OBJECT AND EVENT DETECTION FOR SEMANTIC ANNOTATION AND TRANSCODING

M. Bertini[1], R. Cucchiara[2], A. Del Bimbo[1], A. Prati[2]
[1]D.S.I. – Università di Firenze – Italy
[2]D.I.I. – Università di Modena e Reggio Emilia – Italy

ABSTRACT

Video annotation provides a suitable way to describe, organize, and index stored videos. On the other hand, transcoding aims at adapting content to the user/client capabilities and requirements. Both cues are now mandatory, given the tremendous demand of multimedia access from remote clients, in particular nowadays that new terminals with limited resources (PDAs, HCCs, Smart phones) have access to the network. In this paper we propose an unified framework to define event-based and object-based semantic extraction from video to provide both semantic video annotation for video stored and semantic on-line transcoding from live cameras. Two case studies (highlights' extraction from soccer videos for the annotation and people behavior detection in domotic application for transcoding) and corresponding experimental results are reported.

1. INTRODUCTION

Annotation and transcoding are two very common keywords in multimedia. Annotation is the process of manual or automatic association of meaning's description to multimedia data (in our case videos); it is the initial step for defining many multimedia services such as efficient video storing, content-based retrieval, query-based information extraction, retrieval by example and so on. Nowadays the problem of extracting high-valued semantics from the video in a way and with a language that could meet the user's requirements and wishes has to be addressed and solved. Certainly the user knows the *object* he wants to see, i.e. the subject of the action; moreover, talking of videos, the user owns the more powerful concept of *event* that correlates the object and the action. Events and associated objects must be exploited in the video description to improve the level of semantics embedded into the data.

Thus, the annotation provides a suitable way to describe, organize and index video archives, adding or, better, extracting and emphasizing the describable semantics that could be further used to reply to the user's queries. The main goal of the researches in this field is to define an automatic approach, guided by a suitably defined ontology. In particular, research in the field of detection and recognition of sport highlights in videos is motivated by the strong interest shown by broadcasters, who are interested in systems that ease the process of annotation of the huge amount of live and archived video materials. Among the many sports types, soccer is for sure one of the most relevant - at least in Europe. Within scope of the ASSAVID project, a number of tools supporting automatic annotation of sport videos were developed. In this context, the *sport highlights* are the *events* used in the annotation. The problem of detection and recognition of highlights in sport videos is an active research topic. Among sports that have been analyzed so far, we can cite soccer ([1, 2]), tennis ([3]), basketball ([4]), baseball ([5]), American football. In this paper we report on our experience in the classification of soccer highlights, using an approach based on temporal logic models.

On the other side, the user would like to directly access to multimedia data in the best way according with its capabilities, in terms of display, processing resources, and bandwidth. This problem, well known as the supply of UMA (Universal Multimedia Access), is very critical due to the large amount of different clients (such as PDA, HCC, smart phones and so on). In this framework, the transcoding process aims at providing a suitable way to change the multimedia format according with the user requirements and constraints.

Semantic transcoding assumes that the user does not want to access to all the data, but only to the data semantically useful. In this case, the possibility to select *not only what to see* but also *what to see better* should be a winning strategy. Therefore, semantic transcoding techniques are spreading [6, 7, 8].

In this paper, we propose to define the video semantics in terms of *events* and *objects* depicted in the scene. The concept of event is very important in the annotation phase to help indexing the video, whereas for the transcoding phase is more frequent that the user expresses his preferences by means of which object(s) is interested in. However, as well as the annotation can be based also on the definition of objects of interest, so the event can be important for the transcoding phase if the user does not want to see hours of video. In this case, by means of event(s), the user can select only parts of the video: the transcoding at "object-level" will be then applied only to those parts.

To this aim, we define the *classes of relevance*, i.e., the set of events and objects in which the user is interested, and their corresponding weights.

In general, a *class of relevance C* is defined as a pair $C = <o_i , e_j>$, where o_i represents an object (as class and not as single instance of the class) and e_j is an event (also in this case as class), selected between the set of objects O and events E detectable by the system:

$$O = \{o_1, o_2, \dots , o_n, *\} E = \{e_1, e_2, \dots , e_m, *\}$$

where * is a special class (for both the objects and the events) that indicates all the objects/events.

2. EVENT-BASED SEMANTIC ANNOTATION OF SOCCER VIDEOS

Inspection of tapes totaling over 20 hours of video showed that producers of videos use a main camera to follow the action of the game; since game action depends on the ball position, there exists a strong correlation between the movement of the ball and camera action. The main camera is positioned along one of the long sides of the playing field. Our method has been tested using several soccer videos containing a wide range of different video editing and camera motion styles, as produced by several different international broadcasters. Considering a variety of styles is of paramount importance in this field, as otherwise the system lacks robustness. In fact, videos produced by different directors display different styles in the length of the shots, in the number of cameras, in the editing effects.

Identification of the part of the playing field currently framed and camera action are among the most significant features that can be extracted from shots taken by the main camera; these features can be used to describe and identify relevant game events. Typical actions featured by the main camera are: pan, tilt, and zoom. Pan and tilt are used to move from a part of the playing field to another one, while zoom is used to change the framing of the subject.

Events that we have been elected for thorough investigation are E = {forward launches (FL), shots at goal (SG), turnovers (TO), placed kicks (PK)}. Placed kicks comprise penalty kicks, free kicks next to the goal box, and corner kicks. These are typical highlights shown in TV news and magazine programs summarizing a match, even if these actions do not lead to the scoring of a goal.

The soccer playfield has been divided in 12 zones, 6 for each side (Fig. 1). The features used to recognize the playfield zones are playfield lines and the playfield shape. They can be the objects that are used to detect events or to describe and annotate the scene for user requirements. They can be classified in the set of objects O = {F, O, R, C, M}, where F = playfield shape descriptor, O = playfield line orientation descriptor, R= playfield size descriptor, C = playfield corner position, M = midfield line descriptor.

A Naive Bayes classifier has been used to classify each playfield zone *Zx*. Playfield zone classification is performed through the following steps: playfield lines and shape are extracted from the image, descriptors are then calculated and the observation values of the classifier variables are selected. The classifier with the highest confidence value (if above a threshold) is selected.

As above-mentioned, camera parameters are strongly related to ball movement. Pan, tilt and zoom values are calculated for each shot. The curves of these values are filtered and quantized.

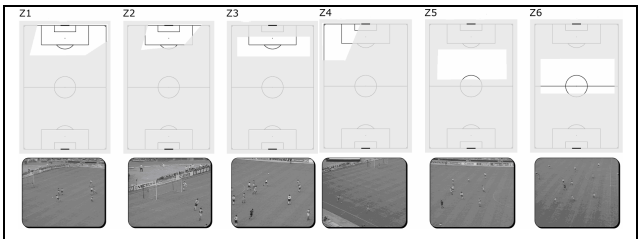

Fig. 1: Playfield zones, Z7 to Z12 are symmetrical

Analysis of the videos has shown that conditions such as a flying ball rather than a change of direction, can be observed from camera motion parameters. Through heuristic analysis of these values, three descriptors of *low*, *medium* and *high* ball motion have been derived. Information provided from playfield zone recognition and camera motion analysis is used to create temporal logic models of the highlights. In fact, the highlights can be characterized by the playfield zone where they happen, and how the action develops through the playfield. For instance, the forward launch requires that the ball moves quickly from midfield toward the goal box. The models of all the actions have been implemented as FSM (finite state machine). Fig. 2 shows the FSM for the goal shot.

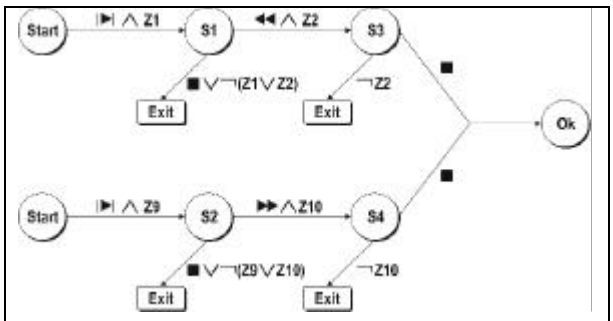

Fig. 2: Shot model: the arcs report the camera motion and playfield zones needed for the state transition.

3. SEMANTIC TRANSCODING WITH CLASSES OF RELEVANCE

The transcoding system is based on a set of techniques used to adapt the video to the user's requirements. Besides

normal transcoding policies that implement spatial reduction (to fit display size), temporal downscaling (to meet bandwidth constraints), and color reduction (to fit display color depth), this module includes many computer vision and image processing techniques able to extract from the video significant features. The set of available techniques defines the set of selectable classes of relevance.

By using the preferences, the user can assign to each class a relevance measure by means of weight (from 0 to 1, where 1 is the more relevant) [7, 9]. As an example, in the case of soccer videos, the user can assign the maximum value of relevance to the shots at goal (as event), or to the shots at goal in the "left corner" (as event plus object). Another example can be a domotic application in which the user is more interested in detecting when a person falls (event – fall – plus object – person). Thus, provided the set of objects O and events E as follows:

O = {foreground objects (FO), foreground people (FP), face of foreground people (FF), background (BG)}
E = {people walking (PW), people sitting (PS), people falling down (PF), people lying (PL)}

the user can select between available classes showing its preferences. For instance, $C=\{C_1, C_2, C_3\}$ with
C_1 = <FP, PL>; C_2 = <FP, *-PL>; C_3 = <BG, *>, and weights = $\{w_1,w_2, w_3\}$ = {0.7, 0.2, 0.1} are possible preferences, i.e., the user is very interested in having the fallen person (C_1) at the best quality possible, while in other situations the person (C_2) can be sent with lower quality.

The transcoding system can be programmed differently for each classes. For example, in the case of class C_1 there should not be any temporal compression (we want to see each frame of a fallen person) and very low coding compression to help recognizing the person. Similarly, in the case of class C_2 only the frames in which the scene changes considerably are sent at high resolution. Finally, for the class C_3 we can use a static or dynamic frame skipping by sending only the background every N frames.

4. EXPERIMENTAL RESULTS ON CASE STUDIES

4.1. Case study of semantic annotation: results on soccer videos

The system has been tested on about one hour of videos, separated in 80 sequences, selected from 15 European competitions. Table I reports the results: it can be noticed that those events that can be easily defined (placed kicks and shots at goal) obtain a good recognition rates. Recognition of shots at goal is very good, specially if we consider that the false detection is due to attack actions near the goal box. Also the good result of forward launches detection is encouraging, since this highlight is usually very similar to other actions.

It must be noticed that while this highlight is not important per se, it is linked to other important highlights, such as counterattacks. The recognition of turnovers is critical since its definition is quite fuzzy; probably adding other features (such as player position) would reduce the false detection.

TABLE I – HIGHLIGHT CLASSIFICATION RESULTS.

	Detect.	Correct	Miss.	False
Fwd. launch	36	32 (95%)	1 (3%)	4 (6%)
Sh. at goal	18	14 (93%)	1 (6%)	4 (6%)
Turnover	20	10 (50%)	3 (5%)	10 (50%)
Placed kick	13	13 (87%)	2 (13%)	0

4.2. Case study of semantic transcoding: results on domotic application for PDAs

As a reference framework for semantic transcoding we used that of domotic applications in which tele-presence and tele-viewing are essential for the safety of disabled people. The scenario we envisioned is that of the staff taking care of a disabled person that uses a PDA to continuously monitor from everywhere his/her state, or only when an event occurs. Since PDAs have limited resources and (typically) a low Internet connection, this system can heavily benefit from the application of video transcoding.

In these experiments, we aim at comparing different transcoding policies in order to demonstrate the effectiveness of our proposal. In previous works [7, 9] we compared our semantic transcoding with classical transcoding policies, i.e. spatial, temporal, color and code downscaling. In this paper we limited our comparison to the case of spatial and coding.

The comparison is made by using a modified version of the PSNR (Peak Signal-to-Noise Ratio) in which we take into account the weights of the classes of relevance. In the definition of the PSNR we substitute the MSE (Mean Square Error) with a WMSE (Weighted MSE) that takes the weights assigned to the classes into account [9]. Moreover, the bandwidth required is used as additional figure for the comparison.

Semantic object-based transcoding consists in using different compression in different classes of relevance as in [7]. Objects are compressed with a compression rate inversely proportional to the interest the user has for that class. Objects of low interest, such as background, can further be modified with a temporal downscaling, for instance sending it only when changes considerably.

Moreover, for low-performance clients, we develop a new transcoding policy, called *semantic spatial*

transcoding that first fits the video with the client's display size, and then compresses the video by using a semantic coding, preserving the resolution and the quality of the video. First two images of Fig. 3 show some results of this method, while the other two images report the view at the same size and bandwidth using only spatial transcoding. In this case, it is possible to view the whole scene, but with very lower quality.

Fig. 3 frames with semantic spatial transcoding (first and second) or spatial transcoding (third and fourth)

Table II reports the results of this comparison. For the classes of relevance, we use three sets of weights. In the first, called w/o semantic, we actually do not have the weights, i.e., there is no semantics in the video relevant for the user. In the second case, the user is very interested in the head/face of the moving person, and slightly interested in the body. The background has no importance. A typical case can be a video-surveillance application that aims at detecting people. In the third setup, the whole moving person (body plus head) is very relevant, but the user is slightly interested in the background too. From the experimental results it is possible to depict that the classical coding policies, such as JPEG (at different compression factor C) and MPEG2 (with high compression) can achieve very good bandwidth occupation but severely affect the quality of the video. This is particularly harmful if further processing steps are required: though the image does not seem deteriorated by human eyes, it can become useless for an image processing task. Table II shows how our semantic coding is able to even abate a little the bandwidth and, at the same time, achieve an excellent PSNR. Fixed spatial downscaling can decrease the bandwidth but it achieves poor PSNR performance.

Finally, the results show that the combination of semantic spatial and semantic coding transcoding can further improve the PSNR performance at a low bandwidth cost (less than 70 kbps w.r.t. the case of only semantic coding) and thus is a good strategy for the domotics applications accessed thru low resource devices, such as PDAs.

5. CONCLUSIONS

In this paper we have presented a system that recognizes soccer events using an approach based on temporal logic. Experimental results are extremely encouraging.

TABLE II. NUMERICAL RESULTS OF THE COMPARISON (ORIGINAL BANDWIDTH=24304.22 KBPS)

Transcoding policy	w/o semantic	w_i=[0, 0.1, 0.9]	w_i=[0. 1, 0.9]	Bandwidth (kbps)
JPEG (C=20)	41,09	38,05	38,42	1342,42
JPEG (C=80)	30,35	27,54	28,05	511,28
MPEG	10,38	9,91	10,44	210,00
Semantic Coding	31,37	29,98	25,58	118,32
Fixed spatial (C=20)	26,89	24,85	26,11	494,34
Semantic spatial+Coding	33,98	31,32	32,23	181,74

The events and the objects detected can be used for an automatic annotation to support user queries or for a semantic transcoding. In this case our experiments confirmed that the exploitation of semantics in the adaptation of videos to user requirements and interests is suitable to abate bandwidth and maintain the quality of the multimedia services.

This work is partially funded by the "Domotics for disability" by Fondazione CRM.

REFERENCES

[1] S.Choi, Y.Seo, H.Kim, K.-S.Hong, "Where are the ball and players? : Soccer Game Analysis with Color-based Tracking and Image Mosaic", in *Proc. of Int'l Conf. Image Analysis and Processing (ICIAP'97)*, 1997

[2] Y. Gong, L.T. Sin, C.H. Chuan, H. Zhang, and M. Sakauchi, "Automatic Parsing of TV Soccer Programs", in *Proc. of the Int'l Conf. on Multimedia Computing and Systems (ICMCS'95)*, Washington, D.C, May 15-18, 1995.

[3] G. Sudhir, J.C.M. Lee, A.K. Jain, "Automatic Classification of Tennis Video for High-level Content-based Retrieval", in *Proc. of the Int'l Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, 1998

[4] S.Nepal, U.Srinivasan, G.Reynolds, "Automatic Detection of 'Goal' Segments in Basketball Videos", in *Proc. of ACM Multimedia*, pp. 261-269, 2001

[5] Y.Rui, A.Gupta, A.Acero, "Automatically Extracting Highlights for TV Baseball Programs", in *Proc. of ACM Multimedia*, 2000

[6] Nagao, K., Shirai, Y., Squire, K., "Semantic annotation and transcoding: Making web content more accessible", *IEEE Multimedia,* vol. 8, 2001, pp. 69–81

[7] Cucchiara, R., Grana, C., Prati, A., "Semantic Transcoding for Live Video Server", in *Proceedings of ACM Multimedia*, 2002, pp. 223-226

[8] Vetro, A., Divakaran, A., Sun, H., Poon, T., "Adaptive transcoding system based on MPEG-7 meta-data", in' *Proc. IEEE Pacific-Rim Conference on Multimedia*, 2002

[9] Cucchiara, R., Grana, C., Prati, A., "Semantic Video Transcoding using Classes of Relevance", *International Journal of Image and Graphics*, vol. 3, no.1, Jan. 2003, pp. 145-169