer. Le plus fin est le bruit inhérent aux images numérisées. Les effets naturels dans la scène correspondent à un second niveau d'incertitude et le grain le plus épais d'imprécision est lié aux objets eux-mêmes.

2.3.1 L'hypothèse du bruit gaussien

Un problème récurrent du traitement d'images est le problème du bruit introduit par la numérisation préalable des images. Une numérisation de basse résolution introduit une faible qualité d'image qui pose problème lors des étapes de détection et de reconnaissance. Or une numérisation de haute résolution pose des problèmes de manipulation d'information, car il s'agit alors de gérer d'importants volumes de données, ce qui est problématique dans l'élaboration de systèmes embarqués ou de système temps réel, par exemple. Par conséquent nous sommes contraint à utilisé des images de basse résolution. L'hypothèse, la plus communément envisagée est celle d'un bruit gaussien appliqué à chaque pixel de l'image. Dans [Baumberg and Hogg, 1995], l'utilisation d'un modèle de ce type durant l'étape de détection permet d'estimer le bruit et ainsi de fiabiliser les résultats. Dans [Du et al., 1993], les auteurs font mention d'un modèle de bruit aléatoire combiné au modèle d'objet qu'ils utilisent. Leur modèle d'objet est défini par un polyèdre 3D, dont chacune des arêtes est soumise à une «contamination aléatoire du bruit», définie par deux degrés de liberté supplémentaires. Il en résulte une intéressante flexibilité du modèle d'objet.

2.3.2 Les phénomènes naturels

L'utilisation d'images réelles pose, en soi, un problème. En effet, de nombreux algorithmes de traitement d'image élaborés en laboratoire se révèlent inadéquats dans le cadre d'images réelles (cf.figure 7). Les phénomènes naturels tel que les ombres d'objets, les réflexions d'objets sur le sol et autres surfaces propices, l'obscurité et les changements d'illumination sont des problèmes majeurs du traitement de séquences d'images naturelles. [Koller et al., 1993] propose, pour faire face aux problèmes d'ombrage, un modèle d'illumination basé sur l'existence d'une source lumineuse à l'infini immobile au cours du temps (durant la séquence). Une ombre théorique de l'objet est alors calculée et prise en compte lors de l'instanciation du modèle d'objet. Cette méthode est très intéressante mais ses hypothèses sont très strictes : elles conviennent, en effet, à des scènes extérieures (source lumineuse à l'infini) et de courte durée(source lumineuse immobile au cours du temps).Dans [Karmann and Brandt, 1989], les auteurs proposent une solution possible aux problèmes de changement d'illumination de la scène par une méthode d'intégration de l'image de référence utilisée.

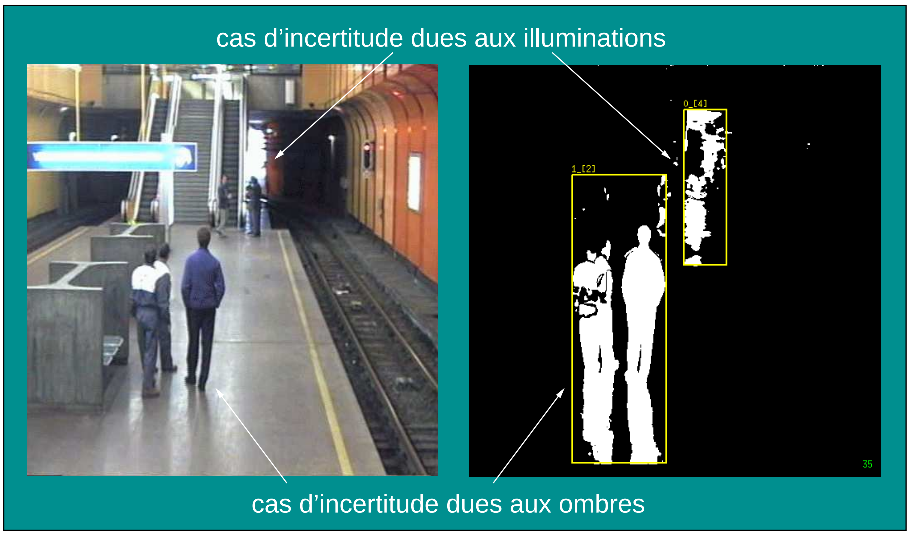

FIG. 7 – *Illustration des problèmes d'incertitudes*

2.3.3 Les incertitudes objet

Les objets mobiles peuvent parfois apparaître comme indésirables. Les objets mobiles naturels tels que les objets mus par le vent (ex: feuillage des arbres) et la pluie doivent être pris en compte. Les artefacts mobiles sont aussi problématiques. Que leur existence corresponde à un phénomène périodique, comme un escalator ou comme un métro (avec une périodicité différente) ou qu'ils révèlent un phénomène chaotique (apparition d'un véhicule et tout autre accident dans l'image), les artefacts mobiles sont un problème sur lequel la littérature est assez discrète.

2.4 Bilan et conclusions

Nous venons de voir, dans cette section la majorité des problèmes que pose l'élaboration d'un système de traitement de séquence d'images pour l'interprétation de scène. Bien que de nature, tout à fait différents, les trois axes, que sont les problèmes d'occultations, les problèmes de suivis multipistes et sont les problèmes d'incertitudes doivent rester au premier plan des préoccupations du concepteur de tels systèmes. Le problème se pose alors de trouver une architecture générale permettant de savoir, à quels étapes de la séquence des traitements, quel problème peut être résolu. Pour concevoir cette architecture, nous allons voir dans la suite quelques éléments de réponses.

FIG. 8 – *La division classique du processus*

3 État de l'art

Le traitement de séquences d'images pour le suivi de personnes est un problème réel qui légitime son unité autant sur le plan théorique que sur le plan applicatif en ceci qu'il cherche à reconstituer par le calcul les évolutions d'une scène donnée au cours du temps. Malgré cela sa résolution soulève une pléïade de problèmes, de domaines, de type d'information, de moyens algorithmiques très différents qui obligent à briser cette unité au profit d'une division des problèmes. Ainsi, les techniques employées pour résoudre les différents problèmes évoqués précédemment pourront plus rigoureusement trouver leur place dans le cours des traitements.
La division du problème, aussi souhaitable qu'arbitraire, est souvent évoquée dans la littérature. Bien que le nombre de sous problèmes évoqués ne soit jamais le même, la réalité qu'il recouvre est souvent semblable. Nous introduirons ici une division en trois points :détection, reconnaissance et suivi. (cf. figure 8)

Le premier point à aborder est que cette division en sous-problèmes amène un élément de réponses quant à la structuration de l'information. La représentation de l'information est un problème en soi, et est hautement contraint par l'utilisation d'algorithmes particuliers. Du fait que les traitements à effectuer sont d'ordre très divers, les niveaux de représentations de l'information le sont aussi. On trouve en premier lieu l'image couleur ou monochrome puisque c'est le point de départ du problème, mais aussi des images binaires, les contours, les régions et d'autres objets plus ou moins abstraits. Il est alors important de définir le niveau de structuration de l'information qui prime dans un sous problème particulier. En fait, le traitement de séquence d'images pour la représentation de scène se caractérise souvent par un double processus. Plus l'on abstrait l'information au cours des traitements, tentant de rendre compréhensible son contenu, plus la taille de ses données diminue. L'information croit en contenu informatif en même temps que décroît la taille de ses données.

Compte tenu de la division du problème envisagé la suite sera divisée en trois points correspondants aux trois sous problèmes: détection, reconnaissance et suivi.

3.1 La détection

La première étape du traitement de séquence d'images est la détection de mouvement témoignant de l'existence possible de régions mobiles dans l'image courante. Ces indicateurs fournissent alors la base des traitements algorithmiques futurs. On trouve dans la littérature trois familles de méthodes caractérisées par la nature de l'indicateur qu'elles extraient : l'approche basée sur les vitesses, l'approche basée sur l'extraction de primitives et une approche basée l'extraction de régions qualifiées 'd'intérêt'. La suite décrit ces trois familles de méthodes de détection de régions mobiles.

3.1.1 L'approche vitesse

On cherche dans cette approche à qualifier les vitesses dans la scène. Une approche, basée sur l'utilisation des champs de vitesses, est développée dans [Bouthemy, 1988]. Son idée repose sur l'hypothèse d'invariance entre t et $t+dt$. i.e. la fonction d'intensité lumineuse en un point (x, y, t) est identique en $(x + dx, y + dy, t + dt)$. On definit alors une fonction erreur appelée DFD «Deplaced Frames Difference» (1) par:

- (1) $DFD(x, y, t) = f(x, y, t) - f(x + dx, y + dy, t + dt)$
- (2) $DFD(x, y, t) = 0 \ \forall x \ \forall y \ \forall t$

On cherche alors à minimiser la DFD pour tout point (x, y) de l'image à l'instant t donné, pour obtenir une représentation des vitesses 3D projetées dans le plan image. L'équation ainsi obtenue (2) est connue sous le nom de «Équation de Contrainte du Mouvement Apparent (ECMA)». On peut reprocher à cette méthode, outre la lourde phase d'optimisation de la DFD en chaque point de l'image, la mise en oeuvre du procédé.

On trouve aussi dans cette famille la méthode par estimation du flot optique en fonction des variations spatio-temporelles de la fonction d'intensité lumineuse.

3.1.2 L'approche région

L'approche région se caractérise par l'extraction dans l'image courante de régions appelées «blob». Ces régions, définies par un ensemble de pixels connexes, correspondent dans l'image à des zones d'intérêt au regard du critère indicateur, en l'occurence les pixels ayant une valeur différente entre un instant et un autre. L'ensemble de ces méthodes peuvent être formalisées ainsi:

- $I_{resultat} = Combi(I_0, ..., I_t)$ (cf. figure 9)

où $I_{resultat}$ est l'image résultante $I_0, ..., I_t$ l'ensemble des images précédentes, $Comb$ est une combinaison de fonctions parmi : Différence, Abs(valeur absolue), Min, Max, Et(et logique), Ou(ou logique)...

L'instance la plus utilisée [Chleq and Thonnat, 1996] [Baumberg and Hogg, 1995] de cette classe proposé par [Wenstop, 1983] est la méthode utilisant une différence entre l'image courante et une image de référence, souvent I_0.

- $I_{resultat} = Abs(Diff(I_0, I_t))$ (cf. figure 10)

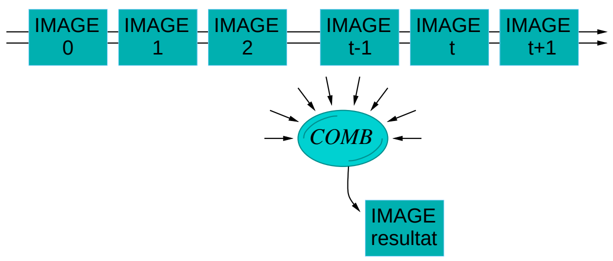

FIG. 9 – *Combinaison d'images*

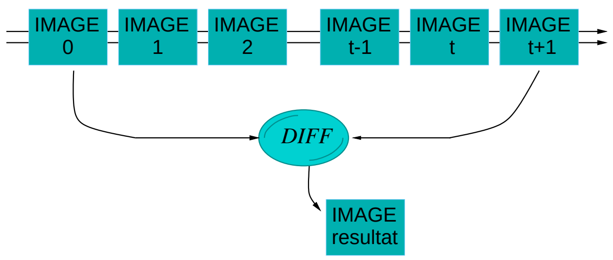

FIG. 10 – *Combinaison d'images par image de référence*

On concède à cette méthode sa simplicité de mise en oeuvre.
L'inconvénient majeur et le problème difficile de cette méthode est la mise a jour de l'image de référence.

L'autre instance très utilisée de la classe de méthode de combinaison d'images est la différence d'images successives [Jain et al., 1979]:

- $I_{resultat} = Max(Abs(Diff(I_t, I_{t-1})), Abs(Diff(I_{t+1}, I_t)))$ (cf. figure 11)

Ces méthodes sont souvent utilisées dans les cas d'applications ne pouvant pas se baser sur une image de référence. Cette méthode est, de la même façon, simple à mettre en oeuvre, mais comporte des inconvénients supplémentaires. En effet, cette méthode ne prend pas en compte les changements et mouvements des régions uniformément colorées, seules les régions texturées sont détectées. [Ricquebourg, 1993]

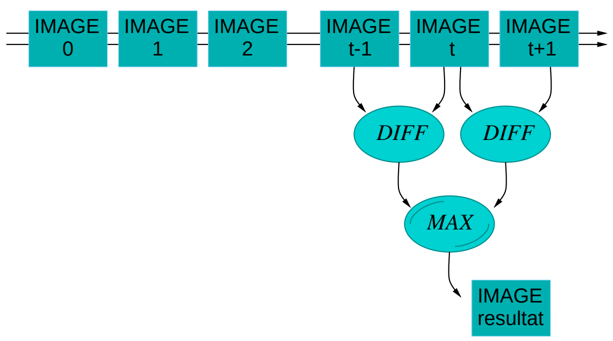

FIG. 11 – *Combinaison d'images par différence d'images successives et Max*

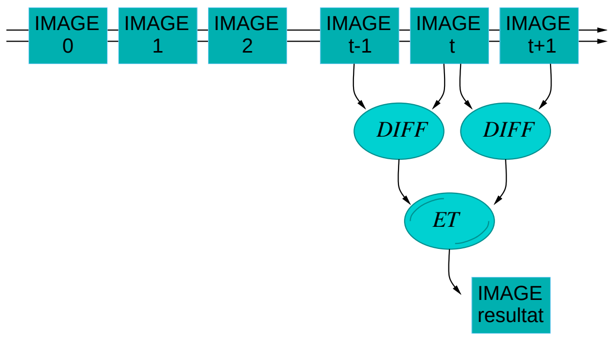

FIG. 12 – *Combinaison d'images par différence images successives et Et*

Dans [Lenz and Gerhard, 1983] l'opérateur Max est remplacé par un «et» logique Et. Cette méthode permet de recueillir un résultat proche des résultat de la détection de contours sans avoir recours au calcul du gradient d'intensité de l'image.

- $I_{resultat} = Et(Abs(Diff(I_t, I_{t-1})), Abs(Diff(I_{t+1}, I_t)))$ (cf. figure 12)

3.1.3 L'extraction de primitives

Cette approche consiste à recueillir dans l'image courante des points particuliers des objets de la scène. Ces points particuliers sont appelés primitives et sont les arêtes, contours ou autre points caractéristiques. Dans [Du et al., 1993] [Koller et al., 1993] les auteurs s'intéressent à la détection des véhicules, ils utilisent pour cela l'extraction de primitives sur les voitures telles que les angles du toit et du capot. Ces primitives sont obtenues grâce à l'analyse de la norme du gradient de l'image et toute autre méthode permettant l'extraction de contours (filtre de Sobel, seuillage par hysteresis, Canny, suivi de contour, etc ...). Cette famille de méthodes possède comme grand avantage de produire des indicateurs ayant un contenu sémantique (i.e. des points précis d'objets), ce qui est un avantage pour les traitements ultérieurs.

3.1.4 Bilan

Le point commun de toutes ces méthodes est donc l'extraction à chaque image de la séquence d'indicateurs qui témoignent des changements dans l'image. Il faut dès lors prendre garde à l'assimilation abusive entre changement et mouvement. Comme nous le verrons plus tard sous certaines hypothèses les deux notions se confondent, mais il existe aussi des cas de changement n'étant pas causé par du mouvement et réciproquement. L'extraction d'indicateurs de changement est donc une étape importante de l'ensemble des traitements sur la séquence d'images car le reste des traitements est contraint par la nature de ces indicateurs. Les algorithmes applicables aux primitives ne le sont pas forcement aux régions, et réciproquement.

3.2 La reconnaissance

Les indices émergeants de la première étape doivent être organisés (filtrage, régularisation, ordonnancement, groupement, etc...). Le but est d'obtenir à la fin de la phase de reconnaissance un ensemble d'objets potentiellement utilisables par les outils traitants les aspects dynamiques. Il s'agit en fait de raffiner, voir établir complètement, la décision sur l'intérêt de zones particulières de l'image. Cette étape pose le problème de ce que l'on cherche à voir dans chacune des images de la séquence. La première difficulté est donc de pouvoir catégoriser ce que l'on veut reconnaître. Le second problème est de savoir comment le reconnaître (i.e. quelles techniques employer?). Nous ne développerons pas ce dernier point; on trouve dans la littérature différentes optiques telles que des contraintes sur l'organisation des indicateurs ou l'optimisation d'une fonction décrivant l'ensemble non-ordonné des indices [Du et al., 1993] [Koller et al., 1993]. Pour le premier problème, de la même façon que nous avons procédé à l'analyse du problème de détection du point de vue de la nature des objets détectés, l'analyse du problème de reconnaissance sera faite selon la nature des objets reconnus. Guidé par les contraintes algorithmiques, les objets à reconnaître sont caractérisés par leur potentiel de déformation intrinsèque, leur rigidité. Les objets rigides sont donc les objets dont la forme ne varie pas au cours du temps, seule la projection dans le plan image peut différer. A une rotation, une translation et une homotétie près un objet rigide conserve sa forme propre. En fait, ce sont les paramètres de ces opérations que les méthodes qui utilisent ces objets visent à estimer pour pouvoir instancier un modèle dans l'image courante. A l'opposé, les objets non rigides ne peuvent pas être considérés de forme constante.

Nous organiserons l'étude des différents modèles d'objets rencontrés selon deux axes : les modèles d'apparence d'objets et les modèles d'objets réels.

3.2.1 Les modèles d'apparence d'objets

Cette famille de méthodes basées sur l'apparence des objets est souvent employée dans les cas où il est difficile de qualifier l'objet pour des raisons de bruit trop important ou dans les cas d'objets trop complexes ou trop divers. Dans [Bremond, 1997] l'étape de détection fournit un résultat bruité qui ne permet pas l'instanciation d'un modèle d'objet complexe. Le cadre de travail reste donc dans le plan image et les indices en deux dimensions. Dans [Choi et al., 1997], les auteurs utilisent un modèle d'apparence d'objets. Le champ d'application est l'analyse de scène sportive et notamment de football. Plutôt qu'un contraignant modèle d'objet, les auteurs choisissent une décision basée sur l'analyse des couleurs. Dans [Woodfill and Zabith, 1991], les auteurs justifient leur approche par l'impossibilité de définir un modèle d'objet pour leur champs d'études. Il n'y a pas de reconnaissance d'objets mais plutôt reconnaissance de zone. [Baumberg and Hogg, 1995] définit un modèle d'apparence d'objet non rigide complexe pour l'analyse du mouvement humain. Les auteurs définissent un modèle 2D de forme grâce à l'extraction d'un vecteur de points du contour de la projection de l'humain dans le plan image, se basant pour cela sur un modèle de distribution appelés PDM : «Point Distribution Model» emprunté au travaux de Cootes. L'idée est de mesurer la différence dx entre un

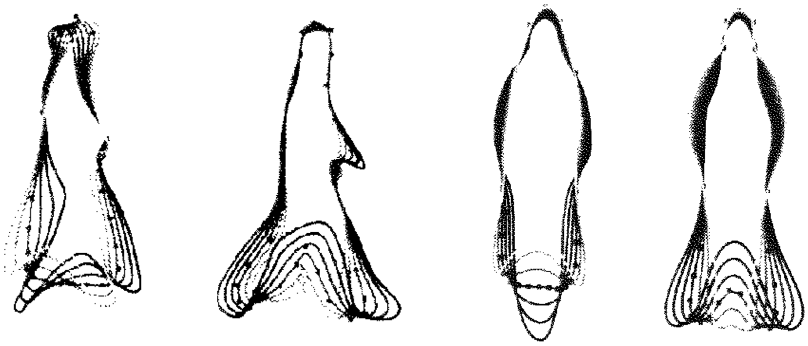

FIG. 13 – *Le model d'apparence d'objet non-rigide de Hogg(95)*

modèle quelconque et le modèle moyen. i.e. $dx = x - x_{moyen}$ où x est un vecteur de points obtenu et x_{moyen} est le vecteur moyen. L'analyse des vecteurs dx donne le mode de variation le plus significatif du modèle. La forme du contour est ensuite optimisée par l'utilisation de b-splines cubiques dont les points de contrôle sont chacun des points du contour définis plus haut. Les résultats sont mesurés sur le critère de compacité du modèle. Des résultats intéressants sont obtenus sur un ensemble d'apprentissage réduit à des personnes marchant latéralement. Sur un ensemble d'apprentissage quelconque les résultats sont plus modestes mais encore utilisables (cf.illustration 13).

L'avantage de l'utilisation de modèles d'objets basés sur l'apparence est de pouvoir être instanciés très tôt dans les traitements, juste après la détection qui, par définition, recueille des indices d'apparence. Le défaut majeur est de ne pas tenir compte de l'aspect volumique ou de caractéristiques intrinsèques de l'objet.

3.2.2 Les modèles d'objets réels

Les modèles d'objets réels utilisent quant à eux des caractéristiques de l'objet. Les méthodes basées sur l'utilisation de modèle d'objets réels exploitent souvent des propriétés plus naturelles. A titre d'exemple les objets rigides, comme les véhicules, peuvent être simplement décris, sur des considérations volumiques. Dans [Koller et al., 1993], l'objet consiste en un polyèdre 3D modélisant un véhicule par douze degrés de liberté. L'ensemble de segments de contours obtenus par l'étape de détection est alors mis en correspondance avec les arêtes du modèle polyèdrique par un processus itératif emprunté aux travaux de [Lowe, 1987]. L'idée est de minimiser un critère, défini par une mesure de distance (distance de Mahalanobis), pour chaque arête du polyèdre (cf.illustration 14).

Les modèles objets non rigides sont tous les objets dont la forme propre n'est pas constante au cours du temps (on entend par là: à une rotation, translation et homotétie près). C'est le cas de la plupart des objets naturels: végétaux, animaux, humains. Le traitement de tels objets requiert l'utilisation de modèles et de techniques particulièrs. Mais le problème lié aux personnes est double et peut être défini ainsi. D'une part, soit une personne donnée, les

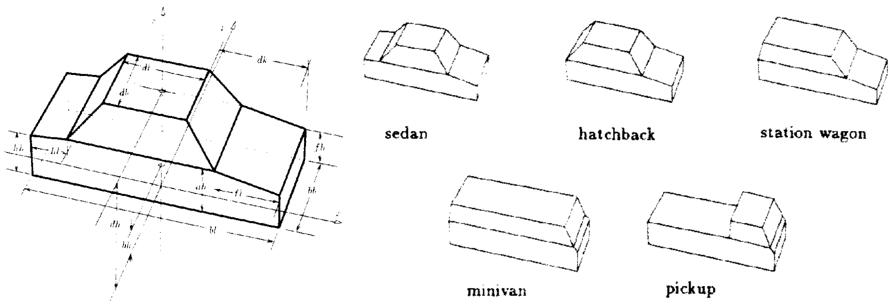

FIG. 14 – *Le modèle d'objet réel rigide de Koller(93)*

caractéristiques la décrivant ne sont pas constantes au cours du temps. En effet il existe une forte variation de formes pendant les déplacements (oscillations des bras et des jambes), voire même une variation de couleur. D'autre part, soient deux personnes données, les caractéristiques les décrivant ne sont pas forcément les même. Ce second point pose un problème quant à la généralité du modèle.

Dans [Atika, 1984], l'auteur utilise un modèle de personne défini par un squelette de six segments (deux bras, deux jambes, torse et tête).

Dans [Chen and Lee, 1992] un modèle de squelette de dix-sept segments est utilisé. Un modèle volumique constitué de quatorze cylindres elliptiques est utilisé dans [Rohr, 1994].

L'avantage de ces modèles d'objets réel est d'aller dans le sens d'une indépendance entre les contraintes du modèle et les algorithmes de reconnaissance. Ceci permet de garder un contrôle sur l'instanciation que ne permet pas les méthode à base de modèle d'apparence d'objets. L'inconvénient est d'éloigner la reconnaissance de l'objet de l'extraction d'indices de mouvement; de nombreuses étapes sont nécessaire pour instancier un modèle d'objet réel par un ensemble de détections.

3.2.3 Bilan

L'organisation des indices émergeant de la détection en objets potentiellement utilisables est un problème, basiquement complexe, rendu encore plus difficile par l'introduction de modèles d'objets. Néanmoins, plus l'on cherche une qualification précise du contenu de la scène, plus l'utilisation de connaissances *a priori* devient cruciale. Comme nous l'avons vu, l'emploi de modèles d'objets permet d'établir, lors de la reconnaissance, la décision sur l'existence ou non d'objet particulier dans la scène. Ce point est primordial dans la mesure où les traitements des étapes suivantes (le suivi en particulier) seront conditionnés par la qualité de la reconnaissance.

3.3 L'appariement temporel et le suivi

Contrairement aux étapes précédentes dont les traitements pouvaient être qualifiés de statique dans le sens où ils avaient pour objet l'analyse d'une seule et unique image, les méthodes qui seront exposées dans la suite ont une dimension supplémentaire à traiter : le temps. Le but des traitements de cette étape est d'inclure les résultats de la détection et de la reconnaissance dans la continuité de la séquence d'images. Pour cela plusieurs techniques sont abordées. Le point discriminant ces méthodes est l'usage fait des mesures courantes (les mesures faite sur l'image courante). Une première famille de méthodes vise à mettre en correspondance ces mesures avec celles obtenues précédemment, à l'opposé d'autres techniques utilisent ces mesures pour vérifier la cohérence d'une prédiction faite grâce à l'ensemble des résultats des images précédentes. La suite est divisée en deux points correspondants à ces deux paradigmes.

3.3.1 La mise en correspondance

La première famille de méthodes évoquées par [Meyer and Bouthemy, 1992] est la mise en correspondance directe. On cherche ici à placer les mesures reçues des étapes précédentes dans la continuité des mesures réalisées sur les images précédentes. Formellement, il s'agit de prolonger la suite des détections par les détections obtenues à l'image courante en minimisant l'erreur commise par cet appariement. Nous détaillerons plus précisément cette approche dans la troisième partie puisque c'est l'orientation prise.

3.3.2 Le filtre de Kalman

L'autre famille très populaire de résolution de problème de suivi est l'estimation par filtre de Kalman. La philosophie générale du procédé s'oppose à la précédente approche par l'utilisation des mesures courantes. Dans ce cas le traitement se fait en deux temps. Premièrement, on génère une prédiction de ce que pourrait être le résultat courant à partir de l'ensemble des résultats précédents, puis l'on compare le résultat prédit avec les mesures réellement obtenues. Les données des étapes ultérieures ont alors un rôle de vérification. Plus précisément, le filtrage de Kalman utilisé par [Baumberg and Hogg, 1995] et [Choi et al., 1997] et [Ricquebourg, 1997] se définit comme un filtre récursif produisant une prédiction de l'état courant E_t du système à partir de l'ensemble des mesures $M_{t-k}, ..., M_{t-1}$ si l'évolution dynamique du système est considéré comme linéaire.

- $\frac{dX(t)}{dt} = F(t) \times X(t) + B(t)$ où

F(t) représente l'évolution du système.
X(t) est un vecteur d'état.
B(t) un terme tenant compte du bruit et des termes non-homogènes souvent nuls.

La résolution de cette équation puis un passage en mesure discrète donne $X(t)$ en fonction de $X(t-1)$ et un terme correspondant à l'évolution dynamique du système. Ce qui permet de calculer, avec une précision dépendant du modèle de mouvement considéré, l'état du système à l'instant t.

L'intérêt est double; obtenir une prédiction à l'étape courante indépendante des mesures réalisées et mesurer la fiabilité du modèle de mouvement grâce aux mesures obtenues. L'inconvénient, outre la délicate mise en oeuvre du procédé, est aussi double. D'une part, l'utilisation du filtre de Kalman nécessite une modélisation de l'évolution dynamique des objets de la scène (ce qui n'est pas trivial puisqu'elle est associée au mouvement humain), d'autre part, comme tout processus itératif sa sensibilité aux valeurs initiales est grande.

3.3.3 Bilan

Une fois encore, l'apport de modèle (de mouvement, cette fois) permet d'augmenter la robustesse des traitements. Toute fois, autant l'intégration d'un modèle d'objet en reconnaissance était concevable, autant l'élaboration d'un modèle de mouvement global d'une scène parait plus délicat.

3.4 Les conclusions

La division des problèmes introduite dans cette section comprend donc trois points que sont la détection, la reconnaissance et le suivi, chacun ayant des hypothèses, des outils algorithmiques, des domaines très différents. Le but de cette division nous l'avons vu est double. D'une part, ceci nous permet de reprendre un par un les problèmes évoqués dans la première section et d'insérer dans le fil des traitements les algorithmes permettant de les résoudre. D'autre part, cette division des problèmes permet de comprendre l'importance d'introduire des modélisations des objets et des mouvements afin d'augmenter la robustesse des traitements de chacun des sous problèmes.

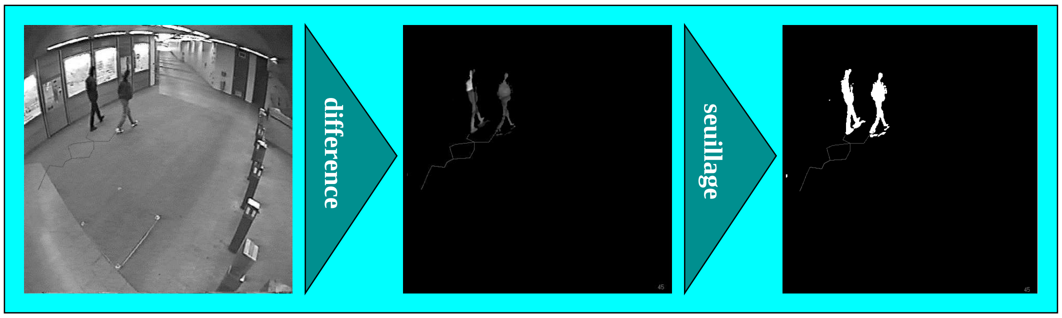

FIG. 15 – *la détection par seuillage paramétré*

4 Solution à base de contexte

Comme nous venons de le voir, l'élaboration d'un système capable de gérer efficacement des séquences d'images pour le suivi de personnes ou la représentation de scènes repose sur une division préalable des problèmes. Cette séparation des différents problèmes nous est imposée (cf. section 1.) par l'introduction, dans les traitements, d'algorithmes capables de résoudre certains problèmes particuliers. Nous allons voir dans cette section, que la résolution des problèmes évoqués dans la section 1 requiert non seulement une classification claire des difficultés et une subdivision des taches, mais aussi un apport de connaissances externes. C'est sur cette idée que repose une nouvelle solution logicielle, élaborée durant le stage, dont le fondement est l'apport de connaissance divisible en deux points que sont l'information externe concernant les humains et l'information externe concernant la scène 3D.

Nous verrons dans un premier temps une solution réalisée en 1996, dans le cadre du projet ESPRIT PASSWORD dont l'architecture fut le point de départ de la réalisation du stage. Ensuite, nous détaillerons cette réalisation ainsi que les différents modèle de connaissances externes qu'elle utilise.

4.1 Le logiciel de départ

Nous allons, dans ce qui suit, décrire le système opérationnel de traitement de séquences d'images pour le suivi de personnes qui est à la base des réalisations actuelles. Ce système constitue un point de départ en matière d'architecture logicielle et en matière de sensibilité aux changement de contexte (cf.figure 2) .

4.1.1 La détection

La technique utilisée pour l'extraction de régions d'intérêt est la méthode de combinaison d'images avec image de référence. Après une phase de seuillage paramétré (cf.figure 15) et une série de filtrages (érosion et dilatation), chacune des images fournit un ensemble de régions de taille et de caractéristiques très diverses.

4.1.2 La reconnaissance

L'étape de reconnaissance utilise un modèle d'objets basé sur l'apparence définie dans le plan image 2D par deux degrés de liberté non-contraints que sont la hauteur et la largeur. L'algorithme décidant de l'existence d'un objet consiste en un appariement global des régions. Les techniques utilisées pour recomposer les régions d'intérêt sont basées sur deux critères: Deux régions dont les boites englobantes se recouvrent sont fusionnées. Les régions faiblement éloignées sur l'axe vertical sont aussi fusionnées. Cette technique a pour but de lutter contre les problèmes de seuillage trop contraint ainsi que les problèmes d'occultations statiques partielles.

4.1.3 L'appariement temporel

La méthode de suivi se base sur un critère de recouvrement. Deux régions de deux images consécutives sont considérées comme provenant du même objet si il existe un recouvrement spatial entre les deux. Il en résulte un graphe temporel G(S, A) des objets de la scène où S constitue l'ensemble des objets détectés à chaque image et A l'ensemble des filiations entre les détections au cours du temps. Cette filiation mettant en évidence la correspondance d'un objet de la scène entre deux images consécutives. Cette représentation très intéressante constitue un des points forts du système.

4.1.4 Les conclusions

Après avoir détaillé l'architecture de ce logiciel à travers la division classique des problèmes exposés dans la section 2, nous allons analyser ses lacunes à travers la grille de lecture présentée dans la section 1. Les problèmes d'incertitudes, tout d'abord, sont globalement gérés par un savant paramétrage. Le choix d'un «bon» paramètre lors des phases de dilatation et d'érosion permet de minimiser l'influence du bruit inhérent aux images. Le choix d'un «bon» paramètre lors de la phase de seuillage de l'image de différence permet aussi de réduire l'effet des ombrages sur le sol. Malheureusement, ce paramètre de seuillage intervient aussi dans la résolution des problèmes liés à l'incertitude objet. Un seuil élevé (30 sur 255) réduit l'effet des ombres mais augmente l'incertitude de la détection des objets eux-meme. Inversement, un seuil bas (15 ou 20 sur 255) donne une détection nette des objets mobiles mais rend le système très sensible aux ombres. Globalement, le problème d'incertitude est contraint par l'utilisation d'un paramètre «aveugle» réglé de façon externe. De plus, l'utilisation de paramètre diminue l'adaptabilité du logiciel, dans le sens où seule l'expertise de spécialistes en vision permet de définir de tels seuils.

Les règles de fusion des régions mobiles (proximité des régions) sont une tentative de gestion des problèmes d'occultations statiques applicables dans des cas de scènes très précises. Cette hypothèse est valide, si les objets de la scène évoluent dans le même plan avec une caméra respectant une certaine distance (une dizaine de mètres) et une certaine inclinaison (environ 45 degrés) par rapport au plan de référence de la scène. Dans les autres cas les règles de fusion ne sont pas très robustes. Les occultations dynamiques ne sont absolument pas gérées.

Les problèmes de suivi sont contingentés par l'utilisation du critère de recouvrement spatial. Si deux régions de deux images consécutives ont des pixels communs alors ils sont considérés comme provenant du même objet. Ce critère est valide dans une séquence d'images avec une fréquence d'acquisition d'image suffisante si l'on considère des personnes marchant normalement. Expérimentalement, ce critère se révèle suffisant pour des scènes peu encombrées, constituées de personnes indépendantes. Dans le cas de scènes encombrées le graphe d'appariement temporel devient rapidement inexploitable. L'existence de groupe de gens ne peut être traitée par ce critère dans la mesure où plusieurs objets partagent fréquemment le même espace du plan image. Les problèmes de mélange et d'éclatement de pistes sont parmi les plus lourds handicaps de ce système.

Pour résumer, le défaut majeur du système est sa faible adaptabilité au changement de scènes. Il est clair, que certaines des orientations prises sont basées sur des hypothèses contextuelles (réparation des occultations statiques par exemple). La non-utilisation de modèle d'objet, bien qu'elle fournisse un résultat non biaisé, est aussi problématique dans le sens où les résultats obtenus sont difficilement qualifiables rendant particulièrement difficile l'étape de suivi (Lorsqu'un éclatement de piste est observé, par exemple, il est difficile de décider si il est légitime ou non). Il s'impose alors de relier ces hypothèses à un cadre précis à partir duquel puisse être qualifiées ces connaissances. De façon générique, ce cadre correspond à celui de connaissances contextuelles. Nous allons donc dans la suite détailler l'organisation de telles connaissances utilisables par un système informatique capable de résoudre plus efficacement les problèmes de ce logiciel.

4.2 Les contributions

La suite constitue la contribution du stage a l'élaboration d'un système de traitement de séquences d'image adaptatif et robuste. Nous avons montrer plus avant la nécessité d'introduire de l'information contextuelle dans les traitements. Nous verrons dans un premier point quelques définitions de ce que nous appellerons par la suite le contexte, puis nous détaillerons le nouveau système.

4.2.1 Le contexte statique

[Strat, 1993] décrit les différents niveaux informations engagées dans le traitement de séquences d'images. L'auteur distingue trois niveaux: les connaissances principales, l'information contextuelle et l'information factuelle. Les connaissances principales sont toujours valides, ne changent pas selon l'application et durant le processus. L'information contextuelle est invariante durant le processus, mais varie en fonction de l'application. Les informations factuelles, enfin, varient non-seulement à chaque application mais aussi durant les traitements. Dans [Bremond, 1997] on trouve une définition plus formelle de la notion de contexte. Selon l'auteur, peut être considérée comme information de contexte, toute information qui :

FIG. 16 – *Les scènes*

- (1) conserve ses valeurs constantes durant le processus.
- (2) ne conserve pas forcément ses valeurs constantes d'un processus à un autre.

Nous prendrons dans la suite, cette définition comme référence. En termes logiciels les données qui vérifient ces critères correspondent à deux notions. La première constitue la généralisation de ce que l'on cherche à reconnaître, c'est à dire des humains. Un modèle d'humain s'impose donc. La seconde notion qui remplit les critères contextuels correspond au décor dans lequel on cherche à reconnaître quelque chose. Cette notion sera appelée « contexte statique».

En pratique, le contexte statique se définit par un ensemble de référentiels et un ensemble d'objets. Les objets sont ceux de la scène réelle considérés comme fixes (cf. figure 16). Les référentiels sont ceux impliqués dans les traitements; on trouve au minimum le référentiel image 2D et le référentiel scène 3D. Les connaissances contextuelles sont alors organisées autour de ces deux thèmes avec tous les processus de mise en correspondance souhaitable. Les processus minimum sont les mécanismes de passage d'un référentiel à un autre et aussi les correspondances des objets réels dans les divers référentiels. L'existence d'une telle organisation apporte la possibilité d'obtenir des informations supplémentaires sur les objets mobiles dans les autres référentiels que celui de l'image 2D.

En pratique, il est alors possible d'obtenir des mesures 3D et des localisations 3D des objets mobiles permettant l'utilisation de modèles non directement liés à l'application échappant ainsi au problème de paramétrisation. La connaissance des mesures 3D peut être utilisée pour les calculs de distances réelles dans la scène, pour les objets mobiles ou statiques. Des informations d'entrée sortie de personne dans la scène peuvent être définies de façon robuste et utilisées de façon stable pour les étapes de reconnaissance et de suivi. L'évaluation de contexte locaux [Intille and Bobick, 1995] rendue possible par l'existence du contexte statique facilite le traitement des occultations partielles dynamiques et statiques.

Cette structuration de l'information contextuelle des scènes permet d'améliorer les possibilités du système au niveau de la robustesse des traitements, comme nous venons de le voir, mais aussi au niveau, de l'adaptabilité des traitements. Le regroupement de l'ensemble des informations de même nature utiles à plusieurs niveaux des traitements diminue la sensibilité du système au changement de cadre. Si l'on souhaite changer la caméra de place pour une même scène seul le plan image va changer. Si l'on souhaite rajouter un caméra dans une scène existante, il suffit de rajouter un référentiel image à l'ensemble de ceux déjà existants. Si un objet considéré fixe dans la scène est modifié, il suffit alors de changer ses propriétés et ses correspondances dans l'ensemble du contexte. L'unification des informations de même nature, concernant le contexte, permet d'augmenter la robustesse et l'adaptabilité d'un système de traitement de séquences d'images.

L'instanciation du contexte statique dans le système réalisé durant le stage est constituée de trois référentiels: un 3D de la scène, un 2D de l'image, un 2D du sol. Le premier est utilisé pour les mesures 3D de l'étape de reconnaissance. Le second est la base de l'étape de détection et le troisième durant l'étape d'appariement temporel. L'utilisation d'un modèle de caméra sténopée [Robert, 1993] permet d'obtenir les matrices de passage en coordonnées linéaires homogènes du référentiel 3D à un référentiel 2D. Une hypothèse supplémentaire, dite du point bas, permet d'évaluer les mesures 3D d'un objet 2D de l'image. Le principe est de prendre le point bas d'une région mobile comme appartenant au plan du sol de la scène.

Les contraintes fortes de l'emploi de contexte statique sont les problèmes de calibration de la caméra ainsi que les problèmes d'acquisition de l'information contextuelle, lourde à mettre en oeuvre. L'hypothèse du point bas n'est pas forcément valide si les problèmes d'ombrage ne sont pas résolus.

La solution proposée, comprenant un module gérant les informations contextuelles, va être détaillé dans la suite de ce document en reprenant l'architecture détaillé dans la seconde section (cf. figure 17).

4.2.2 La détection

Un module spécialisé a été réalisé, intégrant la plupart des méthodes de la famille de combinaison d'images (voir sous-section 3.1.2).

- (1) $I_{resultat} = Abs(Diff(I_0, I_t))$
- (2) $I_{resultat} = Max(Abs(Diff(I_t, I_{t-1})), Abs(Diff(I_{t+1}, I_t)))$
- (3) $I_{resultat} = Max(Abs(Diff(I_t, I_0)), Abs(Diff(I_{t+1}, I_t)))$
- (4) $I_{resultat} = Abs(Diff(I_{t+1}, I_t)))$

Ceci permet de changer d'instance de cette classe par la modification d'un simple paramètre. En pratique, les expériences ont été réalisées par combinaison d'image avec une image de référence car la possibilité d'avoir une image de référence nous était donnée sur toutes les séquences d'images. Les expériences réalisées avec les autres méthodes donnent un résultat moins intéressant pour les algorithmes de reconnaissance utilisés dans la suite. Les méthodes (2) et

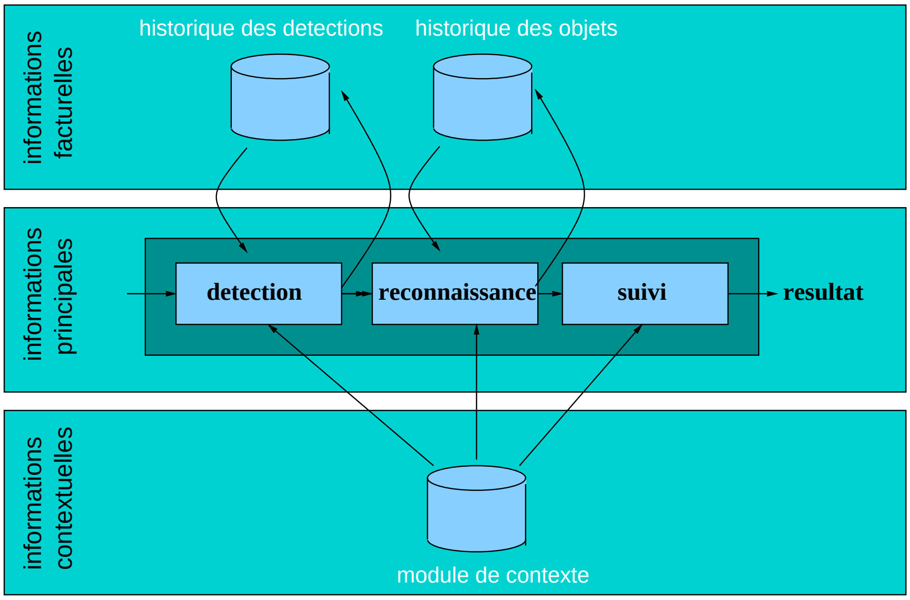

Fig. 17 – *Une solution à base d'informations contextuelles*

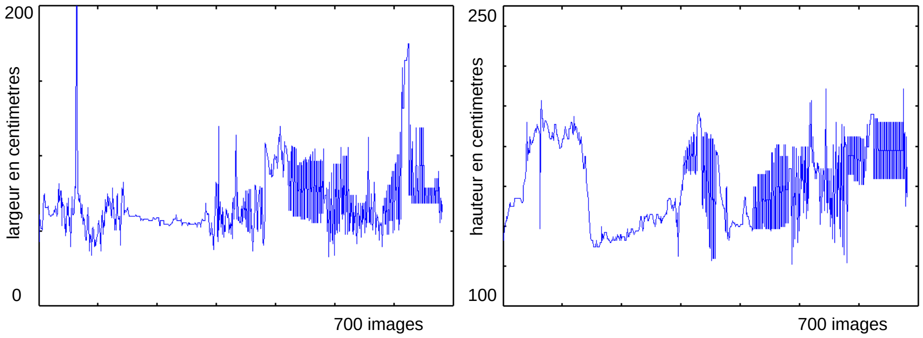

Fig. 18 – *La distribution des valeurs prises par les paramètres du modèle humain*

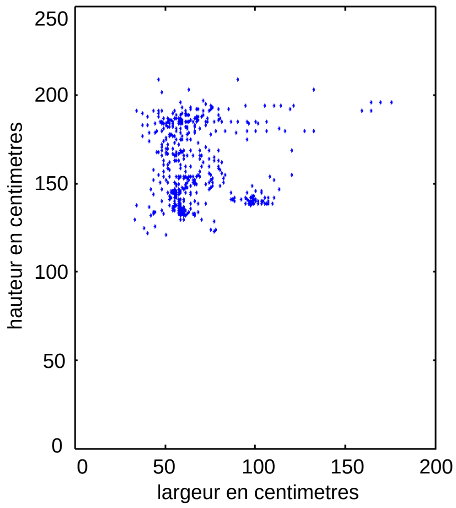

FIG. 19 – *La distribution des valeurs prises par les paramètres combinés hauteur et largeur*

(4) ont tendance à accroître le nombre des détections en segmentant une personne en trois ou quatre régions. On peut expliquer cela par le fait que les objets qui nous intéressent (les personnes) sont souvent vêtus d'habits unis non-texturés difficiles à gérer par ces méthodes. De plus ces méthodes ne sont pas sensibles au personnes immobiles dans la scène, ce qui est parfois le cas dans nos séquences. Quant à la troisième méthode, elle réunit les propriétés des méthodes (1) et (4), surévaluant la aussi le nombre de détection souhaité. Cette méthode peut être intéressante pour définir de vastes régions d'intérêt pourvu qu'elle puisse être couplée avec une méthode de reconnaissance capable de re-sélectionner parmi ses résultats.

4.2.3 La reconnaissance d'humain

L'accent a été mis sur la reconnaissance de personne. Pour cela le système utilise un modèle d'objet à trois paramètres. Le modèle d'objet est défini par sa vitesse, sa hauteur et sa largeur. (cf. figures 18, 19 et 22). Ces paramètres sont considérés dans le référentiel 3D-scène via le module de contexte statique. En ceci, ce modèle ne peut pas être considéré comme un modèle d'apparence mais comme un modèle d'objets réels bien qu'il n'y ait pas de considération volumique. L'instrument de la reconnaissance est un traitement par contraintes de ces mesures. On définit une borne supérieure et une borne inférieure pour chacun des degrés de liberté. Ainsi durant la phase de reconnaissance, on cherche à construire de façon itérative une combinaison de régions dont les mesures 3D ne dépassent pas les bornes supérieures du modèle. Les combinaisons, dont les mesures sont inférieures à la borne inférieures sont ensuite filtrées. Quant au troisième degrés de liberté, nous reviendrons sur son utilisation dans la section suivante concernant l'appariement temporel.
L'algorithme de reconnaissance (cf. figure 20) est donc divisé en deux étapes : une phase de construction et une phase de filtrage.

Phase de construction :
(C1) Choix d'une région R_1
(C2) Choix d'une seconde région R_2
(C3) Test de fusion
Si la fusion R_f de R_1 et R_2 respecte les contraintes Sup du modèle
Alors R_1 et R_2 sont remplacées par R_f
Sinon R_1 et R_2 restent dans la liste des détections
Phase de filtrage :
(F1) Choix d'une région R_3
(F2) Test de validité
Si la région R_3 respecte les contraintes Inf du modèle
Alors R_3 est validée
Sinon R_3 est éliminée de la liste des détections

FIG. 20 – *L'algorithme de reconnaissance*

Quatre instances de cet algorithmes, différant par le critère de choix des étapes (C1), (C2) et (F1) ont été utilisées.

- (1) instance basique : (C1), (C2) et (F1) ne sont pas contraints, les régions sont prises dans un ordre arbitraire fourni par le module de détection.
- (2) «focus of attention» : (C1), (C2) et (F1) ne sont pas contraints, mais l'ensemble des détections est divisé en sous-ensembles de détections définis dans certaines régions d'intérêt dans l'image obtenues grâce au suivi.
- (3) ordonnancement global : (C1) est fait en fonction de la position de la région dans l'image. La région située le plus bas dans l'image est choisi d'abord.
- (4) ordonnancement relatif : (C2) est fait en fonction du résultat de (C1). On choisit en (C2) la région la plus proche de celle choisie par (C1).

Les méthodes basées sur un ordonnancement de l'espace de recherche sont relativement similaires dans leur approche et dans leurs résultats. Ces deux méthodes ont l'avantage de fusionner prioritairement des régions proches les unes des autres. Ceci correspond à l'intuition que deux régions proches appartiennent plus facilement au même objet que deux régions éloignées. Ceci optimise donc la construction des objets mais est assez sensible à la proximité de deux personnes dans l'image. La méthode basée sur les «focus of attention» a comme avantage de réduire l'espace de recherche de façon drastique au départ de l'application de l'algorithme à certaines régions de l'image définies par les résultats du suivi, mais cette méthode est très sensible aux pertes de suivi. En effet, si une piste est inopportunément perdue, alors il n'y aura pas de région d'intérêt associée et donc pas de construction d'objet à l'étape suivante. L'utilisation d'un tel algorithme est problématique en soi. Bien que celui-ci limite l'explosion combinatoire lors de la recherche d'une solution en acceptant la première solution acceptable par les contraintes du modèle, il semble que

resultats obtenus sur 700 images	valeurs des moments en centimetres		pourcentage de valeurs dans [moy - n*ec.type, moy + n*ec.type]		
	moyenne	ecart type	n=1	n=2	n=3
largeur	68	15	65	87	96
hauteur	160	18	42	97	100
vitesse	13	14	80	90	94

FIG. 21 – *Les statistiques du modèle d'humain*

de nombreuses solutions plus intéressantes échappent à l'algorithme, ce qui est préjudiciable aux étapes suivantes.

L'avantage de ce modèle est l'utilisation de paramètres n'étant pas liés à l'application (i.e. les bornes Inf et Sup sont définies de façon externe.) Ceci augmente donc la robustesse et l'adaptabilité du système. Les inconvénients sont multiples. Tout d'abord l'utilisation de mesures générées par le module de contexte statique rend la décision prise sur la qualification d'un objet sensible aux erreurs faites par ce module. En pratique, on s'aperçoit que la calibration joue un rôle prépondérant. Compte tenu du matériel expérimental, notre modèle est trop imprécis (cf.figures 18, 19, 22 et 21). En effet, le seul paramètre convenable est la hauteur pour laquelle 3 % d'erreur nous autorise un intervalle de valeurs égal à $[144, 196]$ (en centimètres), ce qui ne parait pas inexploitable. Le paramètres moyen est la vitesse (en centimètres par images à 5 images seconde). Avec trois écarts type 94 % des valeurs sont traitées. Ce qui autorise 6 % d'erreur de reconnaissance, soit un intervalle égal à $[0, 65]$. Ces valeurs ne paraissent ni inexploitables, ni aberrantes. Toute fois, il aurait été souhaitable d'obtenir un intervalle plus petit (au moins pour la borne Sup) pour les besoin du suivi (cf.sous-section 4.2.4). En revanche, le paramètres de largeur n'est pas satisfaisant. 4 % d' erreurs nécessite un intervalle égal à $[23, 113]$ ce qui est trop. cette étalements des valeurs s'explique par un fort écart type (15 centimètres, soit plus d'un quatre de la moyenne) causé par la non-rigidité des objets á reconnaître (l'oscillation des bras et des jambes).

Comme le montrent les différents graphiques de données recueillis à partir de sept cents images, les trois degrés de liberté associés à notre modèle ont une trop forte variance pour servir de critère de décision fiable. Nous verrons dans les perspéctives (sous-section 5.2) des pistes pour palier à ce problème.

4.2.4 L'appariement temporel

Comme nous l'avons vu, les problèmes posés par l'étape de suivi du logiciel de départ étaient importants. Nous avons réalisé dans ce logiciel un module de suivi indépendant qui réalise l'appariement temporel selon différentes politiques.

- (1) méthode basée sur la proportion de recouvrement
- (2) méthode basée sur la similarité des régions

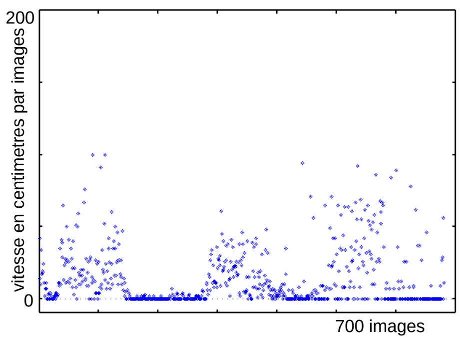

FIG. 22 – *La distribution des valeurs prises par le paramètre dynamiques du modèle humain*

- (3) méthode hybride basée sur (1) et (2)
- (4) méthode basée sur la localisation 3D

La méthode basée sur la proportion de recouvrement consiste à décider que deux régions proviennent du même objet si le pourcentage de recouvrement dépasse un certain seuil. La méthode basée sur la similarité des régions consiste à décider comme provenant du même objet deux régions dont les boites englobantes sont semblables. Cette similarité entre deux boites englobantes est calculée grâce à une fonction faisant intervenir la largeur et la hauteur des régions. La méthode hybride réunit les conditions des deux méthodes précédentes soit par un Et logique soit par un Ou logique. Le point commun de ces méthodes est l'utilisation des caractéristiques 2D du modèle, la décision de suivi est prise dans le plan 2D image. Ces méthodes fournissent un résultat intéressant dans les cas d'objets proches de la caméra et relativement bien détectés. Ces méthodes n'utilisant pas d'information 3D sont indépendantes de la qualité du module de contexte. Par contre, elles n'ont aucun recours pour traiter les occultations. De plus les objets à traiter n'étant pas rigides leur variation d'apparence rend les critères d'appariement 2D peu fiables dans le cas général, notamment dans les scènes encombrées.
La méthode (4) utilise des paramètres 3D du modèle pour apparier les objets. Cette méthode de suivi est basée sur l'utilisation d'une matrice d'ambiguité (cf. figures 23) que l'on definit de la façon suivante : Soit deux images consécutives I_{t-1} et I_t dans lesquelles on a reconnu respectivement les ensembles d'objets $E_{t-1} = (O_0, ..., O_o)$ et $E_t = (N_0, ..., N_n)$. On definit une matrice $(n \times o)$ telle que C_{ij} soit le critère d'appariement temporel entre O_i et N_j.

- $C_{ij} = 0$ si la distance entre O_i et N_j est supérieure à la borne Sup du paramètre vitesse du modèle d'humain.
- $C_{ij} = 1$ si la distance entre O_i et N_j est inférieure à la borne Sup du paramètre vitesse du modèle d'humain.

La matrice obtenue est une matrice $(n \times o)$ définie dans $(0, 1)$ où la valeur 1 signifie une filiation entre deux objets. Cette filiation est dite non-ambiguë si à

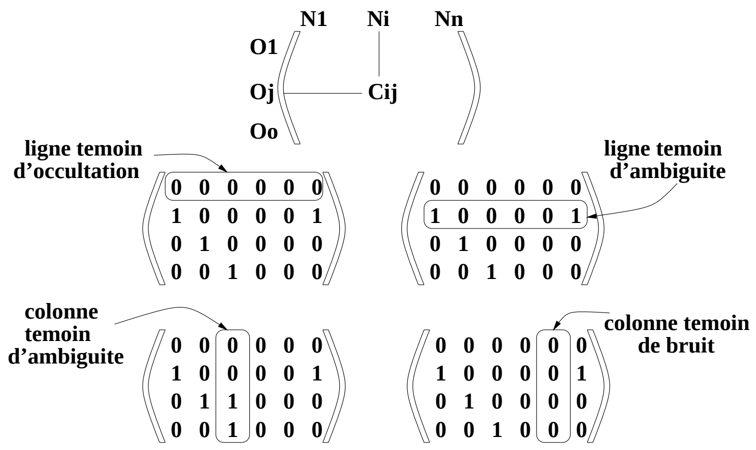

FIG. 23 – *Matrice d'ambiguité*

un objet de I_{t-1} correspond au plus un objet de I_t et réciproquement. Tous les autres cas sont dits ambigus. Cela signifie qu'une matrice idéale est une matrice telle qu'il existe au plus un 1 sur une ligne et sur une colonne. Les traitements associés à l'appariement temporel visent à transformer la matrice obtenue afin de la rendre idéale. Rendre idéale cette matrice, lever les ambiguïtés, revient alors à reconnaître des motifs qualifiant certains problèmes. Par exemple, une colonne N_i de valeurs nulles signifie que l'objet N_i n'a pas pu être apparié avec un objet de l'image précédente. Une ligne O_j de valeurs nulles signifie que cet objet n'a pas trouvé d'appariement dans l'image courante. De même une ligne ou une colonne contenant plus de une valeur égale à 1 signifie que l'objet en question peut s'apparier avec plusieurs objets distincts.

Le suivi n'est pas un problème totalement résolu. Bien que l'utilisation d'informations contextuelles rende intelligible les problèmes liés à l'appariement temporel, il est clair qu'un suivi robuste doit être réalisé à partir d'information 3D.

La figure 24 illustre ce propos en montrant les résultats comparés des logiciels avec ou sans module de contexte statique. A gauche, le logiciel de départ (sans information contextuelle), ne gérant pas les problèmes d'occultations dynamiques, produit un graphe temporel inexploitable. A droite, le logiciel utilisant un module d'information de contexte est capable de résoudre les problèmes d'occultations dynamiques.

4.3 Les conclusions

Nous pouvons maintenant admettre que l'utilisation d'informations contextuelles ne relève ni du gadget ni d'un raccourci de traitement. Son recours est essentiel à la qualification de la plupart des problèmes évoqués dans notre grille de lecture décrite dans la première et la seconde section. En effet la plupart des difficultés à résoudre proviennent du manque d'information 3D et d'information de la scène en général, que se soit sur la nature des objets qui nous intéressent ou sur le décor, l'environnement du traitement.

FIG. 24 – *Comparaison des résultats des logiciels avec ou sans contexte*

5 Conclusions et perspectives

5.1 Les conclusions

En conclusion, nous avons, dans ce mémoire, essayé de définir une grille de lecture du problème de traitement de séquence d'images pour le suivi de personnes. Cette grille se constitue de deux axes : un point de vue thématique des difficultés rencontrées et un point de vue séquentiel. Le premier point regroupe les problèmes liés à l'emploi d'information issue de caméra monoculaire, les problèmes liés au traitement dynamique et les problèmes liés aux différentes formes de bruit. Le second point de vue nous a amené à diviser le processus en sous taches que sont la détection, visant à extraire les indicateurs du mouvement, la reconnaissance des objets de la scène et enfin le suivi de ceux-ci. C'est à la lumière de cette division que la conclusion selon laquelle certains problèmes sont insolubles sans recourir à de l'information externe a été établie. Nous avons donc dans une dernière section analysé la nature de cette information qu'est l'information contextuelle et réalisé un système implémentant ces notions. Bien que tous les problèmes ne soient pas tous réglés, il semble que certains éléments de réponses ont été apportés, en ce qui concerne, par exemple, les caractéristiques minimales d'un modèle d'objets (le notre n'étant pas assez contraint ni prédictif), en ce qui concerne aussi les aspects dynamiques, qui paraissent devoir être traités dans un référentiel 3D de la scène ou 2D du sol, pour obtenir une robustesse suffisante du traitement des ambiguïtés. Enfin cette analyse du domaine nous a permis de définir certaines pistes non traitées, dont l'étude pourrait amener un gain en termes de robustesse et d'adaptabilité.

5.2 Les perspectives

- Agrégation de techniques

Comme il était suggéré dans le passage traitant de la détection, chacune des classes de méthodes était plus ou moins sensible à certains problèmes. L'idée, selon laquelle une agrégation contrôlée des différentes techniques fournirait un résultat plus robuste, parait être sensée.

- Seuillage contextuel

L'apport d'informations contextuelles dans les traitements, nous l'avons vu, augmente la qualification des problèmes intrinsèques au domaine, notamment dans les étapes de reconnaissance et de suivi. Il se pourrait alors que cette information puisse être utilisée dès la première étape notamment en fournissant de l'information de couleur ou de niveau de gris lors du seuillage de "l'image-combinaison", afin de prendre en compte les problèmes de contrastes.

- Modèle d'humain

Comme nous l'avons vu l'utilisation du modèle dont nous nous étions doté se révèle trop peu contraignant et trop peu prédictif pour le champ d'application qui nous intéresse. Une avancée significative sur ce point serait l'élaboration d'un modèle plus fin d'humain tel que le modèle de [Baumberg and Hogg, 1995].

On peut aussi imaginer un modèle dynamique capable d'ajuster ces paramètres au cours du temps.

- Évaluation automatique

L'évaluation automatique des étapes du processus est un concept emprunté au domaine du pilotage de programmes. Cette technique parait s'appliquer au traitement de séquence d'images, garantissant ainsi que le passage de données contrôlées d'une étape des traitements à l'autre. L'avantage pourrait être double. D'une part une évaluation permettrait de filtrer certaines erreurs à chaque niveau des traitements. D'autre part une évaluation permettrait d'automatiser le réglage des paramètres.

- Gestion du graphe temporel

Le graphe des détections, en ce qu'il constitue la sortie finale du système, semble pouvoir être l'objet de traitements particuliers. En effet, si l'on considère acquise la sémantique associée à ce graphe, des traitements de raffinement simples et efficaces peuvent s'opérer. A une échelle de temps plus grande que celle utilisée pour le processus de détection (de l'ordre de la seconde par exemple), le contrôle de ce graphe grâce à des indices tel que le nombre de personnes dans la scène, fiabiliserait certainement les résultats sans les retarder excessivement.

FIG. 25 – *Botticelli, Sandro. «Le retour de Judith, v. 1472»*

Références

[Atika, 1984] Atika, K. (1984). Image sequence analysis of real world human motion. *Pattern recognition*, 17(1):73 – 83.

[Baumberg and Hogg, 1995] Baumberg, A. and Hogg, D. (1995). An adaptative eigenshape model. In *British Machine Vision Conference BMVC*, Birmingham.

[Bouthemy, 1988] Bouthemy, P. (1988). Modèles et méthodes pour l'analyse du mouvement dans une séquence d'images. *Tecnique et Science Informatique*, 7(6):527–546.

[Bremond, 1997] Bremond, F. (1997). *Environnement de resolution de problemes pour l'interpretation de sequences d'images*. PhD thesis, INRIA - Universite de Nice Sophia-Antipolis.

[Chen and Lee, 1992] Chen, Z. and Lee, H. (1992). Knowledge-guided visual perception of 3d gait from a single image sequence. *IEEE Transactions on systems, man and cybernetic*, 22(2):336–342.

[Chleq and Thonnat, 1996] Chleq, N. and Thonnat, M. (1996). Realtime image sequence interprtation for videosurveillance. In IEEE, editor, *International Conference on Image Processing*, pages 801 – 804, Lausanne, Switzerland.

[Choi et al., 1997] Choi, S., Seo, Y., Kim, H., and Hong, K. (1997). Where are the ball and players? soccer game analysis with color-based tracking and image mosaik. In *ICIAP*.

[Du et al., 1993] Du, L., Sullivan, G., and Baker, K. (1993). Quantitative analysis of the viewpoint consistency constraint in mode-based vision. In *International Conference of Computer Vision*, pages 632–639, Berlin.

[Intille and Bobick, 1995] Intille, S. and Bobick, A. (1995). Closed world tracking. In *5th International Conference on Computer Vision ICCV*, Cambridge.

[Jain et al., 1979] Jain, R., Martin, W., and Aggarwal, J. (1979). Segmentation throught the detection of changes due to motion. *Computer Graphics and Image Processing*, 2:13–34.

[Karmann and Brandt, 1989] Karmann and Brandt (1989). Detection and tracking of moving objects by adaptative background extraction. In *Internation Conference on Image Processing (ICIP)*, pages 1051–1058.

[Koller et al., 1993] Koller, D., Daniilidis, K., and Nagel, H. (1993). Model-based object tracking in monocular image sequence of road traffic scenes. *International Journal of Computer Vision*, 3(10):257–281.

[Lenz and Gerhard, 1983] Lenz, R. and Gerhard, A. (1983). Image sequence encoding using scene analysis and spacio-temporal interpolation. In *Image* *sequence processing and dynamique scene analysis*, pages 264 – 274. T.S. Huang ed.

[Lowe, 1987] Lowe, D. (1987). Three-dimensional objet recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395.

[Meyer and Bouthemy, 1992] Meyer, F. and Bouthemy, P. (1992). Region-based tracking in an image sequence. In *European Conference on Computer Vision ECCV*, pages 476–484.

[Nagell, 1988] Nagell, H. (1988). The representation of situations and their recognition from image sequences. In *RFIA*, pages 1221–1229, Lyon-Villeurbanne.

[Pentland, 1995] Pentland, A. (1995). Machine understanding human action. In *7th International Forum on of Frontier of Telecommunication Technology*, Tokyo.

[Ricquebourg, 1993] Ricquebourg, Y. (1993). Segmentation et suivi d'objets mobiles par modèles structurels adaptatifs. Master's thesis, Institut national des sciences appliquées de Rennes.

[Ricquebourg, 1997] Ricquebourg, Y. (1997). *Analyse de mouvements articulés: mesure et suivi 2D; application à la télésuveillance.* PhD thesis, Universite de Rennes I.

[Robert, 1993] Robert, L. (1993). *Perseption stereoscopique de courbes et de surface tridimensionnelles. Application a la robotique mobile.* PhD thesis, Ecole Polytechnique.

[Rohr, 1994] Rohr, K. (1994). Toward model-based recognition of movement in image sequences. *Computer Vision, Graphique and image processing: image understanding*, 59(1):94–115.

[Strat, 1993] Strat, T. (1993). Employing contextual information in computer vision. In *DARPA93*, pages 217–229.

[Wenstop, 1983] Wenstop, O. (1983). Motion detection for image information. In *3th Scandinavian conference on image analysis*, pages 381 – 386, Tromso, Norway.

[Woodfill and Zabith, 1991] Woodfill, J. and Zabith, R. (1991). An algorithm for real-time tracking of non-rigid objects. In *AAAI*, Stanford.