EXPLOITATION OF TEMPORAL DEPENDENCIES OF DESCRIPTORS TO EXTRACT SEMANTIC INFORMATION

A. Bonzanini, R. Leonardi, P. Migliorati,

DEA - University of Brescia , Via Branze, 38 - 25123, Brescia, Italy
Tel: +39 030 3715433; fax: +39 030 380014
e-mail: pier@ing.unibs.it

ABSTRACT

Nowadays the problem of semantic video indexing is of great interest due to the wide diffusion of large video databases. In this paper we present a semantic video indexing algorithm based on finite-state machines that exploits the temporal dependencies of low-level descriptors extracted from the MPEG compressed bit-stream. The performance of the proposed algorithm have been evaluated considering the semantic indexing of soccer games video sequences, and the simulation results have shown the effectiveness of the proposed algorithm.

1. INTRODUCTION

Nowadays we are able to store and transmit large quantities of multimedia documents, thanks to the development of devices with large storage and transmission capabilities and to the use of efficient data compression techniques.

The need of an effective management of this material has brought to the development of semantic indexing techniques [1] [2] useful for various applications, such as, for example, electronic program guides.

Figure 1 illustrates how a man uses its cognitive skills to face the semantic indexing problem, while an automatic system can face it in two steps [3]: in the first step some low-level indices are extracted in order to represent low level information in a compact way; in the second step we need a decision-making algorithm to extract a semantic index from the low-level indices.

The problem of low-level descriptors extraction is widely discussed in the literature [4], whereas the problem of finding decision-making algorithms is less addressed [5]. Furthermore it is quite clear that this problem can be partially solved only for particular types of sequences, i.e., with specific program categories.

In this paper we present a semantic video indexing algorithm based on finite-state machines that exploits the temporal dependencies of low-level descriptors extracted from the MPEG compressed bit-stream.

In particular we have applied the proposed algorithm to the analysis of soccer video sequences in MPEG-2 compressed format. For these type of sequences the semantic content can be related to the presence of interesting events such as, for example, goals, shots to goal, and so on. These events can be found at the beginning or at the end of the game actions. A good semantic index of a soccer video sequence could be therefore a summary made up of a list of all game actions, each characterized by its beginning and ending event. Such a summary could be very useful to satisfy various types of semantic queries.

We have chosen three low-level descriptors which represent the following characteristics: (i) lack of motion, (ii) camera operations (represented by pan and zoom parameters) and (iii) the presence of shot-cuts. We have then studied the correlation between these descriptors and the semantic events defined above, and we have found that they are really meaningful.

The proposed algorithm exploits this correlation in order to detect the presence of goals and other relevant events in soccer games video sequences.

The simulation results have shown the good performance of the proposed algorithm.

The paper is organized as follow. In Section 2 we discuss the selected low-level descriptors, whereas in Section 3 the proposed algorithm is presented. In Section 4 we report some experimental results, and final conclusions are given in Section 5.

2. THE LOW-LEVEL DESCRIPTORS

As previously mentioned, we have considered three low-level descriptors associated to each P-Frame, which represent the following characteristics: (i) lack of motion, (ii) camera operations (represented by pan and zoom parameters) and (iii) the presence of shot-cuts.

Lack of motion has been evaluated by thresholding the mean value of motion vector module μ, given for each P-frame by

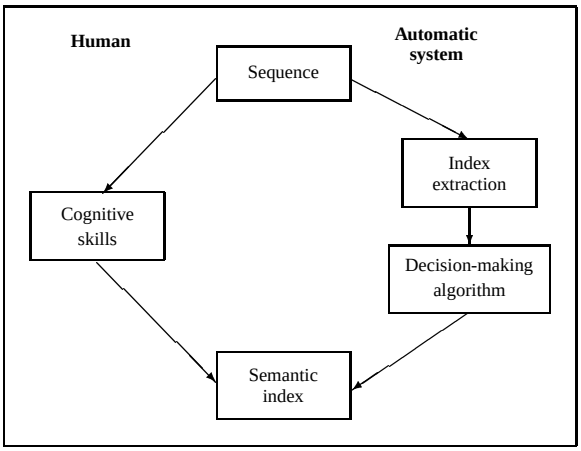

Fig. 1. Semantic video indexing problem solution.

$$\mu = \frac{1}{MN - I} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \sqrt{v_x^2(i,j) + v_y^2(i,j)} \qquad (1)$$

$$\mu < S_{\mathrm{no-motion}} \qquad (2)$$

where M and N are the frame dimensions (in MacroBlocks), I is the number of Intra-Coded MacroBlocks, v_x and v_y are the horizontal and vertical components of motion vectors and $S_{\mathrm{no-motion}}$ is the threshold value.

We have then evaluated the correlation between this parameter and semantic events for various threshold values, and we have found that with the value of $S_{\mathrm{no-motion}} = 4$ we are able to detect 65 over 92 semantic events with 60 false detections (requiring the presence of lack of motion for at least 3 P-frames) (see Fig. 2). This arises from the fact that qualitatively we have found lack of motion before events at the beginning of game actions or after events at the end of game actions.

Camera motion parameters, represented by horizontal "pan" and "zoom" factors, have been evaluated using a least-mean square method applied to P-frame motion fields [6]. We have detected fast horizontal pan (or fast zoom) by thresholding the pan value (or the zoom factor), using the threshold value S_{pan} (or S_{zoom}).

Exploiting the correlation between pan and zoom factors and semantic events, we have noticed that we can find fast pan in correspondence with shots toward the goalkeeper or fast ball exchanges, and fast zoom in correspondence with interesting situations according to the perception of camera operator. By requiring the presence of fast pan OR fast zoom for at least 3 P-frames, we are able to detect 53 over 103 semantic events with 67 false detections (see Figures 3, 4) [9] .

In our implementation, shot-cuts have been detected using only motion information too. In particular, we have used

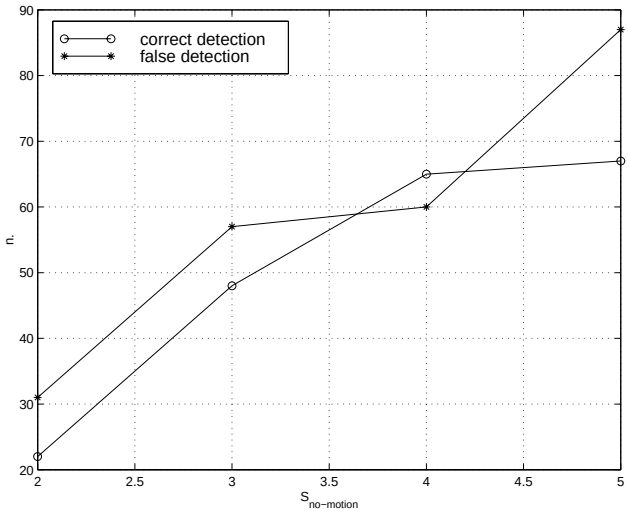

Fig. 2. Number of significant events detected (over a total of 92 significant events) and false detections, by setting various values to $S_{\mathrm{no-motion}}$.

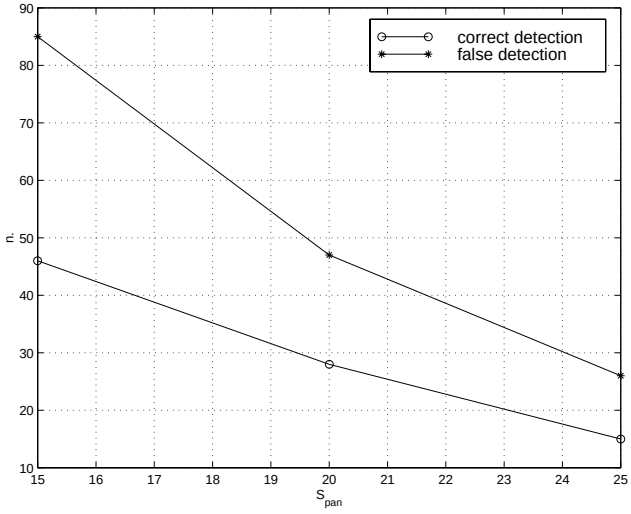

Fig. 3. Number of significant events detected (over a total of 103 significant events) and false detections, by setting various values to S_{pan}.

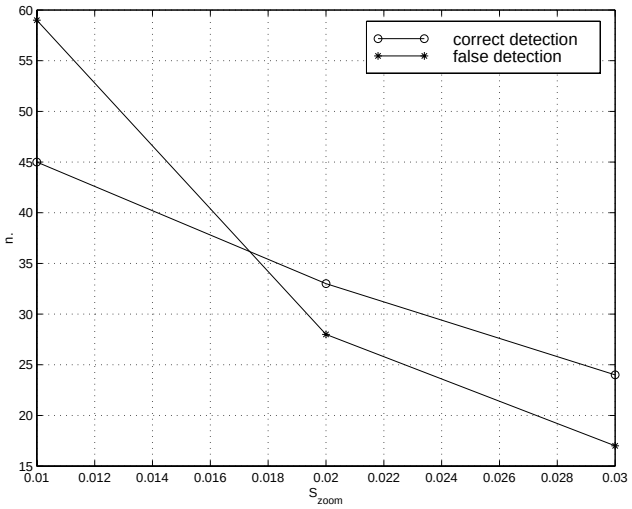

Fig. 4. Number of significant events detected (over a total of 103 significant events) and false detections, by setting various values to S_{zoom}.

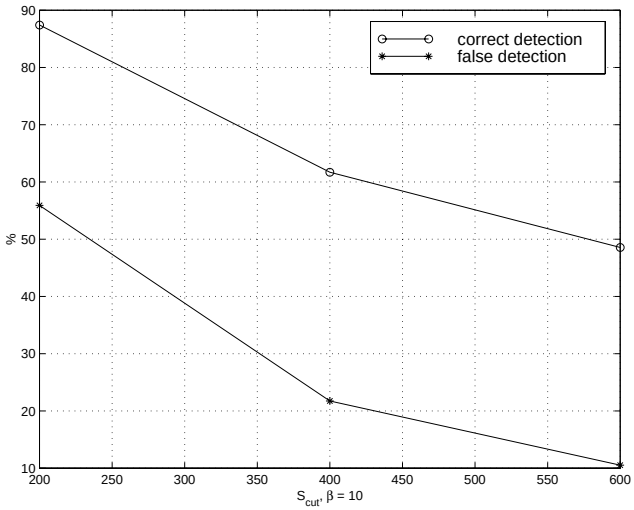

Fig. 5. Percentage of shot-cuts detected (over a total of 177 significant events) and false detections, by setting various values to S_{cut} and $\beta = 10$.

the sharp variation of the above mentioned motion descriptors and of the number of Intra-Coded Macroblocks of P-frames [7] [8].

To evaluate the sharp variation of the motion field we have used the difference between the average value of the motion vectors modules between two adjacent P-frames. This measure is given by

$$\Delta\mu(k) = \mu(k) - \mu(k-1) \tag{3}$$

where $\mu(k)$ è the average value of the motion vectors modules of the P-frame k, given by Eq. 1.

This parameter will assume significantly high values in presence of a shot-cut characterized by an abrupt change in the motion field between the two considered shots.

This information regarding the sharp change in the motion field have been suitably combined with the number of Intra-Coded MacroBlocks of the current P-frames, as follow

$$\mathrm{Cut}(k) = \mathrm{Intra}(k) + \beta\Delta\mu(k), \tag{4}$$

where $\mathrm{Intra}(k)$ shows the number of the Intra-Coded MacroBlocks of the current P-frame, and β è is a proper weighting factor.

When this parameter is greater than a prefixed threshold value S_{cut}, we assume there is a shot-cut [9]. The performance of this methods with various values of β and S_{cut} are presented in Figures 5, 6.

3. THE PROPOSED ALGORITHM

As it can be seen, the above mentioned low-level descriptors are not sufficient, individually, to reach satisfying results (all the semantic events detected with few false detection)[9].

To find particular events, such as, for example, goals or shot toward goal, we have exploited the temporal evolution of the low-level descriptors in correspondence with such events. We have noticed that in correspondence with goals we can find fast pan or zoom followed by lack of motion followed by a shot cut. The concatenation of these low level events have therefore been detected and exploited using the finite-state machine shown in Fig. 7.

From the initial state SI the machine goes into state S1 if fast pan or fast zoom is detected for at least 3 consecutive P-frames. Then, from state S1 the machine goes into final state SF, where a goal is detected, if a shot-cut is detected; from state S1 it goes into state S2 if lack of motion is detected for at least 3 consecutive P-frames. From state S2 the machine goes into final state SF if a shot-cut is detected, while it returns into state S1 if fast pan or zoom is detected for at least 3 consecutive P-frames (in this case the game action is probably still going on). Two ”timeouts” are used to return into initial state SI from states S1 and S2 in case nothing is happening for a certain number of P-frames (corresponding to about 1 minutes of sequence).

4. SIMULATION RESULTS

The performance of the proposed algorithm have been tested on 2 hours of MPEG2 sequences containing the semantic events reported in Table 1. As we can see from Table 1, almost all live goals are detected, and the algorithm is able to detect some shots to goal too, while it gives poor results on free kichs.

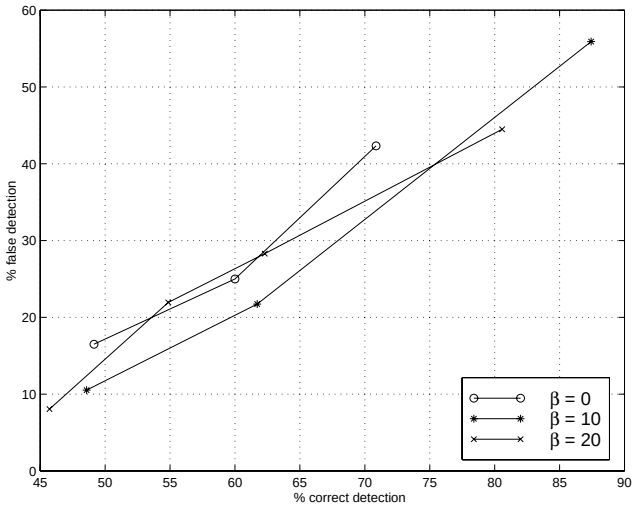

Fig. 6. Percentage of shot-cuts detected (over a total of 177 significant events) and false detections, by setting various values to β and to S_{cut}.

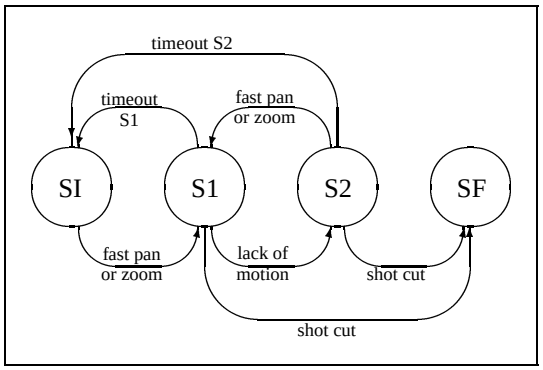

Fig. 7. The proposed algorithm.

The number of false detection is quite relevant, but we have to take into account that these results are obtained using motion information only, so these false detection will probably be eliminated using other type of media information.

5. CONCLUSIONS

In this paper we have presented a semantic video indexing algorithm based on finite-state machines that exploits the temporal dependencies of low-level descriptors extracted from the MPEG-2 compressed bit-stream. In particular we have applied the proposed algorithm to the semantic indexing of soccer games video sequences, obtaining very interesting results. Current research is devoted to the extension of the proposed algorithm to detect salient semantic events in other categories of video programmes.

Events	**Present**			**Detected**		
	Live	Replay	Total	Live	Replay	Total
Goals	20	14	34	18	7	25
Shots to goal	21	12	33	8	3	11
Penalties	6	1	7	1	0	1
Total	**47**	**27**	**74**	**27**	**10**	**37**
False						**116**

Table 1. Performance of the proposed algorithm.

6. REFERENCES

[1] Yao Wang, Zhu Liu, Jin-cheng Huang, "Multimedia Content Analysis Using Audio and Visual Information", To appear in Signal Processing Magazine.

[2] C. Saraceno, R. Leonardi, "Indexing Audio-Visual Databases Through a Joint Audio and Video Processing", International Journal of Imaging Systems and Technology, Vol. 9, No. 5, pp. 320-331, Oct. 1998.

[3] R. Lagendijk, "A Position Statement for Panel 1: Image Retrieval", Proc. of the VLBV99, Kyoto, Japan, pp. 14-15, October 29-30, 1999.

[4] A. Ferman, S. Krishnamachari, A. Tekalp, M. Abdel-Mottaleb, R. Mehrotra, "Group-Of-Frames/Pictures Color Histogram Descriptors for Multimedia Applications", Proc. of IEEE International Conference ICIP-2000, Vancouver, Canada, pp. 65-68, September 10-13, 2000.

[5] R. Zhao, W.I. Grosky, "From Features to Semantics: Some preliminary Results", Proc. of IEEE International Conference ICME2000, New York, NY, USA, 30 July - 2 August 2000.

[6] P. Migliorati, S. Tubaro, "Multistage Motion Estimation for Image Interpolation", Signal Processing: Image Communication, Vol. 7, pp. 187-199, July 1995.

[7] Yining Deng, B. S. Manjunath, "Content-Based Search of Video Using Color, Texture, and Motion", Proc. of IEEE International Conference ICIP-97, Santa Barbara, California, USA, pp. 534-536, October 26-29, 1997.

[8] Thomas Sikora, "MPEG Digital Video-Coding Standards", IEEE Signal Processing Magazine, Vol. 14, No. 5, September 1997.

[9] A. Bonzanini, R. Leonardi, P. Migliorati, "Semantic Video Indexing Using MPEG Motion Vectors", Proc. EUSIPCO'2000, pp. 147-150, 4-8 Sept. 2000, Tampere, Finland.