Modélisation du mouvement global sur une base d'ondelettes : application à l'indexation de séquences vidéo

Eric Bruno, Denis Pellerin

Laboratoire des Images et des Signaux (LIS)
INPG, 46 Av. Félix Viallet 38031 Grenoble Cedex, France
bruno,pellerin@lis-viallet.inpg.fr

Résumé – Ce papier décrit un nouveau modèle de mouvement global basé sur les ondelettes B-splines. L'utilisation d'un tel modèle permet d'estimer directement le mouvement global suivant un schéma multirésolution, et cela sans segmentation *a priori* de la scène. Ce modèle permet d'estimer des flots optiques de façon fiable et robuste. Les coefficients d'ondelettes des niveaux de résolution les plus faibles permettent aussi de définir des descripteurs du mouvement global pour indexer des vidéo par leur contenu dynamique. Nous avons testé notre approche sur des vidéos représentant différentes activités humaines. Les résultats obtenus nous montrent que les descripteurs de mouvement sont pertinents pour classer les séquences d'images selon le type d'activité qu'elles contiennent

Abstract – This paper describes a framework to estimate global motion model based on B-spline wavelets. The wavelet-based model allows optical flow to be recovered at different resolution levels from image derivatives. By combining estimation from different resolution levels in a coarse to fine scheme, our algorithm is able to recover accurate optical flow. The wavelet coefficients of the model at low resolution levels also provide features to index video. As an example, we have considered video sequences containing different human activities. Wavelet coefficients are efficient to give a database partition related to the kind of human activities.

1 Introduction

Le problème de l'indexation de vidéos par le contenu consiste à décrire une séquence directement par ses attributs statiques (couleurs, textures, formes), dynamiques (mouvement, activité) ou encore audio.

Le mouvement est souvent un critère objectif et approprié pour l'indexation de vidéos. Une première analyse consiste en général à partionner la vidéo en plans. Ce découpage fournit un ensemble de sous-séquences d'images dont le contenu temporel est cohérent. Chaque plan est ensuite analysé de manière à générer des descripteurs de mouvement.

La plupart des approches exploitent une segmentation au sens du mouvement qui repose sur des modèles paramétriques 2D [7]. Les paramètres du modèle de mouvement sous-jacent à la segmentation sont alors des descripteurs caractérisant le contenu dynamique de la vidéo. Ces techniques, du fait de la segmentation au sens du mouvement, sont peu fiables dans le cas où le mouvement dans la scène est particulièrement complexe (spots publicitaires, sport...).

Des solutions visant à caractériser *globalement* le mouvement par des attributs statistiques sur la texture temporelle ont été proposées dans [6, 8]. Ces descripteurs du mouvement global apportent une information qualitative sur le contenu dynamique de la séquence mais ne fournissent pas une mesure directe du mouvement. De fait, il est difficile de donner une signification explicite à ces descripteurs, tels que la nature du mouvement de la caméra ou d'un objet dans la scène.

L'approche proposée dans ce papier tente de réunir les avantages de la modélisation paramétrique du mouvement et de la description globale du mouvement. Ainsi, sans segmentation préalable au sens du mouvement, nous cherchons à definir des descripteurs globaux permettant de remonter de façon précise au mouvement entre les images.

Notre approche consiste à modéliser spatialement le mouvement sur toute l'image par une base d'ondelettes B-splines. L'estimation du modèle ne demande aucune segmentation préalable. Les coefficients d'ondelettes obtenus sont des descripteurs du mouvement global qui permettent une analyse et une indexation des vidéos.

2 Modélisation globale du mouvement

La modélisation globale du mouvement consiste à représenter en tout point de l'image le flot optique par *un seul modèle paramétrique*.

Soit $\mathbf{V}(x,y,t) = (v_x, v_y)$ le champ de vecteurs vitesses entre 2 images $I(t)$ et $I(t+1)$ ($\mathbf{V}$ défini sur toute la surface Ω de l'image). $\mathbf{V}(x,y,t)$ peut être approximé sur Ω par une combinaison linéaire de fonctions de bases :

$$\mathbf{V}(x,y) = \sum_k \mathbf{c}_k \phi_k(x,y) \quad \text{avec } \mathbf{c}_k = (c_x^k, c_y^k) \qquad (1)$$

Le modèle paramétrique de mouvement est défini par la base $\{\phi_k\}_k$. Le choix des fonctions ϕ_k détermine la capacité du modèle à décrire correctement le mouvement dans sa globalité.

Les modèles polynomiaux (affines, quadratiques...), largement utilisés en analyse du mouvement, ne peuvent modéliser un flot optique complexe dans sa globalité et nécessitent une segmentation préalable du mouvement [2].

Des chercheurs ont formulé le problème de la modélisation du mouvement global comme un problème d'interpolation [10, 11]. Bien que le modèle soit défini globalement, l'influence de chaque fonction de base est très local (le support spatial des noyaux interpolant). Les paramètres du modèle $\mathbf{c}_k$ sont une version sous-échantillonée du flot optique réel et ne sont donc pas directement des descripteurs du mouvement global.

Nous avons choisi de modéliser le mouvement global par des ondelettes pour plusieurs raisons :

- les ondelettes sont parfaitement adaptées pour la représentation de fonctions continues par morceau.
- cette décomposition fournit une représentation très compacte, un signal étant en général décrit par relativement peu de coefficients d'ondelettes.
- la décomposition en séries d'ondelettes opère une analyse multirésolution du signal, ce qui permet une description du mouvement à plusieurs échelles.

Parallèlement à nos travaux, Wu *et. al.* [12] ont récemment proposé un modèle de mouvement global défini par une base d'ondelettes de Cai-Wang (IJCV 2000). Notre approche diffère sur les aspects suivants: la nature des ondelettes, la construction de la base, l'estimation des paramètres et enfin la stratégie multirésolution utlisée.

3 Modèle de mouvement basé sur les séries d'ondelettes B-splines

3.1 Décomposition en séries d'ondelettes

Toute fonction $f(x) \in L^2(\mathbb{R})$ peut être développée en une somme pondéré de fonctions de bases :

$$f(x) = \sum_k c_{l,k}\phi_{l,k}(x) + \sum_{j\geq l}\sum_k d_{j,k}\psi_{j,k}(x) \qquad (2)$$

$\phi_{j,k}(x) = \phi(2^j x - k)$ et $\psi_{j,k}(x) = \psi(2^j x - k)$ sont respectivement la fonction d'échelle et l'ondelette, dilatée par un facteur j et translatée par k.

Soit V_j, avec $j \in \mathbb{Z}$, une famille de sous-espaces vectoriels fermés de $L^2(\mathbb{R})$. La famille de fonctions d'échelles $\{\phi_{j,k}(x)\}_k$ est une base de V_j. V_j contient toutes les fonctions de $L^2(\mathbb{R})$ au niveau de résolution j. Le niveau de détails suplémentaires est obtenu par la famille d'ondelettes $\{\psi_{j,k}(x)\}_k$, qui est une base de W_j. Ainsi, l'approximation de $f(x)$ au niveau j peut être obtenue en utilisant uniquement une combinaison linéaire de fonctions d'échelles :

$$f_j(x) = \sum_k c_{j,k}\phi_{j,k}(x) \qquad (3)$$

Nous désirons modéliser le flot optique $\mathbf{V}(x,y)$ par des fonctions d'échelles 2D. Généralement, le passage du 1D au 2D s'effectue par le produit tensoriel suivant :

$$\Phi_{j,k_1,k_2}(x,y) = \phi(2^j x - k_1)\phi(2^j y - k_2) \qquad (4)$$

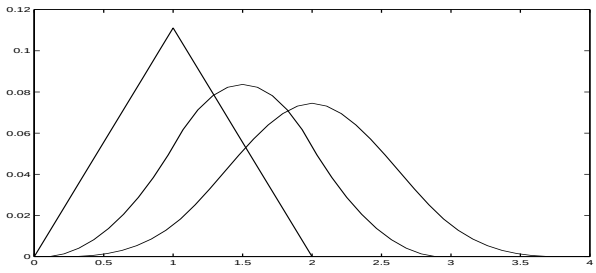


FIG. 1: *a)Fonctions d'echelles B-splines de degré 1, 2 et 3 pour $j = 0$*

Les indices j, k_1, k_2 representent respectivement l'échelle et le décalage horizontal et vertical. Le mouvement global peut alors être approximé au niveau de résolution j par :

$$\mathbf{V}_j(\mathbf{p_i},t) = \sum_{k_1,k_2=0}^{2^j-1} \mathbf{c}_{j,k_1,k_2}\,\Phi_{j,k_1,k_2}(\mathbf{p_i}) \qquad (5)$$

3.2 Les ondelettes B-splines

Afin de mesurer un flot optique lisse et régulier, l'ondelette utlisée doit être aussi lisse et symétrique que possible. La famille d'ondelettes *B-splines* sont des fonctions ayant une regularité et une symétrie maximum. De plus, ayant un support compact, les ondelettes B-splines ne posent pas de problèmes de bords critiques (contrairement à l'analyse de Fourier par exemple).

Une fonction B-spline de degré $N-1$ est le résultat de la convolution de N fonctions "portes" :

$$\phi_0^{N-1}(x) = (B * B * \ldots * B)(x) \qquad (6)$$

avec $B = \left(\frac{1}{2}, \frac{1}{2}\right)$. La figure 1 représente les fonctions B-spline 1D de degré 1, 2 et 3 que nous utiliserons dans la suite de ce papier. Une fonction d'échelle B-spline au niveau de résolution $j+1$ est définie par l'équation de dilatation :

$$\phi_{j+1}^{N-1}(x) = 2^{1-N}\sum_{k=0}^{N}\left(\begin{array}{c} N \\ k \end{array}\right)\phi_j^{N-1}(2x-k) \qquad (7)$$

$\phi_0^{N-1}(x)$ est défini sur $[0, N]$. Comme nous travaillons sur un signal à support fini, la fonction d'échelle $\phi_0^{N-1}(x)$ est dilatée afin d'être définie sur le support du signal.

4 Estimation des coefficients d'ondelettes

Considérons une séquence d'images $I(\mathbf{p_i}, t)$ avec $\mathbf{p_i} = (x_i, y_i) \in \Omega$. L'*hypothèse de conservation de la luminance* stipule que le niveau de gris d'un point physique de la séquence ne varie pas au cours du temps, c'est à dire :

$$I(\mathbf{p_i}, t) = I(\mathbf{p_i} + \mathbf{V}(\mathbf{p_i}, t), t+1) \qquad (8)$$

où $\mathbf{V}(\mathbf{p_i}, t)$ est le flot optique entre les deux images $I(\mathbf{p_i}, t)$ et $I(\mathbf{p_i}, t+1)$.

Considérons le modèle de mouvement défini par l'expression (5) à un niveau de résolution j. La mesure du mouvement global à cette résolution consiste à estimer le

vecteur des paramètres du mouvement $\theta_j = [\mathbf{c}_0 \dots \mathbf{c}_N]^T$ à partir de la séquence d'images. Ceci est obtenu par la minimisation d'une fonction de coût robuste [9] appliqué sur l'ensemble du support de l'image Ω :

$$E = \sum_{\mathbf{p}_i \in \Omega} \rho\left(I(\mathbf{p_i} + \mathbf{V}_j(\mathbf{p_i}, t), t+1) - I(\mathbf{p_i}, t), \sigma\right)$$

et

$$\theta_j = \text{argmin}_{\theta_j}(E) \tag{9}$$

La fonction $\rho(\cdot, \sigma)$ est un M-estimateur de Geman-McLure et le paramètre σ détermine la robustesse du processus d'estimation.

L'utilisation d'un M-estimateur comme norme d'erreur est nécessaire afin de tenir compte des erreurs dues à la violation de l'hypothèse de la conservation de la luminance ou à la non validité en tout point de Ω du modèle de mouvement. L'estimation de θ_j est alors robuste aux données aberrantes.

La minimisation est effectuée progressivement en modélisant le mouvement par une base d'ondelettes à un niveau de résolution grossière, puis par des bases dont la résolution est de plus en plus fine. Ce schéma multirésolution diffère de ceux généralement employés, qui consistent à estimer le flot optique sur des versions sous-échantillonnées de l'image. Ici la taille du support reste fixe, mais la résolution du modèle varie.

Cette estimation récursive est menée par la méthode des *moindres-carrés pondérés itérés (MCPI)* [9]. Cette procédure de minimisation, adaptée aux propriétés spécifiques des ondelettes, est détaillée dans une précédente communication [3].

5 Résultats

L'estimation du vecteur de paramètres du mouvement θ permet d'obtenir un flot optique de qualité, comme nous le montrons brièvement dans la première partie de cette section. Cette représentation compacte du mouvement permet aussi de définir des descripteurs du mouvement global pour l'indexation de vidéos. Nous présentons en deuxième partie un résultat de classification obtenu sur une base de vidéos.

5.1 Estimation du mouvement

La séquence *Baltrain* (Fig 2.a) présentée en exemple, possède un mouvement global complexe. Le mouvement est induit par le déplacement de la caméra vers la droite ainsi que le déplacement du calendrier, du train, du ballon et du mobile. Les figures 2.b, c et d représentent le flot optique estimé à différents niveaux de résolution. Le mouvement estimé est proche du mouvement réel (caméra, calendrier, train, ballon), alors qu'il est défini par seulement 16×16 paramètres (au niveau $j = 4$). Une mesure quantitative de la précision de l'estimation a été obtenue sur la séquence artificielle *Yosemite*, où l'erreur angulaire moyenne [1] est égale à 4.4^o.

5.2 Indexation de vidéos par les coefficients d'ondelettes

5.2.1 Définition des descripteurs de mouvement

Les coefficients d'ondelettes estimés à partir de chaque paire d'image $(t, t+1)$ d'une séquence S forment un vecteur de paramètres de mouvement $\theta_t = [\mathbf{c}_1 \dots \mathbf{c}_N]$. Nous définissons un descripteur de mouvement associé à une sequence S contenant M images comme le centre de gravité de l'ensemble des vecteurs de paramètres du mouvement estimés sur S :

$$\mathbf{\Theta}_S = \frac{1}{M} \sum_{t=1}^{M} \theta_i \tag{10}$$

L'ensemble des descripteurs $\mathbf{\Theta}_S$ pour K vidéos décrivent un espace de descripteurs du mouvement :

$$\Omega = \{\mathbf{\Theta}_{S_i}\}, \;\; i = 1 \dots K \tag{11}$$

La dimension de Ω est égale au nombre d'ondelettes utilisées dans le modèle de mouvement (typiquement 32, 64, 512,...). L'*analyse en composante curviligne (ACC)* [5] permet de projeter les éléments de Ω dans un espace de dimension inférieur. La projection de Ω sur un plan nous permet de juger de la pertinence des descripteurs de mouvement Θ_S pour classer et indexer des vidéos par le mouvement.

5.2.2 Résultats expérimentaux

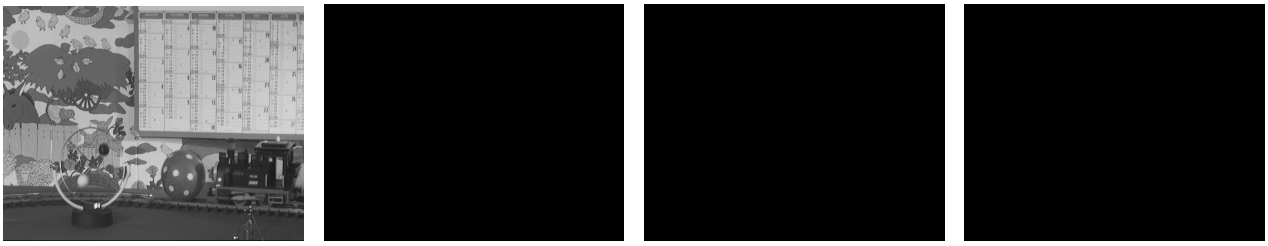
Nous avons testé notre approche sur une base contenant 30 vidéos représentant six types d'activités humaines [4](Figure 3.a): *up*, *down*, *left*, *right*, *come* and *go* (fréquence d'acquisition : $10Hz$, nombre d'images par séquence : 10). A chaque séquence d'images est associé un descripteur de mouvement estimé par une base d'ondelettes B-splines de degré 1 au niveau 2. La base de vidéos est alors représentée par un espace de descripteurs de dimension 32.

La figure 3.b) représente la projection 2D de Ω. Ce résultat montre que les coefficients d'ondelettes sont pertinents pour regrouper les vidéos par rapport au type d'activités qu'elle contiennent.

6 Conclusion

Nous avons présenté un nouveau modèle de mouvement basé sur les ondelettes B-splines. L'utilisation de cette modélisation paramétrique permet à la fois d'estimer de façon précise le flot optique dans les séquences d'images et de définir des descripteurs de mouvement pertinents pour l'indexation de vidéos.

Les perspectives concernent plus particulièrement le problème de l'indexation de vidéos. Nous devons définir de façon plus précise les descripteurs de mouvement correspondants à une séquence (l'utilisation du seul centre de gravité calculé sur l'ensemble des images n'est pas satisfaisante). Un deuxième point concerne l'extraction d'informations de plus haut niveau à partir des coefficients estimés, tel que le déplacement de la caméra et le mouvement des objets.



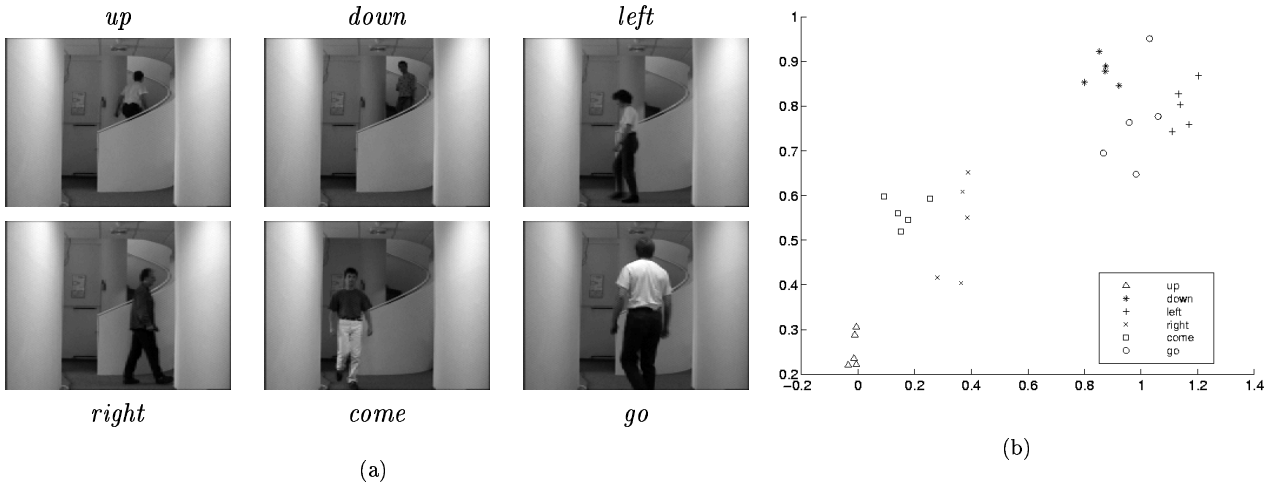
(a) Image de la séquence *baltrain*

(b) Flot optique estimé à $j = 2$

(c) Flot optique estimé à $j = 3$

(d) Flot optique estimé à $j = 4$

FIG. 2: *Image de la séquence* Baltrain *et mouvement global estimé au niveau a) 2, b) 3 et c) 4. Le modèle est défini par des B-splines de degré 3.*



(a)

(b)

FIG. 3: *a) Extraits de la base de données et b) projection 2D de l'espace de descripteurs de mouvement obtenue par ACC*

Références

[1] J. Barron, D. Fleet, and S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 1(12):43–77, 1994.

[2] M. Black and A. Jepson. Estimating optical flow in segmented images using variable-order parametric models with local deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):57–92, July 1996.

[3] E. Bruno and D. Pellerin. Global motion model based on b-spline wavelets : application to motion estimation and video indexing. In *Proc. of the 2nd Int. Symposium. on Image and Signal Processing and Analysis, ISPA'01*, June 2001.

[4] O. Chomat. *Caractérisation d'éléments d'activités par la statistique conjointe de champs réceptifs*. PhD thesis, Institut National Polytechnique de Grenoble, 2000.

[5] P. Demartines and J. Herault. Curvilinear component analysis: A self-organising neural network for non linear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154, 1997.

[6] R. Fablet and P. Bouthemy. Statistical motion-based retrieval with partial query. In *Proc. of the 4th Int. Conf. on Visual Information Systems, VISual99*, volume 1929, pages 96–107, November 2000.

[7] M. Gelgon and P. Bouthemy. Determining a structured spatio-temporal representation of video content for efficient visualization and indexing. In *Proc. 5th Europeen Conf. on Computer Vision, ECCV'98*, Freiburg 1998.

[8] R. Nelson and P. Polana. Qualitative recognition of motions using temporal texture. In *CVGIP: Image Understanding*, volume 1, pages 78–89, July 1992.

[9] J.-M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(14):348–365, December 1995.

[10] S. Srinivasan and R. Chellappa. Noise-resiliant estimation of optical flow by use of overlapped basis functions. *Journal of the Optical Society of America A*, 16(3):493–507, March 1999.

[11] R. Szeliski and J. Coughlan. Spline-based image registration. *International Journal of Computer Vision*, 22(3), 1997.

[12] Y. Wu, T. Kanade, C. Li, and J. Cohn. Image registration using wavelet-based motion model. *International Journal of Computer Vision*, 38(2), 2000.