Soccer Videos Highlight Prediction and Annotation in Real Time

M. Bertini, A. Del Bimbo, and W. Nunziati

Università di Firenze, 50139 Firenze, Italia
bertini, delbimbo, nunziati@dsi.unifi.it

Abstract. In this paper, we present an automatic system that is able to forecast the appearance of a soccer highlight, and annotate it, based on MPEG features; processing is performed in strict real time. A probabilistic framework based on Bayes networks is used to detect the most significant soccer highlights. Predictions are validated by different Bayes networks, to check the outcome of forecasts.

1 Introduction

Sports videos are particularly important because of the extremely large audience: broadcasters produce huge amount of video that cover sport events every day. Broadcasters need to select the most relevant sequences from sports video documents ("highlights") for different purposes:

- Archiving for later reuse (posterity logging)
- Production of programmes in short time (live logging)
- Selective transmission of events to handheld devices and smart phones (real time semantic transcoding)

The latter purpose is becoming more and more interesting since this service is now provided by several mobile phone companies. In order to be able to provide an effective selection of the most interesting events of a video there is need of an automatic annotation system, that should perform real time analysis of an incoming video stream, marking the beginning of a possibly interesting sequence (containing an highlight) and then signaling the end of the interesting sequence. Possibly the system should be capable to even forecast a highlight (e.g. to provide real-time services such as mobile phone access) that will last some seconds with a certain probability. The probability may be used by the final user in order to select only certain highlights that are forecast with a minimum probability.

Since we are addressing a system that forecasts an highlight all the processing has to be performed in real-time. To this end we have considered a set of cues and a system architecture that allows to perform RT processing. A method to extract rapidly visual cues is to use features extracted from the compressed domain, e.g. MPEG motion vectors and MPEG DC components of DCT blocks.

In Sect. 2 we report on previous work done in the field of sport video annotation. Discussion of the usage of Bayesian Networks for our particular task is provided in Sect. 3. The description of our proposed approach is provided in Sect. 4. Results are discussed in Sect. 5, and conclusions in Sect. 6.

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 637–644, 2005.

2 Previous Work

Automatic sports video annotation has been addressed by several authors, with increasing attention in the very recent years. In particular rule-based modelling of complex plays for basketball is presented in [1] and in [2]. In this latter paper, basketball game shots are classified into one of three categories and basket highlights are detected from this classification, observing the occurrence of appropriate sequences of instances of these classes. In [3] Bayes networks have been used to model and classify American football plays using trajectories of players and ball. However, trajectories are entered manually, and not automatically extracted from the video stream. Kijak et al. [4] have used multimodal features to analyze tennis sports video structure. Models are used to integrate audio and visual features and perform stochastic modelling. Visual cues are used to identify the court views. Ball hits, silence, applause and speech help to identify specific events like scores, reserves, new serves, aces, serves and returns. Annotation of soccer videos has been addressed by a large number of researchers. Choi et al. [5] detect and track the ball and the players in the video sequence. The metric position of the players on the playfield is estimated using an approximation of the perspective planar transformation from the image points to the playfield model. In [6], the playfield is divided into several distinct zones. The framed zone is identified using patterns of the playfield lines which appear in the image. The ball position is also used to perform detection of shot on goal and corner kick events. In [7] MPEG motion vectors are used to detect events. In particular, they exploit the fact that fast imaged camera motion is observed in correspondence of typical soccer events, such as shot on goal or free kick. Recognition of relevant soccer highlights (free kicks, corner kicks, and penalty kicks) has been presented in [8]. Low level features like the playfield shape, camera motion and players' position are extracted and Hidden Markov Models are used to discriminate between the three highlights. More recently, in [9], Ekin et al. have performed event detection in soccer video using both shot sequence analysis and visual cues. In particular, they assume that the presence of highlights can be inferred from the occurrence of one or several slow motion shots and from the presence of shots where the referee and/or the goal post is framed. In [10] a system based on FSMs, that detects several different soccer highlights such as shot on goal, placed kicks, forward launches and turnovers, using visual cues has been presented. Ball trajectory is used by Yu et al. [11]. In order to detect the basic actions and compute ball possession by each team. Kalman filter is used to check whether a detected trajectory can be recognized as a ball trajectory. Experiments report detection of basic actions like touching and passing. Examples of detection of basic highlights in volleyball, tennis and soccer are reported. [12] has reported on detection of Formula 1 highlights using a multimodal fusion of cues and dynamic Bayes networks.

3 Probabilistic Highlight Modeling

Soccer highlights have a loosely defined structure. To capture the high intra–class variation that characterize the visual appearance of these events, we modeled

highlights using Bayesian networks (BNs). Bayesian networks [13] are directed acyclic graphs whose nodes represent random variables and whose edges correspond to direct dependencies between the variables. These dependencies are represented in a quantitative manner through conditional probability distributions. Among the reasons that make BNs appealing for our problem, the following are the most important:

- Factorization of the joint probability model. BNs represent the joint probability distribution defined by all possible points in the feature space into local, conditional distributions for each variable given its parents.
- Reasoning under missing observation. A BN is always able to produce an output, using all the evidence available. It does not require explicitly that all the observations are available. Moreover, even if observations are non–synchronized, the network still produce a valid output, hence different pieces of evidence can be gathered over time.
- Probabilistic output. The output of a BN is usually the posterior probability of an unobserved node, given the observations. This output can be directly related to user–centered preferences and needs.

For our particular task, a remarkable additional advantages of using BNs, stems from the causal interpretation that is usually associated to an edge in a BN. This give us a method to translate our knowledge into valid models. A top–down approach is adopted, which correspond to see observable features as directly "generating" higher level semantic events. We begin by defining a random variable for each of our observed feature, and a boolean random variable for the "highlight" node, which will tell us eventually whether an highlight is occurring or not. Directly connecting feature (input) variables to the output would result in a single conditional cpt, that would require us to specify a large number of parameters. To further factorize the joint probability distribution, we introduce additional, intermediate–level variables on which observed feature have a direct impact. To keep inference computation tractable for exact inference algorithms, we avoid to introduce cycles in the underlying undirected graph, ending in the reversed tree–like structure of fig.3 and 4.

Model parameters (i.e., probabilities in the CPTs) have been learned from labeled examples in a supervised way, as follows. Given N the number of a labeled example and $n(x)$ the number of time we observed event x, we use the following estimates for prior and conditional probabilities respectively:

$$P(x) \leftarrow n(x)/N \quad P(x|y) = P(x,y)/p(y) \leftarrow n(x,y)/n(y)$$

4 Real-Time Annotation

MPEG videos are used in order to extract as much visual features as possible from the compressed domain, to speed up the processing. In particular the system has been tested using MPEG-1 and MPEG-2 videos. Output of the BNs is used to detect interesting highlights, associating a confidence number to the beginning

and end of sequences that may contain a highlight, and thus allowing end users to set a sensitivity threshold to the system. In fact in the envisioned use case, where forecast highlights are transmitted to a handheld device, some users may prefer to get only very probable highlights, e.g. to reduce the costs related to video transmission, while other users may prefer to see more actions, accepting false alarms.

Only visual features are used by the system, since audio features may not be always available. The features may be divided in two groups: *compressed domain features*, that are extracted directly from the MPEG video stream:

- Motion vectors: MPEG motion vectors are used to calculate indexes of camera pan and tilt, and an index of motion intensity (see Fig. 1);
- Playfield: YUV color components are used to extract and evaluate the playfield framed.

and uncompressed domain features, that are extracted from images

- Players: players are extracted using previous knowledge of team colors (to improve precision) from uncompressed I frame: the ratio of pixels of the two teams is the cue used by the Bayes networks.
- Playfield lines: playfield lines are extracted from the uncompressed I frame: they are filtered out based on length and orientation.

The ratio of playfield framed allows to classify frames in three types (see Fig. 2): long, medium and close shot. The playfield area framed is classified in three zones, using the histogram of line orientation: left, center and right.

(a) (b)

Fig. 1. *a)*: original frame; *b)*: Motion vectors and average motion vector (long red line)

Evidence and inference are computed for each MPEG GOP (12 frames, i.e every 1/2 second in PAL video standard). If the highlight is predicted in the following 6 seconds (12 GOPs) the video is processed by the Bayesian validation networks. Conditional probabilities are updated every 2 secs. Four networks are used to predict highlights: two networks to predict attack actions (left-right) and two networks to predict placed kicks (left-right).

Fig. 2. *a)*: long; *b)*: medium; *c)*: close shot;

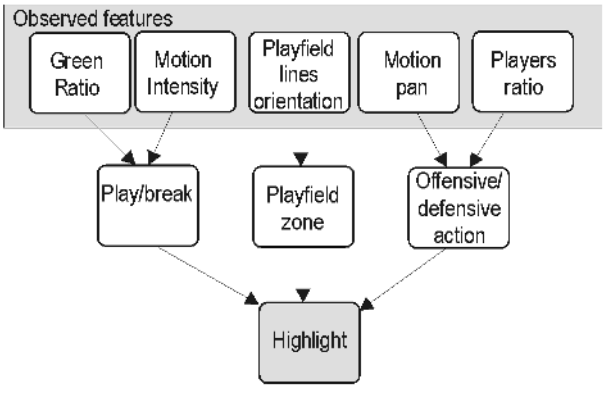

Fig. 3. Bayesian network used for highlight prediction (placed kick and attack action)

Fig. 3 shows the structure of the Bayesian network used for prediction; it is interesting to note that placed kicks are characterized by an initial break phase. The system is able to detect if the predicted action is concluded by a shot on goal. In fact when there is a shot on goal typically there is a sequence composed by 3 phases:

1. Fast panning of main camera toward goal post (Long Shot);
2. Zooming on the player who kicked the ball (Medium Shot or Close Shot);
3. View of the crowd or close up of the trainer (Close Shot).

The sequence is about the same in both cases of a scored goal or of a near miss, and the features that are extracted are the same. It must be noted that due to the soccer rules only the referee can decide if a ball that enters a goal post scores a goal; thus we can simply detect the presence of shot on goal, and not a goal.

To detect the shots on goal two networks are used, one for the left, and one for the right side of the playfield. The networks have the same structure, while the conditional probability tables of the nodes change every 2 seconds following the three typical phases described before. Fig. 4 shows the structure of the Bayesian network used for shot on goal detection.

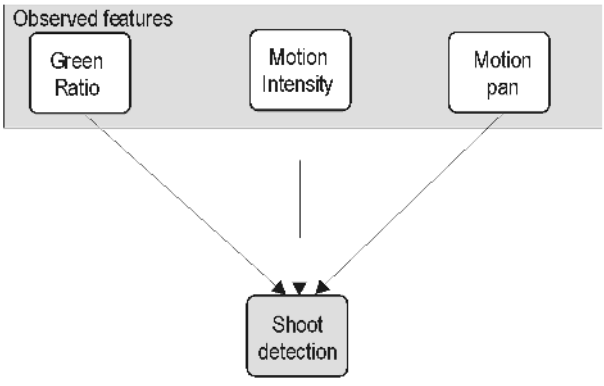

Fig. 4. Bayesian network used for shot on goal detection

The workflow of the system is: feature extraction from P and I frames, feature quantization and the prediction of the Bayes network is evaluated. If the evidence is above the user defined threshold then the shot on goal recognition is activated.

5 Experimental Results

The video stream used for the test set are MPEG-1 and MPEG-2 videos at 25 frames per second (PAL standard) and with a resolution that is respectively of 360×288 and 720×576. The GOP length is 12 frames. 268 case examples (~ 90 min) collected from World Championship 2002 and European Championship 2004 have been used to test the annotation system.

- 172 highlights that have been concluded with a shot on goal (SOG): 134 attack actions (AA) and 38 Placed kicks (PK)
- 54 highlights that have not been concluded with a shot on goal (NSOG): 51 attack actions and 3 Placed kicks
- 42 Other Actions (OA)

Table 1 and 2 report precision and recall figures, and a breakdown of the classification of SOG, NSOG and OA actions, and attack actions and placed kicks. The average number of frames between the prediction and the appearance of a SOG action is 74,2 (~ 3 sec. for a PAL system).

Typically the best results in terms of prediction of SOG are obtained in the case of attack actions or penalty kicks: in the latter case the prediction is performed when the large view of the player that is going to kick is shown. The lower precision of placed kick detection is due to cases of free kicks that are quite far from the goal box area; in this case the area is framed after kicking the ball, and the number of frames between prediction and the actual event is the lowest. Corner kick are less critical because of the typical large view, but since usually prediction starts after the kick the number of frames between prediction and event is quite low. Among the critical attack actions, that cause misses and

misclassifications there the cases in which the attacker does not directly kick the ball toward the goal post, but rather waits or makes small range assists to other team mates.

Analysing table 1, it must be noted that some results of the proposed system, while still not being the expected ones, are still acceptable. E.g. if a SOG is predicted but not recognized is still an acceptable result w.r.t the prediction requirements. The same applies to a NSOG that is predicted and then recognized a SOG. In fact these two types of errors affect only the validation of the forecast.

Table 1. Annotation performance of SOG, NSOG and OA. $\star$: expected behaviour; †: acceptable results; ‡: bad results.

Highlight type	Predicted and SOG recognized	Predicted and SOG not recog.	Not predicted	Precision	Recall
SOG	$151/172^{\star}$	$13/172^{\dagger}$	$8/172^{\ddagger}$	0.96	0.88
NSOG	$7/54^{\dagger}$	$43/54^{\star}$	$4/54^{\ddagger}$	0.74	0.80
OA	$0/42^{\ddagger}$	$2/42^{\dagger}$	$40/42^{\star}$	0.77	0.95
			Avg.	**0.83**	**0.88**

Table 2. Annotation performance of attack actions and placed kicks

Highlight type	Correctly detected	Misclassified/ missed	Precision	Recall
AA	163/185	22/185	0.98	0.88
PK	37/41	4/41	0.63	0.91
		Avg.	**0.83**	**0.88**

6 Conclusions

In this paper we have reported the results of real time annotation system, applied to soccer videos, that forecast the appearance of highlights in real-time, it classifies also the type of highlights and the presence of shots on goal. Our future work will deal with a refinement of the proposed system, extending and specializing the types of highlights that may be forecast, and extending the system to other types of sports.

Acknowledgments

This work has been partially funded by the European VI FP, Network of Excellence DELOS (2004-06). Authors would like to thank Filippo Conforti for his help.

References

1. W. Zhou, A. Vellaikal, and C.C.J. Kuo, "Rule-based video classification system for basketball video indexing," in *ACM Multimedia 2000 workshop*, 2001, pp. 213–126.
2. S. Nepal, U. Srinivasan, and G. Reynolds, "Automatic detection of 'goal' segments in basketball videos," in *Proc. of ACM Multimedia*, 2001, pp. 261–269.
3. S.S. Intille and A.F. Bobick, "Recognizing planned, multi-person action," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 414–445, March 2001.
4. E. Kijak, G. Gravier, L. Oisel, and P. Gros, "Audiovisual integration for tennis broadcast structuring," in *CBMI 2003*, Rennes France, 2003, pp. 421–428.
5. S. Choi, Y. Seo, H. Kim, and K.-S. Hong, "Where are the ball and players? soccer game analysis with color-based tracking and image mosaic," in *Proc. of Int'l Conference on Image Analysis and Processing*, 1997.
6. Y. Gong, L.T. Sin, C.H. Chuan, H. Zhang, and M. Sakauchi, "Automatic parsing of tv soccer programs," in *Proc. of IEEE Int'l Conference on Multimedia Computing and Systems*, 1995, pp. 15–18.
7. R. Leonardi and P. Migliorati, "Semantic indexing of multimedia documents," *IEEE Multimedia*, vol. 9, no. 2, pp. 44–51, April-June 2002.
8. J. Assfalg, M. Bertini, A. Del Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using hmms," in *Proc. of Int'l Conf. on Multimedia and Expo (ICME2002)*, 2002.
9. A. Ekin, A. Murat Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796–807, July 2003.
10. J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," *Computer Vision and Image Understanding*, vol. 92, no. 2-3, pp. 285–305, November-December 2003.
11. X.Yu, C. Xu, H.W. Leung, Q. Tian, Q. Tang, and K. W. Wan, "Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video," in *ACM Multimedia 2003*, Berkeley, CA (USA), 4-6 Nov. 2003 2003, vol. 3, pp. 11–20.
12. M. Petkovic, V. Mihajlovic, and W. Jonker, "Multi-modal extraction of highlights from tv formula 1 programs," in *Proc. of IEEE Int'l Conference on Multimedia & Expo*, 2002.
13. F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer, 2001.