Scorebox Extraction from Mobile Sports Videos using Support Vector Machines

Wonjun Kim, Jimin Park, and Changick Kim*
School of Engineering, Information and Communications University,
119 Munji street, Yuseong-gu, Daejeon, 305-714, Republic of Korea

ABSTRACT

Scorebox plays an important role in understanding contents of sports videos. However, the tiny scorebox may give the small-display-viewers uncomfortable experience in grasping the game situation. In this paper, we propose a novel framework to extract the scorebox from sports video frames. We first extract candidates by using accumulated intensity and edge information after short learning period. Since there are various types of scoreboxes inserted in sports videos, multiple attributes need to be used for efficient extraction. Based on those attributes, the optimal information gain is computed and top three ranked attributes in terms of information gain are selected as a three-dimensional feature vector for Support Vector Machines (SVM) to distinguish the scorebox from other candidates, such as logos and advertisement boards. The proposed method is tested on various videos of sports games and experimental results show the efficiency and robustness of our proposed method.

Keywords: Mobile device, Scorebox, Information gain, SVM

1. INTRODUCTION

With rapidly increasing multimedia mobile services, e.g., DVB-H or DMB, sports video is one of the most popular contents consumed by small mobile device users. In sports videos, scorebox has an important role by delivering time and score information to viewers. However, since most of serviced videos are directly resized to fit small screen due to cost reasons, it may deliver the viewers uncomfortable experiences in understanding contents. Therefore, it is thought that extracting and magnifying the scorebox in sports videos may help deliver better viewing experience to small mobile devices users. To this end, it is essential to design an efficient algorithm to extract the scorebox from similar candidates such as logos and advertisement boards regardless of various font size, style, color as well as position of the scorebox.

While there have been many research papers for analysis of sports video, especially soccer video, little work has been done for scorebox extraction. There are some important characteristics of scorebox. The scorebox is a closed caption region distinguished from the surrounding region. It generally appears at the same place in a scene throughout entire video. Since it is a superimposed graphic region, relatively saturated colors tend to be used with a small standard deviation [1]. Moreover, it is observed that contrast within the scorebox is very high.

Most of previous methods are based on these properties. Huang *et al*. [2] assume that a scorebox often appears at the bottom part of scene for a while and then disappears. They check the difference of intensity at the bottom part of the adjacent frames to detect the scorebox appearance. In [3], authors examine the change of the color distribution in horizontal direction and then find the difference between the ground and caption region. Texture-based approaches have been also used to detect the scorebox. In [4], authors consider a textual description of what is said by the commentator during a match. They use the informative keywords, such as scores, logo, 1-0, 1-2, and so on. Chen *et al*. [5] basically use the DCT coefficient, especially AC energy, to detect the caption region. The above methods perform well as far as there is only one caption in a scene and it is regarded as a scorebox. In this sense, the proposed method in [6] can be a good solution. They treat the caption box based on the assumption that there is high color correlation and low motion activities within the caption box. A caption template is generated by using fonts and digits information. These days, it is observed that content providers tend to insert more than one caption. Then the challenge is to distinguish the scorebox from other captions which is not considered in the existing methods.

* ckim@icu.ac.kr; phone 82 42 866-6168; fax 82 42 866-6245; vega.icu.ac.kr/~ckim

Applications of Digital Image Processing XXXI, edited by Andrew G. Tescher,
Proc. of SPIE Vol. 7073, 70730P, (2008) · 0277-786X/08/$18 · doi: 10.1117/12.797775

Unlike previous work, we use various attributes of scorebox to compute the optimal information gain and top three ranked attributes in terms of information gain are selected as a three-dimensional feature vector for SVM to distinguish the scorebox from other candidates. The rest of this paper is organized as follows: We present the socrebox determination by using the information gain and SVM classifier in Section 2. The experimental results on various sports videos are presented in Section 3, followed by conclusion in Section 4.

2. PROPOSED METHOD

Our goal is to extract the scorebox from other candidates in a scene. The proposed method can be divided into two phase, which are learning and detecting phases. The overall procedure is shown in Fig. 1. Each module in Fig. 1 is explained in the following subsections.

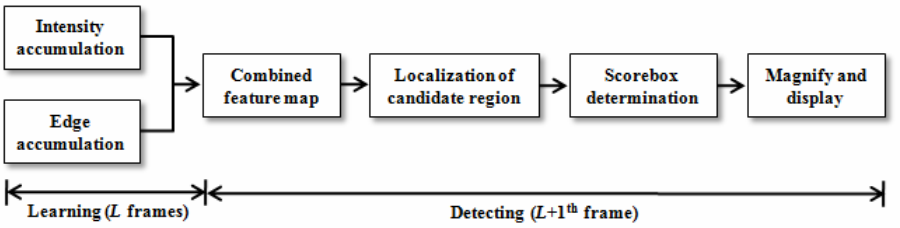


Fig. 1. The overall procedure for the scorebox extraction

2.1. Combined feature map

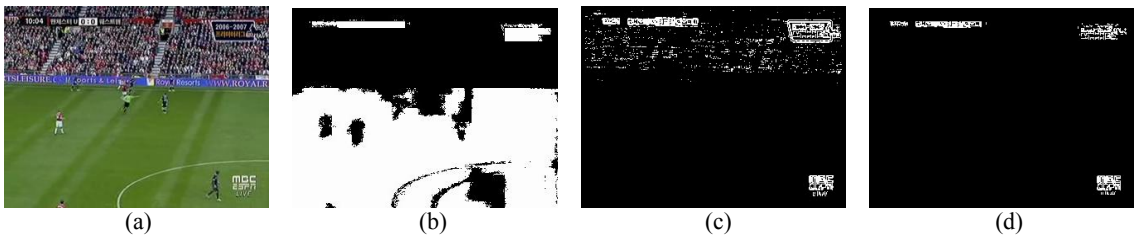
The combined feature map is generated using collected intensity and edge information during the learning period. Since the scorebox appears at the same place through a video sequence, the pixel value in the scorebox would barely change along with time. We define the accumulated intensity map, *AIM* as follows: If the difference of intensity between frames is smaller than 5 at the pixel (x, y), the weight value of pixel (x, y) is increased by one. After the learning period, if the accumulated weight value is larger than threshold value, *AIM* at (x, y) is set to one. Otherwise, it is set to zero.

However, since sports videos such as soccer and tennis include static background, edge information is also taken into account in our method. Edge information can be easily obtained by convolving an image with the Sobel mask. To binarize an edge map at each frame during learning period, Otsu's method [7] is used before accumulated edge map generation. We define the accumulated edge map *AEM* as follows: If a pixel value of edge map is same as previous one at the same position, the weight value for the pixel is increased by one. After the learning period, if the accumulated weight value is larger than a threshold value, $AEM(x, y)$ is set to one.

Finally, we generate the combined feature map by comparing two maps. Let *CFM* denote the combined feature map and is defined as follows.

$$CFM(x,y) = \begin{cases} 1, & if\ AIM(x,y) = AEM(x,y) = 1 \\ 0, & otherwise \end{cases} \quad (1)$$

The result of combined feature map is shown in Fig. 2. There are three captions in the image. Since the ground region tends to have zero edge value, we can easily extract caption regions by using combined feature map.

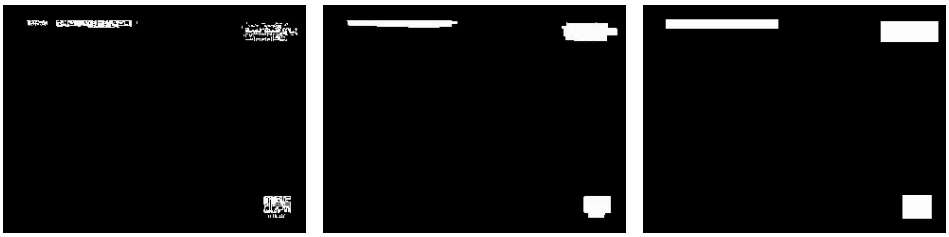


(a) (b) (c) (d)

Fig. 2. (a) Soccer video. (b) Accumulated intensity map. (c) Accumulated edge map. (d) Combined feature map.

2.2. Candidate region localization

The combined feature map can serve as a useful indicator for scorebox extraction. To generate the connected components, we first generate a linked map as shown in Fig. 3-(b). If a gab between two non-zero points in a same row is shorter than 20, it is filled with 1. We eliminate generated connected components, if it is smaller than a pre-defined value based on the observation of minimum size of scoreboxes. Then each connected component is reshaped to have smooth boundaries. Since it is reasonable to assume that the caption is generally in rectangular shapes, a rectangular box is generated by linking four points, which correspond to (min_x, min_y), (max_x, min_y), (min_x, max_y), (max_x, max_y) taken from the link map shown in Fig. 3-(b). The refined caption regions are shown in Fig. 3-(c).



(a) (b) (c)

Fig. 3. (a) Combined feature map. (b) Linked map. (c) Smoothed caption map.

2.3. Scorebox determination

In this subsection, we present a method to distinguish the scorebox from other captions. First, the optimal information gain is computed for each extracted caption based on eight attributes of the scorebox. Then the top three ranked attributes in terms of information gain are selected to construct a three-dimensional feature vector for SVM. The features are used as inputs to SVM based on Radial Based Function (RBF) kernel to distinguish the scorebox from other candidates, such as logos and advertisement boards. Each step will be explained in detail as follows.

2.3.1. Feature vector selection by information gain

To select the scorebox among candidates, useful attributes of the scorebox need to be investigated. In our proposed method, we investigate eight attributes, which are the average of hue, saturation, intensity, ratio of width and height, size, position, density of edge, and magnitude of wavelet coefficient. Color and geometry as well as texture information of scorebox are expected to be reflected well by using these attributes. To find the relevance of the attributes for determining the scorebox, the concept of information gain is employed [8]. First, we determine the value of each attribute. Then the information gain on the corresponding attribute is obtained using computed value. The workflow for the scorebox extraction is shown in Fig. 4. The best attribute value for classifying candidates can be determined efficiently by the information gain.

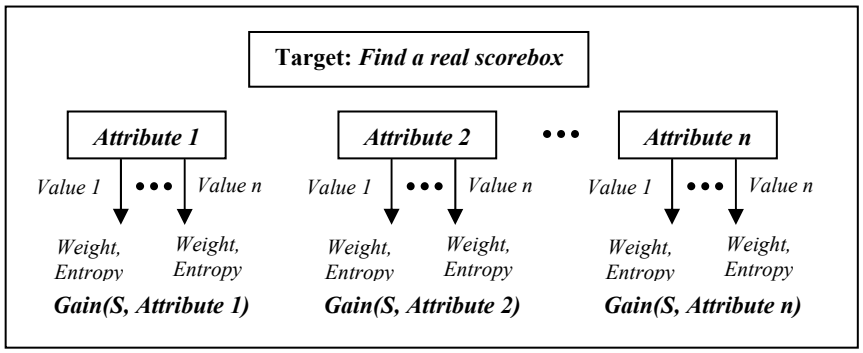


Fig. 4. The concept of determination process based on information gain

The detailed explanation of Fig. 4 is as follows. First, we determine appropriate values for each attribute. In our method, two or three values are used to represent each attribute. For example, the average of hue is assigned into one of three values whereas the average of saturation and intensity are assigned into one of two values. All the values used in our method are shown in Table 1.

Table 1. Value assignments for each attribute

Attribute	Range for values	Value
Average of hue (*AOH*)	0 ~ 0.17 OR 0.83 ~ 1 0.17 ~ 0.5 0.5 ~ 0.83	Reddish Greenish Blueish
Average of saturation (*AOS*)	0 ~ 0.2 0.2 ~ 1	Diluted Pure
Average of intensity (*AOI*)	0 ~ 0.4 0.4 ~ 1	Dark Bright
Ratio of width and height (*RWH*)	0 ~ 0.25 0.25 ~ 1	Short Long
Size (*SIZE*)	0 ~ 0.15 0.15 ~ 1	Small Large
Position (*POS*)	0.1, 0.25, 0.4 0.55, 0.7, 0.85	LT, RT, T LB, RB, B
Density of edge (*DOE*)	0 ~ 0.6 0.6 ~ 1	Simple Complex
Magnitude of wavelet coeff. (*MWC*)	0 ~ 0.07 0.07 ~ 1	Low High

All the attributes are normalized between 0 and 1 using the maximum value in each attribute. Value of position is determined using two points, which are (min_x, min_y) and (max_x, max_y). Video scene is divided into four equal-sized blocks. If both points belong to left-top part, the value of position is left-top (LT). If one belongs to left-bottom and the other belongs to right-bottom, the value of position is bottom (B). The density of edge can be obtained from the ratio between the number of edges and size of each caption. Magnitude of wavelet coefficient is computed by root of summation absolute value of horizontal and vertical coefficients of Haar transform as follows,

$$MWC_i = \frac{\sum_{j \in C_i} \sqrt{HL_j^{\,2} + LH_j^{\,2}}}{128 \times S_i}, \; 1 \le i \le P, \tag{2}$$

where *HL* and *LH* denote the horizontal and vertical coefficients of Haar transform in caption *C*, respectively. *S* and *P* denote the size of each caption and the number of extracted captions.

We calculate all the attributes from 35 captions, which can be regarded as training examples. Various types of caption inserted in the sports videos are used in our method. The result is shown in Table 2. In the tenth column, the identity of the caption is provided as SB for scorebox and NSB for others by visual observation. Note that values for each attribute are determined by using Table 1. Based on the Table 2, we can calculate the information gain. The entropy is firstly defined as below,

$$Entropy(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-, \tag{3}$$

where *S* denotes the set of extracted captions. p_+ and p_- denote the probability of positive and negative examples of target concept, respectively. The information gain can be represented by using entropy as follows,

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v), \tag{4}$$

where A denotes the attribute (i.e., *AOH*, *AOS*, *AOI*, *RWH*, *SIZE*, *POS*, *DOE*, and *MWC*) and S_v denotes the subset of S for which attribute A has value v. The weight is defined as $\frac{|S_v|}{|S|}$. *Values*(A) is the set of all possible values for attribute A. In other words, *Values*(A) can be represented as *Values*(A) = {Short, Long} for attribute *RWH* and values of other attributes can also be represented in the same way. We calculate the information gain in each caption. The information gain for each attribute is shown in Table 3. The computing process of information gain for *RWH* is also shown in (5). We can see that the top three ranked attributes are *RWH*, *POS*, and *MWC*. A three-dimensional feature vector composed of the selected attributes is put into SVM classifier.

Table 2. Training examples for the target concept *SCOREBOX*

Training caption	*AOH*	*AOS*	*AOI*	*RWH*	*SIZE*	*POS*	*DOE*	*MWC*	Identity
1	Reddish	Diluted	Bright	Short	Large	RT	Complex	High	NSB
2	Reddish	Diluted	Dark	Long	Large	LT	Complex	High	SB
3	Blueish	Diluted	Dark	Long	Large	T	Simple	Low	SB
4	Greenish	Diluted	Dark	Long	Large	LT	Simple	Low	SB
5	Greenish	Diluted	Dark	Short	Large	RT	Simple	High	NSB
6	Greenish	Pure	Bright	Short	Small	RB	Complex	High	NSB
7	Reddish	Pure	Dark	Long	Small	LT	Complex	High	SB
8	Greenish	Diluted	Bright	Short	Large	LB	Complex	High	NSB
9	Greenish	Diluted	Dark	Long	Large	LT	Simple	Low	SB
10	Greenish	Pure	Dark	Short	Large	RT	Simple	Low	NSB
11	Reddish	Diluted	Dark	Long	Large	LB	Complex	High	SB
12	Reddish	Diluted	Bright	Short	Large	LT	Complex	High	NSB
13	Greenish	Diluted	Dark	Long	Large	RT	Complex	Low	SB
14	Blueish	Diluted	Dark	Short	Large	RT	Complex	High	NSB
15	Reddish	Diluted	Dark	Long	Large	LT	Simple	Low	SB
16	Greenish	Diluted	Bright	Short	Small	RB	Complex	High	NSB
17	Blueish	Diluted	Dark	Short	Small	LT	Simple	Low	NSB
18	Greenish	Diluted	Dark	Long	Small	RT	Complex	Low	SB
19	Greenish	Pure	Bright	Short	Small	RB	Complex	High	NSB
20	Greenish	Diluted	Dark	Long	Large	LT	Complex	Low	SB
21	Greenish	Diluted	Bright	Short	Large	LT	Complex	Low	SB
22	Greenish	Diluted	Bright	Long	Large	LT	Complex	Low	SB
23	Blueish	Diluted	Dark	Short	Large	RT	Complex	High	NSB
24	Reddish	Diluted	Dark	Short	Large	LT	Complex	Low	SB
25	Greenish	Pure	Bright	Short	Small	RB	Complex	High	NSB
26	Blueish	Diluted	Bright	Short	Large	RT	Complex	High	NSB
27	Reddish	Diluted	Dark	Long	Large	B	Simple	Low	SB
28	Reddish	Diluted	Dark	Long	Large	LT	Complex	High	SB
29	Blueish	Diluted	Dark	Short	Large	RT	Complex	High	NSB
30	Greenish	Diluted	Bright	Short	Small	RB	Complex	High	NSB
31	Blueish	Diluted	Bright	Short	Large	LT	Complex	High	SB
32	Reddish	Diluted	Dark	Short	Large	LT	Complex	Low	SB
33	Greenish	Pure	Dark	Short	Large	RB	Simple	Low	SB
34	Reddish	Diluted	Dark	Short	Large	LT	Complex	Low	SB
35	Reddish	Diluted	Bright	Short	Small	RT	Complex	High	NSB

$$
\begin{aligned}
&Values(RWH) = Short, Long \\
&S = [19+,16-] \\
&S_{Short} \leftarrow [6+,16-], S_{Long} \leftarrow [13+,0-] \\
&Gain(S, RWH) = Entropy(S) - \sum_{v \in \{Short, Long\}} \frac{|S_v|}{|S|} Entropy(S_v) \\
&\quad = Entropy(S) - (22/35) Entropy(S_{Short}) - (13/35) Entropy(S_{Long}) \\
&\quad = 0.9947 - (22/35)0.8454 - (13/35)0 \\
&\quad = 0.463
\end{aligned} \tag{5}
$$

Table 3. Information gain for each attribute

Attribute	*AOH*	*AOS*	*AOI*	*RWH*	*SIZE*	*POS*	*DOE*	*MWC*
Information gain	0.087	0.027	0.174	0.463	0.107	0.377	0.016	0.295

2.3.2. Scorebox extraction using SVM classifier

The SVM can provide a good generalization performance on classification problems without incorporating domain knowledge [9]. Since the input vectors cannot be properly classified with a linear decision function in most classification problems, we use the RBF kernel for mapping input vectors to high dimensional feature space. In order to perform training, 35 captions shown in Table 2 are used for training samples. After training, support vector, Lagrange multiplier α, and biased factor b are obtained. By using these values, we can obtain the label of **x** by observing the sign of classifier f when the text vector **x** is given as follows,

$$
f(\mathrm{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathrm{s}_i, \mathrm{x}) + b, \ \ where \ \ K(\mathrm{x}, \mathrm{y}) = e^{-\|\mathrm{x}-\mathrm{y}\|^2 / 2\sigma^2}. \tag{6}
$$

Here **x** and **s** denote the test vector and support vector obtained from training, respectively. N and y denote the number of support vectors and the label of support vector, respectively. The value of σ is 1 in our method, which denotes the width of basis function. If the value of f is larger than 0, then the label of test vector **x** is assigned to +1 and corresponding caption is determined as the scorebox. The classification result of three captions in Fig. 3-(c) is shown in Table 4. We can see that the first one is extracted as the scorebox in terms of the label of feature vector.

Table 4. Classification result

Feature vector	f	label
(0.992, 0.1, 0.077)	2.419	+1
(0.196, 0.25, 0.074)	-0.117	-1
(0.081, 0.55, 0.435)	-1.911	-1

3. EXPERIMENTAL RESULTS

The result of scorebox extraction on various sports videos is shown in Fig. 5. Scoreboxes are successfully extracted regardless of shape, color, and size. Although there exist several advertisement boards and logos in each image frame, the proposed algorithm shows robustness to such false positions.

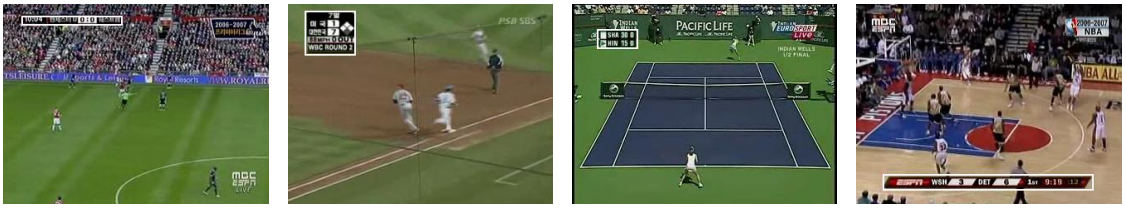


Fig. 5. Result of scorebox extraction on various sports videos. Extracted scorebox is represented using white rectangle.

To show the efficiency of proposed method, we compare the classification accuracy of proposed method with other cases, which can be represented by using the different dimension of feature vector and the different type of decision function. First, One-dimensional feature vector is compared with three-dimensional feature vector used in our method. The best attribute, *RWH*, is used as a one-dimensional feature vector. Secondly, linear decision function, which can be represented by linear hyperplane on the feature vector space, is compared with nonlinear decision function based on radial basis function. 32 test captions are used to analyze the classification accuracy. The comparison result is shown in Table 5. We can see that the performance of proposed method is better than other cases.

Table 5. Performance comparison according to SVM setting

SVM setting	Classification accuracy
3-D feature vector + nonlinear function	93.75 %
3-D feature vector + linear function	81.25 %
1-D feature vector + nonlinear function	78.13 %

The framework for evaluating performance has been implemented by using Visual Studio 2003 (C++) under FFMpeg library, which has been utilized for MPEG decoding. All sports videos for the experiment are encoded with the image size of 320×240 and composed of more than 300 frames. Soccer, baseball, tennis, and basketball video in Fig. 5 are used for performance evaluation. The learning period L is set to 100 frames and the threshold value for *AIM* and *AEM* is set to 80, respectively. The experiments were performed on the PC (Core2Duo 1.8GHz). The result of performance evaluation is shown in Table 6. Since the decreasing rate of processing time is insignificant, we can say that the proposed method is very efficient.

Table 6. Evaluation of processing time

Total frames	300 frames
Processing time without SB extraction	32.37 fps
Processing time with SB extraction	32.22 fps
Decreasing rate of speed	0.48 %

The extraction accuracy evaluates how many percents of all ground-truth scorebox are extracted. In this paper, recall and precision defined as below are used for evaluating extraction rate,

$$\text{Recall} = \frac{Card(SB_n \cap SB_{n,GT})}{Card(SB_{n,GT})}, \text{Precision} = \frac{Card(SB_n \cap SB_{n,GT})}{Card(SB_n)}, \quad (7)$$

where SB_n denotes the extracted scorebox and $SB_{n,GT}$ denotes the actual scorebox based on the ground truth in the n^{th} frame. Recall and precision are estimated on the basis of the results shown in Fig. 6. Based on these results, we can see that the accuracies are very high.

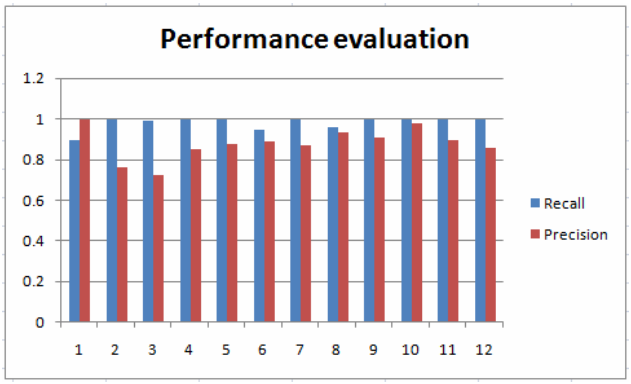


Fig. 6. Extraction accuracy based on recall and precision.

4. CONCLUSION

An efficient and robust method for scorebox extraction from sports videos has been proposed in this paper. We use the combined feature maps based on accumulated intensity and accumulated edge information to extract scorebox candidates. Linked maps are generated to make connected components for each candidate and then each connected component is reshaped to have smooth boundaries. We carefully select eight attributes for the scorebox. Based on those attributes, the optimal information gain is computed and top three ranked attributes in terms of information gain are selected as a three-dimensional feature vector for SVM to distinguish the scorebox from other candidates. The extracted scorebox can be magnified for better understanding of the game at viewer's request. To validate the performance of our method, sports videos with diverse scoreboxes are evaluated. The proposed method is very useful for mobile applications to provide magnified scorebox for better delivering viewing experience. Our future work is to extract the score and time information from the extracted scorebox for more advanced and intelligent applications.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center support program supervised by the Institute of Information Technology Advancement. (grant number IITA-2008-C1090-0801-0017)

REFERENCES

[1] R. Jacques and L. V. Eycken, "Graphic/non-graphic segmentation for multi-standard compression," in *Proc. Electronic Imaging Conference*, pp. 615-624, Jan. 1998.
[2] 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.
[3] H. S. Yoon, Y. L. Bae, and Y. K. Yang, "A soccer image sequence mosaicking and analysis method using line and advertisement board detection," *ETRI Journal*, vol. 24, pp.443-454, Dec. 2002.
[4] C. G. M. Snoek and M. Worring, "Time interval maximum entropy based event indexing in soccer video," in *Proc. IEEE Int. Conf. Multimedia and Expo*, vol. 3, pp. 481-484, July 2003.
[5] D. Y. Chen, M. H. Hsiao, and S. Y. Lee, "Automatic closed caption detection and filtering in MPEG videos for video structuring," *Journal of Information Science and Engineering*, vol. 22, no. 5, pp. 1145-1162, 2006.

[6] H. C. Shih and C. L. Huang, “A robust superimposed caption box content understanding for sports videos,” *IEEE Int. Symposium on Multimedia*, pp. 867-872, Dec. 2006.
[7] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. Syst. Man Cybern*., vol. 9, no. 1, pp. 62-66, Mar. 1979.
[8] T. M. Mitchell, *MACHINE LEARNING*, International ed., McGraw-Hill: Singapore, 1997, pp.52-76.
[9] Christopher J. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Mining and Knowledge Discovery*, vol. 2, pp. 121 - 167, 1998.