Enhanced Sports Video Shot Boundary Detection Based on Middle Level Features and a Unified Model

Bo Han, Yichuan Hu, Guijin Wang, Weiguo Wu, and Takayuki Yoshigahara

Abstract — *The performance of shot boundary detection algorithms using low level features can hardly fulfill the requirement of automatic sports video analysis, which is a promising module in Personal Video Recorders. Two kinds of middle level features are proposed in this paper to effectively enhance the shot boundary detection. One kind of the features is extracted from the projection of the Dominant Color mask; the other kind is extracted from the reliable block-based Motion Vectors. These novel features, together with the Region Color Histograms feature, are integrated into a unified model which employs Support Vector Machines to detect both cuts and gradual transitions. Experiments on diversified soccer video sequences demonstrate that our scheme outperforms the existing algorithms and is quite competent for the targeted applications*[1].

Index Terms — Shot boundary detection, scene change detection, middle level features, unified model, sports video analysis, video indexing.

I. INTRODUCTION

With the prevalence of large capacity multimedia data storage/playback devices, such as Personal Video Recorders, there is a great demand for the embedded function of effective video content indexing, browsing and retrieval. Sports constitute a most important content type among the video sequences recorded by the consumers. Due to their enormous market appeal, some automatic sports highlight detection modules have already been implemented in high-end Personal Video Recorders [1].

Because of its popularity throughout the world, soccer is the most representative sports type in the highlight detection application in Personal Video Recorders. After the live broadcast, most audiences are only interested in the highlights, such as goals, shots on goal, placed (including corner, free, and penalty) kicks, and red card events. From the statistics of 12 soccer matches, the highlights only occupy about 1/10 in each match (100 minutes around). Therefore, the analysis in this paper is mainly focused on soccer video; while our scheme can be easily extended to other sports types, such as baseball, tennis and basketball.

Generally, sports video analysis can be based on the three modalities (i.e. textual, auditory, and visual). Textual information, such as closed caption, can greatly simplify the analyzing process [2]; while it always lags behind the other information, and is usually unavailable. As is discussed in [3], auditory information does not indicate any specific highlight, unless the arduous task of understanding the speech from a noisy background could be fulfilled. Therefore, visual information analysis is quite necessary for sports highlight detection applications in Personal Video Recorders.

Shot boundary detection is usually the first step toward automatic video indexing and browsing. It searches for and recognizes the visual discontinuities caused by the transitions, to segment a video stream into elementary uninterrupted content units for subsequent high-level semantic analysis. Despite the long research history and numerous proposed techniques, shot boundary detection is not completely solved [4]. And since "sports video is arguably one of the most challenging domains for robust shot boundary detection" [5], the inadequate performance of general techniques [6] will inevitably negatively affect the subsequent tasks.

There are two kinds of shot transitions, i.e. abrupt changes (also called cuts) and gradual ones, which are presented mainly in the forms of fade (seldom appears in nowadays sports video), dissolve, and wipe. Traditionally, if there exist frames that are sandwiched by the adjacent shots but belong to neither of them, the transition is called a gradual one; otherwise, it is called a cut.

In the literature, "region color histograms" is the most recommended feature for cut detection [7] [8] [9]. J. Bescos, et al proposed a unified model for shot boundary detection, and reported excellent performance on a relatively large data set [10]. A. Ekin, et al proposed the feature "dominant color proportion" for shot boundary detection in sports video, and obtained much better results than general detectors [5]. Recently, Support Vector Machine (SVM) has been successfully adopted as a statistical machine learning approach to automatically construct a decision hyper-plane during the training procedure [11] [12]. While the algorithms in [11] and [12] only adopt general features, and their performances, though relatively high, are not satisfactory for sports video analysis applications. In our implementation, which integrated the aforementioned model and features, we found that the following problems are frequently encountered.

[1] This work was supported in part by the National Natural Science Foundation of China under Project 60472028 and by the Research Fund for the Doctoral Program of Higher Education under Grant No. 20040003015.

B. Han and Y. Hu are with Institute of Image and Graphics, Department of Electronic Engineering, Tsinghua University, Beijing 100084 China (e-mail: {hanb02, hyc00}@mails.tsinghua.edu.cn).

G. Wang is with Department of Electronic Engineering, Tsinghua University, Beijing 100084 China (e-mail: wangguijin@tsinghua.edu.cn).

W. Wu is with Sony China Research Laboratory, Beijing 100080 China (e-mail: weiguo.wu@sony.com.cn).

T. Yoshigahara is with Information Technologies Laboratories, Sony Corporation, Tokyo Japan (e-mail: takayuki.yoshigahara@jp.sony.com).

Contributed Paper
Manuscript received July 15, 2007

Fig. 1. (a) An abrupt shot transition with an interlace frame (the middle one); (b) three frames selected during a gradual shot transition (dissolve), they are 5 frames apart in the video sequence.

Fig. 2. (a) Three frames selected during a fast camera pan (b) three frames selected during a graphics wipe. The interval between each adjacent frame pair is 9 in the video sequences.

1) When the last frame before the transition and the first frame after the transition both have the field as background, their features, region color histograms and dominant color proportions, are very similar (see Fig. 1); this often results in a "miss".

2) During a fast camera pan, which is used to track a running player with a close-up view, the region color histograms change as fast as those during a gradual transition (see Fig. 2); this often results in a "false".

Feature extraction is essential for the performance of shot boundary detection algorithms [7] [8] [9] [13] [14]. As shot transition is a semantic level concept in nature, the low level features used in existing algorithms are incapable of solving the problems above. Although object level features are ideal to address the semantic meaning, video object analysis still remains a complex and challenging task today.

In this paper, we designed two novel kinds of middle level features to fully characterize the object level features in sports video, such as the field region, player movement, etc. One kind is based on dominant color segmentation; the other is based on motion vector filtration. Further, these features, together with the feature of region color histograms, are integrated into a unified model based on the ideas in [5] and [10]. Finally, the SVM based classifiers, which are obtained after a training stage, make the decision for both abrupt transitions and gradual ones.

Our experiments are carried out on a relatively large data set consists of 12 video sequences, of which each corresponds to half of a soccer match. 6 of them are used to train the classifiers, and the other 6 are used to test our scheme. The experiments demonstrate that our scheme achieved a considerable improvement over the algorithms mentioned above.

The rest of this paper is organized as follows. Section 2 and section 3 propose middle level features, which is based on dominant color segmentation and motion vector filtration, respectively. Object level discrimination of these features is also demonstrated here. Section 4 presents the overall SVM based structure of our scheme. Section 5 exhibits the experimental results. Section 6 concludes the paper.

II. Dominant Color (DC) Middle Level Features

Due to large proportion of field (or court) area, which is characterized by a specific uniform color, in most frames in sports video, DC segmentation has been widely adopted as an effective tool for shot analysis, camera view classification and object detection, etc [5] [15] [16] [17] [18]. In our scheme, based on the proposed DC segmentation algorithm, the extracted features, F_{DC1}, F_{DC2}, and F_{DC3}, can fully reflect the regions and sizes of objects in the frames that are dominated by the field (or court) region.

A. DC Segmentation

Because the fields in broadcast soccer video always have a greenish tone (may vary from cyan to yellow, actually), the DC stably lies in a compact region in the hue channel of the HSV (Hue-Saturation-Value) [15], or HSI (Hue-Saturation-Intensity) color space [5] (See Fig. 3). As has been stated in the literature, the DC distribution may vary under the circumstances of different stadiums, weathers, or lighting conditions; so the DC model must be adaptively extracted. In [5] and [15], the DC needs to be learned at start-up for each video clip using the cumulative histogram. The thresholds adopted for segmentation are heuristically determined [5] or Gaussian distribution assumption based [15]. While it should be noted that the DC distribution may also vary due to different camera views in soccer video (This is partly caused by the adaptability of digital video cameras.) So it is not sound to adopt a single DC for all the frames.

In the proposed scheme, we make the DC frame adaptive, thus avoid the trouble of start-up learning and the problem of game start detection. It is effective and robust because:

1) In a video frame with the field (or court) as its background, the portion of the DC pixels is large enough for DC extraction.

2) Our DC model is based on the HSV histogram statistics which is obtained from a large number of video clips ("The best goals of UEFA Champions League in season 04-05" and "Top 100 goals of the world in 2006").

It can be seen from Fig. 3 that the DC can not be modeled using any well-formed distribution. While we have two observations in the HSV color space:

One is the possible range (Hue: [20, 55]; Saturation: [16, 255]; Value: [32, 207]) of soccer video DC, where Hue has 180 levels, both Saturation and Value have 256 levels.

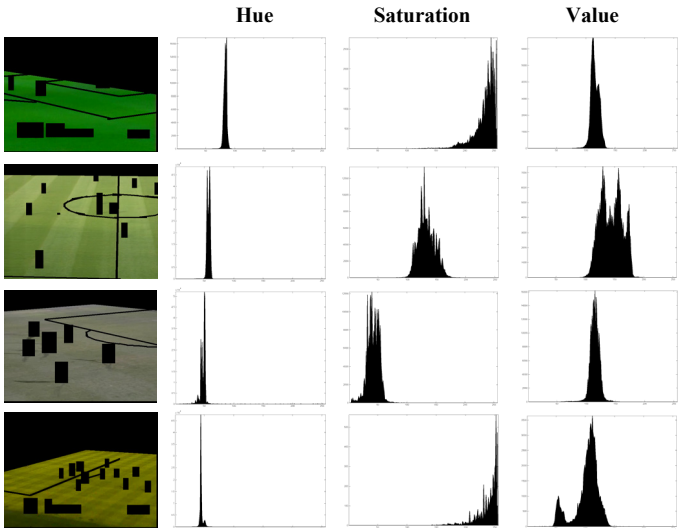

Fig. 3. Soccer field DC sample frames, and their histograms in the HSV color space. To obtain the statistics, non-DC regions are manually masked with black in the original frame.

The other is that the DC occupies no more than 16 bins in the Hue histogram, and no more than 160 bins in the Saturation histogram.

Based on these observations, the DC can be extracted. Then the binary DC mask of this frame (tagged as t), $M_{DC}(t)$, is generated via marking the DC pixels with 1 and others with 0. See Fig. 4 for some DC segmentation results.

The *DC proportion* feature is defined as in (1), where N_H and N_W denote the frame height and frame width, respectively.

$$F_{DC1}(t) = \sum_{i=1}^{N_H}\sum_{j=1}^{N_W} M_{DC}(t,i,j) \Big/ (N_H \cdot N_W) \tag{1}$$

If $F_{DC1}(t)$ is smaller than a specific threshold, such as 0.2, we will regard the extracted color as non-dominant and set $F_{DC1}(t) = 0$, just as when the frame is dominated by a non-greenish color.

B. Feature Definitions

If the feature $F_{DC1}(t) > 0$, two novel middle level features, $F_{DC2}(t)$ and $F_{DC3}(t)$, will be extracted from the mask $M_{DC}(t)$, which results from our DC segmentation. $F_{DC2}(t)$ mainly characterizes the *soccer field region* captured by the camera; and $F_{DC3}(t)$ mainly represents the *relative player size* in the frame.

As is known, such features can be extracted via region based 2-D analysis, which is quite case-specific and computationally complex. It is found that the projections of $M_{DC}(t)$ contain plenty of information that reflects object level features. Thus, we base the novel features on the normalized projection vectors, to make our algorithm robust and practical.

The original horizontal projection is calculated as follows.

$$P_H^o(t,i) = \sum_{j=1}^{N_W} M_{DC}(t,i,j) \Big/ N_W,\ i = 1,2,\cdots,N_H \tag{2}$$

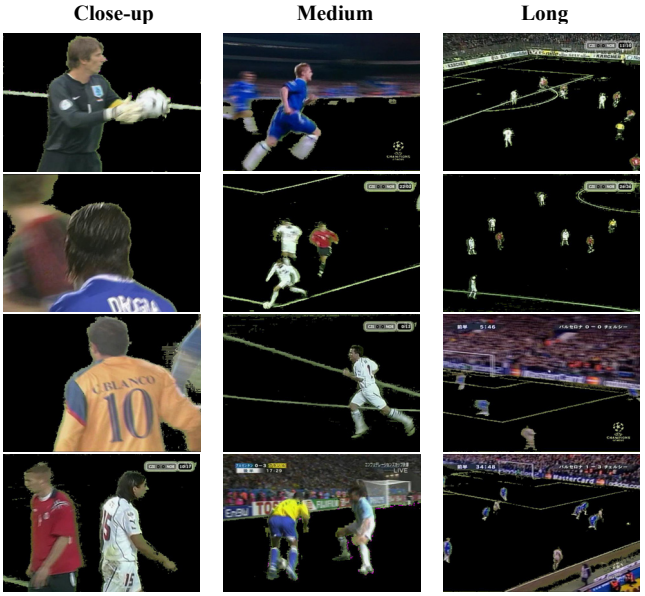

Fig. 4. Examples of DC segmentation result. The black pixels represent the DC region. Frames in the three columns belong to the Close-up, Medium, and Long shot types, respectively.

Fig. 5. Examples of soccer field region detection results. In the bottom row, the lower-corner regions are detected and excluded from the relative player size feature extraction.

To facilitate feature extraction, it is down-sampled using median filter to obtain the 10-element-vector $P_H(t)$. $F_{DC2}(t)$ is extracted from $P_H(t)$. We prescribe that $P_H(t)$ must contain no less than 3 successive elements that correspond to the field-region. Because such successive elements are larger than other elements, the $P_H(t)$ elements can be classified into two clusters (See Fig. 5 for some classification results.) Then, $F_{DC2}(t)$ is defined based on the classification.

$$F_{DC2}(t) = \underset{i \in Field-region}{mean} P_H(t,i) - \underset{i \in Out-of-field-region}{mean} P_H(t,i) \tag{3}$$

The vertical projection calculation is also based on the classification.

$$P_V^o(t,j) = \underset{i \in Field-region}{mean} M_{DC}(t,i,j),\ j = 1,2,\cdots,N_W \tag{4}$$

Then, similar to $P_H(t)$, a 30-element-vector $P_V(t)$ is obtained via down-sampling (4) using median filter.

As introduced, $F_{DC3}(t)$ is designed to characterize the relative player size in the frame. In our algorithm, the player region consists of the non-DC pixels inside the field region. While in the case of lower corner view (see Fig. 5) in soccer video, the bottom out-of-field region may be considered as a player region in the previous classification. So we must distinguish this case and eliminate the effect of this region.

In this case, the 3 bottom-most elements of $P_H(t)$ mainly result from the bottom out-of-field region, which is right-angled triangular. Thus, the value of the 3 elements is considered as a function of the index, and the best fitting line function is obtained using the Least Squares method. Taking the line parameters and the fitting error as the features, we found that this case can be perfectly distinguished from other cases. The classifier is trained using 10 lower corner view frames and many other kinds of frames. If such a bottom right-angled triangular region is detected, $P_H(t)$ will be re-calculated based on the line function. This is equivalent to regarding pixels in this region as DC pixels.

Now $F_{DC3}(t)$, which is defined as the sum of relative player width and relative player height, will be computed as follows.

$$F_{DC3}(t)=1-\max_{i\in bottom-most\ 4\ elements} P_H(t,i)+ \underset{j\in Non-player-region}{mean} P_V(t,j)-\underset{j\in Player-region}{mean} P_V(t,j) \tag{5}$$

Any soccer video Close-up view frame with $F_{DC1}(t) > 0$ has a distinctive property, that all of the bottom-most 4 elements of $P_H(t)$ are relatively small, because the player region occupies a large proportion in the bottom half of the frame. So the former part of the definition is designed to represent the relative player width in the frame.

The latter part of the definition, which represents the relative player height, is based on the player region classification. First please note that if a bottom right-angled triangular region is detected, the $P_V(t)$ elements that are affected by this region will not be used in (5) (See Fig. 5 for the effective region for relative player height computation.)

We prescribe that $P_V(t)$ may contain no more than two player regions, of which each consists of several successive elements. Because elements in such regions are smaller than other elements, the $P_V(t)$ elements can be classified into two clusters. It is obvious that the latter part will be much smaller if the frame presents a Long view.

In order to demonstrate the effectiveness of our features for differentiate between camera views with $F_{DC1}(t) > 0$, we will compare it with the features proposed in [5] for shot classification. Because our features reflect object level information in the frame, which is essential for shot classification, our algorithm performs much better (Fig. 6). As is shown, the three shot types can be clearly classified by the black dashed lines in the proposed feature space.

The philosophy of making use of these features in shot boundary detection is: the two shots on both sides of a shot boundary have a high probability of belonging to different shot types; hence their frames usually have diverse DC features.

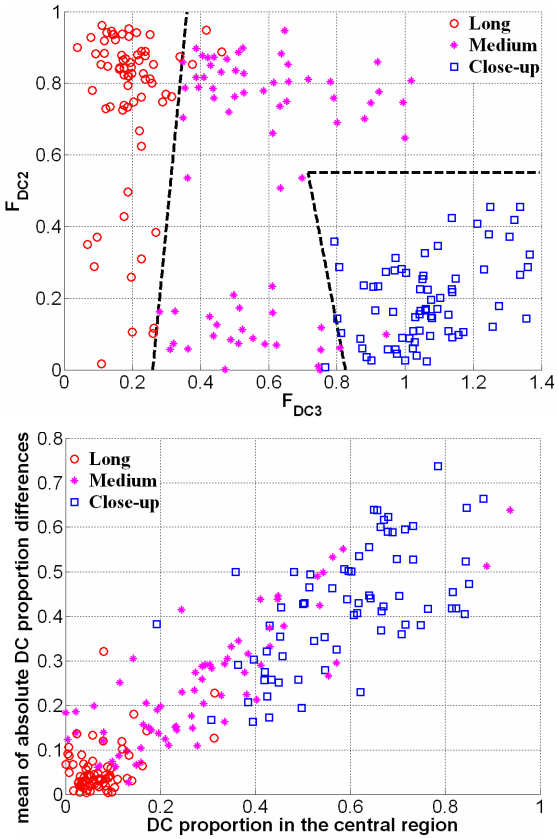

Fig. 6. Frame-view classification comparison. For each of the 3 types, i.e. Long, Medium, and Close-up, 70 samples with $F_{DC1} > 0$ are selected from different sequences. The top one demonstrates their distribution in the feature space proposed in this paper; while the features proposed in [5] are adopted in the bottom one.

III. Motion Vector (MV) Middle Level Features

Besides the DC features, which has been exploited in last section, motion vector is another effective low level feature, from which middle level features could be extracted to characterize object level features [15] [19]. In our work, this kind of feature is designed to mainly reflect the texture and motion of the objects in sports video.

In video compression, the obtained MVs are not guaranteed to represent the true motion of the corresponding blocks, especially in sports video. For instance, the cases of rapid changing content, large low-textured areas and image blur caused by camera operations would make the MVs arbitrary and unreliable. Thus, we propose a MV filtration scheme [20] to eliminate such unreliable MVs for video data analysis.

Most existing approaches extract features from the magnitude and direction of the MVs in a frame [15] [19]. These features can reflect the characteristics of camera operation and object movement, with the prerequisite that all the MVs are reliable. We argue that the distribution of reliable

MVs is much more important than the magnitude and direction of MVs for shot boundary detection. Two features, denoted by F_{MV1} and F_{MV2}, will be defined here to help solving the two problems listed in section 1.

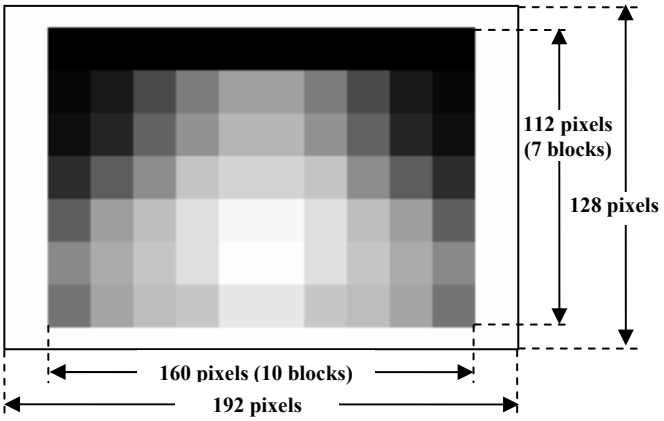

Fig. 7. Illustration of the weight matrix. (the grayscale of blocks represent their reliability probability in fast camera pan clips, white = 1, black = 0.)

In order to reduce the computational cost, each frame is first down-sampled to 192×128, and the block size of 16×16 is selected. The motion compensation on the marginal columns and rows is considered unreliable because of possible camera operations. Thus, motion search is only carried out for the central 10×7 blocks (Fig. 7).

It is obvious that few reliable MVs exist on the shot boundary of a cut. Hence, to help solving problem 1) in section 1, we extract the feature *Proportion of Reliable MVs* for each frame as follows.

$$F_{MV1}(t) = \frac{1}{70}\sum_{i=1}^{7}\sum_{j=1}^{10} M_{MV}(t,i,j) \tag{6}$$

where (i, j) is the block coordinate in the frame; M_{MV} is a 7×10 binary mask matrix, of which each element represents whether the MV of that block is reliable.

Observing Fig. 2, we can hardly distinguish between the two cases from temporal change of the color based features. While, you will see that this problem can be solved using the second MV based feature that will be extracted next.

When the camera is panning fast to track a running player with a close-up view, the background usually becomes blurred because it is out of focus and moving fast relative to the camera; while the foreground object remains clear and locates stably in the central region of the frame. As a result, the blocks with reliable MVs are densely distributed in the central region of the frame. In contrast, such blocks, if exist, are arbitrarily distributed during a gradual shot transition.

Hence, to help solving problem 2) in section 1, we extract another feature *Centrality of Reliable MVs* for each frame as follows.

$$F_{MV2}(t) = \sum_{i=1}^{7}\sum_{j=1}^{10}\left[W_{MV}(i,j)\cdot M_{MV}(t,i,j)\right] \Big/ \sum_{i=1}^{7}\sum_{j=1}^{10} M_{MV}(t,i,j) \tag{7}$$

where W_{MV} is a weight matrix with the same size as $M_{MV}(t)$. W_{MV} is obtained beforehand by temporally averaging the $M_{MV}(t)$ of 20 fast camera pan clips (about 1 second long each) collected from the training video data. Thus, each element of W_{MV} represents the probability of the case that the corresponding block has a reliable MV. Fig. 7 demonstrates the W_{MV} in the form of an image, where the intensity is in proportion to the probability. So the frames inside a fast camera pan clip are assumed to have larger $F_{MV2}(t)$ values than those inside a gradual shot transition clip.

Finally, our scheme also exploits MVs themselves, as in others' works. The *Average of Reliable MVs* for each frame, which is a two-component-vector, is defined as follows.

$$F_{MV3}(t) = \sum_{i=1}^{7}\sum_{j=1}^{10}\left[MV(t,i,j)\cdot M_{MV}(t,i,j)\right] \Big/ \sum_{i=1}^{7}\sum_{j=1}^{10} M_{MV}(t,i,j) \tag{8}$$

IV. CLASSIFICATION BASED ON SVM

As is stated in [4], a typical shot boundary detection algorithm is a specific combination of three elements, i.e. feature, metric and threshold.

Besides the novel middle level features proposed in the former two sections, the extensively adopted Region Color Histograms (RCH), which compose the third kind of features of our framework, will be briefly introduced in this section.

Based on the ideas in [5] and [10], all the features are integrated into a unified model, in which both cuts and gradual transitions can be well distinguished.

The final step is pattern classification in a high-dimensional decision space. A classifier structure like the decision tree [10] is suitable for problems with few features. While it needs several thresholds to be manually selected in our scheme, thus makes the training complicated and impractical. Therefore, SVM is adopted to automatically construct a decision hyper-plane. LibSVM [22] is used to find the best parameters for the C-SVC with Radial Basis Function kernel, and then to train the classifier.

A. Difference of RCH

In the literature, we found that the RCH is the most recommended feature [7] [8] [9] for cut detection. And similar features of color histograms have proven also effective for gradual transition detection [10] [11] [12]. Thus, the RCH is chosen as a basic feature in our framework.

The algorithm in this subsection may be considered as a modified version of that proposed in [21]. Since the RCH should not change rapidly due to camera operations, i.e. pan,

tilt, and zoom, which are frequently used in sports video, the area of each region should be relatively large. In our algorithm, each frame is uniformly divided into 6 regions, in a 3 rows by 2 columns pattern (Fig. 8). The width of each region is much larger than its height because pans appear much more frequently than tilts in soccer video. For each region, a normalized color histogram with 64 bins (2 bits for each color channel) is computed in the CIE Lab space, which is recommended for histogram comparison [8] [14].

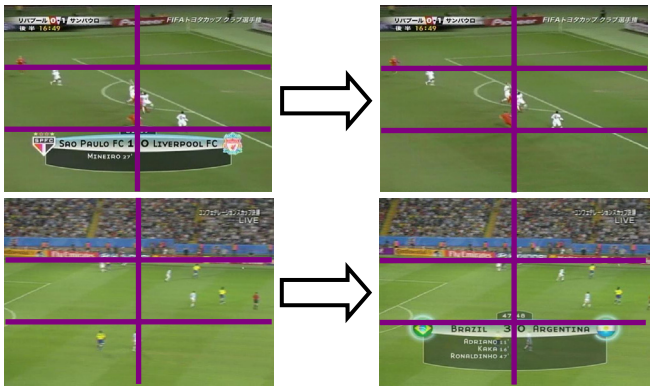

Fig. 8. Each frame is segmented into 6 regions for RCH calculation. If a large logo area in the bottom region(s) abruptly appears or disappears, the corresponding bottom region(s) will be excluded from the calculation of D_{RCH} for cut detection.

The bin-to-bin difference is adopted for histogram comparison between two frames as follows, where t is the tag of a frame, s is the interval between two frames, H denotes the normalized color histogram, the subscript i indicates the corresponding region, and j denotes the bin index.

$$D_i(t,s)=\frac{1}{2}\sum_{j=1}^{64}\left|H_i(t,j)-H_i(t-s,j)\right|, i=1,2,\cdots,6 \tag{9}$$

$$D_{RCH}(t,s)=\underset{i}{mean}\, D_i(t,s) \tag{10}$$

In our experiments, it is found that false cut detections are often caused by the case when a large logo suddenly appears or disappears, and most of such logos locate in the bottom 2 regions (Fig. 8). We propose a clustering based algorithm to elimination such effects. The 6 $D_i(t, s)$, together with the value 0, are first clustered into two classes. If the cluster of larger values only consists of histogram differences of the bottom 2 regions, the values in this cluster will not be used for the computation of the overall RCH difference in (10).

B. Abrupt Transition Detection

For interlace TV broadcast signal, a cut may contain one mixed frame between the adjacent shots (see Fig. 1). And this frame can not be decomposed via down-sampling, because of the video compression effect. Thus, a transition should also be regarded as a cut if there is only one frame sandwiched by the adjacent shots but belongs to neither of them. In our scheme, to judge whether a cut exists before frame t, we only compare it with frame $t-2$, rather than frame $t-1$.

As the first step of the classification framework, the following condition is checked on the input frame pair.

$$C_{DC}(t,s):\left(F_{DC1}(t)>0\right)\wedge\left(F_{DC1}(t-s)>0\right) \tag{11}$$

As is stated in [5], the frame pair being very similar is much more probable if $C_{DC}(t,2)$ is true. So we trained two classifiers separately, as shown in Fig. 9.

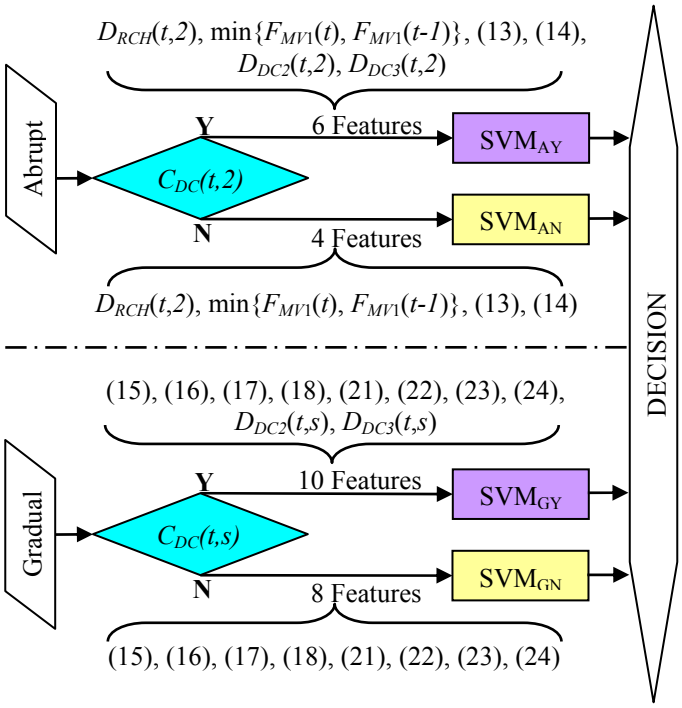

Fig. 9. Feature classification framework for shot boundary detection. For both abrupt transitions and gradual ones, two SVMs are trained separately according to the condition in (15).

If and only if $C_{DC}(t,2)$ is true, the following differences, $D_{DC2}(t,2)$ and $D_{DC3}(t,2)$, will be used (in SVM$_{AY}$ in Fig. 9).

$$\begin{cases} D_{DC2}(t,s)=\left|F_{DC2}(t)-F_{DC2}(t-s)\right| \\ D_{DC3}(t,s)=\left|F_{DC3}(t)-F_{DC3}(t-s)\right| \end{cases}, if\ C_{DC}(t,s) \tag{12}$$

$D_{RCH}(t,2)$ and min$\{F_{MV1}(t), F_{MV1}(t\text{-}1)\}$ are always adopted in abrupt transition detection. They should be used together with information from the neighborhood so as to adaptively determine the corresponding thresholds. So the other two features are defined as follows, where N_N denotes the neighborhood range. Here we do not use too many neighborhood features as in [10], to avoid the problem of possible over-fitting.

$$\overline{D}_{RCH}(t,2)=\frac{1}{N_N}\max\left\{\sum_{i=2}^{N_N+1}D_{RCH}(t-i,2),\sum_{i=2}^{N_N+1}D_{RCH}(t+i,2)\right\} \tag{13}$$

$$\overline{F}_{MV1}(t)=\frac{1}{N_N}\min\left\{\sum_{i=2}^{N_N+1}F_{MV1}(t-i),\sum_{i=2}^{N_N+1}F_{MV1}(t+i)\right\} \tag{14}$$

C. Gradual Transition Detection

Similar as our cut detection scheme, if and only if $C_{DC}(t,s)$ is true, the two DC features in (12) will be made use of. And according to this condition, two classifiers for gradual transition detection are trained separately, as in Fig. 9.

The following three MV features, say, the average Proportion, Centrality, and Amplitude of reliable MVs, are always adopted in gradual transition detection.

$$\overline{F}_{MV1}(t,s)=\frac{1}{s}\sum_{i=0}^{s-1}F_{MV1}(t-i) \tag{15}$$

$$\overline{F}_{MV2}(t,s)=\sum_{i=0}^{s-1}F_{MV1}(t-i)\cdot F_{MV2}(t-i)\Big/\sum_{i=0}^{s-1}F_{MV1}(t-i) \tag{16}$$

$$\overline{F}_{MV3}(t,s)=\left\|\sum_{i=0}^{s-1}F_{MV1}(t-i)\cdot F_{MV3}(t-i)\Big/\sum_{i=0}^{s-1}F_{MV1}(t-i)\right\| \tag{17}$$

In [10], frames ordering information is exploited to model the frame difference patterns generated by gradual transitions. The detection of gradual transitions of length s is carried out via searching for the peak-like pattern in the sequence of $D_{RCH}(t, s)$ with a sliding time window of $2s + 1$.

However, we found that the features proposed in [10] will miss most of the gradual transitions in nowadays sports video. This is mainly because most of the patterns generated by gradual transitions are no longer peak-like, but plateau-like or even "M"-like. Fig. 10 exhibits two examples.

Therefore, five frame ordering features are proposed in this paper to better characterize the patterns generated by gradual transitions. Suppose we are examining whether frame t is the first frame of a shot, and t - s is the last frame of the previous shot, the features will be extracted from the sequence of { $D_{RCH}(t\text{-}s, s)$, …, $D_{RCH}(t+s, s)$ } as follows. Note that in practice, we only use several transition length search steps, for example, 12, 18, 24, 30, and 36 when the frame rate is 29.97.

First, the mean boundary value is defined as follows.

$$\left[D_{RCH}(t-s,s)+D_{RCH}(t+s,s)\right]/2 \tag{18}$$

All the other 4 feature definitions are based on the left peak position and right peak position, which are defined as follows.

$$\begin{cases}P_L(t,s)=\arg\max_{i\in[t-s,t]}D_{RCH}(i,s)\\ P_R(t,s)=\arg\max_{i\in[t,t+s]}D_{RCH}(i,s)\end{cases} \tag{19}$$

Then, one of the peak positions may be adjusted as follows to enable our model to characterize the peak-like pattern at the same time.

$$\begin{cases}P_L(t,s)=P_R(t,s),\ if\ \left[P_L(t,s)=t\right]\wedge\left[P_R(t,s)\neq t\right]\\ P_R(t,s)=P_L(t,s),\ if\ \left[P_R(t,s)=t\right]\wedge\left[P_L(t,s)\neq t\right]\end{cases} \tag{20}$$

The following three features represent the mean peak value, the relative distance between the peaks, and value of the valley point between the peaks, respectively. It is obvious that for a gradual transition, (21) is expected to be much larger than (18).

$$\left[D_{RCH}(P_L(t,s))+D_{RCH}(P_R(t,s))\right]/2 \tag{21}$$

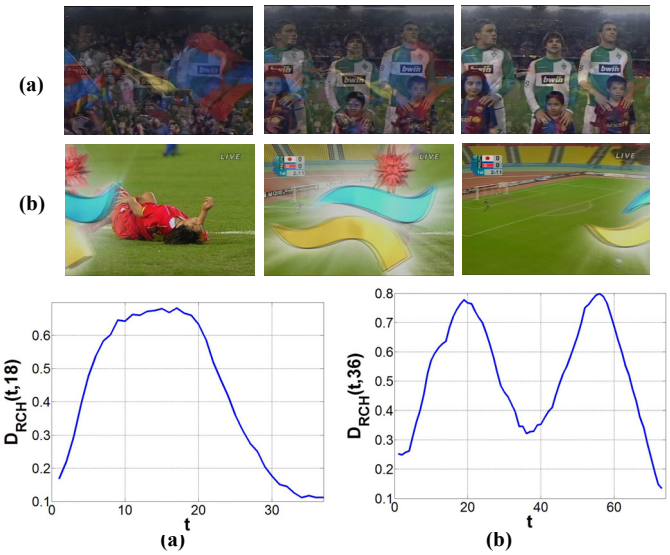

Fig. 10. (a): sample images during a dissolve (about 18 frames long) and the corresponding D_{RCH} sequence (in a plateau-like pattern). (b): sample images during a graphics wipe (about 36 frames long) and the corresponding D_{RCH} sequence (in an "M"-like pattern).

$$\left[P_R(t,s)-P_L(t,s)\right]/2s \tag{22}$$

$$\min_{i\in[P_L(t,s),P_R(t,s)]}D_{RCH}(i,s) \tag{23}$$

The features above characterize the D_{RCH} pattern between the two peak positions. At the same time, for a gradual transition, the sequence is expected to be steadily increasing between $t - s$ and $P_L(t,s)$; and to be steadily decreasing between $P_R(t,s)$ and $t + s$. So, the last frame ordering feature is defined as follows.

$$\sum_{i=t-s+1}^{P_L(t,s)}\max\left\{0,-D'_{RCH}(i,s)\right\}\Big/2\left[D_{RCH}(P_L(t,s),s)-D_{RCH}(t-s,s)\right]+\sum_{i=P_R(t,s)+1}^{t+s}\max\left\{0,D'_{RCH}(i,s)\right\}\Big/2\left[D_{RCH}(P_R(t,s),s)-D_{RCH}(t+s,s)\right] \tag{24}$$

where the difference is defined as:

$$D'_{RCH}(t,s)=D_{RCH}(t,s)-D_{RCH}(t-1,s) \tag{25}$$

V. Experimental Results

Our experiments are carried out on a relatively large data set consisting of 12 video sequences (Table I), of which each corresponds to one half of a soccer match. 6 of them (with

gray background) are used to train the SVMs, and the other 6 are used to evaluate our scheme. For all the sequences, the frame size is 704×480, and the frame rate is 29.97fps. These sequences are diversified enough for examining the proposed algorithm, since they were shot in different stadiums and edited by different TV stations.

TABLE I
SOCCER VIDEO DATA IN OUR EXPERIMENTS

	Date	Tournament	Half	Length
1	20050406	UEFA CL	2	0:44:59
2	20061206	UEFA CL	1	1:03:45
3	20050629	FIFA CC	2	1:00:56
4	20051117	FIFA WC	1	0:43:06
5	20061213	FIFA Club WC	2	0:46:01
6	20061213	Asian Games	1	0:43:02
1	20061206	UEFA CL	2	1:07:45
2	20060124	UEFA CL	1	0:51:19
3	20061101	UEFA CL	2	0:50:53
4	20060616	FIFA WC	1	0:47:39
5	20061214	FIFA Club WC	2	0:46:22
6	20061206	Asian Games	1	0:46:10

Because the sequences are recorded in recent years, they have more complicated editing effects than the data collected many years ago, such as most of the MPEG test sequences; and this makes the shot boundary detection more difficult. Compared to other types of gradual transitions, fades are not so difficult to detect because of the distinctive characteristic of a monochrome frame [7]. While in nowadays broadcast soccer video, we found that fades seldom appear, and wipe types other than graphics wipes are not used any more. Thus, the performance of the existing techniques on our data should not be expected to be as good as those reported in the literature.

To evaluate the performance of our scheme, we adopt the *F1* measure, which is defined as follows.

$$F1 = 2 \cdot Recall \cdot Precision / (Recall + Precision) \tag{26}$$

where Recall and Precision are defined as:

$$\begin{cases} Recall = hit/(hit + miss) \\ Precision = hit/(hit + false) \end{cases} \tag{27}$$

The experimental results on the test sequences are shown in Table II. To demonstrate effectiveness of the middle level features, we trained other two SVMs with only the RCH features as input, for the detection of cuts and gradual transitions, respectively. These results are shown in Table II with gray background.

Please note that video shot transitions should not be too close to each other or overlap (for gradual ones). The results in Table II are obtained after removing some of the false detections of the SVMs using such apriori knowledge.

The proposed middle level features bring in considerable enhancement (nearly 2 percent) to the overall shot boundary detection performance. And we can find that the improvement on gradual transition detection (nearly 9 percent) is much more obvious. While the cut detection performance using only the RCH feature is already so high (see the results on test sequence 4, 5, and 6) that it is very hard to make it better.

TABLE II
EXPERIMENTAL RESULTS ON 6 TEST SEQUENCES

		#	*F1*	
			RCH only	Our scheme
1	Cut	318	0.970	0.981
	Gradual	116	0.778	0.824
	Overall	434	0.922	0.939
2	Cut	284	0.941	0.951
	Gradual	80	0.815	0.857
	Overall	364	0.915	0.929
3	Cut	360	0.979	0.985
	Gradual	156	0.800	0.914
	Overall	516	0.931	0.964
4	Cut	325	0.994	0.992
	Gradual	114	0.632	0.767
	Overall	439	0.917	0.940
5	Cut	211	0.986	0.986
	Gradual	49	0.814	0.891
	Overall	260	0.956	0.967
6	Cut	382	0.990	0.994
	Gradual	65	0.714	0.814
	Overall	447	0.954	0.966
Overall	Cut	1880	0.977	0.983
	Gradual	580	0.759	0.848
	Overall	2460	0.932	0.951

Via the high performance on all the sequences, our scheme proved competent for automatic sports video analysis applications. Moreover, the computational cost is acceptable. The processing speed for a MPEG4 soccer video stream, of which the frame size is 704 by 480, is about 25 to 30 FPS (obtained in the MS Visual Studio development environment on a standard PC with P4-1.7GHz CPU and 512MB RAM).

VI. CONCLUSION

Shot boundary detection is an essential step in sports video highlight detection application in Personal Video Recorders. To avoid the shot boundary detection errors that often occur when using existing techniques, we propose two novel kinds of middle level features, which are integrated into a unified model, together with the region color histograms feature. The proposed scheme achieved superior performance and proved robust on a relatively large soccer video data set. And it can be easily extended to index video streams of some other sports types, such as baseball, tennis, and basketball. It is suitable to apply the scheme in Personal Video Recorders, in respect that the middle level features can be extracted in real time.

Furthermore, the DC features can be adopted for sports video shot classification. Because the graphics object always has a stable motion direction during a wipe, F_{MV3} can be adopted to differentiate graphics wipes from other gradual

transition types in sports video. This will provide a very helpful clue for replay detection, which is one of the key components in sports video analysis [23].

REFERENCES

[1] I. Otsuka, K. Nakane, A. Divakaran, K. Hatanaka, and M. Ogawa, "A highlight scene detection and video summarization system using audio feature for a personal video recorder," *IEEE Trans. Consumer Electronics*, vol. 51, no. 1, pp. 112-116, Feb. 2005.

[2] N. Nitta, N. Babaguchi, and T. Kitahashi, "Generating Semantic Descriptions of Broadcasted Sports Videos Based on Structures of Sports Games and TV Programs," *Multimedia Tools and Applications*, vol. 25, pp. 59-83, 2005.

[3] M. Bertini, A. D. Bimbo, and W. Nunziati, "Common Visual Cues for Sports Highlights Modeling," *Multimedia Tools and Applications*, vol. 27, pp. 215-228, 2005.

[4] A. Hanjalic, "Shot-boundary detection: unraveled and resolved?" *IEEE Trans. Circuits and Systems for Video Technology*, vol. 12, no. 2, pp. 90-105, Feb. 2002.

[5] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization", *IEEE Trans. Image Processing*, vol. 12, no. 7, pp. 796-807, July 2003.

[6] J.-R. Kim, S. Suh, and S. Sull, "Fast scene change detection for personal video recorder," *IEEE Trans. Consumer Electronics*, vol. 49, no. 3, pp. 683-688, Aug. 2003.

[7] R. Lienhart, "Reliable transition detection in videos: a survey and practitioner's guide," *International Journal of Image and Graphics*, vol. 1, no. 3, pp. 469-486, 2001.

[8] U. Gargi, R. Kasturi, and S. H. Strayer, "Performance characterization of video-shot-change detection methods," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, no. 1, pp. 1-13, 2000.

[9] J. S. Boreczky and L. A. Rowe, "Comparison of video shot boundary detection techniques," *Journal of Electronic Imaging*, vol. 5, no. 2, pp. 122-128, April 1996.

[10] J. Bescos, G. Cisneros, J. M. Martínez, J. M. Menéndez, and J. Cabrera, "A Unified Model for Techniques on Video-Shot Transition Detection," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 293-307, Nov. 2005.

[11] H. Feng, W. Fang, S. Liu, and Y. Fang, "A new general framework for shot boundary detection and key-frame extraction," *Proc. ACM SIGMM Int. Workshop Multimedia Information Retrieval*, Nov. 2005, pp. 121–126.

[12] J. Yuan, J. Li, F. Lin, and B. Zhang, "A unified shot boundary detection framework based on graph partition model," *Proc. ACM Multimedia*, Nov. 2005, pp. 539-542.

[13] S. Lefevre, J. Holler, and N. Vincent "A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval," *Real-Time Imaging*, vol. 9, pp. 73-98, 2003.

[14] I. Koprinska, and S. Carrato, "Temporal video segmentation: a survey," *Signal Processing: Image Communication*, vol. 16, no. 5, pp. 477-500, 2001.

[15] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with domain knowledge and hidden Markov models," *Pattern Recognition Letters*, vol. 25, pp. 767-775, 2004.

[16] J. Assfalg, M. Bertini, C. Colombo, A. D. Bimbo, and W. Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," *Computer Vision and Image Understanding*, vol. 92, pp. 285-305, 2003.

[17] L.-Y. Duan, M. Xu, Q. Tian, C.-S. Xu, and J. S. Jin, "A unified framework for semantic shot classification in sports video," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1066-1083, Dec. 2005.

[18] C.-L. Huang, H.-C. Shih, and C.-Y. Chao, "Semantic Analysis of Soccer Video Using Dynamic Bayesian Network," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 749-760, Aug. 2006.

[19] G. Xu, Y.-F. Ma, H.-J. Zhang, and S.-Q. Yang, "An HMM-based framework for video semantic analysis," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 15, no. 11, pp. 1422-1433, Nov. 2005.

[20] Y. Hu, B. Han, G. Wang, and X. Lin, "Enhanced shot change detection using motion features for soccer video analysis," *Proc. IEEE Int. Conf. Multimedia and Expo*, July 2007, pp. 1555-1558.

[21] A. Nagasaka, and Y. Tanaka, "Automatic video indexing and full-video search for object appearances," in: E. Knuth, L.M. Wegner (Eds.), *Visual Database Systems II*, Elsevier, Amsterdam, 1992, pp. 113-127.

[22] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, 2001. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[23] B. Li, J. H. Errico, H. Pan, and I. Sezan, "Bridging the semantic gap in sports video retrieval and summarization," *J. Vis. Commun. Image R.*, vol. 15, pp. 393-424, 2004.

Bo Han received the B.S. degree with honor in Information Engineering from Zhejiang University, Hangzhou China in 2002. He is currently a Ph.D. candidate in Department of Electronic Engineering, Tsinghua University, Beijing China. From March 2005 to July 2005, he was an intern at Intel China Research Center, Beijing China. From November 2005 to April 2006, he was an intern at Sony Intelligent Systems Research Laboratory, Tokyo Japan. His research interests are in the area of digital image/video processing and analysis, and mainly focus on video content understanding.

Yichuan Hu received the B.S. degree in Information and Electronics Engineering from the Department of Electronic Engineering, Tsinghua University, China in 2004. He is currently a master's candidate in Department of Electronic Engineering, Tsinghua University. His research interests are in the area of video compression and communications, video processing and analysis.

Guijin Wang (M'05) received the B.S. and Ph.D. degree (with honor) from the department of Electronic Engineering, Tsinghua University, China in 1998, 2003 respectively, all in the major of Signal and Information Processing. From 2003 to 2006, he has been with Sony Information Technologies Laboratories as a researcher. From Oct., 2006, he has been with the department of Electronic Engineering, Tsinghua University, China as an associate professor. He has published over 20 International journal and conference papers, hold 1 patent and 3 pending patent application. He is the session chair of IEEE CCNC'06, the reviewers for many international journals and conferences. His research interests are focused on wireless multimedia, mesh network, network protocol design, image & video processing, visual compression, and etc.

Weiguo Wu is a Distinguished Researcher of Sony Corporation. He received a B.E. degree from Zhejiang University, China in 1982 and a Ph.D. in Systems and Information Engineering from Yamagata University, Japan in 1996. From 1982 to 1990, he worked as an assistant and then as a lecturer at Zhejiang University, China. In 1990 he came to Japan as a visiting researcher at Yamagata University. He joined Sony Corporation, Tokyo, Japan in 1998. From September 1998 to October 1998, he visited the Robotics Institute, Carnegie Mellon University as a visiting researcher for his joint research project. From 2002 to 2005, he was an associate editor of the Journal of the Institute of Image Information and Television Engineers, Japan. In 2007, he moved to Sony China Research Laboratory, Beijing, China where he is currently the Assistant General Manager. His research interests include computer vision, human machine interaction, statistical learning and video analysis.

Takayuki Yoshigahara is a senior researcher at the Information Technologies Laboratories, Sony Corporation, Tokyo, Japan. He received a ME degree and BE in Applied Physics from Waseda University, Tokyo, Japan, in 1990 and 1988, respectively. He joined Sony Corporation, Tokyo, Japan, in 1990. From 1994 to 1996, he was a visiting research scientist at the Robotics Institute, Carnegie Mellon University, Pittsburgh. His primary research interests are robot vision, video analysis and human-machine interaction.