NEWS SPORTS VIDEO SHOT CLASSIFICATION WITH SPORTS PLAY FIELD AND MOTION FEATURES

De-Hong Wang[1], *Qi Tian*[1], *Sheng Gao*[1], *Wing-Kin Sung*[2]
[1]Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
Email:{dehong, tian, gaosheng}@i2r.a-star.edu.sg
Department of Computer Science, School of Computing,
[2]National University of Singapore, Kent Ridge, Singapore 117543
Email:ksung@comp.nus.edu.sg

ABSTRACT

In this paper a novel sports news video shot classification method has been proposed. First two features based on motion and color are constructed and extracted from video shots: play field color ratio for specific types of sports, background motion and consistency ratio, then they are combined to generate an 11-dimension shot feature to feed into a C4.5 decision tree for shot classification. Based on our video data sets - the sports news video from the CNN Headline News video used in the TRECVID 2003, 7 predefined video shot classes were defined: 4 types of sports field video (basketball, baseball, ice hockey and golf), and sports news lead-in/lead-out, text, and others. Sports news video segments from 15 half-hour CNN News video were used for the training and testing. A performance of average precision and recall 88%, 82% has been achieved, respectively. The proposed method can be further developed and used to search news video for individual sports news and sports highlights.

1. INTRODUCTION

The extensive amount of multimedia necessitates content-based video indexing and retrieval methods. Sports video, due to rich spatial-temporal patterns and having tremendous commercial potentials, has been widely studied. However, published papers seldom cover sports news video classification. The challenges come from the content varieties of same sports and short clip length.

On the other hand, news story segmentation is one of the tasks of TREC 2003. A critical technique for story segmentation is anchor shots detection [1]. However in CNN headline news the sports segment is presented without anchor person coming out; so currently the entire sports segment was identified as a single story in each video [1], which is not helpful for searching specific sports news. Some methods need to be explored to segment the whole sports story into different clips, each clips just cover one kind of game. Similar with [1], in this paper, we assume that shots boundaries are ground truth, then conduct research on sports news shot classification.

Many works have been done for classifying shots extracted from one kind of sports [2-6]; these are quite different from sports news shot classification in which shots may come from various sports and also non-sports such as lead-in/out(denoted as LEDS), text caption and people talking.

Other people studied sports genre classification [7-10] i.e. classifying sports video file into some predefined classes. Considering that different sports often present different motion patterns, some researchers attempt to catch motion patterns from motion vector [6-8]. However their methods may not work when the length of the video clip is short since the motion pattern becomes unstable at that time. So some people think color may by more significant. Assfalg et al [10] classified sports video clip by color histogram of the shot keyframes. However, keyframes may not contain significant part of the field. Xavier et al [8] use dominant color of each frame to replace the color histogram. However, they only pick one color for each frame, so they may not differentiate sports with the same field color, for instance, golf and baseball. Moreover, the introduction of "do not care" color makes some shots such as pitching in baseball unclassifiable.

In addition, some of the above papers indicate that classification accuracy can be improved by filter non-field shots [7,10]. Assfalg et al [10] proposed to differentiate sports field shot with player shots and audience shots using edge features, Saur et al [7] proposed a method to differentiate wide-angle from close-up by computing camera motion parameter and intra-macroblock in a P frame. However their method may not work in classifying close-up shots because the former makes too strict background assumption for close-up shots and the latter assumes that camera moves in wide-angle shots.

In this paper we proposed two kinds of features to address the above problems. The first is three field color ratio namely yellow (for basketball), green (for baseball and golf), white (for ice hockey). The second is background motion and consistency motion ratio; the latter is the ratio of inner macroblock (MB) (Ref. Figure 3)

0-7803-8554-3/04/$20.00 ©2004 IEEE.

whose motion is consistent with the background. We compute these two features once for every four frames. Based on them, a 11-dimension feature vector is calculated for each shot; using them, a decision tree is employed to classify the sports news shots.

To demonstrate the effectiveness of our methods, we select TRECVID 2003 dataset to conduct our experiment. Basketball, ice hockey, baseball and golf, which compose of 90% of sports field shots in CNN headline news, are four predefined classes of field shots. Other three classes are: LEDS, which is the fixed introduction and ending of the sports news; text, which is text caption shots; and non-field sports shots including close-up to people, surroundings of the field and audience.

The rest of the paper is organized as follows: Section 2 describes feature extraction and classification. Experiment results and analysis are given in Section 3. Finally we conclude the paper and list future work in Section 4.

2. FEATURE EXTRACTION

2.1. Play field color ratios

For each sports type, the color of the playing field is either fixed or it varies in a small set of possibilities [10]. The most common field colors in sports video are green (golf, baseball, etc), brown/yellow (basketball, volleyball, etc), and white (ice hockey etc). Therefore, we define these three field colors at current stage. After frame decoding we conduct lighting compensation as [11], then quantize the 24-bit colors into standard 256 colors. By learning from some example areas, three color sets are determined, namely yellow set (denoted by Y), green set (G) and white set (W). In our experiment, there are 20 elements in Y, 12 in G, and 8 in W.

When a frame is quantized into 256 colors, representing it with $C(i, j)$; let $Y(i, j)$, $G(i, j)$, $W(i, j)$ denote the binary mark of yellow, green and white color respectively, where (i, j) is the coordinate of the pixels in the frame.

Since most of field shots are wide-angle or middle-angle shots, we can remove other non-field yellow pixels in field shot by conducting morphological operation. Figure 1 shows an example. Figure 1(a) is an original frame of a basketball game and Figure 1(b) shows the yellow binary mark before morphological operation.
The last result is given in Figure 1(c). Then the yellow field color ratio can be calculated as follows,

$$y\_ratio = \sum_{i=1}^{W}\sum_{j=1}^{H} Y(i, j)/(W * H),$$

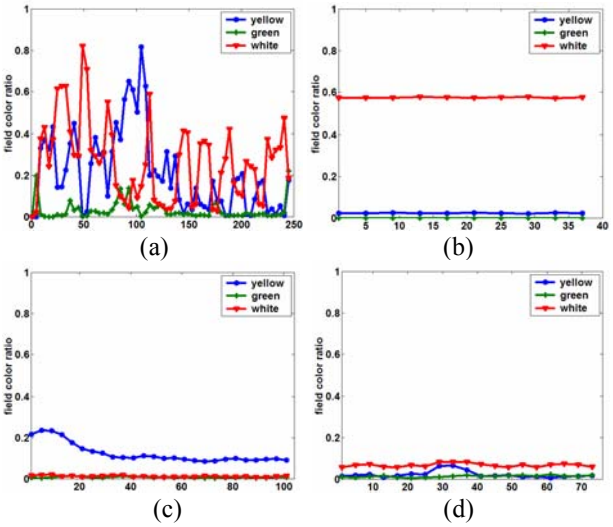
where W is the width of the frame, H is the height of the frame. Similarly, by replacing $Y(i,j)$ with $G(i,j)$ or $W(i,j)$, we can get green color ratio g_ratio and white color ratio w_ratio too.

(a) (b) (c)

Figure 1: Yellow mark extraction result, (a) Basketball frame (b) Initial yellow mark (c) Final yellow mark

Field Color ratio of a shot

Three field color ratios are computed once every 4 frames, let it be a column vector (3x1). So for a shot we can get a matrix (3xm), $m = \lfloor n/4 \rfloor$, where n is the number of frames in the shot. Figure 2 shows some typical examples of sports news shot. The ratios of LEDS shot vary greatly and frequently; on the contrary, the ratios of text are very stable. Yellow is the largest ratio color in basketball, but the value is about 0.15 when the field is small part of view. No ratio is great than 0.1 in some non-field shots.



(a) (b) (c) (d)

Figure 2 field color ratios of typical shot. X-axis is frame number (a) LEDS (b) text (c) basketball field (d) non- field shot

2.2. Background motion and ratio

We extract motion features from P frame in compressed MPEG video streams. We observed that in most of the close-up shots of sports news, movement of foreground is often quite different from that of background. Moreover, since the camera is very "near", even a slight movement also causes big differences in motion vector. On the other hand, to differentiate some baseball shots with golf shots, background motion is a critical factor. For example, passing ball in baseball field shot also has very high green ratio like most of the shots in golf. But, in order to track the ball, the camera in the baseball shot moves quickly which is quite rare in golf shots. Therefore, we proposed a method to extract a new motion feature as follows.

As Figure 3 illustrates, we partition macro blocks of a frame into 4 parts. Apart from the lowest 3 rows (Ads.

MB) which always show some non-relevant information in CNN headline news and the periphery macroblocks, the neighbor of the periphery macroblocks are background MB since most of the time, they belong to the background. The remains are inner MB.

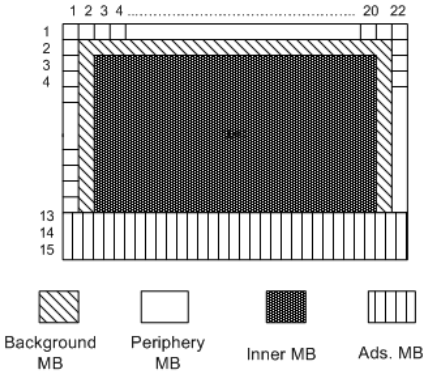


Figure 3 Macroblocks partition in a frame

Put all the background MB motion vectors into a set, discard the outliers, we compute the average of the remain motion vectors in two directions and denote them as the background motion vectors (mvx_{bg}, mvy_{bg}).

Then the motion consistency ratio of inner MB is computed as follows,

$$mv\_ratio = \sum_{i=1}^{W\_inn} \sum_{j=1}^{H\_inn} cons(i,j) / (W\_inn * H\_inn)$$

$$cons(i,j) = \begin{cases} 1 & if\ non-IntraMB\ and\ \ D < TH \\ 0 & otherwise \end{cases}$$

$$D = [mvx_{inn}(i,j) - mvx_{bg}]^2 + [mvy_{inn}(i,j) - mvy_{bg}]^2$$

where $mvx_{inn}(i,j)$ is the motion vector of an inner MB whose position is the ith row and the jth column of the inner MB area. Intra MB is a kind of macroblock which is not coded through motion compensation. *TH* is a threshold, which is set to be 8 in out experiment. *W_inn* and *H_inn* are the width and the height of the inner MB area in unit of macroblocks.

Background motion and ratio of a shot

Similar to color features, we can get motion features of a shot. Figure 4 shows some examples. Figure 4(a) and 4(b) is a shot of close-up of a talking man, sometime he turns his head during the conversation. So we can see that background motion is relatively small while consistency ratio is lower than 0.8 many times. 4(c) and 4(d) are two shots with high green ratio, 4(c) is baseball shot while 4(d) is golf shot. Obviously, the background motion values of baseball change more greatly than that of golf.

2.3. Compact shot features and classification

Although there are obvious patterns in the field color ratio curves and background motion and ratio curves, the features matrices are not suitable for shot classification since their size (column) changes greatly with the shot length and the dimension is still high. So we compute the characteristic parameters of each feature, then a compact feature vector can be figured out for each shot.

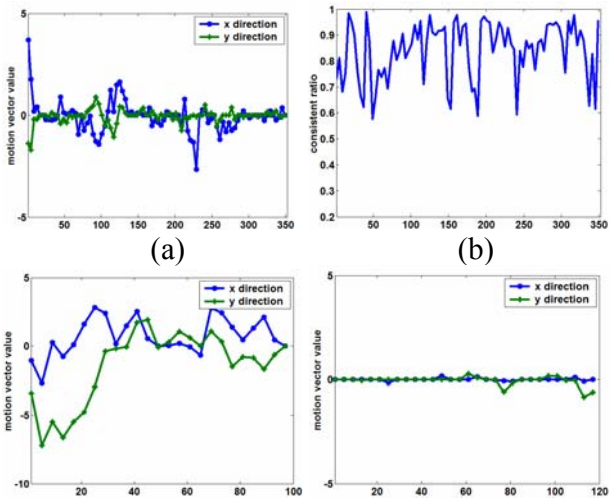


Figure 4 motion features of typical shot. X-axis is frame number. (a) and (b) a close-up of a talking man (c) baseball shot (passing ball) (d) golf

Let $y\_ratio(i)$ represents yellow ratio features of a shot, where $i = 1, ..., m$, the meaning of m is same as Section 2.1. We compute the mean and the variance of the yellow ratio features as follows:

$$m_y = \sum_{i=1}^{m} y\_ratio(i) / m$$

$$v_y = \sum_{i=1}^{m} (y\_ratio(i) - m_y)^2 / (m-1)$$

Similarly, the mean and the variance of green, white ratio features m_g, v_g m_w, v_w and that of background motion vector $m_{vx}, v_{vx}, m_{vy}, v_{vy}$ are also calculated.

Different from above features, for the last motion feature i. e. consistency ratio of inner MB, we do not compute its mean and variance. Observing that in close-up shots, this value is often below a threshold *THR*, we compute the compact features like following formula; where i, m is the same as above, $mv\_ratio(i)$ is the consistency ratio features of a shot. *THR* is set to 0.8 in our experiment.

$$n_{mvr}(i) = \begin{cases} 1 & mv\_ratio(i) > THR \\ 0 & otherwise \end{cases}$$

$$r_{mv} = \sum_{i=1}^{m} n_{mvr}(i) / m$$

Now for a shot, we get 11 features. These features are used to classify the shots by a C4.5 decision tree.

3. EXPERIMENT RESULTS

We use TRECVID 2003 video to conduct our experiment. The sports video segments are extracted from 15 days' CNN headline news from February to April 1998. The sports video shots in this dataset change greatly both in content and length. The length of the shots ranges from 1 second to 15 seconds. The basketball shots almost include all NBA fields and most of time only a part of field is yellow. The baseball shots cover many kind of baseball segments such as pitching, running to base, passing ball etc. Golf shots also contain shots from first stroke to pushing the ball into the hole, tracking the ball etc. The non-field shots cover the talking people, the surroundings of the sports field, close-up to people in the field or besides the field, wide-angle of audiences etc.

All shots are encoded with TMPEG, the GoP is set to IPPP. Color features are extracted from each I frame and motion features from the first P frame of each GoP.

We separate the segments into two parts, half of them for training and validation, while the other for testing, i.e. training and testing shots come from different segments. A C4.5 decision tree is selected as classification tools.

The results are given in Tables 1 and 2. Table 1 is a confusion matrix where each row shows the classification of ground truth. For example, the first row shows that only 1 LEDS shot is misclassified into a non-field shot (``Others" in the table) while 17 of them are correctly classified. Table 2 shows the classification performance, including Precision (denoted by P), Recall (denoted by R), and F-measure (denoted by F)

Table 1 Confusion matrix of test data

	L	T	O	Bk	Hk	Bs	G
LEDS	17	0	1	0	0	0	0
Text	0	11	0	0	0	0	0
Others	0	1	67	0	1	0	0
Basketball	0	0	0	29	0	1	0
Hockey	0	1	2	1	11	0	0
Baseball	1	0	6	1	0	16	1
Golf	0	0	3	0	0	0	3

Tabble 2 Performance of classification

Class	L	T	O	Bk	Hk	Bs	G
P	0.944	0.846	0.848	0.935	0.92	0.94	0.75
R	0.944	1.000	0.971	0.967	0.73	0.64	0.50
F	0.944	0.917	0.905	0.951	0.82	0.76	0.60

The tables show that the classification of LEDS, text, and basketball shots are quite good. However baseball and golf shots are easy mixed with "others" shots. We find that the typical miss-classified shots are zooming in to a player in the golf field and tracking baseball from field to audience. Obviously these shots cover both field and non-field segments. The macro precision is 0.883, macro recall is 0.822. Moreover, apart from non-field shots, the classification accuracy of field shots is very high.

4. CONCLUSIONS

In this paper we present the compact color and motion features for sports new shot classification. C4.5 decision tree is employed to classify the shots. Although the dataset (CNN headline sports news) is very challenge both in content and shot length, our method achieved promising result, the macro precision is 0.88, macro recall is 0.82. This work can be extended to segment and to classify sports news, especially suitable for non-sportscaster news story segmentation. In the future, we will explore more robust features to improved the performance and extend current work to more sports genre such as tennis, soccer, volleyball etc. and conduct sports news story segmentation.

REFERENCES

[1] Lekha Chaisorn, Tat-Seng Chua, Chin-Hui Lee: A Multi-Modal Approach to Story Segmentation for News Video. World Wide Web 6(2): 187-208 (2003)

[2] D. Zhong, S.-F. Chang, "Structure Analysis of Sports Video Using Domain Models," In Proc. of ICME 2001.

[3] C.W. Ngo, T.C. Pong, and H.J. Zhang, "On Clustering and Retrieval of Video Shots", In Proc. of ACM ultimedia 2001, pp. 51-60, 2001.

[4] Ling-Yu Duan, Min Xu, Tat-Seng Chua, Qi Tian et al, A mid-level representation framework for semantic sports video analysis. ACM Multimedia 2003: 33-44

[5] P. Xu, L. Xie, S.-F. Chang et al, Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video, In Proc. of ICME 2001

[6] D. D. Saur, Y. P. Tan, S. R. Kulkarni and P. J. Ramadge, ``Automated Analysis and Annotation of Basketball Video," SPIE Vol. 3022, Sep. 1997.

[7] E. Sahouria and A. Zakhor. Content analysis of video using principal components. IEEE Transactions on CSVT, 9(8):1290-1298, 1999.

[8] X. Gilbert, H. Li, and D. Doermann. Sports Video Classification Using HMM. ICME, pages 345-348, 2003

[9] Shinobu Hattori et al "A content based video classification semantic description extraction" SCI2003

[10] J. Assfalg, M. Bertini, C. Colombo, and A. D. Bimbo, "Semantic Annotation of Sports Videos," IEEE Multimedia 9(2): 52-60, 2002.

[11] Rein-Lien Hsu, Mohamed Abdel-Mottaleb, Anil K. Jain: Face Detection in Color Images. IEEE PAMI 24(5): 696-706 (2002)