Chapitre IV

La détection du mouvement

1. Introduction

L'analyse du mouvement se distingue fortement de l'analyse d'images statiques. Premièrement et d'un point de vue pratique, on constate que la contrainte *temps-réel* s'impose presque toujours (pour des applications embarquées par exemple). D'un point de vue traitement du signal, l'analyse du mouvement intègre des *informations temporelles* extraites à chaque instant à partir d'au moins deux clichés de la séquence, voire encore de la séquence toute entière[40]. Les signaux en étude ont à la fois une dimension spatiale et une dimension *temporelle*. Enfin et d'un point de vue algorithmique, l'analyse du mouvement gère des modèles parfois très complexes de la scène observée (modèles d'éclairement variable pour les surfaces des objets tridimensionnels en mouvement, etc.).

On se propose ici de montrer comment on peut appliquer efficacement le formalisme bayesien étudié au Chap.I au problème de la *détection du mouvement*. Il s'agit d'une première étape conceptuellement élémentaire mais qui peut s'avérer cruciale pour aborder correctement une analyse plus approfondie du mouvement. Comme il a été vu à la fin du Chap.I (§4), l'application du formalisme bayesien aux problèmes de traitement d'images bas niveau réalise une "traduction" de ceux-ci en termes de *fonction d'énergie à minimiser*. Les algorithmes de détection du mouvement que l'on étudiera par la suite conduisent tous à la définition de fonctions d'énergie n'ayant pas de propriétés remarquables susceptibles de simplifier cette tâche d'optimisation (convexité de la fonction et continuité du domaine de définition des variables par exemple). Etant donné la contrainte temps-réel, il est important de résoudre ce problème au moins à cadence vidéo, sans qu'il y ait pour autant une dégradation significative de la qualité de la détection. La difficulté du problème a poussé la plupart des chercheurs utilisant des stations de travail conventionnelles à proposer des modèles très simplifiés (taille réduite du champ à optimiser, limitation du nombre de paramètres à estimer) et à se limiter à

[40] Dans ce cas le traitement sera dit en *temps-différé*, car la durée du traitement dépend de la taille de la séquence, et cette durée dépasse en général le temps imparti au traitement dans une application *temps-réel*.

l'utilisation d'algorithmes d'optimisation déterministes. Cependant, on sait que les procédures d'optimisation stochastiques de type recuit simulé, bien qu'extrêmement gourmandes en calcul donnent théoriquement les meilleurs résultats, indépendamment du modèle de détection choisi. Voilà qui est intéressant : ne pourrait-on pas utiliser un processeur du type proposé au chapitre précédent pour rendre compte de ce problème sans compromettre la contrainte *temps-réel*?

Notre objectif est de démontrer la possibilité d'intégrer un processeur de type PPOS dans une chaîne de traitement d'images *à cadence vidéo*[41]. Nous avons choisi en l'occurrence le problème de la détection de mouvement. Il faut tout d'abord -et c'est l'objet de ce chapitre- établir un modèle suffisamment simple pour pouvoir être implanté avec des éléments optoélectroniques conventionnels, sans qu'il y ait pour autant une perte significative de la *qualité* du traitement par rapport aux modèles complexes mais *temps-différés* qui tournent sur station de travail.

2. L'analyse du mouvement

2.1 Introduction

L'analyse du mouvement est une tâche extrêmement complexe impliquant en même temps le développement de capteurs de vision plus ou moins spécialisées ainsi qu'une capacité de calcul très importante capable de traiter un flot de données-images continu. On ne peut expliquer que l'évolution naturelle ait permis le développement de systèmes de vision aussi complexes que l'œil et le cortex visuel que par l'énorme avantage biologique que de tels systèmes confèrent aux organismes les possédant ; il est évident que dans le règne animal (de *animalis* = animé, qui *bouge*), l'information du mouvement est un indice précieux indiquant la présence de la proie ou du prédateur.

Mais le champ du mouvement ne saurait être complètement déterminé par la seule connaissance de la séquence d'images. En toute rigueur, le champ des vitesses est une information supplémentaire à l'interprétation de la scène (comme c'est le cas par exemple dans l'analyse d'écoulements en mécanique de fluides où l'on combine résultats de simulations et expériences parfois complexes et astucieuses -coloration des fluides, méthodes optiques interférentielles- pour essayer de *voir directement* le champ de vitesses). D'un point de vue physique, l'image directe "optique" et l'image indirecte (parfois inaccessible) du champ des vitesses décrivent mieux *ensemble* le phénomène réel, en reconstruisant plus ou moins bien son espace de phases[42].

---

[41] En vision artificielle, le terme *temps-réel* est souvent rattaché à la notion de *cadence vidéo*. La cadence de traitement doit être de l'ordre de 25Hz ; ceci conduit, pour des images à 256 niveaux de gris et de taille 256x256 pixels, à un flux de données en entrée qui est de l'ordre de 13Mb/s.

[42] Remarquons toutefois que l'estimation du champ de vitesses physique dans une scène n'est pas nécessairement le but de l'estimation du mouvement pour tout système de vision ; par exemple, *l'ombre*

Voici le problème qui nous intéresse : en se limitant au seul capteur *d'images*, existe-t-il un moyen de reconstruire *l'espace des vitesses 3-D réel* ? La réponse est certainement - et dans une large mesure- affirmative, puisque nous mêmes sommes capables d'interpréter très efficacement une séquence d'images arrivant sur notre rétine pour en déduire le mouvement réel des objets qui nos entourent. Ceci est possible moyennement un certain nombre d'hypothèses *a priori* et des *modèles* complexes plus ou moins inconscients, innés et affinés par l'expérience quotidienne.

2.2 Hiérarchie fonctionnelle dans les systèmes de vision.

Voici un liste des étapes ou échelons que l'on distingue classiquement à la fois par le niveau *d'interprétation* et par la *complexité du modèle* impliqué dans l'analyse du mouvement :

- *La détection du mouvement.* C'est l'étape la plus élémentaire, visant à distinguer les zones fixes et les zones mobiles d'une séquence d'images. Le champ estimé (appelé carte de détection) est de nature binaire.

- *L'estimation du mouvement.* C'est une étape qui peut encore être traitée dans le cadre de l'analyse *bas niveau* ; il s'agit de déterminer, pour chaque point de l'image, un vecteur vitesse ou déplacement. La carte des vitesses ainsi construite est appelé *flot optique*.

- *La segmentation du mouvement.* Le but est de segmenter l'image en zones homogènes de mouvement. Ces zones vont correspondre vraisemblablement aux différents objets mobiles de la scène ; c'est pourquoi cette opération utilise non seulement les résultats plus ou moins directs du flot optique, mais intègre aussi des modèles des objets mobiles.

- *L'interprétation du mouvement* qualitative (s'agit il d'une translation, d'une rotation) ou quantitative (évaluation d'un torseur cinématique tridimensionnel), la reconnaissance des formes des objets mobiles, etc..

Notons au passage qu'une hiérarchie similaire se retrouve dans le système de vision des vertébrés associé à une surprenante modularité fonctionnelle : tout se passe comme s'il existaient *plusieurs* sens de vision indépendants (mouvements, couleurs, reconnaissance de formes spécifiques - visages, expressions, caractères écrits, etc), dont les résultats fusionneraient pour créer une représentation haut niveau (et unique) du monde (voir *fig.*I.2). C'est une remarque intéressante qui peut nous aider à comprendre

---

mouvante d'un lion qui ne se trouve pas dans notre champ de vision ne donne pas lieu à des changements du champ de vitesses *réel* de la scène, mais peut être une information précieuse pour en déduire le mouvement *physique* de celui-ci - et agir en conséquence.

que le sens de la vision (comme la plupart des processus cognitifs élaborés) est loin d'être un processus instantané et uni-modulaire comme on aurait eu tendance à le croire jusqu'à il n'y a pas si longtemps[43]. Il résulte au contraire de l'association d'un grand nombre de modules et de stratégies certainement "découvertes" par hasard et accumulées par voie d'une longue évolution darwinienne. Le résultat est un véritable *patchwork* heuristique (*cf. fig.*I.2). A fortiori, il est absurde de chercher un modèle mathématique unique pouvant rendre compte en même temps de tous ces processus différents.

3. La détection du mouvement

3.1 Introduction.

Nous allons consacrer le reste du chapitre au problème bas niveau de *détection du mouvement*. Une bonne détection des régions en mouvement peut s'avérer indispensable pour aborder correctement les étapes plus en aval de l'analyse (segmentation des régions en mouvement, amélioration significative de la qualité du calcul classique du *flot optique* -par exemple [Horn81]). Par ailleurs, comme on l'a remarqué en conclusion au Chap.I, les opérations bas niveau en traitement d'images ont beau être élémentaires, elles ne sont pas pour autant celles qui engendrent le moins de calculs - bien au contraire. A elles seules, ces deux considérations suffisent déjà à justifier pleinement que nous ayons porté notre intérêt sur la détection du mouvement comme application potentielle pour notre processeur dédié. En effet, les algorithmes d'optimisation stochastique fournissent théoriquement de bien meilleurs résultats que les algorithmes d'optimisation déterministes : on peut donc espérer l'obtention d'une *carte ou masque* de détection qualitativement très amélioré. D'autre part il ne faut pas oublier que l'un des objectifs de notre recherche sur les processeurs optoélectroniques stochastiques est précisément de démonter leur capacité à traiter correctement les énormes charges de calcul engendrées par les problèmes de traitement d'images bas niveau *à cadence vidéo*. Mais ce n'est pas tout : la détection du mouvement est à elle seule une opération intrinsèquement intéressante et qui peut s'avérer utile dans un grand nombre de tâches automatiques (contrôle routier, suivi de cibles, compression d'images, etc.). C'est pourquoi, bien que conduisant à une interprétation sommaire de la scène, la détection du mouvement a

---

[43] C'est peut être l'influence d'une conception dualiste qui aurait relégué à l'*âme* toute la responsabilité des processus cognitifs de haut niveau. Pour Descartes par exemple, l'interprétation d'une image est un fait *immédiat* et qui ne souffre pas de description : après avoir décrit assez précisément la formation (optique) des images sur la rétine et l'acheminement de celles-ci "par l'entremise des nerfs optiques, dans la superficie interieure du cerveau" ("*Les passions de l'âme*", *art.*XXXV), il continue : "puis de là, par l'entremise des esprits dont ces cavités sont remplies, ces images rayonnent en telle sorte vers la petite glande que ces esprits environnent, que le mouvement qui compose chaque point de l'une des images, tend vers le mesme point de la glande [...] au moyen de quoy les deux images qui sont dans le cerveau n'en composent qu'une seule sur la glande, qui agissant immediatement contre l'ame, luy fait voir la figure de cet animal." (Pour un développement passionnant de la conception "fonctionnaliste" moderne, voir l'ouvrage de Daniel.C.Dennett, "Consciousness explained").

suscité très tôt de nombreuses recherches et des réalisations motivées par les applications industrielles, civiles ou militaires.

3.2 Description du problème.

La détection du mouvement consiste à distinguer dans une séquence d'images, les *objets mobiles* par rapport aux objets fixes -et en l'occurrence par rapport à un *fond* fixe. Cette classification s'effectue en affectant une *étiquette* à chacun des points de l'image, et cela pour chaque image de la séquence. Dans le cas le plus simple, un étiquetage binaire suffit conduisant à un *masque binaire des objets mobiles* noté *MB* (pour chaque point de l'image, ou bien celui-ci est fixe, ou bien il est mobile)[44] ; mais des classifications plus exhaustives peuvent se révéler utiles comme point de départ pour des traitement de plus haut niveau (le *multi-étiquetage* permet de distinguer plusieurs objets différents en mouvement).

La détection du mouvement s'appuie sur l'étude des variations temporelles de la fonction de luminance. Deux hypothèses simplificatrices sont généralement admises :

- d'une part, on suppose le capteur fixe (autrement il faudrait redéfinir la notion de mouvement dans une scène) ;
- d'autre part on suppose que l'éclairage de la scène observé est quasi constant (en principe il ne devrait pas y avoir d'ombres, mais si tel est le cas elles seront traitées comme s'il s'agissait d'objets mobiles).

Dans ces conditions, toute variation temporelle de l'intensité est nécessairement liée soit au mouvement, soit à la présence de bruit. A partir de là, plusieurs approches ont été utilisées pour tenter de reconstruire le *MB*.

---

[44] Une remarque tout de suite : le *masque binaire des objets mobiles* (MB) et la *carte des changements temporels de l'image* (CCT - voir plus loin) sont des choses bien différentes. L'obtention de cette dernière n'est qu'une étape élémentaire dans la construction beaucoup plus élaboré du masque binaire des objets mobiles.

4. Obtention du masque des objets mobiles (MB)

4.1 Utilisation d'une image de référence.

L'idée est de disposer d'une image, dite *image de référence* qui contient exclusivement le fond fixe de la séquence à analyser. A partir de là, tout objet présent dans l'image analysée mais absent de l'image de référence sera considéré comme un objet en mouvement. En pratique, il suffit de calculer la différence entre l'image courante et l'image de référence et de binariser le résultat (ce qui permet du coup un traitement élémentaire du bruit). Cependant, l'obtention de l'image de référence est loin d'être évidente. Une mise à jour est indispensable afin de tenir compte des éventuelles modifications du fond (changement d'éclairement ou changement de "statut" d'un objet, qui peut cesser de se déplacer ou au contraire commencer à bouger).

Plusieurs stratégies ont été proposées pour l'obtention de l'image de référence. Les plus simples se basent sur l'enregistrement direct de la scène dépourvue d'objets mobiles [Marvall95]. Si cela n'est pas possible (c'est à dire dans la plupart des cas), on peut essayer de reconstruire petit à petit l'image de référence à partir de la séquence d'images en cours d'observation [Jain79]. La construction et la mise à jour de l'image de référence peut se faire par filtrage temporel récursif [Bulas93]. Par ailleurs, certains auteurs proposent un filtrage plus élaboré tel que le filtrage prédictif de Kalman [Karman90].

En résumé, si la méthode peut paraître simple, la difficulté d'obtention de l'image de référence restreint son utilisation à des cas particuliers (la technique est surtout utilisée dans l'analyse des scènes routières [Koba87]). Un point intéressant est pourtant le fait que la mise à jour continue de l'image de référence autorise naturellement la prise en compte d'éventuelles variations d'éclairement.

4.2 Utilisation de la carte des changements temporels.

L'introduction d'informations temporelles dans ce modèle se fait par le biais *d'une carte binaire des changements temporels* (CCT), qui découle d'une analyse plus ou moins immédiate de la fonction de luminance par rapport au temps. L'idée est d'obtenir une carte des changements qui se produisent à chaque instant dans l'image, en essayant de minimiser les effets parasites dus au bruit. Comme on le verra par la suite, la carte des changements temporels n'est qu'un premier pas vers la reconstruction du masque des objets mobiles[45].

[45] La carte de changements temporels (CCT) seule n'est autre chose que le résultat plus ou moins bien approché par le "filtre de nouveautés".

4.2.1 Obtention de la CCT

Plusieurs méthodes ont été proposées pour aboutir à la carte des changements temporels :

a) Utilisation de la différence inter-image (DI). La technique la plus simple pour aboutir à la carte des changements temporels consiste à étudier la dérivée temporelle de la fonction de luminance, approchée de façon discrète par la différence entre deux images successives[46] :

$$DI(s,t) = \left| \frac{\partial I(s,t)}{\partial t} \right| = |I(s,t) - I(s,t-1)|$$

Les images réelles n'étant jamais exemptes de bruit, la différence inter-image ne fera que l'amplifier. Néanmoins, étant donné la nature du bruit, celui-ci se verra amplifié *localement* lors de l'opération de dérivation. Il est alors possible de le réduire en réalisant un filtrage passe-bas de la différence inter-images [Diel91].

Toujours à cause du bruit, il est préférable de choisir un *seuil* pour distinguer plus sûrement les points où la variation de luminance est significative (due au mouvement) de ceux où elle est plus faible (donc due au bruit). La carte de changements temporels significatifs est donc obtenue selon :

$$CCT(s,t) = \begin{cases} 0 & si \quad DI(s,t) < \theta \\ 1 & sinon. \end{cases}$$

Cette technique de détection a été abondamment utilisée [Aach93], [Lalan90a]. Remarquons que la carte de changements temporels qui en résulte est binaire, et que le seuil θ est un premier paramètre du modèle. Le choix de ce paramètre peut se révéler crucial dans le cas d'images trop bruitées ou peu contrastées. Aussi, dans certains cas – connaissance a priori de la statistique du bruit des objets en mouvement et du fond -, ce paramètre peut être déterminé de façon optimale. Dans d'autres cas, si l'on connaît seulement le modèle du bruit, ce paramètre peut être estimé en temps réel sur la séquence en traitement.

La méthode de différence seuillée présente l'avantage d'être très simple d'implantation. Cependant, l'opération de binarisation pour un pixel donné ne tient compte que de l'information de mouvement qui lui est associé : autrement dit, le résultat est peu robuste vis-à-vis du bruit. Pour renforcer la robustesse des détecteurs de

---

[46] On suppose tout le long du chapitre que les images à traiter sont à niveaux de gris ; on pourrait cependant utiliser l'information de couleur pour réaliser une détection de changements temporels plus efficace pour les images en couleur : la DI serait alors donné par la norme de la différence entre deux vecteurs 1x3 (RGB), et non pas par la différence entre les *projections* de ces mêmes vecteurs sur l'axe des "niveaux de gris".

changements temporels, certains auteurs ont eu l'idée de prendre la décision relativement à un voisinage centré autour du point, et de s'appuyer sur des tests statistiques : c'est ce qui est développé ci-dessous.

b) Méthode de détection des changements à l'aide des test de vraisemblance. Dans ce genre de méthodes, on considère deux fenêtres carrés centrées sur chaque pixel de l'image pour deux images consécutives. Le test est réalisé en calculant les probabilités associées aux hypothèses suivantes (d'où le nom du test) : (1) les fenêtres ont des distributions de gris identiques et (2) les fenêtres ont des distributions différentes, et dans ce cas on considère qu'il y a eu un changement temporel. Différentes modélisations de la distribution des niveaux de gris dans la fenêtre ont été étudiées : modélisation constante [Skif89], modélisation linéaire et modélisation quadratique [Hsu84]. La carte des changements temporels étant binaire, il faut choisir aussi un seuil de binarisation. La plupart du temps ce seuil est fixé de façon arbitraire, mais dans certains cas il peut être estimé à partir des données statistiques [Aach93].

Si l'on ne connaît pas précisément la statistique des images (scènes de rue par rapport aux scènes répétitives dans une chaîne de production par exemple), alors les masques reconstruits après une simple différence seuillée sont en général moins précis et plus bruités que ceux obtenus à l'aide d'une méthode de vraisemblance.

c) Autres méthodes. Remarquons enfin qu'il existe des méthodes permettant d'intégrer de faibles variations d'éclairement dans le calcul de la carte de changements temporels [Skif89], [Ulst73], ainsi que des méthodes basées précisément sur des modèles d'éclairement [Skif89], [Phong75] permettant d'obtenir des cartes de changement temporels pour des images où les objets changent d'*orientation* par rapport à la source lumineuse.

4.2.2 Reconstruction du MB à partir de la CCT.

La carte de changements temporels ne fournit pas directement le masque des objets mobiles. Outre les détections parasites, la carte des changements temporels fait apparaître quatre zones distinctes (voir *fig*.IV.1) :

(1) : le *fond* ou zone *fixe*, où l'amplitude de la différence des niveaux de gris est faible et donc au dessous du seuil de binarisation ;

(2) et (3) : la zone *d'écho*, et la zone de *recouvrement*, constituées respectivement des pixels qui appartenaient à l'objet mobile et qui ont été découverts et recouverts suite au mouvement. L'amplitude de la différence y est plus importante, et dépasse le seuil de binarisation θ ;

(4) : la zone de *glissement*, pour laquelle la différence des niveaux de gris est faible et reste en dessous du seuil binarisation. Remarquons toutefois que l'objet mobile peut être texturé -et qu'il est en tout cas rarement uniforme en niveau de gris ; l'hypothèse

simplificatrice que nous sommes implicitement en train d'admettre consiste à supposer que les variations de niveaux de gris sur l'étendue spatiale de l'objet sont moindres que celles que l'on obtient par différence entre n'importe quelle partie de l'objet mobile et le fond.

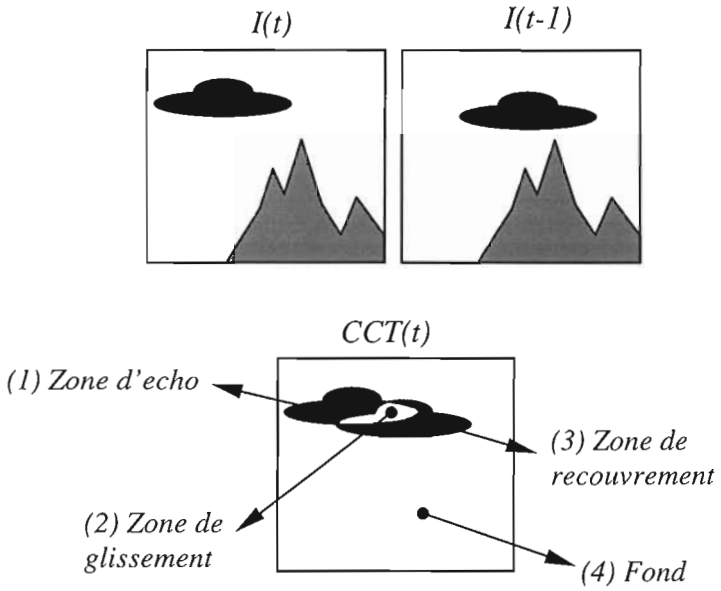

Fig.IV.1: Carte de changements temporels (CCT) obtenue par différence seuillée (ou par test de vraissemblance) entre deux images consécutives.

L'existence de l'ensemble de ces quatre zones dépend d'une part de la fréquence d'acquisition du capteur d'images par rapport à la vitesse l'objet mobile, et d'autre part de la taille des détails de l'objet mobile que l'on doit être en mesure de détecter par rapport à la résolution du capteur d'images. Dans ce qui suit, nous allons nous placer toujours dans l'hypothèse que ces quatre zones existent au moins pour les zones "intéressantes" de l'objet mobile (on dira que la séquence est "bien conditionnée" par rapport aux hypothèses faites au paragraphe §3.2.

> **Remarquons** en particulier que si l'objet mobile possède une texture "fine" (régulière ou pas), et que si la vitesse de déplacement de l'objet est suffisamment élevée par rapport à la cadence du capteur d'images compte tenu de la taille de la texture, alors celle-ci *n'est pas un détail de l'objet mobile* pour lequel la carte des changement temporels puisse fournir les quatre zones en question.

Pour obtenir le MB, il faut traiter la CCT pour (a) éliminer la zone d'écho, (b) intégrer la zone de recouvrement, (c) compléter la zone de glissement et (d) diminuer les détections parasites. Plusieurs stratégies ont été proposées à cette fin :

- Méthode du ET logique [Wiklun87]. Elle consiste a réaliser un ET logique entre les cartes de changement temporels calculées à partir de trois images successives. La méthode est très rapide mais ne donne de bons résultats que s'il n'y a pas de recouvrement entre les positions successives de l'objet mobile (pas de zone de glissement). C'est une condition assez restrictive.

- Certains auteurs proposent l'utilisation d'heuristiques *ad-hoc* pour reconstruire le masque des objets mobiles : moyennant certaines hypothèses plus ou moins arbitraires quant à la taille de l'objet et aux caractéristiques du bruit, [Diel91] propose de regrouper les points isolés avec leurs voisins grâce à un filtre médian 5x5, d'éliminer les zones détectées de trop petite taille, ainsi que de "boucher" les trous à l'intérieur des régions de changement temporel (zones de glissement). La méthode propose donc une régularisation des images provenant d'une connaissance de leur caractéristiques intrinsèques ;

- une méthode plus élégante pour régulariser la carte de changements temporels est d'utiliser la **modélisation markovienne**. Ce formalisme permet la prise en compte de connaissances a priori pertinentes sur la solution cherchée telles que l'homogénéité spatiale et temporelle des masques. Cette approche présente en plus l'avantage de réaliser en même temps le débruitage de l'image et le traitement d'images en mouvement même quand celui-ci est de faible amplitude (remplissage de zones de glissement). Plusieurs variantes du modèle markovien ont été proposées, plus ou moins complexes et plus ou moins particulières, allant par exemple du modèle à vingt paramètres de [Kurians95][47] pouvant traiter des objets uniformes en rotation, jusqu'au modèle simplifié de Dumontier [Dumont96] à trois paramètres dont il sera question au chapitre suivant.

4.3 Techniques diverses pour l'obtention directe du MB.

Il serait sûrement inutile de chercher à trouver un algorithme permettant de traiter tous les cas de figure en ce qui concerne la détection du mouvement. Les astuces au cas par cas permettent de faire beaucoup mieux là où les algorithmes trop généraux ne donneraient que des résultats médiocres. En voici quelques exemples n'utilisant pas la méthode de la CCT ni l'image de référence :

- Les méthodes décrites auparavant ne rendent pas compte du mouvement d'objets de petite taille (détection et suivi de corps célestes en mouvement – planètes, satellites naturels ou artificiels -, suivi de cibles lointaines, etc.). En effet, de par leur taille, les modèles précédemment décrits n'y verront que du bruit. Dans [Blostein91], à chaque pixel sont associés plusieurs trajectoires possibles, calculées à partir de tests statistiques dans l'hypothèse d'un mouvement constant. Une autre approche [Letang93] assimile chaque point de l'image à un signal temporel ; une analyse fréquentielle permet alors d'éliminer les variations temporelles parasites, et de faire ressortir les petits objets qui se déplacent lentement.

---

[47] la profusion de paramètres est symptomatique -à mon avis- d'une tendance à vouloir donner trop de généralité au modèle (*cf.* remarque au §2.2) ; à la fin c'est un système beaucoup plus complexe (nous mêmes) qui finit par choisir la valeur des paramètres en fonction du problème.

- Il existe également des méthodes basées sur la détection des contours en mouvement [Vieren88], [Stelma85]. Elles consistent pour la plupart à utiliser conjointement un détecteur de contours et un détecteur de changements temporels.

- Des stratégies multi-échelles ont aussi été adoptées, par exemple dans [Gil93], où la détection des changements temporels est effectuée sur plusieurs niveaux d'une pyramide d'images. La localisation de l'objet mobile est affinée le long de la pyramide en tenant compte des informations de contour provenant du signe de la différence inter-image.

4.4 Conclusion

D'après ce rapide tour d'horizon des diverses techniques de détection de mouvement, il ressort clairement qu'il n'existe pas une méthode universelle permettant de traiter tous les cas de figure. Soit parce que les besoins sont différents (plus ou moins de détail quant à la forme de l'objet, contraintes de temps de calcul, etc.) soit parce les séquences à traiter proviennent elles-mêmes d'origines très différentes, chaque application nécessite le développement de méthodes spécifiques - ou au moins l'adaptation d'une technique déjà existante. La **régularisation markovienne de la carte de changements temporels** semble toutefois se dégager puisqu'elle permet d'intégrer dans le même modèle des informations d'origines diverses, et de s'adapter à tous les besoins par l'ajustement d'un ensemble de paramètres plus ou moins réduit. Cette méthode présente alors le double avantage d'avoir une certaine généralité, tout en respectant un cadre mathématique assez rigoureux - ce qui permet en outre de mieux appréhender la signification et l'importance de chaque paramètre du modèle. Par la suite, nous allons nous concentrer sur la méthode de détection du mouvement basée sur cette technique.

5. Régularisation markovienne de la CCT

5.1 Introduction

Tout comme grand nombre de problèmes en traitement d'images, la détection du mouvement par traitement de la CCT appartient à la classe de problèmes dits *mal posés* : la carte de changement temporels ne donne qu'un *indice* de la solution recherchée (i.e. le masque binaire des objets mobiles MB). La modélisation markovienne propose de restreindre l'ensemble de solutions possibles par l'introduction d'un certain nombre de *contraintes a priori.*

Plus précisément, la régularisation de la carte des changements temporels par *champs de Markov* et *estimation bayesienne,* consiste à : (1) définir un *modèle d'observation* liant la ou les observations (différence inter-image, carte de changements temporels, etc.) avec le champ d'étiquettes binaire ; (2) modéliser le champ d'étiquettes binaires MB par un

champ de Markov ; et (3) utiliser une procédure d'optimisation - stochastique ou déterministe - pour obtenir l'estimé du *MAP* du champ d'étiquettes final. Pour éviter les confusions, nous allons noter E le champ d'étiquettes binaires considéré comme un champ aléatoire (MRF) et nous allons réserver la notation *MB* seulement pour le résultat final de l'estimation : $MB \equiv \hat{e}^{MAP}$ (*cf.*Chap.I,§2.2).

5.2 Interprétation des variables.

Dans ce qui suit, on reprendra les notations du formalisme probabiliste bayesien et du modèle markovien (*cf.*Chap.I,§3), tout en donnant une interprétation des variables correspondant au problème de la détection du mouvement.

5.2.1 Modèle d'observation

Les données. Soit $\{I(0), I(1),\ldots, I(t)\}$ la séquence d'images à analyser, c'est à dire l'ensemble d'images perçues par le capteur jusqu'à l'instant présent t - dans la pratique, on dispose au plus de quatre ou cinq images en même temps. Comme on vient de le voir, la technique de régularisation de la carte des changements temporels ne s'appuie pas directement sur la donnée image, mais sur la différence inter-image. Par analogie avec les méthodes de traitement d'images statiques pour lesquelles les données à traiter sont directement issues de l'observation, on appellera ce champ *le champ d'observations à valeurs continues O(t)* :

$$O(t) \equiv DI(t) = \left| I(t) - I(t-1) \right| \qquad (1)$$

L'obtention de ce champ implique un prétraitement des données-image. Afin de réduire le bruit, on utilise parfois un filtre passe bas (masque ¼[1 2 1] en ligne et en colonne) avant de calculer la différence. Dans le cas de véritables traitements spatio-temporels, le champ d'observation est constitué lui aussi de plusieurs champs bidimensionnels : $O=\{O(1),\ldots,O(t)\}$. Enfin, certains modèles conservent le signe de la dérivée temporelle, ce qui permet de différencier la zone d'écho de la zone de recouvrement.

Enfin, la carte de changements temporels (CCT) est bien sûr un champ de contraintes aux données (qui peut ou non résulter de la différence inter-image par simple seuillage *cf.*§4.2.1). Par la suite, on appellera ce champ le *champ d'observations binaires*. Il sera noté $\hat{O}(t)$ et l'on a alors :

$$\hat{O}(t) \equiv CCT(t) \qquad (2)$$

Lien entre étiquettes et observations. Le lien entre le champ de mouvement binaire E et les champs d'observations - $O(t)$ et/ou $\hat{O}$(t) - n'est pas immédiat comme c'est le cas par exemple dans le problème de la restauration d'images bruitées (*cf.*Chap.I,§5.1). Tout d'abord, il n'existe pas une définition claire de ce qu'est la *détection du mouvement* (l'ombre d'une automobile en mouvement est-elle bien un objet en *mouvement* qui doit être détecté? le "bruit" présent dans l'image provenant par exemple d'un nuage de

poussière ou de la pluie ne provient-il pas d'un phénomène de *mouvement*?). Le modèle dépend très étroitement de l'objectif recherché, mais celui-ci n'est pas non plus très simple à définir.

On se bornera par la suite à la description détaillée de quelques modèles mettant en jeu de fonctions prédéfinies plus ou moins *ad-hoc* ; dans ces modèles, le lien entre le champ d'étiquettes binaires et le champ d'observation continu est défini formellement par :

$$O = \Psi(E) + B\,,$$

où B est un vecteur de bruit gaussien non corrélé en espace et en temps, et Ψ est une fonction prédéfinie [Lalan90a]. Par ailleurs, on remarquera que dans ces modèles, le champ d'observation binaire (i.e. la carte des changements temporels $\hat{O}(t)$) n'impose pas directement une contrainte sur le champ d'étiquettes, mais sur *la forme des potentiels des cliques* du champ spatio-temporel de Markov. Sans casser complètement un lien temporel, le champ $\hat{O}(t)$ tantôt réduit le poids d'une interaction, tantôt l'augmente : il est l'équivalent d'un *bord* tridimensionnel (invariable pendant le calcul), qui rajoute *a priori* des informations locales sur la forme finale du champ d'étiquettes.

5.2.2 Le champ de Markov (masque binaire des objets mobiles ou champ d'étiquettes E).

Champ de Markov spatio-temporel. Comme on l'a dit auparavant, la détection du mouvement consiste à déterminer l'état de mouvement de chaque pixel de chaque image de la séquence. Il s'agit, comme pour tant d'autres problèmes de traitement bas niveau, d'un problème de *segmentation*. La seule différence réside dans le fait que les données sont à la fois spatiales et temporelles. Pour chaque séquence d'images $I(s,t)$, le résultat de l'estimation serait idéalement une séquence de même longueur, c'est à dire un champ d'étiquettes binaire tridimensionnel : le *masque spatio-temporel binaire des objets mobiles MB(s,t)*, issu d'une segmentation spatio-temporelle des pixels (*fig*.IV.2).

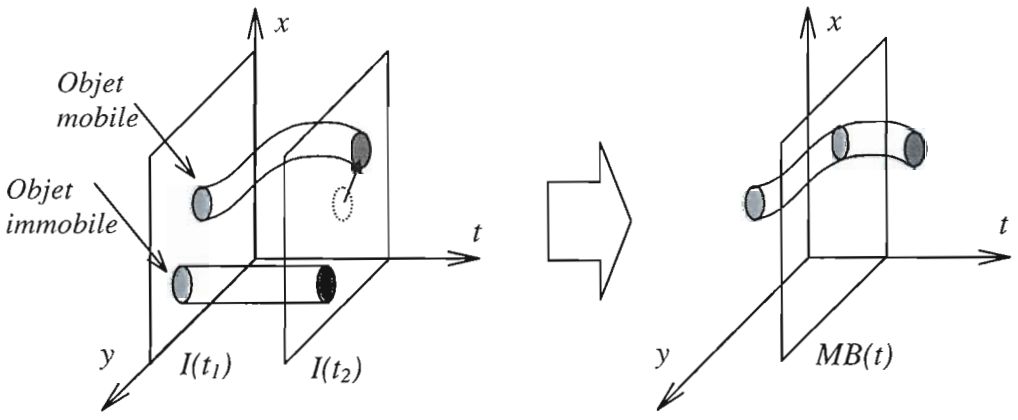

Fig.IV.2 : Resultat de la segmentation spatio-temporelle : le champ d'étiquettes final correspond au masque spatio-temporel binaire des objets mobiles MB.

Cependant, la contrainte temps-réel implique un traitement *causal* : on ne dispose pas de la séquence préenregistrée depuis le début jusqu'à la fin. A chaque instant, les données futures sont inaccessibles. Par ailleurs, pour une question de difficulté d'implantation, toutes les techniques présentées vont se restreindre au traitement d'une tranche temporelle très fine du champ spatio-temporel de labels -voir à un *seul* champ spatial *E(t)*. Les labels en dehors de cette tranche seront considérées comme constants et ne seront pas relaxés pendant la phase d'optimisation.

Forme du voisinage spatio-temporel. La *fig.*IV.3 illustre un exemple de voisinage spatio-temporel : le champ de Markov est défini à tout instant par l'ensemble des trois champs binaires *E(t-1), E(t)* et *E(t+1)*. Un voisinage plus simple que celui de la figure (voisinage d'ordre 1 aux 6 plus proches voisins par exemple) peut favoriser la reconstruction des masques d'objets à contours horizontaux ou verticaux, ainsi que celle des masques des mouvements de translation horizontaux ou verticaux.

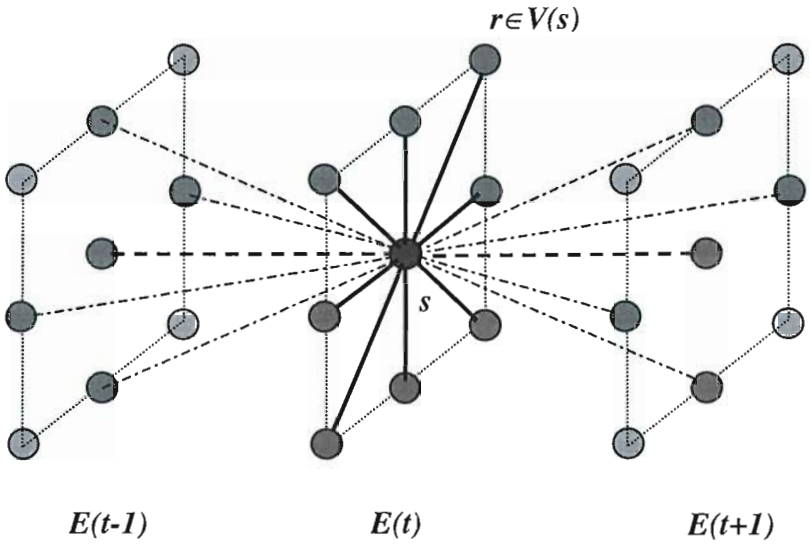

Fig.IV.3 : Voisinage spatio-temporel du site s -aux 18 plus proches voisins- et cliques binaires associées : en gras, les cliques spatiales ; en tirets les cliques temporelles ; en tirets plus fins, les cliques spatio-temporelles.

Potentiel des cliques. Les potentiels des cliques spatiales sont fixés une fois pour toutes et sont invariants par translation spatiale : ils caractérisent le degré de régularisation spatiale du masque binaire du mouvement. Par contre, le potentiel de la clique temporelle peut être contrôlé par la valeur locale (en temps et en espace) du champ d'observation binaire $\hat{O}(t)$. La raison en est que si en un site donné on a mesuré un *changement*, il est probablement lié à un changement de *l'état de mouvement réel*: il faut alors neutraliser la liaison qui force le site présent et passé à avoir la même valeur -ou même en inverser le poids.

On distinguera les vrais modèles spatio-temporels des modèles spatiaux *à contrainte temporelle* : dans ces derniers, il n'existe pas de véritable voisinage temporel. Le champ de Markov est associé exclusivement au champ d'étiquettes courant *E(t)*, tandis que les champs *E(t-1)* - et/ou *E(t+1)* - sont *gelés* pendant la régularisation de *E(t)* et agissent comme des champs de *contrainte*. Dans ces modèles, si les liaisons temporelles restent abusivement représentés par des cliques, la raison en est que l'énergie de contrainte temporelle est souvent déduite par simplification d'un véritable modèle spatio-temporel.

5.2.3 Energie globale, force locale et optimisation.

On a vu au Chap.I,§4 que le choix de l'estimateur MAP rendait le problème équivalent à celui de la minimisation d'une fonction d'énergie ayant deux termes : un terme de *régularisation* et un terme d'attache aux données de *l'observation* (les notations minuscules correspondent aux réalisations des variables aléatoires) :

$$U(e,o) = U_{obs}(e,o) + U_{reg}(e)$$

La fonction d'énergie est parfaitement définie une fois que l'on a choisi le modèle d'observation et la forme du voisinage (topologie et potentiels de cliques). On est alors à même de calculer la *force locale* ou *gradient* au site (s,t) :

$$F(s,t) = -\Delta U(e/o,\hat{o})\Big|_{e(s,t)=0}^{e(s,t)=1},$$

quantité nécessaire à la mise à jour locale (stochastique ou déterministe) du champ d'étiquettes (plus la force définie ci-dessus sera grande, et plus le site aura des chances de voir son étiquette changée à +1, *cf.*Chap.II,§2.2.2). L'opération est répétée pour tous les sites (de façon séquentielle ou parallèle), en variant le paramètre de contrôle T, jusqu'à obtention du champ optimal (recuit simulé, *cf.*Chap.II,§2.3). En raison de l'importante charge de calcul engendrée par l'optimisation stochastique, la plupart des auteurs ont du se satisfaire de méthodes déterministes sous-optimales (ICM par exemple). Il est intéressant de noter que plusieurs d'entre eux (par exemple [Memin93]) choisissent d'estimer *a posteriori* la qualité des résultats du modèle *temps-réel* par rapport à ceux obtenus en utilisant une procédure de recuit simulé en *temps-différé*.

5.3 Description des algorithmes de Lalande et Dumontier.

Les algorithmes développés dans [Dumont96] et [Caplie95] sont des versions simplifiées de l'algorithme de détection du mouvement originellement décrit dans [Lalan90a], véritable modèle spatio-temporel trop gourmand en calcul pour être aujourd'hui correctement implanté sur des ordinateurs séquentiels conventionnels. Le temps de calcul est en réalité la *seule* raison pour laquelle ces deux auteurs ont cherché à simplifier le modèle original, (1) en réduisant de façon conséquente le nombre de variables à calculer (un seul champ d'étiquettes à relaxer au lieu de deux champs couplés,

les modèles ne sont plus spatio-temporels, mais *à contrainte temporelle*), et (2) en utilisant des techniques d'optimisation extrêmement rapides mais loin d'être optimales (méthodes déterministes de type ICM). Les performances se trouvent bien sûr compromises, mais pas au point de ne pas justifier une réalisation expérimentale (carte DSP [Dumont96], simulation d'un réseau VLSI [Caplie95]).

Notre processeur optoélectonique est dédié à l'optimisation stochastique en temps-réel. La rapidité de calcul est telle que l'on pourrait en principe implanter le modèle complet de Lalande sans compromettre ni les résultats, ni même la contrainte de *cadence vidéo*. Cependant, le nombre de variables et le caractère continu de certaines d'entre elles rendraient la tâche beaucoup trop ambitieuse (système optoélectronique compliqué mais pas irréalisable) par rapport à notre objectif, qui - rappelons-le - est de *démontrer* l'intérêt de tels processeurs dédiés aux tâches de traitement d'images bas niveau. C'est pourquoi, l'obtention d'un algorithme de détection du mouvement extrêmement simple devenait une nécessité (un seul champ à relaxer à la fois, priorité sur les modèles à variables *binaires* pour éviter les problèmes d'uniformité d'éclairement, etc.). Nous sommes donc parti du modèle à contrainte temporelle de Dumontier-Caplier, que nous avons cherché à étudier pour en dégager les caractéristiques principales. Notre modèle modifié est extrêmement simple, mais il reprend en fait certaines heuristiques présentes dans l'algorithme original de Lalande et négligés dans l'algorithme de Dumontier-Caplier. Le nouveau modèle proposé semble se comporter au moins aussi honorablement que celui de Dumontier-Caplier, au moins pour les quelques cas de figure que nous avons pu étudier (*cf* Ann.C). Nous nous proposons par la suite de discuter avec plus ou moins de détail des modèles de Lalande et Dumontier-Caplier pour en extraire les idées fortes.

5.4 Un modèle complet mais gourmand en calcul : le modèle de Lalande.

On présentera par la suite le modèle de Lalande [Lalan90a], qui peut être considéré comme le point de départ de notre travail. Il s'agit d'un véritable modèle spatio-temporel quoique de faible "épaisseur temporelle".

5.4.1 Données et modèle d'observation.

Le modèle de Lalande fait appel au champ d'observations $O(t)$ continu et aussi au champ d'observations binaire $\hat{O}(t)$. Dans ce modèle où seul est considéré simultanément le champ de Markov spatio-temporel formé par le *couple glissant* $\{E(t), E(t+1)\}$, la relation entre l'observation $O(s,t)$ (à l'instant t et au site s) et la paire d'étiquettes (aux instants t et $t-1$) s'écrit simplement :

$$O(s,t) = \Psi_L\left(E(s,t), E(s,t)\right) + n\,,$$

(*n* est une variable gaussienne centrée de variance σ^2). Le potentiel Ψ_L est défini par le tableau *tab.*IV.1.

$e_t(s)$	$e_{t-1}(s)$	$\Psi_L \{ e_t(s), e_{t-1}(s) \}$
0	0	0
0	1	m_2
1	0	m_2
1	1	m_1

Tab.IV.1 : Définition du potentiel d'attache aux données pour le modèle de Lalande. On a $0 < m_1 < m_2$.

Le choix de la fonction Ψ_L s'interprète de la façon suivante : si le site était fixe à l'instant *t-1* et reste fixe à l'instant *t* (transition fond-fond), alors il est vraisemblable que *l'observation* non bruitée - la différence inter-image *o(s,t)* - soit *nulle* dans ce site. Lors d'une transition objet-objet (zone de glissement) l'observation est proche d'une valeur m_1 qui dépend des caractéristiques de l'objet, mais qui reste *faible* devant la variation de luminance m_2 entraîné par une transition fond-objet (zone de recouvrement) ou objet-fond (zone d'écho). C'est la traduction formelle des hypothèses faites au §3.2.

La fonction d'énergie associé au processus d'observation (*cf.*Chap.I,§2.3) s'écrit alors (les indices permettent de simplifie l'écriture pour la variable temporelle) :

$$U_{obs}(e_{t-1}, e_t / o_t) = \sum_{s \in S} V_O(e_{t-1}(s), e_t(s), o_t(s)) \ ,$$

$$avec \quad V_O(e_{t-1}(s), e_t(s), o_t(s)) = \frac{1}{2\sigma^2}[o_t(s) - \psi_L(e_{t-1}(s), e_t(s))]^2$$

Commentaires. Le modèle d'observation de Lalande présuppose donc :

(a) que les variations spatiales de luminance sur *l'étendue* de l'objet mobile sont moindres que celles qui existent *entre* l'objet et le fond. Autrement dit, les objets mobiles sont supposés uniformes[48] ; cette hypothèse est plus probablement vérifiée pour des objets de petite taille ;

(b) que le *bruit* est le même pour l'objet et pour le fond. C'est le cas si le bruit provient uniquement du système d'acquisition (et non d'une quelconque variation temporelle de l'éclairement due au mouvement par exemple). Le choix de la statistique gaussienne est certainement arbitraire pour un certain nombre de cas. Remarquons aussi que si le système d'acquisition est tel que $I(t)=I_{originale}(t)+n_{gaussien}$, (hypothèse largement admise) alors le bruit sur l'observation, qui n'est autre chose que la différence inter-image en valeur absolue $O(t)=DI(t)=|I(t)-I(t-1)|$ *n'est pas* lui même gaussien.

---

[48] le modèle aura du mal par exemple à détecter correctement le mouvement - évident pour l'œil - d'un zèbre ou d'un tigre se déplaçant sur une pelouse parfaitement uniforme ; il n'est pas adapté non plus au cas ou fond et objet possèdent la même texture (une raie sur un fond marin de sable par exemple). Mais c'est en réalité une affaire de paramètres : si la résolution augmente, chaque grain de la texture sera considéré comme un objet mobile. Remarquons aussi que pour l'œil, si l'objet ne bouge pas, il passe inaperçu ; mais dès qu'il se déplace il est étonnant de constater la facilité avec laquelle nous sommes capables d'isoler la forme en mouvement.

(c) Enfin, le champ d'observation *binaire* $\hat{O}(t)$ n'intervient pas directement comme une contrainte aux données (il n'engendre pas un terme d'énergie d'attache aux données), mais contrôle les interactions temporelles du champ de Markov (voir ci-dessous).

5.4.2 Champ d'étiquettes et voisinage.

Le champ d'étiquettes est représenté à chaque instant par le couple glissant $\{E(t), E(t-1)\}$. Le voisinage et les cliques binaires associées sont représentées sur la *fig.*IV.4.

Potentiel spatial V_S. Il s'agit d'un potentiel à deux niveaux qui favorise la continuité spatiale du masque :

$$\forall (r,s) \in SxS, \quad V_S(e_t(r), e_t(s)) = \begin{cases} -\beta_S & si \quad e_t(r) = e_t(s) \\ +\beta_S & si \quad e_t(r) \neq e_t(s) \end{cases} \quad avec \quad \beta_S > 0,$$

où β_S est un paramètre constant et global (indépendant du temps et du site spatial considéré). On écrira souvent le potentiel à niveau de type V_S sous la forme équivalente :

$$V_S(e_t(s), e_t(r)) = -\beta_S.(2.e_t(s)-1)(2.e_t(r)-1)$$

Potentiel temporel V_T. La valeur du potentiel temporel est dépendante de la donnée d'observation binaire $\hat{O}(t)$. Le tableau *tab.*IV.2 définit la valeur locale du potentiel temporel.

$e_{t-1}(s)$	$e_t(s)$	$\hat{o}_t(s)$	V_T
0	0	0	$-\beta_T$
0	0	1	$+\beta_T$
1	0	0	$+\beta'_T$
1	0	1	$+\beta'_T$
0	1	0	$+\beta_T$
0	1	1	$-\beta_T$
1	1	0	$+\beta_T$
1	1	1	$-\beta_T$

Tab.IV.2 : Définition du potentiel temporel

Les configurations associées à $-\beta_T$ sont favorisées, les autres défavorisées. β'_T prend une valeur très élevée afin d'éliminer presque certainement la configuration improbable *(1,0,0)*, ou la configuration *(1,0,1)* dont la gestion doit être décalée dans le temps.

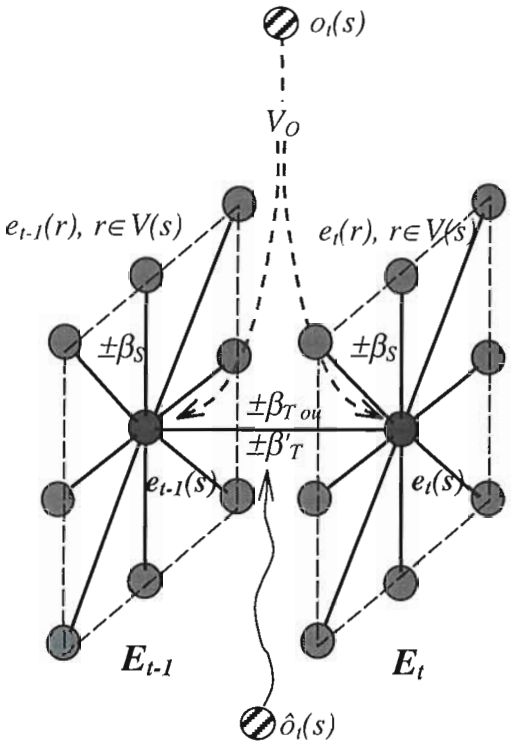

Fig.IV.4 : Voisinage spatio-temporel et cliques binaires associées dans le modèle de Lalande [Laland90].

-Le voisinage spatial est d'ordre 2 ; les potentiels de cliques sont invariants par translation.

-Le voisinage temporel est d'ordre 1. Le potentiel de la clique temporelle est localement contrôlé par la donnée binaire $\hat{o}(s,t)$.

-En pointillés est représentée la contrainte aux données observées $o(s,t)$ et la "clique" associée de potentiel V_O (à valeurs continues).

rem. : les potentiels à niveaux V_S et V_T sont représentés par leurs valeurs ($\pm\beta_S$, $\pm\beta_T$ ou β'_T).

5.4.3 Energie globale et force locale.

Pour résumer, la fonction d'énergie du modèle de Lalande se compose de deux termes :

$$U(e/o_t,\hat{o}_t) = U_{obs}(e_{t-1},e_t/o_t) + U_{reg}(e_{t-1},e_t/\hat{o}_t),$$

où l'on a :

- $U_{obs}(e_{t-1},e_t/o_t) = \frac{1}{2\sigma^2}\sum_{s\in S}[o_t(s) - \psi(e_{t-1}(s),e_t(s))]^2$ est l'énergie d'attache aux données ;
- $U_{reg}(e_{t-1},e_t/\hat{o}_t) = U_S(e_{t-1}) + U_S(e_t) + U_T(e_{t-1},e_t/\hat{o}_t)$ est l'énergie de régularisation spatio-temporelle avec :

$$\begin{cases} U_S(e_t) = \sum_{(s,r)\in C_S} V_S(e_t(s),e_t(r)) = -\frac{\beta_S}{2}\sum_{s\in S}(2.e_t(s)-1)\sum_{r\in V(s)}(2.e_t(r)-1) \text{ et} \\ U_T(e_{t-1},e_t/\hat{o}_t) = \sum_{(s,r)\in C_S} V_T(e_t(s),e_t(r),\hat{o}_t(s)). \end{cases}$$

La force locale correspondante au site s et au temps t est alors composée de trois termes :

$$F(s,t) = F_S(s,t) + F_T(s,t) + F_O(s,t),$$

avec respectivement :

- $F_S(s,t) = \sum_{r \in V(s)} \{-V_S[1, e_t(r)] + V_S[0, e_t(r)]\} = \beta_S \sum_{r \in V(s)} (2.e_t(r) - 1)$, (régularisation spatiale).
- $F_T(s,t) = -V_T(e_{t-1}(s), 1, \hat{o}_t(s)) + V_T(e_{t-1}(s), 0, \hat{o}_t(s))$, (régularisation temporelle).
- $F_O(s,t) = -V_O(e_{t-1}(s), 1, o_t(s)) + V_O(e_{t-1}(s), 0, o_t(s))$, (force de rappel à l'observation).

Commentaires. La forme du potentiel temporel permettrait une implantation électronique-logique immédiate (il s'agit de potentiels à niveaux discrets) ; en effet, on a après calcul pour la force de rappel temporelle au site *(s,t)* :

$e_{t-1}(s)$	$\hat{o}_t(s)$	$F_T(s,t)$
0	0	$\beta_T + \beta'_T$
0	1	$-2\beta_T$
1	0	$\beta_T - \beta'_T$
1	1	$-\beta_T - \beta'_T$

Par contre, même si la fonction prédéfinie Ψ_L du modèle d'attache aux données est très simple (*tab.*IV.1), la force de rappel à l'observation reste *à valeurs continues* ; on a :

$$F_O(s,t) = -\frac{1}{2\sigma^2}\left[\psi_L{}^2(e_{t-1},1) - \psi_L{}^2(e_{t-1},0) - 2.o_t(s).\{\psi_L(e_{t-1},1) - \psi_L(e_{t-1},0)\}\right],$$

soit plus explicitement :

	si e(t-1,s) =1 :	*si e(t-1,s) =0 :*
$F_o(s,t)$	$-\frac{(m_1 - m_2)}{2\sigma^2}[(m_1 + m_2) - 2.o_t(s)]$ $\approx \frac{m_2}{2\sigma^2}[m_2 - 2.o_t(s)]$	$-\frac{m_2}{2\sigma^2}[m_2 - 2.o_t(s)]$

Tab.IV.3 : force de rappel à l'observation dans le modèle de Lalande.

Le *tab.*IV.3 va nous permettre de mieux interpréter la simplification introduite plus loin dans notre modèle (où la force de rappel à l'observation est à valeurs *discrètes* ; voir aussi *fig.*IV.7).

5.4.4 Algorithme d'optimisation.

La minimisation de la fonction d'énergie est effectuée par relaxation déterministe de type ICM. A chaque instant *t*, on obtient le champ estimé *E(t-1)* et une *première* estimation du champ *E(t),* utile pour l'initialisation des champs *E(t)* et *E(t+1)* à l'instant suivant. L'optimisation par couples glissants permet de revenir à l'instant *t+1* sur la

décision prise à l'instant *t*, et ceci permet de mieux gérer les zones de glissement et/ou d'éliminer la zone d'écho. *La fig.*IV.5 donne une synoptique de l'algorithme de Lalande.

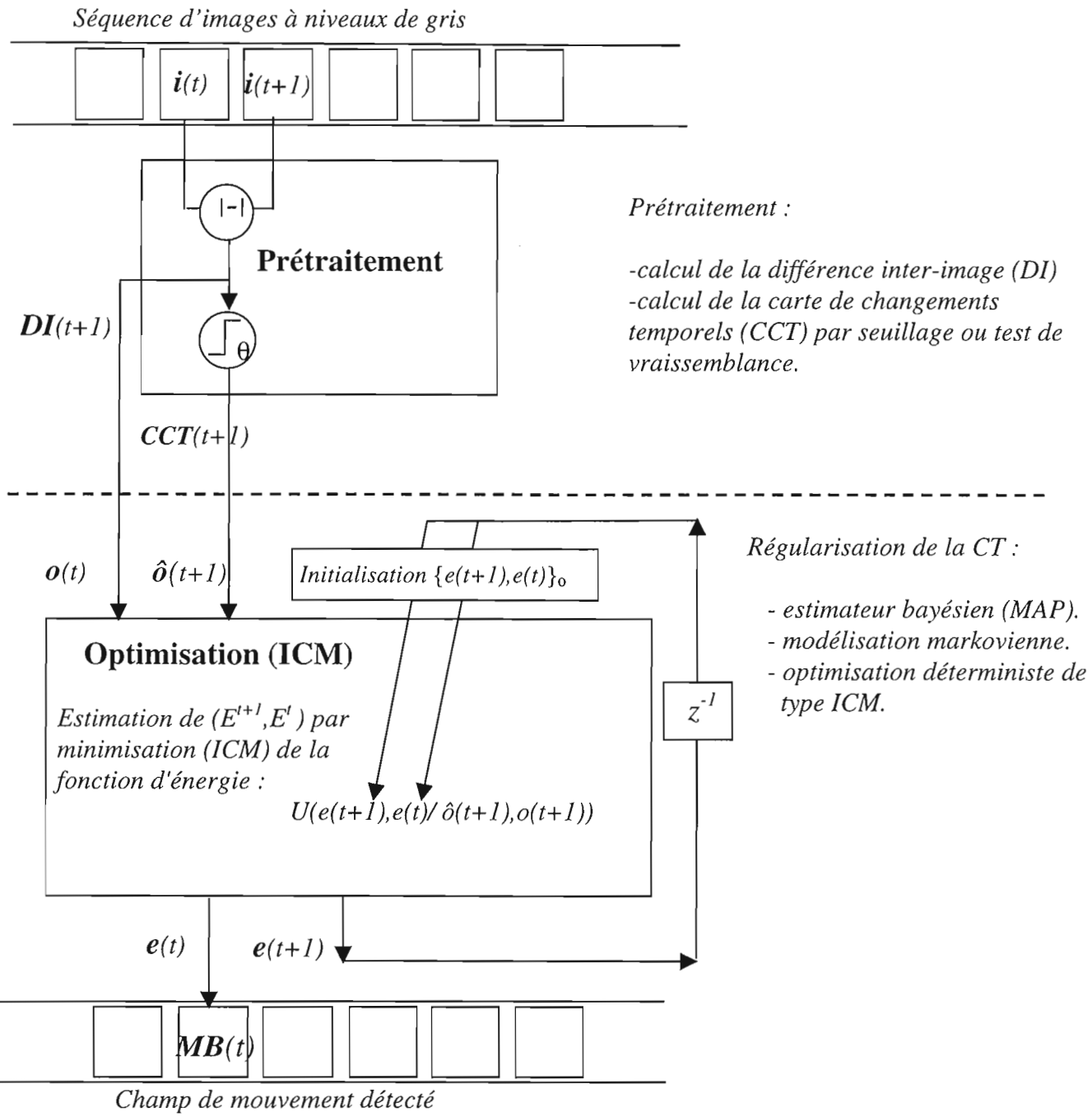

Fig.IV.5 : Schéma synoptique de l'algorithme de détection du mouvement proposé dans [Lalande90]. Le prétraitement se résume au calcul de la CCT binaire par seuillage (ou test de vraisemblance) entre deux images succésives. Le champ de mouvement est estimé par couples glissants. Le calcul est réalisé par optimisation déterministe de type ICM. L'initialisation du couple est réalisé grâce au résultat précédent de l'estimation.

5.4.5 Conclusion

Ce modèle fait apparaître des véritables cliques spatio-temporelles, mais seulement dans une tranche temporelle de maigre épaisseur. L'avantage -qui est de pouvoir un peu mieux gérer les zones de glissement- n'est pas si grand face à la complexité introduite (estimation conjointe de deux champs d'étiquettes). En ce qui nous concerne, puisque la relaxation conjointe de plusieurs champs de d'étiquettes en même temps impliquerait une architecture totalement différente pour notre processeur (plusieurs processeurs tableau en cascadé par exemple), il serait aussi facile d'implanter un modèle spatio-temporel beaucoup plus complet (par exemple comportant cinq à huit coupes temporelles [Caplier95]).

5.5 Le modèle simplifié de Dumontier.

Le modèle de Dumontier [Dumont96] tend de réduire la complexité du modèle de Lalande en ne considérant qu'un champ *E(t)* à estimer à la fois. *E(t-1)* est donc fixe pendant la durée de l'optimisation. Il s'agit d'un modèle markovien spatial *à contrainte temporelle* selon notre définition. On a changé les notations originales et re-interprété quelque peu les termes pour permettre une présentation cohérente des différents modèles.

5.5.1 Données et modèle d'observation

Les données retenues par le modèle de Dumontier sont les mêmes que celles de Lalande (*O(t)* et $\hat{O}(t)$). Toutefois, Dumontier et Caplier [Caplie95] semblent s'accorder sur le fait que le calcul de la CCT en utilisant la méthode de vraisemblance n'apporte pas d'amélioration significative quant au résultat final (robustesse de la modélisation markovienne qui traite le bruit en même temps que la régularisation de la CCT).

Le lien statistique entre observations et étiquettes à l'instant t est décrit par une fonction Ψ_D inspirée de Ψ_L, mais qui ne fait intervenir que le champ d'étiquettes actuel :

$$o_t(s) = \Psi_D(e_t(s)) + n \text{ , avec } \Psi_D(e_t(s)) = \begin{cases} 0 & si \quad e_t(s) = 0 \\ \alpha > 0 & si \quad e_t(s) = 1 \end{cases}$$

L'interprétation est cette fois-ci la suivante : si le site en question appartient à la zone fixe de l'image, il n'y a pas de changement temporel significatif de luminance. En revanche, si le site appartient à une zone en mouvement, on considère qu'il y a changement temporel, et on suppose l'observation proche d'une valeur $\alpha>0$. Cette valeur est estimée de façon heuristique dans [Caplier95]. La conclusion semble être que l'on peut fixer ce paramètre une fois pour toutes quel que soit la séquence sans qu'il y ait de dégradation du résultat de la détection.

L'énergie d'attache aux données s'écrit alors :

$$U_{obs}(e,o)=\sum_{s\in S}V_O(e_t(s),o_t(s)),$$

$$avec \quad V_O(e_t(s),o_t(s))=\frac{1}{2\sigma^2}[o_t(s)-\Psi(e_t(s))]^2$$

Commentaire : L'élimination de *E(t-1)* comme entrée de la fonction Ψ_D est à notre avis une simplification beaucoup trop grande. Il est beaucoup plus raisonnable de considérer (d'après les hypothèses faites au §3.2) qu'un changement temporel est le résultat d'une variation de l'état de mouvement du pixel (zone de recouvrement ou zone d'écho), et non du mouvement lui même (zone de glissement). Lalande considère en effet que cette dernière éventualité conduit à une observation d'amplitude m_2 *beaucoup plus petite* que celle induite par un changement d'état de mouvement ($m_2 << m_1$). L'élimination du champ *E(t-1)* comme champ à estimer ne doit pas conduire à son élimination dans le modèle d'observation, d'autant plus que comme on le verra par la suite, celui-ci est gardé en mémoire pour définir une *contrainte temporelle* sur le champ *E(t)*.

5.5.2 Champ d'étiquettes et voisinage

Comme dit plus haut, le modèle de Dumontier-Caplier ne considère - à chaque instant - qu'un champ de Markov d'étiquettes *E(t)* sans dimension temporelle. La *fig.*IV.6 représente le voisinage de Markov spatial et les contraintes aux données.

Potentiel spatial V_S. La forme du potentiel spatial est la même que pour le modèle de Lalande.

Potentiel temporel passé (V_P) et futur (V_F). Dumontier définit deux potentiels à niveaux, associés aux cliques temporelles pour le passé et pour le futur. Il s'agit d'un abus de langage, puisque les champs *E(t-1)* et *E(t+1)* ne forment pas partie du champ de Markov en cours de relaxation. En réalité, il s'agit de potentiels de *contraintes aux données* :

- On a pour la contrainte au passé :

$$V_P(e_t(s)/e_{t-1}(s))=\begin{cases}-\beta_P & si \quad e_t(s)=e_{t-1}(s)\\ +\beta_P & si \quad e_t(s)\neq e_{t-1}(s)\end{cases} \quad avec \quad \beta_P>0,$$

où *e(t-1)=MB(t-1)* est le masque binaire du mouvement estimé lors de la relaxation à l'instant *t-1*. Il est *figé* à l'instant *t*, et agit comme un champ de contrainte au passé.

- Et pour la contrainte au "futur" :

$$V_F(e_t(s)/e_{t+1}(s))=\begin{cases}-\beta_F & si \quad e_t(s)=e_{t+1}(s)\\ +\beta_F & si \quad e_t(s)\neq e_{t+1}(s)\end{cases} \quad avec \quad \beta_F>0,$$

où *e(t+1)* est estimé très grossièrement, puisqu'il est pris égal à la carte de changements temporels future $\hat{o}(t+1)=CCT(t+1)$.

Commentaire : la contrainte aux données passées semble bien traduire la contrainte de régularisation temporelle du champ spatio-temporel dans le modèle de Lalande. Le terme de "contrainte aux données futures" est quant à lui un peu plus douteux, puisque l'estimée du champ binaire futur *e(t+1)* est en réalité le champ d'*observation* binaire $\hat{O}(t+1)$: on peut se demander alors s'il est vraiment utile d'introduire *deux* contraintes de rappel aux données (V_O et V_F), et en tout cas pourquoi l'une mettrait en jeu la différence inter-image, et l'autre la différence inter-image seuillée (voir *fig.*IV.6).

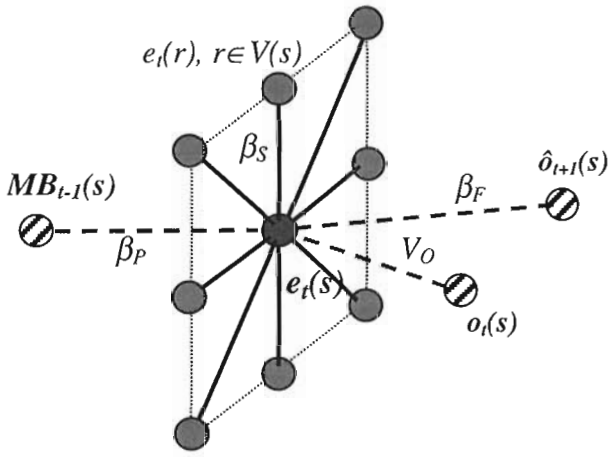

Fig.IV.6 : Cliques spatiales et contraintes aux données dans le modèle de Dumontier. Il est facile de représenter l'énergie d'attache aux données par des potentiel de cliques, mais les 'sites' achurées ne foment pas partie du champ de markov (leur valeurs sont fixes pendant la relaxation du champ e(t)). Les potentiels à deux niveaux V_S et V_F sont représentés par leurs valeurs respectives ($\pm\beta_S$, $\pm\beta_F$).

5.5.3 Energie et force de rappel locale

L'énergie associée au modèle de Dumontier s'écrit finalement :

$$U(e/o_t,\hat{o}_{t+1}) = U_{obs}(e_t/o_t) + U_{reg}(e_t/e_{t-1},\hat{o}_{t+1})$$

où : $U_{obs}(e_t/o_t) = \frac{1}{2\sigma^2}\sum_{s\in S}[o_t(s) - \psi_D(e_t(s))]^2$ est l'énergie d'attache aux données, et

$U_{reg}(e_t/e_{t-1},\hat{o}_{t+1}) = U_S(e_t) + U_P(e_t/e_{t-1}) + U_F(e_t/\hat{o}_{t+1})$ est l'énergie de "régularisation" spatiale et temporelle[49], avec :

$$\begin{cases} U_S(e_t) = \sum_{(s,r)\in C_S} V_S(e_t(s),e_t(r)) = -\frac{\beta_S}{2}\sum_{s\in S}(2.e_t(s)-1)\sum_{r\in V(s)}(2.e_t(r)-1)\ , \\ U_P(e_t/e_{t-1}) = \sum_{s\in S} V_P(e_t(s)/e_{t-1}(s)) = -\beta_P(2.e_t(s)-1).(2.e_{t-1}(s)-1) \text{ et} \\ U_F(e_t/\hat{o}_{t+1}) = \sum_{s\in S} V_F(e_t(s)/\hat{o}_{t+1}(s)) = -\beta_F\sum_{s\in S}(2.e_t(s)-1).(2.\hat{o}_{t+1}(s)-1) \end{cases}$$

[49] Comme expliqué plus haut, du point de vue de la relaxation du champ de Markov, les termes temporels doivent être considérés comme des *contraintes aux données* ; on entend ici *régularisation* temporelle dans le sens particulier signifiant que le *résultat* de la relaxation précédente doit ressembler au *résultat* de la relaxation courante.

La force locale correspondante au site s (au temps t) est alors :

$$F(s,t) = F_O(s,t) + F_P(s,t) + F_S(s,t) + F_F(s,t) \text{ , avec}$$

$$\begin{cases} F_O(s,t) = V_O(1/o_t(s)) - V_O(0/o_t(s)) = \dfrac{\alpha}{\sigma^2}\left[o_t(s) - \dfrac{\alpha}{2}\right], \\ F_P(s,t) = V_P(1/e_{t-1}(s)) - V_P(0/e_{t-1}(s)) = \beta_P(2.e_{t-1}(s) - 1), \\ F_S(s,t) = \sum_{r \in V_S(s)} [V_S(1,e_t(r)) - V_S(0,e_t(r))] = \beta_S \sum_{r \in V_S(s)} (2.e_t(r) - 1) \text{ et} \\ F_F(s,t) = V_F(1/\hat{o}_{t+1}(s)) - V_F(0/\hat{o}_{t+1}(s)) = \beta_F(2.\hat{o}_{t+1}(s) - 1). \end{cases}$$

La force locale fait apparaître quatre termes. Le seul terme provenant véritablement d'une modélisation markovienne est le terme de régularisation spatiale F_S. Les trois autres sont en réalité des termes *d'attache aux données* :

• F_P est un terme d'attache au *résultat de l'estimation précédante e(t-1)=MB(t-1)* (qui est fixe et ne peut pas être re-estimé à l'instant t comme dans le modèle de Lalande). C'est la seule véritable *contrainte temporelle*.

• Les deux autres termes F_O et F_F sont des termes d'attache aux données provenant respectivement du champ d'observation continu *O(t)* et du champ d'observation binaire *Ô(t+1)*. Il est certainement abusif de considérer F_F comme un terme de régularisation *temporelle*, puisque *Ô(t+1)* n'est qu'une estimation très grossière de *MB(t+1)*. Autrement dit, le modèle de Dumontier n'est pas simplement un modèle simplifié de Lalande : une nouvelle définition *du processus d'observation* a été implicitement mise en place, sans qu'il y ait une véritable justification théorique ; on peut se demander en particulier pourquoi prendre un terme d'attache aux données *continu* au temps *t*, et un terme d'attache *discret* à l'instant suivant.

• La force de rappel à l'observation F_O est à valeurs continues. La forme est équivalente à celle du modèle de Lalande seulement dans le cas où le site considéré était fixe à l'instant *t-1* (fond et zone de recouvrement), et l'analogie se poursuit en posant $m_2=\alpha$ (*cf.* deuxième colonne du *tab.*IV.1). La force de rappel joue alors correctement *dans le sens de d'intégration de la zone de recouvrement.* Pourtant, le comportement peut être aberrant si le site était *mobile* à l'instant précédent (zone d'écho et zone de glissement). En effet, si l'observation continue *o(s,t)* est inférieure au seuil $\alpha/2$ (ce qui est vraisemblable pour les sites appartenant à la zone de glissement dans le cas d'objets uniformes), la force de rappel à l'observation tend à *empêcher le remplissage de la zone de glissement* ; et si l'observation dépasse ce seuil (ce qui est certainement le cas pour la zone d'écho), alors la force tend à maintenir l'état de mouvement courant, *rendant difficile l'élimination de la zone d'écho* (*voir fig.*IV.7). En d'autres termes, la force de

rappel à l'observation définie selon le modèle de Dumontier n'est pas adapté aux hypothèses faites au §3.2, mais correspond plutôt au cas d'objets mobiles texturés, bruités et/ou sur lesquels l'éclairement est fortement variable (cas de surfaces changeant d'orientation par rapport à la source lumineuse, par exemple) ; lesdits objets se déplaçant sur un fond constant ayant éventuellement le *même* niveau de gris moyen que l'objet mobile. C'est une situation intéressante, mais qui n'est pas compatible avec celle pour laquelle ont été conçues les autres termes de la force locale.

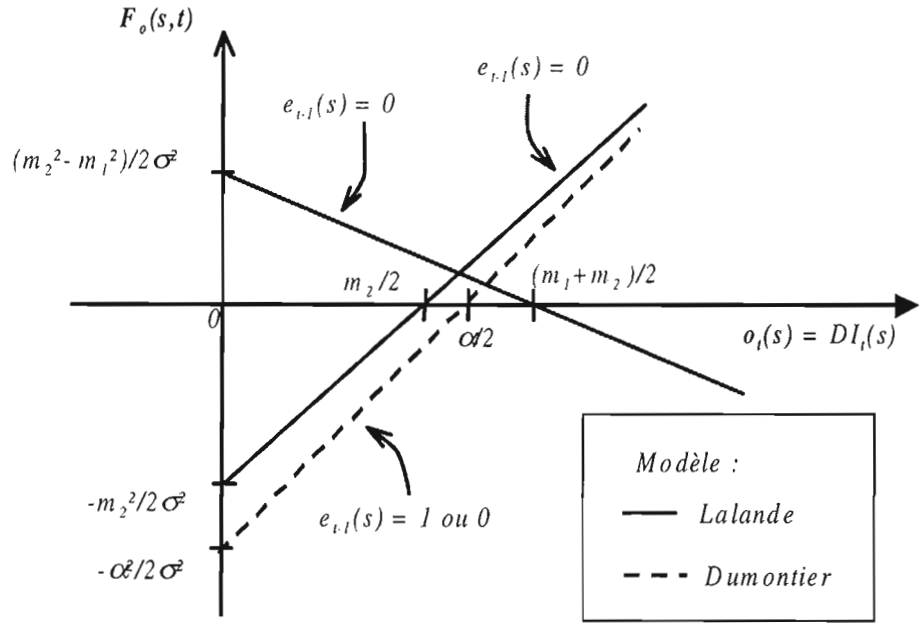

Fig.IV.7 : Force de rappel à l'observation dans le modèle de Lalande et de Dumontier. La force "tire" vers l'état de mouvement 1 si elle est positive, et vers l'état immobile si elle est négative. Le modèle de Dumontier ne distingue pas les cas $e_{t-1}(s)=1$ et 0.

5.5.4 Optimisation

La méthode choisie par Dumontier est l'ICM. Le balayage de l'image est séquentiel, ligne par ligne, et la mise à jour des sites est dite site-recursive (dès qu'un site est étudié, la modification éventuelle de son étiquette est immédiatement prise en compte). La convergence est supposée atteinte lorsque la variation relative de l'énergie d'une itération à l'autre est inférieure à une valeur prédéfinie. Le nombre d'itérations avant convergence varie en fonction de l'amplitude du mouvement entre deux images successives, et de la qualité des champs initiaux. Néanmoins, une étude plus approfondie [Caplier95] montre que dans le cas de séquences d'images pour lesquelles les déplacements restent modérés par rapport à la cadence d'acquisition, il est possible d'arrêter la relaxation après un nombre d'itérations fixé une fois pour toutes sans dégradation sensible des résultats. Quatre itérations sur l'ensemble de l'image semblent alors suffisantes. La *fig.*IV.8 donne une synoptique de l'algorithme de Dumontier pour la détection du mouvement.

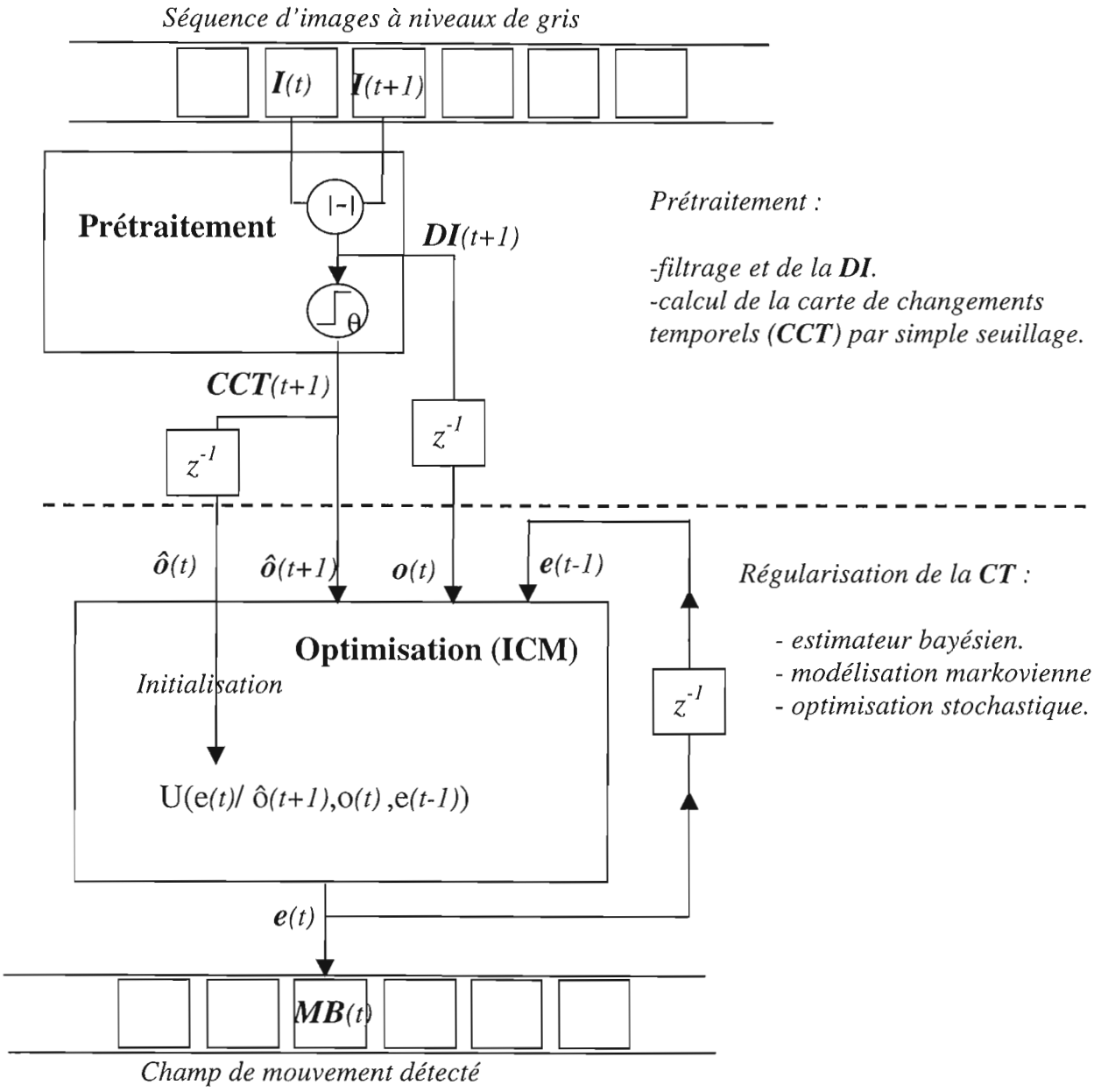

Fig.IV.8 : Schéma synoptique de l'algorithme de détection du mouvement proposé dans [Dumont96]. Le prétraitement se résume toujours au calcul de la CCT binaire par seuillage (avec ou sans pré-filtrage passe bas). Le calcul de l'estimé du champ du mouvement MB(t) est réalisé par optimisation déterministe de type ICM ; celle-ci est très sensible à la valeur initiale du champ e(t) : dans le modèle de Dumontier, celui-ci est initialisé comme ô(t), c'est à dire avec la carte binaire des changement temporels.

5.6 Estimation des paramètres.

Les deux algorithmes décrits nécessitent le choix de deux sortes de paramètres : (a) ceux correspondant exclusivement au modèle de détection du mouvement, et (b) les paramètres de contrôle de la procédure de relaxation (critère d'arrêt). Ainsi, le modèle de Lalande nécessite le choix de 7 paramètres pour le modèle correspondant à la :

1. force de régularisation temporelle : β_T, β'_T et choix du seuil de binarisation θ.
2. force de régularisation spatiale : β_S
3. force de rappel aux données : m_1,m_2 et l'estimation de σ (variance du bruit).
4. Le modèle de Dumontier quant à lui nécessite le choix de 6 paramètres :
5. force de *rappel* temporelle : β_P
6. force de régularisation spatiale : β_S
7. force de rappel aux données observées : α et σ (variance du bruit) pour F_O et seuil de binarisation θ et β_F pour F_F.

En ce qui concerne les paramètres du modèle, Dumontier fait un choix heuristique qui semble convenir pour toutes les séquences en étude, à savoir : β_P= *10,* β_S= *20,* β_F= *30, et* α = *20*. La variance du bruit de l'observation est estimée périodiquement (à partir de deux images consécutives), et le seuil de binarisation est choisi de façon expérimentale en fonction de la séquence analysée (typiquement entre 25 et 45 pour des images à 256 niveaux de gris et pour des scènes de rue en lumière ambiante).

Enfin, le critère d'arrêt est déterminé -aussi bien dans le modèle de Dumontier que celui de Lalande- de façon heuristique (4-5 itérations de la procédure *ICM*).

5.7 Conclusion

Comme expliqué au §4.2.2, pour régulariser la *CCT* il faut (a) remplir la zone de glissement, (b) éliminer la zone d'écho, (c) intégrer la zone de recouvrement et (d) éliminer le bruit -il n'y a pas d'ordre chronologique entre les opérations. On peut repérer ces différentes fonctions dans les termes de la force locale de Lalande (§5.4.3) et Dumontier (§5.5.4):

(a) remplissage de la zone de glissement :

- Modèle de Lalande : c'est le rôle de F_S + F_T (modèle markovien spatio-temporel).
- Modèle de Dumontier : le remplissage est réalisé grâce à F_O à condition que l'objet présente du bruit et/ou une texture fine (autrement, la force de rappel à l'observation agit de manière à *éliminer* la zone de glissement!). Ensuite F_S permet le remplissage par homogénéisation grâce au modèle *markovien spatial* et F_P agit

de façon "indirecte" par *rappel* temporel, une fois le processus de remplissage "amorcé" par F_O et F_S.

(b) élimination de la zone d'écho (transition *objet-fond* : $e_{t-1}(s) = 1$ et $o_t(s)$ grand) :

- Modèle de Lalande: $F_O(s,t)$ (voir *fig.*IV.7).
- Modèle de Dumontier : la valeur de l'observation $o(s,t)$ est certainement importante dans la zone d'écho, ce qui fait que $F_O(s,t)$ fait la promotion (de façon erronée) de l'étiquette *mobile* pour la zone d'écho ; de même, le terme de rappel temporel $F_P(s,t)$, contribue à maintenir l'étiquette erronée. L'élimination de la zone d'écho ne s'appuie donc que sur le terme de rappel au "futur" $F_F(s,t)$ qui est négatif (le champ 'futur' n'est autre chose que la *CCT* future, en principe *nulle* dans la zone d'écho courante). C'est probablement la raison pour laquelle il faut que $\beta_F > \beta_P$.

(c) intégration de la zone de recouvrement (transition *fond-objet* : $e_{t-1}(s) = 0$ et $o_t(s)$ grand) :

- Modèle de Lalande : $F_O(s,t)$ (voir *fig.*IV.7).
- Modèle de Dumontier : $F_O(s,t)$, qui se comporte exactement comme dans le modèle de Lalande (voir *Fig.IV.g*). Par contre, $F_F(s,t)$ n'aide à l'intégration de la zone de recouvrement qu'à condition que la CCT y soit *non nulle,* mais on peut voir que cela se produit seulement si l'objet au temps $t+1$ à fait un "saut" et n'a aucun recouvrement avec l'objet à l'instant t ; enfin, $F_P(s,t)$ contribue *négativement* (c'est son rôle *conservateur*).

(d) élimination du bruit :

- Modèle de Lalande : F_S et F_T (modèle markovien spatio-temporel).
- Modèle de Dumontier : F_S (modèle markovien spatial) et F_P de façon indirecte.

Une certaine *redondance* dans la fonctionnalité des termes est inévitable -la régularisation markovienne traite le bruit en même temps qu'elle complète la zone de glissement-, mais les tendances *antagonistes* devraient idéalement être réduites à celles (aussi inévitables) qui se présentent entre les termes de *rappel* et ceux de *régularisation.* Comme on vient de le voir, ce n'est pas le cas pour plusieurs termes de la force locale dans le modèle de Dumontier ; cependant, grâce à un choix judicieux (mais heuristique) des paramètres, le modèle donne finalement des résultats satisfaisants.

Remarquons enfin que tant dans l'algorithme de Lalande comme dans celui de Dumontier, l'initialisation du champ d'étiquettes (nécessaire au bon fonctionnement de la procédure d'optimisation déterministe) avec la CCT contribue de façon positive pour les opérations (c) et (d), mais de façon négative pour (a) et (b).

6.Modèle proposé pour la détection du mouvement

6.1 Introduction

Nous allons présenter maintenant un nouvel algorithme de détection de mouvement simplifié, inspiré toujours de celui de Lalande, mais dont :

(a) **l'origine de chacun des termes de la force locale est clairement justifiée** par la nécessité de réaliser une au moins des opérations de reconstruction de la CCT énumérés au §4.2.2. Ceci permet en particulier une bonne interprétation physique des paramètres et éventuellement une estimation automatique, non heuristique. Le modèle de Lalande présente une grande clarté dans ce sens sauf pour ce qui concerne le modèle d'observation (voir deuxième commentaire au §5.4.1) ; c'est loin d'être le cas pour le modèle de Dumontier-Caplier (voir commentaire au §5.7). La justification passe par expliciter clairement les hypothèses du modèle ; nous allons nous intéresser aux séquences vérifiant les hypothèses faites au §3.2 (la force de rappel à l'observation du modèle de Dumontier est inadaptée).

(b) **toutes les opérations de reconstruction de la CCT sont représentées** par au moins un terme de la force locale.

(c) **les termes de la force locale ne sont pas en contradiction fonctionnelle** (sauf entre les termes de *rappel* aux données et ceux de *régularisation*, ce qui est naturel étant donné qu'il s'agit d'un modèle de *régularisation contraint)*. Cette "contradiction" est présente dans le modèle de Dumontier-Caplier (comportement antagoniste de F_O, F_F et F_P à l'heure de traiter la zone d'écho).

(d) enfin, et *seulement* pour des raisons d'implantation matérielle (car, a priori, cela appauvrit le modèle) on cherchera aussi à **discretiser les valeurs de la force de rappel à l'observation.**

6.2 Description du nouveau modèle

6.2.1 Données et modèle d'observation.

L'analyse faite au paragraphe précédent conduit tout d'abord à l'élimination de $\hat{o}(t+1)$ comme donnée du modèle (la $CCT(t+1)$ n'est certainement pas une bonne approximation de $MB(t+1)$). A présent on ne considère que l'observation $o(t)$ comme donnée pertinente à l'instant t. Cependant, et dans le seul but de simplifier le codage optique dans le démonstrateur, nous allons prendre la version seuillé du champ $o(t)$, c.à.d. le champ *d'observation binaire* $\hat{o}(t)$ (remarquons toutefois que $\hat{o}_t$ ne représente en aucun cas une estimée du masque des objets mobiles, ni futur, ni passé). On a étoffé ensuite le modèle d'observation qui présentait l'inconvénient de ne pas distinguer les transitions *objet-fond* et *fond-objet* (ce qui induisait au comportement aberrant de la force de rappel à

l'observation dans le modèle de Dumontier). On reprend ainsi la définition -plus sensée- du lien statistique entre observations et étiquettes du modèle de Lalande, à ceci près que l'on choisit le champ d'observation binaire $\hat{o}(t)$ comme donnée du modèle, dégradé par un *bruit de canal N* :

$$\hat{o}_t(s) = \begin{cases} \Psi_A(e_{t-1}(s), e_t(s)), & si\ N = 1 \\ 1 - \Psi_A(e_{t-1}(s), e_t(s)), & si\ N = 0 \end{cases}$$

La fonction Ψ_L est remplacée par Ψ_A binaire, obéissant au tableau *tab.*IV.4 ci-dessous (il s'agit d'une version simplifié du tableau *tab.*IV.1, où m_1=0 et m_2=1) :

$e_t(s)$	$e_{t-1}(s)$	Ψ_A
0	0	0
0	1	1
1	0	1
1	1	0

Tab.IV.4 : Définition du potentiel d'attache aux données pour notre modèle. Le seul paramètre à ajuster est le seuil de binarisation θ, ce qui peut être fait automatiquement à partir des statistiques de bruit de l'image (voir paragraphe suivant).

L'interprétation de la fonction Ψ_A est très simple et sensée : en absence de bruit, $\hat{o}(s,t)=\Psi_A[e_{t-1}(s),e_t(s)]$ et l'observation binaire n'est pas nulle *que s'il y a eu un changement d'état de mouvement.* C'est la situation correspondant à la zone d'écho et à la zone de recouvrement ; compte tenu des hypothèses faites au §3.2, il est vraisemblable que pour un choix approprié du seuil de binarisation, on ait en effet $\hat{o}(s,t)=CCT(s,t)$.

Modèle de bruit.

Le modèle de Lalande considère un vecteur de bruit gaussien s'ajoutant à l'observation continue $O(t)$. Le choix - inspiré de modèles courants en restauration d'images - n'est pas très bien justifié dans le cas de la détection du mouvement. Dans le nouveau modèle proposé il est possible de trouver une interprétation exacte pour le vecteur de bruit : celui-ci n'est plus un bruit gaussien centré, *mais un bruit de canal de taux d'erreur ε.*

Supposons que l'image $I(t)$ issue du capteur est elle même brouillée par un bruit gaussien centré de variance σ^2 (hypothèse facilement vérifiée dans un grand nombre de cas), et que le niveau de gris est m_o pour l'objet et m_f pour le fond[50]. Calculons la probabilité pour que l'évaluation de la *CCT* avec une seuil de binarisation θ conduise à une décision erronée :

[50] En toute rigueur le fond et l'objet mobile peuvent être bruités (bruit constant ou variable dans le temps) et/ou texturés. Nous n'allons pas modéliser ces phénomènes ; nous allons nous restreindre aux cas pour lesquels la dynamique des variations spatiales de niveaux de gris - pour le fond et pour l'objet - reste petite par rapport au bruit de capture.

- Pour une transition *fond-fond* ou *objet-objet* ($\Psi_A = 0$), l'observation $\hat{o}(s,t)=CCT(s,t)$ devrait être *nulle*. La probabilité d'erreur vaut (voir *Fig.IV.h*) :

$$\text{(a)} \quad \Pr[\hat{o}_t(s) = 1/\Psi_A = 0] = \frac{2}{\sqrt{2\pi\sigma_y^2}} \int_{\theta}^{+\infty} e^{-\frac{x^2}{2\sigma_y^2}} dx,$$

avec $\sigma_y^2 = 2.\sigma^2$, correspondant à la variance de la différence $y=I(t,s)-I(t-1,s)$.

- Pour une transition *objet-fond* ou *fond-objet* ($\Psi_A = 1$), l'observation $\hat{o}(s,t)=CCT(s,t)$ devrait être *non nulle*. La probabilité qu'il en soit autrement vaut (voir *fig.*IV.9) :

$$\text{(b)} \quad \Pr(\hat{o}_t(s) = 0 / \Psi_A = 1) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \int_{-\theta}^{+\theta} e^{-\frac{(x-M)^2}{2\sigma_y^2}} dx,$$

avec $M=|m_o-m_f|$.

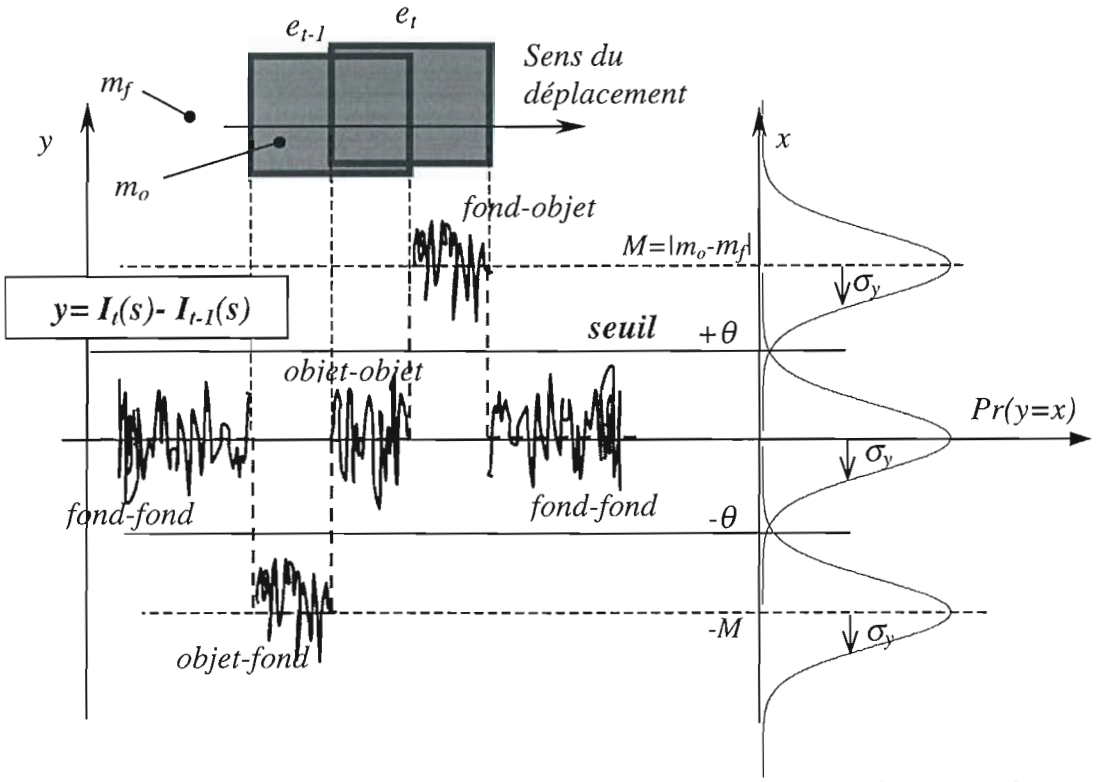

Fig.IV.9 : Interprétation de la relation entre le bruit gaussien de l'image et le seuil de binarisation θ du modèle.

Pour pouvoir utiliser le formalisme du modèle d'observation avec bruit de canal, il faut choisir le seuil de binarisation θ tel que les deux probabilités (a) et (b) soient *égales*[51], ce qui s'écrit:

[51] si la variance du bruit est faible, alors une bonne approximation pour le seuil est évidemment $\theta=M/2$.

$$(a)=(b) \Leftrightarrow \; erfc_{\sigma_y}(\theta)=\frac{1}{2}\left\{erfc_{\sigma_y}(M-\theta)-erfc_{\sigma_y}(M+\theta)\right\},$$

où *erfc* est la fonction d'erreur définie par :

$$erfc_{\sigma_y}(x)=\frac{1}{\sqrt{2\pi}\sigma_y}\int_{-x}^{x} e^{-\frac{t^2}{2\sigma_y^2}}dt = erfc_1\left(\frac{x}{\sqrt{2}\sigma_y}\right).$$

Si l'on respecte ce choix pour le seuil de binarisation, alors on peut définir le taux d'erreur ε d'un bruit de canal N entre l'observation binaire et le champ d'étiquettes selon :

$$\Pr(N=1)=\frac{\varepsilon}{2}=\Pr\left(\hat{o}_t(s)=1/\Psi_A=0\right)=\Pr\left(\hat{o}_t(s)=0/\Psi_A=1\right), \text{ et bien sûr :}$$

$$\Pr(N=0)=\frac{1-\varepsilon}{2}=\Pr\left(\hat{o}_t(s)=0/\Psi_A=0\right)=\Pr\left(\hat{o}_t(s)=1/\Psi_A=1\right).$$

Le taux d'erreur ε est fonction de la variance du bruit gaussien et du seuil de binarisation :

$$\varepsilon = erfc_{\sqrt{2}\sigma}(\theta)=\frac{2}{\sqrt{\pi}\sigma}\int_{\theta}^{+\infty} e^{-\frac{t^2}{\sigma^2}}dt$$

La *fig.*IV.10 montre le résultat de la résolution numérique pour θ de l'équation (a)=(b), et le taux d'erreur de canal correspondant.

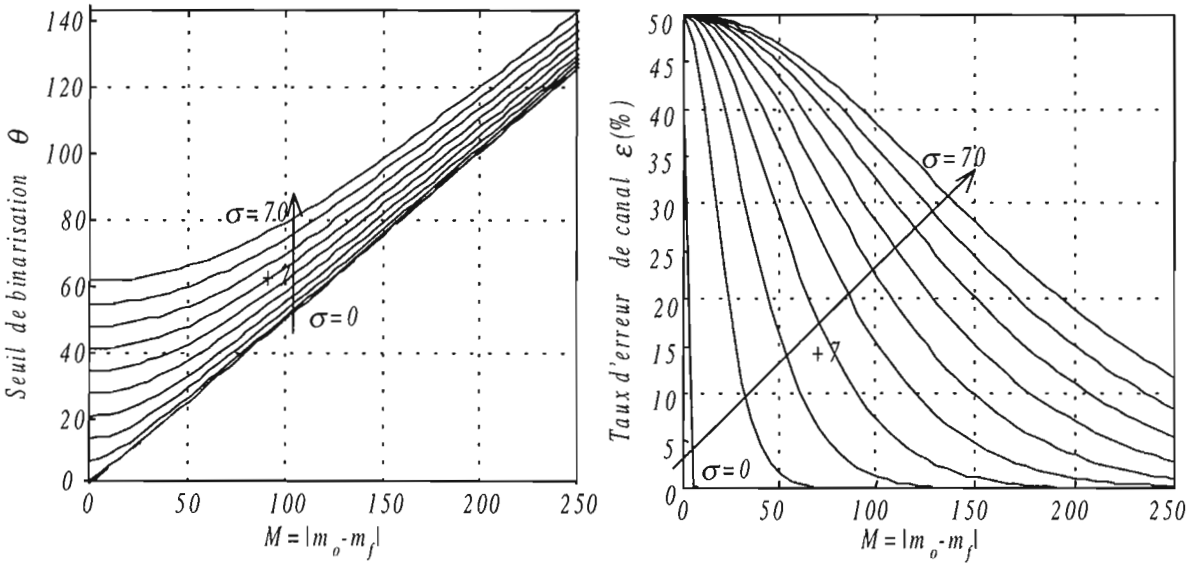

Fig.IV.10 : A gauche, seuil de binarisation optimal en fonction de la différence entre les niveaux de gris du fond et de l'objet, pour plusieurs variances du bruit de capture. A droite, taux d'erreur de canal correspondant pour le modèle d'observation.

Energie de rappel aux données. En notant $\beta_C = \ln\left(\frac{1-\varepsilon}{\varepsilon}\right)$, on peut exprimer la statistique du bruit de canal de façon condensée selon (*cf.* exemple Chap.I,*§6.1*) :

$$\Pr(N=n)=\frac{\varepsilon}{2}\exp\left(\beta_C.\delta(n)\right), \text{ avec } \delta(0)=1 \text{ et } 0 \text{ sinon.}$$

Le terme de vraisemblance marginal s'exprime alors par :

$$\Pr[\hat{O}_t(s)=\hat{o}_t(s)/E_t(s)=e_t(s),E_{t-1}(s)=e_{t-1}(s)]=\frac{\varepsilon}{2}\exp[\beta_C.\delta(\hat{o}_t(s)-\Psi_A(e_t(s),e_{t-1}(s)))],$$

d'où la vraisemblance de l'observation binaire $\hat{O}_t$ (si toutefois on considère les bruits de canal indépendants entre les pixel de l'image) :

$$\Pr(\hat{O}_t=\hat{o}_t/E_t=e_t,E_{t-1}=e_{t-1})=\left(\frac{\varepsilon}{2}\right)^{nxm}\exp\left[\beta_C\sum_{s\in S}\delta(\hat{o}_t(s)-\Psi_A(e_t(s),e_{t-1}(s)))\right]$$

(n x m est le nombre de pixels de l'image). L'énergie d'attache aux données est finalement :

$U_{obs}(e_t/\hat{o}_t,e_{t-1})=\sum_{s\in S}V_c[\hat{o}_t(s),e_t(s),e_{t-1}(s)]$ où le potentiel des "cliques" V_C vaut :

$V_C(e_t(s),\hat{o}_t(s),e_{t-1}(s))=-\beta_C\delta[\hat{o}_t(s)-\psi_A(e_t(s),e_{t-1}(s))].$

Soit, à une constante additive près, et en utilisant la définition de Ψ_A :

$$U_{obs}(e_t/\hat{o}_t,e_{t-1})=-\frac{\beta_C}{2}\sum_{s\in S}(2.e_{t-1}(s)-1).(2.\hat{o}_t(s)-1)$$

On remarquera que e_{t-1} et $\hat{o}_t$ sont des champs de contrainte, *fixes* pendant l'optimisation du champ d'étiquettes courant e_t. La force de contrainte ou rappel aux données locales est calculés selon :

$$F_c(s,t)=-\{V_c[e_t(s)=1,\hat{o}_t(s),e_{t-1}(s)]-V_C[e_t(s)=0,\hat{o}_t(s),e_{t-1}(s)]\},$$

soit : $F_c(s,t)=-\beta_c(2e_{t-1}(s)-1).(2\hat{o}_t(s)-1)$

On retrouve bien l'esprit de la définition de la fonction Ψ_L de Lalande, à savoir que l'on *préfère changer l'état de mouvement seulement quand un changement important est détecté* ; autrement, on préfère garder l'état de mouvement passé (effet mémoire).

Les séquences étudiées par Dumontier et Caplier, correspondent à M≈220-180=40 et σ≈10 ; notre modèle conduit alors à θ=21,5. Par *analogie* avec la forme de la force de rappel locale (voir *fig.*IV.11)[52], on devrait avoir $\theta=\alpha/2$, soit α=43. Caplier [Caplier95] préconise α =*20* pour toutes les séquences, tout en remarquant que la valeur du paramètre n'affecte pas beaucoup le résultat final de la détection.

[52] l'analogie ne concerne pas la *valeur* de la force, seulement est comparé le seuil qui dicte la tendance - positive ou négative - pour fixer à +1 l'étiquette à l'instant suivant.

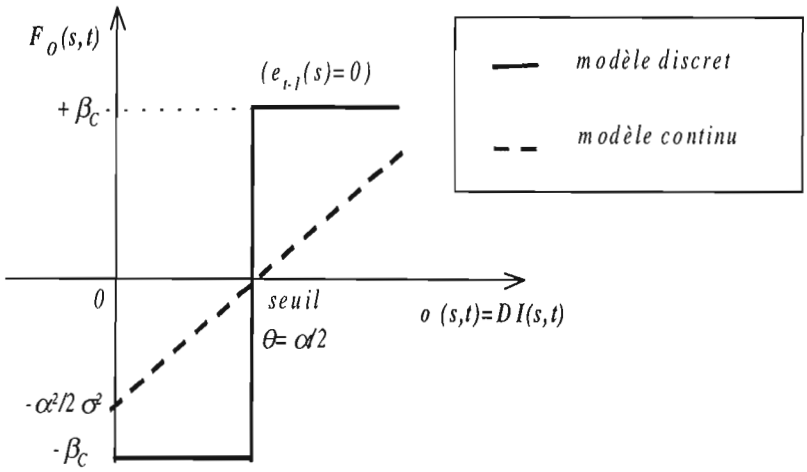

Fig.IV.11 : Force de rappel à l'observation discrète (dans le cas $e_{t-1}=0$) et correspondance entre le seuil de binarisation θ du nouveau modèle et le paramètre α du modèle d'observation continu de Dumontier (voir aussi fig.IV.7).

6.2.2 Champ d'étiquettes et voisinage

La *fig.*IV.12 représente le système de voisinage choisi pour notre modèle. Le voisinage spatial peut être d'ordre un ou deux. Les "cliques futures" du modèle de Dumontier reliant l'observation discrète *ô(s,t+1)* et le champ d'étiquettes courant *e(s,t)* ont été éliminées. Il reste que l'on peut choisir de garder ou non la contrainte temporelle vers le passée car la définition du potentiel Ψ_A réalise déjà une fonction mémoire. Cependant celle-ci est *conditionnelle* : le lien vers le passé est coupé dès que *o(s,t)* dépasse le seuil θ. C'est ainsi que deux modèles *différent*s sont envisageables, avec ou sans terme de *rémanence*.

(A) Modèle avec terme de rémanence

Dans ce modèle, la liaison vers le passé est inconditionnelle (elle ne dépend pas de l'observation). Le terme F_P provenait dans le modèle de Lalande de la *régularisation* temporelle du champ ; il n'est ici qu'une *contrainte* temporelle utile pour limiter le bruit. La force locale s'écrit :

$$F(s,t) = F_S(s,t) + F_P(s,t) + F_C(s,t)\text{, soit :}$$

$$F(s,t) = \beta_S \sum_{r \in V(s)} (2.e_t(r) - 1) + \beta_P.(2.e_{t-1}(s) - 1) - \beta_C (2.e_{t-1}(s) - 1).(2.\hat{o}_t(s) - 1)$$

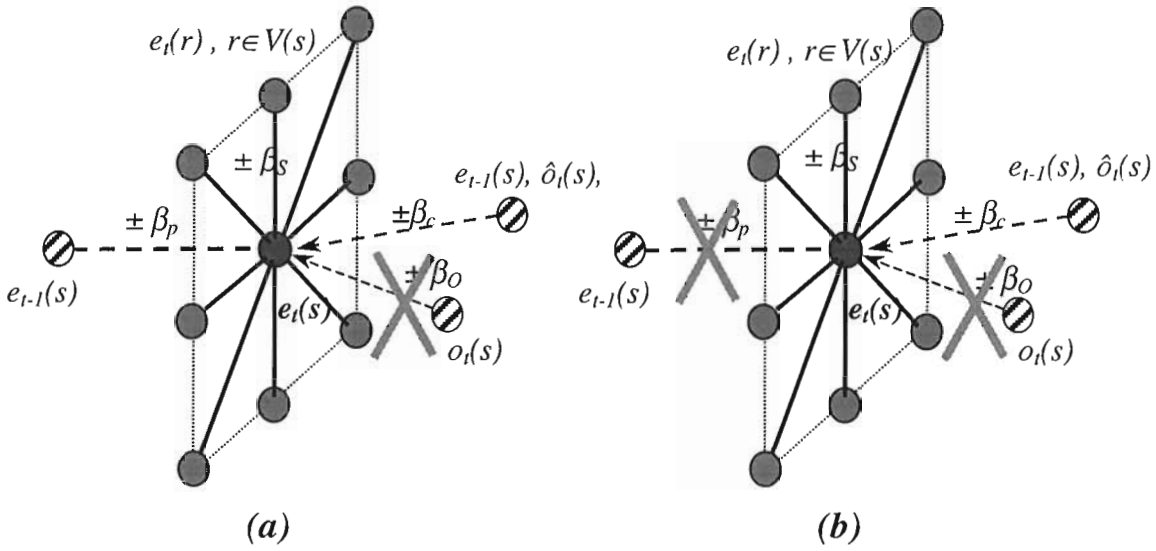

Fig.IV.12 : cliques spatiales et champs de contrainte pour le modèle avec rémanence (a) et sans rémanance (b) (comparer avec le système de voisinage du modèle de Dumontier de la fig.IV.6).

(B) Modèle sans terme de rémanence.

Dans ce modèle on considère suffisante la régularisation *spatiale* pour l'élimination du bruit. Ainsi, aucun terme de la force locale n'opère de façon antagoniste. La force locale s'écrit tout simplement :

$$F(s,t) = F_S(s,t) + F_O(s,t), \text{ soit:}$$

$$F(s,t) = \beta_S \sum_{r \in V(s)} (2.e_t(r) - 1) - \beta_C (2.e_{t-1}(s) - 1)(2.\hat{o}_t(s) - 1)$$

Commentaires. L'idée "si un pixel était en mouvement, on préfère qu'il reste en mouvement, et s'il était fixe, qu'il reste fixe" (effet mémoire), ne s'appliquera *que* pour les pixels pour lesquels il n'y a *pas eu de changement temporel significatif*. C'est désormais le modèle d'observation, à lui seul, qui prend en compte le bruit : s'il y a un *faux* changement temporel (dû au bruit), il sera sans doute local, et la régularisation spatiale se chargera de le corriger (voir illustration du principe dans la *fig.*IV.13 ci-dessous).

Le modèle (B) a toutes les bonnes propriétés annoncés en introduction (§6.1). Tout d'abord, l'origine des termes est clair, et les paramètres (θ, β_O, et β_S) peuvent être déduits, sinon de façon automatique[53], au moins de façon systématique et rationnelle à partir des

[53] Le paramètre β_S caractérise le champ de Markov d'étiquettes final, c.à.d. le *MB* (masque binaire des objets mobiles). L'ajustement de ce paramètre dépend donc des caractéristiques des objets en mouvement dans la scène ; dans les cas simples étudiés (*cf.*Ann.C), le *MB* idéal est constitué à chaque instant d'une

caractéristiques de la séquence d'images. On remarquera aussi que le modèle n'omet aucune des taches nécessaires à la régularisation de la CCT (§4.2.2), et cela sans qu'il y ait de contradiction fonctionnelle entre les termes de la force locale :

- Opérations (a) et (c) : l'opération de remplissage de la zone de glissement et d'élimination du bruit sont réalisées grâce au terme de régularisation spatiale F_S ;
- Opérations (b) et (c) : l'élimination de la zone d'écho et l'intégration de la zone de recouvrement se font grâce au terme de rappel conditionnel F_C.

Enfin, les valeurs de la force locale sont *discrètes*, ce qui permettra de simplifier la conception d'un prototype optoélectronique utilisant des modulateur spatiaux de lumière *binaires* (voir chapitre suivant). C'est donc le modèle retenu pour le démonstrateur optoélectronique qui sera exposé au chapitre suivant.

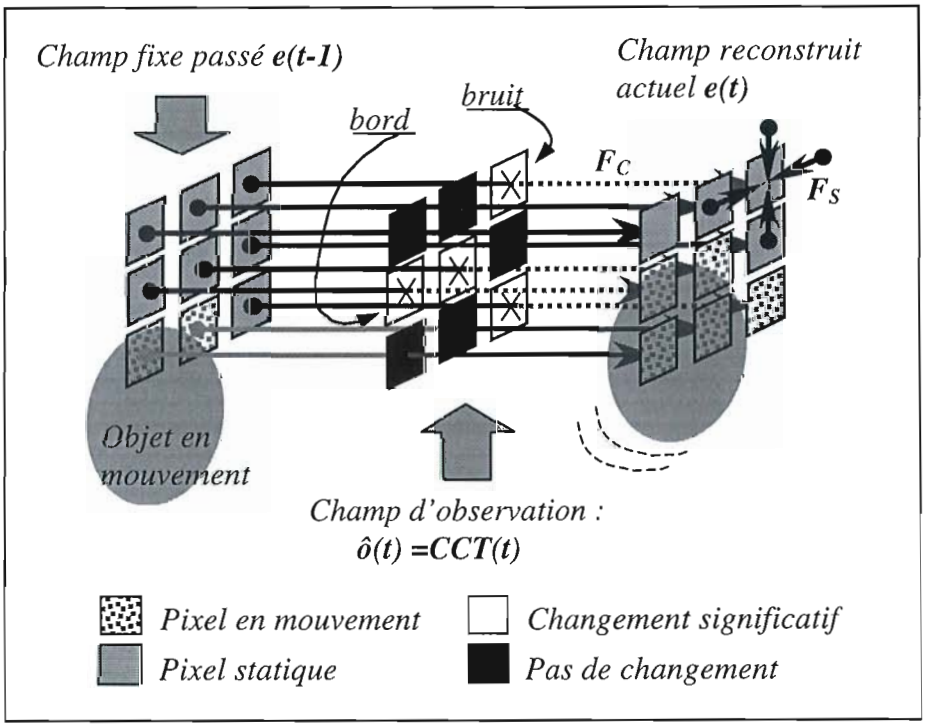

Fig.IV.13 : philosophie de l'algorithme de détection du mouvement proposé. La force spatiale de régularisation (F_S) prend en charge le traitement immédiat du bruit de la carte des changements temporels.

Notons une particularité du modèle (plus ou moins partagé par les modèles de Lalande et Dumontier) : si un objet présent dans la scène mais jusque là immobile commence à se

---

forme plus ou moins connexe sur fond immobile. Ce genre de cartes binaires n'est pas sans rappeler le résultat de la restauration d'images binaires bruitées étudié au Chap.I et II, pour lesquelles on avait $\lambda=\beta_S/\beta_O$ entre 0,4 et 1,2 ; en première approximation on aura alors β_S/β_C compris entre ces deux valeurs, mais une estimation plus précise est théoriquement possible (voir Chap.I,§5.1.2).

déplacer, alors il est facile de voir *que la trace initiale de l'objet immobile ne pourra pas disparaître par la suite* du champ d'étiquettes ; de même si un objet auparavant en mouvement s'immobilise tout à coup (et reste présent dans la scène), *son masque final (avant qu'il ne s'immobilise) ne disparaît pas non plus*. Il s'agit du même problème que peuvent rencontrer des méthodes de détection du mouvement basées sur la mise à jour d'une image de référence (§4.1). En fait, dans un certain nombre de cas ceci n'est pas un inconvénient, soit parce que l'objet mobile rentre et ressort de la scène–cas des voitures dans une route surveillée par une caméra fixe-, soit parce que précisément il peut s'avérer utile de garder la trace des objet mobiles *en puissance*.

Remarque : L'énergie totale à minimiser pour le modèle proposé est finalement :

$$U(e_t / e_{t-1}, \hat{o}_t) = U_{contr}(e_t / e_{t-1}, \hat{o}_t) + U_{reg}(e_t), \text{ où}$$

$$\begin{cases} U_{reg}(e_t) = -\dfrac{\beta_S}{2} \sum_{s \in S} \sum_{r \in V(s)} (2.e_t(s) - 1)(2.e_t(r) - 1), \text{ (\textit{régularisation} spatiale).} \\ U_{contr}(e_t / e_{t-1}, \hat{o}_t) = +\beta_C \sum_{s \in S} (2.e_{t-1}(s) - 1)(2.\hat{o}_t(s) - 1)(2.e_t(s) - 1) \\ \qquad\qquad = -\beta_C \sum_{s \in S} (2.b(s) - 1)(2.e_t(s) - 1), \text{ (terme de \textit{contrainte}).} \end{cases}$$

Nous avons posé dans la formule ci-dessus : $b(s)$ = [$e(s,t-1)$ xor $\hat{o}(s,t)$] ("xor" représente le *OU exclusif* logique). On a bien sûr $e(t-1)=MB(t-1)$ et $\hat{o}(t)=CCT(t)$. Le champ binaire b est donc le résultat d'un traitement binaire très simple sur les champs $MB(t-1)$ et $CCT(t)$. Si l'on laisse de côté le terme de lissage spatial, le minimum énergétique est tout simplement réalisé quand $e(t)=b(t)$; autrement dit, en absence de bruit un procédé simple pour obtenir le masque binaire des objets mobiles consiste à réaliser itérativement l'opération booléenne : $MB(t)$= [$MB(t-1)$ xor $CCT(t)$], avec $MB(0)$ initialisé à 0.

6.2.3 Méthode d'optimisation

La méthode d'optimisation choisie est bien sûr le *recuit simulé* étudié en détail au Chap.II. Le principal avantage est qu'il n'y a pas à se préoccuper de la "propreté" de l'initialisation du champ d'étiquettes avant de lancer la procédure d'optimisation, car l'algorithme converge vers un minimum global de la fonction d'énergie indépendamment de la configuration initiale. Par contre, la charge de calcul est bien plus importante que pour n'importe laquelle des méthodes déterministes, mais c'est précisément le rôle du PPOS étudié au Chap.III que de rendre le calcul d'optimisation possible même à cadence vidéo. La *fig.*IV.14 donne un schéma synoptique de l'algorithme proposé.

Il est intéressant de noter l'analogie entre la force locale correspondant au modèle de débruitage d'images binaires (*cf.*Chap.I,§5.1) et celle correspondant au modèle de détection :

- Détection de mouvement : $F(s,t) = \beta_S \sum_{r \in V(s)} (2.e_t(r) - 1) + \beta_C.(2.b(s) - 1)$

- Restauration d'images binaires : $F(s,t) = \beta_S \sum_{r \in V(s)} (2.e(r) - 1) + \beta_o.(2.o(s) - 1)$

La recherche du *MB* à chaque instant correspond donc au débruitage d'une "image" binaire : le champ *b(t)*=[*MB(t-1) xor CCT(t)*]. Les paramètres du *recuit* seront donc les mêmes pour le débruitage et pour la détection du mouvement. En particulier, on pourra utiliser le concept de *recuit sérieux* pour évaluer les performances du démonstrateur final.

Il est important de noter que la remarque faite au Chap.II.§3.2 à propos de la méthode de débruitage s'applique ici en toute rigueur : il se peut en effet que le résultat du débruitage du champ binaire b(t) soit "meilleur" (au sens de la conservation des discontinuités) pour une méthode d'optimisation déterministe *bien initialisée* que pour l'algorithme stochastique de recuit simulé. L'effet sera d'autant plus évident que les dimensions de l'image traitée resteront modestes (c'est le cas pour les imagettes 24x24 du démonstrateur optoélectronique). La prise en compte des discontinuités dans le modèle de débruitage rendrait sûrement meilleure la qualité du traitement stochastique par rapport aux méthodes déterministes. D'ailleurs, comme expliqué au §4.3, plusieurs techniques de détection du mouvement ont été proposées qui tirent avantage d'une détection conjointe des contours.

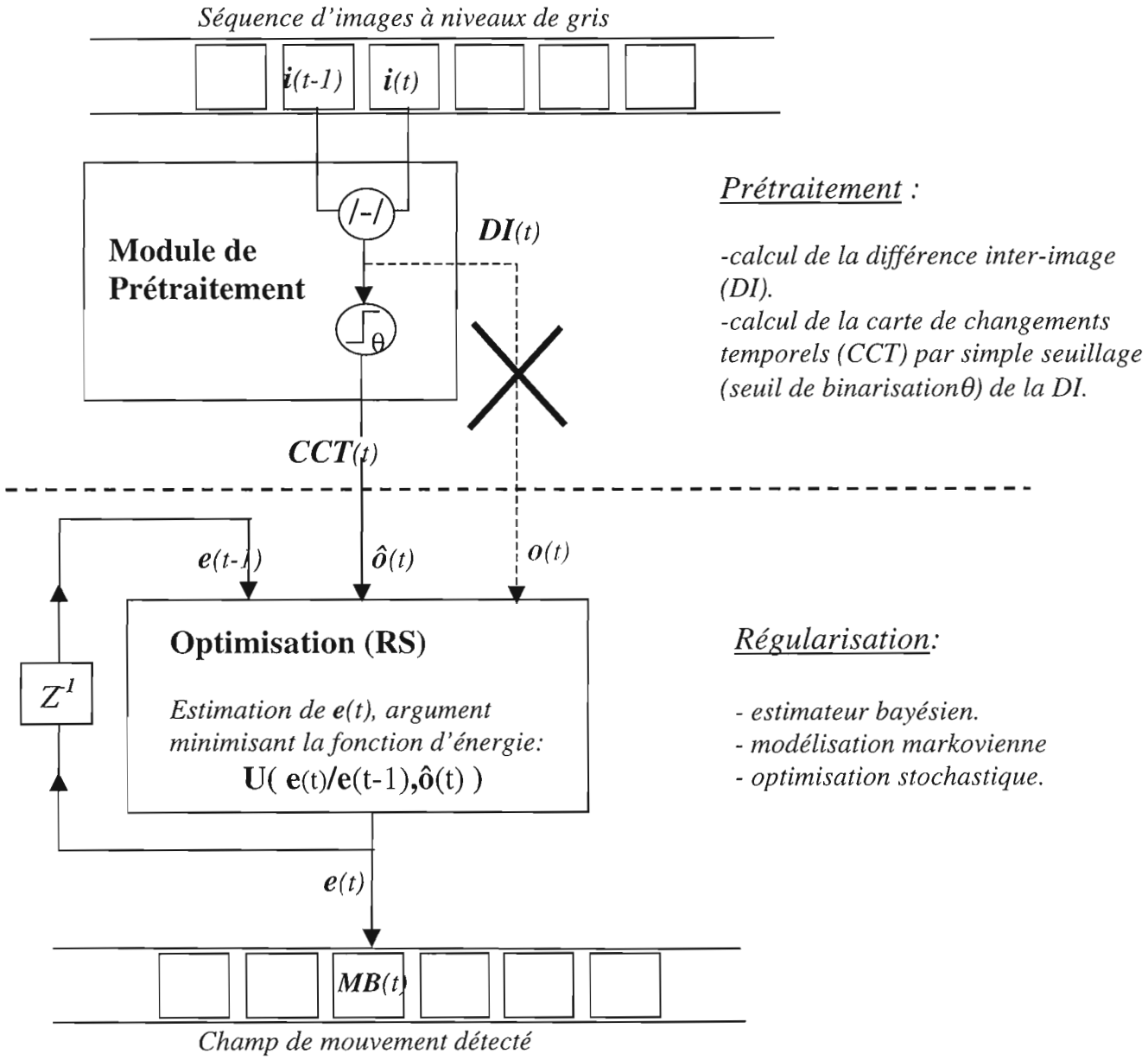

Fig.IV.14 : Schéma synoptique de l'algorithme final de détection du mouvement.

Le prétraitement se résume au calcul de la ***CCT*** *par simple seuillage de la différence inter-image (****DI****) ; celle-ci n'est pas utilisée comme champ de contrainte à l'observation. Le calcul du masque des objets mobiles* ***MB*** *est réalisé par estimation bayesienne et optimisation* stochastique *(et parallèle) -voir Chap.II ; il n'est donc pas nécessaire d'initialiser 'proprement' le champ d'étiquettes.*

Le module d'optimisation sera implanté matériellement grâce au processeur PPOS étudié au Chap.III. Pour simplifier la réalisation optoélectronique (modulateurs de lumière binaires rapides) les champs de contrainte aux données $\mathbf{e}(t-1)$ *et* $\hat{\mathbf{o}}(t)$ *sont* binaires.

7. Simulation et performances

7.1 Qualité du masque des objets mobiles MB.

Nous avons cherché à valider l'algorithme de détection du mouvement proposé sur quelques séquences réelles et synthétiques. On trouvera une description détaillée des séquences dans l'Annexe C ; on se limitera ici à commenter quelques résultats.

7.1.1 Séquences synthétiques.

Nous avons exploré rapidement l'influence du *bruit* du capteur, du *contraste* (entre l'objet mobile et le fond), de la *taille* de l'objet mobile et de la *vitesse* de celui-ci, et enfin l'incidence de la *forme* sur la qualité du masque des objets mobiles, et cela pour le modèle proposé et pour le modèle de Dumontier-Caplier.

Contraste et bruit : (voir en particulier les séquences *Contraste_1* et *Constraste_2* de l'Ann.C). Le taux d'erreur de canal ε est représentatif du rapport signal sur bruit du champ binaire d'*observation* (la CCT) -voir §6.2.1. Il dépend du contraste entre l'objet mobile et le fond, et aussi du bruit du capteur (supposé gaussien). La différence entre les niveaux de gris entre l'objet et le fond est de l'ordre de 20 unités pour les séquences étudiées ; la variance du bruit du capteur vaut $\sigma = 10$ (ces ordres de grandeur correspondent aux niveaux de gris rencontrés dans les séquences "réelles" -*cf* Annexe C). Dans ces conditions, le seuil de binarisation vaut $\theta \approx 12$ et la CCT à un taux de bruit de canal $\varepsilon \approx 23\%$; le terme de contrainte est proche de l'unité $\beta_C = 1{,}27$ (*cf.* §6.2.1).

On constate que pour des séquences "bien conditionnées" par rapport aux hypothèses du modèle (i.e. objet et fond sont uniformes, §3.2), le modèle discret proposé se comporte mieux que le modèle continu de Dumontier-Caplier. En particulier la zone de glissement est correctement reconstruite, ce qui n'est pas bien réalisé par le modèle continu. Ceci est dû au fait que ni l'observation continue $o(t)=DI(t)$ ni l'observation binaire $\hat{o}(t)=CCT(t)$ ne présentent de signal significatif sur le fond et sur la zone de glissement. Chose qui peut paraître paradoxale, les performances du modèle de Dumontier-Caplier s'améliorent un peu quand le rapport signal sur bruit *diminue* (le bruit augmente et/ou le contraste diminue), car des détections parasites apparaissent dans le "vide" de signal laissé dans la zone de glissement, détections qui permettent parfois une amorce de reconstruction -qui sera perpétuée par la suite grâce au terme de rappel temporel. Cette remarque tient aussi pour le reste des séquences étudiées (voir en particulier les séquences *Automobile* avec et sans bruit).

Taille : Une rétine capable de détecter le mouvement en temps-réel pourrait trouver des applications dans le domaine du contrôle routier. Une caméra munie d'un tel capteur pourrait par exemple être placé à la hauteur d'un sémaphore et être dirigé vers un carrefour ; dans ces conditions, il n'est pas absurde d'imaginer un champ de vision d'à peu

près dix mètres de large pour la caméra de surveillance. Imaginons un capteur de 128x128 pixels, ayant à une surface d'environ un centimètre carré. Le grandissement du système optique pour couvrir la largeur du champ doit être de 1cm/10m=10^{-3}.Une voiture mesure environ 3m de long sur 1,5 de large, et donne alors une image sur la rétine d'à peu près 20x10 pixels. Les niveau de gris et le bruit des séquences étudiées sont ici les mêmes que pour la séquence *Contraste_1*.

Pour des objets petits (de l'ordre de deux fois la taille du voisinage, voir séquence *Taille_1*), les deux algorithmes se comportent relativement bien. Pour des tailles supérieures (de l'ordre de 20x10 pixels - voir *Taille_2* et séquences *Automobile*), seul l'algorithme proposé arrive à reconstruire correctement la zone de glissement.

Vitesse : Il n'est pas absurde de supposer que les voitures ont, dans le carrefour, des vitesses comprises entre 10 et 60km/h. Les vitesses des images mobiles sur la rétine sont comprises alors entre 1 et 10 pixels/s. La zone de recouvrement correspond à 3% et 20% de l'objet mobile respectivement (voir séquences *Taille*, *Contraste* et séquences réelles). La séquence *Vitesse* représente la situation particulière dans laquelle il *n'y a pas de zone de glissement*, car l'objet se déplace trop vite entre deux séquences consécutives. On constate un comportement particulier pour l'algorithme de Dumontier-Caplier (le masque de l'objet est accompagné de deux "copies" -passé et futur). Notre modèle souffre d'un autre défaut, à savoir, l'incapacité à éliminer la trace initiale de la zone d'écho (voir séquences). Un phénomène similaire peut avoir lieu si le signal fourni par la différence inter-image sur la zone d'écho est en dessous du seuil de binarisation (par exemple pour un objet dont le niveau de gris ne soit pas uniforme mais évolue spatialement). C'est ce qui arrive dans la séquence *Automobile_1* (les roues de la voiture semblent se "multiplier").

Forme : Le choix de la structure du voisinage markovien doit en principe être fait en fonction des caractéristiques spatiales de l'image (et les caractéristiques du mouvement). Ainsi, un voisinage d'ordre 1 favorise les angles droits et les mouvements parallèles aux axes de la grille ; un voisinage d'ordre 2 favorise les coins arrondis et les mouvements souples. Le modèle de Dumontier-Caplier testé utilise exclusivement un voisinage d'ordre 2. Nous avons testé notre modèle en choisissant un voisinage d'ordre 1, c'est pourquoi les masques reconstruites sur les séquences correspondant à des rectangles en mouvement sont plus fidèles à la forme de l'objet pour notre modèle (voir notamment la séquence *Croisement*). La séquence *Disque* montre l'amélioration du rendu des arrondis grâce au voisinage d'ordre 2.

7.1.2 Séquences réelles

Nous avons cherché à valider le modèle sur quelques scènes de rue, prises grâce à une caméra de vidéoconférence pour PC (QuickCam™). Les conditions naturelles d'éclairement, de contraste, de bruit, etc., ne sont pas contrôlées ni estimées

automatiquement (il faut garder à l'esprit qu'il s'agit juste d'une démonstration). Les séquences présentent un bruit relativement faible ; par contre une grande quantité de facteurs parasites (variations d'éclairement locales, "motifs" internes à l'objet mobile, dégradé du niveaux de gris sur les surfaces) écartent le problème des hypothèses faites au §3.2. Sur ces séquences l'algorithme de Dumontier-Caplier semble plus performant, au moins pour ce qui est de la localisation grossière de la région en mouvement (voir notamment la séquence *Couloir*).

7.1.3 Conclusion.

L'observation des séquences réelles et synthétiques laisse apparaître une certaine *complémentarité* dans les modèles. Tout d'abord, le modèle de Dumontier-Caplier semble s'attaquer directement à la régularisation spatiale de la carte de changements temporels (il est aisé de remarquer que le masque des objets mobiles ressemble beaucoup à une simple version débruitée de la *CCT*). Le principal défaut du modèle, comme on pouvait s'y attendre d'après l'analyse faite au §5.7, est que les régions régulières et uniformes de l'objet en mouvement ne peuvent pas être détectées correctement. Cela tient au choix de la fonction d'observation Ψ_D (*cf.*§5.5), dont le principe est de ne considérer en mouvement que les régions présentant une forte variation d'intensité (supérieure au seuil α). Par contre, le modèle se comporte mieux face aux situations mal conditionnées par rapport aux hypothèses définies au §3.2 (détails trop fins et/ou mouvement trop rapide de l'objet mobile ne permettant pas de définir correctement les quatre zones de la *CCT* - fond, zone de glissement, zone d'écho et zone recouvrement). C'est le cas pour des objets texturés et/ou à niveaux de gris variable dans le temps (changements d'éclairement locaux dus par exemple à la variation de l'inclinaison des surfaces), mais aussi pour des objets mobiles dont la *forme* évolue dans le temps (dans ce dernier cas le masque reconstruit indique bien la *région* mobile, mais avec une perte substantielle de la *forme* de l'objet mobile). C'est bien sûr le cas pour les séquences réelles étudiées (*cf.*Ann.C, séquences *Piétons* et *Couloir* notamment).

Quant au modèle proposé, on remarquera la forte ressemblance entre le masque final et le champ binaire noté *XOR(t)* dans l'Ann.C, issu de l'opération itérative : *XOR(t)*=[*XOR(t-1) xor CCT(t)*]). La ressemblance est d'autant plus grande que le bruit est faible (comparer séquences réelles avec et sans ajout de bruit). En absence de bruit, le champ *XOR* est en toute rigueur équivalent au *MB*, mais en présence de bruit, le champ *XOR* s'écarte de plus en plus du masque binaire *MB*, car le bruit s'accumule irrémédiablement. Autrement dit, le masque binaire des objets mobiles *n'est pas* une simple version débruitée du champ *XOR* - ceci est vrai pour t=0, mais l'avantage de la restauration se fait sentir de plus en plus fortement au fur et à mesure du déroulement de la séquence.

Ceci nous conduit à une remarque d'ordre général : il y a deux situations dans lesquelles la détection du mouvement est possible sans avoir recours à une image de

référence ou à des techniques plus élaborées (reconnaissance des formes, suivi de cibles par corrélation, etc.) :

a) La première correspond à la situation dans laquelle l'objet mobile possède des textures fines et/ou des motifs intérieurs constants[54], et le fond (uniforme ou également texturé) à *le même niveau de gris moyen* que l'objet (la séquence est *totalement* mal conditionnée par rapport aux hypothèses du §3.2). Dans ce cas, le débruitage de la carte de changements temporels fournit une approximation suffisamment bonne du masque binaire des objets mobiles, tandis que la méthode de reconstruction basée sur l'intégration de la zone de recouvrement et l'élimination de la zone d'écho s'avère inefficace (puisque le "saut" de l'objet mobile entre deux images consécutives est plus grand que la "taille" de la texture, et la CCT ne fournit pas les quatre zones décrites au §4.2.2). C'est par exemple le cas de la séquence mal conditionnée *Couloir* dans l'Ann.C.

b) La deuxième situation correspond au cas où objet et fond sont uniformes, mais possèdent *des niveaux de gris moyens différents* (autrement il serait impossible d'effectuer une quelconque détection). La séquence est "parfaitement conditionnée" par rapport aux hypothèses du §3.2. Dans ce cas, la *CCT* seule n'est pas une bonne approximation du *MB* (impossibilité de récupérer la zone de glissement), et la reconstruction du masque des objets mobiles en utilisant la stratégie basée sur l'intégration de la zone de recouvrement et l'élimination de la zone d'écho est justifiée. Cependant, dans une situation de type b) "pure", on peut se demander s'il ne serait plus simple de retrouver le *MB* en effectuant un simple seuillage de l'image permettant de dégager l'objet du fond et en régularisant ensuite le résultat obtenu pour en éliminer le bruit.

En fait, une situation *réelle* pour laquelle on ne connaît pas a priori les caractéristiques de l'image (niveaux de gris de l'objet et du fond, taille des textures, etc.) correspond à des séquences mal conditionnées par rapport aux hypothèses du §3.2, dont les caractéristiques évoluent dans le temps entre les situations a) et b), (à cause des variations d'éclairement et/ou la déformation de l'objet mobile, etc.). Les séquences réelles seront "plutôt de type a)" ou "plutôt de type b)".

Dans une situation "plutôt de type a)", la méthode de détection simple consistant à débruiter la *CCT* se verra considérablement améliorée si l'on rajoute des informations permettant de reconstruire la zone de glissement ; une stratégie pour le faire est proposé dans [Bellon94] consistant à prendre en compte à la fois la *CCT* et des informations provenant d'une image de référence. L'utilisation conjointe de ces deux champs d'observation (de nature très différente) permet selon l'auteur de mieux gérer les zones de glissement.

---

[54] "texture fine et/ou motif intérieur constant" doit être compris ici au sens de "détails pour lequel la CCT ne permet pas d'obtenir l'ensemble des quatre zones définies au §3.2, compte tenu de la vitesse de l'objet, la fréquence de capture d'images et la résolution du capteur" (voir note bas de page 5 au §3.2).

Dans une situation "plutôt de type b)" il est tout d'abord vraisemblable (mais on ne l'a pas testé) que la méthode basée sur le traitement spatio-temporel à partir de la *CCT* soit plus robuste face aux aléas dus au mauvais conditionnement de la séquence qu'un simple seuillage suivi d'une régularisation spatiale, et cela grâce à l'effet "mémoire" propre à la contrainte temporelle. Par ailleurs, la méthode de détection se verra sans doute améliorée si l'on rajoute des informations permettant de "rattraper" les erreurs produits par le mauvais conditionnement de la séquence ; pour ce faire, une stratégie peut consister - comme c'est le cas dans le modèle de Lalande - à analyser le champ d'observation continu $o(t)$ grâce à *deux* seuils m_1 et m_2 (*tab.*IV.1 au §5.4.1) au lieu d'un seul (paramètre θ de binarisation), ce qui permet de distinguer les transitions fond-objet (ou objet-fond) et les transition fond-fond (ou objet-objet) ; l'idée subsiste (simplifiée) dans le modèle de Dumontier, car si l'observation continue $o(t)$ dépasse le seuil α (voir §5.5), le site sera "encouragé" par le terme de contrainte à prendre l'état de mouvement, quel qu'eut été l'état de mouvement passé (voir *fig.*IV.7 et commentaire au §5.5.1).

> **Remarque** : Si l'éclairement est constant, il est toujours possible de ramener une situation de type a) ("mal conditionnée par rapport aux hypothèses du §3.2") à une situation de type b) ("bien conditionnée") en augmentant la résolution du capteur et/ou sa fréquence de capture.

Une remarque pour finir : le modèle de Dumontier-Caplier a été testé en utilisant un algorithme d'optimisation déterministe (*ICM*). La minimisation de la fonction d'énergie du modèle est donc plus ou moins bien approchée (et il n'y a pas de moyen évident de mesurer le biais que cela introduit dans le résultat final de la détection). La comparaison des modèles devrait se faire en utilisant un algorithme d'optimisation performant de type recuit simulé - en temps différé bien sûr. Le modèle de Dumontier-Caplier a été testé par les auteurs en utilisant des techniques déterministes, et l'algorithme final prescrit également la façon d'initialiser le champ d'étiquettes (au moyen de la $CCT(t+1)$, voir *fig.*IV.8).

7.2 Charge de calcul.

Nous n'avons pas réalisé des véritables recuits sérieux (1000 itérations) sur les séquences de test ; nous avons préféré optimiser de façon heuristique le nombre d'itérations du recuit, qui s'élève finalement à une *centaine* par image. La méthode de Dumontier-Caplier utilise une technique de relaxation déterministe (l'*ICM*) comportant entre 4 et 5 itérations par image. Finalement, la charge de calcul correspondant au traitement du problème de la détection du mouvement par recuit simulé est quelque 20 fois plus grande que pour une technique basée sur une méthode déterministe (en tout cas pour des images de l'ordre de 100x100 pixels). On pourra comparer au *tab.*VI.1 du ChapVI,§3.2 les performances respectives de plusieurs réalisations matérielles conçues

pour traiter le problème de la détection du mouvement par régularisation markovienne ; on verra alors que seul un processeur exploitant l'approche du *parallélisme massif* peut aujourd'hui prétendre rendre compte de cette énorme charge de calcul en *temps réel*.

8. Conclusion

Au Chap.III nous avons présenté un prototype de processeur optoélectronique capable de simuler le comportement d'un réseau bidimensionnel de spins (modèle d'Ising 2D). Moyennant un montage optique adéquat, on peut utiliser ce prototype pour traiter le problème du débruitage d'images binaires, et donc à l'occasion le problème de la détection du mouvement (*cf* §6.3). Au chapitre suivant nous allons donc aborder la conception et la réalisation d'un tel montage optoélectronique. L'application visée est bien sûr la détection du mouvement à cadence vidéo (sur des images 24x24, n'oublions toujours pas qu'il s'agit d'un démonstrateur), mais le montage servira aussi pour étudier d'autres aspects fondamentaux des montages optoélectroniques, à savoir les *interconnexions optiques en espace libre*.