Determining a Structured Spatio-temporal Representation of Video Content for Efficient Visualization and Indexing

Marc Gelgon and Patrick Bouthemy

IRISA/INRIA
Campus universitaire de Beaulieu
35042 Rennes cedex, France
e-mail : mgelgon@irisa.fr, bouthemy@irisa.fr
Tel : 33-2.99.84.74.32 Fax : 33-.2.99.84.71.71

Abstract. Efficient access to information contained in video databases implies that a structured representation of the content of the video is built beforehand. This paper describes an approach in this direction, targeted at video indexing and browsing. Exploiting a 2D motion model estimator, we partition the video into shots, characterize camera motion, extract and track mobile objects. These steps rely on robust motion estimation, statistical tests and contextual statistical labeling. The content of each shot can then be viewed on a synoptic frame composed of a mosaic image of the background scene, on which trajectories of mobile objects are superimposed. The proposed method also provides instantaneous and long-term, qualitative and quantitative object motion cues for content-based indexing. Its different steps and the system they form are designed to keep computational cost low, while being able to cope with general video content was aimed at. We provide experimental results on real-world sequences. The structured output opens important possible extensions, for instance in the direction of higher-level interpretation. [1]

1 Introduction and related work

Fast, reliable and convenient access to visual information in still image and video databases is of growing importance in tasks concerning professionals in a variety of fields, as well as emerging services targeted at the general public.

Broadly, still image and video have led to their own direction in the research carried out for accessing content-based information. The major issues and cues have been reviewed in [2,9,14]. In the former field, prototypes such as QBIC

[1] This work was supported in part by DGA (Délégation Générale pour l'Armement - French Ministry of Defense) and AFIRST (French-Israeli Scientific and Technical Research Agency).

[11] are now available, but recent work such as [22], should contribute to largely improving the performance of still image retrieval applications.

This paper is concerned with access to information contained in image sequences. Ideally, a retrieval system should allow two types of accesses. The first one consists in expressing a query to the system, which returns matching entries. The second one consist in viewing the documents, leaving the user to find the relevant information. Indeed, if the query cannot be precisely defined in the terms offered by the interface, or if relevant information has not been correctly indexed, it is necessary to be able to browse in an efficient way through the video. Hence, it is desirable to represent the videos in two ways, an indexed version of video used by a query-type interface, and a version for browsing through an appropriate interface.

Content-based video indexing and efficient content visualization share the necessity for a phase of video structuring. Three levels of analysis can be distinguished. A fundamental and early task is the partitioning of the video into shots, which is mainly done by detecting shot changes. Current systems generally rely on comparison of grey-level or color histograms, computed on successive frames [1,17,26]. Thresholding on the sum of histogram bin differences, or a χ^2 test have been proposed. In [17], a set of histograms are computed on a partition of the image into blocks, and the eight largest differences are discarded, so as to reduce the perturbation causes by camera motion and mobile objects. An experimental comparison of these approaches is presented in [5]. Direct processing of the MPEG bit-stream has been proposed for instance in [19], by computing histograms using the DC components of the DCT related to I-frames. The main issues that arise in this task are the presence of progressive transitions, such as dissolve or wipe effects, strong camera motion, and the presence of mobile objects. The first problem is tackled in [1] and [26] by using different tests for cuts and progressive transitions. Coping simultaneously with all three problems generally involves the use of several dedicated techniques, and then implies tuning of multiple and sensitive parameter values. We exploit here an approach addressing these problems jointly, which we proposed in [6].

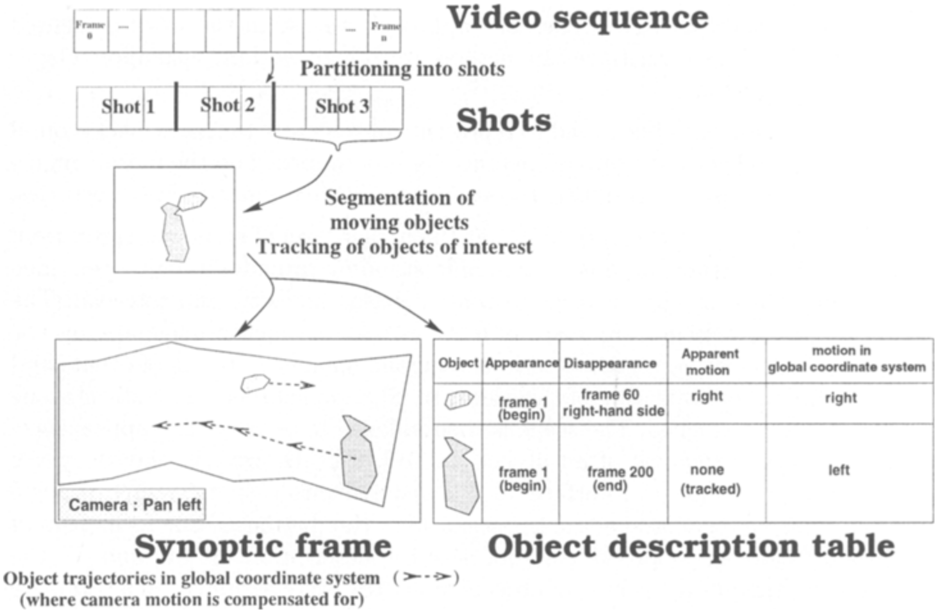
Shot content is often characterized by the estimated type of camera motion during the shot, and by one or several key-frames. Displaying the sequence of these key-frames is a simple way of visualizing the video. Shot content indexing has been proposed in [27], consisting in mapping the shot content on its key-frames, and applying to them still image indexing techniques.

Starting from a partition of a video into shots, some studies aim at summarizing the video even more, doing so-called video skimming, as for instance in [23]. In order to build shot summaries which are more informative than basic key-frames, analysis of the spatio-temporal contents of the shots can be performed. Doing so allows content-based indexing using the determined spatio-temporal structure. In [3], the structured representation of a sequence takes the form of a set of layers, after a joint estimation of the motion models and of their support, considered as a mixture model problem, and the number of models is determined through a MDL criterion. This approach has shown efficient, though a shortcoming of mixture-based approaches is the difficulty to update the number of models along the image sequence. Top-down methods consist in iteratively computing successive dominant motions and assigning their support a single label [16]. Another category of techniques adresses the segmentation issue in a Markovian framework, as a contextual labeling problem [24]. Finally, clustering approaches such as [25], in contrast, start from elementary regions and group them to form the desired regions. The approach we use in the work presented belongs to the Markovian-Bayesian methods, although working also upon a layer of elementary regions.

Mosaic images have been used to represent conveniently the scene background [3, 15], and mobile objects can for instance be superimposed on the mosaic image at several of their positions during the sequence so as to suggest their trajectories.

Research has also been carried out at the higher level of scene interpretation. This step is important, because scene understanding provides a high level video structuring which is very relevant to content-based indexing and retrieval. The spatio-temporal content, supplied as motion-based segmentation maps by the previous phase, can be represented in a symbolic manner in terms of events and spatio-temporal organization of the scene. In [8], mobile objects in a static scene are detected and tracked. The scene is then indexed in terms of e.g. appearance, disappearance or removal of an object. In [10], the 3D scene is played by the user with an appropriate interface, and translated into a spatio-temporal logic formulation so as to characterize the scenario in the database. The formalism of Petri nets has been proposed in [7] for scenario description. In [10] and [7], the location and temporal tracking of objects of interest were assumed. In a complete scheme, however, this high-level analysis relies strongly on the efficiency of the earlier phases.

The work presented in this paper first, proposes an approach for the earlier phases, that is shot change detection and motion-based segmentation, and then derives a set a object and camera motion descriptors and a compact representation of the dynamic content of a sequence. The method is summarized on Figure 1. We first partition the image sequence into shots, using a technique which handles both cuts and progressive transitions with the same test, and copes with the presence of camera motion and mobile objects. Then, for each shot, we extract and track mobile entities along the shot. The method provides, for each shot, a synoptic view consisting of a mosaic image of the background scene, on which trajectories of mobile objects are represented. Besides, qualitative and quantitative camera and object motion descriptors are obtained. The scheme proposed here is unified around a low-cost robust 2D motion model estimator; the various steps of the method exploit its robustness. In comparison with [8] which is targeted at particular scenes (assuming a static camera and disconnected mobile objects), we aim at general sequences by performing a motion-based segmentation. In contrast with [15], we have an explicit segmentation and trajectory of mobile objects, which can be used for further analysis and elaboration of object descriptors. The paper is organized as follows. Section 2 outlines the robust motion model estimator, while Section 3 recalls the method for partitioning a video into shots. Section 4 explains how objects are extracted within these shots. Section 5 describes how a synoptic view of the shots content is derived. Section 6 deals with motion descriptors. Section 7 provides experimental results, and Section 8 contains concluding remarks.


Frame 0
...
...
Frame n
Video sequence
Partitioning into shots
Shot 1
Shot 2
Shot 3
Shots
Segmentation of moving objects
Tracking of objects of interest
Camera : Pan left
Synoptic frame
Object description table
Object trajectories in global coordinate system
(where camera motion is compensated for)

Object	Appearance	Disappearance	Apparent motion	motion in global coordinate system
	frame 1 (begin)	frame 60 right-hand side	right	right
	frame 1 (begin)	frame 200 (end)	none (tracked)	left

Fig. 1. *An overview of the video structuring and representation method : partitioning into shots, locating and tracking mobile objects, and building a still image summary of the shot content showing the trajectories of objects. For indexing or understanding purposes, a table describing each mobile object is also produced, and camera motion is annotated.*

2 Dominant motion estimation

In this work, we manipulate motion information, which support can be either the whole image, or a region in the image. In both cases, motion is represented by 2D affine motion model. The motion model parameters are estimated on the relevant support (a region, or the whole frame) using a motion estimator called *RMRmod* presented in [18]. Since a robust estimator is exploited, only the model accounting for the dominant motion between a pair of successive frames within the considered support is estimated.

Between two successive frames, we first estimate the global dominant motion over the whole image. This measurement is then exploited in the successive phases of the scheme : partitioning of the video into shots, characterization of camera motion, construction of a mosaic image by image warping, and determination of trajectories and object motions in a coordinate system, called synoptic frame in the subsequent, where estimated camera motion is compensated for. Motion model estimation on region supports is used for the motion-based segmentation and tracking phases, and for the derivation of object motion descriptors.

We briefly recall here the general case where the estimation support is denoted by R, which gravity center is denoted $g = (x_g, y_g)$. R can be the whole image or a given region in the image. The displacement vector w_Θ at pixel $p(x, y)$ between two successive frames I_t and I_{t+1} is expressed as :

$$w_\Theta(p) = \begin{pmatrix} a_1 + a_2(x - x_g) + a_3(y - y_g) \\ a_4 + a_5(x - x_g) + a_6(y - y_g) \end{pmatrix} \tag{1}$$

where $\Theta = (a_1, a_2, a_3, a_4, a_5, a_6)$ represents the parameters of the 2D affine motion model. To estimate Θ, an incremental approach is adopted in a multiresolution framework to handle large displacements. Given the current estimate $\widehat{\Theta}^k$, at step k, of the motion parameter vector, we calculate the increment $\widehat{\Delta\Theta^k}$ as :

$$\widehat{\Delta\Theta^k} = \arg\min_{\Delta\Theta^k} \sum_{p_i \in R} \rho(r_i) \tag{2}$$

where the residual r_i is computed using the spatial intensity gradient ∇I and the current motion model estimate $\widehat{\Theta}^k$ as follows :

$$r_i = I(p_i + w_{\widehat{\Theta}^k}(p_i), t+1) - I(p_i, t) + \nabla I(p_i + w_{\widehat{\Theta}^k}(p_i), t+1).w_{\Delta\Theta^k}(p_i) \tag{3}$$

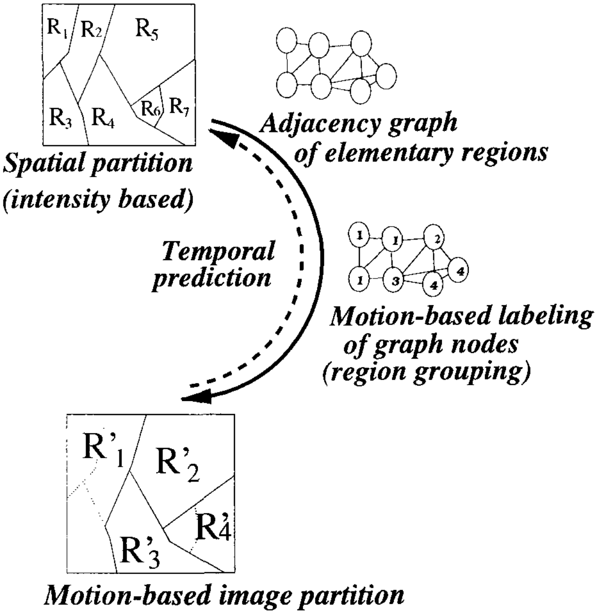
$\rho(x)$ is a hard-redescending M-estimator. Here, we consider Tukey's biweight function. We can then update the estimate $\widehat{\Theta^k}$: $\widehat{\Theta^{k+1}} = \widehat{\Theta^k} + \widehat{\Delta\Theta^k}$. This incremental process is iterated until a stopping criterion is met. This estimator allows us to get an accurate computation of the dominant motion on the support at hand, even if other motions are present.

3 Video partitioning into shots

The method we employ for this early step in structuring an image sequence uses motion models accounting for the global dominant image motion between successive images [6]. The evolution of the size of the associated estimation support enables the detection of both cuts and progressive transitions. More precisely, we consider the variable $\zeta_t = n_d/n_0$, where n_d is the measured size of the estimation support. n_d is provided by the motion estimation phase, as the set of pixels conforming to the global dominant motion. n_0 is the maximum expected support area, and is computed geometrically as the size of the part of $I(t)$ which is likely to be also in $I(t+1)$. This is derived from the dominant motion, supposed to correspond to the camera motion and estimated between t and $t+1$. Observations of the evolution of this variable along the sequence are as follows. Within a shot, ζ_t is close to 1. Between two images of different shots, i.e. at a cut instant, no consistent motion can be estimated, and ζ_t suddenly drops close to 0. In the case of progressive transitions, we observe a less pronounced decrease of ζ_t, but which still enables a correct detection. Significant jumps in ζ_t are detected using the statistical cumulative Hinkley's test [4], involving very little computation, and providing also the temporal bounds of the detected transition.

The key point of this technique is its generality of use in terms of kind of transitions and scene content. First, in contrast with most approaches which perform different tests for cuts and gradual changes, the proposed approach copes with these different kinds of transitions with the same test involving a single threshold which is kept constant [6]. Secondly, the scheme can cope with scenes including mobile objects, even of important size, and important camera motion. The scheme has also been validated on MPEG-1 and -2 reconstructed image sequences.

4 Segmentation-Tracking



R1
R2
R5
R3
R4
R6
R7
Spatial partition (intensity based)
Adjacency graph of elementary regions
Temporal prediction
1
1
2
1
3
4
4
Motion-based labeling of graph nodes (region grouping)
R'1
R'2
R'4
R'3
Motion-based image partition

Fig. 2. *Overview of the segmentation and tracking method. A spatial partition of the image is tracked along the sequence, along with a region-level motion-based partition built upon this spatial partition.*

Once partitioned into shots, we apply to the content of each shot a hierarchical motion-based segmentation method we more extensively describe in [13]. An overview diagram is provided in Figure 2. An intensity-based partition of the image is first built. Affine motion models are robustly estimated on each of these regions, using the approach described in Section 2. This first stage leads to deliberately over-segmented partition, relatively to a meaningful partition in terms of moving objects of interest. Its purpose is to retrieve all relevant boundaries. In a second stage, the elementary regions are grouped on a more meaningful motion-based criterion. The adjacency graph defined on the regions of the elementary partition is considered. The goal is then to assign labels to the nodes of this graph in such a way that regions undergoing similar (resp. different) motions are attributed the same (resp. different) labels. An optimal label configuration is sought for as a statistical contextual labeling problem. An energy function $U(e,o)$ is defined, associated to a graph-level Markovian model, and composed of the sum os local potentials V_1 and V_2 defined on pairs of neighbouring nodes as follows :

$$U(e,o) = \sum_{\gamma_j \in \Gamma} V_1(e(\gamma_j), o(\gamma_j)) + \sum_{\gamma_j \in \Gamma} V_2(e(\gamma_j)) \tag{4}$$

where $e(\gamma_j)$, $o(\gamma_j)$ respectively denote the local label configuration and the observations assigned to a pair of neighbouring nodes γ_j. Γ is the set of all such pairs on the graph. The two potentials relate assignment of labels to motion and geometrical measurements made on the regions. V_1 is defined so as to favour identical label of neighbouring nodes, if the two motion fields estimated on each of the two corresponding regions form a coherent motion field [13]. V_2 is defined so as to favour identical label of neighbouring nodes, also taking into account the degree of geometrical adjacency of the region pair. An HCF procedure is used to perform the minimization of $U(e,o)$. The motion-based regions are derived from the label configuration at convergence. A motion model is then estimated over each motion-based region group, thus characterizing the motion of mobile objects, or of the background region. A temporal prediction-updating technique, both of the spatial partition and of the region-level label configuration, enables a correct updating of spatial and motion boundaries along the sequence, as well as emergence of new motion-based regions, as appropriate.

The advantages of this technique are that, in contrast with merging-only, split-and-merge or clustering methods, and because the topology of the graph is kept unchanged throughout the labeling of its nodes, groups of elementary regions can be re-arranged in a flexible and well-controlled manner during the grouping step and from one frame to the next. Besides, contextual information can easily be incorporated in such a formalism.

5 Camera and object motion description

Section 3 and 4 have described the steps leading to a structured representation of the image sequence content. In this section, we describe how motion descriptors can be obtained for the entities that have been extracted, namely for shots and mobile objects.

First, a qualitative descriptor of camera motion for each pair of successive frames can be provided. To this purpose, the spatial support associated to the global dominant motion model, that has been computed in the shot change detection step, can be again exploited. The parameters of the affine motion model as here expressed in the following basis, corresponding to basic translational, divergence, curl and hyperbolic models, so as to be physically more meaningful:

$$\Theta = (t_1, t_2, div, rot, hyp_1, hyp_2) \qquad \text{with :} \tag{5}$$

$$t_1 = a_1 \qquad t_2 = a_4$$
$$div = \tfrac{1}{2}(a_2 + a_6)\ rot = \tfrac{1}{2}(a_5 - a_3)\ hyp_1 = \tfrac{1}{2}(a_2 - a_6)\ hyp_2 = \tfrac{1}{2}(a_3 + a_5)$$

For the various basic types of camera shooting situations, such as zooming, panning, static camera, only a subset of the six parameters as expressed in this basis should be non zero. In the case of pure horizontal panning, for instance, only parameter t_1 should be non zero. If camera is pure zooming, only the divergence term will be non-zero. In practice, the presence of noise and estimation errors has to be accounted for, and one must shift from an idea of "non-zero" to "significant value".

As regards the problem of determining which motion parameters are significant, we have shown in [12] that likelihood ratio tests tackled this issue in an efficient way, not requiring unstable parameter tuning. We resort to this approach, consisting in testing, for each parameter in Θ, the hypothesis that it is significant, against the hypothesis that it is equal to zero. This is carried out by re-estimating the motion model on the support associated to the dominant motion, while applying the constraint that the considered parameter is equal to zero, and evaluating whether this constrained model explains the data almost as well as the full-affine model. The degree of fitness of the two models to the data are compared using a statistical log-likelihood ratio test. The support on which motion estimation is performed makes the motion characterization technique resilient to mobile objects, even of significant size. Using this test, measurement noise and inadequacy of the model to explain the data can be implicitly taken into account [12].

The application of this significance test to each of the six parameters supplies a binary symbolic parameter vector, which can be mapped onto a set of qualitative motion labels, such as static camera, pan, zoom or sideways, forward or backward traveling. The sign of the significant parameters can also be exploited. It indicates, for instance, the direction of panning (left, right, up, down), whether a zoom or traveling is forward or backward, or the direction of a possible rotation around the optical axis.

The method for qualitative motion labeling, applied to the whole frame when dealing with camera motion, can in the same way be applied to the regions corresponding to the mobile objects extracted by the segmentation and tracking steps. By this means, we obtain, for each mobile object in each frame, a characterization of its motion. The motion-based segmentation step involves the robust estimation of a motion model on every region, and thus of the associated estimation support. We advocated that on the whole frame, utilizing the support of the dominant motion model made the motion characterization technique robust to other mobile objects. In the same way, at region level, we benefit from the respective dominant motion model supports by increasing resilience to minor errors in the determination of region boundaries.

Qualitative description of motion, whether of the camera or of mobile entities in the scene, is well suited for indexing. Indeed, the user may want to retrieve objects on a motion criterion, such as an object going leftwards, or coming towards the camera. However, it is likely the user will express his query in terms of perceived object motion, regardless of the camera motion in the scene. For instance, if he asks for an object going right, it is likely he may be interested in objects going rightwards in a fixed coordinate systen, and consequently also in objects being tracked rightwards by the camera. In many cases, an object motion descriptor obtained after having compensated for camera motion should fulfill better the goal of motion cues for indexing. We thus operate as follows.

Given two successive frames I_t and I_{t+1}, let us denote the estimated global dominant motion parameter vector $\widehat{\Theta^t_{t+1}}$ between these frames, to indicate clearly the temporal aspect of the problem. We built the image $\tilde{I}_{t+1}$ using a back-warping technique and exploiting $\widehat{\Theta^t_{t+1}}$, so as to compensate for camera motion. Assuming the scene is approximately planar and that this plane is not too much slanted, only the projected motion of mobile objects remains between I_t and $\tilde{I}_{t+1}$. The technique described above, applied between I_t and $\tilde{I}_{t+1}$, for each of the relevant regions, provides qualitative "scene related" motion descriptors of significant interest to an indexing system. We also indicate frame appearance and disappearance number, whether a mobile object emerges or disappears within the image, or from an image side, and in this case, from which side. Objects with no apparent translation, while camera translation is significant, are labeled as "tracked by the camera". This information can identify them to indexing or higher-level analysis systems as objects of interest. The part of the image corresponding to the object in a given image frame, can be stored in the database (Figure 1), enabling the use of region-based still image cues.

6 Synoptic views of the dynamic content of a shot

We describe here how, for each shot, a still image can be built that aims at summarizing its content, using the structured spatio-temporal representation and the motion descriptors found.

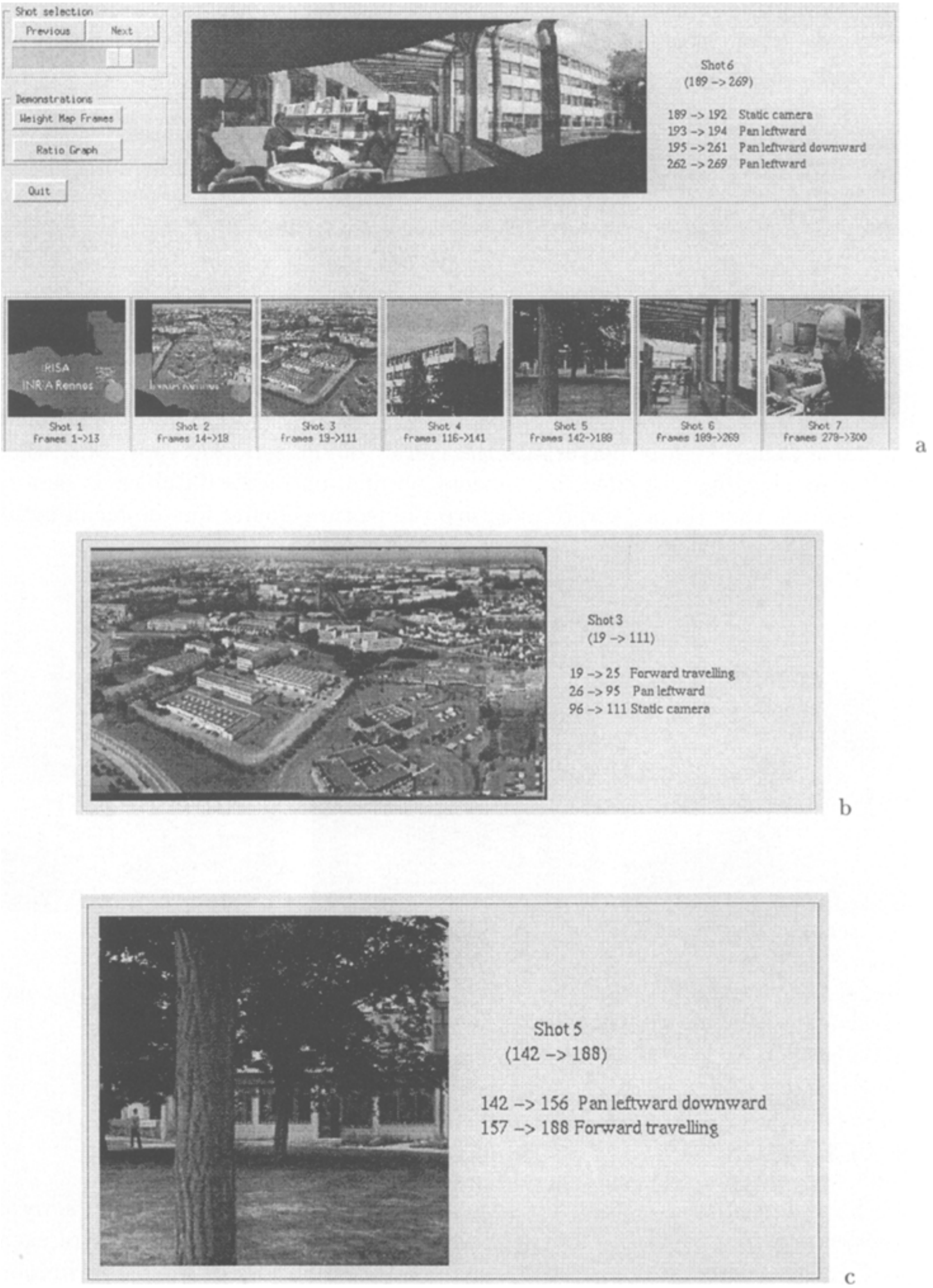
First, the background scene is represented by a mosaic image. Because we wish to build a view for quick visualization, in a context where a user searches for the part of interest in a video, we do not require a mosaic image displaying very low distortion. Such advances have however been recently proposed for instance in [20, 21]. Hence, we use an elementary technique consisting in considering a reference frame I_{t_0}, the first one in the shot, and back-warping all the frames towards a coordinate system related to this first frame, using the global dominant motion model $\widehat{\Theta^t_{t_0}}$, computed by composing $\widehat{\Theta^{t_1}_{t_0}} \dots \widehat{\Theta^t_{t-1}}$. The parts of the image corresponding to mobile objects are withdrawn beforehand. The processing is fast, since the dominant motion model between successive frames is readily available.

We then superimpose on this mosaic the trajectories of mobile objects during the shot. The trajectory of an object is represented as the sequence of position measurements of its gravity center along the shot, also back-warped in the reference frame, so as to coincide with the mosaic image. For objects that are detected to be either partially appearing or disappearing from one side of the image, the gravity center does not approximately correspond to a single physical point along the sequence. For such objects, only the first gravity center measurement is used, the trajectory is built from there on using motion estimates. The mosaic image with the superimposed trajectories suggest in a simple way, yet effective, the contents of the shot. Rotation of objects around their center and objects coming towards or moving away from the camera are annotated with appropriate icons.

7 Results

The method was validated on several real video sequences. The experimental results presented here show the application of the different steps of the method for different sequences, so as to illustrate better their performance.

We report here results for the partitioning into shots and camera motion characterization on a real documentary, that includes various types of camera motions, cuts, dissolves and a special video effect. It also include a mobile person occupying a significant part of the image. The summary created for that sequence is shown on Figure 3, and corresponds to the correct partitioning, with an accurate temporal location of shots. A key-frame per shot is shown, and for any user-selected shot, the synoptic frame is displayed, along with qualitative camera motion estimated during the shot. Results obtained on various MPEG-reconstructed versions of the same sequence at various compression rates showed almost no difference with regard to the original sequence. Unoptimized code leads to a computation time per frame pair of about 30s (image size 360x288) on an Sun UltraSparc workstation. The time-consuming operations are mainly the spatial segmentation updating phase and motion model estimations (one per spatial region and seven per motion-based region). As far as interesting configuration in terms of mobile objects is concerned, we show how a shot content can be structured and represented on a 125 frame sequence called *Mobi.* In this sequence, the camera is panning leftwards and tracking a train, which is pushing a rolling ball. This ball stops rolling from frame 35 to frame 47. On the right, a calendar is being pulled upwards, it stops, and then goes downwards. On the left is a rapidly swinging gyroscope. These mobile objects cause shadow effects. The extracted motion boundaries for frames 1, 60 and 120 are shown in Figure 4. The calendar and the train are correctly separated from the background tapestry. The rolling ball is identified, but its boundaries are unstable over time. It disappears from



Shot 6
(189 –> 269)

189 –> 192 Static camera
193 –> 194 Pan leftward
195 –> 261 Pan leftward downward
262 –> 269 Pan leftward

Shot 1
frames 1->13

Shot 2
frames 14->18

Shot 3
frames 19->111

Shot 4
frames 116->141

Shot 5
frames 142->188

Shot 6
frames 189->269

Shot 7
frames 279->300

Shot 3
(19 –> 111)

19 –> 25 Forward travelling
26 –> 95 Pan leftward
96 –> 111 Static camera

Shot 5
(142 –> 188)

142 –> 156 Pan leftward downward
157 –> 188 Forward travelling

Fig. 3. *Visualization of the video summary (a). A key-frame per detected shot is displayed (bottom row), and, for a user-selected shot, more information about its contents is supplied (mosaic image and sequence of camera motion types). Examples shown correspond to (a) shot 6, (b) shot 3 and (c) shot 5. The first frame and last frame numbers for each shot account for having suppressed frames within progressive transitions.*

frames 37 to 47, and reappears as a new object, because no long-term trajectory association technique is included. The same issue arises for the calendar. A low-cost extension to alleviate these problems for simple cases, considering each object in the scene is to be indexed by still image cues anyway, would consist in matching objects using these cues. Figure 5 shows the created synoptic view. Camera motion is correctly globally characterized as "left panning". The trajectories of the train and the calendar, and the piece-wise trajectory of the rolling ball are drawn as described in Section 6. Because they are temporarily motionless, two distinct objects are considered, both in the case of the ball and the calendar, hence the trajectories on the synoptic view. Descriptors assigned to meaningful objects have been summarized in Figure 6, and describe correctly their behaviour. Though Figure 6 only includes qualitative measurements, the trajectory could also be used for indexing, and retrieval on this cue could be done by a sketch. A few spurious regions are extracted and are not included in the table. They slightly perturbate the clarity of the synoptic view, but, with a view to indexing, the effect on memory occupation in the database is minor, and besides, they do not severely disturb the features found for the meaningful objects.

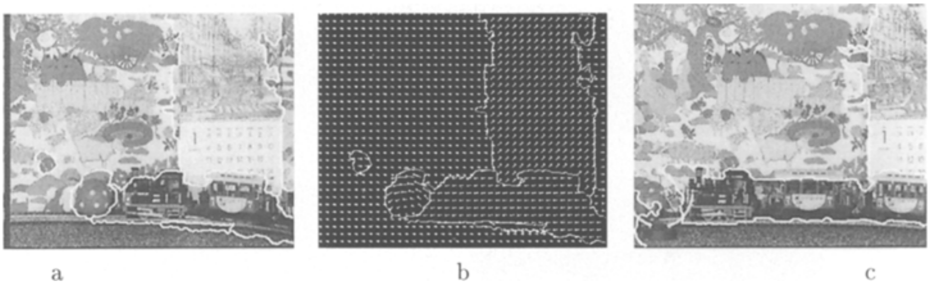


Fig. 4. Mobi sequence : extracted motion segmentation boundaries for frame 4(a) and associated estimated motion(b), motion boundaries for frame 120(c)

8 Conclusion

We have described a method for structuring the content of a video in several steps, namely video partitioning into shots, motion-based segmentation of each shot, and tracking of mobile objects. From this structure, we infer a technique for indexing and quick-viewing of the dynamic content. The method supplies ample descriptions of object and camera motions, that are relevant and useful for indexing. They are both of instantaneous and long-term, qualitative and quantitative nature, by exploiting the extracted motion information and trajectories. The various steps of the scheme exploit an efficient robust 2D motion

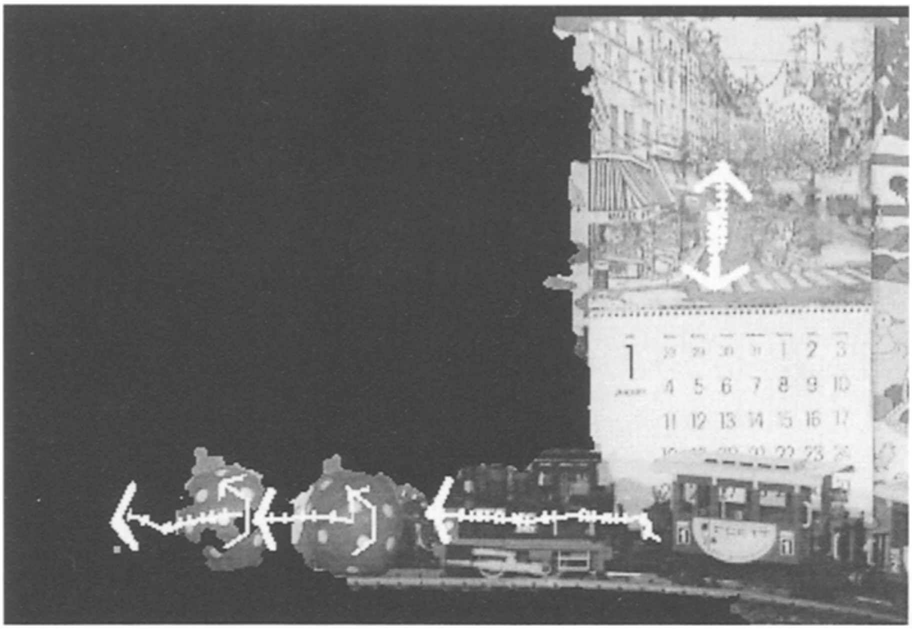


Fig. 5. Mobi sequence : Synoptic summary of the dynamic content of the shot. The background scene is discarded, mobile objects are displayed at their first position in the sequence, and their trajectories are drawn in the global coordinate system. The rotation of the ball is annotated by a curved arrow.

Name	Calendar 1	Calendar 2	Train	Ball 1	Ball 2
Icon					
Appearance frame	1	66	1	1	47
Disappearance frame	56	125	125	33	125
Apparent motion	up-left (1-56)	down-left (60-125)	none (tracked)	clockwise rotation (tracked)	clockwise rotation (tracked)
Global view motion	up (1-56)	down (60-125)	left	left clockwise rotation	left clockwise rotation

Fig. 6. Object motion descriptors

estimator, applied as appropriate on the whole image or on regions. So doing, the same motion measurements have multiple purposes in the structuring and description phases. Overall, the system is designed with a view to providing rich spatio-temporal information, coping with a broad range of scene contents, while keeping computational cost low.

Possible extensions are numerous. First, since objects are delimited, the cues used in still image indexing, such as shape, color, that are usually measured either globally, or on somewhat arbitrary regions, can then be computed on meaningful regions. In another direction, we are currently experimenting a long-term multi-trajectory approach to region tracking. This would enable additional description of spatio-temporal relations between mobile objects, such as occlusions and crossings, which are interesting from the semantic point of view, and improve the quality of long-term motion description in the case of piece-wise trajectories, or in cases as shown for the *Mobi* sequence.

References

1. P. Aigrain and P. Joly. – The automatic real-time analysis of film editing and transition effects and its applications. – *Computer & Graphics*, 18(1):93–103, 1994.
2. P. Aigrain, H.J. Zhang, and D. Petkovic. – Content-based representation and retrieval of visual media : a state-of-the-art review. – *Multimedia Tools and Applications*, 3(3):179–202, November 1996.
3. S. Ayer and H.S Sawhney. – Compact representations of videos through dominant and multiple motion estimation. – *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(8):814–830, August 1996.
4. M. Basseville. – Detecting changes in signals and systems - a survey. – *Automatica*, 24(3):309–326, 1988.
5. J.S. Boreczky and L.A. Rowe. – Comparison of video shot boundary detection techniques. – In *In I.K. Sethi and R.C. Jain, editors, Proceedings of IS-T/SPIE Conference on Storage and Retrieval for Image and Video Databases IV, Vol. SPIE 2670*, pages 170–179, 1996.
6. P. Bouthemy and F. Ganansia. – Video partitioning and camera motion characterization for content-based video indexing. – In *Proc. of 3rd IEEE Int. Conf. on Image Processing*, volume I, pages 905–909, Lausanne, Sept 1996.
7. C. Castel, L. Chaudron, and C. Tessier. – What is going on ? A high level interpretation of sequences of images. – In *4th European Conf. on Computer Vision,*, Cambridge UK, April 1996. – LNCS 1065.
8. J.D. Courtney. – Automatic video indexing via object motion analysis. – *Pattern Recognition*, 30(4):607–625, April 1997.
9. M. De Marsico, L. Cinque, and S Levialdi. – Indexing pictorial documents by their content : a survey of current techniques. – *Image and Vision Computing*, (15):119–141, 1997.
10. A. Del Bimbo, E. Vicario, and D. Zingoni. – Symbolic description and visual querying of image sequences using spatio-temporal logic. – *IEEE Trans. on Knowledge and Data Engineering*, 7(4):609–621, August 1995.
11. M. Flickner et al. – Query by image and video content : the QBIC system. – *IEEE Computer*, pages 23–32, Sept. 1995.

12. E. François and P. Bouthemy. – Derivation of qualitative information in motion analysis. – *Image and Vision Computing*, 8(4):279–287, Nov. 1990.
13. M. Gelgon and P. Bouthemy. – A region-level graph labeling approach to motion-based segmentation. – In *Proc. of Conf. on Computer Vision and Pattern Recognition*, pages 514–519, Puerto-Rico, June 1997.
14. F. Idris and S. Panchanathan. – Review of image and video indexing techniques. – *Jal of Visual Communication and Image Representation*, 8(2):146–166, June 1997.
15. M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu. – Efficient representations of video sequences and their applications. – *Signal Processing : Image Communication*, (8):327–351, 1996.
16. M. Irani, B. Rousso, and S. Peleg. – Detecting and tracking multiple moving objects using temporal integration. – In *Proc. of Second European Conference on Computer Vision*, pages 282–287, Santa Margherita Ligure, Italy, May 1992.
17. A. Nagasaka and Y. Tanaka. – Automatic video indexing and full-video search for objects appearances. – *Visual Database Systems II*, pages 113–127, 1992. – E. Knuth and L.M. Wegner (eds.), Elsevier Science Publ.
18. J.M Odobez and P. Bouthemy. – Robust multiresolution estimation of parametric motion models. – *Jal of Visual Communication and Image Representation*, 6(4):348–365, December 1995.
19. N.V. Patel and I.K. Sethi. – Video shot detection and characterization for video databases. – *Pattern Recognition*, 30(4):607–625, April 1997.
20. B. Rousso, S. Peleg, I. Finci, and A. Rav-Acha. – Universal mosaicing using pipe projection. – In *Proc. of IEEE International Conf. on Computer Vision (ICCV'98)*, pages 945–952, Bombay,India, January 1999.
21. H. Sawhney and R. Kumar. – True multi image alignment and its application to mosaicing and lens distorsion correction. – In *Proc. of Conf. on Computer Vision and Pattern Recognition*, pages 450–456, Puerto-Rico, June 1997.
22. C. Schmid and R. Mohr. – Combining greyvalue invariants with local constraints for object recognition. – In *Proc. of Conf. on Computer Vision and Pattern Recognition*, pages 872–877, San Francisco, USA., June 1996.
23. M.A Smith and T. Kanade. – Video skimming and characterization through the combination of image and language understanding techniques. – In *Proc. of Conf. on Computer Vision and Pattern Recognition*, pages 775–781, Puerto-Rico, June 1997.
24. C. Stiller. – Object-oriented estimation of dense motion fields. – *IEEE Trans. on Image Processing*, 6(2), February 1997.
25. J.Y.A Wang and E.H Adelson. – Representing moving images with layers. – *IEEE Trans. on Image Processing*, 3(5):625–638, September 1994.
26. H.J Zhang, A. Kankanhalli, and S.W. Smoliar. – Automatic partitioning of full-motion video. – *Multimedia Systems*, 1:10–28, 1993.
27. H.J Zhang, J. Wu, D. Zhong, and S.W. Smoliar. – An integrated system for content-based video retrieval and browsing. – *Pattern Recognition*, 30(4):643–658, April 1997.