A Decision Tree-based Multimodal Data Mining Framework for Soccer Goal Detection

Shu-Ching Chen
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199, USA
Email:chens@cs.fiu.edu

Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables
FL 33124, USA
Email:shyu@miami.edu

Min Chen, Chengcui Zhang
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199, USA
Email:{mchen005, czhang02}@cs.fiu.edu

Abstract

In this paper, we propose a new multimedia data mining framework for the extraction of soccer goal events in soccer videos by using combined multimodal analysis and decision tree logic. The extracted events can be used to index the soccer videos. We first adopt an advanced video shot detection method to produce shot boundaries and some important visual features. Then the visual/audio features are extracted for each shot at different granularities. This rich multi-modal feature set is filtered by a pre-filtering step to clean the noise as well as to reduce the irrelevant data. A decision tree model is built upon the cleaned data set and is used to classify the goal shots. Finally, the experimental results demonstrate the effectiveness of our framework for soccer goal extraction.

1. Introduction

Recently, mining information in sports video data, especially soccer videos, has become an active research topic. For soccer video analysis and event recognition, most of the existing work is based on unimodal approaches [2][4][8]. It is easy to see that different modalities have different contributions in soccer goal detection application domain [1]. As a multi-modal approach shows its promise, it also raises the issue of how to handle the rich semantic information contained in large amounts of multimodal features. In [5], a data mining method to detect and recognize soccer highlights using Hidden Markov Model was proposed. However, it cannot identify the goal event and has the problem to deal with long video sequences.

Data mining techniques have long been used to discover interesting patterns from large data sets. In this paper, we proposed a decision tree-based multimodal data mining framework for soccer goal detection. The training data for data mining is the multimodal features (visual and audio) extracted for each video shot. It is shot-based because video shots are the basic indexing unit for video content analysis [7]. In addition, we adopt an advanced video shot detection method, with the advantage of producing some important visual features and mid-level features (e.g., object information) during shot detections. Then, an unsupervised and robust grass area detection method is also proposed with very limited extra effort, which distinguishes our framework from most of the other existing approaches. However, due to the small percentage (e.g., 1%) of the positive samples (goal shots) with the huge amount of negative samples (non-goal shots), domain knowledge utilizing visual and audio clues has been used in our data pre-filtering step to clean the original feature data set in order to provide a reasonable input training data set for the data mining component. To our best knowledge, there is hardly any work addressing this issue. Finally, the decision tree model generated by the data mining process is tested and the overall performance is evaluated by using large amounts of long soccer video sequences with different styles and produced by different broadcasters. Based on our experiments, the results reach 92.3% for both recall and precision, which demonstrates the power of integrating data mining and multimodal processing.

The paper is organized as follows. Section 2 discusses the architecture of the proposed framework. Experimental results are presented in Section 3. Section 4 gives the conclusion.

2. Architecture of the framework

The architecture of our system is shown in Figure 1. As can be seen from this figure, the proposed framework consists of the following three major components:

Video Parsing: Parse the raw soccer video sequences by using a video shot detection

0-7803-8603-5/04/$20.00 ©2004 IEEE

subcomponent. It not only detects video shot boundaries, but also produces some important visual features during shot detection. The detected shot boundaries are passed to feature extraction, where the complete multimodal features (visual and audio) are extracted for each shot.

Data Pre-filtering: Use domain knowledge such as visual/audio clues to eliminate the noise data and reduce the irrelevant data from the original feature set since the ratio of goal shots over the non-goal shots is very small (e.g., 1 goal shot out of 100 shots). By data pre-filtering, the ratio of positive samples over negative samples can be increased to 1:20.

Data Mining: Take the 'cleaned' feature data as the training data, and build a decision tree model suitable for soccer goal detection.

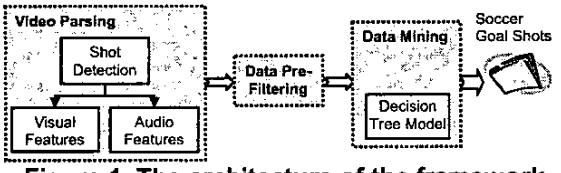

Figure 1. The architecture of the framework

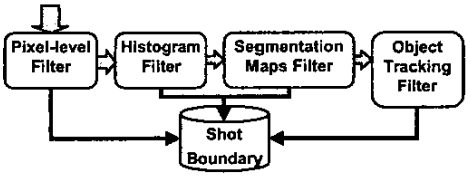

Figure 2. The multi-filtering architecture for shot detection

2.1. Video parsing

2.1.1. Video shot detection. Video shot detection is the first step for video parsing, and the detected shot boundaries are the basic units for video feature extraction. In this study, we improved our previous work [3] by using a multi-filtering architecture including the *pixel-level comparison*, *histogram comparison*, and *segmentation map* techniques (as shown in Figure 2). The first two filters can compensate for each other in reducing the numbers of both false positives and false negatives. In addition, since the object segmentation and tracking techniques are much less sensitive to luminance change and object motion, they are used as the last filter in this multi-filtering architecture, to help determine the actual shot boundaries. The advantages of this method are: 1) It has high precision (>92%) and recall (>98%) values. This overall performance is obtained based on our experiments over more than 1,000 testing shots. 2) It can generate a set of important visual features for each shot during the process of shot detection. Thus the computation for extracting visual features can be greatly alleviated.

2.1.2. Visual feature extraction. In addition to shot boundaries, the process of video shot detection also generates a rich set of visual features associated with each video shot. Among these visual features, *pixel_change* represents the average percent of the changed pixels between frames within a shot, which is output by the first filter (Pixel-Level Filter). The feature *histo_change* indicates the mean value of the histogram difference between frames within a shot, and is output by the second filter (Histogram Filter). Both of the two global features are important indications for camera motions and object motions. Other mid-level features such as the mean (*back_mean*) and the variance (*back_var*) values of the background pixels can be obtained via the segmentation filter. As shown in Figure 3(c)-(d), the background areas (black) and foreground areas (gray) are detected by object segmentation. In global view shots (Figure 3(a) and (c)), the grass areas tend to be detected as the background, while in close-up shots (Figure 3(b) and (d)), the background is very complex and may contain crowd, sign board, etc. Based on our observations, there is a large amount of grass areas present in global view shots (including **goal** shots), while there is less or hardly any grass area in the mid- or the close-up shots (including the cheering shots following the goal shots), which means the average percent of grass areas (*grass_ratio*) in a video shot is a very important indication for classifying shot types (global, close-up, etc.).

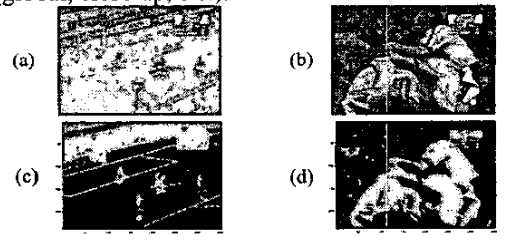

Figure 3. (a) a sample frame from a goal shot (global view); (b) a sample frame from the cheering shot following the goal shot for (a); (c)-(d) object segmentation results for (a) and (b).

We observe that the grass area in a global shot is relatively smooth in terms of its color and texture. Hence, the value of *back_var* < *threshold* will indicate the possible grass area. Then we group the *back_mean* values of all the possible grass areas into a candidate pool, filter off the outliers by taking out those shots that are too short and those whose *back_mean* values are out of a reasonable scope of the average

back_mean, and take the average of the remaining values as the grass detector. A robust method to handle a more complex situation when the grass colors are different between the global shots and the close-up shots caused by the camera shooting scale and lightning condition is also developed. In this case, we select the histogram peak(s) of the values in the candidate pool as the grass detector(s). It should be pointed out that this grass area detection method is unsupervised and the grass values are learned through unsupervised learning within each video sequence, which is invariant to different types of videos.

2.1.3. Audio feature extraction. Both time-domain and frequency-domain audio features are considered in our framework. Since the semantic meaning of an audio track is better represented by the audio features of a relatively longer period, we also explore both the clip-level and shot-level audio features. In this study, we define an audio clip with a fixed length of one second, which usually contains a continuous sequence of audio frames.

The generic audio features are divided into three groups: *volume features* (volume), *energy features* (energy), and *Spectrum Flux features* (sf). For each generic audio feature, the audio files are processed to obtain the audio features at both clip-level and shot-level. The audio data is sampled at a sampling rate of 16,000 HZ. An audio frame contains 512 samples, which lasts 32ms under a sampling rate of 16,000 HZ. Within each clip, the neighboring frames overlap 128 samples with each other. In order to model the energy properties more accurately, four energy sub-bands are also used in this study. In all, 10 audio features (1 volume feature, 5 energy features, and 4 spectrum flux features) are used in our framework.

2.2. Data pre-filtering

Due to the facts that the data amount is typically huge and the ratio of goal shots to non-goal shots is less than 1:100 in this study, an effective data pre-filtering method with the following three major observation rules for mining soccer goal events is proposed.

Rule 1: *As a candidate goal shot, the last three (or less) seconds of its audio track and the first three (or less) seconds of its following shot should both contain at least one exciting point.*

Rule 2: *A goal shot should have a grass_ratio larger than 40%.*

Rule 3: *Within two succeeding shots that follow the goal shot, at lease one shot should belong to the close-up shots.*

The first rule comes from the observation and the prior knowledge that the commentator and crowd become excited at the end of a goal shot. In addition, different from other sparse happenings of excited sound or noise, normally this kind of excitement will last to the following shot(s). This rule removes some of the noise data because, though normally the noise data has high volume, it will not last for long. Rules 2 and 3 are based on the observation that the goal shots belong to the global shots with a high grass ratio and are always closely followed by the close-up shots which include cutaways, crowd scenes and other shots irrelevant to the game without grass pixels.

Our experiments show that 81% of the video shots can be reduced by applying these rules in our proposed data pre-filtering method.

2.3. Mining goal shots using decision trees

In this framework, the decision tree logic is adopted for mining goal shots in soccer videos. The construction of a decision tree is performed by recursively partitioning the training set with respect to certain criteria until all the instances in a partition have the same class label, or no more attributes can be used for further partitioning. An interior node in a decision tree involves testing a particular attribute, and the branches that fork from that node correspond to all possible outcomes of a test. Eventually, a leaf node is formed which carries a class label that indicates the majority class within the final partition. The classification phase works like traversing a path in the tree. Starting from the root, the instance's value of a certain attribute decides which branch to go at each internal node. Whenever a leaf node is reached, its associated class label is assigned to the instance. The algorithm exploited in this study is adopted from the C4.5 decision tree [6].

In the decision tree generation process, the information gain ratio criterion is used to determine the most appropriate attribute for partitioning due to its efficiency and simplicity. Numeric attributes are accommodated by a two-way split, which means one single breakpoint is located and serves as a threshold to separate the instances into two groups. The voting of the best breakpoint is based on the information gain value.

3. Experimental results

3.1. Soccer video data and feature extraction

In our experiments, we collected 27 soccer video files from a wide range of sources via the Internet, with different styles and produced by different broadcasters. The total duration is 9 hours and 28

minutes. Among the total 4,885 video shots, only 41 are goal shots, which constitute only 0.8% of the total shots.

These video files are first parsed by using the proposed shot detection algorithm. Then both visual and audio features are computed and normalized for each video shot via the feature extracting processes. We include **10** audio features and **5** visual features in each feature vector and pass the feature set to the pre-filtering stage. The candidate shots generated by pre-filtering are then used for the data mining stage, which contains much less noise and outliers compared to the original data set. The resulting pool size after pre-filtering is **886**.

3.2. Video data mining for goal shot detection

These 886 candidate shots are randomly selected to serve as either the training data (666 shots, about 75% of the total data) or the testing data (the remaining 220 shots). The training data set contains 28 goal shots; while the other 13 goal shots are included in the testing data set.

• **Construct Decision Tree:** The decision tree is induced by the C4.5 approach based on the training data set. Both visual features (*histo_change*, etc.) and audio features (*volume_mean*, etc.) are used in constructing the decision tree. In addition, we explore another two effective features based on **Rule 1** (specified in Section 2.2). First, for each shot, the peak volumes of its last three-second audio track and its following shot's first three-second track (for short, nextfirst3) are summated as the feature *volume_sum*. Second, the mean volume of its nextfirst3 is acted as another audio feature *volume_nextfirst3*. Total **25** goal shots and **637** non-goal shots are correctly identified (i.e., labeled as "*yes*" and "*non*", respectively). In other words, only three "*yes*" and one "*non*" instances are misclassified.

Table 1. Testing result of goal shot detection

Total	Identified	Missed	Misidentified	Recall	Precision
13	12	1	1	92.3%	92.3%

Table 2. Overall performance

Total	Identified	Missed	Misidentified	Recall	Precision
41	37	4	2	90.2%	94.9%

3.3. Overall classification performance

The generated decision tree model is applied against the testing data set which contains 13 goal shots and 207 non-goal shots. The classification result of the goal shot classification on the testing data set using the proposed framework is shown in Table 1, where both the precision and recall are 92.3% (12/13). Table 2 gives the overall performance when both the training and testing data sets are used. The recall is 90.2% (37/41), and the precision is 94.9% (37/39). As we can see, the result is pretty satisfactory and encouraging.

4. Conclusions

In this paper, we have proposed a framework that uses data mining combined with multi-modal processing in extracting the soccer goal events from soccer videos. It is composed of three major components, namely video parsing, data pre-filtering, and data mining. Our experimental results over diverse video data from different sources have demonstrated that the integration of data mining and multimodal processing of video is a viable and powerful approach for effective and efficient extraction of soccer goal events.

5. Acknowledgement

For Shu-Ching Chen, this research was supported in part by NSF EIA-0220562, NSF HRD-0317692, and the office of the Provost/FIU Foundation. For Mei-Ling Shyu, this research was supported in part by NSF ITR (Medium) IIS-0325260.

6. References

[1] S. Dagtas and M. Abdel-Mottaleb, "Extraction of TV Highlights using Multimedia Features," *Proc. of IEEE International Workshop on Multimedia Signal Processing*, 2001.

[2] 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*, pp. 1040–1043, 2001.

[3] S.-C. Chen, M.-L. Shyu, C. Zhang, and R.L. Kashyap, "Video Scene Change Detection Method using Unsupervised Segmentation and Object Tracking," *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 57-60, 2001.

[4] P. Xu, et al., "Algorithms and System for Segmentation and Structure Analysis in Soccer Video," *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 928-931, 2001.

[5] J. Assfalg, et al., "Soccer Highlights Detection and Recognition using HMMs," *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 825-828, 2002.

[6] J.R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA, 1993.

[7] S.-C. Chen, M.-L. Shyu, C. Zhang, L. Luo, and M. Chen, "Detection of Soccer Goal Shots using Joint Multimedia Features and Classification Rules," *Proc. of International Workshop on Multimedia Data Mining (MDM/KDD'2003)*, pp. 36-44, 2003.

[8] K. El-Maleh, et al., "Speech/Music Discrimination for Multimedia Applications," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2445–2448, Istanbul, Turkey, 2000.