A MARKOV CHAIN MODEL FOR SEMANTIC INDEXING OF SPORT PROGRAM SEQUENCES

R. LEONARDI, P. MIGLIORATI, M. PRANDINI

DEA, University of Brescia
Via Branze, 38,
25123, Brescia, Italy
E-mail: {leon, pier, prandini}@ing.unibs.it

In this paper, we propose a semantic indexing algorithm based on the controlled Markov chain modeling framework. Controlled Markov chain models are used to describe the temporal evolution of low-level visual descriptors extracted from the MPEG compressed bit-stream. To reduce the number of false detections given by the proposed video-processing algorithm, we have considered also the audio signal. In particular we have evaluated the "loudeness" associated to each video segments identified by the analysis carried out on the video signal. The intensity of the "loudness" has then been used to order the selected video segments. In this way, the segments associated to interesting events appear in the very first positions of the ordered list, and the number of false detections can be greatly reduced. The proposed algorithm has been conceived for soccer game video sequences, and the simulation results have shown the effectiveness of the proposed algorithm.

1. Introduction

The design of efficient indexing techniques suitable to retrieve relevant information through audio-visual documents is necessary to enable widespread use and access to richer and novel information sources. Allowing for possible automatic procedures to semantically index audio-video material represents therefore a very important challenge.

To face the problem of semantic indexing, 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.

In this work we have considered soccer games 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 [2], 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 [3] and [4] we have studied the correlation between low-level descriptors and the semantic events in a soccer game. In particular, in [3], 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 [4] 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 video descriptors.

To reduce the number of false detections given by the proposed video-processing algorithm, we have considered also the audio signal. In particular we have evaluated the "loudeness" associated to each video segments identified by the analysis carried out on the video signal. The intensity of the "loudness" has then been used to order the selected video segments. In this way, the segments associated to the interesting events appear in the very first positions of the ordered list, and the number of false detections can be greatly reduced. The proposed algorithm seems promising based on the simulation results obtained in this preliminary study.

This paper is organized as follows. In Section 2 we describe the selected low-level video descriptors, whereas in Section 3 we present the event detection algorithm based on video descriptors. In Section 4 we report some experimental results based on this algorithm. In Section 5 we present the proposed analysis carried out on the audio signal in order to reduce the number of false detection given by the proposed algorithm. Concluding remarks are given in Section 6.

2. The Low-Level visual descriptors

In this section we describe the 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 [4]. 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 for each P-frame. 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 [3]. 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 [5] . 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.

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:

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

where 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 [4].

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

3. The proposed video-processing algorithm

In this section, we briefly describe the controlled Markov chain modelling framework [6], 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 s(t) the random variable representing the state of the controlled Markov chain at time $t \in T := \{0,1,2, ...\}$. At each $t \in T$, the state s(t) takes value in a discrete set S. At time t=0, the initial state s(0) is described in terms of its probability distribution, say P_0, over the space set S.

The evolution of s(t) from time $t \in T$ to time t+1 is governed by a probability of transition. This probability is affected by an input signal, that we denote by u(t), taking value in a discrete input set U. The probability of transition is only a function of the input $u \in U$ applied at time t. By this we mean that s(t+1) is a random variable conditionally independent of all other random variables at times smaller or equal to t, given s(t), u(t). Here we assume a stationary transition probability, i.e.,

$$P(s(t+1) = s' \mid s(t) = s, u(t) = u) = p(s, s', u),$$

$\forall$ s, s' $\in$ S, u $\in$ U, t $\in$ T, where $S \times S \times U \rightarrow [0, 1]$ is the controlled transition probability function. If the input applied to the system keeps constant, say equal to u' $\in$ U, irrespectively of the system evolution, then the controlled Markov chain reduces to a standard Markov chain.

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

$$s(t) = (x(t), q(t)) \in S = X \times Q,$$

where $q(t) \in Q := \{0,1\}$ 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 condition 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 $q \in Q$, the controlled Markov chain reduces to a standard homogeneous Markov chain with state space X.

We denote by $\varphi_{x,q}$ the probability distribution of x(t+1) when x(t) and q(t) take values $x \in X$ and $q \in Q$, 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 $q \in Q$. Then, $\pi_q(x)$ is the probability of x(t) being equal to $x \in X$ 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., $P_0(s) = \pi_0(x)$, if s=(x,0), and 0, otherwise. When a shot-cut event occurs, then the operating mode of the system changes. As for the state component x we suppose that it is reinitialized as a random variable with a certain fixed distribution.
Specifically, we assume that:

$$P ((x,q), (x',q'), 1) = \pi_{q'}(x') , q \neq q'.$$

In our context, T represents the set of time instants associated with the P-frames sequence. As for x(t), it is state of the P-frame observed at time t. In particular, x(t) can take the following values: "LM", "FP", "FZ", "FPZ", and "Other", hence the set X has cardinality 5. The value taken by x(t) is evaluated by means of the low-level descriptors introduced in the previous sections. Fix a time instant t and consider the corresponding P-frame. The state variable x(t) is said to take the value x = "LM" if the descriptor "Lack of motion" is equal to 1. If that is not the case, then, x(t) can take one of the other 4 values. Specifically, 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, x(t) takes the value x= "FZ". In the case when both the "Fast pan" and "Fast zoom" descriptors are equal to 1, x(t) assumes the value x= "FPZ". In all the other cases, 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 modelled 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 X, P_0 and P_1, which govern the evolution of 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 P_0 and 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 x(t), in the obtained P-frame sequences, and finally estimating the probabilities $\varphi_{x,0}$, $\varphi_{x,1}$. 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 x, and iii) determine the likelihood of the sequence of values assumed by s=(x, q) (with 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 obtained results are the following: 8 goals out of 9 (58 false detections), and 10 corner kicks out of 16 are detected (90 false detections).

The number of false detections could seem quite relevant. However, these results are obtained using motion information only, so these false detections can be reduced by using other type of media information (such as audio loudness), as described in the next Section.

5. Processing of the audio signal

To reduce the number of false detection given by the proposed video processing algorithm, we have considered also the audio signal.

To extract some relevant features, we have divided the audio stream of a soccer game sequence in consecutive clips of 1.5 seconds, in order to consider the audio signal quasi-stationary in this window [1][7].

Each clip is then divided in audio-frames of 1024 samples and every audio-frame is overlapped by 512 samples with respect to the previous one. The "loudness" is estimated as the energy of the sequence of audio samples associated to the considered audio-frame. The evolution of the "loudness" in an audio clip follow the variation in time of the amplitude of the signal, and it constitutes therefore a fundamental aspect for the audio signal classification. We then estimate the mean value of the loudness for every clip. In this way we obtain, for each clip, a low-level audio descriptor represented by the "clip loudness".

As previously mentioned, to reduce the false detections given by the proposed algorithm that analyse the video signal, we have used also the audio information. In particular we have evaluated the average value of the "clip loudness" associated to each video segment identified by the video processing algorithm, and then we have ordered the selected video segments accordingly to the average value of the "clip loudness" along the time span of the considered

segment. In this way, the video segments containing the goals appear in the very fist positions of this ordered list. The first simulation results seems to be very encouraging, and the number of false detections appears significantly reduced.

6. 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 visual descriptors extracted from the MPEG-2 compressed bit-stream. Moreover, to reduce the number of false detections given by the proposed video processing algorithm, we have considered also the audio signal. In particular we have evaluated the "loudeness" associated to each video segments identified by the analysis carried out on the video signal. In particular we have applied the proposed algorithm to the semantic indexing of soccer games video sequences, obtaining interesting results.

Acknowledgments

This material is based upon work partially supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. Thomas Sikora, “MPEG Digital Video-Coding Standards”, IEEE Signal Processing Magazine, Vol. 14, No. 5, September 1997.
6. Martin L. Puterman, “Markov Decision Processes”, Wiley, 1994.
7. Yao Y. Wang, J. Huang, Z. Liu, T. Chen, “Multimedia Content Classification Using Motion and Audio Information”, Proc. of IEEE ISCAS'97, Vol. 2, pp. 1488-1491.