ICICS-PCM 2003
15-18 December 2003
Singapore

3C2.2

SEMANTIC LABELING OF SOCCER VIDEO

Haiping Sun, Joo-Hwee Lim, Qi Tian, Mohan S. Kankanhalli*

Institute for Inforcomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
*School of Computing, National University of Singapore, Kent Ridge, Singapore 119260
{haiping, joohwee, tian}@i2r.a-star-edu.sg mohan@comp.nus.edu.sg

Abstract

As traditional shot segmentation may not produce video segments that possess one-to-one correspondence to semantic views, we present an integrated segmentation and classification approach to label soccer video into semantic units in this paper. In our system, each P frame is divided to a 6 by 4 blocks with color and motion features extracted on both block and frame levels. First, a threshold is used to divide the video stream into relatively static parts and active parts. Then every active part is segmented into sub-parts according to 4 view types and the motion features are used to classify segments with Support Vector Machines. Finally, static parts are merged with classified active sub-parts to form labeled segments. Four 10-minute test clips from the World Cup 2002 are used to evaluate our system resulting in a promising classification rate of 79.8%.

1. INTRODUCTION

A lot of effort has been put into video retrieval and classification in the past few years. Low-level features such as color, motion and texture are used, but the results are not satisfactory. Researchers are still looking for effective way to bridge the gap between low-level features and semantic meanings. In [1], a soccer video analysis system was presented to classify soccer video into play/break structure by rules. The broad semantic structure extracted is only a good start. In [2], the color, edge and domain rules were used to detect events in tennis. The color-based adaptive filtering is impressive, but comparing unknown events with well-defined sample events in database is rather simplistic. In [3], the authors used energy redistribution functions and 3 templates to extract motion feature for event detection. The complexity of computing energy for each macro block is high and using motion features alone may make the system less robust.

In this paper, we present a novel system to segment and classify soccer video based on color and motion features (Fig. 1). A key objective is to use the labeled segments for event detection later [4]. There are three main phases, namely, preprocessing, segmentation and classification and post-processing:

1. Preprocessing: a short training video is used to compute field colors automatically.
2. Segmentation and Classification: To do segmentation, the video stream is first divided into relatively static parts and active parts. For static parts, motion features are ignored and key frames are saved. Every active part is further segmented into active sub-parts according to 4 view types (defined in Section 2). In Classification stage, motion features are used to classify (label) segments with Support Vector Machines [5].
3. Post-Processing: static parts are merged with adjacent active sub-parts to form semantic segments with semantic labels.

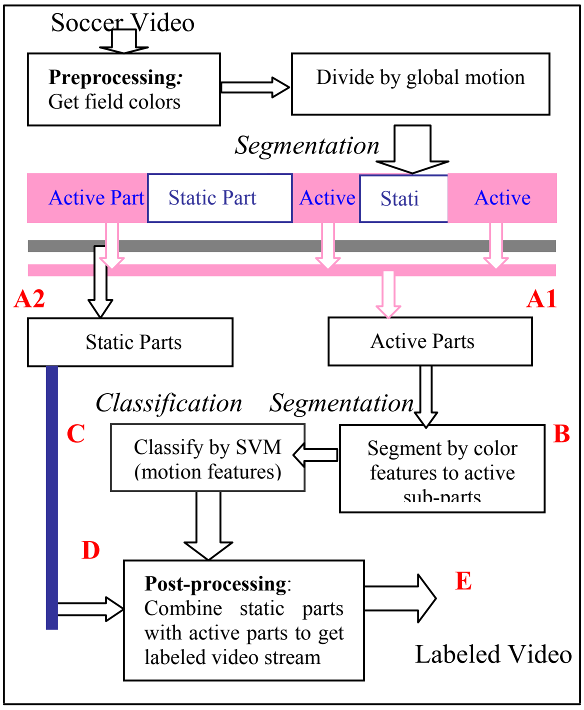

Fig. 1. System architecture of integrated soccer video segmentation and classification

The paper is organized as follow. In Section 2, the motivation and definition of visual keywords for soccer video is presented. Our approach is detailed in Section 3 followed by experimental results in Section 4.

0-7803-8185-8/03/$17.00 © 2003 IEEE

2. VISUAL KEYWORDS AND VIEW TYPES

2.1 Definition of Visual Keywords

Traditionally, the first step to process a video stream is to perform shot segmentation. A shot is defined as a sequence of frames generated between the start and end of a continuous camera operation, and the main purpose of doing this is to simplify computational complexity in processing. But a shot may not correspond well to semantic meaning.

For example, when the image frame sequence in a typical soccer video in Fig.2 is segmented using color histogram into shots, the sequence will be divided into at least two segments due to the significant changes in the backgrounds between two consecutive frames. However as the sequence in Fig. 2 shows the successfully defend by the player, one would prefer to label them as one semantic segment.

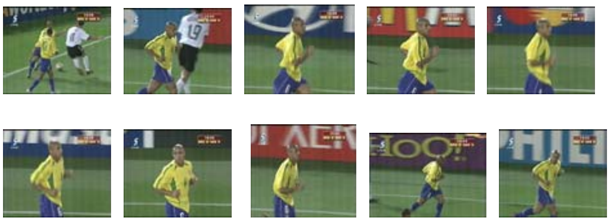

Fig. 2. A frame sequence showing an action of a player

Another example is shown in Fig.3. The whole shot includes three areas in the field: penalty boxes of both side and the area between them. Because most of important events such as shooting, scoring happened within or around penalty box, a sequence of frames including penalty box to show actions happened around it should be considered a semantic segment, which is different from a sequence of frames showing actions around the center circle, which should be regarded as another semantic segment. But traditional method for shot segmentation will not segment in this case. Hence we argue that shots are not the most appropriate semantic units in soccer video.

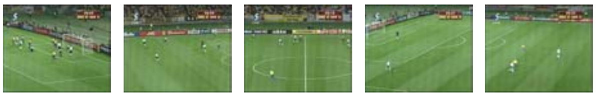

Fig.3. Frames from a shot to present three different areas in the field: two penalty boxes and area between them.

On the other hand, the authors of [6] defined some semantic labels for shots. But some of them are not consistent enough. For example, "Corner Kick" is rather considered as an event than a meaningful label for certain shot. Thus a consistent and comprehensive set of semantic labels is necessary for soccer video.

In this paper, our intent is to define a set of simple and atomic semantic labels called visual keywords for soccer video (Table 1). As an intermediate representation to bridge the semantic gap between low-level features and semantic understanding, these visual keywords can form the basis for event detection in soccer video [4]. Hence the objective of our system is to segment and classify a soccer video stream into semantic units labeled with visual keywords as defined in Table 1.

Table 1. The visual key words considered and the semantic meanings each one stands by

Words	Semantic meanings	Description
AD	Audience	Far view of audience
WA	Fast movement to a penalty box or Fight for ball control	Far view of whole field, active (goal post not visible)
WS	A break happened between two penalty boxes	Far view of whole field, static (goal post not visible)
HA	Move inside or outside a penalty box	Far view of half field, active (goal post visible)
HS	Players are waiting for free kick or corner kick or Break	Far view of half field, static (goal post visible)
MA	Actions such as chasing the ball between players	Mid-range view, active (whole body visible)
MS	Players are waiting for free kick or corner kick	Mid-range view, static (whole body visible)
CP	Close up	Close-up of a player, referee, coach, goalkeeper

2.2 View Types

In [1], 3 types of views: global, zoom-in and close-up are defined. We feel that they are not adequate for soccer video segmentation. We defined 4 view types according to camera shooting positions and ratio of field colors to non-field colors within one frame as shown in Figure 4, and the judging rules to discriminate type II from type III is shown in Fig.5.

We can see that there is almost no green color (field colors) in view Type I and colors are very rich. In view Type II, green colors can mainly be found only at the lower part of a frame (i.e. upper part has more number of colors). In view Type III, field colors are dominant for the whole frame. In view Type IV, the number of colors in certain region of a frame is more than its surrounding. In our system, color histogram, field colors (represented as green color) and non-field colors (represented as black color) are used to recognize them from each other. These 4 view types correspond to 4 different green / non-green frame types.

The mapping relationship between the VKWs and these four view types is shown below in the Fig. 4.

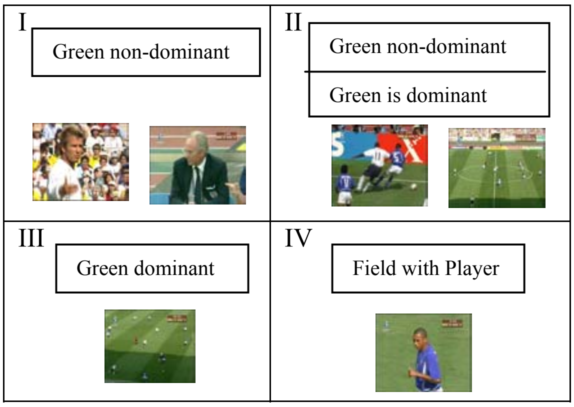

Fig. 4. Four field view types in soccer video

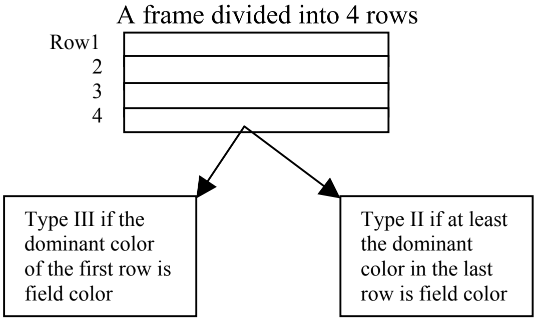
A frame divided into 4 rows
Row1
2
3
4
Type III if the dominant color of the first row is field color
Type II if at least the dominant color in the last row is field color

Fig. 5. The judging rules for discriminate the four types

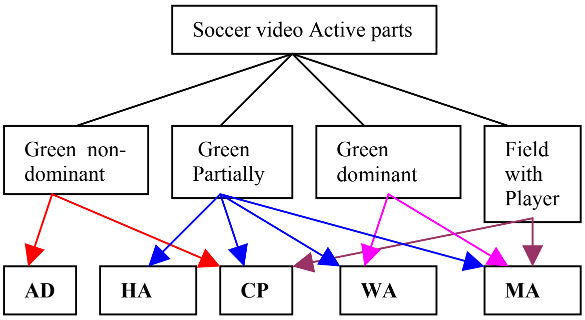

Fig. 6. The mapping relationship between four view types and VKWs

3. SEGMENTATION AND CLASSIFICATION

When extracting features, each P frame is divided into a 4 (rows) by 6 (columns) grid, each of which is called a block. Extracting color and motion features is done on both block level and frame level. We used two games of the FIFA World Cup 2002 (Germany versus Brazil and England versus Brazil) as test data (no replays and commercials).

3.1. Preprocessing

Discriminating field colors from others is not as easy as one may think about because the RGB values may change under different lighting and field conditions or different camera shooting positions. We design a method to solve this problem by forming three tables and the binarization of P frames to green / non-green frames. The formation of these 3 tables is discussed below and the binarization method is discussed later in Section 3.2.2.

First, a table called Green Color Table (GCT) is built manually. All colors perceived by people as field green colors are saved in this table. It is possible that some colors that are actually not field green are also kept in the GCT. Then some sample clips (from view Type II, III, IV) are input for training the system. For the color of a block (this color is in GCT), the system keeps it in the Upper Green Table (UGT) if this block is believed to be colored with field color and is within the upper half of a P frame; or keeps it in the Lower Green Table (LGT) if it is colored with a green color and is within the lower half of a P frame. In order to reduce effects of noise (field green colors could be found in audience too; also a field green color appears different under different camera shooting positions), the size of UGT (m) is set to be larger than that of LGT (n). In our experiments, $m = 11$ and $n = 6$.

3.2. Video Segmentation and Classification

A video stream is first divided into relatively static parts and active parts. For active parts, they are further segmented into sub-parts according to 4 view types using color histogram. Next for those sub-parts, motion features (means and standard deviations of magnitudes and angles of motion vectors at block level and distribution of motion directions) are used to classify them with help of SVM. After this phase, each sub-part is assigned with a Visual Key Word.

Static parts and Active parts:
For each P frame, sum of all motion vectors' magnitudes, *Mag*, is calculated. Setting a certain threshold, a video stream can be divided into relatively static parts and active parts (shown as 'A1' and 'A2' in Fig. 1). The motion features in a static part are ignored and the key frames extracted are considered as its representative. The threshold is determined empirically. In our system, the threshold is set to 60. Shown as 'D' in Fig. 1, static parts are processed again in post-processing phase.

Segmentation by view types and color histogram:
As mentioned above, the 4 view types correspond to 4 different green / non-green frame types. The system

binarizes each P frame according to the following method for all the 24 blocks:

1. Get the dominant color (C_d) of a block;
2. If C_d is in the upper half of a P frame, the block is converted to non-green unless its C_d is in the UGT. If so, it is converted to green color.
3. If C_d is in the lower half of a P frame, the block is converted to non-green unless its C_d is in the LGT. If so, it is converted to green color.

Then for each of the four rows of a P frame, the number of colors (except colors in UGT or LGT) is computed and the decision rules shown in Fig. 7 are used to do segmentation ('B' in Fig. 1).

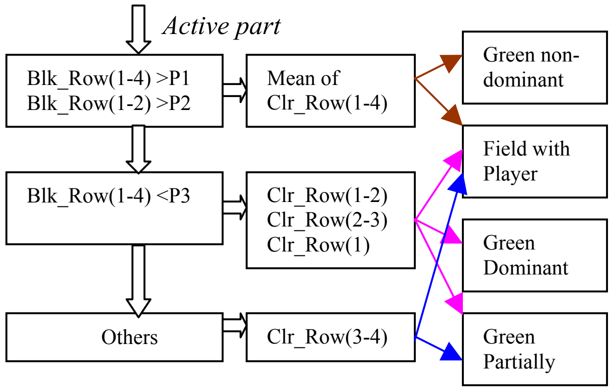

Fig. 7. Segmentation rules for 4 view types.

In Fig. 7, 'Blk_row (i-j)' is the number of blocks considered as colored with field colors from ith row to jth row; 'Clr-Row(m-n)' is the number of colors from mth row to nth row. Pi are parameters obtained from experiments (P1=16, P2=9, P3=6).

For testing purpose, we segmented the two soccer videos into sub-parts and labeled each sub-part according to the VKWs in Table 1 manually and used them as test data. Our experimental results (Table 2) show that green / non- green frames and color numbers are adequate to do segmentation, which can provide a good foundation for further classification.

Table 2 Results of view type classification

	Green non-dominant	Green Partially	Green Dominant	Field with Player
Test samples	80	102	30	64
Correct	76	95	28	60
Percent (%)	95.0	93.14	90.0	93.75

Classification by motion features:
When converting a P frame to a green/non-green frame, the motion features of each block of this frame are also extracted (shown as 'C' in Fig. 1). On the block level, the magnitudes of motion vectors are first mapped into 3 values if the magnitudes of a motion vector are non-zero and the value of the angle of each micro-block is mapped into 8 directions. Next the means and standard deviations of magnitudes and angles of motion vectors within the block are extracted. At the frame level, a direction frequency feature is extracted. That is, in order to describe motion, the direction distribution of all motion vectors within a frame is counted and kept. Motion texture proposed in [7] is a compact representation for motion. It can characterize 6 motions. In our system, the motion features used realize the same effect partially.

The clips from the first halves of the two soccer videos are used as training data and the second halves are used to test the classification by motion. Support vector machine ([5] with the 'multi-classify' option) is adopted as the classifier. Our results on 333 test segments (Table 3) show that our motion features are effective.

Table 3. Result of segment classification

View Types		Accuracy
Green non-dominant	AD / CP	85.7%
Green partially	WA/S / Others	79.1%
	HA/S / Others	81.2%
	MA/S / Others	70.1%
	CP / Others	73.0%
Green is dominant	WA/S / MA/S	93.1%
Field with Player	MA/S / CP	78.4%

From Table 3, we see that it is not easy to recognize MA/S segments from others, because the motion pattern between each two of 'WA/S MA/S' and 'HA/S MA/S' are not discriminative enough. And also, replays may effect the results. For example, given a frame showing a standing player in the field with lots of other players' legs at the upper part of this frame, it is possible to be labeled it as WA/S.

3.3. Post-processing

In this phase, shown as 'D' in Fig. 1, both relatively static parts and active sub-parts processed. Generally speaking, a static part may contain several meaningful sub-parts. So, the system first segments a static part by color histogram. In practice, we adjust the threshold so that a static part contains no more than two sub-parts. The last 5th frame of its left neighbor and 5th frame of its right neighbor are selected as their comparable references. The fifth frame of a static sub-part is extracted as its key frame. The differences between the key frame and comparable references are computed to decide which neighbor a sub-part is to merge with, if the difference is below a threshold. Otherwise, the sub-part is to be abandoned. As a result, segments labeled 'MA/S, 'HA/S', or 'WA/S', are divided into relatively active and static sub-segments by a threshold set manually.

4. EXPERIMENTAL RESULTS

As mentioned above, the segmentation and classification methods are shown to be effective. We used 4 10-minute clips (no replays and commercials) from the second halves of the 2 videos to test the whole system. The results are shown in Table 4.

Table 4. System result of segment classification

	Ground Truth	System Output	Correct	Accuracy
AD	13	13	11	84.6%
WA/S	109	101	79	78.2%
HA/S	56	50	40	80.0%
MA/S	112	104	79	76.0%
CP	71	66	53	80.3%

In Table 4, data in column 'Ground Truth' comes from observation on these 4 clips. Data in 'System Output' are the detection results from the system. Column 'Correct' shows cases that are both detected and classified successfully by the system. Similar to the analysis for Table 3, as the boundaries between each of 'WA/S MA/S' and 'HA/S MA/S' is not clear, recognition of MA/S are not as good as other semantic labels.

Note that there are only 5 classes shown in Table 4. In fact we can set a threshold for WA/S, HA/S and MA/S to differentiate WA from WS, HA from HS and MA from MS. Then we can obtain all the 8 classes. Since the threshold has no effect on classification accuracy,

classifying video stream into 5 classes is sufficient to illustrate the performance of our system, which is very encouraging.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented a novel method for segmenting and classifying soccer video segments using color and motion features. It will form the basis for further event detection in soccer video [4]. The video stream is first divided into relatively static parts and active parts. For active parts, they are segmented into four view types by using color features (green / non-green colors and color histogram). Then, for those sub-parts belonging to one view type, motion features are used to further classify them using SVM. In the post-processing phase, relatively static parts and active sub-parts are processed to produce final labeled segments.

As the system uses color features to segment relatively active parts and if the field colors of a game are very different from those in our test data, the results of segmentation by color will be worse, hence affecting the results of classification by motion features.

In future, we would focus on making the system more robust with more features, such as audio and texture. Last but not least, we would compare different motion representations for better motion features.

6. REFERENCES

[1] P. Xu et al., "Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video", IEEE International Conference on Multimedia and Expo, Tokyo, Japan, Aug, 2001.

[2] D. Zhong and S.F. Chang, "Structure Analysis of Sports Video Using Domain Models", IEEE Conference on Multimedia and Exhibition, Tokyo, Japan, Aug, 2001.

[3] G. Xu, Y.F. Ma, H.J. Zhang, S.Q. Yang, "Motion-Based Event Recognition Using HMM", IEEE Conference on ICPR' 2002.

[4] Y.L. Kang, J.H. Lim, Q. Tian, M. Kankanhalli, "Soccer Video Event Detection with Visual Keywords", *submitted to IEEE PCM'2003.*

[5]SVMTorch http://old-www.idiap.ch/learning/SVMTorch.html

[6] Ling-Yu Duang, Min Xu, Xiao Dong Yu, Qi Tian, "A unified framework for semantic shot classification in sports video", ACM Multimedia, Juan-les-Pins, France, December 2002.

[7] Y.F. Ma and H.J. Zhang, "Motion Texture: A New Motion Based Video Representation", IEEE Conference on ICPR' 2002.