Intégration temporelle pour une détection robuste d'objets mobiles éloignés à partir d'une caméra en mouvement

Temporal Integration for Robust Detection of Distant Moving Objects Observed by a Mobile Camera

par V. REBUFFEL, C. HENNEBERT, J.M. LÉTANG

LETI (CEA - Technologies Avancées)
DSYS - CEN/G - 17, Avenue des Martyrs, F-38054 Grenoble Cedex 9, France

résumé et mots clés

La détection d'objets mobiles dans une scène extérieure perturbée à partir d'une caméra embarquée est un problème difficile, dont la solution ne peut être robuste qu'avec un intervalle temporel d'analyse supérieur à deux images consécutives. On envisage ici deux façons d'intégrer le temps dans le processus de détection : d'une part l'utilisation d'observations temporelles longues pour la décision, d'autre part la mise en œuvre d'un module de prédiction. Ces observations sont typiquement construites à partir d'un filtrage temporel de la séquence d'images à différentes échelles de temps. La prédiction, qui concerne les objets détectés, est mise à jour incrémentalement à l'aide des résultats de la détection à l'instant courant, mais elle permet aussi de guider la détection lors de l'instant suivant. Quand la caméra est fixe, il suffit de considérer les images successives, mais dans le cas d'une caméra mobile, il est nécessaire de compenser préalablement le mouvement apparent de l'image. Les méthodes présentées dans cet article sont validées sur plusieurs séquences réelles et l'apport des différentes solutions est évalué.

Intégration temporelle, Mouvement, Détection d'objets mobiles, Analyse de scène, Capteur fixe ou mobile.

abstract and key words

The detection of moving objects in an outdoor and disturbed scene observed by a mobile camera is a difficult problem, whose solution requires to take into account more than two successive images to achieve robustness. In this paper we present two ways to carry out this temporal integration in the motion detection process : first, the use of long temporal observations for the decision criteria, secondly the implementation of a prediction module. These observations are usually get from temporal filtering at different temporal scales. Prediction of mobile objects is iteratively updated using current detection results, but also controls the detection process of the following instant. When using a static sensor, we have only to consider successive images, but if the sensor itself is moving, we need previously to compensate the apparent motion of the image. The methods that we have developped are tested on various real images, and their interest is evaluated.

Temporal integration, Motion analysis, Detection of mobile objects, Scene analysis, Static or mobile sensor.

1. introduction

Le rôle d'un système de surveillance consiste à observer un secteur donné afin de détecter tout changement le plus tôt possible. A cette mission d'observation s'ajoute généralement une exigence de durée, ainsi qu'une capacité à statuer sur le risque de menace que l'alarme représente. Il s'agit par exemple d'observer une scène extérieure avec un secteur angulaire de 120°, à une distance de 3000m, et ce pendant quelques heures. Pour être opérationnel, un tel système doit donc être robuste, c'est-à-dire peu sensible aux différentes perturbations possibles. Les applications potentielles concernent les domaines civils (zone sensible, nucléaire, transport) ou militaires (champ de bataille). Il s'agit, dans les séquences

d'images correspondantes, de détecter tout objet de quelques pixels ayant un mouvement apparent propre, éventuellement faible. Le facteur dynamique est particulièrement important, la forme de l'objet n'étant guère fiable ni significative vu les distances considérées et les perturbations possibles de l'environnement. On n'est plus ici dans le contexte d'application du schéma classique : détection basée sur le contraste, puis suivi et reconnaissance de cible, mais plutôt dans une situation où l'information de mouvement est nécessaire pour détecter correctement l'objet.

Dans la plupart des systèmes de détection de cibles mobiles par analyse d'une séquence d'images, les cas de non détection sont dus soit à la cible elle-même : contraste trop faible, vitesse trop lente, mouvement particulier, soit à des occultations par des écrans. Ces derniers sont à l'origine de masquages plus ou moins partiels, et résultent d'éléments de la scène comme des rideaux d'arbres ou des fumigènes. Les caractéristiques dynamiques de la cible les plus difficiles a priori pour la détection sont celles d'un mouvement approchant, car il se projette en un mouvement $2D$ très lent dans le plan image, ou d'une trajectoire présentant des manœuvres brusques. Comme il est très important de ne pas perdre la cible, on préfère généralement abaisser les seuils de détection et trier les fausses alarmes obtenues. Ces fausses alarmes résultent de perturbations, souvent spécifiques à l'application considérée. On peut citer les bruits du capteur, les variations d'illumination en longueur d'onde visible, les mauvaises conditions de propagation en infrarouge, mais aussi la présence de variations temporelles qui ne sont pas nécessairement des menaces potentielles, par exemple des mouvements de feuillage ou des reflets. Une mauvaise stabilisation de la caméra dans le plan image (vibrations dans le cas d'une caméra fixe, mouvement mal compensé pour une caméra mobile) est également une cause de fausses alarmes.

Il est évident que l'utilisation d'une intégration temporelle importante dans le processus de détection permettra de mieux détecter les cibles mobiles, notamment celles animées d'un mouvement apparent très lent. Mais elle diminuera aussi le taux de fausses alarmes, soit parce que celles-ci ne seront plus détectées, soit que leurs caractéristiques dynamiques permettront de les trier.
Les images de la figure [3] donnent une idée des scènes traitées. La première séquence représente une scène extérieure comportant un convoi de véhicules se déplaçant lentement à la lisière, elle a été acquise par un capteur visible fixe. La seconde séquence a été acquise avec une caméra infrarouge (8-12μ) mobile. La scène se compose d'un terrain vallonné, un véhicule se déplaçant à une vitesse de 30 km/h à une distance de 2500m environ.

2. intégration temporelle

Réaliser une intégration temporelle, c'est-à-dire construire un processus de détection d'objets mobiles à l'instant t prenant en compte les $n-1$ images précédentes (avec $n > 2$) peut se réaliser de différentes façons.

On peut utiliser ces n images au niveau du pixel, à condition qu'elles soient comparables spatialement, ce qui est vrai en caméra fixe, mais suppose un recalage préalable en caméra mobile. On raisonne alors dans un espace fréquentiel spatio-temporel, les objets en translation dans une image occupant un plan passant par l'origine dans cet espace transformé. Les méthodes relevant de cette approche utilisent des filtres dont le nombre devient vite prohibitif, ou qui sont trop sélectifs en vitesse pour une détection de mouvement [1], [2]. Une solution consiste à ne plus considérer l'aspect spatial, au moins dans un premier temps, et de privilégier l'axe temporel. C'est ce que nous avons retenu [3]. Des points importants doivent être spécifiés : la longueur temporelle n à prendre en compte, l'utilisation des informations temporelles obtenues, l'introduction du facteur spatial dans la détection. Remarquons que plus n est grand, plus l'information aux différentes échelles temporelles est riche mais devra être structurée pour son utilisation dans les critères de détection ultérieurs.

Dans ce qui précède, nous avons présenté l'utilisation des $n-1$ images précédentes de la séquence initiale comme observations pour la décision à l'instant t, mais il est également possible d'exploiter les résultats de détection des m instants précédents (avec éventuellement $m \neq n$). Relève de cette catégorie l'utilisation d'une image de référence. Les résultats antérieurs de détection permettent de prédire la localisation des objets mobiles à l'instant courant. Par l'utilisation d'une formulation statistique du problème, cette prédiction peut alors être combinée avec les observations temporelles de l'instant courant, que ce soit comme un état initial qui est alors affiné, ou comme une observation supplémentaire, ce que nous expliciterons au §5. Une telle approche suppose que la prédiction soit superposable à l'instant courant. Il ne s'agit plus ici de recalage d'images puisque c'est la vitesse de chaque objet mobile relativement à son environnement qui intervient. L'étape de prédiction doit donc intégrer le déplacement des objets. Un filtrage optimal de Kalman est parfois utilisé pour une telle prédiction en estimation de mouvement [4], [5]. Notre contexte est celui de la détection de mouvement, et donc le filtrage prédictif s'appliquera ici plutôt au niveau des objets détectés, donc de régions de l'image, qu'au niveau des pixels.
Cette façon de prendre en compte le passé met en œuvre des calculs plus simples et manipule moins d'information que celle au niveau des pixels. Elle est par contre plus éloignée du capteur, et donc moins capable d'en caractériser les bruits. D'autre part, il faut veiller à ce qu'elle ne soit pas trop conservatrice au niveau des objets détectés et prenne correctement en compte les innovations. Finalement, ces deux approches, résumées dans la figure [1], sont complémentaires, et toutes deux nécessaires pour atteindre une certaine robustesse.

Dans la suite de cet article, nous présentons d'abord une méthode de compensation du mouvement $2D$ permettant, dans le cas d'une caméra mobile, de se ramener à un contexte de caméra fixe. Nous introduisons ainsi la notion de sous-séquences glissantes. Nous analysons alors les filtrages temporels caractérisant des observations temporelles longues, en caméra fixe comme mobile.

L'exploitation de ces observations sera présentée dans le cadre d'une détection bayésienne du mouvement. La mise en œuvre d'un module de prédiction sera enfin étudiée.

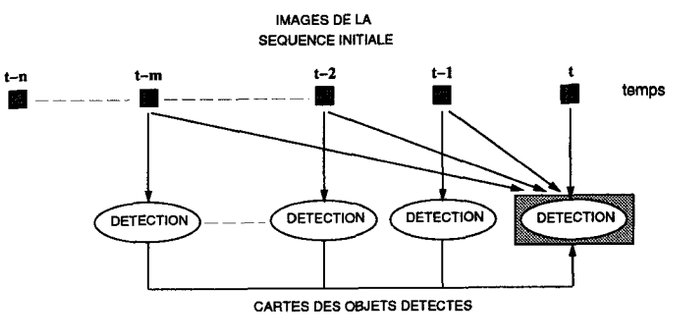


Figure 1. – Intégration temporelle en détection d'objets mobiles.

3. compensation du mouvement dominant apparent

Prendre en compte au niveau de traitements locaux plusieurs images successives suppose évidemment qu'un pixel de l'image corresponde au même point de la scène durant les instants considérés. Ceci n'est pas vérifié pour un capteur possédant un mouvement propre. Or dans notre contexte, la caméra peut être statique (trépied d'observation) ou mobile, cas où il est nécessaire de recaler les images successives utilisées. On peut distinguer les mouvements de veille panoramique ou sectorielle, qui correspondent à des mouvements de rotation pure (tourelle), les mouvements de vibrations d'amplitude faible (dus à un moteur) et les mouvements de déplacement effectif à composante de translation importante (capteur embarqué). La performance des algorithmes de compensation est conditionnée par le type de mouvement de la caméra et le type de scène.

Pour estimer le mouvement dominant entre deux images consécutives, nous avons retenu une méthode différentielle utilisant un modèle paramétrique global du champ des déplacements, inspirée de [6]. Le modèle polynomial du champ des vitesses utilisé est issu des équations de Longuet-Higgins [7] modélisant le mouvement de la caméra par rapport à une scène quelconque, en considérant un modèle au premier ordre de la fonction profondeur de la scène. Il peut s'écrire :

$$\begin{cases} u_\theta = a_0 + a_1 x + a_2 y + a_6 x^2 + a_7 xy \\ v_\theta = a_3 + a_4 x + a_5 y + a_6 xy + a_7 y^2 \end{cases} \quad (1)$$

Ce modèle à huit paramètres est combiné avec l'équation de contrainte du mouvement apparent exprimée localement sous une forme différentielle, et comportant un paramètre ε supplémentaire pour tenir compte d'un changement global de luminance :

$$I_x(x,y).u_\theta + I_y(x,y).v_\theta + I_t(x,y) = -\varepsilon \quad (2)$$

$$\text{avec :} \quad I_x = \frac{\partial I}{\partial x} \quad I_y = \frac{\partial I}{\partial y} \quad I_t = \frac{\partial I}{\partial t}$$

I : Valeur de la luminance au point (x, y),
(u_θ, v_θ) : Modèle du champ des vitesses.

Après introduction de l'expression du champ des vitesses (u_θ, v_θ) (éq.1), dans l'équation de contrainte du mouvement (éq.2), les paramètres a_i cherchés sont obtenus par optimisation aux moindres carrés pondérés. Le calcul est conduit en multirésolution spatiale, afin d'avoir accès à des mouvements apparents d'amplitude variée. Pour cela, on construit une pyramide gaussienne à trois niveaux, les paramètres du mouvement sont d'abord estimés au niveau grossier, puis transmis au niveau inférieur où ils servent à initialiser l'estimation, afin d'obtenir une estimation de plus en plus précise.
Les objets mobiles recherchés étant supposés petits, ils perturbent peu ce calcul global - d'autant moins que la mise en œuvre d'un estimateur robuste permet de diminuer leur influence en les excluant du processus d'estimation. Notons que le modèle retenu est quadratique, les résultats expérimentaux obtenus avec un modèle affine n'étant pas assez précis, sauf dans le cas d'une translation dominante et d'une scène à profondeur affine (de type vue aérienne). Une fois le champ des déplacements estimé entre deux images, l'une des images peut alors être recalée par rapport à l'autre par application de ce champ et interpolation bilinéaire.

De nombreux travaux voisins se trouvent dans la littérature pour la compensation du mouvement entre deux images successives [6], [8]. Dans notre cas, si l'on veut que la décision sur une image utilise les $n-1$ images précédentes, il faut considérer des sous-séquences, de longueur n, glissantes selon l'axe temporel. Les n images de la sous-séquence doivent être toutes recalées sur la même, la première, la dernière, ou la médiane de la sous-séquence, dite image de référence courante. Deux calculs sont possibles pour cela : soit l'estimation du mouvement est faite entre l'image de référence et chacune des images de la sous-séquence, soit l'estimation est faite uniquement entre deux images consécutives, et les déplacements sont cumulés à l'intérieur de la sous-séquence. Dans le cas de mouvements de petite amplitude à moyenne nulle, par exemple des oscillations autour de la position d'équilibre de la caméra, la première solution est meilleure. Par contre, dès que le capteur présente un déplacement effectif, que ce soit en rotation ou en translation, l'erreur obtenue par cumul des champs est inférieure à celle que l'on aurait en estimant les champs entre images non consécutives. En effet, l'amplitude du mouvement apparent peut devenir importante, parfois suffisante pour ne plus justifier des méthodes différentielles, même faisant appel à plusieurs niveaux de résolution spatiale. D'autre part, le mouvement du capteur étant relativement régulier, du moins à l'échelle temporelle considérée, les paramètres trouvés à l'instant

t peuvent servir d'initialisation au calcul de ceux de l'instant suivant, améliorant la vitesse et la précision du calcul de ceux-ci. La seconde solution sera donc retenue. De plus, elle présente l'avantage d'être moins sensible aux changements d'illumination, qui moins perceptibles entre deux images proches, et d'exiger moins de calcul.

A noter qu'en toute rigueur, le cumul des champs de déplacement doit être effectué par somme vectorielle et chaînage des vecteurs déplacement. Nous avons adopté une solution approchée, beaucoup plus rapide, consistant à additionner les paramètres globaux du mouvement, et ce sans observer de dégradation des résultats.

Il est intéressant d'analyser la variation au cours du temps de la qualité des images ainsi compensées. Cette qualité est estimée à l'intérieur de la sous-séquence analysée par l'Erreur Quadratique Moyenne entre deux images recalées (EQM), et ce à différentes échelles temporelles. A chaque échelle, on considère en fait le maximum des EQM des couples d'images possibles. Dans la suite, K est le nombre de niveaux temporels (le niveau k varie de 1 à K) et n le nombre d'images de la sous-séquence ($n = 2^K$). Toutes les images sont recalées dans la sous-séquence relativement à la dernière $I(t)$, et notées $I^r(t-i)$ (avec $i \in 0,..n-1$, et $I^r(t) = I(t)$). La qualité de la compensation est alors quantifiée au niveau k par :

$$Q_k = \max_{i=0,...2^{K-k}-1} (EQM(I^r(t-2^k i), I^r(t-2^k i - 2^k + 1))) \quad (3)$$

par exemple au niveau 2 :

$$Q_2 = max(EQM(I^r(t), I^r(t-3)), EQM(I^r(t-4), I^r(t-7)))$$

L'étude de la pyramide d'images recalées laisse apparaître deux types de comportement. Pour certaines des séquences traitées, l'EQM est quasiment stable à l'intérieur de la sous-séquence, en fait les erreurs d'estimation des champs s'annulent lors de l'addition des déplacements, et donc Q_k est sensiblement le même aux différents niveaux. Mais dans la majorité des cas, l'erreur se dégrade et Q_k augmente avec l'échelle temporelle. On constate expérimentalement que le premier cas correspond en général à un mouvement très lent du capteur, d'amplitude inférieure à 0.3 pixel/image, ou encore à certains types de défauts du capteur.

Cependant, dans tous les cas, l'erreur reste faible, et même très faible pour des mouvements de rotation du capteur. Pour les mouvements plus complexes, la qualité de la compensation dépend du type de scène. Finalement, l'algorithme de compensation développé donne de bons résultats si un mouvement apparent dominant existe, ce qui traduit soit une composante prédominante de rotation du capteur, soit une scène présentant peu de plans de différentes profondeurs. Pour donner un ordre de grandeur dans le cas d'une caméra effectuant une translation transverse de 30km/h, avec une focale usuelle, le déplacement apparent d'un point à 50m sera de 6.7 pixels, à 100m de 3.3 pixels, et à 1000m de 0.33 pixels. Or pour une telle translation le champ devrait être uniforme. Lorsque la scène comporte des régions significatives de profondeurs différentes, on ne peut donc plus appliquer un modèle global de compensation, du moins avec la précision des résultats requis dans notre contexte. Il faut alors réaliser une segmentation de la scène au sens de critères de profondeur et de mouvement apparent, et compenser l'image par régions obtenues. Les travaux dans ce domaine sont encore très prospectifs. On pourra consulter par exemple [9].

4. filtrage temporel

Sur la séquence initiale ou la séquence compensée par sous-séquences glissantes si le capteur est mobile, il est possible de baser la détection à l'instant courant sur n images du passé. C'est le point que nous développons dans cette partie. Les algorithmes de détection de mouvement utilisent les changements temporels. Notre idée est d'évaluer ces changements à plusieurs échelles temporelles. Ces calculs s'effectuent en chaque pixel de la séquence.

Quelle longueur temporelle n prendre en compte? Si n est grand, la méthode sera plus robuste mais plus complexe à mettre en œuvre car il faudra définir comment combiner des informations à de nombreux niveaux différents. D'autre part, en caméra mobile, l'erreur de compensation cumulée devient vite importante avec n, ne serait-ce que parce que la partie de la scène visible dans l'image varie trop. Enfin, des considérations de temps de calcul ainsi que de place mémoire conduisent à prendre n petit, en particulier en caméra mobile pour lequel le glissement des traitements des sous-séquences n'est pas effectué aussi strictement qu'en caméra fixe. Nos essais ont montré qu'une longueur temporelle de 32 images en caméra fixe est un bon compromis, alors qu'en caméra mobile, une longueur de 8 ou 16 images (suivant l'amplitude du mouvement) ne peut être dépassée sans dégradation. Si l'on veut accéder à des mouvements plus lents de l'objet, il est préférable d'augmenter l'échantillonnage temporel que le nombre de niveaux temporels d'analyse.

Afin d'accéder à l'information de changement aux différentes échelles temporelles, il est nécessaire de filtrer le signal. Rappelons que l'on veut caractériser le comportement temporel de chaque pixel de l'image, et non accéder à l'information de mouvement, et donc qu'il s'agit d'une transformation le long de l'axe temporel. Dans le cas d'une caméra fixe, afin de caractériser efficacement le comportement temporel d'un pixel situé sur la trajectoire d'un objet mobile, nous avons retenu une décomposition en ondelettes qui permet une bonne localisation en temps et en fréquence. Nous avons choisi la base de Haar d'ordre 2, qui correspond en fait à des différences de moyennes temporelles d'images. La formule de décomposition au niveau k (k varie de 1 à $K = 5$) donne :

$$D_k = \sum_{i=0}^{2^{k-1}-1} I(t-i) - \sum_{i=2^{k-1}}^{2^k-1} I(t-i) \quad (4)$$

L'idée sous-jacente est que les bruits divers (capteur, variations d'illumination, petits mouvements d'oscillation) correspondent à

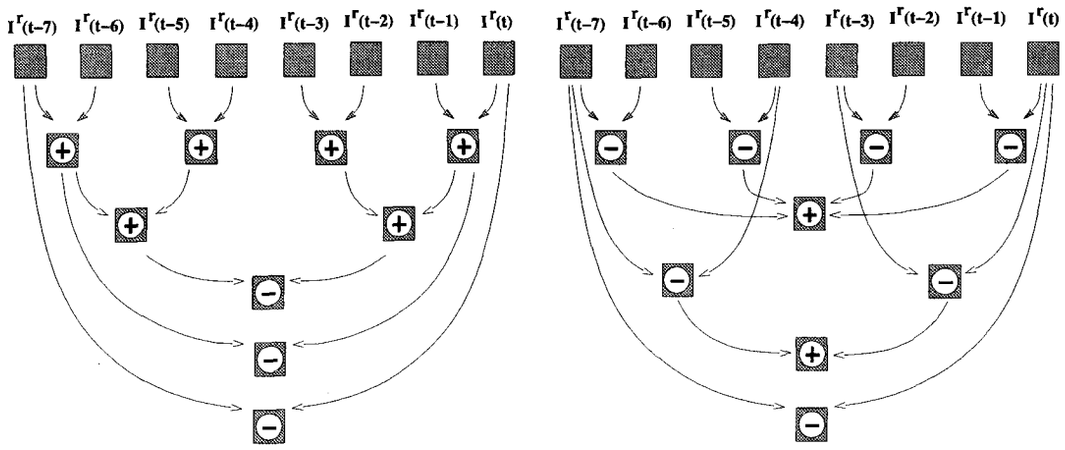


Figure 2. – Schéma des filtrages temporels, passe-bas (à gauche) et passe-haut (à droite).

des hautes fréquences, alors que les objets mobiles recherchés, qui sont lents et réguliers à l'échelle considérée, sont présents aux différentes fréquences : leur contour apparaît dès les hautes fréquences et leur masque complet progressivement aux basses fréquences.

Dans le cas d'une caméra mobile, les bruits divers sont faibles devant les erreurs résiduelles de compensation. Or comme nous l'avons vu au paragraphe précédent, celles-ci ont un comportement temporel variable, qu'il est cependant possible d'estimer. Plutôt que de les caractériser, ce qui est d'autant plus difficile qu'on dispose de moins de niveaux qu'en caméra fixe, on va choisir le filtrage temporel dans le but de les réduire, tout en mettant en évidence les pixels des objets mobiles [10]. Le filtrage est donc choisi en fonction du comportement temporel des erreurs de compensation. Dans le cas où l'erreur de compensation est stable au travers des échelles temporelles, les fausses alarmes qui en résultent apparaissent sous forme de hautes fréquences (plus hautes que les objets mobiles), et on retiendra un filtrage passe-bas pour les réduire :

$$D_k = \sum_{i=0}^{2^{K-k}-1} I^r(t-i) \; - \sum_{i=2^K-2^{K-k}}^{2^K-1} I^r(t-i) \qquad (5)$$

Lorsque l'erreur de compensation augmente avec les échelles, un filtrage passe-haut permettra d'en diminuer les conséquences :

$$D_k = \sum_{i=0}^{2^{K-k}-1} \left(I^r(t-2^k i) - I^r(t-2^k i - 2^k + 1)\right) \qquad (6)$$

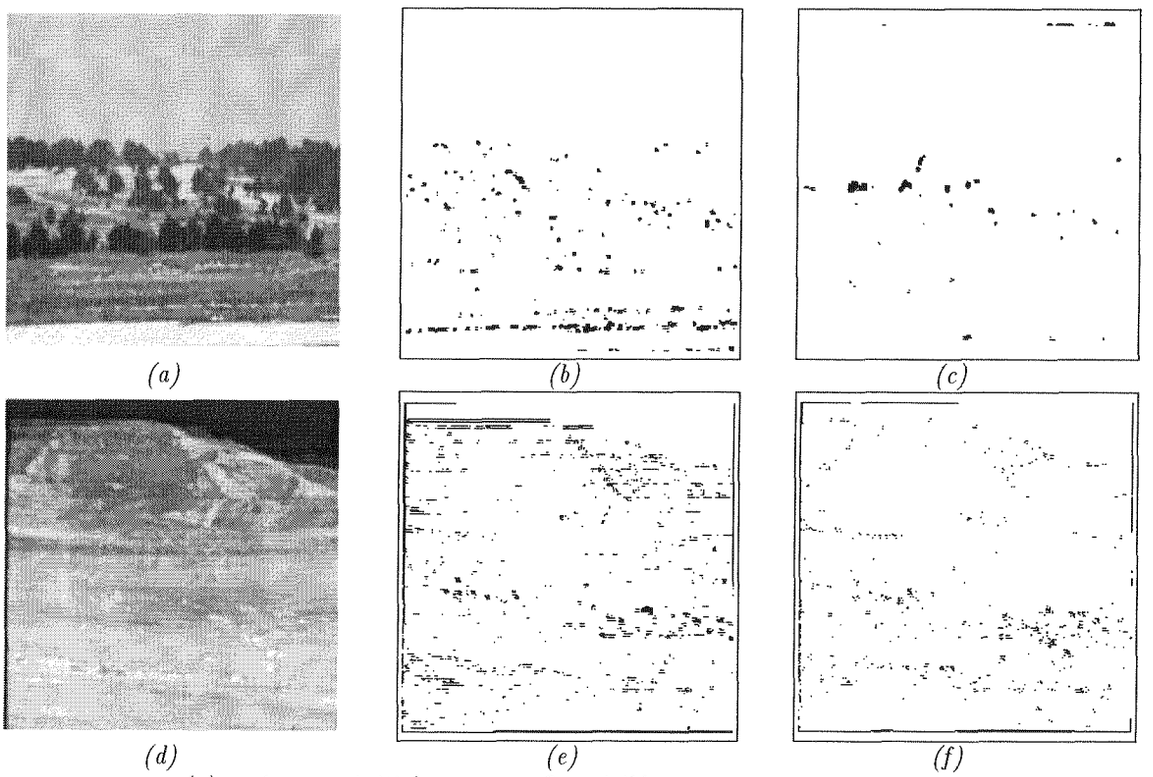
La figure [2] représente ces deux filtrages dans le cas de 3 niveaux. Entre les deux filtrages temporels, le choix est fait automatiquement en fonction d'une mesure basée sur la qualité de compensation multiéchelle présentée plus haut (éq.3) :

$$Q = \frac{Q_K}{Q_1} = \frac{Q_3}{Q_1} \text{ pour 3 niveaux.}$$

Si $Q \approx 1$, c'est-à-dire si les erreurs de compensation correspondent à des basses fréquences, on appliquera le filtrage passe-bas, et si $Q \gg 1$ le filtrage passe- haut. La robustesse du test provient de ce que si Q est à la limite du seuil, l'application d'un filtrage comme de l'autre donne une réponse à peu près identique, en terme de discernabilité des objets par rapport aux erreurs de compensation.

Pour augmenter la robustesse au sens d'une moins grande sensibilité à un bruit ponctuel, les différences d'images intervenant dans les filtres (équations 4, 5 et 6) ne sont en fait pas calculées en un pixel mais sur un voisinage 3×3, suivant un test de vraisemblance bilinéaire à deux hypothèses : présence ou absence de changement temporel [11]. Les réponses des filtres sont en fait les sorties des tests de vraisemblance, et ce à chaque niveau temporel. De plus, en cas de changement d'illumination (détecté par une mesure globale), nous modifions ce test afin de s'abstraire au moins partiellement du changement de luminance. Pour cela, le terme d'ordre 0 du modèle linéaire n'est plus pris en compte, ce qui revient à modéliser la perturbation par une constante additive sur la distribution d'intensité [3].

La figure [3] illustre le filtrage temporel sur deux séquences différentes. La première, acquise en caméra fixe, a donc été traitée par un filtre passe-bande. On a représenté les niveaux 1 et 5. La seconde, en caméra mobile, se comporte en compensation avec une mesure $Q = 1.9$, c'est donc un filtrage passe-haut qui lui a été appliquée, filtrage dont on a illustré les niveaux 1 et 3. (Les niveaux sont représentés sous forme binarisée en sortie du test de vraisemblance). On remarque que le niveau 1 contient essentiellement du bruit alors que les objets apparaissent plus nettement aux niveaux plus élevés.



(a) *(b)* *(c)*

(d) *(e)* *(f)*

(a) : séquence initiale, capteur fixe visible.
(b) et (c) : résultat du filtrage temporel, niveaux temporels 1 et 5.
(d) : séquence initiale, capteur mobile infrarouge.
(e) et (f) : résultat du filtrage temporel, niveaux temporels 1 et 3.

Figure 3. – Différents niveaux de filtrage temporel.

5. utilisation de la séquence issue du filtrage temporel

Les informations issues des filtrages aux différents niveaux temporels doivent être combinées afin de statuer sur la présence d'un objet mobile. Rappelons que l'on se place ici en détection de mouvement, non en estimation, et donc que l'on ne cherche pas à estimer le champ des déplacements apparents de l'image. Afin de pouvoir combiner les différentes informations issues du filtrage temporel tout en assurant des propriétés spécifiques de lissage de la solution, nous avons opté pour un schéma de régularisation bayésienne que nous ne détaillerons pas ici. Le problème de détection est formulé comme une segmentation en trois étiquettes : statique, mobile, et bruit de mouvement. Cette dernière, qui a été ajoutée pour plus de robustesse, correspond essentiellement aux erreurs de compensation dans le cas d'une caméra mobile, et à tous les bruits ou mouvements parasites dans le cas d'une caméra fixe. La solution est obtenue sous la forme de la minimisation d'une énergie qui est une fonction des étiquettes de mouvement. Introduite en détection de mouvement avec une observation basée sur deux images consécutives [12], cette approche a dû être adaptée pour une prise en compte du passé plus importante, c'est-à-dire qu'elle doit intégrer les observations issues du filtrage temporel aux différentes échelles.

Dans le cas d'une caméra fixe, nous avons retenu une énergie qui comporte trois termes, exprimant *(1)* la régularité spatiale locale, *(2)* la cohérence avec les observations temporelles aux différents niveaux, et *(3)* l'adéquation avec un module de prédiction que nous présentons dans la suite. Le premier terme contribue à agglomérer les pixels étiquetés mobiles. Le second terme traduit l'adéquation entre une étiquette donnée et la signature temporelle du pixel analysé, c'est-à-dire la réponse au filtre multiéchelle temporel, après application du test de vraisemblance. Le dernier terme est présenté au §6. Cette méthode, développée dans [13], est résumée dans la figure [4]. Un état initial est obtenu grâce à une heuristique basée sur une forme simplifiée de l'énergie : une règle simple

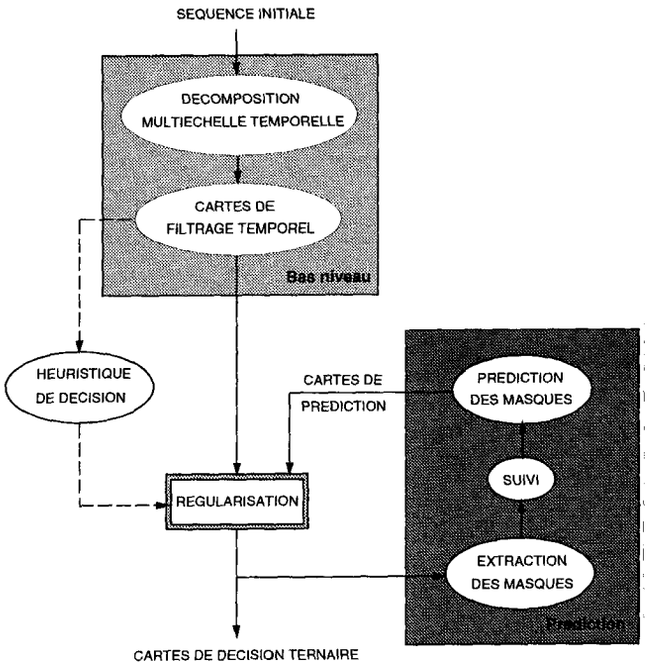


Figure 4. – Détection de mouvement en caméra fixe.

affecte à chaque pixel une première étiquette en fonction du vecteur issu des 5 niveaux du filtrage temporel.

Dans le cas d'une caméra mobile, le même schéma global est appliqué [10], mais le terme d'énergie lié au filtrage temporel est légèrement différent, ce qui provient entre autre du fait que les erreurs ont été réduites par le filtrage plutôt que caractérisées selon leur fréquence comme en caméra fixe. D'autre part une observation supplémentaire basée sur la vitesse apparente résiduelle dans l'image a dû être ajoutée (figure [5]).

Dans les deux cas le processus de relaxation est basé sur un algorithme de type ICM et converge rapidement, en particulier grâce à un bon état initial, et au faible nombre d'étiquettes possibles. La figure [6] présente les résultats de détection sur les deux séquences de la figure [3], sous la forme de superposition des masques d'objets détectés. Le mouvement des objets est trop lent (0.15 pixel/image environ) pour être détecté par des méthodes n'utilisant que deux images consécutives.

6. mise en œuvre d'une prédiction

Nous présentons maintenant un second moyen d'intégrer l'aspect temporel en détection d'objets mobiles, utilisant les résultats de détection des instants précédents. Ici aussi une longueur temporelle intervient. Le but est moins de séparer les objets mobiles des bruits de type capteur, que de mieux détecter les

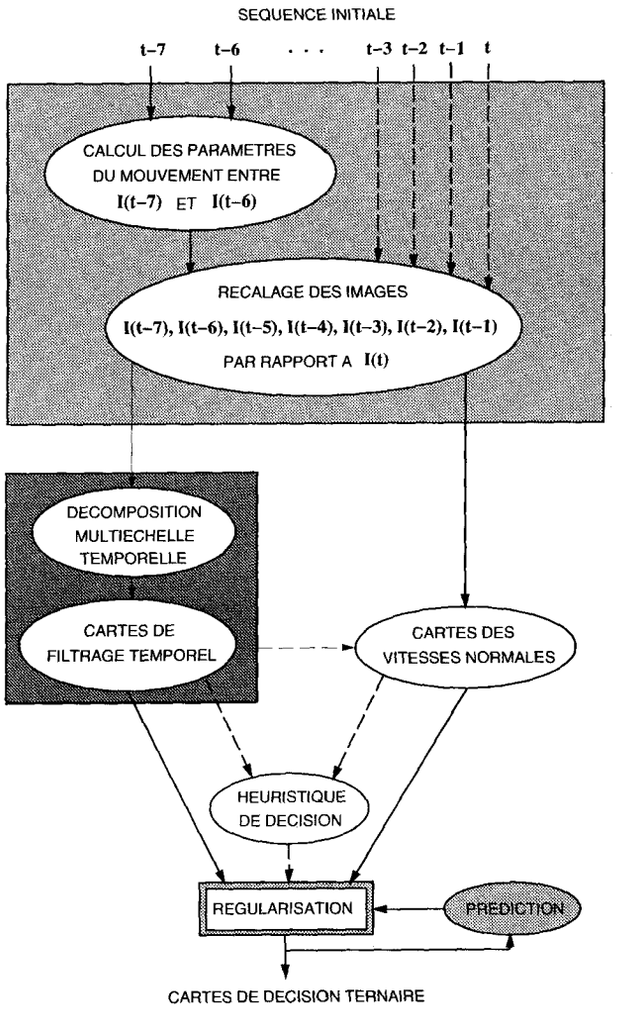


Figure 5. – Détection de mouvement en caméra mobile.

objets perturbés, partiellement masqués, et de les distinguer des fausses alarmes résiduelles. C'est d'ailleurs pourquoi le traitement à ce niveau est le même en caméra fixe ou mobile compensée.

L'idée est de propager dans le temps l'information de détection de mouvement de la carte de décision courante, en prédisant dans la carte suivante les masques des objets déclarés en mouvement à l'aide d'une information élémentaire sur leur mouvement. La position du centre de gravité de chaque masque des objets en mouvement suffit à réaliser cette tâche tant que les objets sont petits, et de mouvement plutôt régulier à l'échelle temporelle considérée, comme les cibles non résolues de notre contexte. L'espace d'état de la prédiction est alors relativement simple. Le formalisme utilisé pour cette procédure récursive temporelle est celui d'un filtre α-β modifié reposant sur les équations de Kalman. Les mesures, qui sont constituées des centres de gravité des masques et calculées par des algorithmes de traitement d'image, sont accompagnées de bruits de mesure évalués également au niveau de l'image.

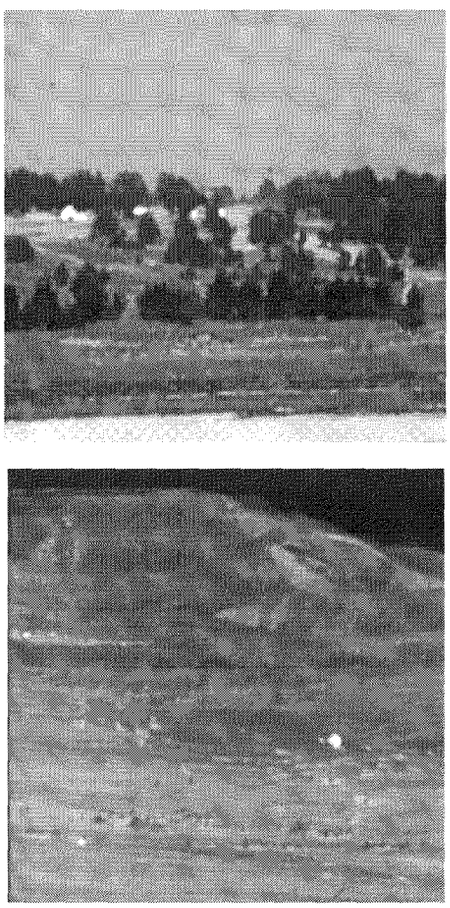


Figure 6. – Séquences de la figure [3], avec superposition des masques détectés.

Il est important de comprendre que les variables d'état ne sont pas tous les pixels de l'image, système nécessaire en estimation incrémentale du mouvement ([14], [5]) mais qui serait trop complexe et non adéquat au niveau dynamique pour la détection de mouvement.

La prédiction temporelle ne fournit que des informations ponctuelles : les centres de gravité des zones connexes en mouvement et leurs incertitudes de position associées. Pour les exploiter dans la phase de détection de mouvement, nous construisons une carte de prédiction temporelle, où l'intensité en chaque pixel est proportionnelle à la probabilité de présence d'un objet mobile. Cette carte est obtenue en déplaçant chaque masque d'objet de son déplacement estimé. Elle est mise à jour incrémentalement à chaque détection. Par ce processus, la longueur temporelle prise en compte est implicitement tout le passé.

Comment utiliser la carte de prédiction dans la détection de l'instant suivant? Rappelons que cette détection est réalisée dans un contexte statistique, par la minimisation d'une énergie fonction des étiquettes de mouvement. La carte de prédiction pourrait être utilisée comme état initial de l'optimisation, mais l'avantage en est faible puisqu'on dispose déjà d'un bon état initial, et de plus la stratégie qui en résulte est trop conservatrice. Il est bien plus intéressant de concevoir la prédiction comme une observation supplémentaire, et de définir un terme d'énergie approprié. On réalise ainsi une fusion entre la prédiction temporelle et les observations issues du filtrage multiéchelle temporel, tout en satisfaisant les propriétés a priori de lissage. Cette fusion intègre bien le fait que la prédiction n'est pas une donnée complètement fiable, et n'est importante que lorsque les observations provenant du filtrage temporel sont perturbées, éventuellement inexistantes. Une telle stratégie permet de ne pas maintenir artificiellement de fausses détections, par exemple dans le cas d'un objet manœuvrant. Il est important de noter que le module de prédiction proposé n'a pas pour objectif d'assurer un suivi complet des objets mobiles. Il réalise effectivement une trajectographie élémentaire des objets, mais dans le but d'aider leur détection aux instants ultérieurs. Ce bouclage de la prédiction vers la détection permet de garder un équilibre entre la prédiction, de nature conservative, et les observations, évaluées à chaque instant. Un module classique de pistage d'un système de surveillance utilise les informations de détection, mais sans un tel bouclage.

Expérimentalement, l'addition du module de prédiction permet de mieux détecter les objets perturbés ou masqués, cas conduisant à des déformations ou fragmentations de masque. La figure [7] illustre la détection d'un char masqué par des feuillages. On observe également une diminution des fausses alarmes. Celles qui subsistent, comme les objets mobiles d'ailleurs, sont associées à des caractéristiques diverses : mesures dynamiques, critères de forme ou au minimum de taille, ainsi qu'une évaluation de la confiance sur le résultat obtenu. Cette confiance est obtenue à partir du filtrage de Kalman et de l'erreur résiduelle de régularisation. Ces différentes mesures peuvent permettre de distinguer objets mobiles de variations temporelles comme fumée ou reflets, et de statuer sur le caractère de menace représenté par l'objet détecté.

7. performances et résultats expérimentaux

L'intégration temporelle apportée a permis de détecter des objets trop lents ou trop perturbés pour être détectés à partir de deux images seulement. Est-il possible d'évaluer les limites du système obtenu?

Rappelons que dans un contexte de surveillance, il est souhaitable de détecter la cible le plus tôt possible. Dans notre méthode les facteurs qui interviennent sont la vitesse, mais aussi le contraste et la dimension de la cible, ainsi que le niveau de bruit de l'image. Ces facteurs sont liés et il n'est pas possible de donner une borne inférieure de discernabilité pour l'un d'entre eux

(a) *(b)* *(c)*

(a) : séquence initiale, (b) et (c) : résultat de la détection, sans (b) et avec (c) module de prédiction

Figure 7. – Capteur infrarouge, deux chars derrière des feuillages.

(a) *(b)* *(c)*

(a) : Capteur infrarouge 8 – 12μ. Piéton se déplaçant de 0.1 pixel/image.
(b) : Capeur visible. Deux véhicules dont un approchant vers le capteur.
(c) : Capteur infrarouge 3 – 5μ. Convoi de véhicules.

Figure 8. – Exemples, avec superposition des masques des objets détectés.

indépendamment des autres. On peut théoriquement estimer le seuil de détectabilité d'un test de changement temporel, puis du filtrage temporel multiéchelle. Il est plus difficile d'évaluer l'apport de la régularisation. En fait le module de régularisation ne change pas les seuils de vitesse ou de contraste caractérisant la détectabilité d'un pixel, mais plutôt sa probabilité d'être détecté. D'autre part, à partir d'une certaine dimension et si l'objet ne présente pas de texture propre, les algorithmes de détection ne sont sensibles qu'aux bords de l'objet, et la détectabilité n'est plus véritablement fonction de la taille à partir d'une vingtaine de pixels. Si l'on veut donner une formule, on peut à l'aide de quelques constatations théoriques, donner la règle suivante valable pour une caméra fixe (D est la dimension de l'objet exprimée en pixels, V sa vitesse en pixel/image, ∇I son contraste en niveaux d'intensité et σ la variance du bruit sur l'image) [15] :

$$\begin{cases} \text{si} \quad D <= 3 & : \quad V \times \nabla I \times D > 3 \times \sigma \\ \text{si} \quad 3 < D <= 16 & : \quad V \times \nabla I \times \sqrt{D} > \sigma \\ \text{si} \quad D > 16 & : \quad V \times \nabla I \times 4 > \sigma \end{cases} \tag{7}$$

Il faut noter que V est la composante de la vitesse de l'objet dans le plan image, considérée comme constante pendant l'intervalle de temps pris en compte, et que l'objet est supposé non perturbé et non texturé. Typiquement, un objet non résolu de 2 pixels, avec un contraste par rapport au fond de 20 niveaux de gris et une vitesse de 0.5 pixel/image sera détecté si le bruit du fond ne dépasse pas 6.6 niveaux d'intensité, ce qui correspond déjà à un bon capteur (par exemple séquence de la figure [8a]). Une cible de 6 pixels, si le bruit de la scène est de 10 niveaux, sera détectée à partir d'une vitesse de 0.2 pixel/image. Des validations expérimentales ont confirmé ces ordres de grandeur par incrustation de cibles de synthèse sur séquences réelles. Les résultats sur scènes et cibles réelles sont très corrects, cohérents avec la formule (7). Il faut noter la difficulté de relier D, la dimension apparente en pixels de l'objet, avec sa dimension réelle, notamment en infrarouge. La figure [8] donne des exemples de détection de cibles très lentes (déplacement inférieur à 0.2 pixels/image) et petites (non résolues).

Dans le cas d'une caméra mobile, le résultat est dégradé environ d'un facteur 4, d'abord par suite d'une moins grande intégration temporelle, mais aussi à cause des erreurs de compensation. Notons cependant que la qualité de la compensation intervient plus dans le taux de fausses alarmes que dans la non détection des cibles.

Des facteurs autres que la dimension et la vitesse de la cible interviennent. Le cas de mouvement manœuvrant pose peu de problème à la détection, d'une part parce que le mouvement n'est généralement pas trop discontinu aux cadences usuelles d'acquisition, d'autre part parce que les innovations sont correctement prises en compte dans la fusion avec la carte de prédiction. Le masquage des objets dégrade évidemment les seuils de détectabilité, mais la détection reste correcte grâce au module de prédiction. Par contre le traitement de cas comme le croisement d'objets, le regroupement apparent à l'horizon, ou la dispersion, doit être résolu par une gestion plus élaborée de l'affectation des mesures (formes connexes détectées) aux filtres (objets précédemment détectés) [16]. La stratégie optimale dépendant de la configuration à traiter, un tel système doit être capable de gérer plusieurs hypothèses.

Le second aspect à quantifier pour évaluer les performances de la méthode est la présence de fausses alarmes. La définition d'une fausse alarme est d'ailleurs difficile, et dépend du contexte. Les bruits les plus usuels (de type capteur) sont éliminés dès le filtrage temporel. Les variations d'illumination provoquent des fausses alarmes au niveau des contours de luminance de l'image, qui sont correctement éliminées dans le cas d'une caméra fixe, mais peuvent subsister en caméra mobile d'autant plus que la compensation préalable est alors elle-même dégradée. D'autre part, il subsiste un certain nombre de fausses alarmes résultant de phénomènes temporels effectifs : fumée, feuillage, reflets.. mais s'agit-il vraiment de fausses alarmes? Un module de haut niveau, utilisant les caractéristiques obtenues (forme, dynamique, trajectoire, confiance), doit pouvoir statuer sur la menace que les alarmes détectées représentent. Spécifique de l'application considérée (environnement, cible, mission), il correspond à la fonction de tri d'un système de surveillance. Le taux de fausses alarmes ne peut en toute rigueur être déterminé qu'en sortie de ce module. Cependant, sur les séquences analysées, on constate que le nombre de fausses alarmes est faible et peut être supprimé par quelques critères simples, notamment de cohérence temporelle.

Nous avons testé notre algorithme sur des séquences visibles comme infrarouges sans le modifier. La méthode ne fait pas d'hypothèses particulières sur le capteur. Si en infrarouge les problèmes de variation d'illumination ne se posent pas, en revanche les images sont moins bien contrastées, surtout si les conditions de propagation sont mauvaises. Notons que la qualité de la compensation est dégradée si l'image est peu contrastée, plus d'ailleurs à cause du manque de points de contraste utilisés pour l'estimation que par le fait que les contrastes sont moins marqués. La compensation est même assez bonne en $8 - 12\mu$ et en proche infrarouge. Les résultats de détection de cibles mobiles sont néanmoins corrects en toute longueur d'onde. Notons que si l'on dispose des deux types de capteur, une fusion des deux au niveau de la détection est envisageable grâce au cadre de régularisation statistique retenu.

8. conclusion

Nous avons donc développé des méthodes de détection d'objets mobiles réalisant une intégration temporelle supérieure à deux images consécutives. Cette prise en compte du passé est faite d'une part au niveau des observations temporelles, par le biais d'un filtrage temporel multiéchelle, d'autre part par l'intégration d'une carte de prédiction dans la détection. Cette détection est réalisée selon un formalisme de régularisation bayésienne qui permet de fusionner information de prédiction et résultats du filtrage. L'intégration ainsi mise en œuvre a permis d'atteindre une réelle robustesse, en abaissant le seuil de détectabilité des cibles (en particulier pour les cibles très lentes, ou masquées), et en diminuant le taux de fausses alarmes. Les longueurs temporelles sous-jacentes, actuellement fixées entre 8 et 32 images, peuvent être déterminées automatiquement, en fonction du mouvement du capteur, ainsi que de la scène analysée.

Finalement, nous avons montré l'intérêt d'intégrer le mouvement, ou plus généralement l'information temporelle, dans un processus de détection des objets mobiles, sortant ainsi du schéma classique "détection par le contraste puis suivi". La mission de surveillance évolue alors du suivi d'une cible identifiée vers une analyse vigilante plus globale, qui peut prendre en compte plusieurs cibles éventuelles, s'adapter aux perturbations de type changement d'illumination, et donc interpréter dynamiquement une scène complexe pendant une longue durée.

Remerciement

Nous remercions P.Bouthemy (IRISA-Rennes) pour son aide dans la conception des méthodes, le CTP(DGA) et Intertechnique pour la fourniture des séquences de test, la DRET(DGA) pour le financement partiel des travaux, en particulier C.Rouchouze pour l'intérêt qu'il a montré à nos études.

BIBLIOGRAPHIE

[1] D.J. Fleet and A.D. Jepson, "Hierarchical construction of orientation and velocity selective filters", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989, Vol.11, n°3, p.315-325.

[2] E. P. Simoncelli, "Distributed representation and analysis of visual motion", *PhD Thesis, Harvard University, M.I.T.*, janvier 1993.

[3] J.M. Létang, V. Rebuffel et P. Bouthemy, "Motion detection based on a temporal multiscale approach", *IEEE International Conference on Pattern Recognition, The Hague*, septembre 1992, p.65-68.

[4] A. Singh, "Incremental estimation of image flow using a kalman filter", *Journal of Visual Communication and Image Representation*, 1992, Vol.3, n°1, p.39-57.

[5] F. G. et D. Dubois, "Approche adaptative de l'estimation du mouvement dans une séquence d'images par filtrage de Kalman temporel", *14ème Colloque GRETSI, Juan-les-Pins*, septembre 1993, p.875-878.

[6] J.M. Odobez et P. Bouthemy, "Robust multiresolution estimation of parametric motion models in complex image sequences", *7th European Conference on Signal Processing, EUSIPCO'94, Edinburgh*, septembre 1994.

[7] H.C. Longuet-Higgins, "The interpretation of a moving retinal image", *Processing Royal Society London, B-208*, 1980, p.385-397.

[8] M.Irani, B. Rousso et S. Peleg, "Detecting and Tracking Multiple Moving Objects Using Temporal Integration", *2nd European Conference on Computer Vision, Santa Margherita Ligure*, mai 1992, p.282-287.

[9] C. Hennebert, V. Rebuffel et P. Bouthemy, "Structuration Spatio-Temporelle d'une Scène Texturée : Segmentation au sens du mouvement et de la profondeur", *RFIA'96, Rennes*, janvier 1996, Vol.2, p.838-847.

[10] C. Hennebert, V. Rebuffel et P. Bouthemy, "Detection of small and slow moving objects in an image sequence", *Scandinavian Conference on Image Analysis, Uppsala*, juin 1995, Vol.1, p.155-162.

[11] Y.Z. Hsu , H.H. Nagel et G. Rekers, "New likelihood test methods for change detection in image sequences", *Computer Vision, Graphics and Image Processing*, 1984, Vol.26, p.73-106.

[12] P. Bouthemy et P. Lalande, "Detection and tracking of moving objects based on a statistical regularization method in space and time", *1st European Conference on Computer Vision*, avril 1990, p.307-311.

[13] J.M. Létang, V. Rebuffel et P. Bouthemy, "Motion Detection Robust to Perturbations : a statistical regularization and a temporal integration framework", *International Conference on Computer Vision, Berlin*, mai 1993, p.21-30.

[14] M.J. Black, "Combining Intensity and Motion for Incremental Segmentation and Tracking Over Long Image Sequence", *2nd European Conference on Computer Vision, Santa Margherita Ligure*, mai 1992, p.485-493.

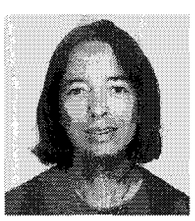
[15] V. Rebuffel, "Rapport de tranche *T*0 du contrat DRET 93465", CEA-LETI, juillet 1995.

[16] V. Rebuffel, "Rapport de tranche *T*0'-lot1 du contrat DRET 93465", CEA-LETI, décembre 1995.

Manuscrit reçu le 20 septembre 1995.

LES AUTEURS

V. REBUFFEL



Véronique Rebuffel est titulaire d'un diplôme d'ingénieur de l'Ecole Polytechnique obtenu en 1981, d'ingénieur de l'Ecole des Mines en 1983. Elle a rejoint le LETI en 1988. Son activité s'est d'abord développée en morphologie mathématique, vision industrielle, et systèmes experts pour la vision par ordinateur. Depuis 1991, elle consacre ses recherches à la vision dynamique : détection de cibles mobiles, estimation de mouvement, analyse dynamique de scènes, mouvement déformable, ainsi qu'aux modèles statistiques en traitement d'images.

C. HENNEBERT

Ingénieur de l'ESIGELEC de Rouen, Christine Hennebert a obtenu en 1993 un DEA de l'INPG de Grenoble, mention Signal-Image- Parole. Elle soutiendra une thèse de doctorat en octobre 1996, thèse actuellement préparée au sein du LETI à Grenoble. Son principal centre d'intérêt est l'analyse du mouvement 2D appliquée à la détection de cibles mobiles dans des séquences d'images de scènes réelles visibles ou infrarouges.

J.M. L'ETANG

Ingénieur en génie électrique de l'INSA Lyon, Jean-Michel Létang a reçu en 1990 le DEA de l'Université Cl.Bernard de Lyon en Génie biologique et médical, puis en 1993 le titre de Docteur de l'INPG de Grenoble, mention Signal-Image-Parole. A la suite de ses recherches doctorales menées au LETI, il a été invité à l'INRS-Télécommunications de Montréal en 1994 et à l'INRIA de Grenoble en 1995. Actuellement aux Laboratoires d'électronique de Philips, ses principaux centres d'intérêt sont la vision par ordinateur, l'analyse du mouvement, les modèles statistiques et l'imagerie médicale.