USER INTERACTION IN CONTENT-BASED VIDEO CODING AND INDEXING

Paulo Correia, Fernando Pereira

Instituto Superior Técnico - Instituto de Telecomunicações
Av. Rovisco Pais, 1096 Lisboa Codex, Portugal
Phone: + 351.1.8418463; Fax: + 351.1.8418472
E-mail: Paulo.Correia@lx.it.pt, Fernando.Pereira@lx.it.pt

ABSTRACT

The current level of request for applications involving content-based video coding and video information retrieval is increasing the relevance of systems able to somehow 'understand' the content of video sequences. This type of systems will enable the provision of services where enhanced user interaction with the visual content is supported. Such an object-based video coder is being specified by ISO under the MPEG-4 project [1], while the standardization of video description capabilities to support content-based video indexing and retrieval is being considered by MPEG-7 [2].

In this paper, we discuss the issue of video content identification and characterization, and the importance of providing the means for the user to interact with the analysis process so that the achieved results can have a meaningful and powerful semantic value.

1 VIDEO ANALYSIS FOR CONTENT-BASED CODING AND INDEXING

To take advantage of the potential, in terms of new applications, created by the emerging MPEG-4 and MPEG-7 object-based coding and indexing standards, video content must be structured as a composition of objects for which a number of features are available. In some cases, these data are made available by the visual information production process itself (e.g. when using chroma keying techniques together with manual annotations for describing contents). In other cases, this type of information is only partly available (or not available at all) and thus some video analysis must be performed.

The objective of the video analysis task usually consists in identifying the relevant objects that compose a scene - *segmentation*, and in extracting relevant features for the individual objects or for the composed scene. The resulting analysis data can be used both for content-based video coding and indexing. To provide the requested content-based functionalities, analysis results should be consistent in time, guaranteeing a correct tracking of the segmentation partition labels, and an appropriate handling of the extracted features.

A list of potentially useful video analysis results is [3]:

- *Segmentation* of the scene;
- *Tracking* of objects along the sequence;
- *Prioritization* of objects;
- *Depth ordination* of objects;
- Spatial and temporal *composition data*;
- Detection of *scene cuts*;
- Detection of the *presence of a certain object* (or type of object) in the sequence;
- *Classification of the scene*, e.g. into sports, life music, news, etc.

Also a number of relevant analysis results associated to each specific object can be listed:

- *Shape* information;
- *Motion* information;
- Adequate *temporal resolution* (object rate);
- Adequate *spatial resolution*;
- Adequate *quality* (e.g. SNR);
- Adequate *scalability* layers;
- Special needs for *protection against errors*;
- *Indexing features* related to size, shape, motion, color, first and last images where the object is present in the sequence, etc.;
- Identification of *scene cuts* and *key (object) images*;
- Information for *sprite generation*.

The lists above contain analysis results that may be useful for coding, for indexing, or for both purposes, depending on the type of application envisioned.

How easy or difficult is the extraction of the desired analysis results is to a great extent conditioned by the amount of *a priori* information available in each specific application case. For some specific applications, it is possible to find a set of automatic analysis tools that perform segmentation and feature extraction as desired. This may be the case of a surveillance application, where the analysis system provides to the video encoder detailed information about the intruder object (even if with a low contour precision), allowing the selective coding of the scene. For this purpose, a simple, real-time, fully automatic analysis scheme based on the detection of moving objects may be employed (see figure 1 b). However, for many applications, only part of the desired analysis results can be automatically achievable. Even if the automatic detection of moving objects with precise contours may be sometimes achieved, e.g. by combining the partial results from tools based on the homogeneity of both motion, and texture (see figure 1), and

a) *b)* *c)* *d)*

Figure 1 - Example of fully automatic segmentation: a) original QCIF image; b) change detection result; c) texture segmentation (based on YUV homogeneity); d) combination of b) with c).

some indexing features, like a description of the trajectory of a moving object, can also be automatically estimated, there are many cases where fully automatic segmentation and feature extraction tools do not provide acceptable results. For example, the performance of automatic tools for the segmentation of a complex video scene, or for its classification (e.g. into indoors, outdoors, or sports, news, movie, etc.), may not be so good, and thus some guidance from the user may be determinant to achieve useful results. We can then conclude that the type of video analysis techniques to use depends very much on the application considered, not only in terms of the choice of the most effective automatic tools, but also on the role that user guidance may assume.

Considering the most relevant applications [4], it is possible to classify them according to the type of video analysis that they need/allow [5]:

- *Real-time, fully automatic* analysis - e.g. videotelephony without any user guidance to the analysis process.
- *Real-time, user guided* analysis - e.g. videotelephony with some user guidance; for example, the user may be allowed to mark on the screen the relevant objects to be identified. This user guidance can be given by the sender or by the receiver, if a back channel is available.
- *Off-line, fully automatic* analysis - corresponds to the situation where a computationally very expensive, not real-time, automatic segmentation or feature extraction process is implemented, e.g. content creation for a database (unlikely case due to the uncertainty of results from the fully automatic processing).
- *Off-line, user guided* analysis - e.g. content creation for a video database where the quality of the analysis results is critical and thus some user interaction, for coding and indexing, is used to guarantee meaningful results.

When dealing with real world video sequences, the achievement of acceptable video analysis results by using only automatic analysis techniques cannot be guaranteed, except for the cases where enough *a priori* knowledge is available. This is however not always the case and, unless a more flexible and powerful video analysis framework is available, notably supporting user guidance whenever possible, the analysis results may be of very poor quality. For all the applications where the quality of the analysis results is critical and some user guidance is possible, such as off-line applications, not allowing the user to interact with the video analysis process would just be a waste of powerful tools, preventing the achievement of the best possible results.

2 TYPES OF USER INTERACTION

From the previous discussion on analysis for content-based video coding and indexing, we can conclude that an effective video analysis framework needs to include not only the best automatic analysis tools available, but also the possibility for the user to guide and improve the automatic process.

Interaction should be limited to the minimum possible, mainly allowing to efficiently set “in track” the automatic analysis process, and to (sometimes) refine the automatically produced results. Interaction is typically performed for key images (usually the first, and eventually those where new objects enter the scene, or where relevant features are triggered), creating “good” analysis seeds that will be tracked along time, thus constraining the posterior automatic analysis. Some automatic algorithms can improve themselves, becoming more efficient, by learning from the user guidance that is supplied to correct their results.

Assuming that user interaction is an essential part of a powerful video analysis framework, two different types of user interaction are then considered:

- *Initial user interaction* - to partly drive the automatic analysis process, allowing to improve its performance for the rest of the time (see figure 2);
- *User refinement* - to let the user supervise the evolution of the analysis results, correcting the undesired deviations whenever needed, and ideally as little as possible (see figure 3).

The adequate form of user interaction to support depends on the application, and may assume very different natures. Among other elements, it needs to take into account the type of segmentation to achieve and the features to extract.

3 USER ASSISTED SEGMENTATION

The interaction of the user with the segmentation process can be done in quite simple ways (like the specification of a few numerical parameters), or it can require a sophisticated user interface (e.g. supporting drawing capabilities).

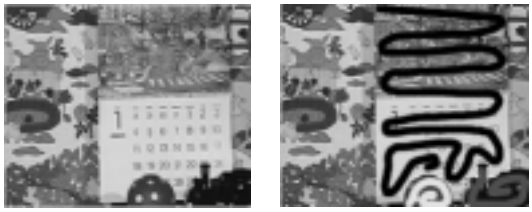

Figure 2 - Initial user interaction to mark the image area corresponding to an inhomogeneous relevant object

Figure 3 - User refinement by merging automatically extracted regions (e.g. by mouse clicking) to define quite inhomogeneous objects

In terms of initial user interaction, the interface with the user may be required to support the following actions:

- Definition of the target number of objects;
- Definition of a set of constraints that the relevant objects must respect, e.g. position in the image, size, shape, orientation, color, type of motion;
- Drawing of a rough outline of the relevant objects over the first original image;
- Marking the relevant objects over the first original image, e.g. by means of crosses or lines;
- Improvement of a fully automatic segmentation for the first image, by merging, splitting and correcting the boundaries of the regions found, in order to identify the desired objects for further tracking.

As for user refinement of segmentation results, and although it should be needed as little as possible, its use may be crucial to help automatic tools at critical time instants, like scene cuts, occlusion periods, light changes, etc. Possible ways of user refinement for segmentation are:

- Merging and splitting automatically detected regions, in order to define relevant objects;
- Identification of new objects;
- Adjustment of regions/objects boundaries.

4 USER ASSISTED FEATURE EXTRACTION

User interaction in the context of feature extraction can assume an even more important role than for segmentation since many high level indexing features (usually related to quite abstract video characteristics) typically require fully manual or, at least, semi-automatic extraction. A typical example is scene classification in the context of off-line content creation for video databases.

Like in the case of segmentation, interaction may serve not only to set the features but also to refine those automatically extracted. The requirements on the user interface (and thus on how sophisticated it needs to be) depend on the applications envisioned. Its complexity may range from the simple input of alphanumerical data to an interface with sophisticated graphical capabilities.

Possible forms of initial user interaction in the context of object-based coding and indexing are:

- Identification of scene cuts;
- Choice of key (object) images, to serve as basis for indexing and coding;
- Identification of the images in which a certain (type of) object appears;
- Setting a priority label for each object;
- Setting the depth order for the objects;
- Setting the desired quality and resolutions for each object;
- Selection of the scalability layers for each object;
- Identification of special error protection needs for each object.

User refinement of features often becomes essential due to the difficulty of automatic tools to reach the desired results, notably when dealing with high level features. Examples of actions to be supported by means of user refinement of extracted features are:

- Correction of automatic content classification;
- Correction of automatically attributed priority labels, scalability layers, resolutions, etc.;
- Correction of automatically attributed indexing information, such as motion trajectories, shape descriptions, etc.;
- Addition to and removal of automatically detected scene cuts.

While there are some forms of user interaction that are only acceptable for off-line applications (e.g. it would be

a) *b)*

Figure 4 - a) User interface of a video analysis framework supporting user interaction showing the original image, the automatic segmentation mask and the mask boundaries overlaid on the luminance of the original image. b) The segmentation mask (two objects) after some user interaction with the automatic result.

quite difficult to manually detect scene cuts or to choose key-frames for indexing, in real-time conditions), there are also ways to introduce user guidance in real-time environments, like the simple marking of (semantically) relevant objects.

5 RESULTS AND CONCLUSIONS

This paper claims that a powerful video analysis framework needs to simultaneously use the best automatic analysis tools available and to allow user interaction. Interaction, whenever supported by the application, permits to further control the analysis process and to refine its results. Thus user interaction appears not as a substitute for mediocre automatic analysis tools but rather as a complement to overcome difficult cases.

In order to show the role and advantages of user guidance in video analysis for coding and indexing, as described in this paper, an automatic video segmentation framework, developed by the authors [6], including texture, motion and tracking analysis modules is taken as an example.

The user interface of the developed video analysis application is displayed in figure 4 *a)*. The displayed image illustrates the user interaction for the refinement of segmentation results (usable both to correct analysis results at any frame and to create an initial partition based on a pre-segmentation mask). Refinement actions supported include split and merge actions on regions, as well as the possibility to correct region boundaries. Also the introduction of new objects is supported. In this environment, the user can interact with the segmentation results in an intuitive way. The application displays the original image, the segmentation partition, and also an additional image with the luminance of the original image and a uniform color for each segmented region. This additional image (together with the zoom option) is very useful for the manual correction of the object boundaries.

The developed application handles one image at a time; this means that whenever the automatic analysis deviates from the desired results, the user may stop the automatic process and open the interaction mode to correct the analysis results before giving again the control to the automatic mode.

The analysis environment was tested with different types of video sequences in terms of motion and detail, such as classified within MPEG-4, and we concluded that user interaction is of major importance in ensuring the achievement of correct analysis results. The interface developed includes a number of generic useful features. However, if a specific application domain is targeted, a set of more tuned user interface modules could be added.

Finally, the analysis environment here described also includes some authoring capabilities, e.g. resize, rotate, change position, etc. allowing to create new scenes, which may be after coded with an MPEG-4 video codec. This codec encodes the video objects previously defined, and also takes benefit of some of the extracted features.

REFERENCES

[1] MPEG Video Group, "Final Text of FCD 14496-2 Visual", Doc. ISO/IEC JTC1/SC29/WG11 N2202, Tokyo MPEG meeting, March 1998

[2] MPEG Requirements Group, "MPEG-7: Context and Objectives", Doc. ISO/IEC JTC1/SC29/WG11 N2207, Tokyo MPEG meeting, March 1998

[3] P. Correia and F. Pereira, "Video Analysis for Coding: Objectives, Features and Methods", 2nd Erlangen Symposium on 'Advances in Digital Image Communication', Erlangen-Germany, April 1997, pp. 101-108

[4] F. Pereira and R. Koenen, "Very Low Bitrate Audio-Visual Applications", Signal Processing: Image Communication Journal, vol.9, nº.1, November 1996, pp. 55-77

[5] P. Correia and F. Pereira, "The Role of Analysis in Content-Based Video Coding and Indexing", Signal Processing, Special Issue on Video Sequence Segmentation for Content-Based Processing and Manipulation, Vol. 66, issue 2, Elsevier, 1998

[6] P. Correia and F. Pereira, "Segmentation of Video Sequences in a Video Analysis Framework", Workshop on Image Analysis for Multimedia Interactive Services, Louvain-la-Neuve, Belgium, June 1997, pp. 155-160