Structuring Video Documents for Advanced Interfaces

Cordelia Schmid¹, Serge Benayoun³, Hélène Bernard³, Pascal Bertolino¹, Patrick Bouthemy², Marc Gelgon², Roger Mohr¹, Fabien Spindler²

¹ INRIA Rhônes-Alpes, IMAG
655, avenue de l'Europe, 38330 Montbonnot, France.
Pascal.Bertolino,Roger.Mohr,Cordelia.Schmid@imag.fr
² IRISA/INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France.
Patrick.Bouthemy,Marc.Gelgon,Fabien.Spindler@irisa.fr
³ Alcatel Alsthom Recherche
Route de Nozay, 91460 Marcoussis, France.
Helene.Bernard,Serge.Benayoun@aard.alcatel-alsthom.fr

Abstract

Video is traditionally sliced in shots, grouped in segments. On top of the shot structure there exist semantic links between objects extracted within a shot, i.e., the person visible here is also appearing there. How to automatically build such a structure is an open problem with already several partial answers. This paper presents the first version of our platform for reaching such a goal. It is mainly devoted to the construction of links between occurrences of objects, based only on image analysis. We shall discuss the construction of the video structure, and from there on two applications: a semi-automatic tool for building interactive videos and an interactive powerful browser.

1 Introduction

This paper addresses the problem of rapid access to visual information in video databases. This requires facilities such as the ability to have a quick overview of the video and focus on parts of interest, or to find occurrences of a particular person, object, or situation. The information added to the raw digital video data hence consists in structuring this data into meaningful entities (shots, people, objects...) and assigning them descriptors, which a database user could then use through a video navigation and query system. This requires that a person should have manually extracted all this information beforehand, which is very tedious or not feasible for very large databases. In this paper a framework and a prototype for indexing video databases is presented: automatic extraction of information aims at leaving the person indexing the video with far less work per video. We will show how the information extracted automatically can by itself be exploitable.

Since the prototype is aimed at a large range of audio-visual material (films, documentaries, TV news, ...), the various elements composing the video content analysis system should be applicable in general scene conditions: special transition effects between shots, mobile camera, several mobile zones of possibly significant area, articulated entities (e.g., people walking). Another requirement is that the algorithms should perform well enough to leave the database creator with only minor corrections to avoid as much tedious editing as possible.

The architecture of our prototype has some similarities with the system built at NTT ([19]). Figure 1 illustrates the different steps: first a the video is segmented into shots, then within each
shot objects are extracted and their corresponding signatures are stored in the index table. From there, the corresponding objects are searched and these correspondences build links within the video. From this final structure, different applications can be constructed.

![Diagram of video structuring process]

Figure 1: The three main steps in video structuring.

### 1.1 Related work

There is a growing number of emerging applications with a concern to provide systems that are integrated, work on real world databases and benefit from state-of-the-art development of computer vision. A review of image and video indexing techniques was recently proposed in [9]. A recent example of an integrated system using colour, texture and motion can be found in [22]. Work carried out towards browsing applications includes for instance cooperation of audio and video information for judicious content-based frame sub-sampling [17]. Strong interest has been devoted to create a mosaic image that captures in a still image the background scene as seen by a moving camera [15]. Abstraction of the dynamic content, i.e. maintaining in the summary important objects and their action or interaction, is a major issue, for which proposals can be found in [2, 10, 7].

It is was pointed out by [20] and [19], that overall structure on the top of the video is the key of future applications like browsing, edition, searching, etc. The system V-active produced by the company ephyx (http://www.ephyx.com) offers the possibility to view a video interactively. However, structure has to be added manual to the video which makes the preparation process tedious. This work does not provide any addition to the kind of structure that can be added to the
video, nor on how it can be exploited. Its key contribution is the use of efficient image processing techniques for partially automating the process of video annotation and structure.

As far as content based video analysis is concerned, extraction and qualification of significant content entities (shots, objects) is necessary. Automatic extraction and tracking of significant objects in videos is a difficult but critical step, that provides rich scene understanding and numerous indexing possibilities (e.g. action-based indexing [6], or retrieval based on motion descriptors [7]).

1.2 Outline of the paper

The first part of the paper briefly presents the image analysis tools used in our prototype. Section 2 describes the temporal partitioning into shots. The extraction of relevant objects in the video is presented in section 3. Moving objects are automatically extracted, based on a motion-based detection scheme. User interaction facilities allow to include static objects. On the extracted objects, local and global colour features are computed; they are stored in a huge index table, allowing us to search for new occurrences of the same physical object (cf. section 4). Using this index table and the motion-based detection results, a preprocessing of the video establishes links between shots, and shot descriptions are added as well.

Two applications are described in section 5. The first one is the automatic construction of an interactive video; as a result of this processing, objects can be designated interactively during the real time display, and from there it is easy to associate an action to this event. The second application is the intelligent browsing. Thanks to this structure, a video can be explored by skipping parts, or searching for an occurrence of a given object. We finally discuss directions for possible improvements.

2 Temporal partitioning of video into shots

The first step consists in delimiting the temporal basic video units, namely shots. Video partitioning into shots is often based on the comparison of color histograms of successive frames [1], possibly provided by the MPEG I-frames [14]. Satisfactory results have been obtained for the detection of transition effects, through the use of a two-level thresholding technique applied to histogram comparison [21], or the modeling of the temporal intensity change law during the transition interval [1]. However, because of possible variations in histograms due to strong camera motion or moving objects, and because there are a large variety of transition effects, test thresholds appear to be delicate to set. The motion-based technique presented in the following enables the detection of both cuts and progressive transitions using the same test, involving a single and fixed parameter. Its design makes it resilient to the difficulties mentioned above. The proposed method is based on the robust estimation of a global dominant apparent motion model between successive frames.

This global dominant motion is assumed due to camera motion. The temporal variations of the proportion of image data conforming this dominant motion model is closely related to the degree of change in scene content. We shall call this proportion $\eta$. Shot changes are detected as jumps in the value of $\eta$ using a statistical hypothesis test. This method enables the detection of both cuts and progressive transitions using the same test with the same parameterization involving in fact only one parameter. Besides, the estimation of the dominant image motion is straightforwardly reused to qualify the shot in terms of camera movement (static, panning, zooming, tracking, ... ) from statistical likelihood tests. For a more detailed description of the approach see [4, 5].

Figure 2 shows the results of partitioning of a video into shots. The mosaic (right side of the
figure) represents each shot by a thumbnail, the temporal median image. The strip (bottom of the figure) shows the shots of the video. Cuts and progressive transitions are differentiated by separate colours. Zooming into the strip allows to see which images belong to a shot. As shot extraction might fail, editing shots is possible in order to correct them.

Figure 2: An example for partitioning a video into shots. ¹

3 Extracting objects

3.1 Extraction of moving objects

Finding objects of interest in an image sequence is a fundamental issue, but still subjective. Considering colour or texture homogeneity to define a so-called object of interest often leads to having far too many detected objects, among other problems. Motion, on the other hand, is more restrictive and, broadly speaking, identifies the interest for an object to the fact that it is animated. We opt for this criterion.

The general case of moving object extraction is carried out through motion-based segmentation, i.e. partitioning of each image into zones with different motions. An alternative is motion-based detection of objects that are mobile relatively to the background scene (i.e. there are two classes, either mobile or conforming to background apparent motion). In this case, the regions of interest are identified to the connected components of the detection. We retain this latter possibility for the following reasons: good accuracy is obtained while maintaining computational cost low. The only

¹The video data was provided by INA, department of Innovation.
loss in generality of the method is concerned with moving objects that are connected in the image. However, this loss can be compensated for by an editing facility. An important point in favour of this detection is that a motion-based segmentation, that generally consider rigid motions, will identify, for instance, each limb of a walking person as a different region. In such a case, detection will extract the person as desired, i.e. as a single region.

The approach, fully described in [13], makes again use of the 2D affine motion model estimated for the shot change detection phase, which attempts to model the image motion due to camera movement. The residual apparent motion is then assumed to be caused by mobile objects. A detection label map is sought for, using a statistical contextual labeling approach involving a Markovian multi-scale model associated to the image grid. In addition to a carefully-designed motion conformity measure, spatial and temporal homogeneity constraints help improve result accuracy and stability and limit false alarms. The developed framework enables temporal tracking of the detected regions, temporal association being carried out by examining the proximity and spatial coincidence of successive regions. The method accounts for merging, splitting, appearing or disappearing regions.

3.2 Extraction of static objects

General object extraction from still images is almost impossible, even if some work on region extraction claims that they are achieving it. In these cases, it is in fact restricted to particular situations (performed on objects on a quite regular background), or with manual initialization of the outline [3].

We therefore limited this first version of the prototype to objects extracted manually by the user in a single image of the video. Once an object is manually extracted from any user-selected frame within the shot, it is automatically tracked for the entire duration of the shot, using the dominant background motion. This manual facility is required, because discrimination of “zone of interest” on the background is somewhat subjective.

Figure 3 displays an example where four objects have been extracted. The moving car is extracted automatically (notice that the associated blob includes the smoke coming out from the car); the two houses and the telephone booth are extracted manually. Editing facilities allow to improve the quality of the automatic detection algorithm.

4 Finding objects

Linking objects between shots is a typical matching problem; we would like to know if object A in shot X is similar to object B in shot Y. To solve this problem, we integrated two different matching methods in our platform. One method matches local characteristics between objects (cf. section 4.1). The other one compares color histograms of the object regions (cf. section 4.2). The results of the two methods are combined into a final matching score; the two methods are complementary and their combination allows to deal with a wide variety of scenes.

4.1 Matching using local characteristics

An object is characterized by a set of local feature vector. These vectors are differential invariants which describe the signal locally, are invariant to image rotations and absorb illumination changes.
These vectors are computed at interest points of the object that is at locations of the object which contain most information. A statistical distance (Mahalanobis) allows to compare feature vectors and to determine potential matches. Spatial coherence with neighbouring points constrains the matches and rules out most of the false matches. Finally, a voting algorithm decides for the most probable match, the second probable and so forth. For a more detailed description of the approach see [16]. The advantage of a local approach is its robustness to image variations. Even if one part of the image changes, the remaining common part can still be matched correctly. In the case of our application, it is possible to match objects correctly even if one of them is only partially visible. However, such an approach requires a minimum amount of structure in the object, else the local descriptors contain no information and it is impossible to distinguish between different objects. For unstructured objects the approach presented in the following works well.

4.2 Histogram comparison

Color histogram comparison has first been introduced by Swain and Ballard [18]. Many authors have since applied and improved this approach, as color has shown to be a very discriminant feature. The basic idea is to compare the color distribution of the pixels in the image. To be invariant to illumination changes, the color values for a pixel are normalized and the distribution is reduced to two dimensions. This distribution is described by a histogram and different objects are compared by histogram intersection. In general, histograms method require a segmentation of the objects. However, in our case we have already extracted the objects. To match different objects, we compute the histogram only for the pixels of each object and then compare those histograms. The histogram
based method will however fail if one of the objects is only partially visible. In this case the method described in section 4.1 will still work correctly.

4.3 Building classes

Once the matching between objects is performed, classes of similar objects are constructed. This is performed by computing the transitive closure of the graph obtained by the relation *is similar to*. Figure 4 displays different classes that are obtained in such a way. As such construction might fail, editing these classes is allowed in order to correct them.

![Figure 4: An example of classes.](image)

5 Applications and experimentations

In this section, we outline the video structure and present two applications of the prototype, one for indexing databases and one for users to look up indexed databases.

5.1 The internal video representation

The internal structure of a video consists of mainly three entities:

- the shots with the associated attributes
  - time position: start and end
- camera motion type
- global image signatures for static shots
- the set of objects appearing in this shot
- additional information not automatically filled (semantic descriptors on the kind of action and so on)

- the object occurrences present in a shot with the attributes
  - the corresponding shot where it appears
  - the kind of motion associated
  - time stamp of entrance and exit within the shot (might be different from the shot start and end)
  - signatures
  - the corresponding class as defined in 4.3
  - additional information not automatically filled by the system (name, etc.).

- the object classes; these classes have a much more semantic meaning; attached to them are
  - the different object occurrences
  - manually provided annotations like an object name, a textual description or a specific sound

Shot classes are not considered yet. Additional objects could be considered in the future like scenes which cluster shots into meaningful groups like it is done for instance in [8].

5.2 Building interactive videos

A first application is the construction of interactive videos. Existing tools need to design on each image of the video the parts which are sensitive to a click, and which event should be associated to them. Existing improvements allow interpolation on the sensitive region between frames.

Using the links between the images make this preprocessing much more easier. During the object extraction process, it is only required to extract an image of a given object we want to be sensitive to and to indicate the associated action. Then all similar views of this object are going to be detected and at this point it is easy to make them sensitive parts too. Of course, as a physical object may have different aspects, this process has to be repeated when the object appears differently. From our experiments, we came to the conclusion that in the worst case, twenty views are usually sufficient for covering efficiently all aspects; however such a result is shape dependent and sometimes, more views are needed.

5.3 Intelligent browsing

Many suggestions have been made for video browsing; most of them extract a few images from different shots and allow to navigate in this limited subset. A much more elegant solution is obtained by the analysis of the sound track and the images in order to extract significant parts of the video [17].
Here we suggest to exploit the previously created links between shots. Using the environment we have built, the user can ask to jump to the next or the previous shot, or he can search for the next/previous occurrence of an object and the associated information is displayed.

Using the links between shots, it is also possible to build higher level structure (see [8]). An easy case is to detect a dialog between two persons, even if they include some short additional illustration during this dialog. Putting such higher level on the top of the existing tool will then allow to browse quickly (from one higher level section to the other), or to browse regularly as described before, or just play part of each shot one after the other, by selecting representative subsequences.

5.4 Experimental results

The video analysis and pattern recognition tools adapted to form the system have previously been validated independently in their own context [5, 13, 16]. The point is then to evaluate their inter-dependent performance as a system. The experiments have been conducted on MPEG-1 reconstructed frames of TV video documents.

They have shown that special transitions between shots are numerous, particularly in advertisements or news, and that the proposed scheme achieves good results even in difficult cases. Mobile object detection is known to be a far more delicate problem, but encouraging results have been obtained, that validate the interest of automatic extraction over fully manual extraction. Computational cost remains reasonable, considering the processing is carried out off-line, the amount of information extracted and the degree of generality in the scenes that can be cope with.

Object identification has been extensively tested for large collections of static images; obtained results are very good (i.e. a recognition rate of more than 90%). Preliminary results for our video application are very good; more extensive testing is under progress. It is yet obvious that there exist cases for which neither our method nor any existing recognition method will work. A possible example is a person observed from a front and a side view. The possibility of editing the results is therefore crucial.

Figure 5 shows an integrated fully automatic example, i.e without manual correction. The moving ship is detected and extracted automatically, the result of the extraction is shown on the right side in the top row. Interest points used for the identification of the object are superimposed on this image (cf. method 4.1). The ship object is compared to a collection of approximatively 700 objects, some of them are displayed in the middle row. The algorithm identifies the object in the second row on the left as the most similar one. Interest points are superimposed on this object for illustration purpose. The identification of a similar object occurrence and its corresponding image allows to create a link.

6 Conclusion, discussion

In this paper an integrated system for constructing and playing a structured video has been presented. An high performance camera motion estimation method is used both to detect shot transitions of different nature and to track moving objects. Static objects (w.r.t. the camera motion) may also be automatically tracked from a manual initialization. Next objects are efficiently classified based on local and global color dependent features which leads to links between objects extracted in different shots. Each step of this process may be verified and corrected if necessary by user-friendly viewing and editing tools. The final overall structure allows efficient interaction with the video.
Figure 5: An example for automatic object extraction and automatic similarity search. Top row: image and extracted moving object. Middle row: some of the objects to which the ship object is compared; interest points are superimposed on the most similar object. Bottom row: the image which corresponds to the most similar object. The identification of a similar object occurrence and its corresponding image allows to create a link.
Future technical directions concern improvements in object tracking which is the critical point in our system, detection of other kinds of shot transition like special effects and extension of our classification method with other image features. We can also foresee the need for a special algorithms for people identification and matching [11, 12], as our tools are generic and therefore less adapted to a specific application. In addition, potential end-user applications like interactive advertising, electronic business, browsing in large video databases or content based video retrieval are going to be evaluated.

7 Acknowledgments

We would like to thank INA, department of Innovation, for providing the video sequences used in this paper.

References


