Abstract

Unlike American football, baseball, tennis and many other sports games, soccer is not a well-structured game. Soccer videos are basically continuous streams with exciting highlights embedded. However, the highlights of the same type are correlated in spatial and temporal feature distributions. In this paper we present an effective scheme to represent soccer scenes with low/mid-level image and sound features. We discuss three aspects of feature design in soccer video indexing system: temporal structure, low/mid-level features, domain specific knowledge. We use the maximum-entropy based machine learning method as a test platform to verify the feature design scheme. The maximum-entropy method can automatically choose the features with more distinguishing power. The feature representation is applied to soccer video indexing. Extensive experiments are conducted and satisfying results are described.

1. Introduction

Videos provide far richer information than static images. At the same time the large volume of videos makes them more difficult to access, archive and process. Therefore, video indexing plays a key role in video processing. We are interested in video representations which take advantage of temporal information to achieve reliable and efficient video indexing. There are always three problems on video representation: 1. How to incorporate temporal information? 2. Are low-level features enough? 3. How much should an approach rely on domain knowledge? In this paper we describe our feature design scheme which statistically includes information from temporal context within a video. We represent videos by low/mid-level image and sound features because only the low-level, or at most some mid-level features, can be extracted efficiently and reliably. Domain knowledge is incorporated in the feature design to improve the performance of video indexing.

The feature design scheme is applied to a soccer video indexing system which is able to extract interesting portions of a game, and classify them into four types: goal, shot-on-goal, corner kick, free kick. The system has been extensively tested on 18 world cup games of 1994, 1998 and 2002. We use the maximum-entropy based method to fuse the features from image and sound streams. The maximum entropy based method works as a test platform to verify the feature design scheme since it has the automatic feature selection capability, which chooses the features with more distinguishing power in the statistical modelling process. The features selected by the maximum-entropy method indicate that: 1. Temporal context information is very important to identify sports highlights because the same type of highlights undergoes similar transitional patterns, 2. Low/mid-level features, such as color, camera motion, line markings distribution and soccer ball trajectory, are able to generate satisfying soccer indexing results, 3. Domain specific features designed for soccer games, such as goalpost detection, can improve the video indexing performance.

1.1 Related Work

There have been research activities that strive to detect highlight scenes and/or analyze game structures for broadcast sports games. Some research studies depend only on image features. These methods generally have limited capabilities of classifying sports highlights. Gong et al. identified the play location in a soccer game by tracing field line markings and identifying ball in key frames [1]. Yow et al. presented soccer highlights on panoramic images by detecting ball, players and goalposts [2]. Xu et al. described a color-based method to segment soccer videos into play and break segments [3]. Utsumi et al. detected and tracked players on soccer fields and they did not perform content analysis [4].

Rui et al. detected baseball highlights using audio-track features alone [5]. They analyzed the pitch and energy level of audio signals to extract the exciting game portions, however, it is difficult to use the same scheme on events classification.

Tovinkere and Qian developed a system for soccer events recognition based on the player and ball positions in the field [6]. Their system required specific devices (microwave signals) or accurate 3D position reconstruction.

Babaguchi et al. worked on collaboration of visual and closed caption streams for American football games [7]. Chang et al. proposed a video indexing method for American football based on image and speech features [8]. These two methods are both based on heuristic rules, which makes them have little room for further extensions. We described
our work on baseball highlights detection and classification using multimedia features in [9].

Assfalg et al. classified soccer highlights into penalty, free kick and corner using Hidden Markov Models based on camera motion and manually annotated player locations [10]. They assumed that the camera motion indicated the ball position. The system was strongly dependent on the features for soccer highlights indexing.

2. Feature Design

Many sports games are well structured, such as baseball, tennis, American football. Each game of these sports is composed of syntactic units, for example, a baseball game consists of 9 innings and each inning has 3 outs for each team. Every unit starts with canonical scenes, such as pitch scene in baseball, serve scene in tennis and scrimmage scene in American football. Comparing with these games, soccer games are continuous streams. The entire game just flows smoothly without syntactic units. There are no canonical scenes except the kickoff at the beginning of each half and after a goal, however, the after-goal kickoff is rarely shown.

To identify the content of a play, intuitively, we need to know the play location, process and result. We classify four types of highlights in soccer games: goal, shot-on-goal, corner kick and free kick. The first two define the results of a play and the other two define the processes which lead to the results. The play location helps to recognize if the play happens close to the goal area, which has a higher probability of being a highlight. These high level concepts could be solved individually by line markings tracing, player tracking, speech recognition which are computation expensive and sensitive to noise. We propose a combination of low-level and mid-level, domain independent and specific features for soccer highlights indexing.

2.1 Low-Level Features

• Color/Edge distribution
  Given the layout of color and edge in a scene shot of a soccer game, it is easy to distinguish if the camera is shooting at the field, bench or audiences. We use the percentages of field pixels (green color) and edge pixels to represent the color/edge distribution.

• Camera motion

We apply a simplified pan-tilt-zoom camera model to capture the camera motion information through each scene shot. This model fits the camera motions for sports broadcast very well. Since scenes in sports games are dynamic, the camera motion is calculated based on the robust estimation scheme. We apply RANSAC [11] to the camera motion computation by randomly choosing a small subset of images to obtain an initial solution where the subset defines a rigid scene, and then identifying the outliers which are the points with dynamic motion. The process is repeated enough times on different subsets and the best solution is the one which maximizes the number of points lying in the rigid scene.

• Line markings distribution
  To avoid the costly computation of line tracing [1], we use hough transform to get a global distribution of line markings which helps to identify play locations. We first extract white edge pixels from grass background. These pixels certainly include line markings, player jerseys, socks, posts and noise. Without identifying the edge pixels corresponding to line markings, we conduct a global hough transform on all of the edge pixels. In our experiments, the resolution is $10^2$ for line direction and 10 pixels for distance from origin to line. We generate 36 parameters: 18 parameters representing the numbers of edge pixels aligned with 18 directions, and the other 18 parameters denoting the numbers of pixels at 18 distances from origin. These 36 parameters provide a global image of line markings distribution. The parameters exhibit different patterns for various field locations, such as lower middle field, upper left corner, middle right goal. To verify the representation, we conduct a side experiment to detect play locations based on these parameters only. We divide the soccer field into 9 portions, and perform a supervised training based on the 36 parameter representation to identify the field location. The precision on test images is about 84% which is good enough for our video indexing system since this feature is not the only one to recognize play locations.

• Pitch and power of audio signals
  We calculate the pitch and power values of the audio stream as [5]. We use the percentage, average, max and range values of pitch and power as audio features. These values are helpful in detecting highlights.

2.2 Mid-Level Features

• Ball trajectory approximation
  Since ball tracking is a difficult and computation expensive problem, we propose an approximation strategy to identify and track the soccer ball together. Our goal is to detect fast moving ball along straight lines.
This is a reasonable assumption for highlight scenes. We first extract small white blobs from individual images, then perform a line fitting across the time domain. The images are first stabilized by camera motion compensation, and then are put together into a 3D volume. We perform a 3D hough transform on the image volume. The most possible line represents the soccer ball trajectory, and its corresponding projection on each image provides the ball detection. This method deals with occlusions, appearance changes and global camera motions very well. It does not require prior segmentation or initialization. The speed, position and trajectory of the ball are the features. We regard these as mid-level features because the computation is based on low-level features, such as color, edge and camera motion. Meanwhile, the trajectory information is represented as 2D locations. It does not imply the high-level concepts of ball/player layout or team strategy.

- **Motionless detection**
  Instead of tracking multiple players to detect the player distribution, such as “human wall” distribution for free kick, which is accomplished mostly by hand [10] or by heuristic rules, we try to detect “motionless” portions in a scene shot. We identify the edge pixels on the field without relative motion. The larger value of the motionless feature implies the higher probability of being followed by a free kick or corner kick highlight. We extract vertical edge pixels on the field and then perform global camera motion compensation. Only the pixels without relative motion are counted.

### 2.3 Temporal Features

Take a closer look at soccer games. The same type of highlights still share similar patterns. The patterns exist in the temporal context of the highlight scenes, but they are not demonstrated in the form of scene transitions as baseball, tennis or American football. For example, a home run highlight in baseball is usually composed of four consecutive scene shots: pitch scene, outfield overview, audience scene and player running scene. Therefore, highlights can be recognized by using concatenated features extracted from consecutive scene shots as input to the statistical modelling, as shown in Figure 1. In comparison, a corner kick highlight in soccer usually covers one or two scene shots showing a global view of the goal area. At some point before the highlight, we may see a scene with the corner flag posts. Similarly, at some point after the goal highlight, cheering players might be shown. Instead of collecting all the features from contextual scene shots, we statistically estimate the temporal information within a fixed period of time before and after the current scene shot, as shown in Figure 1. Empirical results show that 30-40 seconds is an appropriate period of time to extract contextual information.

![](image.png)

Figure 1: Temporal structure comparison of baseball and soccer games. Black denotes the current scene shot, light grey represents its context, dark grey indicates the highlight portions. We use 40 seconds videos before and after the current scene shot to extract temporal information.

We accumulate the temporal features by sampling and statistical voting. We extract one set of low and mid-level features for every 15 frames before and after the current scene shot and within 40 seconds range. The weighted average values are computed. The closer time points have higher weights. Therefore, temporal information is encoded into two sets of features for beforehand and afterward contexts, respectively.

### 2.4 Domain Specific Features

The above features can work on many other sports games. We also design these features which are domain dependent and easy to implement: goalpost, corner flag post and net detection. These features can be efficiently extracted. Not surprisingly, they improve the indexing performances of goal and corner kick categories.

### 3 Statistical Platform

The maximum-entropy method is a statistical technique based on exponential models for selecting and combining features into a predictive model. For every type of highlight, the model is able to assign a probability to each scene shot – the probability that the scene shot is classified as that type of highlight. The probability distribution is chosen by incrementally building a log-linear model that weighs different features of the surrounding context. Beeferman et al. applied the model to text segmentation where they used binary features only [12]. We use real value features to include the uncertainty information of feature extraction.

The goal of statistical modelling is to build a model \( p(h|X_i) \), where \( h \in \{ \text{goal, shot-on-goal, corner kick, free kick, junk} \} \) is a random variable corresponding to the presence of any type of highlight in the context \( X_i \). \( X_i \) is a feature sequence which represents the context of \( i \)th scene shot,

\[
X_i = I_{i-t}, I_{i}, I_{i+t}
\] (1)

This is a reasonable assumption for highlight scenes.
assuming we are considering the \( i \)th scene shot. \( I_i \) denotes the features extracted from the current scene shot, \( I_{i-t} \) represents the features accumulated over the time period \( t \) beforehand and \( I_{i+t} \) represents the afterward. The dimension of \( I_i \) is \( l \), therefore, \( 3 \times l \) feature functions are defined for each context: \( f_{k}(X_i), k = 1, \ldots, 3 \times l \).

The maximum-entropy principle chooses the most uniform distribution from the set of those models \( p(h\mid X_i) \) which agree with the training data. We denote the selected model as \( p_\lambda \),

\[
p_\lambda(h\mid X_i) = \frac{1}{Z_\lambda(X_i)} \exp \left( \sum_k \lambda_{kh} f_k(X_i) \right) \tag{2}
\]

where

\[
Z_\lambda(X_i) = \sum_h \exp \left( \sum_k \lambda_{kh} f_k(X_i) \right) \tag{3}
\]

and \( \lambda_{kh} \)'s correspond to the weights of different features in the integration. We use the Improved Iterative Scaling Algorithm to compute the parameters \( \lambda_{kh} \) [13].

\section*{3.1 Feature Selection}

The maximum-entropy method provides a nice scheme to incorporate automatic feature selection within the exponential models. Each stage of the process is characterized by a set of active features which determine a space of models. The optimal model in this space is the model with the greatest entropy. At each stage of the model-construction process, our goal is to select the candidate feature which, when adjoined to the set of active features, produces the greatest increase in likelihood of the training sample.

\section*{3.2 Test Platform}

The maximum-entropy method chooses the feature with the most distinguishing power each time it performs feature selection. The features selected and their orders indicate the feature importance and independence. In this way we can use the maximum-entropy method as a test platform to verify how much the indexing system is dependent on low-level features, how important the mid-level features are and how necessary the domain specific features are. We can also tell if the temporal contextual features should be included or not.

\section*{4. Performance}

In this section we present the experimental results of applying the feature design scheme to soccer video indexing. 18 world cup games of year 1994, 1998 and 2002 are used. These games were directly recorded from TV. These 18 games include 22 different teams and 8 stadiums, broadcasted by 7 different TV stations. We use 9 games as training data and 9 games as test data.

Given each scene shot, we first cut it into three segments with equal length in order to incorporate the transitions within each scene shot, then generate 242 low-level image features and 21 sound features. We extract 10 mid-level features from the ball trajectory approximation and motionless percentage, such as ball maximum speed, average speed, starting and ending locations within the given scene shot. We generate 24 domain specific features for each scene shot. Each of the three segments has 8 features, such as goalpost location, net pixel percentage, corner flag post size, etc. Since we use the temporal information before and after the current scene shot to capture the context, the dimension of the feature vector is much higher. In our experiments, the dimension of the feature vector is 891 (dimension of \( X_i \) in Equation (1)) while the dimension of the feature vector from each scene shot is 297 (dimension of \( I_i \) in Equation (1)). The number of selected features is determined by validation data. We use one game as validation data (the other 8 games as training data). The feature selection process automatically stops when the validation data achieve the best classification results. The selected feature number is 100 in our experiments, including 62 low-level image features, 4 sound features, 15 mid-level features and 19 domain specific features. 18 out of these 100 features are accumulated from the temporal context. Table 1 summarizes the dimension reduction of each type of features.

On average, the recall and precision for highlight classification on 9 test games are 71.31% and 80.12%, respectively. Table 2 lists the recall and precision for each type of highlight. The recall for \textit{free kick} is relatively low because the players have diversified distributions in locations and patterns at free kicks.

To get a better insight of the system performance, we take
Table 1: Feature dimension reduction

<table>
<thead>
<tr>
<th>feature dimension</th>
<th>low-level image</th>
<th>low-level sound</th>
<th>mid-level</th>
<th>domain specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>scene shot $I_i$</td>
<td>242</td>
<td>21</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>context $X_i$</td>
<td>726</td>
<td>63</td>
<td>30</td>
<td>72</td>
</tr>
<tr>
<td>selected features</td>
<td>62</td>
<td>4</td>
<td>15</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 4: Highlight classification results with and without mid-level features on four world cup games

<table>
<thead>
<tr>
<th>highlight</th>
<th>with mid-level features</th>
<th>without mid-level features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>recall</td>
<td>precision</td>
</tr>
<tr>
<td>goal</td>
<td>75.00%</td>
<td>60.00%</td>
</tr>
<tr>
<td>shot-on-goal</td>
<td>74.36%</td>
<td>89.23%</td>
</tr>
<tr>
<td>corner kick</td>
<td>68.52%</td>
<td>77.08%</td>
</tr>
<tr>
<td>free kick</td>
<td>36.84%</td>
<td>43.75%</td>
</tr>
</tbody>
</table>

a closer look at the performance statistics of four test game using the same selected feature set, as shown in Table 3 and the left hand side of Table 4. We also test the same indexing system without mid-level features where the feature number is also 100. Table 4 compares recall and precision of the indexing system with and without mid-level features. It shows that both the recall and precision are dramatically improved (around 25%) knowing the ball information and no-motion statistics.

5. Conclusion

In this paper we describe the feature design scheme to deal with video representation which statistically incorporates the temporal information and achieves satisfying video indexing performance based on low and mid-level features with the minimum requirement on domain knowledge. The maximum-entropy method is used to integrate the features and select the features with more distinguishing power. The features chosen verify our feature design scheme. We demonstrate the experimental results on 18 world cup games from different years played by different teams.

There is still room to improve the indexing results. It is always a tradeoff between performance and computation effort to generate features. Though the experiments show that the mid-level features drastically improve the classification results, the computation cost of these features are 3-4 times higher than the other low-level and domain specific features.

Most features presented in this paper are ready to be applied to other sports games except the domain specific features, such as tennis, American football, baseball and hockey. We are working on a systematic scheme to design domain specific features.

References