Volume 6, Issue 3, March 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Video Mining: An Enhanced Model for Extracting Text from Video

Neeru Mago
Department of Computer Science & Applications
PUSSGRC, Hoshiarpur, Punjab, India

***Abstract**—Video mining consists of the analysis of content-based classification, indexing, and retrieval. The representation, browsing, and visualization of the features in the video is also included in video mining. The text data present in images and video contain certain useful information for automatic annotation, indexing, and structuring of images [1]. Extracting text from video becomes extremely difficult and challenging job due to differences in text style, font, size, orientation and alignment. The low image contrast and complex background may also be the factors for making this task tough. A large number of techniques have been proposed to address this problem. However the problem of text information extraction is not well surveyed. This paper describes the classification and review of these algorithms, discuss performance evaluation, and to point out promising directions in this field. Further, an enhanced model for extracting text from video is also presented in this paper.*

***Keywords**— Text Extraction, Text Localization, Text Recognition, Text Segmentation, Video Text*

I. INTRODUCTION

In many applications like image indexing, document processing, video understanding, video retrieval and video content summary, the most important problem is how to extract text from images or videos [3]. There are three categories of text appearing in images: document text, caption text, and scene text [2]. Document text: A document image (Fig. 1) consists of few graphic components and major part is text. It is generally obtained by scanning printed documents, journals, handwritten historical document and book cover etc. Caption text: At the time of editing, the caption text is artificially superimposed on the image. For example, subtitles and the subject of the image content, etc. It is also called overlay text or artificial text (Fig. 2). Scene text: It contains text on the natural part of the scene image and describes important semantic information such as food containers, street signs, bill boards, banners, advertisements, names of streets, institutes, shops, road signs, traffic information, board signs, nameplates and text on vehicle etc (Fig. 3). It is more difficult to detect scene text as compared to caption text; since scene text can have any orientation and may be distorted by the perspective projection therefore.

There are many sources of video contents like internet, TV broadcasting and surveillance cameras. These videos include texts of any type such as scrolling texts or caption text made by artificial overlaying after recording and scene text embedded in backgrounds. Very useful information can be extracted from the text embedded in images. An exact and meaningful keyword can be provided from the text extracted from video clips which helps in indexing and summarizing the content of video clip [1].

II. RELATED WORK

There are many techniques and methods developed and used for extracting text from videos and images. In scanned documents, currently available optical character recognition (OCR) systems can achieve almost perfect recognition rate on printed text [4], but it is difficult to precisely recognize text information directly from camera-captured scene images and videos. A more precise model is developed by Lu et al. [3] that defines a dictionary of basic shape codes to perform character and word retrieval without OCR on scanned documents and describes the inner character structure of an image. Local features of character patches can be extracted from an unsupervised learning method in a paper by Coates et al. [5] that links a variant of K-means clustering and united it by cascading sub-patch features of an image. To design a discriminative feature representation of scene text character structure, a complete performance evaluation of scene text character recognition was carried out [8]. Weinman et al. [7] combined the Gabor-based appearance model, a language model related to simultaneity frequency and letter case, similarity model, and lexicon model to perform scene character recognition. A real time scene text recognition and localization method based on extreme regions is proposed by Neumann et al. [1]. Smith et al. [2] used integer programming to build a similarity model of scene text characters based on SIFT and maximized posterior probability of similarity constraints. Mishra et al. [9] adopted conditional random field to combine bottom-up character recognition and top-down word level recognition.

Smith and Kanade [7] used a method that define a scene-change by calculating the difference between two consecutive frames and then applied this scene-change information for text detection. An accuracy of 90% is achieved by this method. Gargi et al. performed text detection with an assumption that when a text caption appears, the number of intra coded blocks in P- and B- frames of an MPEG compressed video increases. Lim et al. developed an extremely simple and fast method with an assumption that text generally has a higher intensity than the background. The number of pixels that are lighter than a predefined threshold value were counted and a significant color difference relative to their neighbourhood was exhibited. A frame with a large number of such pixels was regarded as a text frame. However, problems can still occur with color-reversed text. A brief overview of the existing text extraction methods and their features are listed in Table 1.

Table 1 Existing Text Extraction Methods And Their Features

AUTHOR	METHOD	FEATURES
Lee and Kankanhalli	Coarse search using edge information, followed by connected component generation	✓ Scene text (cargo container) ✓ localization and recognition
Ohya et. al.	Adaptive thresh holding and relaxation operations	✓ Color, scene text (train, signboard, skew and curved) ✓ localization and recognition
Smith and Kanade	3*3 filter seeking vertical edges	✓ Caption text ✓ Localization
Zhong et. al.	CC- based method after color reduction, local spatial variance-based and hybrid method	✓ Scene text (CD covers) ✓ Localization
Yeo and Lin	Localization based on large inter-frame difference in MPEG compressed image	✓ Caption text ✓ Localization
Shim et. al.	Gray level difference between pairs of pixels	✓ Caption text ✓ Localization
Sato et. al.	Smith and Kanade's localization method and recognition-based character extraction	✓ Recognition
Jain and Yu	CC- based method after multi-valued color image decomposition	✓ Color (book cover, web image, video frame) ✓ Localization
Chun et. al.	Filtering using neural network after FFT	✓ Caption text ✓ Localization
Antani et. al.	Multiple algorithms in functional parallelism	✓ Scene text ✓ Recognition
Messelodi and Modena	CC generation, followed by text line selection using divisive hierarchical clustering procedure	✓ Scene images (book covers, slanted) ✓ Localization
Wu et. al.	Localization based on multi-scale texture segmentation	✓ Video and Scene images (newspaper, advertisement) ✓ Recognition
Hasan and Karam	Morphological approach	✓ Scene text ✓ Localization
Li et. al.	Wavelet-based feature extraction and neural network for texture analysis	✓ Scene text (slanted) ✓ Localization ✓ Enhancement ✓ Tracking
Lim et. al.	Text detection and localization using DCT coefficient and macro block type information	✓ Caption text ✓ MPEG compressed video ✓ Localization
Zhong et. al.	Texture analysis in DCT compressed domain	✓ Caption text ✓ JPEG and I-frames of MPEG ✓ Localization
Chen et. al.	Text detection in edge-enhanced image	✓ Caption text ✓ Recognition ✓ Localization
Strouthopoulos et. al.	Page layout analysis after adaptive color reduction	✓ Color document image ✓ Localization
Jung	Gabor filter-like multi-layer perception for texture analysis	✓ Color ✓ Caption text ✓ Localization

III. TEXT EXTRACTION FROM VIDEOS

In fig 4, an enhanced model for extracting text from video is presented. Various stages in the process of extracting text from video are shown. The input to the system can be a real time video or a video from stored database.

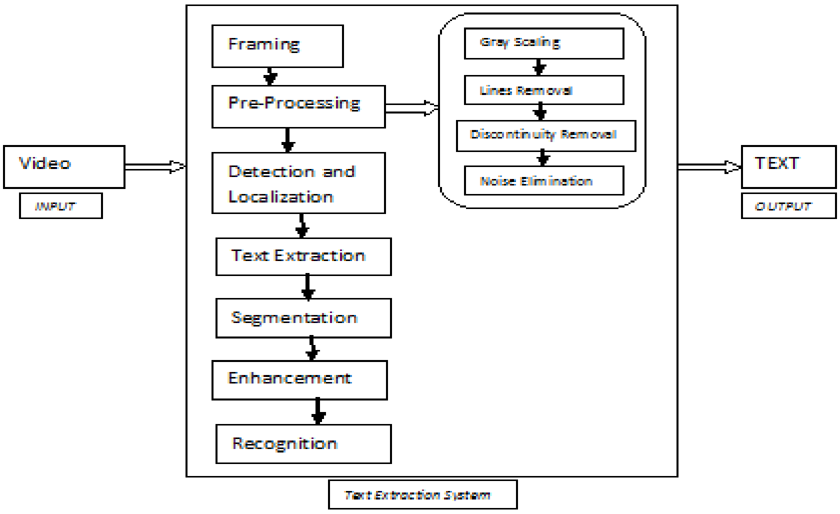

Fig 4: Process Of Text Extraction From Video

A. Framing of Video

Java Media Framework (JMF) can be used to capture the contents of media and frame the video. JMF is a framework that handles streaming media in Java programs. It is an optional package of Java 2 standard platform that provides a unified architecture. For managing the acquisition, processing and delivery of time-based media, JMF provides a messaging protocol. Various features of JMF are:

1) Enables Java programs to present or playback multimedia contents.
2) Capture audio through microphone and video through camera.
3) Perform real-time streaming of media over the Internet.
4) Process media such as changing media format, adding special effects.
5) Store media into a file.
6) Provides a platform-neutral framework for handling multimedia.

The system takes a video containing text as an input in this stage. By using JMF, usually at the rate of 1 frame per second, the video is then framed into images. Depending on the speed of the video i.e. on the basis of fps (frames per second), this rate could be increased or decreased. The images signal is then scaled at a particular resolution and saved at the specified location on the hard disk drive which is then given as an input to the next stage [7].

B. Pre Processing

A scaled image (the input from the previous stage) is then converted into a gray scaled image. **The first stage of the pre-processing part is formed by this gray scaled image.** This process first considers the RGB color contents of each pixel of the image and then converts them to grayscale. The purpose of converting a colored image to a gray scaled image is the easier recognition of the text appearing in the images. Since after gray scaling, the image is converted to a black and white image containing black text with a higher contrast on white background, it becomes easy to understand and recognize text on the image. **The second stage of pre-processing is lines removal**. A video may include some noise i.e either a horizontal fluctuation or a vertical fluctuation. Thus, we need to remove these horizontal and vertical lines across the screen on the video for the successful recognition of the text appearing in the frames. This is done by clearing all pixels i.e by changing pixel color from black to white. This stage will only remove lines if the video frame contains any horizontal and vertical fluctuations. Otherwise, the image will remain same [6]. **The third stage of pre-processing is discontinuities removals** that are created during the second stage of pre-processing. If the horizontal and vertical fluctuations appeared exactly where the text is present, then it creates discontinuities between the text appearing in the video frame. This makes the recognition of the text very difficult. In this stage, scanning of each pixel from top left to bottom right takes place and each pixel and all its neighbouring pixels are considered. If a pixel under consideration is white, and all the neighbouring pixels are black, then that corresponding pixel is set as black. All the black neighbouring pixels indicate that the pixel under consideration is cleared at the lines removal stage [3]. **The final output of pre-processing stage is noise elimination**. This is again carried out by scanning each pixel from top left to bottom right and taking into consideration each pixel and all its neighbouring pixels. [2].

C. Detection and Localization

In detection stage, existence of text in the image is determined since there is no prior information on whether or not the input image contains any text. However, the number of frames containing text in a video is much smaller than the number of frames without text. The text detection stage aims to detect the presence of text in a given image. The frame containing text is selected from shots elected by video framing. Since the portion occupied by a text region relative to the whole image is usually small, a very low threshold values are needed for scene change detection. This can be a simple and efficient solution for video indexing applications. After detection, the localization stage includes localizing the text in the image. In other words, the text present in the frame was tracked by identifying boxes or regions of similar pixel intensity values.

D. Text Extraction

After determining the location of text and generating bounding boxes around it, text is extracted i.e. segmented from the background. The various techniques of text extraction are as follow:

1) Region based Method: The properties of the color or gray scale in the text region or their differences to the related properties of the background are used in Region-based method. There is very little variation of color within text and this color is sufficiently different from text's immediate background. Text can be obtained by limiting the image at intensity level in between the text color and that of its immediate background. This method is not robust to complex background. This method is further divided into two sub-approaches: connected component (CC) and edge based.
   - .i. CC based Method: CC-based methods use a bottom-up approach. It combines smaller components into successively larger components until all regions are identified in the image. To merge the text components using the spatial arrangement of those components, a geometrical analysis is required. It filters out non-text components and the boundaries of the text regions are marked. This method locates text quickly but it fails for complex background.
   - ii. Edge based Method: Regardless of color/intensity, layout, orientations, etc, edges are a reliable feature of text. Edge based method is focused on high contrast between the text and the background. The three distinguishing properties of text embedded in images can be used for detecting text. They are density, edge strength and the orientation variance. Edge-based text extraction algorithm is a general-purpose method. It can effectively localize and extract the text from both document and images but this method is not robust for handling large size text.
2) Texture based Method: This method is based on the fact that text in images has discrete textural properties which differentiate them from the background. This method is based on Gabor filters, Wavelet, Fast fourier transform (FFT), spatial variance, etc. They are used to detect the textual properties of the text region in the image. This method has the ability to detect the text in the complex background. The only drawback of this method is large computational complexity in texture classification stage.
3) Morphological based Method: It is a topological and geometrical based method for image analysis. For character recognition and document analysis, morphological feature extraction techniques have been efficiently applied. It is used to extract important text contrast features from the processed images. These features are invariant against various geometrical image changes like translation, rotation, and scaling. Even after the lightning condition or text color is changed, the feature still can be maintained. This method works robustly under different image alterations

E. Segmentation

In this stage, the text tracking and extraction techniques are applied. After the text is localised, text tracking step deals with separation of text pixels from background pixels. The output of this step is a binary image where black text character appears on a white background. This stage includes extraction of actual text region by dividing pixels with similar properties into segment. Further, it discards the redundant portions of frame.

F. Enhancement

As the text region usually has low-resolution and is prone to noise Enhancement of the extracted text is required. Thereafter, OCR can be used to recognize the extracted text.

G. Recognition

In this stage, actual recognition of extracted characters is performed by combining various features extracted in previous stages. The characters contained in the image are compared with the pre-defined neural network training set and depending on the value of the character appearing in the image, the character representing the closest training set value is displayed as recognized character.

IV. CONCLUSION

In this paper, a comprehensive survey of text extraction methods from images and video is provided. In the literature review, we have studied a large number of algorithms proposed by many authors; but no single method can provide satisfactory performance. Every application has large variations in character font, size, texture, color, etc. There are various sources for text information extraction in images (e.g., color, texture, motion, shape, geometry, etc). It is always beneficial to merge these different sources to enhance the performance of a text information extraction system. Several applications have already been presented such as an automatic video indexing system. Many researchers have already investigated text localization, text detection and tracking for video images. However, their text extraction results are not up to the mark for general OCR software. Hence we propose an advance text extraction system by introducing text enhancement stage for low quality video images and for more adaptability. An enhanced text extraction system can be used for any type of image, including both scanned document images and real scene images through a video camera.

REFERENCES

[1] Mona Saudagar, S. V. Jain "A study of multi-oriented text recognition in natural scene images", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 12, December 2014.

[2] Divya gera, Neelu Jain, "Comparison of Text Extraction Techniques-A Review", International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2015.

[3] Jayshree Ghorpade, Raviraj Palvankar, Ajinkya Patankar and Snehal Rathi, "EXTRACTING TEXT FROM VIDEO", Signal & Image Processing : An International Journal (SIPIJ) Vol.2, No.2, June 2011.

[4] M. Prabaharan, K. Radha, "Text Extraction from Natural Scene Images and Conversion to Audio in Smart Phone Applications", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 1, January 2015.

[5] Sneha S. Kapse, Prof. Pravin Kshirsagar, "Text Based Video Indexing and Retrieval by Using DLER Technique", International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 4, Issue 7, July 2015.

[6] Sumit R. Dhobale, Prof. Akhilesh A. Tayade, "A survey on Text Retrieval from Video", International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 3, Issue 11, November 2014.

[7] Rosy K Philip, Gopu Darsan, "A survey on Text Extraction Techniques in Complex Images and Videos", International Journal of Advanced Technology in Engineering and Science Volume No 03, Special Issue No. 01, March 2015.

[8] Prajakta D. Sawle, Chetan J. Shelke, "Text Extraction in Android Mobile Application using Character Descriptor", International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 4, April 2015.

[9] Kadhim Mahdi Al-Musawi, "Arabic Text Extraction from Video Film", International Journal of Comp Sci and Mobile Comp., Vol.4 Issue.5 May-2015, pg. 1117-1123.

[10] A Survey Keechul Jung, Kwang In Kim, Anil K. Jain , "Text Information Extraction in Images and Video".

[11] R.Bhavadharani et al, "A Dynamic Approach to Extract Texts and Captions from videos", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April-2014, pg. 1047-1052.

[12] Shilpi Rani, Rakesh Kumar Yadav, "An Efficient Method for Texst Extraction from Colored Images", International Journal of Computer Applications (0975 –8887)Volume 96–No.13, June 2014.

[13] Deepika Sood, Baljit Singh, "Extraction of Text From Video Clips", International Journal of Advances in Science Engineering and Technology , ISSN: 2321-9009 Volume-1, Issue-3, Jan.-2014.

[14] Shivananda V. Seeri, J. D. Pujari and P. S. Hiremath, "Multilingual Text Localization in Natural Scene Images using Wavelet based Edge Features and Fuzzy Classification", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),Volume 4, Issue 1, January-February 2015 .

[15] S.J.Wamane, T.A.More, "Embedded Technology Based Image and Video Data Extraction", International Journal of Innovative Research in Science, Engineering and Technology Volume 3, Special Issue 4, April 2014.

[16] Samabia Tehsin, Asif Masood and Sumaira Kausar," Survey of Region-Based Text Extraction Techniques for Efficient Indexing of Image/Video Retrieval", .J. Image, Graphics and Signal Processing, November 2014 in MECS.

[17] Latika R. Desai, Poonam B. Kadam and Swati Shinde, "Review on Text Detection Methodology from Images", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014.