International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
332 – 337

Super Imposed Method for Text Extraction in a Sports Video

Tabish Sayyed, Dinesh Barai, Snehal Kande

Department of Computer Engineering,
Bapurao Deshmukh College of Engineering, Sewagram,Wardha

Abstract:- Video is one of the sources for presenting the valuable information. It contains sequence of video images, audio and text information. Text data present in video contain useful information for automatic annotation, structuring, mining, indexing and retrieval of video. Nowadays mechanically added (superimposed) text in video sequences provides useful information about their contents. It provides supplemental but important information for video indexing and retrieval. A large number of techniques have been proposed to address this problem. This paper provides a novel method of detecting video text regions containing player information and score in sports videos. It also proposes an improved algorithm for the automatic extraction of super imposed text in sports video. First, we identified key frames from video using the Color Histogram technique to minimize the number of video frames. Then, the key images were converted into gray images for the efficient text detection. Generally, the super imposed text displayed in bottom part of the image in the sports video. So, we cropped the text image regions in the gray image which contains the text information. Then we applied the canny edge detection algorithms for text edge detection. The ESPN cricket video data was taken for our experiment and extracted the super imposed text region in the sports video. Using the OCR tool, the text region image was converted as ASCII text and the result was verified.

Keywords:- *Video Retrieval, Video Annotation, Sports video, video summarization, Video text information and super imposed text*

******

1. INTRODUCTION

Increasing availability of video data has rekindled interest in the problems of how to index video information automatically and how to browse and manipulate them efficiently. Traditionally, the images and video sequences have been manually annotated with a small number of key word descriptors after visual inspection by the human reviewer. This process can be very time consuming. Text Information retrieval from video images (Video Annotation) has become an increasingly important research area in recent years for the video information retrieval and video mining applications. Detection and recognition of text captions embedded in image frames of videos is an important component for video retrieval and indexing. Video text detection and extraction is an important step for information retrieval and indexing in video images [1]. Recognizing video text information directly from video provides unique benefits. They are (i) It is very useful for describing the contents of video sequence; (ii) it can be easily extracted and compared to other semantic contents; (iii)

Extracted text is exactly synchronized with the image data when the event occurs; (iv) manual logging may not be feasible for large collection of archived videos; (v) it enables applications such as keyword-based image search, automatic video logging, and text-based image indexing; (iv) It is an important component for the automatic annotation, indexing, and parsing of images and videos. For example, the sports videos text information displays valuable game information [2], such as scores and players. The valuable text information extracted from sports videos is called key captions text, and they can be used for video highlights or content search.

Video text may be divided into two types: *Scene text* (which exist in the real-world objects and scenes) and *Superimposedtext or Graphic text (*which are added during editingprocesses) [1][3][4]. Scene text appears within the scene which is then captured by the recording device. It is showing naturally in scenes like text on cloth, street signs, bill boards, and text on vehicle and etc. The appearance of the text is typically incidental to the scene content and only useful in applications such as in navigation, surveillance or reading text appearing on the known objects rather than general indexing and retrieval. It is difficult to detect and extract since it may appear in a virtually unlimited number of poses, size, shapes and colors. Super imposed text is mechanically added into the video frame to supplement the visual and audio content, and is often more structured and closely related to the subject then the scene text is. The superimposed text is one powerful source of high-level semantics. If these text occurrences could be detected, segmented, and recognized automatically, they would be a valuable source of high-level semantics for indexing and retrieval. An example of the superimposed text includes headlines, Keyword summaries, time and locations stamps, names of the people and sports scores. The superimposed

 ISSN: 2454-4248
332 – 337

text is the most reliable clue for enable users to quickly locate their interested content in an enormous quantity of video data, many research efforts have been put into video indexing and summarization. Three are basically three reasons: 1) it is closely related to the current content of video;

2) It is has distinctive visual characteristic; and 3) the state-of-art optical character recognition (OCR) techniques are far more robust than the existing speech analysis techniques and visual object analysis. It has a number of functions which differ between the domains. The extraction of the superimposed text in sports video is very useful for the creation of sports summary, highlights and etc [5]. It is also useful for identification of channel logos of television media and flash news can be extracted which is displayed in the bottom of the news video frame.
Text in video sequences can exhibit many variations with respect to the following properties [6]:

(i). Geometry:

a). Size: The text size assumptions can be made depending on the application domain.

b). Alignment: The characters in the super imposed text appear in clusters and usually lie horizontally, although sometimes they can appear as non-planar texts as a result of special effects. This does not apply to scene text, which can have various perspective distortions. Scene text can be aligned in any direction and can have geometric distortions.

c). Inter-character distance: characters in a text line have a uniform distance between them.

(ii). Color: The characters in a text line tend to have the same or similar colors. However, video images and other complex color documents can contain „text strings with more than two colors (polychrome)" for effective visualization, i.e., different colors within one word.

(iii).Motion: The same characters usually exist in consecutive frames in a video with or without movement. This property is used in text tracking and enhancement. Caption text usually moves in a uniform way: horizontally or vertically. Scene text can have arbitrary motion due to camera or object movement.

(iv).Edge: Most super imposed and scene texts are designed to be easily read, thereby resulting in strong edges at the boundaries of text and background.

(v). Compression: Many digital images are recorded, transferred, and processed in a compressed format.

The remainder of this paper is organized as follows: Section 2 introduces the background of superimposed video text detection and extraction on sports videos. Section 3 describes the methodology used to detect regions of super imposed text in a sports video. Section 4 presents experimental results. Section 5 draws some conclusions.

2. BACKGROUND

From observation on TV program, normally the superimposed text is displayed only a single line. The super imposed text falls in the two categories. (i). Moving text in the video frame. i.e., the advertisement messages/text and flash news may be moving from left to right or right to left. In general, text motion can divided into three classes. There are static, simple linear motion (for example, scrolling movie credits) and complex nonlinear motion (for example, zooming in/out, rotation, or free movement of scene text), respectively. (ii). Non-moving Text in a video frame, i.e, the sports score card display changing the score display without moving.

The algorithms for text detection can be classified in two categories [7], [8] those working on the compressed domain and those working on the spatial domain. Compressed domain includes both in the compressed and in the semi-compressed domain. It is based on the localization of static characters over moving background taking into account the macro-blocks belonging to P frames Moreover it assumes that text has horizontal geometry, that it does not occupy the whole frame and that it has to appear at least in three frames. These three features allow the algorithm to isolate macro-blocks and to determinate if the macro blocks are candidates to contain text. Both recall and precision are high in those sequences with moving background and static text, like sports sequence (e.g. score in a football match). But it cannot be used in sequences containing moving text or static background. Semi-compressed domain algorithms don't work directly with macro -blocks but analyzing the Discrete Cosines Transform components.

Spatial domain methods work with the pixel values and positions are called methods in the spatial domain and they can be classified according to the following image features. They may be generally grouped into four categories: i). Connected component methods; ii). Texture classification methods; iii). Correlation based methods and iv). Edge detection methods. The connected component methods detect text by extracting the connected components of monotonous colours that obey certain size, shape, and spatial alignment constraints. The texture-based methods

 ISSN: 2454-4248
332 – 337

treat the text region as a special type of texture and employ conventional texture classification method to extract the text. The correlation based methods are those that use any kind of correlation in order to decide if a pixel belongs to a character or not. Edge detection methods have been increasingly used for caption extraction due to the rich edge concentration in characters. The edge based text detection method can classify into two categories, one is in compressed domain and the other is in pixel domain. In compressed domain, it utilizes the Discrete Cosines Transform (DCT) coefficients to measure the horizontal and vertical intensity variation of each DCT block in an I-frame. In pixel domain, it employs Canny, Sobel or Gaussian filter to perform edge detection.

The text detection problem involves locating regions in a video frame that contain text. It refers to the determination of the presence of text in a given frame (normally text detection is used for a sequence of images). Text Region Extraction is the process of determining the location of text in the image. The video text detection is based on the special characteristics of the text such as contrast, color, font size, font shape, orientation and stationary location.

A sport video is one of the complex video databases. It is a popular component in any broadcast television media. The superimposed text always displays in bottom or top of the sports video. The text information provides valuable game information such as scores, players, and so on. This video text information plays an important role for sports video understanding, summarization, and retrieval.

Some researchers have investigated for video text information. Keechul Jung, Kwang In Kim, and Anil K. Jain

[6] presented the survey of text information extraction in Images and Videos. Xingquan Zhu,., Xindong Wu, Ahmed K. Elmagarmid, Zhe Feng, and Lide Wu, [9] presented a method for extracting the super imposed text region. Then, they applied the video text to generation of hybrid sequence stream. From the Hybrid stream, they were extracted the sports events and summary. Dongqing Zhang and Shih-Fu Chang [10] have developed a novel system for baseball video event detection and summarization using superimposed caption text detection and recognition. Christian Wolf and Jean-Michel Jolion [11] presented an algorithm to localize artificial text in videos using a measure of accumulated gradients and morphological processing. Jovanka Malobabiæ, Noel O'Connor, Noel Murphy, Sean Marlow [12] presented an algorithm for detection and localization of artificial text in video using a horizontal difference magnitude measure and morphological processing.

3. METHODOLOGY

In this paper, we provide a robust detection method for super imposed text in sports videos. We presented an efficient algorithm for super imposed text region extraction of super imposed text in video. The extraction of super imposed text in a video frame consists of six steps. They are, 1. Video Frame Extraction; 2. Key Frame Extraction; 3. Grayscale Conversion; 4. Cropping the Video Image; 5. Canny Edge Detection; 6. Text region Retrieval.

The video consists of sequence of images (video frames). In the first step, we extracted all frames in the video and saved as JPEG images. To reduce the number of frames, in the second step key-frame selection was performed. We have applied the Histogram technique for selecting the key frames. The selected the key frames have difference greater than given threshold value. The threshold value was selected depending on the video domain. Key frame is the frame which has major difference from previous frame. It uses adjacent frame subtraction with threshold value. The next process is grayscale conversion. It uses the grayscale conversion equation gs = red*0.5+green*0.3+blue*0.2 to convert the image into grayscale images. The images which are converted into gray scale are easy to process the image. It usually gray of the color picture to elevate processing speed and decrease resource occupation when processing images. Then next step is to crop the images into one tenth of its original size. Because of the score cards in the sports video are visible only in bottom of frame or top of the frame. The region consists of inning details, score, out, ball count, player name. The regions are called as important regions. The other regions of the image are not useful. So, we discarded the image regions. The processing time and memory required was also reduced in cropped image process. The nature of the cropping varies with the nature of the video database. The cropped Images were stored and further processes were carried out.

The Canny Edge Detection algorithm was applied for the extraction of the text region in the cropped image. In three ways canny edge detection is efficient than the ordinary edge detectors.

A. Good detection - the algorithm marks as many real edges in the image as possible.
B. Good localization - edges marked will be as close as possible to the edge in the real image.
C. Minimal response - a given edge in the image only be marked once, and where possible, image noise will not create false edges.

The canny edge detection algorithm is easy to implement, and more efficient than other algorithms. From this edge

 ISSN: 2454-4248
332 – 337

detected images, text region is identified. The flowchart of the proposed method is illustrated in Fig. 1.

The extracted regions can then be used as input to an appropriate Optical Character Recognition system which produces index-able keywords.

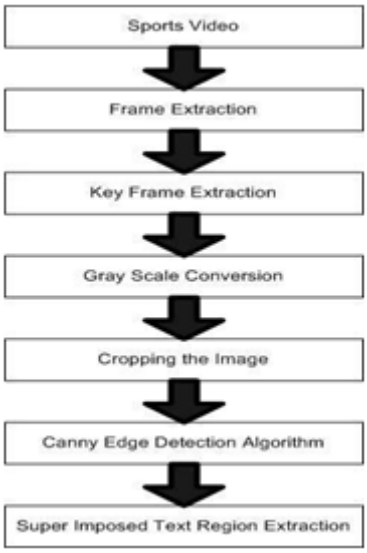

Fig 1: An example of text detection in pixel domain by edge based method

The Optical Character Recognition engine takes an image as the input and yields an ASCII string result. Text embedded in video (closed caption) is a powerful keyword resource in building video retrieval and annotation system. Generating captions or annotations automatically for video is a challenging task. It enables text-based querying and content summarization. The video text plays important role in the video data mining applications. The proposed algorithm is one of efficient algorithm for extraction the video caption text. It applied in the other videos like Movies (the translated texts displayed in the bottom of the page) and News video. The extracted text is closely related to the frame or video sequence. So, the text is useful for video data mining tasks.

4. EXPERIMENTS

The superimposed text extraction system receives an input in the form of a still image or a sequence of images. We have collected the ESPN sports video database for about five minutes video for the superimposed text extraction. The system was developed using Java and Java Media Framework. First we extracted the key frames using the Histogram based image comparison shown in the Fig. 2.

Fig 2: Key Frames

Then key frame images are converted as Grayscale images for easy process. Fig 3 shows the grayscale images.

ISSN: 2454-4248
332 – 337

Fig 3: Grayscale Images

Then one by tenth of the gray scale images were cropped for the super text extraction shown in the Fig 4. Because of the sports channel contains the score in the bottom of the screen. Then the canny edge detection algorithm is applied on the cropped images to extract the text area. The Fig 5 shows the edge detected images.

Fig 4: Cropped Images

Fig 5: Canny Edge Detection

Finally, the super imposed text region extracted and is shown in Fig 6.

Fig 6: Text Region

The extracted text images can be transformed into ASCII text using OCR tool. The ASCII text is used to video indexing and retrieve the video.

To evaluate the performance of the proposed super imposed text extraction approach, we adopt two widely used metrics such as, Recall-Precision and Accuracy is presented in the Table 1.

 ISSN: 2454-4248
332 – 337

Table 1: Performance Results

	Correct = Y	Correct = N
Assigned =Y	A= 42	B= 8
Assigned =N	C= 6	D= 9
Performance Results		
Precision	(A/A+B)	0.82
Recall	(A/A+C)	0.88
Accuracy	((A+D) / (A+B+C+D))	0.79

It is worthwhile to mention that the current system is not optimized and the performance is expected to improve if the implementation details are fine-tuned. The main purpose of showing these results is to demonstrate that the proposed system offers a promising direction and worth further direction.

5. CONCLUSION

There are numerous applications of a video text information extraction system, including vehicle license plate extraction (surveillance video data); text based video indexing, video content analysis and video event identification. In this paper, we proposed a robust extraction method of text information in sports videos. Our proposed algorithm can only detect the videotext in the boundary of the image. In general, there is little valuable information in non-play shots because these non-play shots include scenes when the game is temporally halted by scoring, no run play, or time-out, and do not carry any meanings related to the content of a game. In future research, we have to extract only the play shots. We can combine the audio information such as applause and cheering to generate more exciting sports summaries. However, the text extraction results are inappropriate for general OCR software: text enhancement is needed for low quality video images and more adaptability is required for general cases (e.g., inverse characters, 2D or 3D deformed characters, polychrome characters, and so on). Video text has some temporal features:

(1). the same text usually appears over several continuous frames. (2). motive characters have a dominant translation direction: linearly horizontal or vertical direction. (3). when text appear or disappear, text region color change significantly in adjacent frames. The Temporal features are also considered for video super imposed text detection.

6. REFERENCES

[1]. Luo B., Tang X., Liu J., and Zhang H.-J., "Video caption detection and extraction using temporal information," In ICIP'03, pages 297-300, 2003, Barcelona, Spain, September.

[2]. Cheolkon Jung and Joongkyu Kim, "A Novel Approach for Key Caption Detection in Golf Videos Using Color Patterns," ETRI Journal, Volume 30, Number 5, pp-750-753, October 2008.

[3]. David Crandall, Sameer Antani, Rangachar Kasturi, "Extraction of special effects caption text events from digital video," International journal on document analysis and recognition, Volume: 5, pp:138- 157, 2003.

[4]. Huiping Li Doermann, D. Kia, O.," Automatic text detection and tracking in digital video," IEEETransactions on Image Processing, Volume: 9, Issue: 1 On page(s): 147-156, Jan 2000.

[5]. Cheolkon Jung, Su Young Lee, Joongkyu Kim , "Robust Detection Of Key Captions For Sports Video Understanding," IEEE International Conference on Image Processing (ICIP 2008), pp-2520-2524, 2008.

[6]. Keechul Jung, Kwang In Kim, and Anil K. Jain. "Text information extraction in images and video: a survey," International Journal of Pattern Recognition, Volume: 37, Number:5, pp:977–997, 2004.

[7]. Miriam León, Antoni Gasull , "Text Detection In Images And Video Sequences," Source:http://gps-tsc.upc.es/imatge/pub/ps/IASTED05_Leon_Mallo_Gasull.pdf

[8]. Miriam Le´on, Sergio Mallo and Antoni Gasull, "A Tree Structured-Based Caption Text Detection Approach," proceedings of the fifth IASTED International Conference Visualization, Imaging, and Image Processing , pp 220-226, September 7-9, 2005, Spain.

[9]. Xingquan Zhu., Xindong Wu, Ahmed K. Elmagarmid, Zhe Feng, and Lide Wu, "Video Data Mining : Semantic Indexing and Event Detection from the Association Perspective," IEEE Transactions of Knowledge and Data Engineering., Vol: 17, No : 5, 2005.

[10]. Zhang and S.F. Chang, "Event Detection in Baseball VideoUsing Superimposed Caption Recognition," Proc. ACM Int"l.Conf. on Multimedia, pp. 315-318, 2002.

[11]. Christian Wolf and Jean-Michel Jolion, "Extraction and recognition of artificial text in multimedia documents," Pattern Analysis and Applications (PAA), Volume 6, Number 4, February 2004 , pp. 309-326(18), Springer

[12]. Malobabić, Jovanka and O'Connor, Noel E. and Murphy, Noel and Marlow, Seán "Automatic detection and extraction of artificial text in video," In: WIAMIS 2004 - 5th International Workshop on Image Analysis for Multimedia Interactive Services, pp:21-23 April 2004, Lisbon, Portugal.