Video structuring, indexing and retrieval based on global motion wavelet coefficients

E. Bruno[1], D. Pellerin[2]
[1]Computer Vision and Multimedia Laboratory, University of Geneva
25 rue du Général Dufour, 1211 Geneva 4, Switzerland
[2] Laboratoire des Images et des Signaux, INPG
46 av. Félix Viallet, 38031 Grenoble Cedex, France
e-mail: Eric.Bruno@unige.ch, denis.pellerin@lis.inpg.fr

Abstract

This paper describes an approach for video structuring and indexing. It relies on motion wavelet coefficients directly estimated from image sequence. These coefficients provide a multiscale characterization of optical flow. They allow to define dominant *and* local *motion descriptors, respectively related to camera and object displacements. We use dominant motion descriptors to perform a temporal segmentation of the sequence. Shots extracted are characterized in term of dominant motion properties and indexed by using descriptors related to local motion content. These operations allow to retrieve shots, by example queries, according to only dynamic content of the scene and not camera displacements.*

1. Introduction

Video databases are growing so rapidly that most of the contained information is becoming inaccessible. A valuable tool in the management of visual records is the ability to automatically ”describe” and index the content of video sequences. Such a facility would allow recovery of desired video segments or objects from a large video databases. Efficient use of stock film archives and identification of specific activities in surveillance videos are usually quoted as potential applications.

Motion-based video indexing requires to extract motion features which are relevant to characterize video content. These features are then analyzed to recover the temporal structure of video (corresponding to various motion properties) and index elementary video shots according to some descriptors. These stages should facilitate higher level tasks, such as video browsing, editing or retrieval.

Our approach for motion-based video indexing relies on previous works which concern global motion estimation between two images using wavelet-based parametric model [1, 7]. Contrary to polynomial motion models, largely used in motion-based video indexing [4, 5], such a model can be directly applied over the whole image without any prior and unreliable segmentation stage. The estimated motion parameters (wavelet coefficients) then provide a robust, global and meaningful description of motion content.

The wavelet coefficients of global motion model are efficient to classify video sequences [1], and we propose in this paper to use them to perform, according to dominant and local motion properties, a temporal partition and index of image sequences. This aim is achieved by defining, from the optical flow wavelet coefficients, features related to *dominant* and *local* motion. Dominant motion features are temporally segmented using an hierarchical classification modified to take into account temporal relation between the motion features extracted. Each video shot, related to dominant motion, is then indexed by using a variance measure of wavelet coefficients related to local motion. Experiments on motion-based retrieval are performed to validate our approach.

2 Optical flow wavelet coefficients estimation

In this section, we briefly outline the algorithm that we have developed to estimate motion wavelet coefficients. Further details can be found in [1].

Let us consider an image sequence $I(\boldsymbol{p}_i, t)$ with $\boldsymbol{p}_i = (x_i, y_i) \in \Omega$ the location of each pixel in the image. The *brightness constancy assumption* states that the image brightness $I(\boldsymbol{p}_i, t+1)$ is a simple deformation of the image at time t

$$I(\boldsymbol{p}_i, t) = I(\boldsymbol{p}_i + \mathbf{v}(\boldsymbol{p}_i), t+1), \qquad (1)$$

where $\mathbf{v}(\boldsymbol{p}_i, t) = (u, v)$ is the optical flow between $I(\boldsymbol{p_i}, t)$ and $I(\boldsymbol{p}_i, t+1)$. This velocity field can be globally modeled as a coarse-to-fine 2D wavelet series expansion from scale L to l

$$
\begin{aligned}
\mathbf{v}_{\boldsymbol{\theta}}(\boldsymbol{p}_i) &= \sum_{k_1,k_2=0}^{2^L-1} \boldsymbol{c}_{L,k_1,k_2}\Phi_{L,k_1,k_2}(\boldsymbol{p}_i) \\
&+ \sum_{j\geq L}^{l}\sum_{k1,k2=0}^{2^j-1}\Big[\boldsymbol{d}^H_{n,k1,k2}\Psi^H_{j,k1,k2}(\boldsymbol{p}_i) \\
&+\boldsymbol{d}^D_{n,k1,k2}\Psi^D_{j,k1,k2}(\boldsymbol{p}_i) + \boldsymbol{d}^V_{n,k1,k2}\Psi^V_{j,k1,k2}(\boldsymbol{p}_i)\Big]\,, \qquad (2)
\end{aligned}
$$

where $\Phi_{L,k_1,k_2}(\boldsymbol{p}_i)$ is the 2D scaling function at scale L, and $\Psi^{H,D,V}_{j,k_1,k_2}(\boldsymbol{p}_i)$ are wavelet functions which respectively represent horizontal, diagonal and vertical variations. These functions are dilated by 2^j and shifted by k_1 and k_2. The coarsest level corresponds to $L = 0$ whereas l defines the finest details that can be fitted by the motion model.

In order to recover a smooth and regular optical flow, we use *B-spline* wavelets, which have maximum regularity and symetry. The degree of the B-spline determines the approximation accuracy.

The motion parameter vector $\boldsymbol{\theta}$, which contains wavelet coefficients $\boldsymbol{c}_{L,k_1,k_2}$ and $\boldsymbol{d}^{H,D,V}_{j,k_1,k_2}$ for all j, k_1, k_2 is estimated by minimizing an objective function

$$
\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \sum_{\boldsymbol{p}_i\in\Omega} \rho\left(I(\boldsymbol{p}_i + \mathbf{v}_{\boldsymbol{\theta}}(\boldsymbol{p}_i), t+1) - I(\boldsymbol{p}_i, t)\right), \qquad (3)
$$

where $\rho(\cdot)$ is a robust norm error (M-estimator). The minimization step is achieved using an incremental and multiresolution estimation method [6].

The wavelet-based motion model enables to estimate for successive frames an accurate optical flow defined by its wavelet coefficients [1]. The motion wavelet coefficient vector $\boldsymbol{\theta}$ also provides a compact and meaningful motion description usefull to characterize video's dynamic contents according to camera displacements and object motions.

3. Multiscale motion characterization

The optical flow multiscale analysis performed by the wavelet motion model (2) allows to separate global information, related to dominant motion (assumed to account for camera displacement), to local information, more related to object displacements.

By considering only wavelet coefficients of the largest scale ($L = 0$), one can define dominant motion descriptors

$$
\boldsymbol{\theta}_{cam} = \left[\boldsymbol{c}_0, \boldsymbol{d}^H_0, \boldsymbol{d}^V_0, \boldsymbol{d}^D_0\right]. \qquad (4)
$$

The function subspace spanned by $\boldsymbol{\theta}_{cam}$ contains any polynoms of degree n, where n is the degree of the B-splines used to model optical flow. Then, a model based on B-splines of degree 1 allows to recover affine dominant motions, of degree 2, quadratic dominant motions, etc. . .

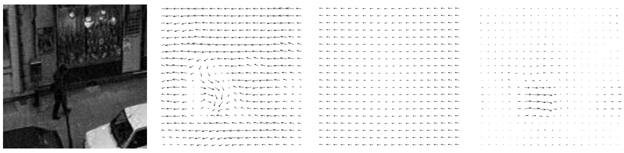
(a) (b) (c) (d)

Figure 1. *Dominant and local motion extraction: a) frame from sequence, b) estimated optical flow , c) dominant motion reconstructed from $\boldsymbol{\theta}_{cam}$ and d) local motion reconstructed from $\boldsymbol{\theta}_{obj}$.*

On the other hand, finer scale wavelet coefficients define descriptors associated to object motions

$$
\boldsymbol{\theta}_{obj} = \left[\boldsymbol{d}^H_{j,k_1,k_2}, \boldsymbol{d}^V_{j,k_1,k_2}, \boldsymbol{d}^D_{j,k_1,k_2}\right], \qquad (5)
$$

for $j = 1, \cdots, l$ and $(k_1, k_2) \in [0, 2^j - 1]^2$.

Note that these two descriptors only provide *qualitative* informations about dominant and local motions and not an accurate estimation.

Figure 1 presents an example of motion content handled by $\boldsymbol{\theta}_{cam}$ and $\boldsymbol{\theta}_{obj}$. The image sequence represents a pedestrian walking to the right. The dominant motion, due to the camera displacement, is a translation to the left of the scene. The optical flow estimated by wavelet model (2) is presented in figure 1.b). Optical flows reconstructed from $\boldsymbol{\theta}_{cam}$ and $\boldsymbol{\theta}_{obj}$ are respectively shown in figure 1.c) and d) and exhibit the camera displacement and the pedestrian motion.

4. Video structuring and indexing

Content-based video indexing primarily requires to recover the temporal structure of video corresponding to elementary shots. Then each shot could be indexed with descriptors related to local motion in order to characterize the dynamic content of the scene and not camera displacements.

In the following, we present a method to recover these shots according to dominant motion similarities and to index them by using local motion descriptors.

4.1. Hierarchical temporal segmentation

Once motion wavelet coefficients have been estimated for each frame f_i of a sequence S containing M frames, we obtain a feature space Ω_S spanned by the motion feature vectors $\boldsymbol{\theta}_i$, $i = 1, \ldots, M$. To temporally segment the feature spaces Ω_{cam} (spanned by $\boldsymbol{\theta}_{cam}$), we consider a hierarchical classification with a temporal connexity constraint.

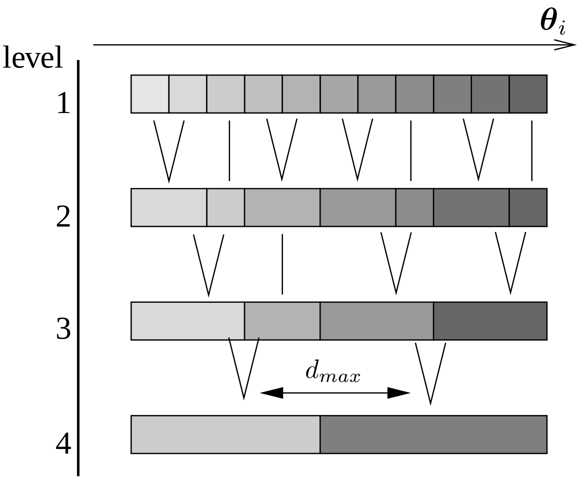

Figure 2. *Hierarchical temporal segmentation of an image sequence. Each frame f_i of a sequence S is indexed by a motion-based feature vector $\boldsymbol{\theta}_i$ which is hierarchically classified with a temporal connexity constraint.*

The *ascendant hierarchical classification* (AHC) [2] is an incremental clustering process used with efficiency for image and video database management [3]. Let us consider the feature space Ω. First, the algorithm merges the closest pairs of vectors $\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i\pm1}$ (according to a L_2 norm) to form new clusters associated to their center of gravity $\overline{\boldsymbol{\theta}}$. A vector $\boldsymbol{\theta}_h$ is kept as individual cluster when $\min_h \parallel \boldsymbol{\theta}_h - \boldsymbol{\theta}_{h\pm1} \parallel > d_{max}$, where d_{max} is a predefined distance threshold. This procedure is iterated, until no cluster can be merged, for the lowest level to the upper one in the hierarchy (figure 2).

The hierarchy's higher level provides a temporal partition $\{P_1, P_2, \dots P_N\}$ of the video sequence according to dominant motion similarities.

4.2. Video shot indexing

We need now to index extracted shots $\{P_1, P_2, \dots P_N\}$ according to local motion content. The vector $\boldsymbol{\theta}_{obj}$ provides an accurate local motion description (motion magnitude, orientation and localization). This is not very useful since we are interested in characterizing general properties of the dynamic content present in video shots. For each frame i, we thus consider a variance measure of the wavelet coefficients in the different subbands of the representation

$$\boldsymbol{\sigma}_i = \left[\sigma_1^H, \sigma_1^D, \sigma_1^V, \sigma_2^H, \dots, \sigma_l^V\right], \\ \text{with } \sigma_j^{H,D,V} = \sum_{k_1,k_2=0}^{2^j-1} \left|\boldsymbol{d}_{j,k_1,k_2}^{H,D,V}\right|^2, \qquad (6)$$

where l is the finest scale level used in (2).

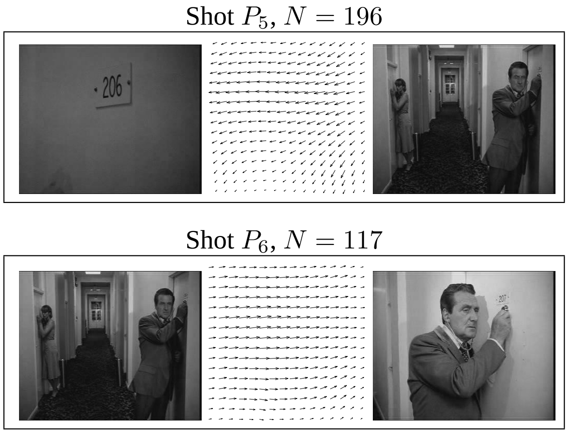

Figure 3. *Example of shots (P_5 and P_6) provided by the hierarchical segmentation of Ω_{cam} estimated on "The avenger" sequence. The shot's first frame is displayed in the left column, the reconstructed optical flow from $\overline{\boldsymbol{\theta}}_{cam}$ in the center column and the shot's last frame in the right column. The value of N is the size of P.*

We define the motion-based descriptor associated to the shot P_k of N frames as the center of gravity of $\boldsymbol{\sigma}_i$

$$\overline{\boldsymbol{\sigma}}_k = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\sigma}_i. \qquad (7)$$

This feature vector allows to index each shot of the sequence according to local motion activities for various scales and orientations.

5. Results

We have carried out experiments on an image sequence of 8000 frames extracted from the action movie "The avenger". For each frame, wavelet coefficients of optical flow are estimated by using a motion model based on B-splines of degree 1 with 3 scale levels ($L = 0$ and $l = 2$ in relation (2)). We have then extracted shots according to global motion similarities and indexed them with local motion descriptors (7). Retrieval operations with example queries are conducted in order to confirm the index stage relevance.

These operations lead to extract 233 shots containing 29 frames on average (from minimum 2 frames to maximum 287 frames). Figure 3 shows extracted shots P_5 and P_6 with their associated optical flows reconstructed from the center of gravity $\overline{\boldsymbol{\theta}}_{cam}$. These two successive shots present opposite global motions, with for the first a camera panning to the left and for the second a camera panning to the right of the scene.

Retrieval by example query operations are simply performed by searching, for a given shot P_k, the closest shots

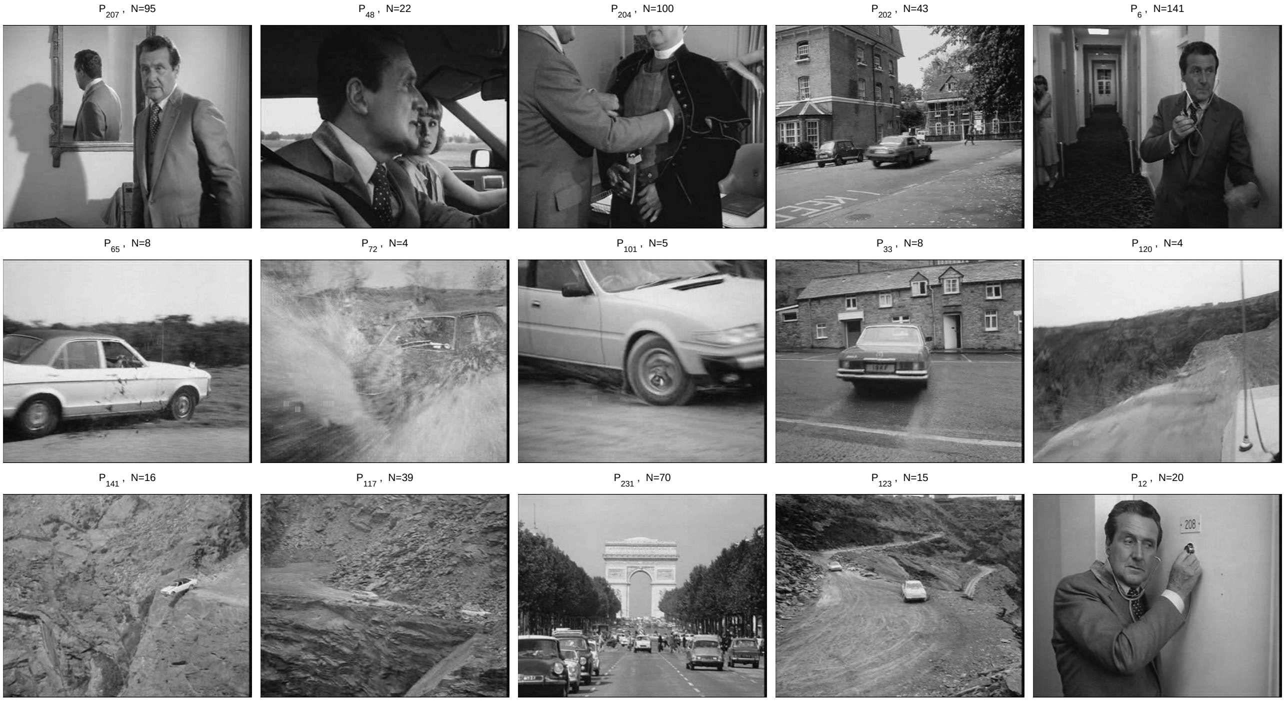

Figure 4. *Example of motion-based retrieval operations. First left column displays queries and the four other present the first four retrieved shots. Each shot is represented by its median frame.*

P_j according to an Euclidean distance between $\overline{\boldsymbol{\sigma}}_k$ and $\overline{\boldsymbol{\sigma}}_j$. This very scheme provides results displayed in figure 4. For each query (left column), the four first answers are sought (shots are represented by their median frame). The three queries present various levels of activity and motion scale. The first example presents a close up scene with low activities. The answers fall in the same class of motion. The second query involves close up scenes with high activities. The proposed answers present close motion properties. The last query contains very local motion activities, and again the shots retrieved present similar motion content (only the actor's hand is moving in the fourth answer).

6. Conclusion

We have described an original approach for motion-based video structuring and indexing. It relies on motion wavelet coefficients directly estimated from two successive frames of the sequence. Wavelet coefficients allow to extract informations related to camera displacements and object motions. Exploiting this property, we have proposed to structure video according to global motion similarities and to index each extracted shots with local motion descriptors. This indexation stage allows to retrieve shots with closed local motion content. We have obtained promising results on a large sequence extracted from an action movie.

In future works, we will have to find solutions to manage efficiently temporal variations of descriptors. Actually, since our approach is based on descriptor temporal means (hierarchical segmentation, definition of local motion descriptors), usefull information contained in this dimension is lost.

References

[1] 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.

[2] E. Diday, G. Govaert, Y. Lechevallier, and J. Sidi. Clustering in pattern recognition. *Digital Image Processing*, pages 19–58, 1981.

[3] R. Fablet and P. Bouthemy. Statistical motion-based retrieval with partial query. In *Proc. of the 4th Int. Conf. on Visual Information Systems, VISUAL'00*, volume 1929, pages 96–107, November 2000.

[4] 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.

[5] M. Irani and P. Anandan. Video indexing based on mosaic representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 86(5):905–921, May 1998.

[6] 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.

[7] Y. Wu, T. Kanade, C. Li, and J. Cohn. Image registration using wavelet-based motion model. *International Journal of Computer Vision*, 38(2):129–152, July 2000.