Attractive Events Detection in Soccer Videos Based on Identification of Shots

Fang Tian[1] and Ping Shi[2]

Abstract The necessity of efficient identification and classification of shots in a soccer game ascends with the increasing popularity and quantity of soccer videos. This paper proposes an effective and real-time identifying system for attractive events, namely shooting shots in soccer video clips. Instead of directly accessing shooting shots, this system identifies close-up shots and audience shots, which generally accompany with shooting shots. Field shots are concerned in this system as well, since this sort of shots is frequently used to describe formations of a team, indicating its strategy. This system begins with the selection of key frames. The ratio of soccer pitch color pixels in the key frame is then calculated to detect long shots. By counting the amount of edge pixels, the differentiation between close-up shots and audience shots will be completed. In order to increase the accuracy, the approach of Key Region Detection is introduced. This innovative method is proven by experiments to be comparatively precise. Experimental results validate the effectiveness of this system.

Keywords: field shots, close-up shots, audience shots, key frame.

1 Introduction

The abstraction of soccer video enjoys high level of popularity, due to its ability to satisfy the requirements of audience, academic researchers and commercial campaigners. Approaches of soccer video abstraction could be generally divided into two groups: the low-level feature based analysis method and the high-level feature based analysis method.

The low-level feature based analysis method focuses on video and audio characteristics in a clip. The commonly used video characters include color and texture information of a frame. D. Yow, et al.[1] detected appealing events in a

[1] Fang Tian (✉)
College of Information Engineering, Communication University of China, Beijing, China 100024, e-mail: rousongsong@126.com

[2] Ping Shi
College of Information Engineering, Communication University of China, Beijing, China

soccer video by extracting color and texture features from frames. S. Intille, A. Bobick[2], and V. Tovinkere, et al.[3] studied the movement tracks of players and the football to implement the detection of attractive events. B.X. Li, et al.[4] used different camera languages by analyzing slow-motion shots, close-up shots and cutting shots, then detected the change of scoreboard which was added in a soccer video to locate the attractive events precisely.

Audio information refers to the sound of audience and commentators. The yelling and applause of audience always take place when events like shooting or scoring happen. The pitch changes of commentators generally indicate the occurrence of special events in a game. Z. Xiong, et al.[5] utilized the features defined by MPEG7 and classified audio information during a soccer video. R. Radhakrishnan, et al.[6] applied Gaussian Mixture Model (GMM) to classify audio information into seven types, then extracted appealing events using Markov Model. The low-level feature based approaches are relatively easy to complete, whereas its fallibility should not be neglected.

To bridge the gap between the low-level features of an image and the comprehension of human beings, high-level feature based method for video abstraction is introduced. The kernel idea of this approach is to use attention modals to filter results of low-level feature based approaches with the help of subjective model. For instance, shooting shots bring audience with different subjective perspectives due to the diverse patterns of shooting, and whether the shooting is successful or not. J.Q. Yu, et al.[7] introduced the degree of attractiveness to evaluate various subjective perspectives of shots.

This paper proposes a real-time system for identifying attractive events by identifying field shots, close-up shots and audience shots. This system begins with the selection of key frames. The ratio of soccer pitch color pixels (abbreviated as SPC pixels in this paper) in a key frame is then calculated to detect field shots. After gray conversion and median filtering, edge feature of these frames is extracted to differentiate close-up shots and audience. In order to achieve a better identifying result, an innovative method of Key Region Detection is presented. Experimental results validate the efficiency of this system.

The paper is organized as follows: a classification of shots in soccer videos is introduced in Section 2; the algorithm of this identification system and the improvement are elaborated in Section 3; experimental results and analysis are presented in Section 4 and conclusions are drawn in Section 5.

2 Classification and Features of Shots in Soccer Videos

According to the result of summarizing large number of soccer video clips, shots in a soccer game could be roughly classified into four categories: field shots, medium shots, close-up shots and audience shots. Samples of these four kinds of shots are shown in Fig. 1.

Field Shot Medium Shot Close-up Shot Audience Shot

Fig. 1 Classification of Shots in Soccer Videos

Field shots indicate shots which include large areas of the grassland in a soccer pitch. Most players and the formation of both sides are involved. These shots are broadly used when events of offense and defense occur. Field shots are characterized by its large ratio of SPC pixels. Accordingly, the percentage of SPC pixels could be calculated and by comparing this percentage with an empirical threshold, a field shot could be identified.

Medium shots reveal the interaction between certain players on both sides, such as dribbling, shooting, goal keeping and conflicts. Medium shots are relatively flexible, for they lack a fixed camera grammar: their background could vary from large numbers of audience to a predominant ratio of the pitch. However, the ratio of SPC pixels in medium shots is generally lower than that in field shots, but higher than that in other two types. The region of SPC pixels in key frames of field shots, medium shots, close-up shots and audience shots are shown in Fig. 2.

Field Shot Medium Shot Close-up Shot Audience Shot

Fig. 2 Ratio of SPC Pixels in Field Shot, Medium Shot, Close-up Shot and Audience Shot

In Fig. 2, the SPC pixels are set black, while the other pixels are set white. It is obvious that the ratio of pitch color pixels vary greatly in these four types of shots.

Audience shots always take place after a shooting or a verdict event. The attitude of the spectators infers the atmosphere of the stadium, and mark attracting events during a game. Audience shots describe large numbers of spectators as a whole, which are normally interpreted as large amount of edge information. Therefore, by analyzing edge information of a frame, the number of audience in would be intimated, and the purpose of identification is achieved.

Close-up shots provide detailed description of facial expressions or physical movements of players and coaches, implying their emotional reaction towards a certain event. Close-up shots are commonly dominated by a large ratio of skin pixels, while its edge information is commonly less than that of audience shots, as shown in Fig. 3. Thus the discrimination between audience shots and close-up shots could be achieved by edge information extraction.

Close-up Shot | Edge Detection Result of Close-up Shot | Audience Shot | Edge Detection Result of Audience Shot

Fig. 3 Edge Detection of Close-up Shot and Audience Shot

3 The Algorithm

In this section the algorithm of this identification system is elaborated. The first part of this section briefly introduces the procedure of the system; the second and the third parts describe the key algorithms applied in this system. The last part emphasizes on the innovative improvement for more precise results.

3.1 Flow Diagram

The flow diagram of this detection system is shown in Fig. 4.

Step One: Preprocessing. The main procedure in this step is to extract the key frame from a clip. A soccer game can last as long as two hours, giving rise to large redundancy in each scene. If extracting features from every frame of a scene directly, the efficiency of detection would be greatly reduced. During the preprocessing step, key frames are primarily extracted and are comprehensive enough to represent events in every scene.

Step Two: Preliminary Sifting. The second step uses color histograms to count the ratio of SPC pixels. As explained in Section 2, the ratio of SPC pixels in a field shot is the highest among all shot types, while this ratio is relatively low in close-up shots and audience shots. By comparing this ratio with a proper threshold, the identification of field shot could be achieved.

Step Three: Identification of Close-up Shots and Audience Shots. As shown by Fig. 3, audience shots always contain large amount of edge information, whereas close-up shots contain relatively less. After extracting edges in a frame with Sobel

operators and calculating the ratio of edge information, close-up shots and audience shots can be distinguished.

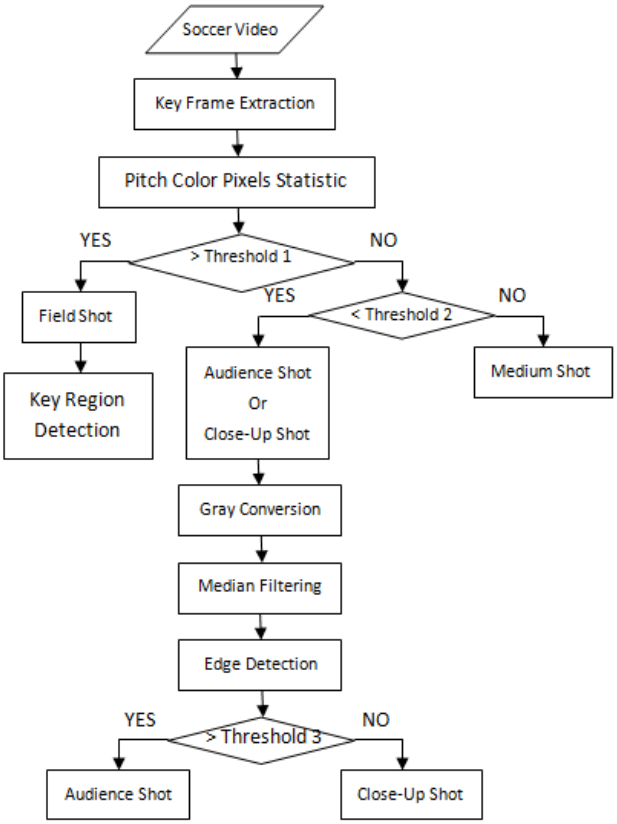

Fig. 4 Flow Diagram

3.2 Key Frame Extraction

Key frames are static images extracted from original videos. They embrace major contents of a scene, thus are commonly used for video processing. This system preliminary chooses the first frame of a shot as the key frame, then calculates the color histograms of every frame in the present shot. When the distance between a certain frame and the previous key frame is longer than the threshold, this present frame is regarded as the new key frame. The distance of histograms are calculated according to equation 1~4.

$$D(k, j) = \sqrt{D_R^2(k, j) + D_G^2(k, j) + D_B^2(k, j)} \quad (1)$$

$$D_R(k, j) = \sum_{i=0}^{255} \{R(j,i) - R(k,i)\} \quad (2)$$

$$D_G(k, j) = \sum_{i=0}^{255} \{G(j,i) - G(k,i)\} \quad (3)$$

$$D_B(k, j) = \sum_{i=0}^{255} \{B(j,i) - B(k,i)\} \quad (4)$$

D(k, j) is defined as the distance between pixels at position *(k, j)* in two frames. $D_R(k, j)$, $D_G(k, j)$, $D_B(k, j)$ separately represent the differences of pixel numbers with the same *R*, *G*, *B* values in tow frames.

3.3 Identification of Field Shots

As shown in Fig. 2, field shots are characterised by its high ratio of SPC pixels, while this ratio is pretty low in audience shots and close-up shots. Medium shots has a relatively high ratio of SPC pixels, whereas not as high as field shots. Accordingly, the identification of field shots is completed by calculating the ratio of SPC pixels in a chosen key frame.

The *R*, *G*, *B* values of a frame are easily obtainable, but relatively difficult to directly indicate human-perceivable colour information. This system first converts the *R*, *G*, *B* values into *H*, *S*, *V* values, where *H* represents the hue, *S* refers to the saturation, and *V* indicates the value. Then the system calculates the ratio of pixels whose H value is within the range of soccer pitch colour. Finally by comparing this ratio with Threshold 1, which could either be preset or be adjusted by users, field shots could be identified: when the ratio is large than the threshold, the shot from which this key frame is extracted could be regarded as field shots; if not, the second comparison with Threshold 2 is made.

When the ratio is smaller than threshold 1 but larger than Threshold 2, the shot could be regarded as medium shot. Otherwise, this shot represents either close-up shots or audience shots, and the frame would be further processed.

3.3 Identification of Close-up Shots and Audience Shots

Close-up shots and audience shots could be distinguished by edge detection. In order to obtain detection results of a higher accuracy, the process of gray conversion and median filtering are introduced. Gray conversion serves as the basis of the subsequent procedures, and median filtering helps to eliminate impulse noise caused by gray conversion.

Edge detection is applied by this system as the kernel of discrimination between close-up shots and audience shots. Commonly used operators of edge detection

include Sobel Operators, Kirsch operators, Roberts Operators and Laplacian Operators. Results of the key frame of an audience shot processed by four operators mentioned above are shown in Fig. 5.

Fig. 5 Edge Detection Implemented by Different Operators

As demonstrated in Fig. 5, images processed by Sobel Operators contain more evident edge information than those processed by other operators. Moreover, Sobel Operators are comparatively simple and efficient. This system utilizes the isotropic Sobel Operators, which enhances the accuracy of edge detection.

The method of binaryzation is employed to emphasize the edge information of the key frame. Pixels representing the edges are set white while other pixels are set black. By calculating the ratio of white pixels and comparing it with Threshold 3, audience shots and close-up shots could be identified: the ratio which is larger than Threshold 3 indicates the original shot to be an audience shot, while the ratio smaller than Threshold 3 refers to a close-up shot.

3.4 Improvement

Statistical data shows that medium shots and close-up shots tend to be classified as field shots by the system, when the backgrounds of these shots are the football pitch. Experimental results shows that the ratio of SPC pixels in a field shot ranges from 0.63 to 0.95, while this ratio in a close-up shot with the background of the pitch ranges from 0.59 to 0.74. The overlapping range cannot be distinguished by a single threshold.

In order to solve this problem, this system defines a “Key Region” of a key frame, which contains 35% of the whole image, as marked by a red rectangle in Fig. 6. The key region in a close-up shot always consist of large area of sport shirts and the ratio of pitch colour pixels is correspondingly low.

Fig. 6 Key Regions in Close-up Shots and Field Shots

Accordingly, when preliminary taking a shot as a field shot, this system further counts the SPC pixel ratio in the key region. When satisfying the condition of Threshold 4, the shot could be confirmed as a field shot.

4 Experimental Results

Two indexes are applied to measure the validity of this system: recall ratio and precision ratio, as defined in equation 5 and equation 6.

$$R_r = \frac{N_a}{N_a + N_m} \tag{5}$$

$$P_r = \frac{N_a}{N_a + N_w} \tag{6}$$

R_r and P_r signifies the recall ratio and the precision ration separately. N_a, N_m, N_w represent the number of shots which are accurately identified, missed, and inaccurately identified respectively.

The result of testing 195 shots segmented from a soccer video, which includes 60 field shots, 60 close-up shots, 60 audience shots and 15 medium shots, are shown in Table 1.

Table 1 Detection Results

Types	N_a	N_m	N_w	P_r	R_r
Field Shots	60	0	4	93.75%	100%
Audience Shots	60	0	2	96.77%	100%
Close-up Shots	57	3	0	100%	95%

The inaccuracy of field shots detection lies in the interference of medium shots with the background of the soccer pitch. However this fallibility has been strikingly reduced by Key Region Detection. The inaccuracy of audience shots results takes place when one frame contains the background of nets or advertising boards with abundant edge information. Despite of these, the system could process video clips in real time and efficiently detect different types of shots with a satisfactory accuracy.

5 Conclusions

This paper proposes an effective and real-time system to identify field shots, audience shots and close-up shots in a soccer video, which indicate attractive events of the game. The innovative improvement of Key Region Detection is employed to increase the accuracy of the system. Experimental results validate the efficiency of this system. With processing approaches of a high simplicity, the performance of this system is satisfactory.

Further work includes the detection of soccer nets and advertising boards to increase the anti-interference performance of the system.

References

1. D. Yow, et al. Analysis and presentation of soccer highlights from digital video. In Proceedings Asian Conference on Computer Vision, 1995.
2. S. Intille and A. Bobick. Recognizing planned, multi-person action. Computer Vision Image Understand, 2001, 81(3):414-445.
3. V. Tovinkere, et al. Detecting semantic events in soccer games: towards a complete solution. In proceedings IEEE Conference Multimedia Expo, 2001.
4. B.X. Li,et al. Bridging the semantic gap in sports video retrieval and summarization. J. Vis. Commun. Image R., 2004(15):393-424
5. Z. Xiong, R. Radhakrishnan, A. Divakaran and T.Huang, Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework, Proceedings of International Conference on Multimedia and Expo, 2003, Vol.3, pp.401-404
6. R. Radhakrishnan, Z. Xiong, A. Divakaran and Y. Ishikawa, Generation of Sports Highlights using a Combination of Supervised and Unsupervised techniques in the Audio Domain, Proceedings of the 2003 Joint Conference of the 4th International Conference on Information, Communications and Signal Processing and 4th Pacific Rim Conference Multimedia, 2003, Vol.2, pp.935-939
7. J.Q. Yu, M. Yang, Y.F. He, Highlights Extractions Method for Soccer Video Based on Exciting Degree, The 3rd Joint Conference on Harmonious Human Machine Environment, 2007.