Jersey number detection in sports video for athlete identification

Qixiang Ye[1,2*], Qingming Huang[2], Shuqiang Jiang[1,2], Yang Liu[3], Wen Gao[1,2]

[1]Institute of Computing Technology of Chinese Academy of Sciences, China
[2]Graduate School of Chinese Academy of Sciences, China
[3]Department of Computer Science, Harbin Institute of Technology China

ABSTRACT

Athlete identification is important for sport video content analysis since users often care about the video clips with their preferred athletes. In this paper, we propose a method for athlete identification by combing the segmentation, tracking and recognition procedures into a coarse-to-fine scheme for jersey number (digital characters on sport shirt) detection. Firstly, image segmentation is employed to separate the jersey number regions with its background. And size/pipe-like attributes of digital characters are used to filter out candidates. Then, a *K*-NN (*K* nearest neighbor) classifier is employed to classify a candidate into a digit in "0-9" or negative. In the recognition procedure, we use the Zernike moment features, which are invariant to rotation and scale for digital shape recognition. Synthetic training samples with different fonts are used to represent the pattern of digital characters with non-rigid deformation. Once a character candidate is detected, a SSD (smallest square distance)-based tracking procedure is started. The recognition procedure is performed every several frames in the tracking process. After tracking tens of frames, the overall recognition results are combined to determine if a candidate is a true jersey number or not by a voting procedure. Experiments on several types of sports video shows encouraging result.

Keywords: Jersey number, sports video, character recognition.

1. INTRODUCTION

Sports video analysis for summarization and retrieval purpose has emerged as a hot research area since the turn of the century. In the literatures, although many efforts have been made on sports video structure analysis and exciting events extraction[1-8], athlete identification is rarely investigated. However in real applications, users often care video clips about their preferred athletes, such as "all video clips about No.7-David Beckham". To identify an athlete in video clips, jersey number detection is an effective and precise way. Compared with athlete identification by other method liked face recognition, jersey number is more feasible since it is made of limited digital characters (from '0' to '9') and is general in even different types of sports videos. In this paper, we propose a scheme for athlete identification in sports video by detecting jersey number. To our best knowledge, this is the pioneer work on character detection on sports shirt for athlete identification task.

In the literatures, video overlay text detection and scene text detection can be seen as the most relevant works. In Wu et al.[9] use nine second-order Gaussian derivatives to extract vertical strokes in horizontal aligned text regions. Lienhart and Wernicke[10] locate text in images and video frames using the image gradient feature and a neural network classifier. Jain and Yu[11] propose a classical text detection algorithm based on connected component analysis. The method can detect text in web images, video frames and some document images such as newspapers. The probability of missing text is minimized at the cost of increasing false alarms. Li et al.[12] use mean, second- and third-order central moments in wavelet domain as the texture features and a neural network classifier is applied for text block detection. Gao et al.[13] combine the edge information with color layout analysis to detect scene text.

[*] Further author information: send correspondence to Qixiang Ye, Qingming Huang, Yang liu or Shuqiang Jiang by email: {qxye, qmhuang, sqjiang}@jdl.ac.cn; This work is a part of "SPISES: SPorts vIdeo Summarization and Enrichment System" at "http://www.jdl.ac.cn/en/project/spises/SPISES.htm".

Visual Communications and Image Processing 2005, edited by Shipeng Li,
Fernando Pereira, Heung-Yeung Shum, Andrew G. Tescher, Proc. of SPIE Vol. 5960
(SPIE, Bellingham, WA, 2005) · 0277-786X/05/$15 · doi: 10.1117/12.632735

From the works review we can see that either gradient or texture or structure can be use to discriminate text line with other things. Compared with the general text detection works, jersey number detection is much more difficult because there is no discriminative texture or structure information to discriminate it with the rest of the word. Variance arises from font, size, orientation and non-rigid deformation of digital characters can aggravate this problem. Furthermore, jersey number can even be blurred from athlete's fast motion or be occluded by other objects. Considering all these problems, we must use the feature combination method to discriminate digital characters with the rest of the world. Then segmentation, track and recognition procedures are combined in a coarse-to-fine scheme to perform the jersey number detection task. In the method, candidates are firstly located by filtering the regions obtained from image segmentation procedure with size and pipe-like attributes. In the procedure, spatial adjacent digital character regions are connected to form a 2-bit jersey number that is larger than '10'. Then the candidates are tracked in successive video frames with a modified SSD (smallest square distance) tracking algorithm. In the recognition procedure, to cope with elastic deformation problem, we use the synthetic training samples to simulate non-rigid deformation of jersey number. The recognition result will be feedback to the detection procedure to discriminate character and non-character patterns. In the tracking process, the candidates are recognized every several times by a *K*-NN (*K* nearest neighbor) classifier and the overall recognition result in the tracking process are combined to determine whether a candidate is a digital character or not . The flow chart of the proposed method is as follows.

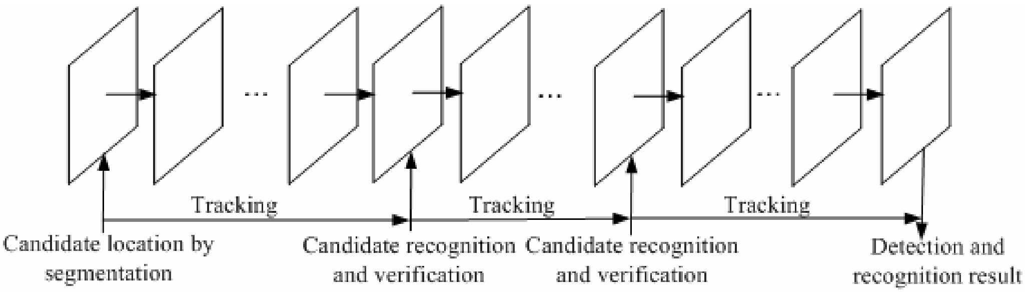

...
...
...
Tracking
Tracking
Tracking
Candidate location by segmentation
Candidate recognition and verification
Candidate recognition and verification
Detection and recognition result

Figure 1: Jersey number detection scheme.

The rest of the paper is organized as follows. Section 2 is the character location algorithm in video frame including candidate location, candidate filtering and recognition-based verification. We present the Jersey number tracking procedure in Section 3. Experimental results are presented in Section 4. The paper is concluded with a discussion of future work in Section 5.

2. JERSEY NUMBER DETECTION IN VIDEO FRAME

In this section, we present the jersey number detection algorithm in which candidate location, candidate filtering and recognition-based verification are described respectively.

2.1 Candidates location

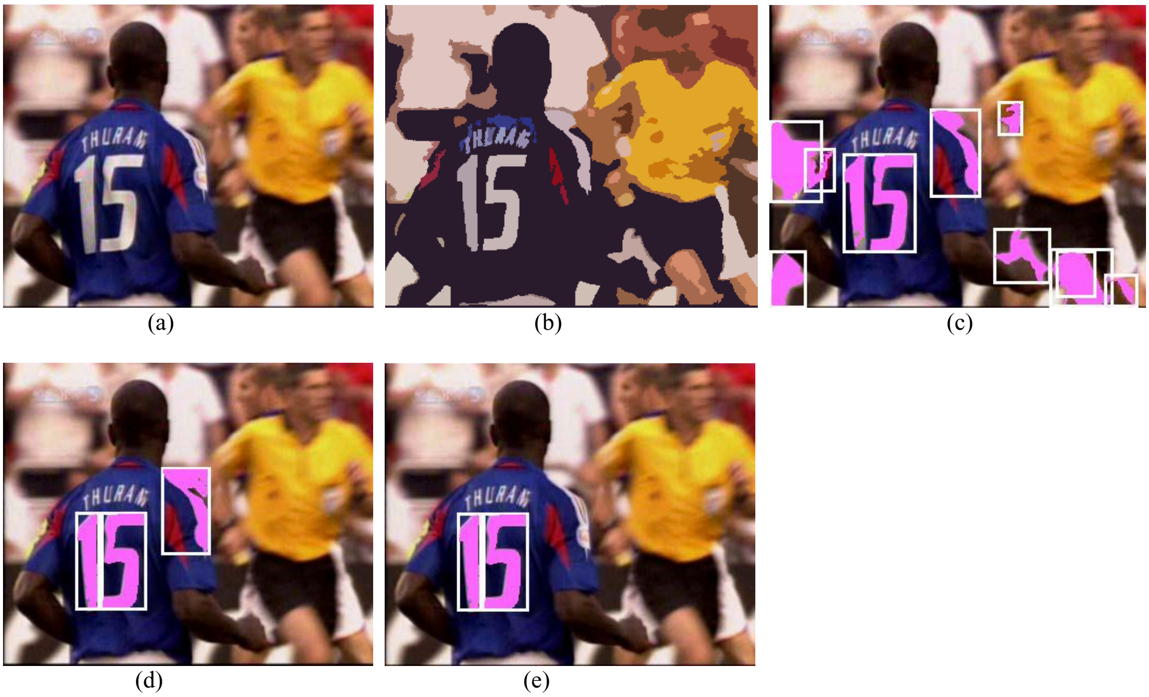
By observing that jersey number region must have a large color contrast with its background, image segmentation is first employed to separate the jersey number from its background. In the segmentation processing, we use GLVQ (Generalized learning vector quantization)[14] algorithm to clustering image pixels into limited color- homogeneous regions. Given the GLVQ algorithm, the primary problem is to decide the color cluster number N_{Quan}. In this paper, we proposed a method to decide N_{Quan} by color variance analysis in the whole video frame. Supposing that there are M $w_1 \times w_2$ window in a frame and each window contains n pixels, we define S_m as the color coarseness of a window, which represent the color variance of the window. Then we can use the average coarseness of a frame S_{avg} to calculate color quantization number. S_m and S_{avg} are calculated as follows.

$$S_m = \left(\frac{1}{n}\sum_{i=0}^{n-1}\left\|\vec{x}_i - \vec{x}_{mean}^{(m)}\right\|^2\right)^{\frac{1}{2}}, S_{avg} = \frac{1}{M}\sum_{m=0}^{M-1} S_m \tag{1}$$

where $\vec{x}_i$ is color values of pixel *i* in $w_1 \times w_2$. In experiments, we set w_1 and w_2 and the 1/10~1/20 of frame width and height. $\vec{x}_{mean}^{(m)}$ is the average color values in a window. $\|\cdot\|$ is Euclidian distance. The larger S_{avg} is the larger N_{Quan} should be. Then we use the following function to decide N_{Quan}.

$$N_{Quan} = \alpha \cdot S_{avg} + 1 \tag{2}$$

where α is a coefficients which is set as 0.5 in experiment. Once N_{Quan} is decided, we use GLVQ algorithm to clustering pixels into N_{Quan} color cluster. After the spatial connection analysis, the frame is segmented into color-homogeneous regions. Fig.2a is the result of segmentation.



(a)

(b)

(c)

(d)

(e)

Figure 2: Jersey number detection and recognition process. (a) original image, (b) segmentation result, (c) candidates after size verification, (d) candidates after PLC verification (e) result after recognition feedback.

The segmented regions are presented by $R_i\{position\,(l_i, r_i, b_i, t_i), size\,(w_i, h_i), color\,(\vec{c}_i), shape\,(\vec{z}_i)\}$, where $position\ (l_i, r_i, b_i, t_i)$ is the position attribute and l_i, r_i, b_i, t_i are the left, right, bottom and top of the outline rectangle respectively. $size(w_i, h_i)$ is the size attribute of R_i and w_i, h_i are the height and width of the outline rectangle respectively (as shown in Fig. 2a). $color\ (\vec{c}_i)$ is the mean color vector of R_i in LUV color space, As a 2-bit jersey number region may be split, two regions (R_i, R_j) will be merged to form a 2-bit jersey number if they are adjacent in spatial with $color\ (\vec{c}_i)$ - $color\ (\vec{c}_j) < t_c$. $shape(\vec{z}_i)$ is the shape attribute represented by Zernike moments feature.

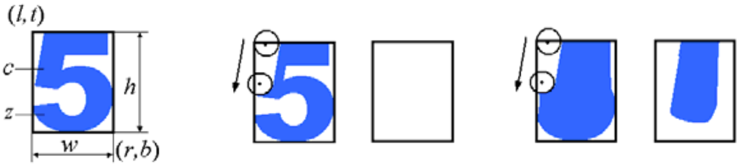
2.2 Candidates filtering

For R_i, $size(w_i, h_i)$ is firstly used to eliminate apparent false alarms. Regions whose width/height values are too small/large to be jersey number are firstly discarded. The regions whose area (compared with the outline rectangle area) is too small to be digital characters are also discarded. In experiments, the regions whose width/height/area whose value is smaller than 0.01/0.01/0.001 frame width/height/area are set as small regions.

Then, morphological "open" operation is used to capture the pipe-like attribute of jersey number. By observing that all digital characters are made of pipe-like components (PLC) (Fig.2a) and the PLC can be eaten off by an "open" operation of morphological methods. After the "open" operation, if R_i has no foreground pixels (Fig. 2b), we consider that it's made of PLC and then kept it as candidate. Otherwise it is regarded as a negative and removed from the candidates. The radius for the "open" operator is calculated by

$$radius_d = t_d \cdot h_i \tag{3}$$

where t_d is selected as 0.2 in the experiments empirically. Fig.3d is a result after PLC verification.



(a) (b) (c)

Figure 3: Jersey number's attributes. (a) is the size attribute, (b) is the "open" operation on a jersey number to verify pipe-like attribute and (c) "open" operation on a non-digit.

2.3 Recognition-based verification

After the PLC verification, there may still exist false alarms (as shown in Fig.2d) since all pipe-like regions can be detected as digital character. The feedback from the recognition procedure is used to reduce these false alarms. In the recognition procedure, the scale, rotation, character font, and non-rigid deformation problems are need to be considered.

To solve the rotation and scale problems, we employ Zernike moments[15], which is invariant to both rotation and scale, as the primary features for R_i, which is represented by $shape(\vec{z}_i)$. $shape(\vec{z}_i)$ contains 60 coefficients with 5 radius in 12 orientations. 13 kinds of representative fonts (including "Times New Roman", "Arial" etc.) are selected to represent digital characters in all fonts.

The most challenge problem in jersey number detection is the non-rigid deformation which is difficult to formulate even by a very complex mathematic function. In this paper, a simple but effective method is developed to cope with this problem by producing synthetic training samples with some representative non-rigid deformation simulation. Printed characters are stuck on a cloth and deformed by a fanner. Then training samples of similar deformation with digital characters in real conditions are produced. For each font, man-made training samples are clustered by a K-mean algorithm into 50 clusters and the centers are selected as positive training samples. Then we obtain 650 (13×50) training samples for each of the digits (0-9). Nearly ten thousands of pipe-like negative samples are selected from real video frames to represent non-digit patterns. Some positive and negative samples are illustrated in Fig.4.

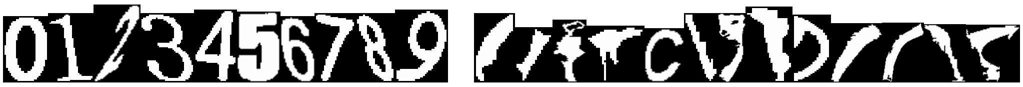


Figure 4: Examples of man-made positive samples for ten classes and negative samples.

A K-NN classifier is employed to classify a segmented region into one of the 11 classes based on the idea that similar observations belong to similar classes. We do not use complex classifiers like SVM, Neural network etc. in the classification process because that these classifiers maybe cannot work well when the feature presentation is sparse and inconstant. While *K*-NN classifier can work relatively stable in these conditions. Street distance is selected in the classification procedure for its better performance than that of the Euclidian distance and Vector angle distance by experiments.

3. CHARACTER TRACKING

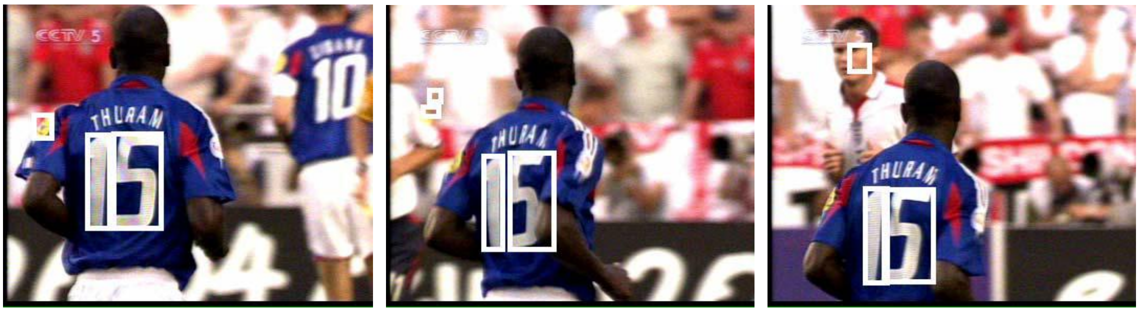
Once a character candidate is detected, the tracking procedure is started. Following the method in[12], we use a simple but effective SSD (smallest square distance)-based tracking algorithm. Character motion is a quite complex process including simple motion (linear motion) and complex motion (zooming in/out, rotation and elastic deformation). Our goal is to design a scheme to efficiently track text with both simple and complex motions.

The SSD-based tracking is based on image region information. Supposing that the original position of a character outline rectangle R is $\vec{p}$, as either the character move or deform the matched position of this character can be coarsely obtained by

$$\vec{p}^{\,'} = D \cdot \vec{p} + \vec{d} \cdot \tag{4}$$

In (4) D may be a complex matrix which contains the zoom in/out of the R and elastic deformation. $\vec{d}$ is a displacement vector in image plane. We define D as "the $position\ (l_i, r_i, b_i, t_i)$ of a region will be refined as $position^{'}(l_i - \delta, r + \delta_i, b_i - \delta, t_i + \delta)$ frame by frame, and then the pixels in the $position^{'}(l_i - \delta, r + \delta_i, b_i - \delta, t_i + \delta)$ will be re-segmented to get a new region candidate $R_i^{'}$." To segment each candidate outline rectangle, we use the color center of the original character as a criteria to detect foreground pixels. The pixels whose Euclidian distance to the color center is smaller that a threshold t_c will be set as foreground pixel to form the new region. And other pixels are set as background. The outline rectangle of the new region is set as a matched rectangle. This procedure will make the complex motion of a character region be simplified. After processing the complex motion, the algorithm will search around the regions in next frame to find a region of the smallest square distance.

In the tracking procedure, recognition is performed every 5 frames. The recognition will be performed on the re-segmented foreground regions. After tracking tens of frames, the overall recognition results are combined to determine whether a candidate is a true jersey number or not by a voting scheme. Some candidates recognized as characters in few frames will be discarded as false alarms. The integration of recognition result in successive frames further reduces the false alarms.

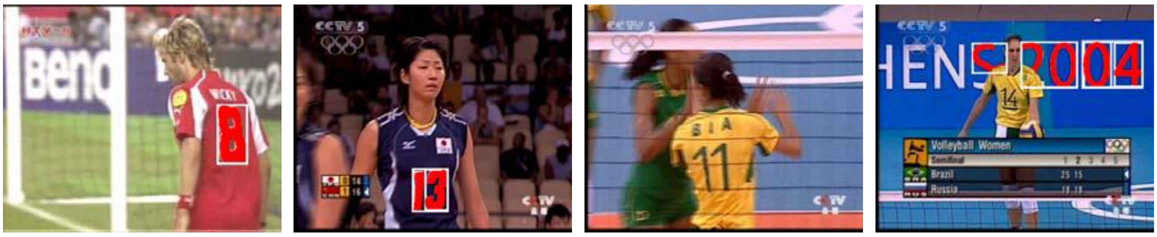


10th frame 31th frame 51th frame

Figure 5: Examples of tracking process.

4. EXPERERIMENTAL RESULTS

We prepare a dataset containing 200 video frames captured from soccer, basketball and volleyball games. The dataset contains jersey number in various conditions including illumination change, elastic deformation and complex background. In Fig.6, we illustrate some detection results. Despite of some failure examples, encouraging result is obtained. In Fig.6a, a jersey number “8” of large elastic deformation is detected and recognized correctly. Characters “1” and “3” are detected in Fig.6b to form the 2-bits jersey number “13”. In Fig.6c, a jersey number is missed because the color clustering algorithm cannot correctly separately the character regions with its background when there boundaries are blurred with the move of players or camera. In Fig.6d, there are some false alarms for that the method cannot discriminate jersey number or general digital characters in scenes. The true jersey number is missed for the same reason as in Fig.6c.



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

Figure 6: Examples of detection results.

Precision and *Recall* rate are used to quantitatively evaluate the experiments. They are defined as

$$Recall = \frac{Number\ of\ correctly\ detected\ jersey\ number}{Number\ of\ jersey\ number}, \tag{5}$$

$$Precision = \frac{Number\ of\ correctly\ detected\ jersey\ number}{Number\ of\ detected\ jersey\ number}. \tag{6}$$

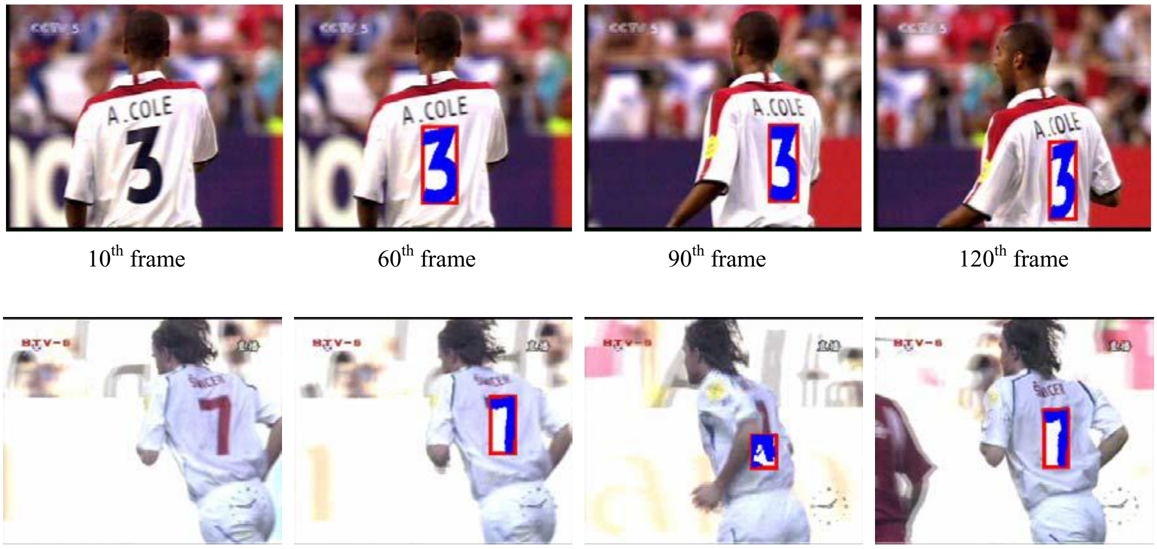
Given the testset, the detection performance is given in the first line of table 1.

Table 1: Detection performance in video frames and video clips.

Data	Recall	Precision	Speed on Pentium IV CPU (frames/s)
200 video frames	62.0%	83.8%	0.8
30 video clips	76.7%	86.9%	18

It can be seen from the result that in video frames, the performance of jersey number detection is not quite good. Then temporal information integration will improve the performance.

For the text tracking experiment, quantitative evaluation of tracking accuracy is not easy because of the lack of groundtruth data. We just evaluate the tracking algorithm by subjectively observation. Some of the tracking results are given in the following figure.



10th frame 60th frame 90th frame 120th frame

10th frame 51th frame 81th frame 111th frame

Figure 7: Experiments of character tracking process.

From the examples we can see that the tracking procedure is effective even for characters of complex motion. In the first line of example, the character “3” is well tracked even it has deep rotation transformation. In the second line, the character “7” is missed at 81th frame because the character is partly occluded. In the 37th frame, the character is matched again. Since the recognition is performed on lots of tracking frames, missing in few frames will not affect the final recognition and detection performance.

5. CONCLUSION AND FUTURE WORKS

A method for jersey number detection for athlete identification in sports video is proposed in this paper. The combination of color, size and shape features has made the jersey number detection task possible. The method can be used to help identify player in sports video automatically and then contribute to the sports video content analysis.

In spite of the present work so far, the detection performance of the proposed method need to be further improved. The main reason that a jersey number is missed is that the candidate location algorithm cannot locate it. That is for the reason that present color clustering algorithm cannot separate the characters from its background well when there is illumination variance or drape in the character region. In the future work, more effective image segmentation algorithm should be integrated to improve the candidate location performance.

ACKNOWLEDGEMENT

This work is partly supported by NEC Research China on “Context-based Multimedia Analysis and Retrieval Program” and “Science 100 Plan” of Chinese Academy of Sciences.

REFERENCE

1. A. Hanjalic, Generic approach to highlights extraction from a sport video,” in *Proc*. International Conference on Image Processing, 2003.
2. L.Y. Duan, M. Xu, T. Chua, Q. Tian, C.S Xu, “A mid-level representation framework for semantic sports video analysis,” ACM Multimedia Conference, 2003.

3. N. Babaguchi, Y.Kawai, T. Kitahashi, "Event based indexing of broadcasted sports video by intermodal collaboration," IEEE Transactions on Multimedia, Vol.4, No.1, pp.68-75, March 2002.
4. Y. Gong, T.S. Lim, and H.C. Chua, "Automatic parsing of TV soccer programs," in *Proc.* IEEE International Conference on Multimedia Computing and Systems, May, 1995.
5. 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. 24, Issue 15, December 2004.
6. Ekin, A. Murat Tekalp, and R. Mehrotra, "Automatic soccer video analysis and summarization," IEEE Transactions on Image Process, Vol.12, Issue.7, July 2003.
7. R.Leonardi, P.Migliorati, and M.Prandini,"Semantic indexing of soccer audio-visual sequences: a multimodal approach based on controlled Markov chains," IEEE Transactions on Circuit and System for Video Technology Vol.14, No.5. May 2004.
8. J. Assfalg, M. Bertini, C. Colombo, A. D. Bimbo, W.r Nunziati, "Semantic annotation of soccer videos: automatic highlights identification," Computer Vision and Image Understanding, Special Issue on Video Retrieval and Summarization, Vol.92,Issue 2/3, November/ December 2003.
9. V. Wu, R. Manmatha, and E. M. Riseman, "Textfinder: an automatic system to detect and recognize text in images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20 pp.1224-1229, 1999.
10. R. Lienhart and A. Wernicke, "Localizing and segmenting text in images and videos," IEEE Transactions on Circuits and Systems for Video Technology, Vol.12, pp.256-268, 2002.
11. A.K. Jain and B. Yu, "Automatic text location in images and video frames," Pattern Recognition. Vol.3, pp.2055-2076, 1998.
12. H. Li, D. Doermann, and Omid Kia, "Automatic text detection and tracking in digital video," IEEE Transactions on Image Processing, Vol.9 pp.147-156, 2000.
13. J.Gao, J.Yang, "An adaptive algorithm for text detection from natural scenes," in *Proc.* of the 2001 IEEE conference on Computer Vision and Pattern Recognition, December 2001.
14. A. Sato, K. Yamada, "Generalized learning vector quantization," Advances in neural information processing system 8, MIT Press, Cambridge, MA, pp. 423-429. 1996.
15. A. Khotanzad, Y.H. Hong, "Invariant image recognition by Zernike moments," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.12, pp489-497, May 1990.