DETECTION OF SOCCER GOAL SHOTS USING JOINT MULTIMEDIA FEATURES AND CLASSIFICATION RULES

Shu-Ching Chen
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199, USA
305-348-3480
chens@cs.fiu.edu

Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
305-284-5566
shyu@miami.edu

Chengcui Zhang, Lin Luo, Min Chen
Distributed Multimedia Information System Laboratory
School of Computer Science
Florida International University
Miami, FL 33199, USA
305-348-6885
{czhang02,lluo0001,mchen005}@cs.fiu.edu

ABSTRACT
As digital video data becomes more and more pervasive, the issue of mining information from video data becomes increasingly important. In this paper, we present an effective data mining framework for automatic extraction of goal events in soccer videos. The extracted goal events can be used for high-level indexing and selective browsing of soccer videos. The proposed multimedia data mining framework first analyzes the soccer videos by using joint multimedia features (visual and audio features). Then the data pre-filtering step is performed on raw video features with aid of domain knowledge, and the pre-filtered data are used as the input data in the data mining process using classification rules. The proposed framework fully exploits the rich semantic information contained in visual and audio features for soccer video data, and incorporates the data mining process for effective detection of soccer goal events. This framework has been tested using soccer videos with different styles as produced by different broadcasters. The results are promising and can provide a good basis for analyzing the high-level structure of video content.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database Applications – Multimedia Data Mining.

General Terms
Algorithms, Experimentation.

Keywords
Event detection, multimodal data mining, integrated data mining in multimedia information systems, frameworks for multimedia data mining.

MDM/KDD'03, August 27, 2003, Washington, DC, USA.

1. INTRODUCTION
With the increasing amount of digital video data, mining information from video data for efficient searching and content browsing in a time-efficient manner becomes important. Motivated by the strong interest of automatic annotation of the large amount of live or archived sports videos from broadcasters, research towards the automatic detection and recognition of events in sports video data has attracted a lot of attentions in recent years. Soccer video analysis and events/highlights extraction are probably the most popular topics in this research area.

Xu et al. [9] proposed a soccer video segmentation method capturing the play/break segments according to whether the soccer ball is in play. Tovinkere and Qian proposed a hierarchical E-R model based on 3D data of the locations of players and ball, trying to model the semantic meaning and domain knowledge of soccer [3]. A set of rules are generated to determine whether an event happens or not. However, though the authors claimed that their approach is a complete solution towards soccer event identification, only two kinds of soccer events (deflection and save) are analyzed and discussed. In [10], a method to detect and recognize soccer highlights using Hidden Markov Model was proposed, in which each model is trained separately for each type of event. As shown in their preliminary results, this method can detect and recognize free kick and penalty event. However, it cannot identify the goal event and has the problem to deal with long video sequences. In [4], the authors focused on the presentation of soccer highlights by using mosaic images with no discussion on event detection and classification. Another approach that used a binary mask to detect the players and ball within the playing field to extract the playing field assuming that the playing field is always green was proposed in [12]. Recently, the approaches using multi-modal analysis have drawn increasing attentions [1][2]. In [1], a multi-modal framework using combined audio/visual/text cues was presented, together with a comparative analysis on the use of different modalities for the same purpose. However, the use of the textual transcript is not always available though it contains rich semantic information for event identification. Also, the grass-area-ratio was used in their paper as an important visual clue to identify the close-up segments following event segments. However, the proposed grass detection method is not robust because it needs training data for each new

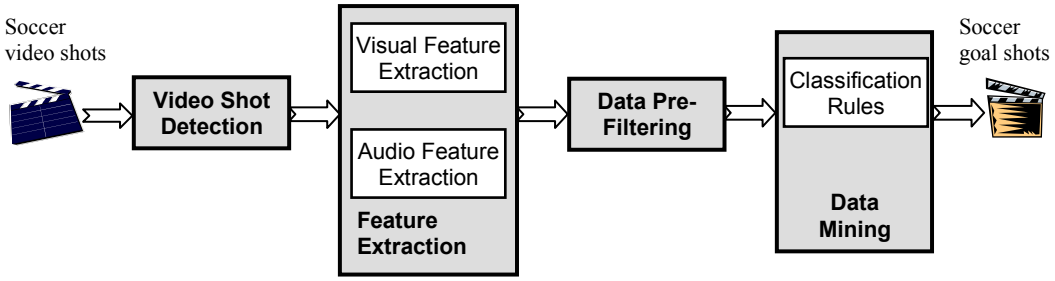

Figure 1. The architecture of the proposed framework

video sequence, which can only be manually done. Moreover, the precision value for goal event detection is poor.

In this paper, we proposed a shot-based multi-modal multimedia data mining framework for the detection of soccer goal shots. Multiple cues from different modalities including audio and visual features are fully exploited and used to capture the semantic structure of soccer goal events. The architecture of our system is shown in Figure 1. The raw soccer video sequences are parsed by a video shot detection component. The detected shot boundaries as well as some important visual features produced during shot detections are passed to the feature extraction component, where the complete set of visual features and audio features are extracted for each shot. There are two reasons to apply data pre-filtering before the actual data mining process: (1) the features may contain a lot of noisy and inconsistent data, and (2) in this specific application domain, the goal shots only constitute a very small amount of the video data (e.g., only 1 or 2 goal shots in a 40-minute long soccer video). The ratio of the positive samples over the negative samples is so small, which makes it unfeasible to directly apply the data mining techniques to the original feature set. In our framework, we propose a data pre-filtering method for mining soccer goal shots by fully utilizing the specific domain knowledge. The 'cleaned' feature data are then passed to the data mining component, where a classification model is trained by using 2/3 of the whole data set and tested using the rest 1/3 of the data. We have evaluated the performance of the proposed framework by using a large amount of soccer video data with different styles and different broadcasters. The experimental results demonstrate the effectiveness and great potential of our proposed framework.

The contributions of the proposed framework are summarized as follows:

- First, the advanced video shot detection method proposed in this work can not only output the shot boundaries, but also generate some important visual features during the process of shot detection. Moreover, since object segmentation is an embedded sub-component in video shot detection, the higher-level semantic information, such as the grass areas which serves as an important indication in soccer goal detection, can be derived from the object segmentation results. Thus just a small amount of work needs to be done in order to extract the visual features for each shot, which distinguishes our framework from most of the other existing approaches.
- Second, the proposed data pre-filtering step is critical in order to apply the data mining techniques to this specific application domain when considering the small percentage of the positive samples (goal shots) compared to the huge amount of negative samples (non-goal shots) in soccer video data. To our best knowledge, there is hardly any work addressing this issue.
- Third, when choosing the proper data mining technique, we take into consideration the special requirements of the specific application domain. Since the number of goal shots within a video is relatively small, missing a goal shot is not desired. Thus we choose to use the PRISM classification rule algorithm [5], which intends to cover all the positive samples in the training data set while not introducing false positives.

The paper is organized as follows. In the next section, we discuss audio/visual feature extraction. Section 3 presents the data pre-filtering and classification rule-based data mining for soccer goal shot detection. Experimental results are presented and analyzed in Section 4. Section 5 concludes our study.

2. VIDEO FEATURE EXTRACTION

2.1 Visual Feature Analysis and Extraction

2.1.1 Video Shot Detection

In this framework, the visual feature extraction is based on video shots. In our previous work [6], an effective and unsupervised video shot detection method using object segmentation and object tracking was proposed. In this study, we further improved both of its effectiveness and efficiency by combining *pixel-level comparison* and *histogram comparison* into the process of shot detection. The multi-filtering architecture for this method is shown in Figure 2. In the traditional pixel-level comparison approach, the gray-scale values of the pixels at the corresponding locations in two successive frames are subtracted and the absolute value is used as a measure of dissimilarity between the pixel values. If this value exceeds a certain threshold, then the pixel gray scale is said to have changed. The percentage of the pixels that have changed is the measure of dissimilarity between the frames. This approach is computationally simple but sensitive to digitalization noise, illumination changes and object moving. As a means to compensate for this, histogram comparison is incorporated into this method to reduce the false positives detected by the pixel-level comparison. Furthermore, since the object segmentation and tracking techniques are much less

sensitive to the above factors, they are used as the last filter in this multi-filtering architecture, to help determine the actual shot boundaries when both pixel comparison and histogram comparison failed. In other word, we apply the segmentation and object tracking techniques only when it is necessary for the sake of efficiency. According to our experiments on a large amount of video sequences (over 1000 testing shots), the overall performance of this shot detection method is promising in terms of precision (>92%) and recall values (>98%). With such a solid performance, only very little manual effort is needed to correct the false positives and to recover the missing positives.

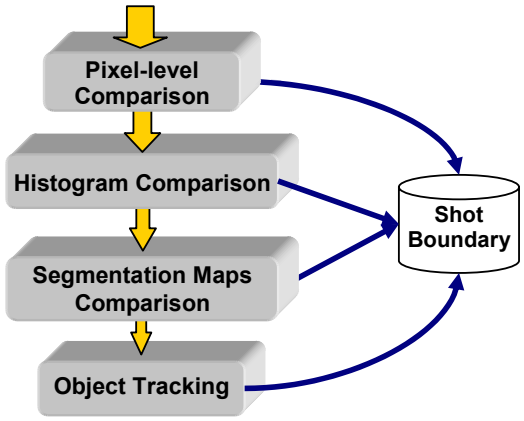

Figure 2. The workflow of the proposed shot change detection method

2.1.2 Visual Feature Extraction

The proposed video shot detection method can not only detect shot boundaries, but also produce a rich set of visual features associated with each video shot. For examples, the pixel-level comparison can produce the percent of changed pixels between consecutive frames, while the histogram comparison provides us the histogram difference between frames, both of which are very important indications for camera motions and object motions. In addition, the object segmentation can even produce the higher level of semantic information such as the object locations and areas. By taking these advantages brought by video shot detection, we include the following 5 visual features in our data mining framework for soccer goal detection:

Feature Name	Description
pixel_change_pencent	The average percent of the changed pixels between frames within a shot
histo_change	The mean value of the histogram difference between frames within a shot
grass_ratio	The average percent of grass areas in a video shot
background_var	The mean value of the variance of background pixels
background_mean	The mean value of the background pixels

Among the above 5 visual features, *pixel_change_pencent* and *histo_change* are obtained directly during the pixel-level comparison and histogram comparison for video shot detection, while the *background_var* and *background_mean* are obtained via object segmentation and can be used to obtain another domain-specific feature -- *grass_ratio*, which is a very important indication for classifying shot types (global, close-up, etc.) according to the video shooting scale. As we can see from Figure 3 (a)-(b), a large amount of grass areas are present in global 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). Another computable observation is that the global shots usually have a much longer duration than the close-up shots.

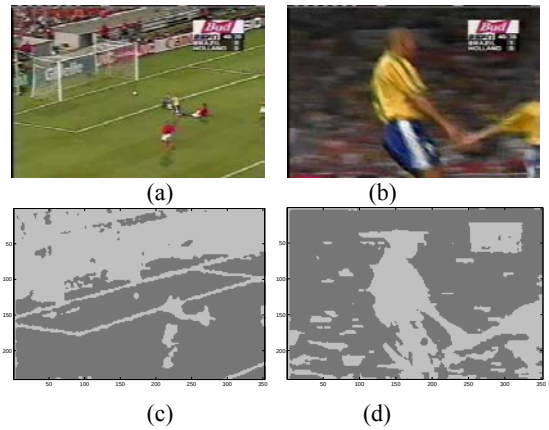

Figure 3. (a) a sample frame from a goal shot (global view); (b) a sample frame from the cheering shot following the goal shot for (a); (c) object segmentation result for (a); (d) object segmentation result for (b).

In the proposed framework, the *grass_ratio* is obtained via three steps:

1. Within each shot, draw a set of video frames at 50-frame interval and do object segmentation for them. By object segmentation, the background (grass, crowd, etc.) areas and foreground areas (player, ball, etc.) are detected as shown in Figure 3 (c)-(d), where foreground areas are marked with the gray color and background areas are marked with the black color. As can be seen from this figure, in global view shots, the grass areas tend to be detected as the background, while in close-up shots, the background is very complex and may contain crowd, sign board, etc.
2. Check the *background_var* of the background areas for each shot, if the *background_var* < *threshold*, indicating the possible grass area, then put its corresponding *background_mean* into a candidate pool containing possible grass values.
3. Once all the possible grass values are collected, filter off the outliers in the candidate pool by taking out those shots that are too short and those shots whose *background_mean* values are out of a reasonable scope of the average *background_mean.* Generate the histogram for the grass values in the 'purified' candidate pool. Based on our observations on a large set of video data, there are two possible situations in the histogram: (1) there is only one peak in the histogram, indicating the good video quality and stable lightning conditions, and (2) there are multiple peaks in the histogram, which correspond to the difference in grass

colors between the global shots and the close-up shots caused by camera shooting scale and lightning condition. In situation (1), the only one peak is selected as the grass pixel detector to calculate *grass_ratio*, while in situation (2), multiple peaks within a reasonable range are all selected as grass detectors. Base on our experiments, most of the multi-peak cases are caused by two different shooting scales (global, close–up) as shown in Figure 4. Figure 5 shows the detected grass areas for three sample images from different types of shots (close-up, global, etc.), and the results are very good.

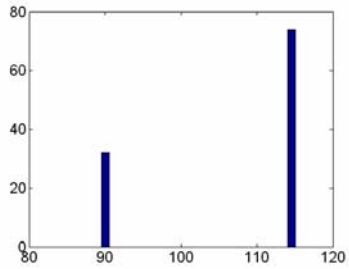

Figure 4. The histogram of the candidate grass values for a 20-minute long soccer video. Two peaks correspond to two major types of shooting scales in the video data – global and close-up.

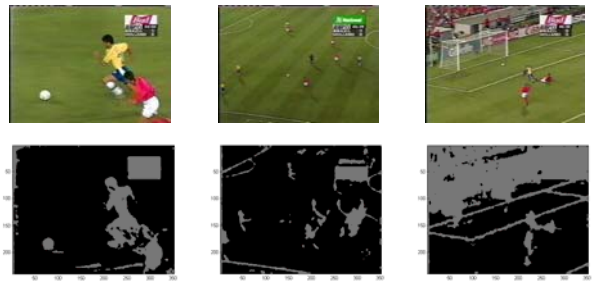

Figure 5. Detected grass areas (black areas) for 3 sample video frames from different types of shots

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 videos.

As a brief summary to the visual feature extraction:

1. First, it takes advantages of the advanced video shot detection process proposed in Section 2.1.1, such that only limited extra effort is needed in order to extract visual features for each shot.
2. Some high-level semantic information, such as the grass area ratio, can also be obtained automatically by using the object segmentation component in video shot detection.
3. This method has great potential to provide more high-level semantic information such as the locations of players and balls, and the spatio-temporal characteristics among objects as well.
4. Necessary data normalization is done within each video sequence. By doing this, the values of each visual feature are normalized to a [0, 1] range.

2.2 Audio Feature Analysis and Extraction

A variety of audio features have been proposed in the literature that can serve the purpose of audio tracks characterization [7] [8]. Generally, audio features can be classified into two major groups: time-domain features and frequency-domain features. And with respect to the analysis requirements of specific applications, audio features may be extracted in different granularities such as frame-level and clip-level.

Our framework exploits both time-domain and frequency-domain audio features. In order to investigate the semantic meaning of an audio track, high-level features that represent the characteristics of a comparable longer period are necessary. In our case, we explore both clip-level features and shot-level features, which are obtained via the computation and analysis of the finer granularity features such as frame-level features.

2.2.1 Audio Feature Analysis

The generic audio features utilized in our framework can be divided into three groups, namely, *volume related features*, *energy related features*, and *Spectrum Flux related features*.

Volume Related Features

Volume is one the most frequently used and simplest audio features. As an indication of the loudness of sound, volume is very useful for soccer video analysis. Four volume-based features used are:

Feature Name	Description
volumn_mean	The mean value of the volume
volumn_std	The standard deviation of the volume, normalized by the maximum volume
volume_stdd	The standard deviation of the difference of the volume
volume_range	The dynamic range of the volume, defined as $(\max(v) - \min(v)) / \max(v)$

Energy Related Features

Short time energy means the average waveform amplitude defined over a specific time window. To model the energy properties more accurately, energy characteristics of sub-bands are explored as well. Four energy sub-bands are identified, which covers respectively the frequency interval of 1HZ-(fs/16)HZ, (fs/16)HZ-(fs/8)HZ, (fs/8)HZ-(fs/4)HZ and (fs/4)HZ-(fs/2)HZ, where *fs* is the sample rate.

Feature Name	Description
energy_mean	The mean RMS energy
sub1_mean	The average RMS energy of the first sub-band
sub3_mean	The average RMS energy of the third sub-band
energy_lowrate	The percentage of samples with RMS power less than 0.5 times the mean RMS power
sub1_lowrate	The percentage of samples with RMS power less than 0.5 times the mean RMS power of the first sub-band
sub3_lowrate	The percentage of samples with RMS power less than 0.5 times the mean RMS power of the third sub-band
sub1_std	The standard deviation of the mean RMS power of the first sub-band energy

Spectrum Flux Related Features

Spectral Flux (Delta Spectrum Magnitude) is defined as the 2-norm of the frame-to-frame spectral amplitude difference vector.

Feature Name	Description
sf_mean	The mean value of the Spectrum Flux;
sf_std	The standard deviation of the Spectrum Flux, normalized by the maximum Spectrum Flux;
sf_stdd	The standard deviation of the difference of the Spectrum Flux, which is normalized too;
sf_range	The dynamic range of the Spectrum Flux.

2.2.2 Audio Feature Extraction

For each generic audio feature, we process the audio files and obtain the features at both clip-level and shot-level. The audio data is separated from the video data and sampled at a sampling rate of 16,000HZ, i.e., 16,000 audio samples are generated for an one-second audio track.

We use audio clips with a fixed length of one second, which usually contains a continuous sequence of audio frames. An audio frame is defined as a set of neighboring samples which last about 10~40ms. In our experiments, an audio frame consist of 512 samples, which last 32ms under the circumstance of a sampling rate 0f 16,000 HZ. Within each clip, the neighboring frames overlap 128 samples with each other.

The steps to obtain the abovementioned audio features are as follows:

1. Firstly, basic information such as the duration and the total number of frames of each video file is collected;
2. For each video file:
   a. Process the video file through the video shot boundary detection algorithm to identify its video shots;
   b. Separate the corresponding audio track from the original video file. Then:
      i. For each video shot, calculate the fifteen generic audio features of its first three seconds and last three seconds $firstVec_i$ and $lastVec_i$;
      ii. For each video shot, calculate the fifteen generic audio features vector SV_i;
      iii. obtain the normalized shot-level feature vector $NormSV_i$ via

$$NormSV_i = (SV_i - \min(SV_i))/(\max(SV_i) - \min(SV_i))$$

3. GOAL SHOT DETECTION

3.1 Data Pre-Filtering

Once the proper video features and audio features have been extracted, the data mining techniques can be applied to identify the goal shots. However, these features may contain noisy and inconsistent data which were introduced during the video production process. Moreover, the data amount is typically huge and the ratio of goal shots to non-goal shots is only 1:100 in our case. It would be difficult for the data mining process to capture the small portion of useful information from the huge amount of other irrelevant information. In the worst case, the goal shots may be treated as noise and ignored by the mining process. Therefore before performing the actual data mining process, for the sake of accuracy and efficiency, a pre-filtering process is needed to clean data and select a small set of candidate goal shots using domain knowledge. Here domain knowledge is defined as the empirically verified or proven information specific to the application domain that is served to reduce the problem or search space [11]. In this section, we present this pre-filtering process using some computable observation rules on the soccer videos, which can be classified into two categories, namely audio rules and visual rules.

3.1.1 Audio Rules

In the soccer videos, the sound track mainly includes the foreground commentary and the background crowd noise. Based on the observation and prior knowledge, 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). Thus the duration and intensity of sound can be used to capture the candidate goal shots as defined in the following rule:

- **Audio 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.

Here the exciting point is defined as a one-second period whose volume is larger than 60% of the highest one-second volume in this video. It is worth mentioning that actually this volume threshold can be assigned to an even greater value for most of the videos. However, based on our experiments, 60% is a reasonable threshold since the number of the candidate goal shots can be reduced to 17% of the whole search space while including all the goal shots. In addition, this rule performs as a data cleaning step to remove some of the noise data because, though normally the noise data has high volume, it will not last for long.

3.1.2 Visual Rules

As mentioned earlier, we have two basic types of shots, close-up shots and global shots, for soccer videos based on the ratio of the green grass area. We observe 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, as shown in Figure 6. Figure 6 (a)-(c) capture three consecutive shots starting from the goal shot (Figure 6 (a)), and Figure 6 (d)-(f) show another three consecutive shots where Figure 6 (d) is the goal shot. As can be seen from this figure, within two consecutive shots that follow the goal shot, usually there is a close-up shot (Figure 6 (b) and (f), respectively).

According to these observations, two rules are defined as follows:

- **Visual Rule 2**: A goal shot should have a grass ratio larger than 40%.
- **Visual Rule 3**: Within two succeeding shots that follow the goal shot, at lease one shot should belong to the close-up shots.

Figure 6. Goal shots followed by close shots: (a)-(c) three consecutive shots in a goal event. (b) is the close shot follows (a) the goal shot; (d)-(f) another goal event and its three consecutive shots, (f) is the close shot follows (d) the goal shot.

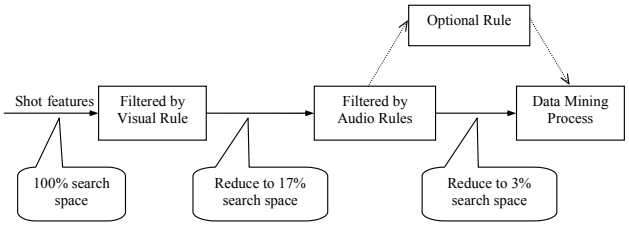

Figure 7. Pre-filtering processes

Note that the threshold defined in Rules 2 can be altered to a higher value for most of the videos. However, our experiments show that 82% of the candidate pool obtained after applying Rule 1 can be reduced using Rule 2 and Rule 3, which means that only 3% of the whole search space is remained as the input for data mining process. In addition, according to the prior knowledge, a goal shot normally lasts more than three seconds, which can be used as an optional filter called Optional Rule. In our case, since the search space has been dramatically reduced, this rule has small effects. In summary, the workflow as well as the performance of the pre-filtering process is illustrated in Figure 7.

3.2 Data Mining Process Using Classification Rules

In this study, we concentrate on discovering the semantic structures, if there is any, of the goal event in soccer videos, where such tasks fall into the area of data mining. In our framework, we choose to use the classification rule approaches to perform data mining process because this approach only induce absolutely accurate rules. On the other hand, the classification methods based on divide-and-conquer techniques normally are unable to achieve that. Since the number of goal shots within a soccer video file is comparatively much less than the number of non-goal shots, misclassifying a goal shot as non-goal shot is not desired and should be avoided. So the preferred classification approach should be able to identify all goal shots correctly even with the tradeoff that some non-goal shots (false positives) may be introduced into the results.

Figure 8 illustrates the workflow of the data mining process. Data after the pre-filtering process forms the training data set for the data mining process. Each data entry contains the shot- level audio and visual features plus the class label. The shot-level features are extracted in the feature extraction stage. A "YES" or "NO" class label is assigned to each shot manually; if there is a goal event, that shot is tagged as "YES". The data first go through the discretization process which operates to convert the numeric attributes in the training data into nominal ones, and then the PRISM algorithm [5] performs the rule generation task regarding the input training data. Rules can be generated with respect to each interested class, and can be applied against the testing data to identify the positive samples of the desired class later.

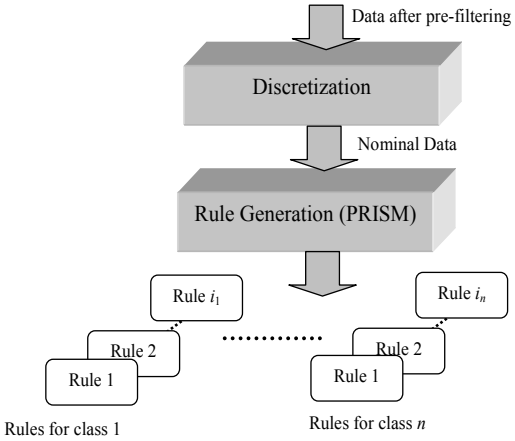

Figure 8. Workflow of mining goal shots in soccer video

3.2.1 Data Discretization

All the visual and audio features are numeric data. For the classification algorithm to accommodate numeric attributes, discretization is an indispensable step which converts the numeric attributes into nominal ones. A discretization scheme is used in our framework to perform the numeric-to-nominal conversion upon numeric attributes. The scheme first sorts the training instances with respect to the specific numeric attribute, and then places a breakpoint at the place where the class label changes to create the partitions. However, to avoid generating a large number of partitions in the discretization process, a minimal number of instances of the majority class that each partition has (e.g., 2 in our implementation). For example, as shown in Table 1, given a sequence of numeric values of attribute volume_mean, if we place the breakpoints wherever the class labels change, it will result in seven partitions as given in List 1. By introducing the constraint of minimum number of two instances of the majority class in each partition into the discretization scheme, three partitions shown in List 2 are produced.

Table 1. Discretization example

Volume_mean	0.2422	0.4000	0.4170	0.4753	0.5126	0.5326	0.5902	0.6172	0.6234	0.6381	0.6905
List 1	no	yes	no	no	no	yes	no	no	yes	yes	no
List 2	no	yes	no	no	no	yes	no	no	yes	yes	no

3.2.2 PRISM Approach

As a simple yet powerful covering approach to generate rules, the PRISM approach [5] is used in the proposed framework. The PRISM approach induces a set of perfect rules for each class with maximum accuracy. By "perfect", it means those rules cover all and only those positive instances for each class in the training data set. For each class, the algorithm starts with an empty rule which covers all instances, and then progressively restricts it by adding unexplored conditions that seem the most useful until the rules satisfy the correct classification criteria. The usefulness of condition is measured by the accuracy formula p/t, where t is the total number of instances covered by the current constructed rule and p is the number of positive instances of the desired class. If there are more than one attribute-value pairs with the same p/t value, then the one with the greatest coverage is chosen. The construction of the current rule stops if all the samples induced so far for that rule are of the same class, otherwise it continues to induce the rest of the rule from the reduced sample set using the remaining attributes. The same procedure applies for all the classes.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

4.1 Video Data Source

Soccer game video files used in our experiments were collected from a wide range of sources via the Internet. After excluding those video files that either have poor digital quality or do not contain any goal scene, there are nine video files left, with different styles and produced by different broadcasters. Those video files last from several minutes to half an hour. The corresponding sound tracks have been extracted from the original video files. These video files and audio tracks serve as the test bed for the proposed framework.

4.2 Video Data Statistics and Feature Extraction

Information Collecting

Information such as the total number of video frames and the duration are necessary and can be obtained by using video editing software. The average duration and number of frames are about 16 minutes and 26,000 frames, respectively.

To facilitate the extraction of audio and visual features that represent higher level semantic meanings, shot boundaries need to be located, which is achieved by parsing the video files by using the proposed shot detection algorithm. Because of the good performance of the shot detection algorithm (with >92% precision and >98% recall value), only little effort is needed to correct the shot boundaries. Averagely, those soccer video files contain about 150 shots. The detailed statistics of all the video files are listed in Table 2.

Feature Extraction and Instances Filtering

Both visual and audio features are computed for each video shot via the feature extracting processes presented earlier. We include fifteen audio features and four visual features in each feature vector. Pre-filtering techniques are applied to reduce the noise and outliers in the original data set, which generates the candidate shots pool for the data mining stage. The resulting pool size after per-filtering is 39.

Table 2. Detailed statistics of all the video data files

Files	Frame#	Shot#	Duration (sec)	Goal#
File 1	30,893	148	1,235.69	2
File 2	20,509	83	820.36	1
File 3	23,958	93	958.11	4
File 4	46,471	230	1,858.90	3
File 5	48,153	346	1,926.07	2
File 6	14,483	104	579.14	1
File 7	14,612	96	584.45	1
File 8	24,244	114	969.59	1
File 9	13,122	106	524.88	1
Total	***236,445***	***1,320***	***9,457.19***	***16***

4.3 Video Data Mining for Goal Shot Detection

These 39 candidate shots are randomly selected to serve as either the training data or the testing data. In our experiments, 27 shots (about 2/3 of the total data) are used as the training data set and the reaming 12 shots as the testing data set. The training data set contains 9 goal shots; while the other 6 goal shots are included in the testing data set.

The rules induced by the PRISM approach using the training data set are listed in Table 3.

Table 3. Rules generated by the PRISM approach

For "*NO*" class	
Rule 1	`histo_change is >=0.146133`
Rule 2	`sub3_lowrate is [0.640041,0.740794)`
Rule 3	`sub1_std is [0.734992,0.826217)`
Rule 4	`volume_mean is [0.712471,0.750064)`
Rule 5	`energy_mean is <0.269879`
For "*YES*" class	
Rule 1	`sub1_std is [0.663026,0.734992)`
Rule 2	`energy_mean is [0.764049,0.959954)`
Rule 3	`sf_range is [0.960759,0.992845)`
Rule 4	`volume_stdd is >=0.946543`

One problem of the classification rule approach is that the generated rules may not exclude each other and there is no ordering in the rules, and hence sometimes a testing instance may be classified to different classes. Such situations are called conflicts. In order to better resolve the conflict situation and

achieve more accurate goal shot classification performance, we further refine the classification strategy that is constructed by applying domain knowledge and guided by the special performance requirements for this specific application domain.

- Firstly, usually the goal shots belong to global shots, and later replay shots review the previous goal event in a closer view. Global shots in general have small or little change in color histogram and pixels. After investigating the set of induced rules, we found that rule 1 for the "NO" class, `histo_change is >=0.146133`, implies that clue. Therefore, this rule will be the last rule to classify the testing instance to determine its class label.
- Secondly, one desired property of the proposed framework is that the number of missed goal shots should be as small as possible. The reason is quite obvious: the observation of soccer video shows that the number of goal shots is much smaller than the number of non-goal shots; if a correct goal shot is misclassified, all the effort is in vain. Subsequently, in the case of conflicts, we consider the rules for the "YES" class are more important than rules 2 to 5 for the "NO" class (except rule 1 for the "NO" class), which means that as long as the testing instance can be classified to "YES" via one of the rules in the "YES" class, it will be considered as a goal shot. Such a class label will be changed to non-goal only if the testing instance can also be classified by rule 1 for the "NO" class.

Hence, when classifying the goal shots, the testing instance will be first examined by rules 2 to 5 for the "NO" class, followed by the rules for the "YES" class, and finally by rule 1 for the "NO" class to determine its class.

4.4 Overall Performance

The remaining 1/3 of the available shots are used as the testing data, among them there are six goal shots. Classification rules produced by the PRISM algorithm upon the training data are applied against the input testing data. Each testing instance is measured using our classification strategy as mentioned above, and a corresponding class label is assigned to the instance.

As shown in Table 4, the classification result is pretty satisfactory and encouraging. All six goal shots were correctly identified as "YES", while there was one non-goal shot misclassified as the "YES" class. Therefore, the recall value is 100%, and the precision value is 86% (6/7).

Table 4. Testing result of goal shot classification

Total Goal Shots	6
Identified Shots	6
Missed Shots	0
Misidentified Shots	1
Recall	100%
Precision	86%

Table 5 shows the overall performance of the goal shot classification using the proposed framework. Both training and testing data are used. The recall is 100%, and the precision is about 93% (15/16).

Table 5. Overall performance of goal shot classification

Total Goal Shots	15
Identified Shots	15
Missed Shots	0
Misidentified Shots	1
Recall	100%
Precision	93.75%

5. CONCLUSIONS

In this paper, we have presented a new multimedia data mining framework for the detection of soccer goal shots by using combined multimodal (audio/visual) features and classification rules. The output results can be used for annotation and indexing of the high-level structures of soccer videos. Specifically, we adopt an advanced video shot detection method, with the advantage of producing important visual features and even high-level semantic features during shot detections. Then by exploring the unique domain knowledge in soccer video data, the collected audio/visual features are cleaned by data pre-filtering, in order to provide a reasonable input data set for the proposed data mining process. The classification rules generated by the data mining process are further refined to resolve the conflicts in rule matching. Our experiments over diverse video data from different sources have demonstrated that our framework is very effective in classifying the goal shots for soccer videos.

6. ACKOWLEDGEMENT

This research was supported in part by NSF CDA-9711582, NSF EIA-0220562, and the office of the Provost/FIU Foundation.

7. REFERENCES

[1] Dagtas, S., and Abdel-Mottaleb, M. Extraction of TV highlights using multimedia features. IEEE International Workshop on Multimedia Signal Processing, 2001.

[2] Zhu, W., Toklu, C., and Liou, S.-P. Automatic news video segmentation and categorization based on closed-captioned text. In IEEE International Conference on Multimedia & Expo, 1036–1039, Tokyo, Japan, 2001.

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

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

[5] Cendrowska, J. PRISM: an algorithm for inducing modular rules. International Journal of Man-Machine Studies, 27, 4, 349-370, 1987.

[6] Chen, S.-C., Shyu, M.-L., Zhang, C., and Kashyap, R.L. Video scene change detection method using unsupervised segmentation and object tracking, IEEE International Conference on Multimedia and Expo (ICME), 57-60, August 22-25, 2001, Waseda University, Tokyo, Japan.

[7] Liu, Z., Wang, Y., and Chen, T. Audio feature extraction and analysis for scene segmentation and classification. Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology, 1998, 20, 1/2, 61-80, 1998.

[8] Wang, Y., Liu, Z., and Huang, J. Multimedia content analysis using both audio and visual clues. Signal Processing Magazine, 17, 12 - 36, Nov. 2000.

[9] Xu, P., Xie, L., Chang, S.-F., Divakaran, A., Vetro, A., and Sun, H. Algorithms and system for segmentation and structure analysis in soccer video. IEEE International Conference on Multimedia and Expo, Tokyo, Japan, Aug. 22-25, 2001.

[10] Assfalg, J., Bertini, M., Bimbo, A. D., Nunziati, W., and Pala, P. Soccer highlights detection and recognition using HMMs. IEEE International Conference on Multimedia and Expo 2002.

[11] Witten, H., and Frank. E. Data mining-practical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers, 1999.

[12] Gong, Y., Sin, L.T., Chuan, C.H., Zhang, H., and Sakauchi, M. Automatic parsing of TV soccer programs. In Proceeding of IEEE Multimedia Computing and Systems, Washington D.C., 1995.