A Novel Model-based Segmentation Approach to Extract Caption Contents on Sports Videos

[§]*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.

ABSTRACT

The study proposes a novel scheme to extract and recognize the caption contents of various sports captions. A caption extraction process based on an iteratively temporal averaging technique is used to detect and locate a caption region in a series of video frames. Moreover, a caption-content extraction process based on caption identification and model-based segmentation processes is used to extract accurately the contents of various sports captions. Finally, some low-quality character images extracted from the caption contents are recognized using a commercial OCR. Experimental results show that the proposed model-based segmentation approach is very efficient to extract the contents of the various sports captions. Furthermore, the recognition performance from the application of the segmentation approach can be improved about 7.72% in test numeral set, compared to the projection-based segmentation method.

1. INTRODUCTION

As the number of archived sports videos increases, efficiently browsing and skimming the content of these sports videos is important. An automatic caption recognition system applied to sports videos is helpful for analyzing and understanding the video content, because a sports caption has a lot of important information about a sports game. However, sports captions embedded in videos of many games have variations in caption size, location, shape, layout, and color. Moreover, extracting the caption contents, including textual and graphic information, is difficult, because the sports captions are highly condensed, as shown in Figure 1.

Some caption extraction approaches have been developed to exploit temporal redundancy and improve the performance of the caption extraction. Furthermore, they are generally classified into two categories according to the order of use of spatial and temporal information. The first category based on caption verification [1] and tracking [4] techniques initially uses spatial image analysis to detect the video caption, and then temporal information to verify and track the video caption. The other category based on multi-frame integration [3] and averaging [2] techniques initially exploits temporal redundancy to enhance the quality of the video frames and increase the accuracy of the location of the caption location. Then, the enhanced frames are further analyzed using spatial information to extract video captions.

Another problem is that segmenting and recognizing various items in the sports caption, such as characters, numerals, and graphic information are difficult. Moreover, the performance of the recognition depends significantly the quality of the segmentation processing. A good segmentation approach ensures good recognition performance. In the previous studies, some video caption segmentation approaches based on projection profile [5, 6] and recognition-based [7] techniques are utilized to segment the text of the video captions into various character components. However, the captions of sports videos differ from these of news and commercial advertisement videos because the former not only contain textual information, but also graphic information, such as the base status of a baseball game. Therefore, the mentioned segmentation approaches may not be suitable for processing the complex and compact captions of the sports videos.

This study proposes a novel caption recognition scheme, including caption extraction, caption identification, model-based segmentation and content recognition approaches for various sports captions. In the caption extraction stage, an iteratively temporal averaging approach is proposed before the spatial-image analysis is performed to enhance the image quality, reduce noisy disturbances, and overcome any complex background variation, thereby significantly facilitating the spatial-image analysis. Then, a simple binarization process based on the global mean and the standard deviation of the gray level of the averaged video image is utilized to determine a threshold range. This effect may yield some holes or disconnectivity in the binary image because some captions have a translucent and blurred background. Therefore, a morphological processing [8], including the line dilation and hole-filling operations is applied to fill the holes and remedy the disconnectivity, such that the caption region is a completely connected component. In the caption identification stage, each connected component is used to extract such geometrical features as size, location, shape and layout. A caption identification process considers these features to identify each connected component as one type of the captions. In the caption-content recognition stage, once an input caption is identified, a novel model-based segmentation approach, rather than a projection-based segmentation approach [5, 6], is applied to accurately extract the caption contents. The segmented numeral images are further enhanced by a size normalization process and then recognized using a commercial OCR system. Finally, a voting technique [9] is adopted to reduce recognition errors.

The remainder of this work is organized as follows. Section 2 describes a caption extraction approach, which uses an iteratively temporal averaging technique. Next, Section 3 presents the caption identification and recognition processes, which use a learning-based classification and model-based segmentation techniques, respectively. Section 4 summarizes the experimental results. Finally, Sections 5 draws conclusions.

2. CAPTION EXTRACTION

In order to automatically extract various caption contents in sports videos, these captions must be detected and located from a series of video frames. An iteratively temporal averaging technique, based on the temporal consistency of caption appearance before spatial-image analysis, is proposed to detect and locate a stable caption region accurately. The caption extraction process is described as follows.

(1) When arbitrary sports videos encoded in MPEG1 format are used as input sources, a sequence of image frames from the videos is captured at two frames per second such that the captured video frames have more variations in the video content.

(2)In the initial capturing process, the intensity of 20 video frames is averaged and the global mean and standard deviation of the gray level of the averaged video frames is calculated.

(3)If the standard deviation exceeds a threshold, then the next 20 video frames are further captured, and all of the video frames so far captured are again used to produce a new averaged video frame.

(4)The intensity of the averaged video frame, except in the caption regions, is approximately uniform because the change of each pixel is random over a long time. Moreover, the standard deviation (*STD*) of the intensity should approach a stable value such that the difference value (*Th1*) between the previous and present *STD* approaches a small value (In the experiment herein, *Th1*=0.1).

(5)The averaged video frame A(x,y) is binarized using an automatically obtained a threshold range. The binary image B(x,y) is defined by

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

where *M* and *STD* represent the mean and standard deviation of the intensity of the averaged video frame, and *k* is a user-defined parameter (In the experiment herein, *k*=2.2) .

(6)Each connected component labeled in the binary image can be considered to be a caption candidate when the size of the connected component is limited within a specific range. The range of sizes from *Th2* to *Th3* can be used to eliminate some of the false candidates (In the experiment herein, *Th2* =1000 and *Th3* =10000). Figure 2 presents the results of the caption extraction process; (a) an averaged video frame; (b) a binary image; (c) two complete connected components and (d) the extracted caption region enclosed by the contour.

3. CAPTION IDENTIFICATION AND SEGMENTATION

Once each caption candidate has been extracted, its type must be determined using an identification process. The identification process exploits such features of each candidate caption as its size, location, shape, and layout, as well as a linear discrimination function [10] applied to classify each candidate into a particular caption type. A caption model constructed in advance for each type of caption is associated with some attributes of the caption contents, including size, position, color, and meaning. As a caption region has to be accurately extracted into a set of caption contents, the attributes of the caption model are applied to perform a segmentation process.

3.1. Extracting features of sports captions

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

A. Normalized size of each caption candidate:

$$f^A=\frac{10A_c}{A}, \quad (2)$$

where A_c and A represent the areas of the caption candidate and the video frame, respectively. A factor of 10 is set by experimental observation.

B. Normalized location of each caption candidate:

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

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

C. Normalized shape of 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 $|FD_i|$ is the absolute value of the *i*th component of the Fourier descriptors. The *FD*s are calculated by the Fourier transform of the coordinates of the contour of the caption candidate, using the contour Fourier method [11]. For instance, $|FD_1|$ denotes the absolute value of the first non-zero frequency component of the descriptors. We select 15 components of *FD*s such that reconstruction error and time cost are a good compromise.

D. Normalized layout of each caption candidate:

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

where E_{ij} is the number of edge pixels in the *i*th block of the candidate (*i*=1,2,…,4) and the *j*th class, which may be a horizontal (*j*=1), right-diagonal (*j*=2), vertical (*j*=3), left-diagonal (*j*=4) or junction (*j*=5) classes. The details are described as follows. Firstly, the edges of the caption region in the averaged video frame, as shown in Fig. 3(a), are detected using the Canny edge detection approach [12], and the contour of the caption region is removed by applying logical AND operation between the contour and edge of the caption images, as shown in Figs. 3 (b) and (c), respectively. The remaining edge pixels comprise the layout of the caption as shown in Fig. 3(d). The edge caption region is divided into four uniform blocks. The index value (*IV*) of each edge pixel is given by

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

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

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

3.2. Caption classification

The classification performance of the linear discrimination function [10] exceeds 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 type is given by

$$g_i(X) = V_i^T X + V_{i0}, \quad \textbf{(8)}$$
$$V_i = S_w^{-1}\mu_i \text{ and } V_{i0} = -\tfrac{1}{2}\mu_i^T S_w^{-1}\mu_i,$$

where μ_i and S_w denote the mean vector of the feature set X in the ith caption type and the within-class scatter matrix, respectively. These parameters are determined in advance from the training patterns of the collected caption types.

3.3. Model-based segmentation and content recognition

Once an input caption has been identified, the caption contents, including textual and graphic information, must be further extracted for semantic analysis and understanding. Therefore, a model-based segmentation approach, rather than a projection-based segmentation approach [5, 6], is proposed to accurately extract the caption region into a set of caption contents. First, each caption model, including the attributes of the caption contents, is constructed in advance as presented in Fig. 4(a). The attributes contain the size, the position, the background color and the meaning of the caption contents, such as score, inning, ball count, base, team name, and others in baseball sports games. Therefore, the model-based segmentation process is implemented to segment the caption region into a set of caption contents according to the position and the size of each element of the caption. Finally, the background color of the caption contents is employed again to verify the type of caption. If the attributes of the segmented caption contents differ from the constructed attributes of the constructed caption models, then the type of caption may have been incorrectly identified. Figure 4(a) and 4(b) respectively presents an example of the caption model and the individual caption data segmented using the model-based segmentation approach. Following the segmentation process, some of the caption contents, including numeral and graphic information, are further recognized to elucidate the semantics of the videos. Numerals are recognized by performing a binarization process, based on the background color of the segmented numeral images. A threshold for each pixel is given by $T = bg - k$, where bg and k are the gray value of the background and the specific parameter, respectively. Furthermore, a size normalization process based on a bilinear interpolation technique is used to stretch the size of the numeral images. Such enhanced numeral images can be fed into a commercial OCR Development Kit Asprise for the recognition process. Finally, a voting post-processing [9] is adopted to select a series of numerals from the recognition results obtained from the consecutive video frames, to correct few errors in numeral recognition, since the appearance of these numerals is temporally constituent. In the classification of graphic information, the background color in the graphic region is called a classification factor, and distinguishes to the type of graphic information.

4. EXPERIMENTAL RESULTS

Various sports videos encoded in MPEG1 from the TV were collected to evaluate the performance of the proposed approach. A database of the sports videos included 23 baseball, 20 basketball, ten rugby, and nine soccer sports games with various caption types. Each video game was randomly clipped into ten 3min video segments at different time. In the caption identification process, six captions of each type taken from the database were used as training data, and the remainder was used as testing data.

4.1. Performance analysis

In Section 2, the average precision rate was 95.43% and the average recall rate was 89.35%, establishing the effectiveness of the caption extraction approach. Experimental testing that the proposed extraction approach detected most captions because the captions appear stably in the video segments.

In Section 3, Table 1 presents the average identification rate of various sports videos to evaluate the effectiveness of the caption identification process. The average identification rate for 372 sports video segments is 92.88% for the training set; that for 248 sports video segments is 83.78% for testing set. Some captions cannot be identified because the caption shape is changed to present additional information. Additionally, the layout feature may be sensitive to the changes of caption contents causing the feature variation to become large. Finally, OCR is naturally used to evaluate the model-based segmentation approach because the accurate segmentation process can improve the recognition performance. Therefore, the proposed segmentation approach is compared with the projection-based segmentation approach [5, 6], presented in Table 2. The horizontal rectangular captions, as shown in Fig. 1(c), were considered to evaluate the comparative performance, because some captions, as shown in Fig. 1(a), can not be used by the projection-based segmentation approach. The correct, damage, and miss rate of the segmentation process are the percentage of segmented numerals that are actually correct, lose some parts of the numerals, and are not successful, respectively. The accuracy rate was determined by means of the resulting OCR for numeral recognition. The recognition performance from the application of the model-based segmentation process can be improved about 7.72% in testing set.

4.2. Discussion

This section summarizes and discusses several important observations regarding the experimental performance of the caption detection and identification processes. First, the proposed learning-based approach is a flexible way to learn various caption styles on sports videos, without the need of the given caption style [13]. Furthermore, the extraction and identification processes may fail for the following reasons. (1) False caption detection may be caused by the commercial advertisements, because some of the objects within the advertisement segments have the characteristics of the sports captions in the videos. (2) A caption may be missed because the caption appears for short or discontinuous period. (3) The identification errors may be caused by the change of caption content. This effect is likely to cause a large variation in the layout. Finally, the algorithm works well in the video frames with complex background because the video content with more variations is benefit to speed up the process of the caption detection. Besides, if the sports video with slow camera motion, it may take more computational time to function.

5. CONCLUSIONS

This work proposes a novel model-based segmentation approach in digital sports videos, which identifies various sports captions to extract and recognize their contents accurately. The proposed approach uses the identification process to classify an input sports captions into one of the sports caption types, and then corresponding caption model is applied to extract the caption contents accurately. Additionally, the caption region must be detected first to identify the sports captions efficiently. A caption approach, exploits the temporal consistency of the caption appearance before spatial-image analysis. The proposed approach helps to provide a high tolerance to noisy and complex-background video frames because the use of a sequence of video frames is superior to working on a single video frame at a time.

Finally, the proposed approaches to extracting and recognizing captions were tested successfully using various sports captions. The learned-based identification approach presents significant advantages: it is straightforwardly extensible to new sports captions; it is more easy and accurate to extract caption contents.

The authors will extend this study to extract highlight events of sports games or summarize the sports video content by analyzing the caption contents. Moreover, we will combine visual and auditory features to improve the capability of video understanding for various sports games.

6. REFERENCES

[1] Ming Luo, Xuesheng Bai, Guangyou Xu, "Content-based analysis and indexing of sports video," Proceedings of SPIE vol. 4676, p223-231, 2002.

[2] D. T. Chen, J, M. Odobez, Herve Bourlard, "Text Detection and Recognition in Images and Videos," Pattern Recognition" vol. 37, pp. 595-608, 2004.

[3] Rongrong Wang, Wanjun Jin, Lide Wu. "A Novel Video Caption Detection Approach Using Multi-Frame Integration," International Conference on Pattern Recognition, Vol. 1, pp. 449-452, 2004.

[4] David Crandall, Sameer Antani, Rangachar Kasturi, "Extraction of special effects caption text events from digital video," International Journal on Document Analysis and Recognition, Vol. 5, No. 2-3, pp. 138 – 157, 2003.

[5] X. Tang, X. Gao, J. Liu, and H. Zhang "A spatial-temporal approach for video caption detection and recognition," IEEE Transactions on Neural Networks(NN), special issue on intelligent multimedia processing, vol. 13, no. 4, pp. 961-971, July, 2002.

[6] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video OCR for digital news archives," Int. Workshop on Content-based Access of Image and Video Database, pp. 52-80, 1998.

[7] Datong Chen, J. M., 'Video text recognition using sequential Monte Carlo and error voting methods," Pattern Recognition Letters, Vol. 26, No. 9, pp. 1386-1403, 2005.

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

[9] H. Bunke and P. S. P. Wang, "Handbook of Character Recognition and Document Image Analysis," World Scientific, Singapore, pp. 79-101, 1997.

[10] K. Fukunaga, "Introduction to Statistical Pattern Recognition," Second Edition, Academic Press, New York, 1990.

[11] C.T. Zahn and R.Z. Roskies, "Fourier descriptors for plane closed curves," IEEE Trans. Computers, vol. 21, no. 3, pp. 269-281, 1972.

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

[13] 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.

(a) (b) (c)

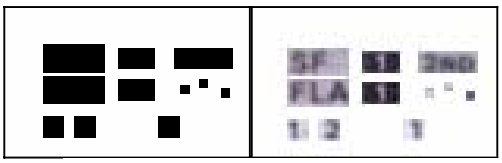
Fig. 1. Some sports captions from (a) baseball games (b) basketball (c) football games.

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

Fig. 2. The results of the caption extraction process; (a) a series of video frames (b) an averaged video frame; (c) a binary image; (d) an enhanced image; (e) an extracted caption image.

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

Fig. 3. The results of the contour and layout extraction process; (a) a caption region; (b) a contour image; (c) an edge image; (d) the layout of the caption image.



(a) (b)

Fig. 4. (a) A caption masking model; (b) the extracted caption data.

Table 1. The performance of the caption identification process.

Sports Videos	Baseball	Basketball	Rugby	Soccer
IR for training set	96.77%	96.55%	90.7%	87.5%
IR for test set	86.02%	81.61%	87.5%	80%

IR: Identification Rate

Table 2. The comparative results of the segmentation approaches.

	Correctness	Damage	Miss	Accuracy
Projection -based	72.5%	8.4%	19.1%	77.48%
Model-based	100%	0%	0%	85.2%