Dix-septième colloque GRETSI, Vannes, 13-17 septembre 1999

Classification de mouvement des objets

Jean MOTSCH[1], Henri NICOLAS[1]

[1]IRISA/INRIA, Projet TEMICS
Campus universitaire de Beaulieu
35042 RENNES CEDEX, FRANCE
Jean.Motsch@irisa.fr, Henri.Nicolas@irisa.fr

Résumé – Le développement de la norme MPEG-4 met en avant l'importance des manipulations de vidéo basées objet telles que la composition de scène vidéo 2D. Le développement d'algorithmes semi-automatiques permettant de réaliser ce type d'applications nécessite l'utilisation d'informations pertinentes sur le comportement des objets vidéo dans la séquence originale. Pour cela, cet article propose une méthode pour découper une séquence en segments temporels où chaque segment correspond à une orientation objet caméra constante ou non. Ensuite, une vue clef peut être extraite pour chaque segment. Cette décomposition d'une séquence d'objet vidéo est basée sur l'étude et la représentativité de chacun des paramètres d'un modèle affine de mouvement.

Abstract – MPEG-4 puts video object based manipulation forward, such as 2D video scene composition. Semi-automatic algorithms designed to perform that task use meaningful informations about the behaviour of the video object in the original sequence. This article proposes a method to split the sequence in temporal segments where the orientation object–camera is constant. Then, a key view can be extracted for each segment. This video object sequence's decomposition relies on the meaning of the affine motion model parameters.

1 Introduction

MPEG-4 a récemment introduit le concept d'objet vidéo [2, 9] dans le cadre de la normalisation. Aux formes de blocs de MPEG-2 succèdent donc les objets de forme quelconque. L'objet vidéo est constitué de l'ensemble de ses projections temporelles, appelées objets vidéo plans.

Cette vision des séquences vidéo introduit de façon naturelle le développement d'applications liées à la manipulation, l'édition et la composition de séquences vidéo. Le principal secteur d'activité concerné est celui de la post-production de vidéo numérique.

Un exemple d'utilisation des objets vidéo consiste à mélanger des objets provenant de différentes sources vidéo [7]. Des applications de ce type existent déjà, soit professionnelles, soit grand public [10], mais elles utilisent principalement la manipulation des objets vidéo par un opérateur et restent proches du trucage traditionnel.

La littérature existante présente de nombreuses méthodes automatiques ou semi-automatiques permettant d'extraire plus ou moins précisement les objets vidéo de leur séquence d'origine [1, 5]. Par contre, à l'exception des méthodes manuelles, peu de méthodes ont été proposées permettant de les manipuler efficacement dans le cadre applicatif mentionné précédemment.

Afin de faciliter la manipulation des objets vidéo, il est utile de posséder des informations qualitatives et quantitatives fiables sur le comportement des objets vidéo manipulés.

Les éléments pouvant servir à caractériser les objets vidéo sont nombreux: forme, texture, couleur ou mouvement apparent. Le mouvement apparent, ou sa représentation par un modèle paramétrique, permet l'extraction d'informations discriminantes sur l'évolution temporelle d'un objet vidéo [3]. En effet, il est possible de relier partiellement le mouvement tridimensionnel réel aux paramètres d'un modèle de mouvement 2D. Ainsi, le suivi de l'évolution des paramètres de mouvement d'un objet peut permettre une classification typologique du mouvement de l'objet, et donc l'obtention d'une segmentation temporelle de la séquence d'objets vidéo.

Pour cela, il peut être intéressant de dissocier deux aspects: d'une part, les plans pendant lesquels l'orientation objet/caméra est constante de ceux pour lesquels l'orientation relative de l'objet par rapport à la caméra se modifie. Cette dernière phase correspond à des mouvements complexes, comme des rotations non planes, difficiles à représenter avec un modèle paramétrique simple. Enfin, une représentation sous forme de graphe d'état permet alors de synthétiser la séquence d'objet vidéo, comme le montre la figure 1.

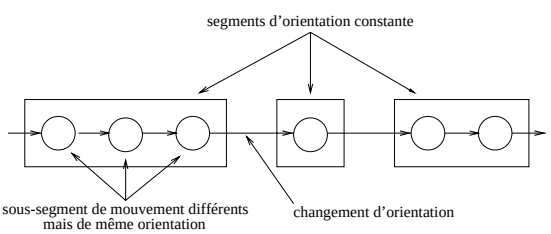

FIG. 1: *représentation par graphe d'état d'un objet vidéo*

Dans ce contexte, nous présentons ici une méthode pour découper la séquence vidéo en utilisant les paramètres de mouvement affine. Cette méthode repose sur une estimation paramétrique contrainte suivie d'un test de vraisem-

blance sur chaque coefficient du modèle et permet de déduire les composantes du mouvement réellement significatives.

2 Décomposition de la séquence de VOPs

La décomposition en segment d'orientation constante ou non est effectuée en analysant le mouvement paramétrique affine estimé sur le support du VOP, tout en supposant que les mouvements plans (translation, divergence, rotation plan) ne modifient pas l'orientation de l'objet par rapport à la caméra. Elle repose sur le caractère significatif ou non des coefficients d'un modèle affine de mouvement et elle s'applique en deux temps. D'abord, choisir le modèle le plus simple « globalement » pour décrire le mouvement de l'objet. Ensuite, chaque coefficient du modèle sélectionné est étudié sur la base d'une estimation contrainte et d'un critère de vraisemblance.

2.1 Du mouvement 3D aux paramètres affines

Considérons le mouvement rigide d'un objet plan d'équation $Z = Z_0 + Z_X X + Z_Y Y$ perçu par une caméra sténopée. En négligeant les termes d'ordre 2 et plus, le mouvement des points $\mathbf{p}$ de l'objet peut s'écrire

$$\boldsymbol{d_\theta}(\mathbf{p}) = \begin{pmatrix} t_x \\ t_y \end{pmatrix} + \begin{pmatrix} k+h_1 & h_2-\theta \\ h_2+\theta & k-h_1 \end{pmatrix} \begin{pmatrix} x-x_g \\ y-y_g \end{pmatrix}$$

où (x_g, y_g) est la position du centre de gravité, et $\Theta = (t_x, t_y, k, \theta, h_1, h_2)$ représente la translation (t_x, t_y), la divergence k, la rotation plan θ et les termes hyperboliques h_1 et h_2 [6].

Le modèle affine permet de décrire correctement les mouvements plans. Dans le cadre des hypothèses énoncées précédemment, les paramètres du mouvement affine dépendent du mouvement réel tridimensionnel de la façon suivante

$$\begin{aligned} t_x &= f\frac{U}{Z_0} + fR_Y \\ t_y &= f\frac{W}{Z_0} - fR_X \\ k &= -\frac{W}{Z_0} - \frac{U}{2Z_0}(Z_X + Z_Y) \\ \theta &= R_Z - \frac{VZ_X + UZ_Y}{2Z_0} \\ h_1 &= \frac{U}{2Z_0}(Z_Y - Z_X) \\ h_2 &= -\frac{UZ_Y + VZ_X}{2Z_0} \end{aligned}$$

où U,V et W sont les valeurs des translations selon X, Y et Z respectivement, R_X, R_Y et R_Z sont les coefficients de rotation, et f la focale de la caméra.

Supposons l'objet parallèle au plan image. Z_X et Z_Y sont nuls et seuls les paramètres t_x, t_y, k et θ sont présents. En revanche, dès que l'objet modifie son orientation par rapport à la caméra, les termes Z_X et Z_Y sont non nuls et h_1 et h_2 sont également présents. Ces deux derniers paramètres peuvent donc servir à détecter un changement d'orientation de l'objet considéré par rapport à la caméra.

2.2 Estimation du mouvement paramétrique

Les paramètres des modèles sont estimés de façon indépendante en utilisant la méthode décrite dans [8], pour le modèle en translation ($T = (t_x, t_y)$), le modèle affine simplifié ($AS = (tx, ty, k, \theta)$) et le modèle complet (A).

L'estimation de mouvement repose sur une approche incrémentale multi-résolution robuste qui autorise la prise en compte de grands déplacements. Elle utilise la minimisation de l'erreur quadratique de reconstruction sur un support d'estimation défini par l'approche robuste. Ainsi, le support réel d'estimation, version seuillée de la carte de pondération, n'est pas le même pour tous les modèles estimés.

La sélection du modèle (T, AS ou A) se fait sur la base du minimum de résidu après estimation. Pour assurer une compétition correcte entre les modèles, les tailles des supports réels doivent être à peu près similaires, pour éviter des modèles estimés sur insuffisamment de points. Cette sélection évite de prendre systématiquement le modèle complet. En effet, l'estimateur utilise alors tous les degrés de liberté du modèle, et du bruit peut être intégré dans les coefficients. La comparaison des modèles permet de circonvenir à cette situation.

2.3 Signification des coefficients

L'étape précédente fournit le nombre maximum de paramètres (modèle à 2, 4 ou 6 coefficients) ainsi qu'un support d'estimation réel, différent du support de l'objet vidéo. Il s'agit maintenant d'étudier la signification de chaque paramètre du modèle choisi. En effet, les paramètres non significatifs ne sont pas nécessairement nuls, en particulier en raison du bruit et des erreurs d'estimation. Pour prendre une décision sur la pertinence ou non des paramètres, nous utilisons une approche statistique basée sur des tests de vraisemblance [4]. Chaque paramètre du modèle est testé en considérant deux hypothèses. La première, H_0, suppose que la composante testée est significative, tandis que la seconde, H_1, suppose qu'au contraire, elle est nulle. Les autres paramètres du modèle de mouvement affine ne sont pas considérés.

Soient $\boldsymbol{\theta_{m_0}}$ et $\boldsymbol{\theta_{m_1}}$ les vecteurs des paramètres associés à H_0 et H_1 respectivement. Il s'agit d'estimer les deux vecteurs de paramètres correspondant aux deux hypothèses: $\boldsymbol{\theta_{m_0}}$ est le modèle complet (à 2, 4 ou 6 coefficients) tandis que $\boldsymbol{\theta_{m_1}}$ est le modèle pour lequel la composante à tester est nulle (donc à 1, 3 ou 5 paramètres). Étant donné que le support d'estimation est connu, l'estimation est effectué par moindres carrés, tout en gardant l'aspect multi-résolution. Cette simplification allège grandement la charge de calcul.

Pour chaque hypothèse, la valeur de la fonction de vraisemblance (f) est calculée. En supposant les résiduels indépendants et gaussiens centrés, on peut montrer que le

rapport de vraisemblance se met sous la forme :

$$\ln \frac{f(\boldsymbol{\theta}_{\boldsymbol{m}_1})}{f(\boldsymbol{\theta}_{\boldsymbol{m}_0})} = \left(\frac{\sigma^2_{m_0}}{\sigma^2_{m_1}} \right)^{n/2} \quad (1)$$

où $\sigma^2_{m_l}$ est la variance estimé a posteriori correspondant aux vecteurs de paramètres $\boldsymbol{\theta}_{\boldsymbol{m_l}}$, et n est le nombre de points du support d'estimation.

Pour décider de l'hypothèse retenue, on utilise le test de log-vraisemblance suivant :

$$\ln \frac{f(\boldsymbol{\theta}_{\boldsymbol{m}_1})}{f(\boldsymbol{\theta}_{\boldsymbol{m}_0})} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda \quad (2)$$

Si ce rapport est plus faible qu'un seuil λ, la composante est déclarée significative, sinon elle est considérée nulle. Pour une composante significative, le signe de la valeur du coefficient peut fournir une information supplémentaire sur le sens du mouvement (à droite ou à gauche, se rapproche ou s'éloigne, etc.).

Ces test sont effectués sur chaque composante de $\boldsymbol{\theta}$. Ainsi, on peut réduire le nombre de paramètres de mouvement et ne conserver que ceux représentatifs du type de mouvement de l'objet vidéo considéré.

2.4 Aspect temporel

Afin d'assurer une homogénéité temporelle, un filtrage médian temporel est effectué sur chaque composante du modèle de mouvement. Cela revient à estimer que si une composante est significative entre les instants $t-1$ et t, et les instants $t+1$ et $t+2$, elle l'est aussi entre les instants t et $t+1$.

2.5 VOP-clef

Une représentation compacte et efficace de la séquence d'objet vidéo peut être construite à partir de la décomposition obtenue par la méthode précedente. Pour chaque état de cette représentation, une vue clef est extraite. Dans les résultats présentés dans la section 3, la vue clef est l'objet situé au milieu des segments découpés. Le choix de l'objet vidéo clef peut aussi être basé sur des critères dépendants de l'application visée, comme par exemple, l'objet ayant la meilleure résolution, l'objet ayant une taille donnée, etc.

3 Expérimentations et résultats

La méthode de classification proposée a été mise en œuvre sur des séquences synthétiques et naturelle. Les séquences synthétiques ont été obtenues par lancé de rayons, en utilisant le mouvement d'un cube texturé. La séquence naturelle est « hall monitor ». Toutes sont au format CIF.

D'abord, nous présentons les résultats obtenus avec des séquences synthétiques pour lesquelles le mouvement est simple : translation, translation et arrêt, divergence et rotation plan. Les résultats sont résumés dans les tableaux 1, 2, 3 et 4, respectivement. L'expérimentation indique que la discrimination est plus aisée en terme de variance pour les mouvements de translation que pour les mouvements de rotation plan et de divergence. Ceci est dû principalement à la petitesse des termes du modèle de mouvement. Une solution pour pallier cette insensibilité pourrait être de considérer des paires d'images plus éloignées temporellement.

paire	1–2	2–3	3–4	4–5	5–6
	6–7	7–8	8–9	9–10	10–11
modèle	T	T	A	A	A
	T	A	A	A	A
mouvement	t_x	t_x	t_x	t_x	t_x
	t_x	t_x	t_x	t_x	t_x

TAB. 1: *mouvement de translation*

paire	1–2	2–3	3–4	4–5	5–6
	6–7	7–8	8–9	9–10	10–11
modèle	T	T	A	A	A
	T	A	A	A	A
mouvement	t_x	t_x	t_x	t_x	
		t_x	t_x	t_x	t_x

TAB. 2: *mouvement de translation et arrêt*

paire	1–2	2–3	3–4	4–5	5–6
	6–7	7–8	8–9	9–10	10–11
modèle	A	AS	A	AS	A
	AS	A	A	AS	AS
mouvement	DIV	DIV	DIV	DIV	DIV
	DIV	DIV	DIV	DIV	DIV

TAB. 3: *mouvement de divergence*

Le tableau 5 présente les résultats obtenus avec une rotation qui fait changer l'orientation relative objet caméra. Il apparaît que, si la présence des termes hyperboliques permet de repérer le changement d'orientation, elle n'est pas stable sur toute la séquence. Cela provient du fait que le seuil λ est le même pour toutes les composantes. Des expérimentations sont en cours pour rendre adaptatif le choix du seuil.

Enfin, des résultats sur des séquences synthétiques plus longues sont présentés sur les figures 2 et 3 où les valeur de signification (0 ou 1) des 6 paramètres de mouvement sont présentées après la phase de comparaison. Pour la première séquence, trois segments ont été détectés, et ils correspondent à la trajectoire réelle de l'objet, ie. mouvement plan simple, orientation non constante et translation. Pour la deuxième séquence, trois segments ont aussi été détectés, dont un segment rotation. Pour la séquence réelle « hall monitor » , l'interprétation est difficile mais les termes hyperboliques détectés correspondent à un changement d'orientation dans la scène.

4 Conclusion et perspectives

L'utilisation directe des coefficients des modèles affine de mouvement permet de segmenter une séquence d'objet vidéo. Elle autorise également la classification des différents segments obtenus.

paire	1–2	2–3	3–4	4–5	5–6
	6–7	7–8	8–9	9–10	10–11
modèle	A	A	A	A	A
	A	A	A	A	A
mouvement	ROT	ROT	ROT	ROT	ROT
	ROT	ROT	ROT	ROT	ROT

TAB. 4: *mouvement de divergence*

paire	1–2	2–3	3–4	4–5	5–6
	6–7	7–8	8–9	9–10	10–11
modèle	A	AS	A	AS	A
	AS	A	A	AS	AS
mouvement	t_x	t_x DIV	t_x DIV h_1	t_x D R h_{12}	t_x DIV
	t_x D R		t_x ROT	t_x DIV h_1	t_x

TAB. 5: *mouvement de « roll »*

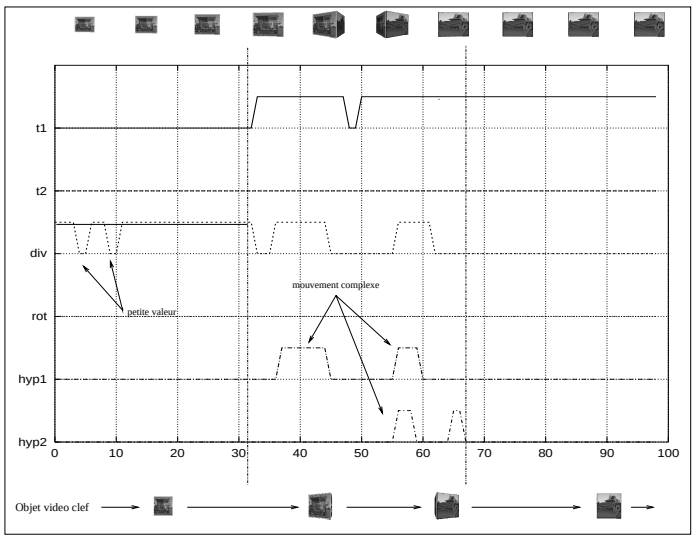

FIG. 2: *résultats sur la première séquence « cube mobile ».*

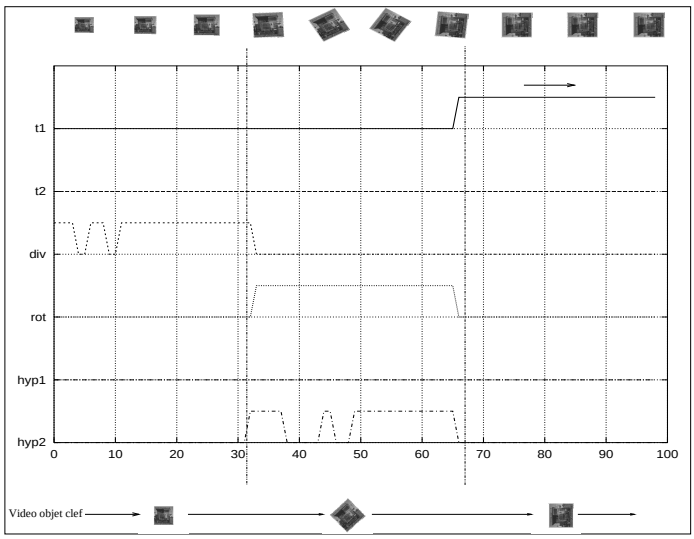

FIG. 3: *résultats sur la deuxième séquence « cube mobile ».*

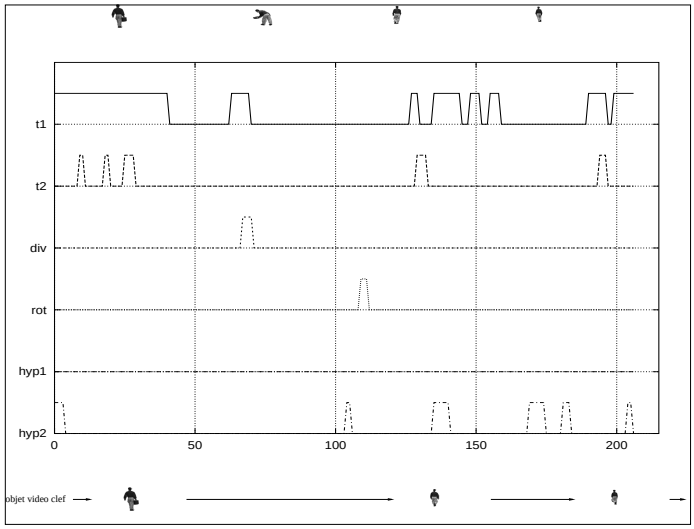

FIG. 4: *résultats sur la séquence « hall monitor ».*

La principale perspective de ce travail est l'utilisation d'un horizon temporel plus large afin de pallier la détection de petits mouvements parasites. L'ajout d'un critère relevant de la texture peut également permettre la détection des changements abruptes d'orientation.

Références

[1] Castagno (R.), Ebrahimi (T.) et Kunt (M.). – Video segmentation based on multiple features for interactive multimedia applications. vol. 8, n° 5, september 1998, pp. 562–571.

[2] Chiariglione (L.). – MPEG and multimedia communications. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, n° 1, february 1997, pp. 5–18.

[3] Courtney (J.). – Automatic video indexing via object motion analysis. *Pattern Recognition*, vol. 30, n° 4, 1997, pp. 607–625.

[4] François (E.) et Bouthemy (P.). – Derivation of qualitative information in motion analysis. *Image and Vision Computing*, vol. 8, n° 4, november 1990, pp. 279–287.

[5] Guo (J.), Kim (J.) et Kuo (C.-C.). – Fast and accurate moving object extraction technique for mpeg-4 object-based video coding. *In : Visual Conference on Image Processing*, pp. 1210–1221. – january 1999.

[6] Nicolas (H.) et Labit (C.). – Global motion identification for image sequence analysis and coding. *In : Proc. of the Intl. Conf. on Acoutic Speech and Signal Processing*, pp. 2825–2828. – may 1991.

[7] Nicolas (H.) et Motsch (J.). – Very low bitrate coding using hybrid synthetic/real images for multi-sites videoconference applications. *In : Visual Conference on Image Processing*. SPIE, pp. 1330–1341. – 1997.

[8] Odobez (J.-M.) et Bouthemy (P.). – *Direct incremental model-based image motion segmentation for video analysis*. – Rapport technique n° 1129, IRISA, septembre 1997.

[9] Sikora (T.). – The MPEG-4 video standard verification model. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, n° 1, february 1997, pp. 19–31.

[10] Simon (J.). – Commotion 1.5 – retouche vidéo. *Univers Mac*, no83, octobre 1998, pp. 160–162.