A Novel Approach for Key Caption Detection in Golf Videos Using Color Patterns

Cheolkon Jung and Joongkyu Kim

ABSTRACT—This paper provides a novel method of detecting key captions containing player information in golf videos. We use the color pattern of captions and its repetition property to determine the key captions. The experimental results show that the proposed method achieves a much higher accuracy than existing methods.

Keywords—Key caption detection, color pattern, text information.

I. Introduction

Text in images and videos contains useful information that can help a machine to understand content. Text is very important for the automatic annotation, indexing, and parsing of images and videos [1]-[3]. For sports videos, text information is very important because it displays valuable game information, such as scores and players. Therefore, it is important to efficiently find the key texts containing valuable game information from sports videos. These key texts are called key captions, and they can be used for video highlights or content search. These key captions often use a pre-formatted template. For example, key captions display inning/score/ball/out counts in baseball video, scores in soccer videos, and players' names in golf videos.

Conventional methods of detecting key captions assume that the location of the key captions is fixed during a game [4]-[6]. Therefore, these key captions are determined by their location property. However, in the case of golf, the location of the key captions containing player information is not fixed as shown in Fig. 1. Here, key captions with player information in Figs. 1(a) and (b) are located at the top-left and top-right positions, respectively; thus, key captions cannot be detected by using the location property. To determine the key captions in golf, we use a color pattern and its repetition property instead of the location property.

(a) (b)

Fig. 1. Key captions in golf videos: (a) key caption with player information located at the top-left position and (b) key caption with player information located at the top-right position.

II. Key Caption Detection

In general, captions in videos contain many edge components and typically last 2 to 10 seconds. These captions can be detected by using this edge and temporal information [7]. The color pattern of a caption region is obtained by the dominant color descriptor (DCD) from MPEG-7 standard descriptors. Five color descriptors are defined in MPEG-7 covering different aspects of color and application areas. The DCD characterizes an image or image region in terms of a small number of dominant color values and some statistical properties related to those. This descriptor provides a compact description of the representative colors of an image [8], [9]. This descriptor consists of the representative colors, their

Manuscript received May 2, 2008; revised May 29, 2008; accepted June 17, 2008.

Cheolkon Jung (phone: + 82 31 290 7199, email: ckjung@ece.skku.ac.kr) and Joongkyu Kim (email: jkkim@skku.edu) are with the School of Information & Communication Engineering, Sungkyunkwan University, Suwon, Rep. of Korea.

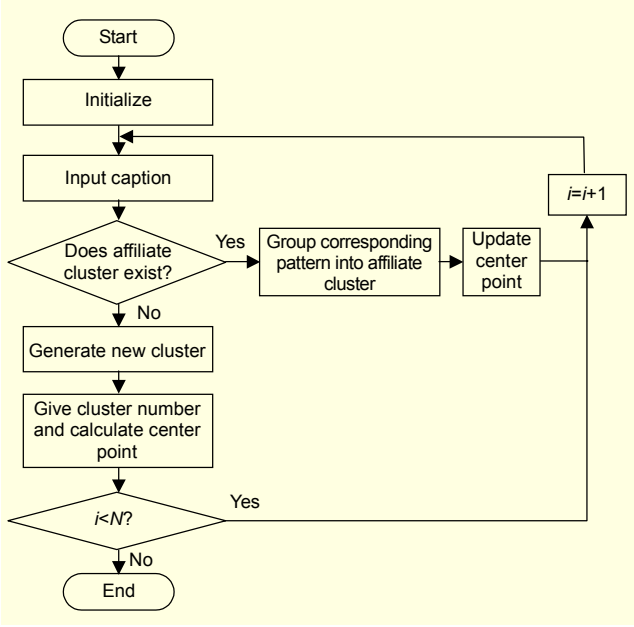

Fig. 2. Flow chart of the proposed method.

percentages in the region, the spatial coherency of dominant colors, and the color variance for each dominant color. It is defined by

$$F = \{\{c_i, p_i, v_i\}, s\}, \quad (i = 1, 2, \cdots, M), \tag{1}$$

where c_i, p_i, v_i, and s are i-th dominant color, percentage value, color variance, and spatial coherency, respectively. The spatial coherency s is calculated by selecting each representative color and calculating the per-coherency as the average connectivity of the pixels using a 3×3 masking window within the cluster. Next, a weighted average of these values is computed using the percentages p_i as the weights, leading to the spatial coherency. The color variances v_i are computed as variances of the pixel values within each cluster.

Consider two dominant color descriptors, $F_1=\{\{c_{1i}, p_{1i}, v_{1i}\}, s_1\}$, ($i$=1, 2,⋯, M_1) and $F_2=\{\{c_{2i}, p_{2i}, v_{2i}\}, s_2\}$, ($i$=1, 2,⋯, M_2). Ignoring the optional variance parameter and the spatial coherence, the dissimilarity $D(F_1, F_2)$ between the two descriptors can be computed as

$$D^2(F_1, F_2) = \sum_{i=1}^{M_1} p_{1i}^2 + \sum_{j=1}^{M_2} p_{1j}^2 - \sum_{i=1}^{M_1}\sum_{j=1}^{M_2} 2a_{1i,2j} p_{1i} p_{2j}, \tag{2}$$

where the subscripts 1 and 2 in all variables stand for descriptions F_1 and F_2, respectively, and $a_{k,l}$ is the similarity coefficient between two colors c_k and c_l. The coefficient $a_{k,l}$ has the range of [0, 1] and is negatively proportional to $D^2(F_1, F_2)$.

Figure 2 shows the key caption detection procedure using the color pattern in detail. A cluster number 1, for example, is given to an initial dominant color value obtained in initialization, and a center point of a corresponding cluster is stored together with a number 1 of a color pattern grouped into an affiliate cluster. When a caption is input, whether an affiliate cluster corresponding to the dominant color value obtained by the DCD exists is determined. The distance between the dominant color value and the center point of a cluster is calculated by (2). When the distance corresponds to the affiliate cluster, the dominant color value is clustered into the same group, a corresponding center point is updated, and the number of grouped color patterns is increased by 1. The same process is performed with respect to a subsequent index. If the dominant color value is not clustered into any clusters, a new cluster will be generated, and a center point and the number of the corresponding cluster will be given. This process is performed until an index i becomes equal to the maximum number of input color patterns N.

Several clusters are produced by the proposed clustering method. Other regions excluding the key captions are classified into one or more small clusters because the number of these regions is relatively small compared with that of key captions containing player information. The key captions are determined by selecting the dominant cluster of maximum size among the existing clusters.

III. Experimental Results

The experiments were performed by using a PC (CPU: Intel Pentium 4, 2.4 GHz) with VC++6.0. We tested eight games for 12.5 hours with 720×480 MPEG-2 videos from the channels SBS Sports, KBS Sports, and MBC ESPN. Performance results for the eight videos are given in Table 1.

The performance was measured on the key caption level. For a quantitative evaluation, we define that a detected key caption is correct on the condition that the intersection of the detected key

Table 1. Performance evaluation results in golf videos. Experiments are performed by using a PC (CPU: Intel Pentium 4 2.4 GHz) with Microsoft Visual C++6.0.

Name	Recall	Precision	Length (h)
LPGA 1	0.83	0.99	1.5
LPGA 2	0.77	1.00	1.0
CJ Nine Bridge 1	0.91	1.00	1.0
British Open 1	0.66	0.95	1.5
British Open 2	0.77	0.91	1.0
British Open 3	0.83	0.92	3.0
British Open 4	0.75	0.81	2.0
CJ Nine Bridge 2	0.96	1.00	1.5
Total	0.82	0.95	12.5

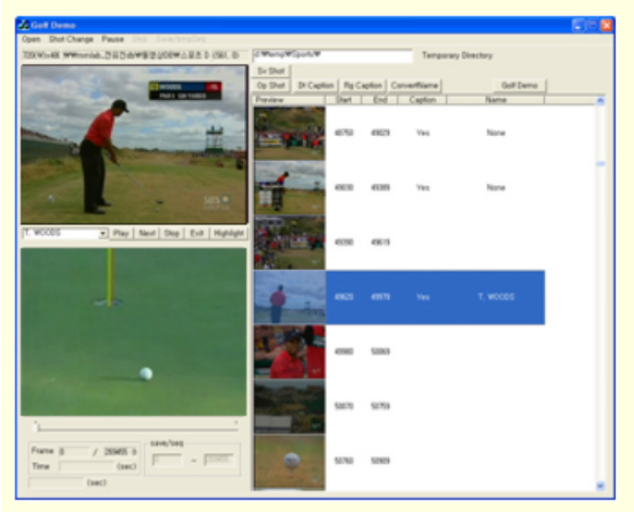

Fig. 3. Golf navigation system using the proposed method.

caption region (DKCR) and the ground truth key caption region (GKCR) covers more than 90% of the DKCR and 90% of the GKCR. The GKCRs in the testing video clips were localized manually [10].

The detection results were compared with the ground truth in terms of two measures: recall rate (*recall*) and precision rate (*precision*). They are obtained as follows:

$$recall = \frac{N_C}{N_G}, \tag{3}$$

$$precision = \frac{N_C}{N_C + N_F}, \tag{4}$$

where N_F, N_C, and N_G are the numbers of falsely detected, correctly detected, and ground truth key captions, respectively. The average recall and precision rates for the 8 golf videos are 0.82 and 0.95, respectively (see Table 1). We do not compare the performance of our method to that of other methods because conventional approaches [4]-[6] cannot detect key captions containing player information which appear in different positions. These results show that our method can efficiently extract player information from golf videos. The user interface of the golf navigation system based on the proposed method is shown in Fig. 3. The top-left window in this figure is the "play" window of a golf video. The bottom-left window shows play shots by a favorite player. The right window is the navigation window that provides the player information of each shot.

IV. Conclusion

We have proposed a novel method to detect key captions containing player information in golf videos. To detect key captions efficiently, the method uses the color pattern of captions and its repetition property instead of a fixed location property. With the proposed method, the recall and precision rates of key caption detection are 82% and 95%, respectively. These results show that our method can detect key captions well, even if their location is not fixed during the game. In the future, we will further investigate the fusion of audio and visual domain features to model highlights.

References

[1] D. Chen, J.M. Odobez, and H. Bourlard, "Text Detection and Recognition in Images and Video Frames," *Pattern Recognition*, vol. 37, 2004, pp. 595-608.

[2] K. Jung, K.I. Kim, and A.K. Jain, "Text Information Extraction in Images and Video: A Survey," *Pattern Recognition*, vol. 37, 2004, pp. 977-997.

[3] N. Dimitrova et al., "Applications of Video Content Analysis and Retrieval," *IEEE Multimedia*, vol. 9, 2002, pp. 43-55.

[4] D. Zhang and S.F. Chang, "General and Domain-Specific Techniques for Detecting and Recognizing Superimposed Text in Video," *Proc. IEEE Int'l. Conf. on Image Processing*, 2002, pp. 593-596.

[5] D. Zhang and S.F. Chang, "Event Detection in Baseball Video Using Superimposed Caption Recognition," *Proc. ACM Int'l. Conf. on Multimedia*, 2002, pp. 315-318.

[6] E. Kim et al., "A Video Summarization Method for Basketball Game," *Lecture Notes on Computer Science*, vol. 3767, 2005, pp. 765-775.

[7] C. Jung and J.K. Kim, "Automatic Superimposed Text Localization from Video Using Time Information," *Journal of Korea Information and Communications Society*, vol. 32, 2007, pp. 834-839.

[8] B.S. Manjunath et al., "Color and Texture Descriptors," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, 2001, pp. 703-715.

[9] L. Cieplinski, "MPEG-7 Color Descriptors and Their Applications," *Lecture Notes on Computer Science*, vol. 2124, 2001, pp. 11-20.

[10] M.R. Lyu, J. Song, and M. Cai, "A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, 2005, pp. 243-255.