VIDEO SHOT DETECTION BASED ON LINEAR PREDICTION OF MOTION

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 (LIS), 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 shot detection. It relies on the temporal analysis of frame to frame optical flow measurements. These measures, which are motion wavelet coefficients directly estimated from image sequence, exhibit discontinuities when shot boundaries occur. These transitions are detected by considering the temporal trajectories of the *linear prediction errors* of wavelet coefficients. Experimental results and performance evaluation using videos with reference segmentations are presented to demonstrate the efficiency of the approach.

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 is usually quoted as potential applications.

In order to achieve a reliable video description, the primary requirement is the structuration of the video into elementary shots. It consists in detecting *transition effects* between *homogeneous segments* (shots). This video partitioning step enables to provide content-based browsing of video. This stage should facilitate higher level tasks, such as video editing or retrieval.

Most of techniques to detect cuts and transitions related to special effects (fade-in, fade-out, dissolve,…) are based on image information such as histogram comparison techniques or temporal image difference [1]. However, false alarms may still occur in case of important camera motion or in the presence of mobile objects, leading to an undesirable over-segmentation of the video stream. In these cases, motion estimation provides a more intrinsic information to analyse video content. It enables to retrieve video dynamic contents, such as temporal or spatio-temporal structures, camera displacements [2] or visual actions or events [3].

In this paper, we exploit the fact that shot boundaries correspond to optical flow temporal discontinuities. To measure the discontinuity of the entire spatial pattern of motion, a global wavelet-based parametric motion model is estimated [4, 5]. The estimated motion parameters (wavelet coefficients) then provide a global, robust, compact and meaningful description of the motion content. We then analyse the temporal trajectories of the motion parameter *linear prediction errors*. This technique enables to enforce discontinuities and to make the shot detection easier. The algorithm is evaluated on large videos, containing various motion contents and many transitions.

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 [4].

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

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

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

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

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

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

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 [4]. The motion wavelet coefficient vector $\boldsymbol{\theta}$ also provides a compact and meaningful motion description. This approach was previously used with success to characterize video's dynamic contents according to camera displacements and object motions [7].

In the video shot detection context, a coarse motion estimation is enough to determine if the image temporal variations is due to motion or scene change. As a consequence, we estimate a low resolution motion model, corresponding to $l = 3$ in relation (2). The dimension of the motion parameter vector $\boldsymbol{\theta}$ is then equal to 128.

3. SHOT DETECTION BASED ON LINEAR PREDICTION OF MOTION WAVELET COEFFICIENTS

Shot boundaries are characterized by scene changes. These transitions could be sudden (cuts) or span several frames (in the case of fade in/out, dissolve or wipe transitions). When the transitions occur, the *brightness constancy assumption* (1) assumed to estimate motion, fails. The magnitude of the estimated optical flow wavelet coefficients contain large errors and suddenly grow up (figure 1).

An obvious approach to detect shot boundaries is to find when wavelet coefficient vectors $\boldsymbol{\theta}$ have a norm greater than a predefined threshold. However, we found it not suitable in practice, since $||\boldsymbol{\theta}||$ have similar magnitude when large motions or shot boundaries occur. This case is displayed in the figure 2 (up), which represents the temporal evolution of $||\boldsymbol{\theta}||$, estimated between frame 5890 to 5990 of a TV news program. A scene change occurs around frame 5900, whereas frames 5940 to 5980 (approximatively) contain large motions. In these two regions, the motion wavelet coefficient vectors have close magnitude.

Hence, rather than only considering wavelet coefficients magnitude, we use also the linear prediction (LP) error of $\boldsymbol{\theta}$. This technique, used by Rui and Anadan to segment visual actions in image sequences [8], is as follow:
for a stationary signal $\{x_1, x_2, \cdots, x_N\}$, the LP technique consists in predicting a value x_n using previous data:

$$x_n = \sum_{i=1}^{p} a_i x_{n-i} + e_n, \tag{4}$$

where p is the order of the predictor, a_i's are the LP coefficients and e_n is the LP error. The a_i's coefficients are estimated by minimizing the squared LP error e_n^2, yielding a set of p equations which are of the form

$$\sum_{i=1}^{p} a_i \gamma_{i,j} = \gamma_{0,j} \text{ , for } j = 1, \cdots, p, \tag{5}$$

where $\gamma_{i,j} = E_n(x_{n-i}\ x_{n-j})$ is the signal autocorrelation at lag $|i - j|$.

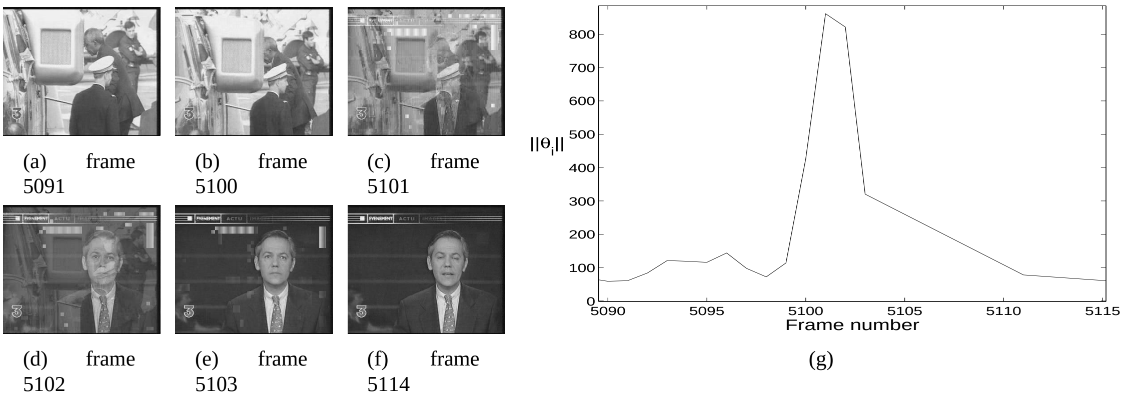
In our problem, the data sequence is $\boldsymbol{\theta}_n$, where the subscript $n = 1, \cdots, N$ denotes that $\boldsymbol{\theta}_n$ is estimated between the frames n and $n + 1$. The vector $\boldsymbol{\theta}_n$ contains all wavelet coefficients, and we consider the prediction error for each coefficient sequence independently.

Because the estimated coefficients are not temporally stationary, we use the technique of fixed-length windowing: the signal is split into short ($\backsim$ 10 frames), equal and overlapping segments in the expectation that, over small intervals, stationarity can be assumed.

Let be $\boldsymbol{e}_n = (e_{1n}, \cdots, e_{ln})$ the prediction error vector containing the LP errors estimated from all wavelet coefficient sequence at time n :

$$\boldsymbol{e}_n = \boldsymbol{\theta}_n - \sum_{i=1}^{p} a_i \boldsymbol{\theta}_{n-i}. \tag{6}$$

The prediction error vector norm $||\boldsymbol{e}_n||$ is large when wavelet coefficients have discontinuities, and small elsewhere. From the LP error curve, peaks that correspond to shot boundaries can easily be detected. Figure 2 (bottom)



(a) frame 5091 (b) frame 5100 (c) frame 5101

(d) frame 5102 (e) frame 5103 (f) frame 5114

(g)

Figure 1: *(a-f) Excerpt of a TV news program sequence involving one dissolve transition. (g) Temporal evolution of the wavelet coefficient vector norm* $||\boldsymbol{\theta}_n||$.

shows the temporal evolution of $||\boldsymbol{e}_n||$ computed from the vectors $\boldsymbol{\theta}_n$ (upper plot in figure 2). The place having large prediction error corresponds to the shot boundary, while the segment with large motion (frames 5940 to 5980) has small LP errors.

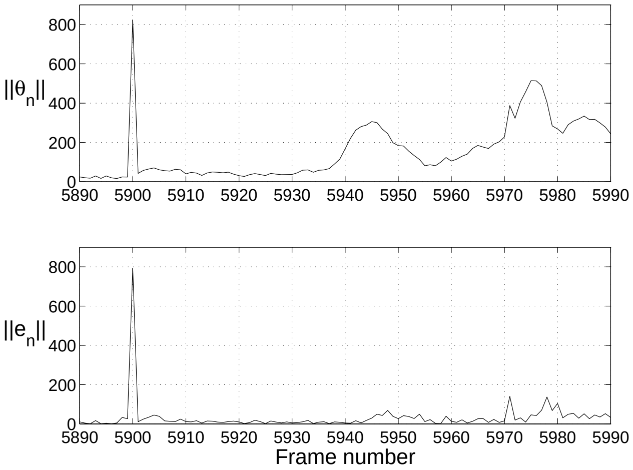


Figure 2: *Temporal evolution of the wavelet coefficient vector norm* $||\boldsymbol{\theta}_n||$ *(up) and the prediction error vector norm* $||\boldsymbol{e}_n||$ *(bottom).*

Then, by thresholding $||\boldsymbol{e}_n||$, we can detect without ambiguity peaks corresponding to shot boundary locations. The threshold is adaptively determined by computing the *median absolute deviation* (MAD) of $||\boldsymbol{e}_n||$ [9].

$$\sigma = \alpha \cdot \text{median}_n(||\boldsymbol{e}_n|| - \text{median}_j(||\boldsymbol{e}_j||)|). \qquad (7)$$

The MAD provides a variance robust estimation of $||\boldsymbol{e}_n||$, where the largest LP errors are considered as outliers. This quantity enables to define a threshold which is insensitive to peak magnitudes. In our experiments, α is set to 15.

4. EXPERIMENTAL RESULTS

We have evaluated our algorithm on two excerpts of videos extracted from the AIM (Multimedia Indexation Action) corpus[1]. A reference temporal segmentation, available for each video, enables to precisely quantify the quality of the shot detection.

The first video deals with TV news (figure 1.a to f) and has 5700 frames. It exhibits 41 transitions, including dissolve effects (between scenes with presenter and reports) and cuts (whithin reports).

The second video is a TV serie, named *"The Avengers"* (figure 3) and has 4000 frames consisting of 71 cuts and 1 dissolve effect. The video contains a large number of shots with various motion contents. It goes from interior scenes, with low activities, to action scenes with high activities and large motions. Figure 3 displays the first 30 shots detected by the algorithm.

Results of shot detection are shown in table 1 in terms of correct, false and missed transitions detected by the method. "Missed" column corresponds not only to missed transitions, but also to progressive effects that are not perfectly aligned, since the "Correct" criterion is based on a "sufficient" overlapping of the two corresponding segments. The length of the common portion must be at least R_{MIN} times the length of the longest segment and R_{MAX} times the length of the shortest segment. Values for R_{MIN} and R_{MAX} are respectively $1/3$ and $1/2$ [10]. For the TV news video, 3 of the 4 "missed" transitions correspond to dissolve effects that have been effectively detected, but less than $1/3$ of the reference duration.

[1] This corpus was developed within the French inter-laboratory research group ISIS and the French National Institute of Audiovisuel (INA) webpage: http://www-asim.lip6.fr/AIM/corpus/aim1/indexE.html

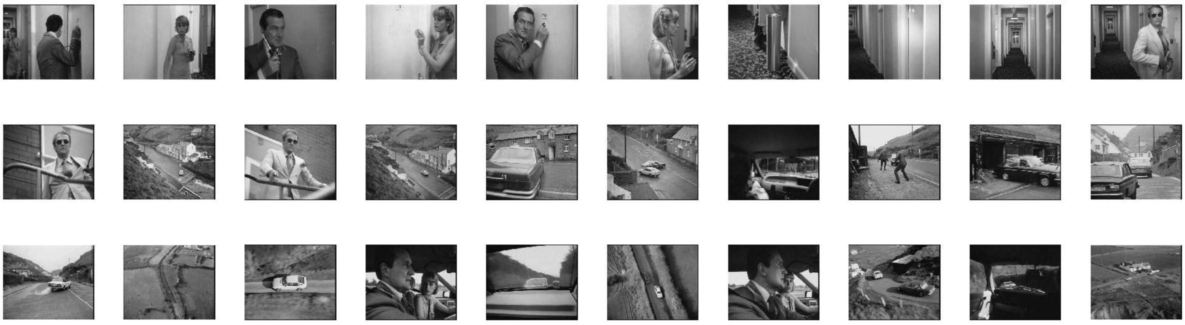


Figure 3: *The first 30 shots detected from the* The Avengers *video. Each shot is represented by its median frame.*

Video	Total frames	Correct	False	Missed
TV news	5700	37	0	4
Avengers	4000	72	1	0

Table 1: *Performance of shot detection on TV news and* The Avengers *videos.*

5. CONCLUSION

We have described an original approach for motion-based video shot detection. It relies on motion wavelet coefficients directly estimated from two successive frames of the sequence. Shot boundaries induce temporal discontinuities of the wavelet coefficients. These discontinuities are enhanced by computing the frame to frame *linear prediction errors* of wavelet coefficients. Experimental results on real videos with reference temporal segmentations exhibit good performances.

Previous experiments have shown that motion wavelet coefficients also enable to characterize the dynamic content of videos in terms of camera displacement and scene activity. We are currently exploring the issue of shot characterization in term of motion contents (shot boundaries, camera displacements and scene activity) in the aim of developing a complete algorithm for motion-based video structuring and indexing.

6. REFERENCES

[1] U. Gargi, R. Kastury, and S. Antani, "Performance characterization and comparison of video indexing algorithms," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR'98*, Santa Barbara, June 1998, pp. 559–565.

[2] P. Bouthemy, M. Gelgon, and F. Ganansia, "A unified approach to shot detection and camera motion characterization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 7, pp. 1030–1044, October 1999.

[3] L. Zelni-Manor and M. Irani, "Event-based analysis of video," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR'01*, Kauai Mariott, Hawai, December 2001, vol. 2, pp. 123–130.

[4] 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.

[5] Y.T. Wu, T. Kanade, C.C. Li, and J. Cohn, "Image registration using wavelet-based motion model," *International Journal of Computer Vision*, vol. 38, no. 2, pp. 129–152, July 2000.

[6] J.-M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348–365, December 1995.

[7] E. Bruno and D. Pellerin, "Video structuring, indexing and retrieval based on global motion wavelet coefficients," in *Proceedings of International Conference of Pattern Recognition (ICPR)*, Quebec City, Canada, August 2002.

[8] Y. Rui and P. Anandan, "Segmenting visual actions based on spatio-temporal motion patterns," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR'00*, Hilton Head, SC, June 2000, vol. 1, pp. 111–118.

[9] P. Meer, D. Mintz, and A. Rosenfeld, "Robust regression methods for computer vision: a review," *International Journal of Computer Vision*, vol. 1, no. 6, pp. 59–70, April 1991.

[10] R. Ruiloba, P. Joly, S. Marchand-Maillet, and G. Quénot, "Towards a standart protocol for the evaluation of video-to-shots segmentation algorithms," in *European Workshop on Content Based Multimedia Indexing, CBMI'99*, Toulouse, France, October 1999, pp. 41–48.