Extracting Caption Content on Sports Videos by Caption Identification

[§]Yih-Ming Su and [*]Chaur-Heh Hsieh

[§]*Department of Electronic Engineering, I-Shou University, Kaohsiung County, Taiwan.*
[*]*Department of Information Engineering, I-Shou University, Kaohsiung County, Taiwan.*

ymsu@isu.edu.tw

Abstract-*The paper describes a novel caption extraction scheme to detect unconstrained captions on sports videos and identify the captions for extracting the caption content directly. A caption detection process based on a multi-frame averaging approach is applied to locate a reliable caption region accurately. Furthermore, a caption-content extraction process based on caption identification and model masking approaches is applied to directly extract the caption content without a conventional segmentation process. Experimental results show that the proposed caption detection approach is very efficient to provide a high tolerance to noisy and complex-background video frames. Furthermore, a learning-based approach for the caption identification process is capable of generalizing the learning knowledge to identify various sports captions and get an average identification rate of 89.99% for testing data.*

Keywords: Caption Detection, Caption Identification

1. Introduction

An automatic caption extraction scheme applied in sports videos is helpful to analyze and understand the video content because a sports caption carries a lot of important information of sports games. However, some major problems are that various sports captions embedded in many sports games have different sizes, locations, shapes, and layouts. Moreover, extracting the caption data, including textual and graphical information, is difficult due to highly compact data layout, as shown in Figure 1. Finally, some blurred and translucent captions may make the extraction process difficult. Therefore, it is essential to extract reliable and stable caption data for further video understanding.

Most of video studies [1-4] on caption detection and extraction use the process of temporal-information verification after spatial-image analysis to enhance extraction performance. The spatial-image analysis approaches, using connected component [2], edge detection [3], and texture analysis [4] techniques, may be sensitive to noise and complex background such that the caption can not be located accurately. Moreover, they usually assume that the captions have a high contrast against video background. Finally, the sports captions being different with news captions not only contain text information, but also graphic information. Therefore, extracting unconstrained caption styles from various sports videos is a difficult task because we have to detect various sports captions with different sizes, locations, shapes and layouts.

To cope with the above problems, this study proposes a novel caption extraction scheme based on multi-frame averaging, caption identification and model masking approaches for unconstrained sports captions. Firstly, in order to release noise disturbance and overcome complex-background variation, a multi-frame averaging technique based on temporal consistency of caption appearance is initially used to obtain a reliable image data before spatial-image analysis, instead of working on one frame at a time. Then, a simple binarization process employing the global mean and standard deviation of the averaged video image is designed to determine a threshold range. Due to some captions with translucent background, this effect may produce some holes or disconnectivity in the binary image. Therefore, a morphological processing [8] including the line dilation and hole-filling operations is used to fill the holes and correct the disconnectivity such that the caption region is a complete connected component. Each connected component is used to extract some geometrical features, including size, location, shape, and layout. These features are applied to determine the caption type. Once an input caption style is identified, a model masking technique is applied to directly extract the caption data without a conventional segmentation processing [5].

The remainder of this paper is organized as follows. Section 2 describes a caption detection approach, using a multi-frame average technique. Next, Section 3 presents the caption identification and extraction processes, using a learning-based and model masking techniques. Section 4 summarizes experimental results. Finally, concluding remarks are

made in Section 5.

2. Caption Detection and Location

In order to automatically extract various caption contents in sports videos, we have to first detect the caption and locate the caption region. Normally, the super-imposed sports captions remain stable for a certain amount of time and appear in the same position. We can use these properties to detect and locate the captions. Additionally, in order to release noise disturbance from video transmission and overcome the variation of complex-background images, a multi-frame averaging technique based on using temporal consistency of caption appearance before spatial-image analysis is used to detect and locate a stable caption region accurately. Figure 2 shows the flowchart of the caption detection process, which is explained as follows.

(1) When arbitrary sports videos encoded in MPEG1 format are as input sources, a sequence of image frames from the videos is captured with 2 frames per second.

(2) In the initial capture processing, 20 video frames are used to average the intensity of all video frames and calculate the global mean and standard deviation of the averaged video frame.

(3) Another 20 video frames are again captured in the next time, and all video frames including previous and present captured video frames are used to produce a new averaged video frame.

(4) The intensity of the averaged video frame except caption regions will be approximately uniform because the change of each pixel is random over a long time. Moreover, the standard deviation (STD) of the intensity should be gradually stable.

(5) A binarization process for the averaged video frame A(x,y) is performed with a threshold range obtained automatically. The binary image B(x,y) is defined by

$$B(x,y)=\begin{cases}0\,(\text{black}) & \text{if } (M-2.2*STD)\le A(x,y)\le(M+2.2*STD)\\ 1\,(\text{white}) & \text{others}\end{cases} \quad (1),$$

where M and STD denote the mean and standard deviation of the intensity of the averaged video frame.

(6) Each connected component labeled in the binary image can be regarded as a caption candidate when the size of the connected component is limited at a specific range.

Figure 3 shows the results of caption detection process: (a) an averaged video frame; (b) a binary image; (c) two complete connected components remedied by line dilation and hole-filling processes [8]; (d) the detected caption region enclosed by the contour.

3. Caption Identification and Extraction

Once each caption candidate is detected by the caption detection process, the candidate has to further be identified to confirm the caption type. In order to perform an identification process, some features, including size, location, shape, and layout, are extracted from each caption candidate, and the linear discrimination function [6] as a classifier is used to classify each candidate into one of all caption types. Then, in order to extract caption data without a conventional segmentation process, each caption masking model corresponding to its caption type is constructed in advance. Finally, the logical AND operation between an identified caption region and corresponding caption masking model is performed to extract the caption content directly.

3.1 Feature Extraction of Sports Caption

An efficient feature extraction approach based on geometrical characteristics of each caption candidate is used to extract the size, location, shape, and layout features of the candidate. The following caption features are described as follows.

A. Normalized size feature for each caption candidate:

$$f^{A}=\frac{4A_c}{A}, \quad (2)$$

where A_c and A denote the area of the caption candidate and the video frame, respectively.

B. Normalized location feature for each caption candidate:

$$f^{X}=\frac{x}{W},\ f^{Y}=\frac{y}{H}, \quad (3)$$

where (x,y) denote the coordinates of the centroid of each caption candidate, and (W,H) represent the width and height of a video frame, respectively.

C. Normalized shape feature for each caption candidate:

$$f^{S}=\left[\frac{|FD_2|}{|FD_1|}\ \frac{|FD_3|}{|FD_1|}\ \frac{|FD_4|}{|FD_1|}\ \ldots\ \frac{|FD_{15}|}{|FD_1|}\right], \quad (4)$$

where $|FD_i|$ denotes the absolute value of the ith component of Fourier descriptors. The FDs are calculated by Fourier transform of the coordinates of the contour of the caption candidate, using the contour Fourier method [7]. For example, $|FD_1|$ denotes the absolute value of the first non-zero frequency component of the descriptors.

D. Normalized layout feature for each caption candidate:

$$f^{L}=\left[\frac{E_{11}}{E_1}\ \frac{E_{12}}{E_1}\ \frac{E_{13}}{E_1}\ \frac{E_{14}}{E_1}\ \frac{E_{15}}{E_1}\ \frac{E_{21}}{E_2}\ \frac{E_{22}}{E_2}\ \ldots\ \frac{E_{44}}{E_4}\ \frac{E_{45}}{E_4}\right], \quad (5)$$

where E_{ij} denotes the number of the edge pixels in

the jth classes, including horizontal(j=1), right-diagonal (j=2), vertical (j=3), left-diagonal (j=4), and junction (j=5) classes, and E_i denotes the total number of edge pixels in the ith block of the caption candidate (i=1,2,…,4). More details are described as follows. Firstly, the edges of the caption region in the averaged video frame, as shown in Fig. 4(a), are detected by the Canny edge detection approach [9], and the contour of the caption region is removed by logical AND operation between the edge and contour caption frames, as shown in Fig 4 (b) and (c), respectively. The remaining edge pixels are comprised of the layout of the caption as shown in Fig. 4(d). In order to add local characteristics, the edge caption region is divided into 4 uniform blocks. Each edge pixel is categorized into 5 classes, using a 3*3 filtering window, as shown in Fig. 5. The index value (IV) of each edge pixel is calculated by

$$IV = \sum_{i=0}^{8} w_i z_i, \quad w_i = 2^i, \tag{6}$$

where the *w*'s are weighting coefficients and the *z*'s are the values of edge pixel and its 8-neightborhood pixels ($z_i \in [0,1]$). The edge pixels are classified by

$$edgeclass = \begin{cases} horizontal\ ,\ IV \in [1,9,16,17,18,33,44], \\ right-diagonal\ ,\ IV \in [2,32,34], \\ vertical\ ,\ IV \in [4,36,64,66,68,72,132], \\ left-diagonal\ ,\ IV \in [8,128,136], \\ junction\ ,\ IV \in others\ , \end{cases} \tag{7}$$

3.2 Caption Classification

According to the analysis of Kimura et al., [6], the classification performance using the linear discrimination function [10] is better than that of using the Euclidean distance and city block distance functions. Therefore, the linear discrimination function adopted to classify an input caption into a caption category is described by

$$\begin{aligned} &g_i(X) = V_i^T X + V_{i0}, \\ &V_i = S_w^{-1}\mu_i \text{ and } V_{i0} = -\tfrac{1}{2}\mu_i^T S_w^{-1}\mu_i, \end{aligned} \tag{8}$$

where μ_i and S_w denote the mean vector of the feature set X in the i^{th} caption category and within-class scatter matrix.

3.3 Caption-Content Extraction

Once an input caption is identified in the identification process, the caption data including textual and graphical information, has to be extracted for semantic analysis and understanding. Therefore, we propose a model masking technique to directly extract the caption data without a conventional segmentation process. Firstly, each caption masking model including the attributes of the caption data is constructed in advance as shown in Fig. 6(a). The attributes contain the size, position, color, and meaning of each caption data, such as score, inning, ball count, base, team name, etc., for baseball sports games. A logical AND operation between the input caption and masking model is performed to segment the caption region into a set of caption data. Finally, the attributes of the caption data are used to again check the caption type such that the actual caption type is again confirmed. If the attributes of the extracted caption data are different from the constructed attributes of the constructed caption models, the identified caption type may be wrong. Fig. 6 (a) displays an example of the caption masking model and (b) the individual caption data extracted by the model masking approach.

4. Experimental Results

To evaluate the performance of the proposed approach, we have collected various sports video games encoded by MPEG1 format from TV channels. A database of the sports videos includes 18 baseball, 10 basketball, 9 rugby, 6 soccer, 4 tennis, 3 volleyball, 2 badminton sports games with different caption styles. Each video game is randomly clipped into three 3-min video segments at different time. In the caption identification process, each caption type with two video segments was used as training data, and the remainder was used as test data.

4.1. Performance analysis

In Section 2, a multi-frame averaging technique, based on temporal consistency property, is efficiently applied to noisy and complex-background video frames, as shown in Fig. 7(a). The result in Fig. 7(b) indicates that the technique is robust against the noise and complex background. To further prove the effectiveness of the caption detection approach, we found the average precision rate of 96.5% and the average recall rate of 88.22%, as shown in Table 1. From the experimental testing, most captions are detected by our detection approach because the captions are stable to appear in the video segments. Moreover, we found that 1-min appearance time of the caption is demanded for detecting the sports captions because the change of each pixel in the averaged video frame except to the caption region is random over a long time. Meantime, the false alarm happened because other captions such as some captions for commercial advertisement and player information may appear in longer time than the sports captions.

In Section 3, to assess the effectiveness of the caption identification process, the average identification rate of various sports videos is shown in

Table 2. The average identification rate for 84 sports video segments is 97.68% for training set; the average identification rate for 24 sports video segments is 89.99% for testing set. Some captions fail to be identified because the caption shape may be changed to display additional information. Additionally, the layout feature is sensitive to the change of caption data such that the feature variation becomes large.

4.2. Discussion

This section summarizes and discusses several important observations regarding the experimental performance of the caption detection and identification processes. Firstly, the effect of using the standard deviation (STD) of the intensity of the averaged video frame is considered in the caption detection process. In Fig.8, the convergence curve in terms of the mean and STD was found to be stable when the number of video frames is large enough. In order to automatically detect the caption, the number of the video frames is determined when the change of the STD value is stable. Although the caption detection process takes more computational time, this process just do one time at the game beginning, and then the subsequent caption data can be directly extracted during sports video games. Secondly, the proposed learning-based approach provides a flexible way to learn various caption styles on sports videos instead of the specific caption style handling [5]. Finally, the following reasons may cause the failure of the detection and identification processes. The detection error may be caused by the inadequate detection time. The identification error may be caused by the similar caption styles from different sports videos. Moreover, not all types of caption styles can be accurately identified under number-limited caption samples where the number of training samples is inadequate.

5. Conclusions

This study proposes a novel learning-based approach to identify various sports captions in order to directly extract the caption content. The proposed approach uses the identification process to classify an input sports captions into one of the sports caption types, and then corresponding caption masking model is used to directly extract the caption content without a conventional segmentation process. Additionally, in order to identify the sports captions efficiently, the caption region has to be detected firstly. A caption detection process, based on a multi-frame averaging approach, first uses temporal consistency of caption appearance before spatial-image analysis. The proposed approach is helpful to provide a high tolerance to noisy and complex-background video frames because using a sequence of video frames to function is superior to working on a single video frame at one time. Finally, the proposed approaches for caption extraction have been tested successfully with various sports captions. The performance of the caption identification process can be achieved by the identification rate of 89.99% for testing set. Moreover, the performance of the caption detection process can be achieved by the average precision rate of 96.5% and recall rate of 88.2%.

In the future, we will extend our study to extract highlight events of sports games or summarize the sports video content by further recognizing the caption content.

References

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

[2] X. Tang, X. Gao, J.H. Liu, and H. J. Zhang, “A Spatial-Temporal Approach for Video Caption Detection and Recognition,” IEEE Trans. On Neural Networks, vol. 13, No. 4, pp. 961-971, 2002.

[3] R. Lienhart and F. Stuber, “Automatic Text Recognition in Digital Videos,” Proceedings of SPIE Image and Video Processing IV, 2666, pp.180-188, 1996.

[4] D. Crandall, S. Antani, and R. Kasturi, “Extraction of Special Effects Caption Text Events from Digital Video,” Inter. J. on Document Analysis and Recognition, vol. 5, pp. 138-157, 2003.

[5] D. Zhang, R. K. Rajendran, and S. F. Chang, “General and Domain-Specific Techniques for Detecting and Recognizing Superimposed Text in Video,” Inter. Conf. on Image Processing, pp. 22-25, 2002.

[6] B. M.Mehtre, M. S. Kankanhalli, and W. F. Lee, “Shape Measures for Content Based Image Retrieval: A Comparison,“ Information Processing & Management, Vol. 33, No. 3, pp. 319-337, 1997.

[7] K. Kimura, T. Wakabayashi, S. Tsuruoka, and Y. Miyake, “Improvement of handwritten Japanese Character Recognition Using Weighted Direction Code Histogram,” Pattern Recognition, Vol. 30, No. 8, pp. 1329-1337, 1997.

[8] Soille, P., Morphological Image Analysis: Principles and Applications, Springer-Verlag, pp. 173-174. 1986.

[9] J. F. Canny, “A computation approach to edge detection.”IEEE Trans. PAMI, Vol.8 No. 6, pp.679-698, 1986.

[10] K. Fukunaga, “Introduction to statistical pattern recognition,” Academic Press, Inc. New York, Second

Edition, 1990.

Fig. 1. A sports caption superimposed in the video frame

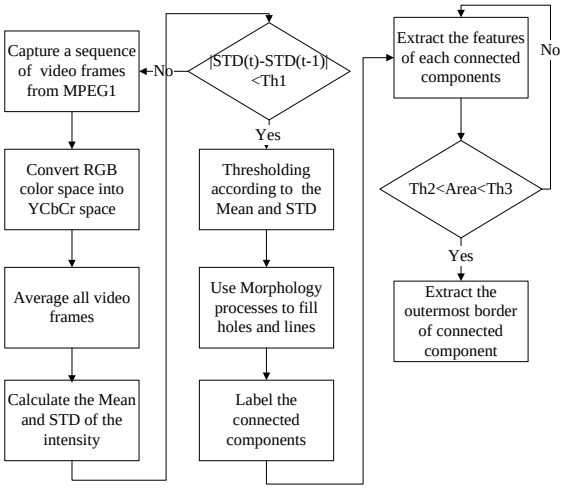
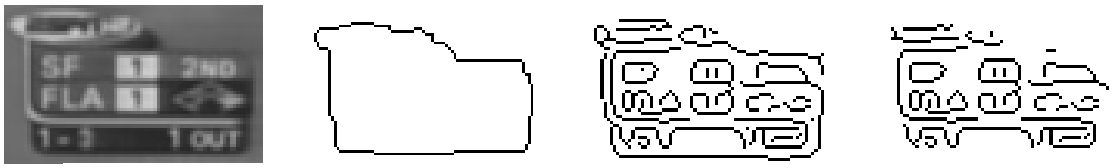


Fig. 2. Flowchart of the caption detection process

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

Fig. 3. The results of the caption detection process; (a) an averaged video frame; (b) a binary image; (c) a hole-filling image; (d) a detected caption image.

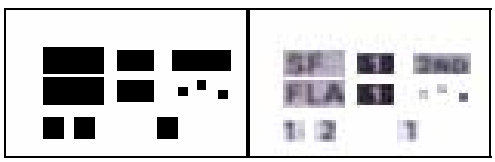


(a) (b) (c) (d)

Fig. 4. The results of the contour and layout extraction process; (a) a caption region; (b) a contour image; (c) a hole-filling image; (d) a detected caption image.

w_4	w_3	w_2
w_5	w_0	w_1
w_6	w_7	w_8

Fig. 5. Illustration of weighting coefficients



(a) (b)

Fig. 6. (a) A caption masking model; (b) the extracted caption data.

(a) (b)

Fig. 7. (a) A noisy and complex-background video frame; (b) a detected caption region in the averaged video frame.

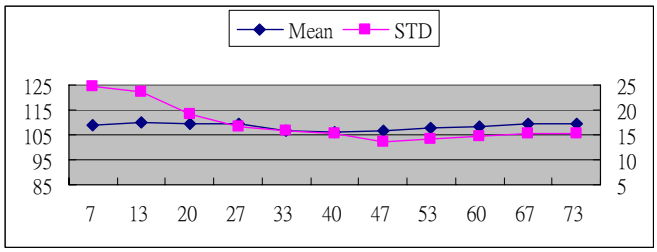


Fig. 8. The convergence curves of the averaged video frames

Table 1. The performance of the caption detection process.

Sports Videos	Baseball	Basketball	Rugby	Soccer	Tennis	Volleyball	Badminton
# video segments	18*3	10*3	9*3	6*3	4*3	3*3	2*3
# correct	31	27	25	16	7	9	5
# missed	2	3	2	1	4	0	1
# false alarm	2	0	0	1	1	0	0
Recall	93.93%	90%	92.6%	94.1%	63.6%	100%	83.3%
Precision	93.93%	100%	100%	94.1%	87.5%	100%	100%

Table 2. The performance of the caption identification process.

Sports Videos	Baseball	Basketball	Rugby	Soccer	Tennis	Volleyball	Badminton
IR for training set	94.4%	95%	94.4%	100%	100%	100%	100%
IR for test set	88.89%	80%	77.8%	83.3%	100%	100%	100%

IR: Identification Rate