Electronic Letters on Computer Vision and Image Analysis 7(1):16-25, 2008

Shot Classification in Broadcast Soccer Video

Xiaofeng Tong, Qingshan Liu, Hanqing Lu

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
P.O.Box 2728, Beijing China 100080
xftong@gmail.com, {qsliu, luhq}@nlpr.ia.ac.cn

Received 31 May 2007; revised 12 September 2007; accepted 14 March 2008

Abstract

In this paper, we present an effective hierarchical shot classification scheme for broadcast soccer video. We first partition a video into replay and non-replay shots with replay logo detection. Then, non-replay shots are further classified into Long, Medium, Close-up or Out-field types with color and texture features based on a decision tree. We tested the method on real broadcast FIFA soccer videos, and the experimental results demonstrate its effectiveness..

Key Words: Shot classification, Replay detection, Sports video.

1 Introduction

In recent years, sports video analysis has received increasing interests due to its tremendous commercial. A popular scheme of sports video analysis has three steps: 1) parsing the video into shots; 2) classifying shots into several types with certain semantics; 3) inferring high-level events with shot transition context. In this paper, we focus on shot classification of soccer video.

A lot of work has been done on shot classification in soccer video. The grass-ratio and non-field area distribution are often used as important features in most previous work [1][2][3][12][13] to discriminate the shots of Long, Medium, and Close-up view. In [4], they used the moments of colour and shape as features to identify shot types, and modelled the shot temporal transition pattern with a Hidden Markov Model (HMM) to detect semantic events. Duan *et al* [5] extracted shape and motion information to categorize shots with Support Vector Machine (SVM) classifiers. Replay as a special scene of sports video, also plays an importance role in semantic event analysis. H. Pan *et al* [9] utilized an HMM to infer replay scenes, in which a zero-crossing measure was considered for the frequency and the amplitude of the fluctuations of adjacent frame differences. However, a single frame inside a replay must be pinpointed in advance. Later, they proposed another replay detection method based on replay-logo [14]. Kobla *et al* [15, 16] used the macro-block, motion and bit-rate information in compressed domain to detect replay scenes. Duan *et al* [17] developed a logo transition detection, in which the mean shift algorithm was used to seek the mood of logo transition, and then a replay scene was determined by a pair of logo-transitions. Wang *et al* [18] utilized

Correspondence to: qsliu@nlpr.ia.ac.cn

Recommended for acceptance by Xuelong Li

ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

context information of the concurrence of replay and other shots types to detect the replay scenes. The above replay detection methods can be categorized into two classes: replay-logo pairing and scene context modelling. However, the former scheme needs a reliable replay-logo (given or automatically learned); while the latter is not robust because the motion pattern is insufficient to intrinsically characterize replay.

In this paper, we present an effective shot classification framework of soccer video. Figure 2 illustrates the proposed hierarchical framework. Replay scene is a special scene that re-plays an interesting scene with slow motion pattern from different views. Though a replay scene may contain several shots, it is often used like a single shot for event detection, so we call it replay shot. We first learn the replay-logo by gradual logo transition and unsupervised clustering. Then we detect the replay shots based on the property that a replay shot is often sandwiched by two replay-logo transitions at the start and end points. We further categorize non-replay shots into the Long, Medium, Close-up, and Out-field types with intrinsic colour and texture features, and the Close-up shots are classified into two sub-types: close-up with field background (*CloseFB*) and close-up with non-field background (*CloseNFB*). Figure 1 shows some examples.

The rest paper is organized as follows: View representation, feature extraction, and classifier design are presented in Section 2. Replay shot detection is given in Section 3. Section 4 reports experiments. Conclusions are finally drawn in Section 5.

(a) Replay (b) Long (c) Medium

(d) CloseFB (e) CloseNFB (f) OutField

Figure 1: Examples of predefined shot types

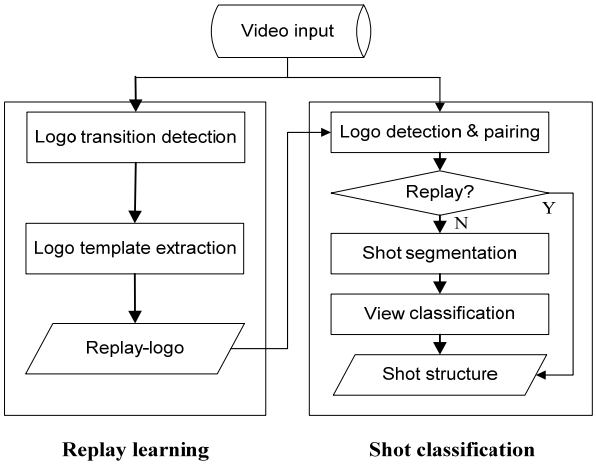


Figure 2: System framework

2 View Representation and Classification

In sports video, almost all the shots are cut except replay transition, so for non-replay shots, we first use a two-threshold colour histogram method as in [19] to detect shot boundaries, and then classify them into the defined classes.

2.1 Predefining and Representing Views

We predefine five main types of shots: Replay, Long, Medium, Close-up (including *CloseFB* and *CloseNFB*) and Out-field shots (Figure 1). Their characteristics are described below:

- ***Replay*** shot: Replay is usually utilized to play back an interesting or important segment with a slow-motion pattern. It is sandwiched by two logo transitions at the start and end points. The logo appears and disappears gradually during the transition.

- ***Long*** shot: A long shot displays a global view of the game field as shown in Figure 1(b). It is captured by a camera at a long distance. It is often used to show the play status, such as, team line-up, play position, and long pass. In a long view, the ration of the field area to the whole image is high, and the size of objects (players) within the field is small.

- ***Medium*** shot: A medium shot is a zoom-in view of a specific part of the field, as in Figure 1(c). It is usually used to follow players who are competing intensely or show a scene before a placed kick. In a medium view, the size of objects in the play-field is bigger than that in a long view, but smaller than that in a close-up. In another point, the field ratio is also medium.

- ***Close-up*** shot: A close-up shot usually shows above-waist view, especially the face expression of one person, and it often aims at a focal character who is the leading actor of current event. There are two sub-types of close-up shots: close-up with field background (*CloseFB*) and with non-field background (*CloseNFB*), as shown in Figure 1(d) and 1(e) respectively. The former has a higher field ratio than the latter. Additionally the *CloseNFB* often has complex texture because its background is audience.

- ***Out-field*** shot: An out-field shot displays the audience or other persons out of the game field. Figure 1(f) shows an example. It indicates a break caused by highlight, such as audience cheer view after a goal. In a out-field shot, the field ratio is small, but the texture is complicated.

2.2 Designing Classifier

Based on above description, we construct a decision tree [8] to classify the shots with visual features. The tree structure is shown in Figure 3. At level 0, the video is partitioned into the replay shots and the non-replay segments by replay-logo detection. Then the non-replay segments are divided into the *Long*, *Medium*, *Close-up*, and *Out-field* types. We first label each frame in a shot, and then assign the shot type by the voting scheme. At level 1, the field-ratio (*Field*) is used to discriminate the {*CloseFB* ∪ *Long* ∪ *Medium*} views and the {*Out-field* ∪ *CloseNFB*} views. At the left branch of level 2, a head area under some constraints is applied to separate the *CloseFB* and the {*Medium* ∪ *Long*} views. At the right branch, a texture feature of the gray level co-occurrence matrix (GLCM) is utilized to distinguish the *Out-field* and the *CloseNFB* views. At level 3, the scale of objects in the field (*Scale*) is used to distinguish the *Medium* views from the *Long* views.

Both the *CloseFB* and the *CloseNFB* belong to the *Close-up* views, while they have different characterization. A *CloseFB* has simple background (grass) and a distinct big head. A *CloseNFB* has complex texture with a low grass-ratio, and the detection of big head is not robust due to noise (skin area) in its background. Therefore, we detect a big head to identify the *CloseFB*, and analyze the texture complexity to verify the *CloseNFB*.

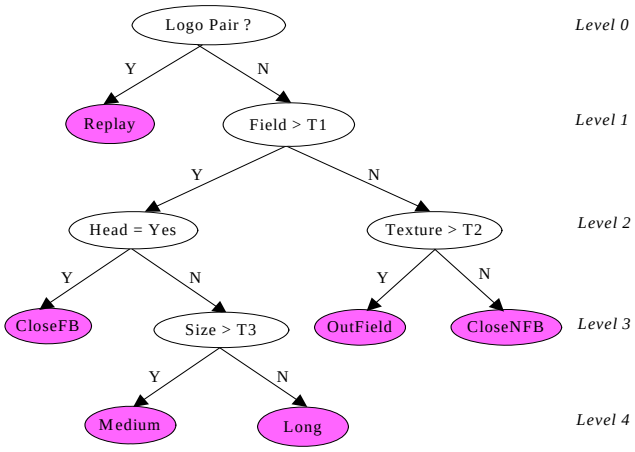


Figure 3: Decision tree for shot classification

2.3 Feature Extraction

2.3.1 Field-ratio

Though the dominant colour of the soccer field is green, there exist some variations in different stadium and under different illumination. We analyze the colour histograms of hue and saturation in HSV colour space based on the beginning part of the video, and take account of a slight range centred the peak of histogram as the dominant colour of the field. Denote the means of the hue and the saturation of the field as H_{mean} and S_{mean}, respectively. The distance from a pixel $f(i, j)$ to the field dominant colour is measured by a cylinder metric.

$$\theta = \left| H(i, j) - H_{mean} \right|, \tag{1}$$

$$d_{hsv}(i, j) = \sqrt{S^2(i, j) + S_{mean}^2 - 2 \cdot S(i, j) \cdot S_{mean} \cdot \cos(\theta)}\,, \tag{2}$$

where $H(i, j)$ and $S(i, j)$ are the hue and the saturation components respectively. If $d_{hsv}(i,j) < T$ (T is a threshold), the pixel $f(i, j)$ is classified into the field. The area ratio of the field and the whole image is taken as the field-ratio. We call the field-ratio Field for simplicity.

After image segmentation by the dominant color, morphological operation and connect-component analysis are performed to smooth and get the field region. Figure 4 shows an example, in which a frame with shadow is well segmented.

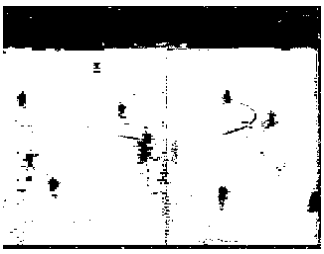
(a)  (b)  (c) 

Figure 4: Field segmentation. (a) Original image, (b) Segmented image and (c) extracted field region

2.3.2 Texture

The gray-level co-occurrence matrix (GLCM) is used for texture description due to its capability in classifying stochastic textures. We use the GLCM contrast as the texture feature in this paper.

2.3.3 Head area

Each *Close-up* view has a big head. Our head area detection method is based on skin color and connect-component analysis.

1) Skin detection

We use a multi-variable Gaussian model [10] for skin detection. The probability of a given pixel lies in skin distribution is

$$p(x \mid skin) = \frac{1}{2\pi \left\| \Sigma_s \right\|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_s)^T \Sigma_s^{-1}(x-\mu_s) \right] \tag{3}$$

where $x = [r, g]^T$ denotes the chrominance vector in RGB colour space, $r = R/(R+G+B)$, $g = G/(R+G+B)$, μ_s and Σ_s represent the mean vector and the covariance matrix of the training pixels respectively. A similar model is developed for non-skin pixels $p(x \mid \neg skin)$. We use the Bayesian rule to determine whether or not a pixel belongs to the skin area as:

$$\begin{aligned} p(skin \mid x) &= \frac{p(x \mid skin)p(skin)}{p(x \mid skin)p(skin) + p(x \mid \neg skin)p(\neg skin)} \\ &= \frac{p(x \mid skin)}{p(x \mid skin) + p(x \mid \neg skin)} \quad \propto p(x \mid skin) \end{aligned} \tag{4}$$

If $p(skin \mid x) > T_s$, it is regarded as a skin pixel.

2) Head area verification

Connect-component analysis on the segmented skin image is used to verify the head area. A qualified head area in a Close-up view must satisfy the constraints: (a) scale (height of detected area) should exceed a threshold; (b) solidity (ratio of the area to the min-max box) should be higher; (c) orientation (direction of the major axis) should be within a certain range. Figure 5 shows an example.

Figure 5: Examples of Head area detection. (a) Original image; (b) skin detection; (c) head area (surrounded by a red ellipse)

2.3.4 Object size

Size of the objects in the field is used to discriminate the Long and the Medium views, because it approximately expresses the distance from the camera to the objects. The object scale estimation is composed by three steps: (1) field segmentation; (2) convex contour extraction of the field, which involves region filling, neighbor regions connection, and convex contour tracking [11]; (3) object segmentation and scale estimation. Two examples are displayed in Figure 6.

Assuming the average height of the objects is H_o, and the height of the field in an image is H_f. The size of the objects is defined as H_o / H_f.

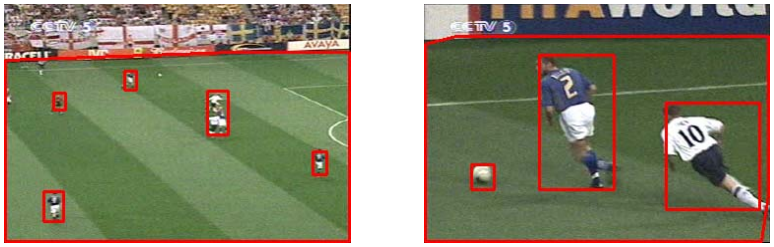


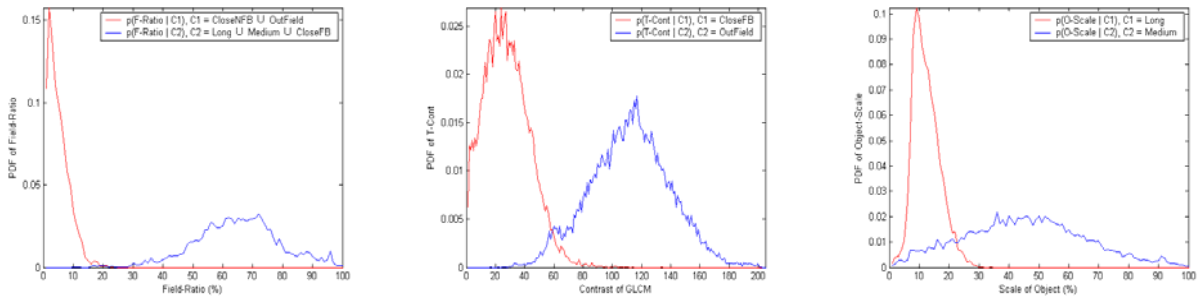
Figure 6: Contours of field and objects (surrounded by red polygon)

2.4 Views Discrimination Properties

In order to obtain proper discriminative thresholds, we test the method on a large quantity of images and get the statistical properties of the features vs. view classes. According to the Bayesian theorem, it can be deduced through the prior probability and the class conditional probability density. To determine whether a view belongs to class ω_1 or ω_2, we compare their posterior probabilities under the observation o:

$$\frac{p(\omega_1|o)}{p(\omega_2|o)} = \frac{p(o|\omega_1)p(\omega_1)/p(o)}{p(o|\omega_2)p(\omega_2)/p(o)} = \frac{p(o|\omega_1)p(\omega_1)}{p(o|\omega_2)p(\omega_2)} \tag{5}$$

The prior probabilities, $p(\omega_1)$ and $p(\omega_2)$, are estimated from the training data. The class conditional probabilities, $p(o|\omega_1)$ and $p(o|\omega_2)$, are obtained by statistics on features (grass-ratio, GLCM contrast, big head and object-scale). Figure 7 shows the conditional probability densities with the observations of the Field, the *T_Cont*, and the *O_Scale* respectively.



(a) (b) (c)

Figure 7: Features Statistics over different classes. (a) F-Ratio for {CloseNFB ∪ Out-field} (red) and {Long ∪ Medium ∪ CloseFB} (blue). (b) T-Cont for CloseNFB (red) and OutField (blue). (c) O-Scale for Long (red) and Medium (blue).

3 Replay Detection

There are logo transitions both at the start and the end of the replay segments. During the transition, a highlighted logo wipes in from one side and out from another side of TV screen. It typical lasts 0.5 - 0.8 seconds or 15-24 frames (Figure 8). We try to learn the logo and use it to detect other replays in the same video. We first detect the logo transitions, and extract the logo template. Then we detect the other logos with template matching. Finally we determine the replays with logo pairing.

Figure 8: A logo transition (5 of 24 frames are displayed)

3.1 Automatic Detection of Logo Transition

Logo transition detection is performed on the differences between the frames, which are characterized by the mean square difference of intensity (*MSD*). In order to smooth the differences between the frames, we carry out a median filtering on them. Figure 9 (a) and (b) illustrates an original *MSD* sequence and a processed sequence after median filtering respectively. Logo transition detection is performed through plateau-like pattern detection over the *MSDs*. The detailed procedure is described below [7]:

1) Compute the differences between the frames with the *MSD*, and conduct the median filtering operation.

2) Check the difference. If the difference exceeds a threshold T_l, go to step 3), otherwise go back to step 1) for the next one.

3) Count the number C_d of consecutive frame-differences that exceed the threshold T_l . If the counter C_d exceeds a certain threshold T_n, a logo-transition is determined. Otherwise go back to step 1) for the next frame.

In a logo-transition, there is one frame that contains the clearest and the most complete logo. The logo is highlighted and located at the middle part of the frame, and it is vertically symmetrical. With this prior information, we take the frame with the minimum pixel-wise intensity difference of vertical symmetry as the replay-logo candidate frame.

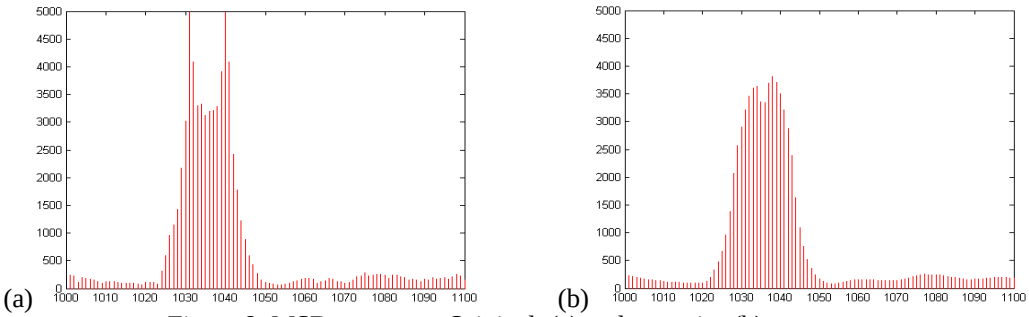


Figure 9: MSD sequence. Original (a) and posterior (b) sequence

3.2 Extraction of Logo Template

Because the logo is superimposed upon a frame, a logo candidate contains both the logo and the background information. In order to eliminate the background, we extract n logo candidates and take their average image as the logo template. The detailed procedure is:

1) Extract n logo candidates from n logo transitions, $f_1, f_2, ..., f_n$. All of them are taken account of a cluster.

2) Compute the cluster centre f_c ,

$$c = \arg\min_i \{\sum_j d(f_i, f_j)\}, \quad i = 1, 2, ..., n \tag{6}$$

where $d(f_i, f_j)$ is the distance (dissimilarity) between f_i and f_j .

3) Take the average of those candidates near to the centre f_c as the logo template $LT()$

$$LT(m,n) = \frac{1}{K}\sum_{k=1}^{K} f_k(m,n), \quad d(f_c, f_k) < \tau$$
$$m = 0,1,...,M-1; \quad n = 0,1,...,N-1 \tag{7}$$

Figure 10 shows two examples. On is from the World Cup of FIFA 2002 soccer video, and another is from an Olympic game video.

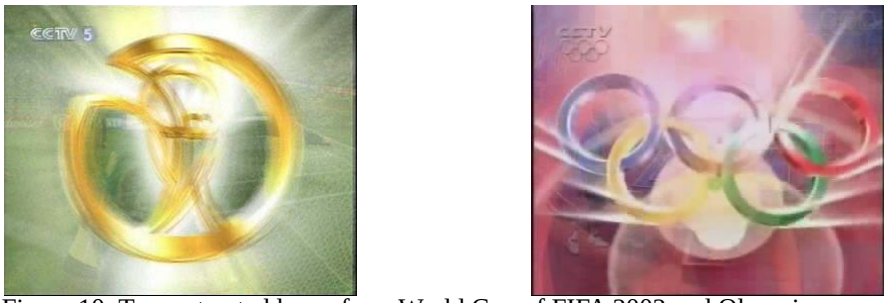


Figure 10: Two extracted logos from World Cup of FIFA 2002 and Olympic games

3.3 Detection of Replay by Logo Template

With the logo template is extracted, we detect the logos by image matching with color and shape features. During the process, we use a sliding local window to examine the 2*w+1 successive frame differences. Let $\{f_i, i=1,2,\cdots,N\}$ be the sequence of frames, $D_i = d(f_{LT}, f_i)$ is the distance between the logo template and the frame f_i. $\{D_i, i=1,2,\cdots,N-1\}$ is the frame difference sequence. We declare a detected logo at the frame f_l if

1) The difference D_l is the minimum within a symmetric sliding local windows of size 2*w+1, i.e., $D_l <= D_j$, $j = l-w,\cdots,l-1,l,l+1,\cdots,l+w$, and

2) D_l is the smaller than a fix threshold, and

3) The difference of invariant moments between f_l and the logo template is small.

Two adjacent logos compose a pair. A scene sandwiched by a logo pair is deemed to be a replay if its length is smaller than a predefined threshold (this threshold is got by offline statistics).

4 Experiments

In the experiments, we first evaluate the performance of shot classification on frames, and then we investigate it on shots. The first experiment is to evaluate the visual feature extraction and discrimination. The second is to measure the real performance on shots, including replay-logo learning and detection.

4.1 View Classification Results

The testing data includes 6000 *Long* views, 2000 *Medium* views, 2000 *CloseFB* views, 2000 *CloseNFB* views, and 2000 *Out-field* views. These frames are captured from real broadcast game videos of the FIFA 2002 world cup. The confusion matrix is reported in Table 1. From the result, we can see that the *CloseNFB* has the best performance. The reasons are three folds: 1) this type of view is simple; 2) it has obvious discriminative appearance; 3) the features are easy to be computed. Most of the false classification of this type are classified into the *Out-field*. It indicates that the field segmentation and the field-ratio feature are robust than the texture feature. The second best item is the *Long* views, and the missing instances are all fall into the *Medium* type. The error is due to the threshold decision. Correspondingly, there are some *Medium* views are mistakenly regarded as the *Long* views. The performance of the *Medium* shots detection is the lowest. The *Medium* type is located at the last level of the decision tree, and there has a little accumulated errors. In addition, the object scale is not robust to discriminate the *Long* view and the *Medium* view, for it is hard to well segment the objects.

	Long	Medium	CloseFB	CloseNFB	Out-field
Long	0.9203	0.0797	0.0	0.0	0.0
Medium	0.1445	0.8485	0.0060	0.0	0.0010
CloseFB	0.0	0.0865	0.9135	0.0	0.0
CloseNFB	0.0010	0.0	0.0065	0.9210	0.0710
Out-field	0.0360	0.0255	0.0	0.0405	0.8980

TABLE 1: Result of Views Classification

4.2 Shot Classification Results

This group experiment is conducted on two complete soccer games, Cameroon vs. Germany (C_G) and England vs. Sweden (E_S) in the FIFA World Cup 2002. The ground truth is elaborately labeled manually. The performance is evaluated by Precision and Recall, and the experimental results are shown in Table 2.

From the results, there is no false alarm but a little missing for the replay detection. The missing comes forth due to: 1) the extracted logo is the mean of several candidates, so it has a blur background (see Figure 10). 2) We use the whole image to measure the similarity between the candidates and the replay-logo, which degrades the real similarity. The *Long* shot type has the second best performance, and the performance of the

Medium type has the lowest detection rate. The most false alarms of the *Out-field* are from the *CloseNFB*. Most false items in the *Medium* actually come from the *Long* and the *CloseFB* shots, which are due to inaccurate object size estimation and big head detection.

Game	Shot Type	Total	Correct	False	Prec.(%)	Recall (%)
C-G	Replay	41	40	0	100	97.6
	Long	294	278	11	96.2	95.6
	Medium	128	107	23	82.3	83.6
	Close-up	190	181	10	94.8	95.3
	Out-field	9	9	2	81.8	100.0
E-S	Replay	45	42	0	100.0	93.3
	Long	295	276	6	97.9	93.6
	Medium	139	127	35	78.4	91.4
	Close-up	216	196	4	98.0	90.7
	Out-field	15	15	5	75.0	100

TABLE 2. Result of Shot Classification

5 Conclusions

In this paper, we presented an effective shot classification method for soccer video. First, we identified the replay shots by automatic replay-logo learning, detection, and pairing. Then, for non-replay shots, we categorized them into the *Long*, *Medium*, *Close-up*, and *Out-field* shots through intrinsic features discrimination. In the future, we will consider shot transition context and couple with HMM to further enhance the performance of classification. We will also mine semantic events based on shot classification.

References

[1] P. Xu, L. Xie, S.F.Chang, A. Divakaran, A.Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," *Proc. IEEE Int'l Conf. On Multimedia & Expo*. 2001, Tokyo, Japan, Aug 22-25, 2001.

[2] A. Ekin, A. Tekalp, "Automatic soccer video analysis and summarization", *SPIE Storage and Retrieval for Media Database*, California, USA, Jan. 2003.

[3] S.Chen, M.Shyu, C.Zhang, L.Luo, and M.Chen, "Detection of Soccer goal shots using joint multimedia features and Classification Rules," *ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining 2003 Workshop*, Wednesday, Aug. 27, 2003.

[4] R. Dahyot, N. Rea, and A.C. Kokaram, "Sport Video Shot Segmentation and Classification," *Visual Communications and Image Processing 2003*, Lugano, Switzerland, July 8-11, 2003.

[5] L. Duan, M. Xu, T. Chua, Q. Tian and C. Xu, "A mid-level representation framework for semantic sports video analysis", *Proc. ACM Multimedia 2003*, Berkeley, USA, Nov 2-8, 2003.

[6] X. Yu, L. Duan, Q. Tian, "Shot classification of Sports Video Based on Features in Motion Vector Field", *The 3rd IEEE Pacific-Rim Conf. on Multimedia Proceedings*, Dec. 16-18, 2002, pp. 253-260,

[7] X. Tong, H. Lu, Q. Liu, "A Three-Layer Event Detection Framework and Its Application in Soccer Video", *Proc. of IEEE Int'l Conf. On Multimedia & Expo 2004*, Taiwan, June 27-30, 2004.

[8] Richard Duda, Peter. Hart, David Stork, "Pattern Classification", John Wiley & Sons, New York, Oct. 2000, pp.394-411.

[9] H.Pan, P.Beek, and M.Sezan, Detection of slow motion replay segments in sports video for highlights generation, *Proc. of IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May 2001, pp. 1649-1652.

[10] J. Terrillo, M. Shirazi, H. Fukamachi and S. Akamatsu, "Comperative Performance of Different Skin Chrominance Models and Chrominance Spaces for the Automatic Detection of Human Faces in Color Images", *Proc. of IEEE Int'l Conf..on Automatic Face and Gesture Recognition*, France, March, 2000.

[11] http://www.intel.com/research/mrl/research/opencv/

[12] L. Wang, M. Lew, and G. Xu, "Offense based temporal segmentation for event detection in soccer video", *Workshop on Multimedia Information Retrieval (MIR)*, New York, USA, Oct, 2004.

[13] J. Assfalg, M. Bertini, A. D. Bimbo, W. Nunziati, and P. Pala, "Soccer highlights detection and recognition using HMMs", *Proc. IEEE Int'l Conf. On Multimedia & Expo, 2002*.

[14] H. Pan, B. Li, and M. Sezan, "Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, May 13 - 17, 2002, pp. 3385-3388.

[15] V. Kobla, and D. Doermann, "Detection of slow-motion replays for identify sports videos," *Proc. of the IEEE Third Workshop on Multimedia Signal Processing*, 1999, pp. 135-140

[16] V. Kobla, D. DeMenthon, and D. Doermann, "Identification of sports videos using replay, text, and camera motion features," *Proc. of the SPIE Conference on Storage and Retrieval for Media Database*, Vol. 3972, Jan, 2000, pp. 332-343.

[17] L. Duan, M. Xu, Q. Tian, C. Xu, "Mean shift based video segment representation and applications to replay detection," *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp.709-712.

[18] J. Wang, E. Chng, C. Xu, "Soccer replay detection using scene transition structure analysis", *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005, pp. 433-437.

[19] H.J Zhang, A. Kankanhalli, and S. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, Vol. 1, No. 1, pp. 10-28, 1993.