MODELLING OF VISUAL FEATURES BY MARKOV CHAINS FOR SPORT CONTENT CHARACTERIZATION

R. Leonardi, P. Migliorati, M. Prandini

Dept. of Electrical Engineering for Automation - University of Brescia
Via Branze, 38 - 25123, Brescia, Italy
e-mail: {leon, pier, prandini}@ing.unibs.it

ABSTRACT

The problem of semantic indexing of audio-visual documents is of great interest due to the wide diffusion of large audio-video databases. In this paper, we propose a semantic indexing algorithm based on the controlled Markov chain modelling framework. Controlled Markov chain models are used to describe the temporal evolution of low-level descriptors extracted from the MPEG compressed bit-stream. The proposed algorithm has been conceived for soccer game video sequences, and seems promising based on the simulation results obtained in this preliminary study.

1. INTRODUCTION

Effective navigation through audio-visual documents is necessary to enable widespread use and access to richer and novel information sources. Design of efficient indexing techniques to retrieve relevant information is another important requirement. Allowing for possible automatic procedures to semantically index audio-video material represents therefore a very important challenge. Such methods should be designed to create indices of the audio-visual material, which characterize the temporal structure of a multimedia document from a semantic point of view [1].

To face the semantic indexing problem, a man uses its cognitive skills, while an automatic system can face it by adopting a two-step procedure: 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, a decision-making algorithm is used to extract a semantic index from the low-level indices [2].
The problem of low-level descriptors extraction is widely discussed in the literature [3], whereas only a few contributions address the decision-making issue [4]. Moreover, the solution to this issue seems to depend on the considered specific program category.

In this work we have considered soccer video sequences. For this program category, the semantic content can be related to the occurrence 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. Therefore a good semantic index of a soccer video sequence could be the list of all game actions, each one characterized by its beginning and ending event. Such a summary could be very useful to satisfy various types of semantic queries.

The problem of automatic detection of semantic events in sport games has been studied by many researchers. In general the objective is to identify certain spatio-temporal segments that correspond to semantically significant events. In [5], for example, a method that tries to detect the complete set of semantic events which may happen in a soccer game is presented. This method uses the position information of the player and of the ball during the game as input, and therefore needs a quite complex and accurate tracking system to obtain this information.

In [6] and [7] we have studied the correlation between low-level descriptors and the semantic events in a soccer game. In particular, in [6], it is shown that the low-level descriptors are not sufficient, individually, to obtain satisfactory results (i.e., all the semantic events detected with only a few false detections). In [7] we have therefore tried to exploit the temporal evolution of the low-level descriptors in correspondence with semantic events, by proposing an algorithm based on a finite-state machine. This algorithm gives good results in terms of accuracy in the detection of the relevant events, whereas the number of false detections remains still quite large.

In this work we present a semantic video indexing algorithm using controlled Markov chains to model the temporal evolution of low-level descriptors.
We have chosen certain low-level descriptors, which represent the following characteristics: (i) lack of motion, (ii) camera operations (pan and zoom parameters), and (iii) presence of shot-cuts.
The proposed algorithm seems promising based on the simulation results obtained in this preliminary study.

The paper is organized as follows. In Section 2 we describe the selected low-level descriptors, whereas in Section 3 we present the event detection algorithm. In Section 4 we report some experimental results, and make some final remarks in Section 5.

2. THE LOW-LEVEL DESCRIPTORS

In this section we describe the three low-level binary descriptors adopted in the proposed algorithm. These descriptors, associated to each P-frame, represent the following characteristics: (i) lack of motion, (ii) camera operations (pan and zoom parameters), and (iii) the presence of shot-cuts, and are the same descriptors used in [7]. Each descriptor takes value in the set $\{0, 1\}$.

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

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

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. The threshold value has been set equal to 4. The descriptor assumes value 0 when the threshold is exceeded.

Camera motion parameters, represented by horizontal “pan” and “zoom” factors, have been evaluated using a least-mean squares method applied to P-frame motion fields [8]. We have then evaluated the value of the descriptor “Fast pan” (“Fast zoom”) by thresholding the pan value (zoom factor), using the threshold value 20 (0.002). In this case, the descriptors assume value 1 when the threshold is exceeded.

In this work, shot-cuts have been detected using only motion information as well. In particular, we have used the sharp variation of the above mentioned motion parameters, and of the number of Intra-Coded Macroblocks of P-frames [9] [10].

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

$$\Delta\mu(k) = \mu(k) - \mu(k-1)$$

where $\mu(k)$ is the average value of the motion vectors modules of 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 has been suitably combined with the number of Intra-Coded MacroBlocks of the current P-frames, as follows

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

where $\mathrm{Intra}(k)$ is the number of the Intra-Coded MacroBlocks of the current P-frame, and β is a weighting factor set to 20. When this parameter is greater than a prefixed threshold set to 700, we say that a shot-cut has occurred [7].

In the next section, we describe the proposed algorithm where the temporal evolution of these low-level descriptors is modeled by a controlled Markov chain.

3. THE PROPOSED ALGORITHM BASED ON CONTROLLED MARKOV CHAIN MODEL

In this section, we briefly describe the controlled Markov chain modelling framework [11], and then detail the controlled Markov chain model adopted in our context.

The components of a controlled Markov chain model are the state and input variables, the initial state probability distribution, and the controlled transition probability function. Here, we consider homogeneous models with state and input variables taking values in finite sets.

Denote by $\mathbf{s}(t)$ the random variable representing the state of the controlled Markov chain at time $t \in \mathcal{T} := \{0, 1, 2, \ldots\}$. At each $t \in \mathcal{T}$, the state $\mathbf{s}(t)$ takes value in a discrete set $\mathcal{S}$. At time $t = 0$, the initial state $\mathbf{s}(0)$ is described in terms of its probability distribution, say P_0, over the space set $\mathcal{S}$. The evolution of $s(t)$ from time $t \in \mathcal{T}$ to time $t+1$ is governed by a probability of transition. This probability is affected by an input signal, that we denote by $\mathbf{u}(t)$, taking value in a discrete input set $\mathcal{U}$. The probability of transition is only a function of the input $u \in \mathcal{U}$ applied at time t. By this we mean that $\mathbf{s}(t+1)$ is a random variable conditionally independent of all other random variables at times smaller or equal to t, given $\mathbf{s}(t)$, $\mathbf{u}(t)$. Here we assume a stationary transition probability, i.e.,

$$\mathrm{P}\left(\mathbf{s}(t+1) = s' \mid \mathbf{s}(t) = s, \mathbf{u}(t) = u\right) = p(s, s', u),$$

$\forall s, s' \in \mathcal{S}, u \in \mathcal{U}, t \in \mathcal{T}$, where $p : \mathcal{S} \times \mathcal{S} \times \mathcal{U} \to [0, 1]$ is the *controlled transition probability function*.

If the input applied to the system keeps constant, say equal to $\bar{u} \in \mathcal{U}$, irrespectively of the system evolution, then the controlled Markov chain reduces to a standard Markov chain with transition probabilities $\{p(s, s', \bar{u}),\ s, s' \in \mathcal{S}\}$. This can be easily shown as follows

$$\begin{aligned}
&\mathrm{P}\left(\mathbf{s}(t+1) = s' \mid \mathbf{s}(t) = s\right) \\
&\quad = \sum_{u \in \mathcal{U}} p(s, s', u)\, \mathrm{P}\left(\mathbf{u}(t) = u \mid \mathbf{s}(t) = s\right) = p(s, s', \bar{u}),
\end{aligned}$$

$\forall t \in \mathcal{T}, s, s' \in \mathcal{S}$, where we used the property that for each $t \in \mathcal{T}$, $\mathbf{u}(t)$ is independent of all the random variables up to and including time t, and $\mathrm{P}\left(\mathbf{u}(t) = \bar{u}\right) = 1$.

In our context, $\mathbf{u}(t)$ is introduced to model the occurrence of a shot-cut event. The control set is in fact defined as $\mathcal{U} = \{0, 1\}$, and if a shot-cut event happens at time t,

then $\mathbf{u}(t) = 1$, otherwise $\mathbf{u}(t) = 0$.
We suppose that the occurrence of a shot-cut event causes the system to change dynamics. In order to model this fact, we describe the state of the system as a two-component state, i.e., $\mathbf{s}(t) = (\mathbf{x}(t), \mathbf{q}(t)) \in \mathcal{S} = \mathcal{X} \times \mathcal{Q}$, where $\mathbf{q}(t) \in \mathcal{Q} := \{0, 1\}$ is called the mode of the system. Also, we impose a certain structure on the controlled transition probability function. Specifically, the controlled transition probability function is supposed to satisfy the following condition

$$p((x,q),(x',q'),u) = 0, \text{ if } (u = 1 \text{ and } q = q') \\ \text{or } (u = 0 \text{ and } q \neq q'), \quad \forall x, x' \in \mathcal{X}, q, q' \in \mathcal{Q},$$

which says that a shot-cut event forces the controlled Markov chain to change operating mode, whereas if no shot-cut event occurs, then the controlled Markov chain remains in the same mode.
Note that within a single mode, say $q \in \mathcal{Q}$, the controlled Markov chain reduces to a standard homogeneous Markov chain with state space $\mathcal{X}$ governed by the transition probability function $\{p_q(x,x'), x, x' \in \mathcal{X}\}$, where $p_q(x,x') := p((x,q),(x',q),0)$, $\forall x, x' \in \mathcal{X}$. We denote by $\varphi_{x,q}$ the probability distribution of $\mathbf{x}(t+1)$ when $\mathbf{x}(t)$ and $\mathbf{q}(t)$ take values $x \in \mathcal{X}$ and $q \in \mathcal{Q}$, respectively. Here, we suppose that each one of the two homogeneous Markov chains admits a stationary probability distribution and we denote by π_q the one associated with mode $q \in \mathcal{Q}$. Then, $\pi_q(x)$ is the probability of $\mathbf{x}(t)$ being equal to $x \in \mathcal{X}$ in the long run, when the system remains in mode q.
We assume that at time $t = 0$, when we start observing the system evolution, the system is in mode $q = 0$ and in stationary conditions, i.e., $\mathrm{P}_0(s) = \pi_0(x)$, if $s = (x, 0)$, $x \in \mathcal{X}$, and 0, otherwise.
When a shot-cut event occurs, then the operating mode of the system changes. As for the state component $\mathbf{x}$, we suppose that it is reinitialized as a random variable with a certain fixed distribution. Specifically, we assume that

$$p((x,q),(x',q'),1) = \pi_{q'}(x'), \quad \forall x' \in \mathcal{X}, q, q' \in \mathcal{Q}, q \neq q'.$$

A schematic representation of the introduced model is given in Figure 1. In this figure, the symbol "$\sim$" is used for "distributed according to".

In our context, $\mathcal{T}$ represents the set of time instants associated with the P-frames sequence. As for $\mathbf{x}(t)$, it is state of the P-frame observed at time t. In particular, $\mathbf{x}(t)$ can take the following values: "LM", "FP", "FZ", "FPZ", and "Other", hence the set $\mathcal{X}$ has cardinality 5. The value taken by $\mathbf{x}(t)$ is evaluated by means of the low-level descriptors introduced in Section 2.

Fix a time instant $t \in \mathcal{T}$ and consider the corresponding P-frame. The state variable $\mathbf{x}(t)$ is said to take the value $x =$"LM" if the descriptor "Lack of motion" is equal to 1.

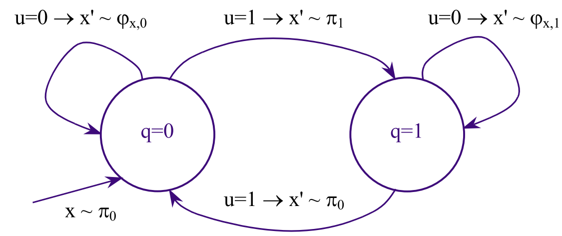

Fig. 1. Controlled Markov chain model.

If that is not the case, then, $\mathbf{x}(t)$ can take one of the other 4 values. Specifically, $\mathbf{x}(t)$ is equal to x="FP" if the value of the descriptor "Fast pan" is 1 and that of the descriptor "Fast zoom" is 0. In the opposite case, i.e., when "Fast pan" is equal to 0 and "Fast zoom" is equal to 1, then, $\mathbf{x}(t)$ takes the value $x =$"FZ". In the case when both the "Fast pan" and "Fast zoom" descriptors are equal to 1, $\mathbf{x}(t)$ assumes the value x="FPZ". In all the other cases, $\mathbf{x}(t)$ is said to take the value x="Other".

We suppose that each semantic event takes place over a two-shot block and that it can be modeled by a controlled Markov chain with the structure described above. Each semantic event is then characterized by the two sets of probability distributions over the state space $\mathcal{X}$ $\mathcal{P}_0 := \{\varphi_{x,0}, x \in \mathcal{X}\}$ and $\mathcal{P}_1 := \{\varphi_{x,1}, x \in \mathcal{X}\}$, which govern the evolution of $\mathbf{x}(t)$ within mode $q = 0$ and $q = 1$, respectively.
Specifically, we have considered 6 models denoted by A, B, C, D, E, and F, where model A is associated to goals, model B to corner kicks, and models C, D, E, F describe other situations of interest that occur in soccer games, such as free kicks, plain actions, and so on. For each event, we have determined the $\mathcal{P}_0$ and $\mathcal{P}_1$ sets of the corresponding model by selecting manually all the pairs of shots related to that event in a set of training sequences, then determining the values taken by $\mathbf{x}(t)$, $t \in \mathcal{T}$, in the obtained P-frame sequences, and finally estimating the probabilities $\varphi_{x,0}$ and $\varphi_{x,1}$, $x \in \mathcal{X}$.

On the basis of the derived six Markov models, one can classify each pair of shots in a soccer game video sequence by using the maximum likelihood criterion. For each pair of consecutive shots (i.e., two consecutive sets of P-frames separated by shot-cuts), one needs to i) extract the sequence of low-level descriptors, ii) determine the sequence of values assumed by the state variable $\mathbf{x}$, and iii) determine the likelihood of the sequence of values assumed by $\mathbf{s} = (\mathbf{x}, \mathbf{q})$ (with $\mathbf{q}$ set equal to 0 before the shot-cut and to 1 after the shot-cut) according to each one of the six admissible models. The model that maximizes the likelihood function is then associated to the considered pair of shots.

4. SIMULATION RESULTS

The performance of the proposed algorithm have been tested considering about 2 hours of MPEG2 sequences containing more than 800 shot-cuts, determined using the algorithm described in Section 2. The sequences contain 9 goals and 16 corner kicks. The results obtained are summarized in Table 1. As shown in this table, 8 goals out of 9, and 10 corner kicks out of 16 are detected.

The number of false detections could seem quite relevant. However, these results are obtained using motion information only, so these false detections could probably be reduced by using other type of media information (such as audio loudness). Also, we made the simplifying assumption that the Markov chain with state $\mathbf{x}$ associated to each mode $q \in \mathcal{Q}$ is in stationary conditions when entering that mode. One could instead introduce a "reset probability distribution" of $\mathbf{x}$ associated to mode q, $q \in \mathcal{Q}$, which should then be estimated from data.

	Occured events	Detected events	False detections
Goal	9	8	58
Corner	16	10	90

Table 1. Performance of the proposed algorithm.

5. CONCLUSIONS

In this paper we have presented a semantic video indexing algorithm based on controlled Markov chain models that exploits the temporal evolution 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 interesting results. Still, further work needs to be done in order to improve its performance by considering other descriptors, related for example to audio information.

6. ACKNOWLEDGMENT

The authors would like to thank Francesca Oliva and Cristiana Molinari for their help on the algorithm simulation.

7. REFERENCES

[1] N. Adami, A. Bugatti, R. Leonardi, P. Migliorati, L. A. Rossi, "The ToCAI description scheme for indexing and retrieval of multimedia documents", Multimedia Tools and Applications Journal, Kluwer Academic Publishers, vol. 14, no. 2, pp. 151-171, 2001. N.

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

[3] Yao Wang, Zhu Liu, Jin-cheng Huang, "Multimedia Content Analysis Using Audio and Visual Information", IEEE Signal Processing Magazine, vol. 17, no. 6, pp. 12-36, Nov. 2000.

[4] 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.

[5] V. Tovinkere, R. J. Qian, "Detecting Semantic Events in Soccer Games: Toward a Complete Solution", Proc. ICME'2001, pp. 1040-1043, August 2001, Tokyo, Japan.

[6] 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.

[7] A. Bonzanini, R. Leonardi, P. Migliorati, "Event Recognition in Sport Programs Using Low-Level Motion Indices", Proc. ICME'2001, pp. 920-923, August 2001, Tokyo, Japan.

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

[9] 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.

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

[11] Martin L. Puterman, "Markov Decision Processes", Wiley, 1994.