International Journal of Advanced Technology in Engineering and Science www.ijates.com
Volume No 03, Special Issue No. 01, March 2015 ISSN (online): 2348 – 7550

A SURVEY ON TEXT EXTRACTION TECHNIQUES IN COMPLEX IMAGES AND VIDEOS

Rosy K Philip[1], Gopu Darsan[2]

[1,2]Dept of computer Science

Sree Buddha College of Engineering, Pattoor, Alappuzha, Kerala, (India)

ABSTRACT

Text data present in images and videos mainly contains useful information for fully understanding it and helps for the automatic indexing, annotation and structuring of images. Extraction of this information involves several steps such as detection, localization, tracking, extraction, and recognition of the text from a given image. Text in images varies with differences in style, size, orientation, and alignment. The low image contrast and complex background makes the problem of automatic text extraction extremely challenging. A number of techniques have been proposed to solve this problem. The purpose of this paper is to classify and review these algorithms, also discuss the performance evaluation, and to point out promising directions for future research. This further open up the possibility for more improved and advanced systems.

Index Terms—Connected Component, CVPR, Edge Detection, Text Extraction

I. INTRODUCTION

In recent years , there is a drastic increase in multimedia libraries. The amount of digital multimedia data is growing exponentially with time. Number of television stations are broadcasting every day. With wide spread of affordable digital cameras and inexpensive memory devices, multimedia data is increasing every second. Ranging from cameras embedded in mobile phones to professional ones, Surveillance cameras to broadcast videos, every day images to satellite images, all these contributes to increase in multimedia data. According to Flickr statistics in 2013, 43 million images per month are uploaded that is in average 1.42 million per day [1]. And according to youtube official announcement, 72 hours of videos are uploaded to the site every minute and watched over 3 billion hours a month [2]. With this dramatic increase in multimedia data, escalating trend of internet, and amplifying use of image or video capturing devices the content based indexing and text extraction have more importance among the researchers.

Text embedded in images contains the useful semantic information regarding the image .This information can be used to fully understand images. Text within an image enables applications such as keyword-based image search, automatic video logging, and text-based image indexing.Extraction of text from images is a very difficult task due to variations in character fonts, styles, sizes and text directions, and presence of complex backgrounds and variable light conditions.Text regions may contain very useful information regarding the image.

Generally, the text present in images can be categorized into three based on the type of images:document images, scene images and born-digital images. Document images are the image-format of the document. The text present in document image is the document text. Born-digital images are the images generated by computer software and are saved as digital images. The text in these images is called as caption text or overlay text. Some researchers specify overlay text as superimposed text or artificial text. The overlay text can be seen as the text scrolling in the news channels. Scene images are the images that contain the text, such as the advertising boards, banners, which is captured naturally when the scene images are taken by the camera, therefore scene text is embedded in the background as a part of the scene taken. Compared withdocument images and born-digital images, the scene images, have more complex foreground/ background, low resolution, compression loss, and severe edge softness. This makes the extraction of text from scene text more difficult. Therefore automatic extraction of texts from images or video is a challenging task and research under this field is still under progress.

Text present in images or videos has certain properties and characteristics as described below

(a)Size: The text can be of variable size.

(b)Alignment: The caption texts appear in clusters and usually in horizontal direction. But this does not apply to scene text that has various perspective distortions. The Scene text may be aligned in any direction.

(c)Edge: An edge in the images is the most reliable feature of text regardless of its color, intensity, orientations, etc.

(d)Color: In a simple image the characters in a text usually have the similar or same color. This property mainly makes use of connected component-based approach for text detection. But the video images and other complex color documents usually contain text strings with two or more colors.

(e)Motion: The same text or characters usually appears in consecutive frames present in a video with or without movement. This property is used in text enhancement and tracking. Caption text move in a uniform way: horizontally or vertically. While, scene text possess an arbitrary motion due to camera or object movement.

(f)Compression: The digital images are usually processed in a compressed format. If one can extract text without decompression then it is possible to develop a faster Text Information Extraction system.

II RELATED WORKS

In this section, the recent techniques focused on text detection and localization are reviewed and then the results are discussed.

2.1 Region -Based Technique

Region-based techniques use the properties of the text region such as color or gray-scale or their differences with the corresponding properties of the background. This method mainly uses a bottom-up approach .It groups the small components into larger components until all regions corresponds to text are identified in the image. A geometrical analysis is done to merge the text components using the spatial arrangement of the components so that non-text components are filtered out. The boundaries of the text regions are then marked.

This algorithm is divided into three phases:

i. *Text candidate spotting*: it attempt to separate text from background of the image.

ii. *Text characteristics verification*: text candidate regions are grouped together to discard those regions wrongly selected.

iii. *Consistency analysis for output:* regions representing text are modified to obtain a more useful character representation as input for an OCR

2.1.1Connected component based methods

CC-based methods are widely used due to their relatively simple implementation. Most of the CC-based methods consist of mainly four processing stages: (a) preprocessing, which includes color clustering and noise reduction, (b) CC generation, (c) filtering out non-text components, and (d) component grouping. A CC-based method could segment a character into multiple CCs, especially in the cases of polychrome text strings and low-resolution and noisy video images. Further, the performance of a Connected Component -based method gets severely affected by component grouping, such as a projection profile analysis or text line selection. In addition, several threshold values must be calculated to filter out the non-text components, and these threshold values are dependent on the image/video database.

J.Gllavata*et.al*[3] proposed a connected component based approach for the text extraction .It is based on color reduction technique and OCR is used for character recognition .It will only detect text with horizontal alignment. Low quality images will not be processed accurately. Zhong et al. [4] used a CC-based method, which uses color reduction. In that they quantize the color space using the peaks in a color histogram in the RGB color space. This is based on the assumption that the text regions cluster together in this color space and occupy a significant portion of an image. Each text component goes through a filtering stage using a number of heuristics, such as area, diameter, and spatial alignment. The performance of this system was evaluated using CD images and book cover images.Kim *et al.* [16] use transition map to detect overlay text.

Lienhart et al. [5] regard the text regions as CCs with the similar or same color and size, and motion analysis is applied to enhance the text extraction results for a video sequence. The input image given is segmented here based on the monochromatic nature of the text components by using split-and-merge algorithm. They primarily focused on caption text, such as credit titles, pre-title sequences, and closing sequences, which exhibit a higher contrast with the background. This made it easy to use the contrast difference between the boundaries of the detected components and its background in the filtering stage. Finally,they used a geometric analysis, including the width, height, and aspect ratioto filter out remaining non-text components. Based on experiments using 2247 frames, their algorithm extracted 86% to 100% of all the caption text.

2.1.2Edge Based Technique

Edges are a reliable feature of text rather than the color/intensity, layout, orientations, etc. The three characteristics of text embedded in images which can be used to distinguish the main features of the detecting text are edge strength, density and orientation. Edge-based algorithms are general-purpose method, which can quickly and effectively localize and extract the text from both document and indoor or outdoor images.

Xiaoqing Liu *et al*[6] method consists of mainly three stages: candidate text region detection, text region localization and character extraction. In its first stage, they used the magnitude of the second derivative of intensity as a measurement of edge strength, and this allowed a better detection of intensity peaks that normally characterizes text in the images. Edge detector is carried out by using a multiscale strategy, in which the multiscale images are mainly produced by Gaussian pyramids after successively applying low-pass filter and down-sample the original image reducing the image in both vertical and horizontal directions. In the second stage, characteristics of clustering can be used to localize text regions. In the third stage, the existing OCR engine where used.This method isnot sensitive to image color/intensity. It can handle both printed and document images effectively. It mainly analyses texts in the form of blocks. But the small image regions and stroke are mis-identified as text in areas containing large characters.Xin Zhang *et al*[7] used the features such as color and edge to extract the text from the video frame. Here, two methods are combined, and is called as color-edge combined algorithm, to remove text background. One of the method that is combined is based on the exponential changes of text color, called Transition Map model, the other one uses the text edges of different gray level image. This algorithm is robust to the image with multilingual text. To improve the efficiency of this method, the edge feature is added to remove background and then edge detection is performed on each color image using Canny operator and some Morphology operation. Finally the background of text is removed with the help of Transition Map model.

Fig. 1. Xiaoqing Liu's Algorithm

The main difference between the method by Sato et al. [8] and the other edge-based methods is the use of recognition-based character segmentation. In those they use character recognition results to make decisions on the segmentation of individual characters, and thus improves the accuracy of character segmentation. The processing time from detection to recognition was less than 0.8 seconds for a 352×242 image.

Anthimopoulos et al. [16] proposed a two-stage methodology for text detection mainly in video images. In its first stage, the text lines are detected based on the Canny edge map of the image. In the next stage, the result is refined using the sliding window and a SVM classifier is trained on features obtained by a new Local Binary Pattern-based operator (eLBP) which describes the local edge distribution. The whole algorithm is used in multiresolution fashion enabling detection of characters for a broad size range.

2.2Texture-based methods

Texture-based methods use the fact that texts in images have distinct textural properties which distinguish them from the background. The techniques mainly based on Gabor filters, Wavelet, FFT, spatial variance, etc. that can be used to detect the textural properties of a text region in an image.

A texture based method has been applied to vehicle license plate localization by Park *et al*. [9]. They used a time delay neural network as the texture discriminator in the HSI color space. Two time delay neural networks are used as horizontal and vertical filters and each receives the HSI color values for a small window of an input image and it decides whether or not the window contains a license plate number. After combining the two filtered images,the bounding boxes for license plates are located based on projection profile analysis.Mao et al. [10] proposed a texture-based text localization method using Wavelet transform.In this work, Harr Wavelet decomposition is used to define the local energy variations in the image at several scales. Binary image,that is acquired after thresholding the local energy variation, is then analyzed by connected component-based filtering using geometric attributes such as size and aspect ratio. All the text regions, which are detected at several scaleswere merged to give the final result.

2.3Text Extraction in Compressed Domain

Most of the digital images and videos are usually stored, processed, and transmitted in a compressed form, based on this idea that the text extraction methods that directly operate on images in MPEG or JPEG compressed formats have recently been presented. These methods only just require a small amount of decoding, and thereby it results in a faster algorithm. Moreover, the DCT coefficients and motion vectors in an MPEG video are also useful in text detection [11]. Bergen and Heeger [17] developed a parametric texture synthesis algorithm which can synthesize a matching texture, given a target texture.

Zhong*et al*. [12] worked on a method for localizing caption texts in JPEG images and I-frames of MPEG compressed videos. In this they used DCT coefficients to capture the textural properties,which includes the directionality and periodicity of local image blocks. The results obtained are then refined using morphological operations and connected component analysis. The authors says that, it is very fast (0.006 seconds to process a 240 × 350 image) and has a recall rate of 99.17% and false alarm rate of 1.87%. However as each unit blockis determined as text or non-text, precise localization results could not be generated.

2.4 Morphological Based Text Extraction

Morphology is a geometrical based approach for image analysis. It is used to extract important text features from the processed images. The feature still can be maintained, despite the change in the lighting condition or text color. Algorithm by Rama Mohan et al[13]The method considers that edge detection is more effective in text extraction. Basic operators of mathematical morphology are used to perform the edge detection. The algorithm is used to find out the connected component. By considering the gray levels of the components their variance is found out for each connected component, when components are found then labeling is done. After selecting the components whose variation is less than threshold value the text can be extracted . This method consists of four steps:

a) Edge extraction

b) Text candidate region formation

c) Labeling of text candidate regions

d) Elimination of non text region

Algorithm by Jui-Chen Wu et al.: Jui-Chen Wu [14] presented a text line extraction algorithm for extracting text regions from jumbled images. The method defines a set of morphological operations for extracting important contrast regions. The main steps of this algorithm are : a) Feature extraction: The relative contrast between texts and their background is an important feature for text line detection. A novel morphology-based scheme for extracting the high contrast feature for locating all possible text lines is used for feature extraction. b) Text candidate selection: a labeling technique is used to select all possible text lines from the analyzed image. c) Candidate verification: After candidate selection a verification process is carried out. The text verifications done on the basis of regularities of character size, the ratio between character width and height, and the period of characters.

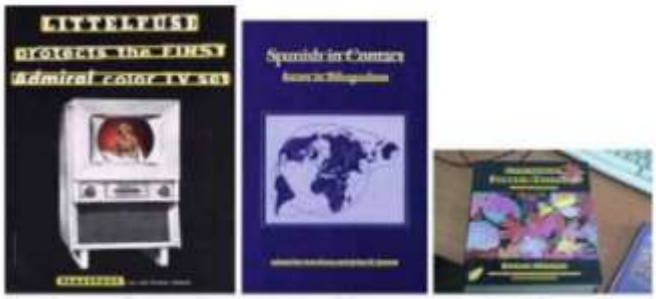

Fig.2 Jui-Chen Wu's Algorithm

2.5Performance Evaluation

There are several difficulties related to performance evaluation in almost all research areas based on computer vision and pattern recognition (CVPR). The empirical evaluation of CVPR algorithms is a means of measuring the ability of algorithms to meet a given set of requirements. Although various studies in CVPR have investigated about the issue of objective performance evaluation, there has been very little focus on the problem of Text information extraction in images and video. This section discuss the current evaluation methods used for Text information extraction and highlights several issues in these evaluation methods. There are different performance evaluations to estimate the algorithm for text extraction. Most ofthe approaches quoted here used Precision, Recall and F-Score metrics to evaluate theperformance of the algorithm. Precision, Recall and F-Score rates are usually computed based on thenumber of correctly detected characters (CDC) in an image,it is used in order to evaluate the efficiencyand robustness of the algorithm. The performance metrics are as follows:

2.5.1 False Positives

False Positives (FP) or False alarms are those regions in the image that are actually not characters of a text, but have detected by the algorithm as text.

2.5.2 False Negatives

False Negatives (FN) or Misses are those regions in the image which are actually text characters, but have not been detected by the algorithm.

2.5.3 Precision rate

Precision rate (P) can be defined as the ratio of correctly detected characters to the sum of correctly detected characters plus false positives.

2.5.4 Recall rate

Recall rate (R) can be defined as the ratio of the correctly detected characters to sum of correctly detected characters plus false negatives.

2.5.5 F-score

F-Score is definedas the harmonic mean of recall and precision rates.

We have provided a comprehensive survey of text information extraction in images and video. Even though a large number of algorithms have been proposed in the literature, no single method can provide satisfactory performance in all the applications due to the large variations in character font, size, texture, color, etc. There are several information sources for text information extraction in images (e.g., color, texture, motion, shape, geometry, etc). It is very advantageous to merge the various information sources to enhance the performance of a text information extraction system. But there is an issue that it is, not clear as to how to integrate the outputs of several approaches. There is a clear need for a public domain and representative test database for objective benchmarking. The lack of a public test set makes it difficult to compare the performances of competing algorithms, and creates difficulties when merging several approaches.

For caption text, significant progress has been made and several applications, such as an automatic video indexing system, have already been presented. However, their 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). Very little work has been done on scene text. Scene text can have different characteristics from caption text. For example, part of a scene text can be occluded or it can have complex movement, varies in size, font, color, orientation, style, alignment, lighting, and transformation.

Although many researchers have already investigated text localization, text detection and tracking of video images is required for utilization in many real applications (e.g., mobile handheld devices with a camera and real-time indexing systems). A text-image-structure-analysis, which is analogous to a document structure analysis, is needed to enable a text information extraction system to be used for any type of image, including both scanned document images and real scene images through a video camera. Despite the many difficulties in using TIE systems in real world applications, the importance and usefulness of this field continues to attract much attention for researchers.

III CONCLUSION

In this paper various text extraction techniques are discussed and analyzed. Every approach has its own benefits and restrictions. Though we have large number of algorithms and methods for text extraction from image but none of them provide an adequate output because of deviation in text. Some of the applications of the text extraction system includes Content-based video coding or document coding ,License/container plate recognition ,Text-based image

indexing ,Video content analysis document analysis, Industrial automation ,Wearable or portable computers mobile robot navigation to detect text based landmarks,object identification etc. A text-image-analysis, is actuallyneeded to enable a text information extractionsystem that is to be used for any type of image, includingboth scanned document images and real sceneimages through a video camera. Despite of the manydifficulties caused in using TIE systems in real worldapplications, the importance and usefulness of thisfield continues to attract much attention

REFERENCES

[1] Michel, F.How many photos are uploaded to Flicker every day, month, year?, http:// www.flickr.com/photos/franckmichel/6855169886/, Dec 5, 2013.

[2] Official Blog: It‘s YouTube‘s 7th Birthday, http://youtube-global.blogspot.com/2012/05/its-youtubes- 7th-birthday-and-youve.html, Oct 2013.

[3] J. Gllavata, R. Ewerth, and B. Freisleben, “A robust algorithm for text detection in images,” *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, pp.611– 616, ISPA, 2003.*

[4] Yu Zhong, KalleKaru, and Anil K. Jain, Locating Text In Complex Color Images, Pattern Recognition, 28 (10) (1995) 1523-1535

[5] R. Lienhart and W. Effelsberg, Automatic Text Segmentation and Text Recognition for Video Indexing, Technical Report TR-98-009, PraktischeInformatik IV, University of Mannheim, 1998.

[6] Xiaoqing Liu, JagathSamarabandu, “ Multiscale Edge-Based Text Extraction from Complex Images,” *International Conference on Multimedia and Expo, pp.1721-1724, 2006*

[7] XinZhang,Fuchun Sun, Lei Gu, “*A Combined Algorithm for Video Text Extraction”,2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*,2010.

[8] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, Video OCR for Digital News Archive, Proc. of IEEE Workshop on Content based Access of Image and Video Databases, 1998, pp. 52-60.

[9] S. H. Park, K. I. Kim, K. Jung, and H. J. Kim, Locating Car License Plates using Neural Networks, IEE Electronics Letters, 35 (17) (1999) 1475-1477

[10] W. Mao, F. Chung, K. Lanm, and W. Siu, Hybrid Chinese/English Text Detection in Images and Video Frames, Proc. of International Conference on Pattern Recognition, 2002, Vol. 3, pp. 1015-1018.

[11] Karin Sobottka, Horst Bunke and HeinoKronenberg,”Identification of Text on Colored Book and Journal Covers”, *Document Analysis and Recognition*, 20-22 ,1999.

[12] Yu Zhong, Hongjiang Zhang, and Anil K. Jain, Automatic Caption Localization in Compressed Video, *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, (4) (2000) 385-392*

[13] Rama Mohan Babu, G., Srimaiyee, P. Srikrishna, A., “Text Extraction From Hetrogenous Images Using Mathematical Morphology” *Journal Of Theoretical And Applied Information Technology © 2005 -2010.*

[14] Jui-Chen Wu · Jun-Wei Hsieh · Yung-Sheng Chen,”Morphology-based text line extraction” Machine Vision and Applications 19:195–207 DOI 10.1007/s00138-007-0092-0,2008

[15] M. Anthimopoulos, B. Gatos, I. Pratikakis, “A two-stage scheme for text detection in videoimages,” *Image and Vision Computing, Elsevier, Vol. 28 , No. 9 , pp.1413–1426, 2010.*

[16] WonjunKim,Changick Kim, "A New Approach for Overlay Text Detection and Extraction From Complex Video Scene," *IEEE Transactions on Image Processing, V.18 , No.2, pp. 401 – 411, 2009.*

[17] D.J. Heeger And J.R. Bergen. "Pyramid-Based Texture Analysis/Synthesis" *,In proceedings Of ACM Conf. Comp. Graphics (Siggraph), Volume 29, Pages 229{233, Los Angeles, Ca, 1995.*

[18] Shivakumara,P. , Sreedhar,R.P. , TrungQuyPhan , Shijian Lu , Chew Lim Tan , "Multioriented Video Scene Text Detection Through Bayesian Classification and Boundary Growing*," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 22 , No.8 pp.1227 – 1235, 2012.*