Scene Classification for Baseball Videos Using Spatial and Temporal Features

Mao-Hsiung Hung, Chaur-Heh Hsieh, Ying-Chung Zhu

Dept. of Information Engineering, I-Shou University
Dahsu , Kaohsiung, 840, Taiwan, R.O.C.

Abstract

Correct classification of various kinds of scenes in sport videos is essential for higher-level content analysis such as event detection. The paper presents a novel technique for the classification of the typical scenes of baseball videos. The spatial color features are employed to detect pitch scene first. The temporal features derived from a shot are utilized to classify infield and outfield scenes in the second stage. Simulation results indicate that high accuracy (more than 90 %) of classification is achieved.

Keywords: scene classification, baseball video

1. Introduction

Recently, there have been increasing researches that attempt to achieve higher-level content analysis of sport videos. Correct classification (detection) of various kinds of scenes (views) in sport videos is essential for further content analysis such as event detection.

In [1] Zhong and Chang proposed an efficient multi-stage view recognition technique to identify canonical views such as pitch scene in baseball and serve scene in tennis. Pei and Chen [2] designed two heuristic hierarchies based on domain knowledge to classify semantic scenes in tennis video and baseball video respectively. In [3], three types of scenes of soccer games were classified according to grass ratio of video frames. In [4] an online learning framework was presented to detect pitch scene in baseball using a small number of pre-labeled training samples.

All the schemes mentioned above classify scenes using the spatial features of the keyframe extracted from a video shot. Our investigations indicate that the spatial features of a keyframe works well only for a shot with low motion activity such as low camera motion. For a shot with high motion activity, the spatial features of every frame in the shot changes a lot with time. As a result, only spatial feature of the keyframe is not enough to achieve good classification accuracy for these high motion scene shots. Sport videos often contain many fast changed shots caused by high camera motion. These shots correspond to some high speeded sport events, for examples, players running and a ball flying on the field. Thus, features over time called temporal features are important to classify the high motion scene shots. This paper presents a new method of scene classification for broadcasted baseball videos, which uses spatial and temporal feature to characterize three scene types including pitch, infield and outfield. Experimental results indicate that high accuracy of classification is achieved.

2. Scene Classification Using Spatial and Temporal Features

Three types of scenes are defined in this paper: pitch, infield and outfield, which are directly related to the events of in-play of baseball games. The pitch scene is generated when the camera focuses on the home plate region to capture pitching and battering; thus the camera motion is small. As a result, a keyframe is good to represent the all frames in the shot. On the contrary, the outfield and infield scenes are obtained by controlling cameras to follow the moving ball, and the actions of fielder and runners. Many camera motions, such as panning and zooming, happen in a scene shot. Therefore, the spatial features of a keyframe are not enough to characterize the two scenes, and thus causing poor performance in the classification. In summary, spatial features of a frame are suitable to detect pitch scenes, whereas temporal features should be included in the detection of outfield and infield scenes. In this work, *k-means* clustering scheme that employs color features of a frame is used to detect the pitch scene first. Then the remaining scenes are classified into infield or outfield scenes using a SVM classifier, which utilizes the temporal features of a shot.

2.1 Pitch Scene Detection

Grass and soil colors are two dominant colors in a baseball field. Assume that the two colors are Gaussian distribution with their respective parameters means and standard deviations. We collect grass and soil pixels as training data, and estimate mean μ and standard deviation σ of each color. Then we use the range μ–2σ to μ+2σ to detect the grass and soil pixels respectively. The ranges for grass and soil colors obtained experimentally are shown as Table 1.

Table1 Ranges of grass color and soil color

	Grass color	Soil color
Y	55~118	64~127
Cb	96~117	108~129
Cr	119~132	126~153

To explore the local characteristics of a video frame, we divide the whole frame into 16 blocks. The upper 4 blocks are discarded because of lack of field colors, as shown in Fig.1. Each block is encoded with 3-bits code as defined in Eq. (1), and explained as follows. If the number of grass pixels of a block is greater than 30 % of the total number of pixels of the block, b_0 is set to 1, otherwise set to 0. The other two bits used to represent soil color and other color (neither grass nor soil) are obtained in a similar way, respectively.

$$b_0 = \begin{cases} 1 & \text{if grass\% > 30\%,} \\ 0 & \text{otherwise} \end{cases} \quad b_1 = \begin{cases} 1 & \text{if soil\% > 30\%,} \\ 0 & \text{otherwise} \end{cases}$$
$$b_2 = \begin{cases} 1 & \text{if other\% > 30\%,} \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

The 3-bit code characterizes the color distribution of each block, and thus is called feature code. Table 2 shows the meaning of the feature codes. The feature codes of all blocks in a video frame concatenate a 12-D feature vector, which describes the color feature of the frame. For example, the video frame in Fig. 1(a) is characterized by the feature vector [5 3 3 5 6 7 4 2 3 5 3 1], as shown in Fig.1(c).

Table 2 Three-bit feature codes and their meaning

Code, $b_2b_1b_0$	Meaning
000 (0)	Not allowed
001 (1)	grass pixels dominate
010 (2)	soil pixels dominate
011 (3)	Mixed grass and soil pixels
100 (4)	Other pixels dominate
101 (5)	Mixed grass and other pixels
110 (6)	Mixed soil and other pixels
111 (7)	Mixed grass, soil and other pixels

In this work, the pitch scene is detected by the above feature vector of a frame. More specifically, in training phase, we collect the frames of pitch scene shots, and extract their feature vectors using the above procedure. Then a feature codebook is generated by applying all the feature vectors to *k*-means clustering, as shown in Table 3. In testing, the feature vector of each frame of a shot is matched with the codevectors of the codebook. If the distance of the best match is less than a threshold, the frame is regarded as pitch one. If the percentage of pitch frames in a shot is greater than 20 %, the scene shot is regarded as pitch; otherwise, it is infield or outfield, and the following classification scheme is used to further identify the two scenes.

Table 3 feature codebook generated with *k-means* algorithm

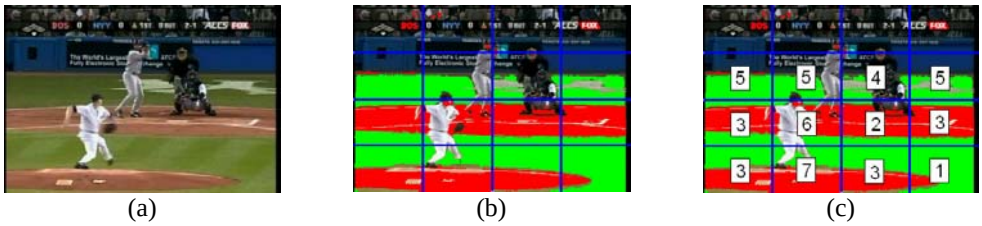
[5 3 3 4 6 3 4 2 3 4 3 1]
[5 3 3 5 3 3 4 3 1 5 3 1]
[5 3 3 5 3 3 4 6 3 4 3 1]

2.2 Infield and Outfield Classification

To investigate the feature vector varies over time, the hamming distances from an input feature vector (v) to three predefined vectors that represent all-grass, all-soil and all-other colors are considered. The predefined vectors and the hamming distances are defined in Eq.2. The three hamming distances of $dist_1$, $dist_2$ and $dist_3$ represent the similarity of an input feature vector v and $v_{\text{all-grass}}$, $v_{\text{all-soil}}$ and $v_{\text{all-soil}}$, respectively.

$$\begin{aligned} dist_1 &= \text{hamming dist}(v, v_{\text{all-grass}}), \\ dist_2 &= \text{hamming dist}(v, v_{\text{all-soil}}), \\ dist_3 &= \text{hamming dist}(v, v_{\text{all-other}}) \end{aligned} \qquad (2)$$

where $v_{\text{all-grass}}$=[1 1 1 1 1 1 1 1 1 1 1 1], $v_{\text{all-soil}}$=[2 2 2 2 2 2 2 2 2 2 2 2], and $v_{\text{all-other}}$=[4 4 4 4 4 4 4 4 4 4 4 4].



(a) (b) (c)

Fig.1 A set of feature codes extracted from a pitch image

For a video sequence, we calculate the three hamming distances for each frame, and consider the results as three signals which are function of time. Fig. 3 is an example which is obtained from a test video containing pitch, infield and outfield scenes. It can be seen from this figure that the same scene present similar signal segment. For example, the two green boxes of infield scenes and two orange boxes of outfield scenes in Fig.2. Therefore, the features of the three signals can be explored to classify infield and outfield scenes.

We use a quadratic equation of $y=ax^2+bx+c$ to approximate a signal segment. Then, each signal segment is replaced by a polynomial curve with particular steepness and tendency. To evaluate the similarity between two polynomial curves, we map a polynomial curve to a point in a polar coordinate system. We first equally divide the polar coordinate into forty pies, and then we obtain 40 radial lines and their angles (0°, 9°, 18°,…, and 360°). Then, we allocate 0°~180° (upper half circle) to polynomial curves where maximum exists, and 180°~360° (lower half circle) to polynomial curves where minimum exists. In the upper half circle, the ratio of resampling points of positive slope to that of negative slope is considered. We assign the ratios of 20:0, 19:1, 18:1,…, 0:20 to 0°, 9°, 18°,…,180° respectively. Similarly, in the lower half circle, the ratios of resampling points negative slope to that of positive slope are 20:0, 19:1, 18:1,…, 0:20. These ratios are assigned to 180°, 189°, 198°,…,360°. As a result, the monotonous increasing and decreasing curves are located at 0° and 180° respectively. The convex and concave curves which are completely symmetric along the vertical axis are located at 90° and 270°. So all polynomial curves can be assigned an angle (θ) according to the above statements, as shown in Fig.3. Regarding the radius (r), we use the mean of absolute value of slopes on the twenty resampling points. As a result, curves with small slope or smooth curves are mapped to the points near the origin of the polar coordinate; whereas curves with large slope are mapped to the places far away the origin.

As mentioned above, each polynomial curve is mapped to one point in the coordinate. Because a scene shot is described by three curve segments, it can be equivalently characterized with three points in the polar coordinate system through the above mapping. The three points form a new 6-D feature vector, denoted by [v_1, v_2, v_3, v_4, v_5, v_6]. In this work, a 2-classs SVM classifier is employed to classify infield and outfield scenes using the feature vector.

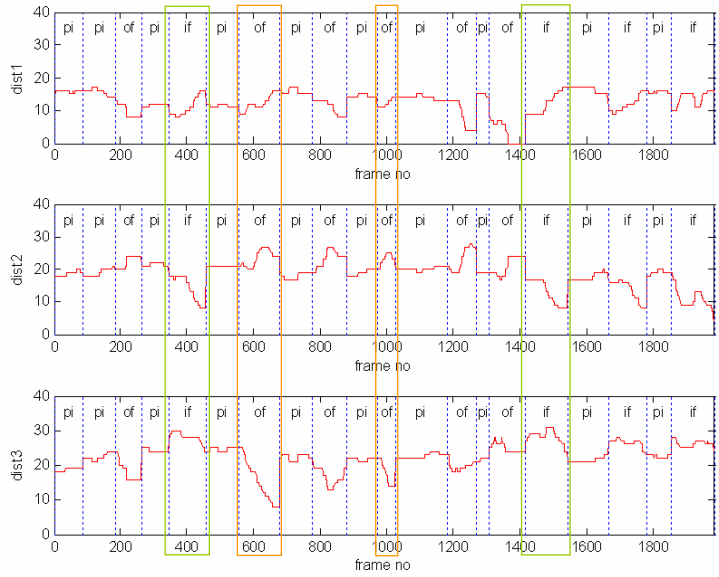


Fig.2 Three hamming distance curves (pi: pitch, if: infield, of: outfield)

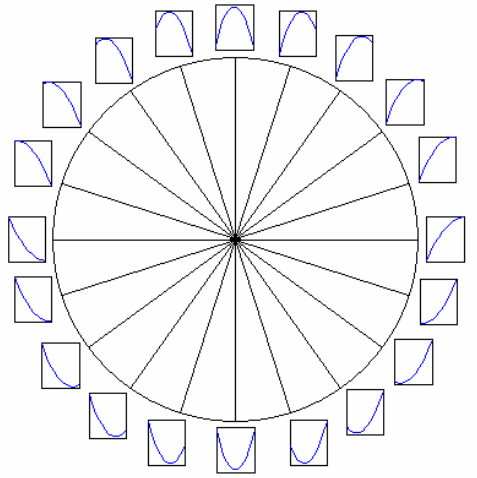


Fig.3 A polar coordinate for polynomial curve mapping

3. Simulation Results

Our classification method contains two stages: (a) detection of pitch scene, and (b) classification of infield and outfield. For the first stage, 316 shots containing 148 pitch and 168 non-pitch scenes are tested. This stage obtains 100% accuracy of detection. .

For the second stage, we collect 169 shots of infield and outfield scenes from three games, denoted by G1, G2 and G3. We choose one game to train SVM classifier, and test the other two games. The classification results of training and testing are listed in Table4. The training results (diagonal cells) achieve above than 97.9% accuracy. The testing results (non-diagonal cells) achieve from 89.55% to 95.08% accuracy.

Table 4 Simulation results of three games

Training\Testing	Game 1	Game 2	Game 3
Game 1	100% (61/61)	93.75% (45/48)	91.67% (55/60)
Game 2	95.08% (58/61)	97.9% (47/48)	91.67% (55/60)
Game 3	93.44% (57/61)	89.58% (43/48)	98.33 (59/60)

4. Conclusion

The paper has presented a new scene classification technique for baseball videos based on spatial and temporal features. Three event-related scenes including pitch, infield and outfield are classified in two stages. Pitch scene is first detected using color feature of each frame of a shot, which achieves 100% accuracy. Then temporal features over a shot are employed to classify infield and outfield scenes in the second stage. This stage obtains more than 90% of average accuracy. The extension of the proposed method to other sport videos will be investigated in the future.

Acknowledgements

This work was supported in part by National Science Counsel Granted NSC 94-2213-E-214-010 and NSC 95-2213-E-214-051, and I-Shou University Granted ISU-94-01-18.

5. References

[1] D. Zhong, S.-F. Chang, “Real-time view recognition and event detection for Sports video,” J. Vis Commun. Image R. 15(2004) 330-347.

[2] S.-C. Pei, F. Chen, “Semantic scenes detection and classification in sports videos,” CVGIP 2003.

[3] P. Xu, L. Xie, S.-F. Chang, A. Divakaran, A. Vetro, H. Sun, Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video, ICME, Tokyo, Aug. 22-25, 2001.

[4] Jun Wu, Xian-Sheng Hua, Jian-Min Li, Bo Zhang, Hong-Jiang Zhang, An Online Learning Framework for Sports Video View Classification, The Fifth Pacific-Rim Conference on Multimedia (PCM 2004), November 30-December 3, Tokyo, Japan, 2004.