SOCCER HIGHLIGHTS DETECTION AND RECOGNITION USING HMMs

J.Assfalg, M.Bertini, A.Del Bimbo, W.Nunziati, P.Pala

University of Florence
Dipartimento di Sistemi e Informatica
Via S. Marta 3, 50139 Firenze, Italy
{assfalg,bertini,delbimbo,nunziati,pala}@dsi.unifi.it

ABSTRACT

In this paper we report on our experience in the detection and recognition of soccer highlights in videos using Hidden Markov Models. A first approach relies on camera motion only, whereas a second one also includes information regarding the location of players on the playing field. While the former approach requires less information, the latter has proven to be more precise. Our experimental evaluation yielded interesting results.

1. INTRODUCTION

Research addressing the problem 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 annotating the huge amount of live and archived video materials. Logging supports the task of editing TV programs, as it enables identification of relevant shots in the broadcasters' archives. Exploitation of such valuable assets is considered a key method for the improvement of production quality in a competitive market. Systems that enable efficient indexing and retrieval by content of video segments are the tools to pursue the aforementioned production quality. Among the many sports, soccer is for sure one of the most relevant—at least in Europe.

Within the scope of the ASSAVID project[1] a number of tools supporting automatic annotation of sports videos are being developed. By gaining insight into the current practice of broadcasters, we were able to identify relevant issues to be addressed, as well as to gain a solid domain knowledge. Broadcasters typically rely on two logging approaches: *production logging*, which is typically carried out live (or shortly after the event) by an assistant producer to select relevant shots to be edited into a magazine or news program that reports on sports highlights of the day (e.g. "Match of the day" or "GrandStand"); *posterity logging*, which is performed by librarians to make a detailed annotation of the video tapes, so as to enable their reuse in the long term, to provide added depth and historical context to recent events (e.g. select best achievements of a sports personality during his/her career).

In this paper we report on our experience in the classification of soccer highlights using Hidden Markov Models. A first approach relies on camera motion only, whereas a second one also includes information on the location of players on the playing field. The methods have been tested using several soccer videos containing a wide range of different styles in the length of the shots, in camera motion, and in editing effects, as produced by several different international broadcasters. We believe that considering such a variety of styles is of paramount importance in this field, as otherwise the system might lack robustness.

2. PREVIOUS WORK

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, 3, 4], tennis [5], basketball [6, 7], and baseball [8]. We review hereafter previous work related to soccer videos.

The work presented by Seo et al. is limited to detection and tracking of both the ball and the players; they do not attempt to identify highlights [1]. In [2], Gong et al. exploit the fact that the playing field is always green for the purpose of extracting it. Successive detection of ball and players is limited to the field, described by a binary mask. To determine position of moving objects (ball and players) within the field, the central circle is first located, and a four-point homographic planar transformation is then performed, to map image points to the model of the playing field. Whenever the central circle is not present in the current frame, a mosaic image is used to extend the search context. In this latter case, the mosaicing transformation is combined with the homographic transformation. This appears to be a fairly expensive approach. Tovinkere and Qian propose a hierarchical E-R model capturing domain knowledge of soccer [3]. This scheme organizes basic actions as well as complex events (both observed and interpreted), and uses a set of (nested) rules to tell whether a certain event takes place or not. The system exploits 3D data on the position of players and ball, which are obtained from either microwave sensors or multiple video cameras. Despite the authors' claim that, unlike other systems, their own works on an exhaustive set of events, only little evidence of this is provided, as only a basic action (*deflection*) and a complex event (*save*) are discussed. In [4], Yow et al. report on the usage of panoramic (i.e. mosaic) images to present soccer highlights: moving objects and the ball are super-imposed on a background image featuring the playing field. Ball, players, and goal posts are detected. However, only presentation of highlights is addressed, and no semantic analysis of relevant events is carried out.

[1] ASSAVID (Automatic Segmentation and Semantic Annotation of Sports Videos) is funded by the European Commission under grant IST-1999-13082.
http://www.bpe-rnd.co.uk/assavid/

0-7803-7304-9/02/$17.00 C2002 IEEE

3. RELEVANT FEATURES

Inspection of tapes, totaling over 20 hours of video footage, showed that producers of videos use a main camera to follow the action of the game; since this is usually taking place around the ball, we can assume that the main camera follows the ball. This implies that a strong correlation exists between the movement of the ball and camera action. The main camera is positioned along one of the long sides of the playing field. Some typical scenes taken with the main camera are shown in Fig. 1.

Figure 1: Example of frames taken from the main camera

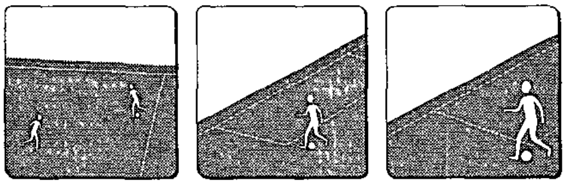


Figure 2: Framing terms: *Very Long Shot*, *Long Shot* and *Medium Shot*

Analysis of videos also showed that actions are almost always taken with a very long shot (VLS) framing, and the framing may change to medium long shot (MLS) through zoom in, and viceversa (see Fig. 2 for examples of different framing terms). Framing terms from medium to *close-up* can not show a team action, since they refer to a single player only, and are usually taken from cameras different than the main camera.

Identification of the part of the playing field currently framed and extraction of camera action are among the most significant tasks that can be performed on shots taken by the main camera. Typical actions featured by the main camera are: *i*) pan, *ii*) tilt, and *iii*) zoom. Pan and tilt are used to move from one part of the playing field to another one; hence, they can be related to the position of the ball. Whereas, zoom is used to change the framing of the subject, and can be used to infer the framing term.

Highlights that have been elected for investigation are: *i*) penalty kicks, *ii*) free kicks next to the goal box, and *iii*) corner kicks. These are typical highlights shown in TV news and magazine programs summarizing a match: they thrill the speakers and the audience, as they embody attack actions in proximity of the goal box area that might eventually lead to the scoring of a goal. Hence, penalty kicks and corners are also often used to calculate statistics supporting the evaluation of the degree of aggressiveness of the two teams.

For our second experiment, we also considered a feature providing a qualitative description of positions of players on the playing field. To this end, three zones have been defined for each of the two halves of the field: the small goal-keeper box (F_1), the goal box (F_2), and the the area between the goal box and the center of the field (F_3). Position of players is then described with 3 fuzzy qualifiers, one for each of the 3 zones (see Fig.3).

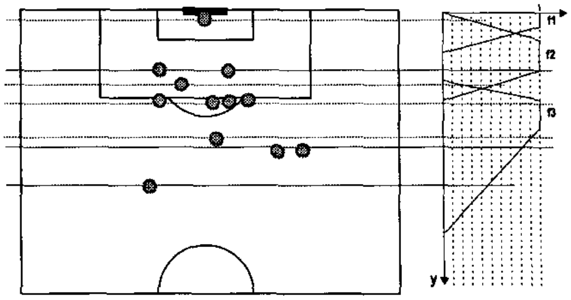


Figure 3: On the left, sample scheme representing the position of players on the field. On the right, the fuzzy qualifiers used to compute the values for the F_1, F_2, and F_3 descriptors: the value for each descriptor F_i is obtained by adding the terms $f_i(y_k)$ obtained for each player k.

Videos used in this study were acquired from Digital Video (DV) tapes, recorded at full PAL resolution, and 25 fps, using a DVLink card on an SGI O2 workstation. Some were retrieved from the BBC Sports archives, while other videos were digitally recorded off-air from other broadcasters.

4. EVENT DETECTION AND CLASSIFICATION

Detection and classification of soccer highlights is accomplished according to a model based approach. We assume that the event space is composed of M classes $E_1, \ldots, E_M$, each class corresponding to one relevant highlight. Each class E_i has one associated model λ_i. We have used Hidden Markov Models since they have proven to be effective in many applications where modeling of temporal sequences is required (such audio analysis and classification [9, 10]). Furthermore, since these models are probabilistic, uncertainty can be managed more effectively than with other models. A HMM is a Markov chain whose state sequence cannot be observed directly, but can be observed through a sequence of observation vectors. Each observation vector is generated by an underlying state with an associated probability distribution. More precisely, a hidden Markov model is defined by a set of states, a set of probabilities for transitions between the states, a set of output values, and a probability distribution of output values on each state. In general, a HMM with N states, and observation vector sequences of length T, is described as follows:

$Q = \{q_t\}^T_{t\pm}$ is the state sequence;

$A = \{a_{ij} = Pr(q_{t+1} = j | q_t = i)\}^N_{i,j=1}$ are the state transition probabilities;

$\Pi = \{\pi_i = Pr(q_1 = i)\}^N_{i=1}$ are the initial state probabilities;

$O = \{o_t\}^T_{t=1}$ is the observation sequence;

and $B = \{b_j(o_t) = Pr(o_t | q_t = j\}$ are the probability distributions for observation o_t given state $q_t = j$.

In the specific case we are analyzing, we can consider that the ball moving across the playing field embodies the hidden process, whilst the values characterizing camera action are the observations. Typically, different patterns of camera motion can be related to different states. For instance, in the model of a penalty, in the first state, when the player is preparing to kick, the camera is supposed to be steady; in the second state, the camera moves fast, towards the goal, trying to follow the ball kicked by the player; during follow-up, corresponding to the last state, the camera slows down. Thus, domain knowledge suggested the usage of three states for each model. Experiments with lower and higher number of states (i.e. 2 and 4) confirmed the validity of the assumption.

Further analysis of the soccer domain guided the choice of the transition model. As highlights that we are going to detect can be regarded as a progression of phases, a *left-to-right* model was selected. Initial weights of the A matrix and Π vector were chosen according to the above considerations, and are as follows:

$$\mathbf{A} = \begin{pmatrix} 0.5 & 0.4 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0 & 0 & 1 \end{pmatrix} \qquad \mathbf{\Pi} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

We adopted a model with discrete observation values: the framing term may assume three values (*VLS*, *LS*, *MLS*); pan and tilt values were quantized in 5 and 2 levels, respectively. A lower resolution is used for tilt, since its presence or absence is more significant than its actual value. The resulting number of observation symbols is 30. Models were trained according to the Baum-Welch algorithm, as described in [9]. Each model was trained separately, using only observation vectors corresponding to the specific event the model should represent. In so doing, if the system is required to manage a new class of highlights, a new HMM has to be introduced, without requiring to re-train existing HMMs.

Once all the HMMs ($\lambda_i, i = 1, \ldots, M$) are trained, detection and classification of an unknown video shot is straightforward. The sequence of observation vectors O is extracted from the video shot and the maximum likelihood function s_i for each model λ_i is computed through the forward-backward algorithm. Then, the following values are computed:

$$\hat{s} = \max_{i=1,\ldots,M}\{s_i\} \qquad \hat{\imath} = \arg\max_{i=1,\ldots,M}\{s_i\}$$

If $\hat{s}$ exceeds a predefined threshold τ_s, then a highlight of class $E_{\hat{\imath}}$ is detected in the video shot. Otherwise, the video shot is considered not to contain any relevant highlight. Usage of a *null* model, instead of applying a threshold, does not represent a viable approach in this context: the variety of non relevant events is enormous, and gathering examples to construct the training set appears a virtually impossible task. Also, adding new models to the system would require the re-training of the *null* model.

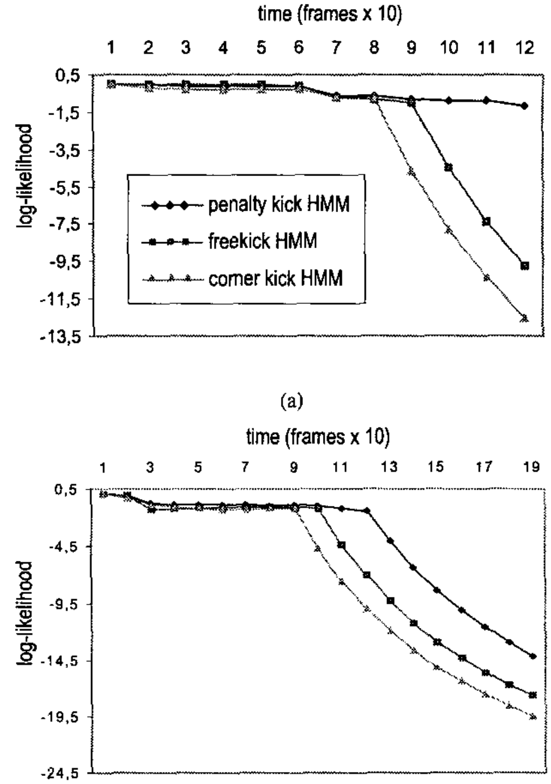
5. EXPERIMENTAL RESULTS

The training set comprised 10 shots for each highlight class, taken from different matches and broadcasters. To test the system 45 video shots were selected, 10 for each of the 3 highlights classes, and 15 that display generic actions taking place next to the goal box. Each shot was tested against the models using the aforementioned procedure. Results are reported in Table 1.

Figs.4-a–b show the behavior of the three models when one highlight is correctly recognized, and when all fail to recognize the highlight, respectively.

	Detected	Correct	Missed	False
Penalty	11	7	3	4
Free kick	12	8	2	4
Corner	11	8	2	3
Other	11	8	2	3

Table 1: Performance figures for the HMMs of the 3 highlight classes show a correct detection rate of approximately 80%. Shots where classified as *Other* if none of the 3 models yielded a satisfactory scoring.



(a)

(b)

Figure 4: The graphs represent the log-likelihood of the observation sequence $\log P(O|\lambda)$, weighted with the observation length, for a) correct recognition and b) missed detection.

Following the first experimental results, the models have then been improved by including information on the position of players. In our preliminary experiments, this information was manually annotated. The fuzzy description has proven to be robust, as outliers (i.e. players not directly involved in the action) did not seriously affected the performance—which, instead, may be the case when relying on the exact localization of players. Besides, description of players' positions have the advantage of yielding an observation variable of fixed length for information whose dimension may

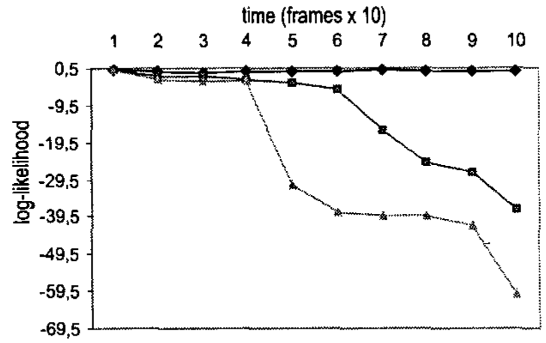


Figure 5: Log-likelihood plot for a sample shot, obtained using models including information on players' positions. Comparison of patterns of the log-likelihood functions for the two approaches—i.e. without and with position information—indicates that the latter allows for a better discrimination of highlights classes.

vary over time (i.e. the number of players framed varies during the shot). By using this data the observation model changes, in that the values of F_1, F_2 and F_3 are not quantized, and therefore represent a continuous variable. Experimental results are shown in Table 2. It can be noted that, using this model, results improve significantly. Moreover, discrimination among highlight classes with these models is much sharper (see Fig.5, where an example of correct recognition is shown).

	detected	correct	missed	false
Penalty	10	9	1	1
Free kick	10	8	2	2
Corner	10	10	0	0

Table 2: Performance improved w.r.t. to the first experiment (up to 90% for penalties and 100% for corners). The poorer performance of free kicks can be accounted to the fact that these are less structured, and may require a richer training set.

Apart from their actual values, what makes figures reported in the Tables 1 and 2 very appealing is the fact that they were obtained with videos collected from a number of different sources (i.e. different matches and broadcasters), indicating the robustness of our approach.

6. FUTURE WORK

In this paper we have proposed two approaches to classify soccer video shots highlights using Hidden Markov Models. Future work might include development of techniques supporting content-based temporal segmentation of so called *iso* feeds (i.e. isolated; not edited, coming from a single camera), which would allow to extend usage of the models to long video sequences. Further, we will investigate the feaseability of using our approach to recognize a wider set of soccer highlights, as well as applying these models to other types of sports. Finally, specific experiments might also be carried out to assess the extent to which quality of the automatically extracted features affects the performance of the models.

7. REFERENCES

[1] Y.Seo, S.Choi, H.Kim, and K.-S.Hong, "Where are the ball and players? : Soccer game analysis with color-based tracking and image mosaick," *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," *Proc. of Int'l Conf. on Multimedia Computing and Systems (ICMCS'95)*, pp. 167–174, 1995.

[3] V.Tovinkere and R.J.Qian, "Detecting semantic events in soccer games: Towards a complete solution," *Proc. of Int'l Conf. on Multimedia and Expo (ICME 2001)*, pp. 1040–1043, 2001.

[4] D.Yow, B.-L.Yeo, M.Yeung, and B.Liu, "Analysis and presentation of soccer highlights from digital video," *Proc. of 2nd Asian Conf. on Computer Vision (ACCV'95)*, 1995.

[5] G. Sudhir, J.C.M. Lee, and A.K. Jain, "Automatic classification of tennis video for high-level content-based retrieval," *Proc. of the Int'l Workshop on Content-Based Access of Image and Video Databases (CAIVD '98)*, 1998.

[6] S.Nepal, U.Srinivasan, and G.Reynolds, "Automatic detection of 'goal' segments in basketball videos," *Proc. of ACM Multimedia*, pp. 261–269, 2001.

[7] D.D.Saur, Y.-P.Tan, S.R.Kulkami, and P.J.Ramadge, "Automatic analysis and annotation of basketball video," *Storage and Retrieval for Image and Video Databases V*, pp. 176–187, Feb. 1997.

[8] Y.Rui, A.Gupta, and A.Acero, "Automatically extracting highlights for tv baseball programs," *Proc. of ACM Multimedia*, 2000.

[9] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.

[10] J. Hui, K. Hirose, and H. Qiang, "A minimax search algorithm for robust continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 688–694, Nov. 2000.