A Geometrical Key-Frame Selection Method Exploiting Dominant Motion Estimation in Video

Brigitte Fauvet[1], Patrick Bouthemy[1], Patrick Gros[2], and Fabien Spindler[1]

[1]IRISA/INRIA , Campus Universitaire de Beaulieu, 35042, Rennes cedex, France
[2]IRISA/CNRS, Campus Universitaire de Beaulieu, 35042, Rennes cedex, France
http://www.irisa.fr/vista

Abstract. We describe an original method for selecting key frames to represent the content of every shot in a video. We aim at spatially sampling in an uniform way the coverage of the scene viewed in each shot. Our method exploits the computation of the dominant image motion (assumed to be due to the camera motion) and mainly relies on geometrical properties related to the incremental contribution of a frame in the considered shot. We also present a refinement of the proposed method to obtain a more accurate representation of the scene, but at the cost of a higher computation time, by considering the iterative minimization of an energy function. We report experimental results on sports videos and documentaries which demonstrate the accuracy and the efficiency of the proposed approach.

1 Introduction and Related Work

In video indexing and retrieval, representing every segmented shot of the processed video by one appropriate frame, called key-frame, or by a small set of key-frames, is a common useful early processing step. When considering fast video content visualization, the selection of one frame per shot, typically the median frame, could be sufficient to avoid visual content redundancies ([1], [2]). On the other hand, key-frames can also be used in the content-based video indexing stage to extract spatial descriptors to be attached to the shot and related to intensity, color, texture or shape, which enables to process a very small set of images while analyzing the whole shot content. Considering only the median image of the shot is obviously too restrictive in that case. The same holds for video browsing. Another important issue is video matching based on feature similarity measurements. When addressing video retrieval, key frames can be used to match the videos in an efficient way. As a consequence, extracting an appropriate set of key-frames to represent a shot, is an important issue.

Several approaches have been investigated to extract key frames. A first category exploits clustering techniques ([3], [4]). Different features can be considered (dominant color, color histogram, motion vectors or a combination of them). Selected images are then representative in terms of global characteristics. Another class of methods consists in considering key frame selection as an energy minimization problem ([5], [6]) that is generally computationally expensive. There are also the sequential methods [7], [12], that somehow consider frame-by-frame differences. If the cumulated dissimilarities are larger than a given threshold, a new key frame is selected. With such methods, the number of selected key frames depends on the chosen threshold value.

P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp. 419–427, 2004.

In this paper, we present an original key frame selection method that induces very low computation time and which is not dependent on any threshold or parameter. Contrary to usual approaches involving a temporal sampling of the shot, its principle is to get an appropriate overview of the scene depicted in the shot by extracting a small set of frames corresponding to a uniform spatial sampling of the coverage of the scene viewed by the moving camera. This method relies on geometrical criteria and exploits the computation of the camera motion (more specifically, of the dominant image motion) to select the best representative frames of the visualized scene. One of the interests of this approach is to be able to handle complex motions such as zooming in the key-frame selection process. Another important feature of our method consists in considering geometrical properties only, which provides an accurate and efficient solution. The remainder of the paper is organized as follows. In Section 2, we present the objectives of this work. Section 3 describes the proposed method called direct method. Section 4 is concerned with an iterative method to refine the previous solution based on an energy minimization. Results are reported in Section 5 and Section 6 concludes the paper.

2 Objectives

Our goal is to account for the complete visualized scene within the shot with the minimal number of key-frames, in order to inform on the visual content of each shot as completely as possible but in the most parsimonious way. Key frames are provided in order to enable fast visualization, efficient browsing, similarity-based retrieval, but also further processing for video indexing such as face detection or any other useful image descriptor extraction.

Camera motion when video is acquired can involve zooming, panning or traveling motion. Therefore, information supplied by the successive frames is not equivalent. For this reason, it is required to choose an appropriate distribution of the key frames along the shot which takes into account how the scene is viewed, while being able to handle complex motions.

The last objective is to design an efficient algorithm since we aim at processing long videos such as films, documentaries or TV sports programs. That is why we do not want to follow approaches involving the construction of images such as mosaic images [11], or “prototype images”, but we want to select images from the video stream only. Beyond the cost in computation time, reconstructed images would involve errors which may affect the subsequent steps of the video indexing process.

3 Key-Frame Selection Based on Geometric Criteria

We assume that the video has been segmented into shots. We use the shot change detection method described in [9] which can handle both cuts and progressive transitions in the same framework. The dominant image motion is represented by a 2D affine motion model which involves six parameters and the corresponding flow vector at point p(x,y) is given by: $\omega_\theta=(a_1+a_2x+a_3y,\ a_4+a_5x+a_6y)$ varying over time. It is assumed to be due to the camera motion, and it is estimated between successive images at each time instant with the real-time robust multi-resolution method described in

[10]. The shot change detection results from the analysis of the temporal evolution of the (normalized) size of the set of points associated with the estimated dominant motion [9].

3.1 Image Transformation Estimation

In order to evaluate the potential contribution (in terms of scene coverage) of every new frame of a given shot, we have to project all the frames in the same coordinate system, e.g., the one corresponding to the first frame of the shot. To this end, we need to compute the transformation between the current frame of the shot and the chosen reference image frame. To do this, we exploit the dominant image motion computed between successive images in the shot change detection step, more specifically the parameters of the 2D affine motion model estimated all along the sequence. The transformation between the current frame I_t and the reference frame I_{ref} (in practice, the first frame of the shot) is obtained by first deriving the inverse affine model between I_t and I_{t-1} from the estimated one between I_{t-1} and I_t, then by composing the successive instantaneous inverse affine models from instant t to instant t_{ref}. Finally, we retain three parameters only of the resulting composed affine motion model, to form the transformation between frames I_t and I_{ref}, that is, the translation and the divergence parameters : $\delta_1 = a_1^{t\to ref}$, $\delta_2 = a_4^{t\to ref}$, $\delta_3 = (a_2^{t\to ref} + a_6^{t\to ref})/2$.

Aligning the successive frames with the three-parameter transformation (i.e., $(x',y')=(\delta_1 + (\delta_3+1)x,\ \delta_2 + (\delta_3+1)y)$) makes the evaluation of the contribution of each frame easier since the transformed frames thus remain (horizontal or vertical) rectangles, while being sufficient for that purpose.

3.2 Global Alignment of the Shot Images

All the successive frames of a given shot are transformed in the same reference system as explained in the previous subsection. The envelop of the cumulated transformed frames forms what we call "the geometric manifold" associated to the shot. Obviously, the shape of this manifold depends on the motion undergone by the camera during the considered shot and accounts for the part of the scene space spanned by the camera. This is illustrated by the example (Fig.1) where the camera tracks an athlete from right to left (with zoom-out and zoom-in effects) during her run-up and high-jump.

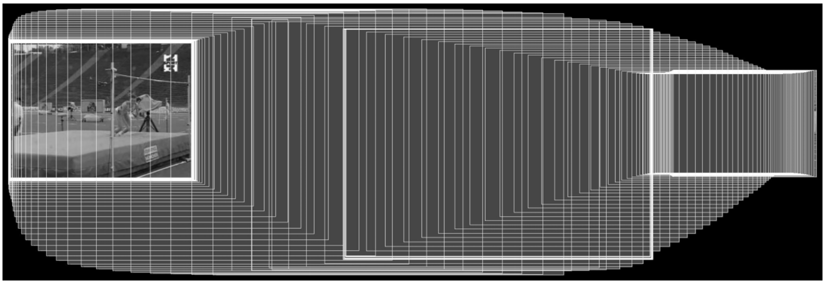

Fig. 1. Geometric manifold associated to a shot (the frames have been sub-sampled for clarity of the display); the last image of the shot is included.

We aim at eliminating redundancy between frames in order to get a representation of the video as compact as possible. We will exploit geometric properties to determine the number of frames to be selected and their locations in the shot.

3.3 Description of the Geometric Properties

We have now to evaluate in an efficient way the scene contribution likely to be carried by each frame. To define them, we consider that the frames have been first transformed in the same reference coordinate system as explained above. Then, every frame involves three kinds of scene information: a new part, a shared part and a lost part (see Fig.2).

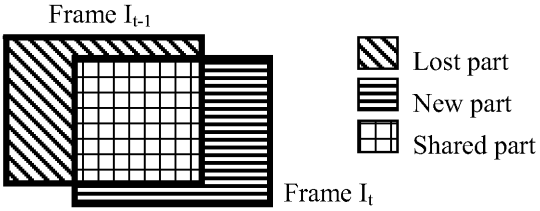

Fig. 2. Definition of the three scene information parts related to frame I_t.

As a matter of fact, we are only interested in considering the geometric aspect of these three sets of scene information. The new part is the part of the scene brought by the current frame and which was not present in the previous frame. Conversely, the lost part is the one only supplied by the previous frame. Finally, the shared part is common to the two successive frames and corresponds to the redundant information. The surfaces of the lost part, of the shared part and of the new part will be respectively denoted by σ_L, σ_S and σ_N.

These respective contributions of the two images to the description of the scene can be translated in terms of information. Let us choose the pixel as the basic information element carried by an image. The information present in the three parts σ_L, σ_S and σ_N. of an image are thus proportional to the number of pixels used to represent these surfaces. In the case of a zoom between the two images, the common portion will be described with more pixels in the zoomed image which thus brings more information than the other one. This is conforming to common sense.

In practice, the computation of these three information quantities requires the determination of polygons (see Fig.2) and the computation of their surface (number of pixels), which requires a low computation time.

3.4 Determination of the Number of Key Frames to Be Selected

The first step is to determine the appropriate number of key frames to select before finding them. We need to estimate the overall scene information, in a geometric sense, supplied by the set of frames forming the processed shot. It corresponds to the surface (denoted Σ_M) of the geometric manifold associated to the shot.

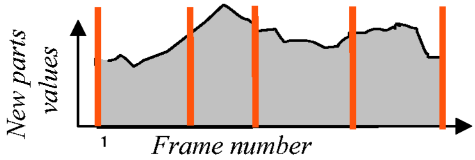

Fig. 3. Plot of the new information parts of the successive frames of the shot. Selecting key-frames (located by vertical segments along the temporal axis) of equivalent geometric contribution amounts to get strips of equivalent size partitioning the grey area equal to the surface Σ_M.

A simple way to compute this surface Σ_M is to sum the new-parts surfaces σ_N of the N_p successive frames of the shot. Selected key-frames are expected to bring an equivalent geometric contribution to the scene coverage (see Fig.3). Then, the number N* of key-frames is given by the closest integer to the ratio between the surface Σ_M and the size of the reference frame $\Sigma(I_1)$ which is given by the number of pixels of the reference image I_1.

3.5 Key-Frame Selection

The N* key-frames to find are determined according to the following criterion. We construct the cumulated function S(k) by successively adding the new scene information supplied by the N_p successive frames of the shot:

$$S(k) = \sum_{j=1}^{k} \sigma_N(j) \quad \text{with} \quad \begin{cases} \sigma_N(1) = \Sigma(I_1) \\ S(N_p) = \Sigma_M \end{cases} \tag{1}$$

The selection principle is to place a new key frame each time the function S(k) has increased of a quantity equal to the expected mean contribution given by Σ_M / N^*. This is equivalent to what is commented and illustrated in Fig. 3. Since we have to finally deal with entire time values, we consider in practice the two frames I_{k-1} and I_k, such that, $k\text{-}1 \le t_i \le k$, where the value t_i is the real position of the i^{th} key frame to select. The selected frame between these two frames is the one corresponding to the cumulated scene information value, S(k-1) or S(k), closest to the appropriate multiple of the mean contribution defined by: $M(i) = i \times \frac{\Sigma_M}{N^*}$. In addition, we take the first frame of the shot as the first key-frame.

4 Key-Frame Selection Refinement

The proposed method provides an efficient way to select appropriate key-frames in one pass as demonstrated in the results reported below. Nevertheless, one could be interested in refining the key-frame localizations, if the considered application requires it and does not involve a too strong computation time constraint. In that case, the solution supplied by the method described in Section 3, can be seen as an initial one which is then refined by an iterative energy minimization method as explained below.

4.1 Energy Function

Let us consider the set of N* key frames as a set of sites: X={x_1, x_2, .., x_{N*}}, with the following symmetric neighborhood for each site x (apart from the first and the last ones): V_x = {x-1, x+1} (1D non-oriented graph). In case this method would not be initialized with the results supplied by the direct method of Section 3, N* would be still determined as explained in subsection 3.4. Let T={t_1, t_2,..,t_{N*}} be the labels to be estimated associated to these sites, that is the image instants to be selected. They take their values in the set {1,..,N_p}, where N_p is the number of frames of the processed shot (with the constraint t_1 = 1).

Let us assume that they can be represented by a Markov model as follows. The observations are given by the scene information parts σ_N = {$\sigma_N(1)$,..,$\sigma_N(N_p)$} and σ_S = {$\sigma_S(1)$,..,$\sigma_S(N_p)$}. We have designed an energy function U(T, σ_S, σ_N) specifying the Markov model and composed of three terms: U1(T), U2(T,σ_S) and U3(T, σ_N). The first term will express the temporal distance between the initial key-frames and the new selected ones. It aims at not moving the key frames too far from the initial ones {$t^0_{x_i}$}. The second term aims at reducing the shared parts between the key-frames while not being strictly null in order to preserve a reasonable continuity. The third term will be defined so that the sum of the new parts of the selected key-frames is close to the surface Σ_M of the shot manifold. The energy function is then given by:

$$U(T,\sigma_S,\sigma_N) = U1(T) + \beta U2(T,\sigma_S) + \gamma U3(T,\sigma_N) \quad \text{with :} \tag{2}$$

$$U1(T) = \sum\nolimits_{x_i} \left| t_{x_i} - t^o_{x_i} \right|, \quad U2(T,\sigma_S) = \left| \sum_{k=1}^{N*} \sigma_S(k) - \frac{N^*\Sigma(I_1)}{\alpha} \right|,$$

$$\text{and} \quad U3(T,\sigma_N) = \left| \sum_{k=1}^{N*} \sigma_N(k) - \Sigma_M \right|$$

β and γ are weighting parameters (automatically set using appropriate relations) controlling the respective contributions of the three terms. α is set according to the value of N* (typically α = 4 in the reported experiments). Let us note that the cliques <x_i, x_{i+1}> are involved in the computation of U2 and U3 through σ_S and σ_N.

We minimize the energy function U(T, σ_S, σ_N) using a simulated annealing technique in order not to be stuck in local minima. We can afford it since we deal with a very small set of sites. We use a classical geometric cooling schedule to decrease the so-called temperature parameter.

5 Experiments

We have processed several videos of different types. Due to page number limitation, we only report in details two representative examples: a sport sequence and a documentary one.

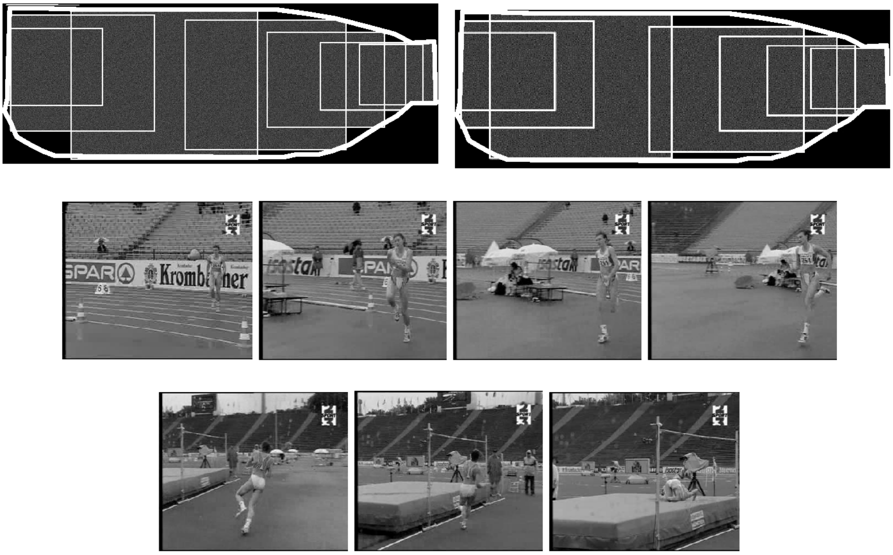

Fig. 4. Results of the direct method (top left) and iterative method (top right). The whole manifold is displayed in Fig.1; the seven key-frames obtained with the iterative method.

The first reported example is a shot of a high jump in an athletics meeting video. It involves a camera panning motion with zoom-in and zoom-out operations, to track the athlete during her run-up and high jump. The corresponding geometric manifold is shown in Fig.1.

We have first applied the direct method described in Section 3; the results are shown in Fig.4 (top left). The number of selected key frames is N*=7. We can notice that they correctly cover the scene viewed in the shot while accounting for the zoom motion and associated changes of resolution.

We have then applied the iterative method introduced in Section 4 and obtained the results displayed in Fig.4. Selected locations of key-frames are slightly modified and redundancy is further decreased as indicated in Table1. In order to objectively evaluate the results, we use the following criteria for performance analysis:

$$\mathrm{Ca} = \sum_{i=1}^{N^*} \sigma_N(i), \quad \mathrm{Cb} = \sum_{i=1}^{N^*} \sigma_S(i). \tag{3}$$

The term Ca, in relation (3), represents the cumulated new information parts of the set of selected key frames. This term corresponds to the estimated coverage of the visualized scene and it must be maximized.

The second criterion Cb evaluates the cumulated intersections of the selected key-frames, their redundancies, and it must be minimized.

This comparison has been carried out on five different sequences and is reported in Table 1. We have also considered an equidistant temporal sampling of the shot with the same number of key-frames.

Table 1. Performance analysis by comparing results obtained with the direct method (Section 3), the iterative method (Section 4) and a temporally equidistant sampling. Values are normalized with respect to results obtained with the latter one. The content of the processed sequences involves: athletics (S1 (see Fig4) and S2), soccer (S3), interview (S4) and documentary (S5).

	Temporal Sampling (Ca, Cb)	Direct method (Ca, Cb)	Iterative Method (Ca, Cb)
S1	(100, 100)	(105.3, 92.7)	(105.5, 92.4)
S2	(100, 100)	(104.7, 94.2)	(107.0, 91.4)
S3	(100, 100)	(101.2, 98.6)	(108.2, 90.9)
S4	(100, 100)	(104.6, 99.8)	(130.5, 98.6)
S5	(100, 100)	(323.0, 63.6)	(327.4, 61.3)

The performance improvement of the proposed methods is clearly demonstrated in Table 1, especially for sequences S4 and S5. For the sequence S5, we also provide the display of the selected key-frames in Fig.5. Our approach is able to adapt the location of the key-frames to the evolution of the camera motion which mainly occurs in the middle of the shot to track the people turning at the cross-road. On the other hand, the camera is mainly static when the two people are approaching in the first part of the shot and are receding in the last part of the shot.

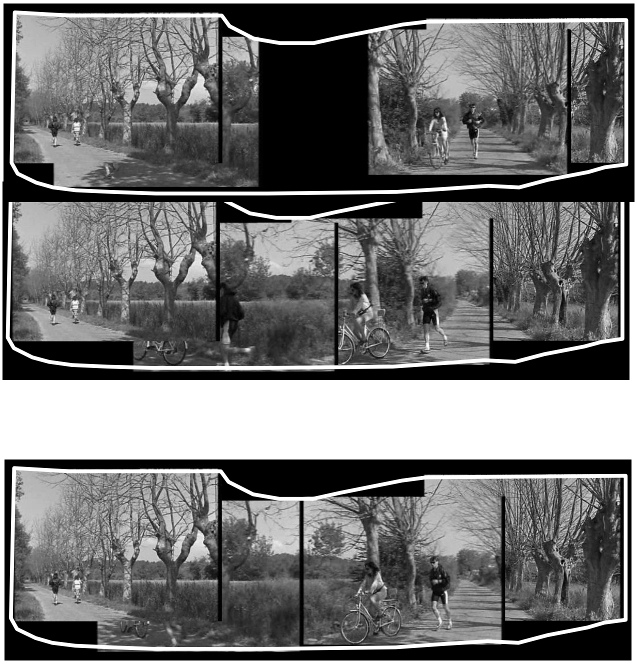

Fig. 5. Comparison of the selection of the key-frames obtained with the three methods applied to the S5 sequence. White line delimits the geometric manifold we want to cover. Top row: temporal sampling method; middle row: direct method, bottom row: iterative method. The images of the selected key-frames are displayed. N*=4.

6 Conclusion

We have presented an original and efficient geometrical approach to determining the number of key-frames required to represent the scene viewed in a shot and to select them within the images of the shot. The image frames are first transformed in the same reference system (in practice, the one corresponding to the first image), using the dominant motion estimated between successive images, so that geometrical information specifying the contribution of each image to the scene coverage can be easily computed. Two methods have been developed. The direct method allows us to solve this problem in one-pass. Results can be further refined by the iterative method which amounts to the minimization of an energy function. Results on different real sequences have demonstrated the interest and the satisfactory performance of the proposed approach. We can choose the iterative method if getting better accuracy is prevailing while computation time constraint is less important.

Acknowledgements. This work was partly supported by the French Ministry of Industry within the RIAM FERIA project. The videos were provided by INA.

References

1. Y.Tonomura, A. Akutsu, K. Otsuji, T. Sadakata: videoMAP and videospaceicon: tools for anatomizing video content, INTERCHI '93, ACM Press, pp 131-141.
2. B. Shahrary, D.C. Gibbon: Automatic generation of pictorial transcript of video programs, Proc. SPIE Digital Video Compression: Algorithms and Technologies, San Jose, CA, 1995, pp. 512-519.
3. Y. Zhuang, Y. Rui, T.S. Huang, S. Mehrotra: Adaptative key frame extraction using unsupervised clustering, Proc 5th IEEE Int. Conf. on Image Processing, Vol.1, 1998.
4. A. Girgensohn, J. Boreczky: Time-constrained key frame selection technique, in IEEE International Conference on Multimedia Computing and Systems, 1999.
5. H.C. Lee, S.D. Kim: Iterative key frame selection in the rate-constraint environment, Image and Communication, January 2003, Vol 18, n°1, pp.1-15.
6. T. Liu, J. Kender: Optimization algorithms for the selection of key frame sequences of variable length, 7th European Conf. on Computer Vision, Dublin, May 2002, Vol LNCS 2353, Springer Verlag, pp. 403-417.
7. M.M. Yeung, B. Liu: Efficient matching and clustering of video shots, Proc. ICIP'95, Vol.1, 1995, pp. 338-342.
8. A. Aner, J. Kender: Video summaries through mosaic-based shot and scene clustering, 7th European Conference on Computer Vision, Dublin, May 2002, Vol LNCS 2353, Springer Verlag, pp 388-402.
9. J.M. Odobez, P. Bouthemy, Robust multi-resolution estimation of parametric motion models. Journal of Visual Communication and Image Representation, 6(4):348-365, Dec. 1995.
10. P. Bouthemy, M. Gelgon, F. Ganansia. A unified approach to shot change detection and camera motion characterization. IEEE Trans. on Circuits and Systems for Video Technology, 9(7):1030-1044, October 1999.
11. M.Irani, P. Anandan: Video indexing based on mosaic representations, IEEE Trans. on Pattern Analysis and Machine Intelligence, 86(5):905-921, May 1998.
12. J. Vermaak, P. Pérez and M. Gangnet, Rapid summarization and browsing of video sequences, British Machine Vision Conf., Cardiff, Sept. 2002.