Extraction de Couches de Même Mouvement Via des Techniques Combinatoires

Motion Layer Extraction of Similar Motion Through Combinatorial Techniques

Romain Dupont [1,2] Nikos Paragios[1] Renaud Keriven[1] Phillipe Fuchs[2]

[1] Atlantis Research Group, CERTIS, ENPC, Marne-La-Vallee, France
[2] Centre de Robotique, ENSMP, Paris, France

{dupont,paragios,keriven}@certis.enpc.fr
fuchs@ensmp.fr

Résumé

On présente une nouvelle technique d'extraction de couches dans une séquence vidéo. Dans ce but, on suppose que la scène observée est composée de plusieurs couches transparentes et que leur mouvement dans le plan 2D peut être approximé par un modèle affine. L'objectif de notre approche est l'estimation de ces modèles de mouvement ainsi que de leur support dans l'image. Notre technique s'appuie sur un processus itératif qui intègre une estimation robuste du mouvement, une formulation MRF, des techniques combinatoires et sur l'utilisation de critères visuels et de mouvements pour extraire les paramètres des modèles de mouvement et les supports des couches. La prise en compte des occlusions ainsi que les techniques adaptives pour détecter de nouveaux objets dans la scène sont aussi considérées. Des résultats prometteurs démontrent les potentiels de notre approche.

Mots Clef

Segmentation, Séquence Vidéo, Couches, Mouvement, Statistiques de Couleurs, MRF, Graphe, Coupe Minimale

Abstract

In this paper we present a new technique to extract layers in a video sequence. To this end, we assume that the observed scene is composed of several transparent layers, that their motion in the 2D plane can be approximated with an affine model. The objective of our approach is the estimation of these motion models as well as the estimation of their support in the image domain. Our technique is based on an iterative process that integrates robust motion estimation, MRF-based formulation, combinatorial optimization and the use of visual as well as motion features to recover the parameters of the motion models as well as their support layers. Special handling of occlusions as well as adaptive techniques to detect new objects in the scene are also considered. Promising results demonstrate the potentials of our approach.

Keywords

Motion Segmentation, Video Sequence, Layer, Color Statistic, MRF, Graph Cut

1 Introduction

La perception du mouvement est une caractéristique importante de la vision biologique et constitue une source d'information très précieuse pour l'analyse bas-niveau et moyen-niveau des séquences vidéos.

La segmentation d'une séquence vidéo en régions de mouvement homogène est un domaine d'étude encore très actif [23, 3, 1, 24, 22, 12, 7, 20] notamment à des fins de vidéosurveillance ou de reconnaissance de gestes. De même, la représentation compacte de la scène en modèles de mouvement avec leurs supports associés constitue une base naturelle pour la compression vidéo [23].

Extraire le mouvement dans une séquence vidéo est une tâche difficile. Une première approche est le calcul de flot optique [16, 2, 10, 5, 22] qui consiste à estimer le mouvement en chaque pixel à partir des contraintes visuelles. C'est un problème encore ouvert aujourd'hui car mal posé, le nombre d'inconnues excédant celui des contraintes visuelles locales. Tout l'enjeu réside dans l'utilisation de contraintes supplémentaires adaptées de lissage et autre techniques sophistiquées pour estimer correctement le flot optique.

Les modèles paramétriques sont une alternative pour

l'estimation du flot optique dense d'une séquence vidéo [19, 5, 6, 22]. On s'appuie alors sur l'hypothèse que le mouvement 2D d'un objet 3D projeté dans l'image suit un modèle paramétrique. Une telle hypothèse est notamment vraie si l'objet 3D est planaire.
L'objectif est ainsi d'extraire les différentes surfaces planaires qui représenteront les *couches* de la séquence vidéo ainsi que les paramètres de leurs modèles de mouvement décrivant leurs déplacements. Dans la littérature, on peut citer diverses approches telles qu'un algorithme de segmentation par K-mean [23] sur les mouvements estimés ou une approche MDL ("Minimum Description Length") [1] pour estimer le nombre de couches présentes dans la scène 3D. Pour cette dernière approche, l'extraction des couches est effectuée via un critère de maximisation de la vraisemblance et une optimisation par l'algorithme EM [13, 1]. D'autres approches [25, 26] s'appuient sur une technique de croissance de régions de même mouvement à partir de graines, dans un cadre combinatoire [26].
Dans cet article, on présente une nouvelle technique itérative pour estimer le support de chaque couche ainsi que leurs paramètres de mouvement respectifs et leurs propriétés visuelles. Ces dernières sont utilisées pour les cas ambiguës où le mouvement ne suffit pas pour estimer les supports. Notre approche estime les mouvements à travers une technique incrémental et robuste qui extrait les supports des couches en utilisant les Champs de Markov Aléatoires (MRF) résolus par graphe (Graph Cut) et l'algorithme d'*alpha*-expansion. A cette fin, plusieurs critères sont considérés : les résidus dus aux mouvements, les propriétés visuelles et les contraintes de lissage temporel et spatial.
L'article est organisé comme suit : en section 2, on introduit brièvement le problème de l'extraction de couches. En section 3, on détaille l'approche itérative pour estimer le mouvement de chaque couche. La section 4 aborde le principe de l'extraction des supports des couches. Enfin, la section 5 présente les résultats de validation et conclut l'article.

2 Décomposition de la scène en couches

Nous considérons une scène statique composée de plusieurs plans 3D, capturée par un observateur mobile sous la forme une séquence d'images 2D. En raison du mouvement propre de la caméra, les zones statiques de la scène sont en déplacement dans l'image. Leurs mouvements 2D dépendent du modèle projectif de la caméra et de la profondeur des différents plans 3D, information inconnue ici.

2.1 Supports de couches

Il s'agit ici de séparer le domaine de l'image en n régions $\mathcal{S}_i, i \in [1, n]$ avec leurs modèles de mouvement correspondants $(\mathcal{A}_i, \sigma_i)$ où $\mathcal{A}_i$ représente les paramètres du modèle et σ_i la variance des résidus associée au modèle (détaillés en section 3). On définit la couche $\mathcal{L}_i, i \in [1, n]$ comme étant le couple $(\mathcal{S}_i, \mathcal{A}_i)$.
Le domaine de l'image Ω est ainsi partitionné en n ensembles disjoints $\mathcal{S}_i$ tels que $\cup_{i=1}^{n} \mathcal{S}_i = \Omega, \quad \mathcal{S}_i \cap \mathcal{S}_j = \emptyset, i \neq j$. Ni le nombre de couches, ni leurs supports, ni leurs paramètres de modèle de mouvement sont connus. Le reste de l'article proposera des méthodes efficaces pour estimer toutes ces inconnues.

2.2 Modèles de mouvement

Il existe de nombreux modèles de mouvement, de divers degrés de liberté, similarité, homographie, quadratique etc. Le modèle affine est un bon compromis entre une faible complexité et une approximation satisfaisante des mouvements des objets de même profondeur. Il consiste en 6 degrés de liberté,

$$\mathcal{A}(x, y) = \begin{pmatrix} a_0 \cdot x + a_1 \cdot y + a_2 \\ a_3 \cdot x + a_4 \cdot y + a_5 \end{pmatrix}$$

Ce modèle décrit précisément le mouvement induit par un objet planaire vu depuis une caméra en déplacement. De plus, les mouvements 2D dus à la projection d'une scène 3D quelconque sujette à des mouvements de rotation, de zoom ou de faible translation sont correctement représentés par ce modèle [3, 17]. De surcroît, quand la profondeur générale de l'objet est supérieure à son épaisseur, le modèle affine permet une précision sub-pixellique.
L'estimation du mouvement consiste à déterminer les paramètres de ce modèle de sorte à établir une correspondance entre les projections 2D des mêmes surfaces 3D entre deux images successives. Cette étape étant critique pour obtenir une segmentation satisfaisante, nous détaillons ici le processus d'estimation de mouvement utilisé pour une région donnée.

3 Estimation du mouvement

Pour estimer le mouvement, on considère usuellement l'hypothèse de luminosité constante d'une image à l'autre [16]. Sous une telle hypothèse, les surfaces lambertiennes et non sujettes à des changement d'illumination globale, l'apparence de la projection 2D de la même surface 3D restera constante dans le temps. Ainsi, si l'on considère le vecteur de déplacement $\mathbf{dx} = (dx, dy)$ du pixel $\mathbf{x} = (x, y)$, on vérifie alors la condition suivante :

$$I(\mathbf{x}; t) \approx I(\mathbf{x} + \mathbf{dx}; t + 1), \ (i)$$
$$I(\mathbf{x}; t) \approx I(\mathbf{x} + \mathcal{A}(\mathbf{x}); t + 1), \ (ii)$$

pour le cas d'un mouvement dense (i) et pour le cas d'un mouvement affine (ii). On peut alors définir le résidu total du au mouvement :

$$E(\mathcal{A}) = \int_{\Omega} |I(\mathbf{x}; t) - I(\mathbf{x} + \mathcal{A}(\mathbf{x}); t + 1)|^2 \, d\mathbf{x} \qquad (1)$$

Pour déterminer les paramètres du modèle affine qui minimiseront la fonction définie dans Eq. 1, on procède en deux étapes : l'initialisation suivie d'un processus itératif pour raffiner l'estimation initiale.

3.1 Estimation initiale

Pour une première estimation des paramètres du modèle affine $\mathcal{A}$, on utilise la forme linéaire de premier ordre de la contrainte locale du flot optique que l'on adapte au cas du modèle affine :

$$\mathcal{A}(\mathbf{x}) \cdot \nabla I(\mathbf{x};t) + \nabla_t I(\mathbf{x}) = 0 \quad (2)$$

où ∇I est le gradient spatial et $\nabla_t I$ le gradient temporel. On cherche alors à minimiser la fonction de coût correspondante :

$$E(\mathcal{A}) = \int_\Omega |\mathcal{A}(\mathbf{x}) \cdot \nabla I(\mathbf{x};t) + I(\mathbf{x};t+1) - I(\mathbf{x};t)|^2 \, d\mathbf{x} \quad (3)$$

via une méthode linéaire standard mais qui est incapable de prendre en compte correctement les larges déplacements entre deux images successives. Pour contourner cette limitation, on considère un processus itératif.

3.2 Raffinements successifs de l'estimation

On s'appuie sur la méthode décrite dans [19] : à chaque itération, connaissant l'estimation courante $\mathcal{A}$ des paramètres du modèle, on estime $\Delta\mathcal{A}$ de sorte à minimiser l'erreur résiduelle suivante :

$$E(\Delta\mathcal{A}) = \int_\Omega [I(\mathbf{x};t) - I(\mathbf{x}+\mathcal{A}(\mathbf{x});t+1) - \Delta\mathcal{A}\nabla I(\mathbf{x}+\mathcal{A}(\mathbf{x});t+1)]^2 \, d\mathbf{x} \quad (4)$$

dont la minimisation est directe en utilisant les méthodes linéaires classiques. Si cette méthode incrémentale améliore le processus d'estimation, elle sera toujours biaisée par la présence de pixels aberrants (occlusions, luminosité localement non constante d'une image à l'autre, etc.) ou par la présence de plusieurs mouvements indépendants.

3.3 Estimation robuste

L'utilisation d'un M-estimateur permet de contourner cette limitation. Cet estimateur robuste assigne des poids $we(\mathbf{x})$ aux contraintes associées à chaque pixel. Ces poids dépendent des erreurs résiduelles de façon disproportionnelle, permettant ainsi de réduire l'influence des pixels aberrants. Par la même occasion, seul le mouvement dominant de la région considérée sera réellement estimé. A cette fin, on définit la fonction d'influence $\psi(x)$ comme suit (estimateur de Tukey [Fig. 1]) :

$$\psi(x) = \begin{cases} x({K_\sigma}^2 - x^2)^2 & \text{if } |x| < K_\sigma \\ 0 & \text{sinon} \end{cases} \quad (5)$$

où K_σ caractérise la forme de la fonction robuste. Les poids $we(\mathbf{x})$ sont alors calculés comme suit ([19]) :

$$we(\mathbf{x}) = \frac{\psi(r(\mathbf{x}))}{r(\mathbf{x})}$$

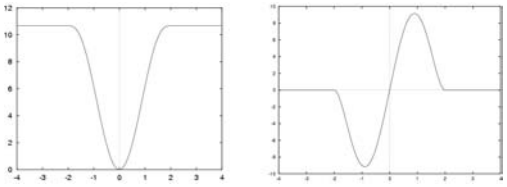

Fig. 1 – Estimateur de Tukey ρ (à gauche) et sa dérivée (fonction d'influence) ψ (à droite)

On peut alors considérer un tel processus pour chaque couche indépendamment qui consiste à minimiser la fonction de coût suivante :

$$E(\Delta\mathcal{A}_1, ..., \Delta\mathcal{A}_n) = \sum_{k=1}^{n} \int_\Omega \mathbf{1}_{\mathcal{S}_i(\mathbf{x})} \rho \left[I(\mathbf{x};t) - I(\mathbf{x}+\mathcal{A}_i(\mathbf{x});t+1) - \Delta\mathcal{A}_i \nabla \, I(\mathbf{x}+\mathcal{A}_i(\mathbf{x});t+1) \right] d\mathbf{x} \quad (6)$$

où $\mathbf{1}_{\mathcal{S}_i}$ est la fonction caractéristique de la région $\mathcal{S}_i$. Si les supports des couches sont connus, on peut alors procéder à l'estimation des modèles de mouvement. Ainsi, avec la méthode robuste décrite ci-dessus, pour une partition donnée de l'image, on est en mesure de déterminer les paramètres des modèles affines et de décrire le mouvement observé. Le problème est maintenant d'avoir une estimation consistante des supports des couches.

4 Extraction des supports des couches

Considérons n couches ainsi que leurs n modèles de mouvement $\mathcal{A}_i$ correspondants. L'extraction des supports consiste à choisir pour chaque pixel de Ω la couche la plus vraisemblable en terme de mouvement parmi les n couches. C'est un problème d'étiquetage où l'on cherche à assigner à chaque pixel $\mathbf{x}$ une étiquette $\omega(\mathbf{x}) \in [1, n]$ selon un critère donné. Notre approche s'appuie sur les critères de mouvements et d'apparence visuelle tout en imposant des contraintes de lissage à la fois spatial et temporel que nous détaillons ci-dessous dans les sous-sections suivantes.

4.1 Critère de mouvement

Considérons la distribution des erreurs résiduelles au sein d'une couche. Sous l'hypothèse que l'estimation du mouvement et des supports est correcte, on peut considérer que les erreurs résiduelles sont dues au bruit, en général gaussien.

Ainsi, le résidu du au mouvement

$$r_i(\mathbf{x}) = \|I(\mathbf{x},t) - I(\mathbf{x}+\mathcal{A}_i(\mathbf{x}),t+1)\| \quad (7)$$

pour la couche $\mathcal{L}_i$ suit une loi normale $G(\mu_i,\sigma_i)$. Par conséquence, d'après le résidu observé, la probabilité pour un pixel donné d'appartenir à la région $\mathcal{S}_i$ s'écrit :

$$p\left(r_i(\mathbf{x})|\mathcal{A}_i,\sigma_i\right) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(\frac{-{r_i}^2(\mathbf{x})}{2{\sigma_i}^2}\right) \quad (8)$$

où σ_i est l'écart-type de chaque support de couche calculé durant l'estimation du mouvement, en utilisant un estimateur robuste qui tolère efficacement jusqu'à 50% de points aberrants :

$$\sigma_i = 1.4826\left\{\underset{\mathbf{x}\in\mathcal{S}_i}{\text{median}}\,|r_i(\mathbf{x})|\right\} \quad (9)$$

De même, on considère que les erreurs résiduelles des pixels appartenant à $\mathcal{S}_i$ sont indépendantes en terme de probabilité. On cherche alors à maximiser la probabilité conditionnelle :

$$\begin{aligned} p_i(r_i(\mathbf{x})|\mathcal{A}_i,\sigma_i) &= p\left(\bigcap_{\mathbf{x}\in\mathcal{S}_i}\{r_i(\mathbf{x})|\mathcal{A}_i,\sigma_i\}\right) \\ &= \prod_{\mathbf{x}\in\mathcal{S}_i} p(r_i(\mathbf{x})|\mathcal{A}_i,\sigma_i) \end{aligned} \quad (10)$$

De plus, on considère que les erreurs résiduelles sont indépendantes entre les différents supports de couches. Ainsi, utilisant la règle de Bayes, le processus d'étiquetage s'appuyant sur le critère de mouvement peut s'écrire :

$$\begin{aligned} p(r_i(\mathbf{x})|\mathcal{A}_i, \quad \sigma_i,\omega) = \\ \prod_{\mathbf{x}\in\Omega} p_{\omega(\mathbf{x})}(r_{\omega(\mathbf{x})}(\mathbf{x})|\mathcal{A}_{\omega(\mathbf{x})},\sigma_{\omega(\mathbf{x})}) \end{aligned} \quad (11)$$

où l'on considère l'hypothèse que les étiquettes sont équiprobables. Maximiser la probabilité à posteriori est équivalent à minimiser l'opposé de la log-vraisemblance d'une telle densité (pour l'image t) :

$$\begin{aligned} E_m^t(\omega) &= -\int_\Omega \log\left[p_{\omega(\mathbf{x})}\left(r_{\omega(\mathbf{x})}(\mathbf{x})|\mathcal{A}_{\omega(\mathbf{x})},\sigma_{\omega(\mathbf{x})}\right)\right]d\mathbf{x} \\ &= \int_\Omega\left(\log\left[\sigma_{\omega(\mathbf{x})}\right] + \frac{{r_{\omega(\mathbf{x})}}^2(\mathbf{x})}{2{\sigma_{\omega(\mathbf{x})}}^2}\right)d\mathbf{x} \end{aligned} \quad (12)$$

Le potentiel le plus faible de cette fonction va alors classer les pixels selon leurs erreurs résiduelles. Cependant, l'estimation du mouvement n'est fiable qu'en présence de structures au sein de l'image. Par conséquence, la classification par mouvement sera ambiguë dans certains cas, notamment dans les régions peu texturées.

4.2 Critère d'apparence visuelle

L'introduction d'un critère d'apparence visuelle pour la classification permettra de contourner de telles ambiguïtés. A cette fin, nous avons choisi d'utiliser une fonction de densité flexible et paramétrique, à savoir une mixture de gaussiennes, pour décrire les propriétés visuelles de chaque couche ;

$$p_i(I) = \sum_{k=1}^{m_i}\pi_{\{i,k\}}\; p_{\{i,k\}}\left(I|\mu_{\{i,k\}},\Sigma_{\{i,k\}}\right) \quad (13)$$

où $p_i()$ est la distribution des couleurs de la i-ième région qui consiste en m_i composantes gaussiennes avec $\pi_{\{i,k\}} \in [0,1]$ étant la probabilité à priori sur la composante k (ou sa proportion dans la mixture) et $(\mu_{\{i,k\}}, \Sigma_{\{i,k\}})$ la moyenne et la matrice de covariance de cette composante. Chaque gaussienne représente généralement l'une des couleurs dominantes de la région considérée (voir Fig. 2 pour un exemple de mixture de gaussiennes). Ces paramètres sont estimés à partir de la distribution observée - ici les pixels de la i-ième région de l'image t - via un algorithme EM [14]. Pour déterminer efficacement le nombre de gaussiennes par mixture, un critère dit MDL (Minimum Description Length [21]) est considéré. Cette paramétrisation de l'apparence visuelle fournit un moyen simple et efficace pour apprendre les caractéristiques visuelles de chaque couche sans la contrainte d'être uni-modal. Ainsi, même les régions avec des caractéristiques colorimétriques variables appartenant au même plan 3D seront regroupées.

Fig. 2 – Exemple d'une mixture de gaussiennes (à gauche de l'image) : ici, la distribution des couleurs de l'image de gauche est représentée par 6 gaussiennes. Chaque gaussienne est représentée par un hexagone qui indique sa proportion en pourcentage, la couleur moyenne (couleur de l'hexagone) et sa variance (sur les bords de l'hexagone) pour les trois couches RVB (ou LAB).

La probabilité à posteriori est une métrique efficace pour diviser le domaine de l'image en régions à partir de leurs apparences visuelles propres. Comme pour le cas du mouvement, on considère que les couches ainsi que les pixels au sein de chaque région sont indépendants en terme de probabilité et que les étiquettes

sont équi-probables, menant à la fonction objective suivante (pour l'image t) :

$$E_v^t(\omega) \quad = \quad -\int_\Omega \log\left[p_{\omega(\mathbf{x})}\left(I(\mathbf{x})\right)\right] d\mathbf{x} \qquad (14)$$

où $p_{\omega(\mathbf{x})}(I)$ est définie par l'équation 13. On cherche alors à minimiser le potentiel de l'addition de ces deux termes (mouvement + apparence visuelle) pondérés par une constante afin d'extraire la partition de l'image la plus appropriée en terme de supports de couches. La méthode de classification présentée permettra de déterminer les supports via un processus de décision indépendant pour chaque pixel selon les similarités entre l'image observée et les propriétés attendues en terme d'apparence visuelle et d'erreurs résiduelles dues au mouvement. Cependant, ce processus indépendant formera de nombreuses discontinuités visuellement choquantes.

4.3 Lissage spatial

On fait alors appel à des contraintes de lissage spatial de sorte que les pixels du même voisinage appartienne à la même couche, ce qui revient à minimiser l'énergie suivante (pour l'image t) :

$$E_s^t(\omega) = \int_\Omega \left[\int_{\mathcal{N}(\mathbf{x})} \mathcal{V}(\omega(\mathbf{x}),\omega(\mathbf{u}))d\mathbf{u}\right] d\mathbf{x} \qquad (15)$$

où $\mathcal{N}(\mathbf{x})$ est le voisinage local du pixel $\mathbf{x}$. Ici, la fonction $\mathcal{V}$ est le modèle de Pott qui est défini comme suit :

$$\mathcal{V}(\omega(\mathbf{x}),\omega(\mathbf{u})) = \begin{cases} 1 & \text{si } \omega(\mathbf{x}) \neq \omega(\mathbf{u}) \\ 0 & \text{sinon} \end{cases} \qquad (16)$$

dont le voisinage local est composé de pixels qui sont 4- ou 8-connexes. Un tel terme favorisera une segmentation constante par morceaux. Cependant, les discontinuités en terme de mouvement correspondant généralement aux discontinuités en terme d'intensité, on tolèrera les discontinuités de classification d'autant plus facilement dans les zones à fort gradient (en général sur les contours des objets) en pondérant le critère de lissage spatial par un terme inversement proportionnel au gradient de l'image [9]. La fonction $\mathcal{V}$ définie ci-dessus est ainsi pondérée comme suit [25] :

$$\mathcal{V}_g = \mathcal{V} \cdot \begin{cases} 4 & \text{si } \|I(\mathbf{x},t) - I(\mathbf{u},t)\| < 4 \\ 2 & \text{si } 4 \leq \|I(\mathbf{x},t) - I(\mathbf{u},t)\| < 8 \\ 1 & \text{sinon} \end{cases} \qquad (17)$$

Ainsi, un contour sera plus vraisemblablement une frontière entre deux régions que ne le sera une zone localement homogène.

4.4 Lissage temporel

On considère ici l'hypothèse que les supports des couches évoluent de façon continuelle d'une image t à l'image $t+1$. On définit ainsi une contrainte de continuité dans l'évolution des supports qui permettra de surcroît d'améliorer la robustesse de l'estimation des supports des couches pour chaque image.
Considérons une séquence d'images $I(;1)$, $I(;2)$, $\cdots$, $I(;\tau)$ ainsi qu'une séquence d'étiquetages (connus ou non) $\omega(;1)$, $\omega(;2)$, $\cdots$, $\omega(;\tau)$. Deux approches sont possibles selon la façon dont l'ensemble de la séquence est traité :

- soit séquentiellement, en considérant uniquement chaque pair d'image $I(;t) \leftrightarrow I(;t+1)$ et en s'appuyant sur les étiquetages de l'image $t-1$ déjà traitée,
- soit en considérant N images successives simultanément $I(;t) \leftrightarrow I(;t+1) \leftrightarrow \cdots \leftrightarrow I(;t+N)$ ($N \approx 3-5$).

Nous verrons en sous-section 5.1 que, pour une image de la séquence donnée, nous utiliserons soit l'approche séquentielle, soit l'approche simultanée.

Approche séquentielle. On considère l'image $I(;t)$ et on fait l'hypothèse que les modèles de mouvement $\mathcal{A}$ ont correctement été estimés pour l'image $I(;t-1)$. On définit alors la fonction de lissage temporel qui prend en compte les modèles de mouvement et les supports de l'image précédente :

$$\mathcal{V}^{\mathbf{t}}(\omega(\mathbf{x};t)) = \qquad (18)$$
$$\begin{cases} 1 \text{ si } \omega(\mathbf{x};t) \neq \omega\left(\mathcal{A}^{-1}_{\omega(\mathbf{x};t-1)}(\mathbf{x});t-1\right) \\ 0 \text{ sinon} \end{cases}$$

où $\mathcal{A}^{-1}_{\omega(\mathbf{x};t-1))}$ est le modèle de mouvement inverse qui établit une correspondance entre les images $I(;t)$ et $I(;t-1)$ pour la couche $\omega(\mathbf{x};t-1)$. On définit alors le terme de lissage temporel séquentiel :

$$E_{ts}(\omega) = \int_\Omega \mathcal{V}^{\mathbf{t}}(\omega(\mathbf{x}))d\mathbf{x} \qquad (19)$$

Approche simultanée. Le critère de lissage temporel proposé ci-dessus ne prend en compte que l'image précédente (dont l'étiquetage est connu) pour assurer une continuité temporelle des supports des couches. On propose ici une prise en compte de plusieurs images simultanément dont l'étiquetage n'est pas encore connu : au lieu de ne considérer qu'un seul couple d'images à la fois, on en considère N, reliées séquentiellement par des contraintes temporelles que nous présentons ci-après.
On considère ici que les modèles de mouvement ont été estimés pour les N images considérées. On estime alors simultanément les supports des couches en rajoutant une contrainte de continuité temporelle des supports entre deux images successives. On écrit la contrainte comme suit :

$$\mathcal{V}^{\mathbf{t}'}(\omega(\mathbf{x};t)) = \begin{cases} 0 & \text{si } \omega(\mathbf{x};t) = \kappa \\ & \Rightarrow \omega\left(\mathbf{x} + \mathcal{A}_\kappa(\mathbf{x});t+1\right) = \kappa \\ 1 \text{ sinon} \end{cases} \qquad (20)$$

On définit alors le terme de lissage temporel simultané :

$$E_{ts'}(\omega) = \sum_{i=t}^{t+N-2} \left[\int_{\Omega} \mathcal{V}^{\mathbf{t}'}(\omega(\mathbf{x}); i) d\mathbf{x} \right] \quad (21)$$

Ainsi, si le pixel $\mathbf{x}$ a pour étiquette κ à l'instant t, on souhaite que son projeté $\mathbf{x} + \mathcal{A}_\kappa(\mathbf{x})$ dans l'image $t+1$ aie aussi l'étiquette κ. Notons que l'inverse n'est pas nécessaire. De plus, l'estimation du mouvement n'étant généralement pas parfaite, nous ne mettons pas de contrainte forte (pénalité infinie par exemple en cas de discontinuité temporelle).

4.5 Classification

Les erreurs résiduelles dues au mouvement, la consistance visuelle et les lissages spatial et temporel peuvent maintenant être considérés ensemble pour extraire la partition optimale des N images :

$$\begin{aligned} E(\omega) &= \sum_{i=t}^{t+N-1} \left[(1-\alpha) E_m^i(\omega) + \alpha E_v^i(\omega) + \beta E_s^i(\omega) \right] \\ &+ \gamma (E_{ts}(\omega) + E_{ts'}(\omega)) \end{aligned} \quad (22)$$

La minimisation de la forme discrète de l'équation 22 peut être déterminée en utilisant plusieurs techniques de complexité diverses comme l'ICM (iterated conditional modes [4]), l'HCF (highest confidence first [11]), le recuit simulé ("simulated annealing" [15]) et l'approche par graphe "Coupe minimale - Flot maximal" ("graph cut" [8]). En raison de son efficacité, l'approche par graphe a été retenue pour déterminer la solution optimale au problème d'étiquetage.

5 Détails de l'implementation

Pour l'énergie (22), calculer directement la solution optimale n'est pas possible en pratique. En effet, ce problème d'étiquetage est NP-complet et il n'existe donc pas d'algorithme de complexité polynomial qui calcule la solution optimale. Cependant, l'algorithme d'α-expansion [8] permet d'obtenir rapidement une approximation satisfaisante de la solution optimale de cette énergie. Son implémentation suit la méthode proposée dans [18]. Suivant les notations de [8], les poids des *t-links* sont déterminés via la fonctin de coût $D_p(i)$ ($i \in [1, n]$) :

$$D_p(i) = (1-\alpha) \underbrace{\left[\log \sigma_i + \frac{{r_i}^2(p)}{2{\sigma_i}^2} \right]}_{\text{terme mouvement}} + \alpha \underbrace{p_i(I(p))}_{\text{terme visuel}} \quad (23)$$

auquel on rajoute $\mathcal{V}^{\mathbf{t}}(i)$ si l'on considère le lissage temporel séquentiel.
Si le critère de lissage temporel simultané $E_{ts'}$ est considéré, le graphe de l'alpha-expansion est transformé en graphe 3D, avec N niveaux dans la troisième dimension si N images sont considérées. On rajoute alors des arêtes d'un niveau à l'autre pour prendre en compte la contrainte de lissage temporel, tel qu'illustré par la figure 3 : si l'on considère le cas de deux couches $\mathcal{L}_1$ (représentée par la source S) et $\mathcal{L}_2$ (représentée par le puit T), de modèles de mouvement associés $\mathcal{A}_1$ et $\mathcal{A}_2$, le pixel p de coordonnées $\mathbf{x}$ peut être projeté soit vers $p1$ de coordonnées $\mathbf{x} + \mathcal{A}_1(\mathbf{x})$ ou vers $p2$ de coordonnées $\mathbf{x} + \mathcal{A}_2(\mathbf{x})$. On ajoute ainsi deux arêtes orientées $p \rightarrow p1$ et $p2 \rightarrow p$ de fort poids. Ainsi, si l'arête $s \rightarrow p$ est coupée (ce qui signifie que p appartient à la couche $\mathcal{L}_2$) alors on pénalise la coupe de l'arête $p \rightarrow p1$ qui correspondrait au cas où la continuité temporelle $\omega(\mathbf{x}; t) = \kappa \Rightarrow \omega(\mathbf{x} + \mathcal{A}_\kappa(\mathbf{x}); t+1) = \kappa$ n'est pas respectée. Et inversement si l'arête $p \rightarrow t$ est coupée.

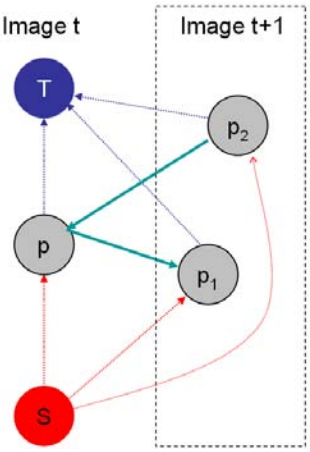

FIG. 3 – Lissage temporel : en vert, les arêtes orientées ajoutées pour prendre en compte la contrainte de lissage. En rouge, les arêtes en provenance de la source S et en bleu, celles vers le puit T. Les pixels $p1$ et $p2$ sont les projections du pixel p dans l'image $t+1$ par les deux modèles de mouvement respectivement $\mathcal{A}_1$ et $\mathcal{A}_2$.

5.1 Aperçu général du programme

Les principaux modules du programme ont été présentés. On détaille maintenant le déroulement général du programme. Pour les N premières images de la séquence, on partitionne leurs domaines d'image de façon régulières en $\mathcal{N}$ blocs (en général, 16 ou 25 blocs). Ces blocs sont utilisés en tant que supports initiaux à partir desquels on estime le mouvement et les propriétés visuelles de chacun d'entre eux pour les N images considérées. L'α-expansion fournit alors une nouvelle partition en régions selon les critères considérés (mouvement, couleurs et contraintes de lissages). Le processus est répété jusqu'à convergence pour les N images. Une fois que les supports des N images sont correctement déterminés, l'initialisation des supports des régions d'une nouvelle image $I(;t)$, $t > N$, est effectuée en projetant les supports de l'image précédente via les

modèles de mouvement estimés dans l'image courante. La contrainte temporelle $E_{ts'}$ est alors ignorée et on n'utilise exclusivement que la contrainte temporelle séquentielle E_{ts} ($N = 1$) (d'une facon générale, si $N = 1$, $E_{ts'} = 0$ et que si $N > 1$, $E_{ts} = 0$).
Une étape critique est l'estimation du nombre de couches : le processus étant initialisé avec un grand nombre de couches, ces dernières seront alors progressivement fusionnées selon le critère suivant. On fusionne deux couches $\mathcal{L}_i$ et $\mathcal{L}_j$ si leurs mouvements sont similaires pour les N images considérées. Comme les paramètres du modèle affine ne définissent pas de façon unique le mouvement sur toute la région, plutôt que de les comparer directement, on considère les flots optiques générés par $\mathcal{A}_i$ et $\mathcal{A}_j$. Ainsi, en utilisant les notions introduites en section 3, le critère de similarité de mouvement s_{ij} entre deux couches $\mathcal{L}_i$ et $\mathcal{L}_j$ s'écrit :

$$s_{ij} = \max_{t\in[1,N]} \left[\frac{1}{|\mathcal{S}_i|} \int_\Omega (\mathcal{A}_i(\mathbf{x}\ t) - \mathcal{A}_j(\mathbf{x};t))^2 \mathbf{1}_{\mathcal{S}_i^t(\mathbf{x})} d\mathbf{x} + \frac{1}{|\mathcal{S}_j|} \int_\Omega (\mathcal{A}_i(\mathbf{x};t) - \mathcal{A}_j(\mathbf{x};t))^2 \mathbf{1}_{\mathcal{S}_j^t(\mathbf{x})} d\mathbf{x} \right]$$

où $|\mathcal{S}_i| = \int_\Omega \mathbf{1}_{\mathcal{S}_i^t(\mathbf{x})} d\mathbf{x}$ et $\mathbf{1}_{\mathcal{S}_i^t}$ est la fonction caractéristique de la région i de l'image t. Si s_{ij} descend sous un seuil donné, les deux couches sont fusionnées pour les N images. De plus, les couches dont leur support est trop petit (et qui entraîne en général une mauvaise estimation du mouvement) sont supprimées.

6 Validation et conclusion

6.1 Comparatifs avec séquences synthétiques

On génère plusieurs séquences synthétiques composées de trois couches texturées ou non, de supports rectangulaires et de mouvements affines propres (cf. figure 5).
Le tableau 1 compare les erreurs de classifications obtenues selon les divers critères proposés dans cet article : avec ou sans le critère de lissage temporel $\mathcal{V}^{\mathbf{t}}$, de lissage spatial pondéré par le gradient $\mathcal{V}_g$ et de l'apparence visuelle E_v. On notera l'importance de ce dernier terme lorsque la séquence comporte des objets non texturés qui permet de lever les ambiguïtés dues au mouvement seul. Notons que le critère E_v n'est pas considéré pour les séquences pleinement texturées, aucun gain de qualité n'ayant été observé. Notons aussi que l'erreur importante constatée pour les séquences peu texturées sans l'utilisation du critère E_v est due au fait que la région peu texturée a été assimilée à l'arrière-plan lors de la classification à cause de la forte ambiguïté.

6.2 Résultats expérimentaux

La validation de la technique et des critères présentés a été effectuée sur deux séquences classiques (figures 6 et 4) afin de permettre une comparaison directe avec la literrature existante. Dans la figure 6, si le calendrier et le train sont bien segmentés, la balle reste cependant sur-segmentée en raison d'un manque de texture et des couleurs similaires à celles de l'arrière-plan et est ainsi classée dans la même couche que celle du train. Pour la séquence de la figure 4, la première image est sur-segmentée mais les images suivantes possèdent le bon nombre de couches. L'arrière-plan est bien distingué du plan intermédiaire (maison et le parterre de fleurs). Le critère d'apparence visuelle E_v a permis de contourner les ambiguïtés dues au manque de textures notamment dans le ciel. Cependant, les branches de l'arbre ne sont pas toujours classées dans la bonne couche. En effet, en raison de la faible distance entre l'arbre et la caméra, comme les branches n'appartiennent pas au

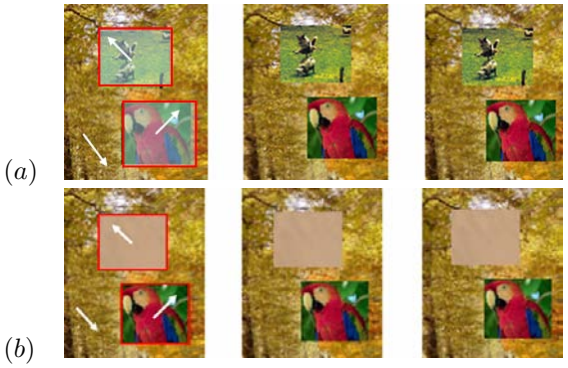

FIG. 5 – Exemples de séquences synthétiques : les 3 couches sont ici délimitées par une bordure rouge. Leurs mouvements sont approximativement indiqués par une flèche blanche. Les séquences (b) sont similaires aux séquences (a) à ceci près, une des trois couches est remplacée par une couche peu texturée.

Paramètres des critères	Err. moyenne
seq (a) : $\alpha = 0, \beta = 1(\mathcal{V}), \gamma = 0$	4.72 %
seq (a) : $\alpha = 0, \beta = 1(\mathcal{V}), \gamma = 1$	4.37 %
seq (a) : $\alpha = 0, \beta = 0.5(\mathcal{V}_g), \gamma = 1$	3.24 %
seq (b) : $\alpha = 0, \beta = 1(\mathcal{V}_g), \gamma = 1$	16 %
seq (b) : $\alpha = 0.3, \beta = 1(\mathcal{V}_g), \gamma = 1$	4.4 %

TAB. 1 – Influence de chaque critère sur la classification. Colonne de droite : proportion des pixels mal classés. La première série ("seq (a)") permet de comparer l'influence du lissage temporel simultané ($N = 5$) et du lissage spatial pondéré ou non par le gradient ($\mathcal{V}_g$) sur les séquences synthétiques texturées (fig. 5a). La seconde série ("seq (b)") permet de comparer l'influence du terme de l'apparence visuelle E_v sur les séquences avec couches peu texturées (fig. 5b).

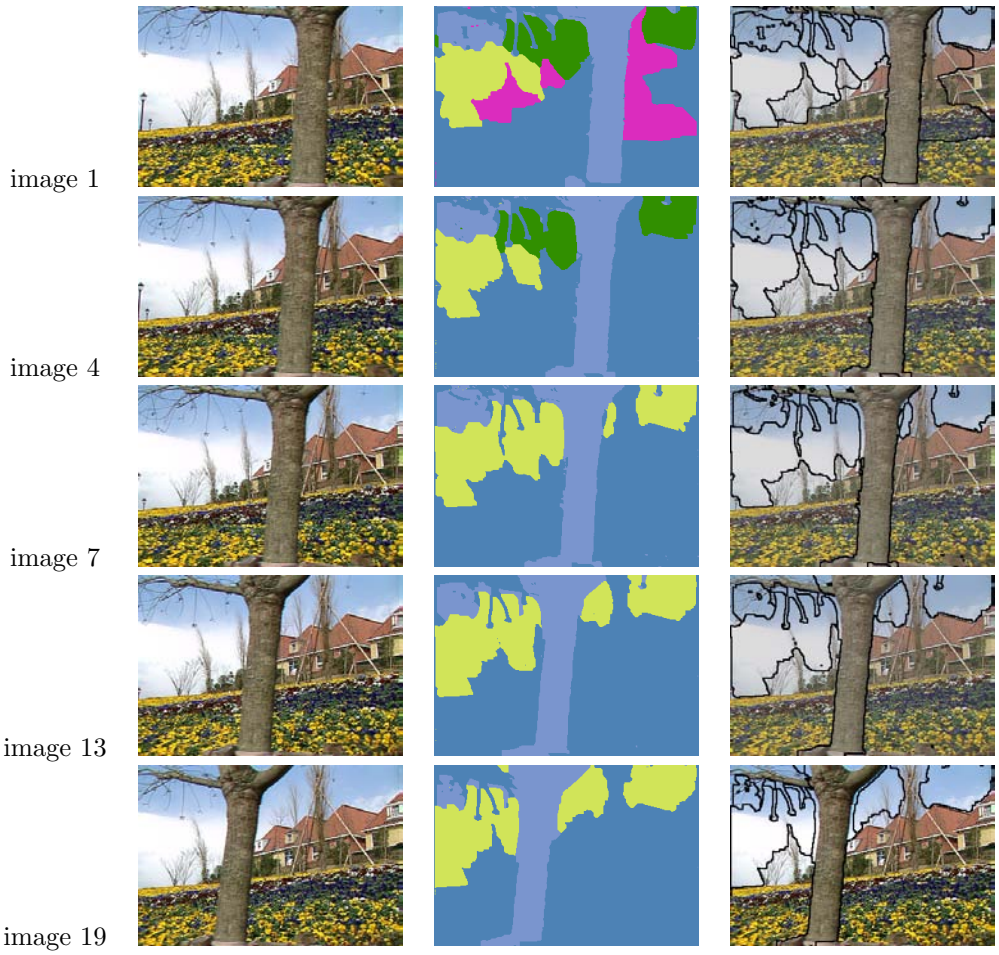

FIG. 4 – Résultats sur la séquence "Flower garden". Première colonne, séquence originale ; seconde colonne, les couches extraites ; troisième colonne, frontières des couches. Principaux paramètres : $\alpha = 0.25$, $\gamma = 1$, $N = 1$ (avec E_{ts} mais sans $E_{ts'}$).

même plan 3D (la profondeur relative est importante au regard de la profondeur de l'arbre), leurs mouvements ne peuvent être correctement représentés avec le même modèle affine que celui du tronc de l'arbre.
La figure 7 présente les résultats obtenus en faisant varier l'influence du critère de lissage temporel simultané $E_{ts'}$. Outre le fait d'avoir une continuité dans les supports des couches à mesure que γ augmente, on note que l'extraction des couches gagne en robustesse. On notera aussi que les deux segmentations de la séquence "flower garden" (fig. 4 et 7), bien que différentes, sont acceptables selon que l'on considère que les maisons appartiennent au plan du parterre de fleurs ou à l'arrière-plan.

6.3 Conclusion

Dans cet article, on a présenté une technique pour extraire les couches composant une scène à partir des modèles de mouvements paramétriques. L'utilisation conjointe d'une estimation robuste des mouvements avec les critères additionnels comme l'apparence visuelle et les contraintes de lissage spatial et temporel permet d'obtenir une segmentation en couches efficace. Les résultats expérimentaux [FIG. 6, 4 et 7] sont prometteurs et démontrent le potentiel de la méthode proposée. L'infuence des divers critères a également été étudiée, montrant notamment l'importance des contraintes temporelles dans l'estimation des couches. D'autres contraintes exploitant la dimension temporelle notamment en ce qui concerne les occlusions sont ainsi à l'étude.

De plus, si l'algorithme détermine rapidement les supports de couches (en moyenne, autant d'itérations sont necessaires que de couches à extraire), la complexité en terme de calculs reste toujours l'une des limitations de l'approche proposée, notamment en ce qui concerne l'estimation du mouvement pour chaque support lorsque ceux-ci sont mal encore définis. Une implémentation en hardware de notre méthode sur un processeur graphique (GPU) est en cours d'investigation.

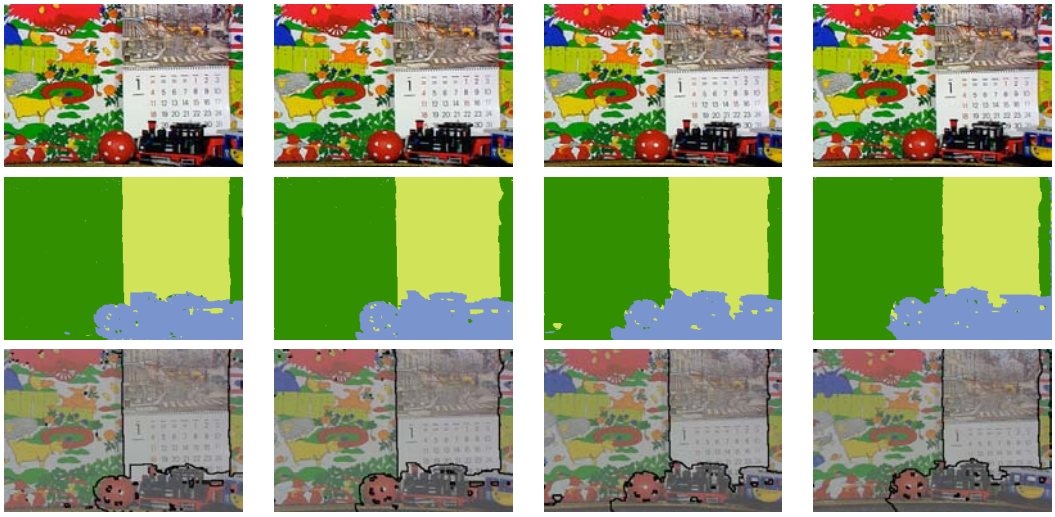

FIG. 6 – Résultats sur la séquence du calendrier (1 image sur 4 est considérée). Chaque colonne représente une image : sont représentées ici les images 0,12,24,36. Première ligne, séquence originale ; seconde ligne, les couches extraites ; troisième ligne, superposition des frontières avec la séquence originale. Ici, il n'y a pas de lissage temporel simultané $E_{ts'}$ ($\gamma = 1$, $N = 1$), on se concentre sur l'influence du critère E_v

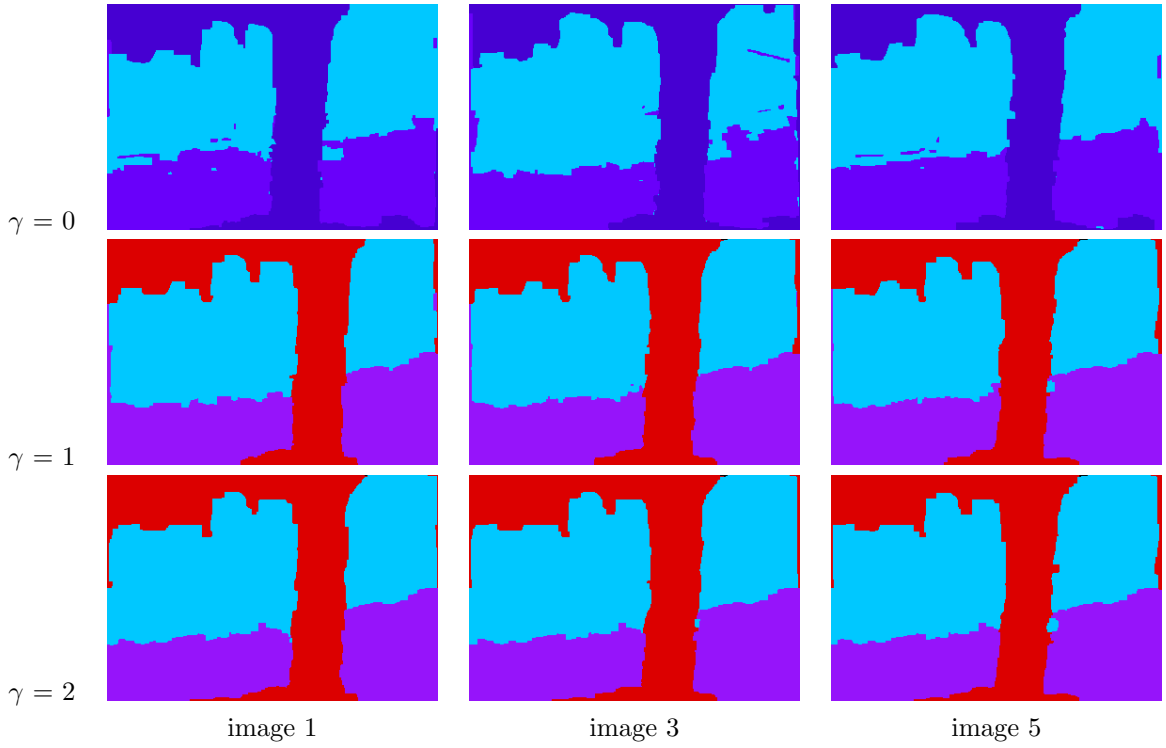

FIG. 7 – Résultats sur les images 1 3 et 5 de la séquence du flowers garden selon diverses valeurs de γ : on considère ici N=5 images simultanément et on ne prend pas en compte le terme d'apparence visuelle E_v ($\alpha = 0$ et $\beta = 1$).

Remerciements

Les auteurs voudraient remercier Oliver Juan pour avoir fourni l'algorithme EM utilisé pour estimer les mixtures de gaussiennes et Yuri Boykov de l'University of Western Ontarion pour les discussions enrichissantes.

Références

[1] S. Ayer and H. Sawhney. Layered Representation of Motion Video Using Robust Maximum-

Likelihood Estimation of Mixture Models and MDL Encoding. In *IEEE International Conference in Computer Vision*, pages 777–784, 1995.

[2] J.L. Barron, D.J. Fleet, S.S. Beauchemin, and T.A. Burkitt. Performance of optical flow techniques. *Computer Vision and Pattern Recognition (CVPR)*, pages 236–242, 1992.

[3] James R. Bergen, P. Anandan, Keith J. Hanna, and Rajesh Hingorani. Hierarchical model-based motion estimation. In *European Conference on Computer Vision (ECCV)*, pages 237–252. Springer-Verlag, 1992.

[4] J. Besag. On the statistical analysis of dirty images. *Journal of Royal Statistics Society*, 48 :259–302, 1986.

[5] Michael J. Black and P. Anandan. The robust estimation of multiple motions : parametric and piecewise-smooth flow fields. *Comput. Vis. Image Underst.*, 63(1) :75–104, 1996.

[6] Michael J. Black and Allan D. Jepson. Estimating optical flow in segmented images using variable-order parametric models with local deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(10) :972–986, 1996.

[7] Patrick Bouthemy and Edouard Francois. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *Int. J. Comput. Vision*, 10(2) :157–182, 1993.

[8] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 :1222–1239, 2001.

[9] Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *ICCV*, pages 105–112, 2001.

[10] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *8th European Conference on Computer Vision (ECCV)*, pages 25–36, 2004.

[11] P. Chou and C. Brown. The theory and practice of bayesian image labeling. *International Journal of Computer Vision*, 4 :185–210, 1990.

[12] Daniel Cremers and Stefano Soatto. Variational space-time motion segmentation. In *International Conference on Computer Vision (ICCV)*, pages 886–893, 2003.

[13] Trevor Darrell and Alex P. Pentland. Cooperative robust estimation using layers of support. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1995.

[14] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[15] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.

[16] B. Horn and B. Schunck. Determinating Optical Flow. *Artificial Intelligence*, 17 :185–203, 1981.

[17] Michal Irani and P. Anandan. A unified approach to moving object detection in 2d and 3d scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(6) :577–589, 1998.

[18] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts ? In *European Conference on Computer Vision-Part (ECCV)*, pages 65–81, 2002.

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

[20] Jean-Marc Odobez and Patrick Bouthemy. Direct incremental model-based image motion segmentation for video analysis. *Signal Processing*, 66(2) :143–155, 1998.

[21] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Ann. Stat.*, 11 :416–431, 1983.

[22] Allan D. Jepson Shanon X. Ju, Michael J. Black. Skin and bones : Multi-layer, locally affine, optical flow and regularization with transparency. In *Computer Vision and Pattern Recognition (CVPR)*, page 307, 1996.

[23] J. Wang and E. Adelson. Representing Moving Images with Layers. *IEEE Transactions on Image Processing*, 3 :625–638, 1994.

[24] Yair Weiss. Smoothness in layers : Motion segmentation using nonparametric mixture estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 520. IEEE Computer Society, 1997.

[25] Jiangjian Xiao and Mubarak Shah. Motion layer extraction in the presence of occlusion using graph cut. In *CVPR (2)*, pages 972–979, 2004.

[26] Ramin Zabih and Vladimir Kolmogorov. Spatially coherent clustering using graph cuts. In *Computer Vision and Pattern Recognition (CVPR)*, pages 437–444, 2004.