Analysis of Text Information Extraction and Card Identification in Video Clips

Khin Thandar Tint, and Kyi Soe

***Abstract*—**Text that appears in images and videos contains important and useful information. Text extraction involves detection, localization, tracking, extraction, enhancement and recognition of the text from the given image. However variation of text due to difference in size, style, orientation, alignment, low image contrast and complex background make the problem of automatic text extraction extremely challenging. Detection and extraction of text in images and videos have been used in many applications. In this paper, edge-based and connected components-based methods are used for text extraction and color averaging algorithm is applied to identify card. It is robust with respect to the font size, style, color, orientation, and alignment of text and can be used in a large variety of application fields. The results from different video clips can be proved that the proposed method can generate very reliable and high accuracy performance.

***Keywords*—**Text extraction, Text detection, Performance Analysis Employee card identification, Connected components

I. Introduction

THE automatic detection and extraction of text in images and videos have been used in many applications. Document text localization can be used in the applications of page segmentation, document retrieving, address block location, etc. Content-based image/video indexing is one of the typical applications of overlay text localization. Scene text extraction can be used in mobile robot navigation to detect text-based landmarks, vehicle license detection/recognition, object identification, etc.

Video text can broadly be classified into two categories: overlay text and scene text. Overlay text refers to those characters generated by graphic titling machines and superimposed on video frames/images, such as video captions [1].

As more and faster digital devices with the larger storage spaces were becoming more popular in days by day, instead of representing information using texts and still images, audio, etc, the information were recorded and stored in video. Hence, the video has now become the most popular media type in daily life. Content-based image indexing refers to the process of attaching labels to images based on their content. Image content can be divided into two main categories: perceptual content and semantic content [2]. Perceptual content includes the attributes such as color, intensity, shape, texture, and their temporal changes, whereas semantic content means objects, events and their relations. A number of studies on the use of relatively low-level perceptual content for image and video indexing have already been reported. Studies on semantic image content in the form of text, face, vehicle, and human action have also attracted some recent interest. Among them, text within an image is of particular interest as (i) it is very useful for describing the contents of an image; (ii) it can be easily extracted compared to other semantic contents, and(iii) it enables applications such as keyword-based image search, automatic video logging, and text-based image indexing.

The proposed paper is also one of the text retrieval works. In this paper, text is detected and retrieved from employee card of CCTV recorded video clip. In the office, all employees and staffs have to work putting the employee card. Here, it needs to make real time monitoring and distinguish between the employees and guests. Hence, the cards are used with different colors, blue for employee and red for guest. Related person's name, occupation, department and etc are printed on all of the cards. The texts are retrieved and the cards are differentiated between employee and guest by their card colors.

Therefore, the proposed system focuses on retrieving texts, identifies card and analyzes performances. The division of this paper is as follows, in Section 2, some related work is given which describes the previous research about text detection and extraction. In Section 3, the proposed method is given. Experimental results are shown and discussed in Section 4. Finally, in Section 5, the conclusion is given.

II. Related Work

A lot of previous works are used for text detection and extraction from videos and images. S. Antani *et al* [3] presented a robust text extraction from video. In this work, they presented update text detection and extraction of unconstrained variety of text from general purpose video. The text detection results from a variety of methods were fused and each single text instance was segmented to enable it for OCR. As video had low resolution and the text often had

Khin Thandar Tint, Student, Faculty of Information and Communication Technology, University of Technology (Yatanarpon Cyber City, Pyin Oo Lwin, Mandalay Division, Myanmar; email : kttoct5@gmail.com

Kyi Soe, Lecturer, Faculty of information and Communication Technology, University of Technology , Yatanarpon Cyber City, Pyin Oo Lwin, Mandalay Division, Myanmar

poor contrast with a changing background, a variety of methods were applied and took the advantage of the temporal redundancy in video to achieve better result in good text segmentation.

Wang et al. [4] described a connected-component based method which combines color clustering, a black adjacency graph (BAG), an aligning-and-merging-analysis scheme and a set of heuristic rules together to detect text in the application of sign recognition such as street indicators and billboards. Uneven reflections result in incomplete character segmentation which increases the false alarm rate in this method.

Kim et al. [5] implemented a hierarchical feature combination method to implement text extraction in natural scenes. However, this method could not handle large text very well due to the use of local features that represents only local variations of image blocks.

Wolf et al. [6] presented an algorithm to localize artificial text in images and videos using a measure of accumulated gradients and morphological post processing to detect the text. The quality of the localized text is improved by robust multiple frame integration. A new technique for the binarization of the text boxes is proposed. Finally, detection and OCR results for a commercial OCR are presented.

Xi et al. [7] implemented a new system for text information extraction from news videos. This method integrates text detecting and text tracking to develop for locating text areas in the key frames or images, together with a scheme to evaluate the performance of the approach. This method enhanced the quality of the detected text blocks by multi-frame averaging, Adaptive thresholding method is applied to binarize the text blocks and recognize the text using an off-the-shelf OCR module.

Gllavata et al. [8] described an efficient algorithm which can automatically detect, localize and extract horizontally aligned text in images (and digital videos) with complex backgrounds. This approach is based on the application of a color reduction technique, a method for edge detection, and the localization of text regions using projection profile analyses and geometrical properties. The output of the algorithm is text box with a simplified background, ready to be fed into an OCR engine for subsequent character recognition. Our proposal is robust with respect to different font sizes, font colors, languages and background complexities. The performance of the approach is demonstrated from different types of video sequences.

Gao et al. [9] developed a three layer hierarchical adaptive text detection algorithm for natural scenes. This method has been applied in a prototype Chinese sign translation system which mostly has a horizontal and/or vertical alignment.

However, in real scenes, due to uneven illumination, reflections and shadows, an image background may contain areas with high spatial intensity variation that do not contain text. Therefore, the proposed system can solve this problem and can successfully detect and extract the desired text from card.

III. Proposed System

The proposed method is based on the fact that edges are a reliable feature of text regardless of color/intensity, layout, orientations, etc. Edge strength, density and the orientation variance are three distinguishing characteristics of text embedded in images, which can be used as main features for detecting text. The proposed method consists of two stages: text detection and extraction and card identification by color.

A. Text Detection and Extraction

There are two methods to detect and extract text information. They are edge-based text region extraction and connected component-based text region extraction.

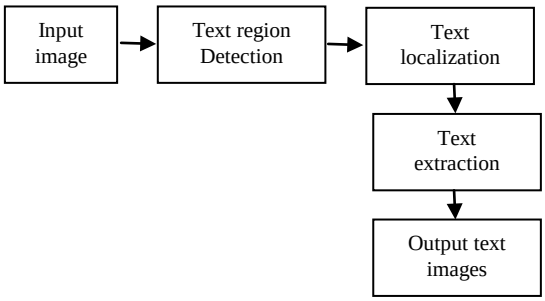

Fig.1 Block diagram of edge based text region extraction

In Fig.1, Gaussian pyramid is constructed by successively filtering the input image with a Gaussian kernel of size 3x3 and down sampling the image in each direction by half. The sample Gaussian pyramid image is next convolved with directional filters at different orientation kernels for edge detection.

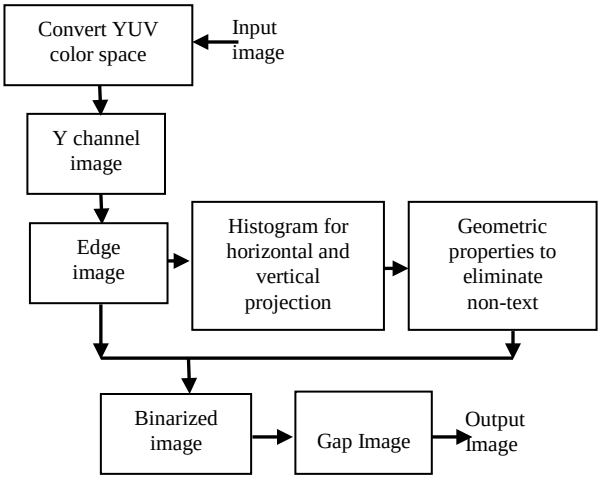

Fig.2 Block diagram of connected component based text region extraction

After convolving, a feature map is created by associated weighted factors. The feature map is the result edge based image. In the text localization, text region is retrieved form background by dilation and erosion. Then, text region is defined and extracted by labeling region and regional props.

In Fig.2, input image is at first convert to YUV color space and then Y channel is used as a gray image to next processes.

It is because the Y channel refers to brightness or intensity of the image whereas the U and the V channels refer to the actual color information. To be more clear edge of text region, every pixel in the image is assigned a weight with respect to its neighbors in each direction.

The resultant image is defined as edge image. From the text image, vertical and horizontal lines are eliminated, and then horizontal and vertical projections are calculated. Candidate text regions are segmented based on adaptive threshold values comparing with the horizontal and vertical projections. Only regions that fall within the threshold limits are considered as candidates for text. After eliminating lines and estimating vertical and horizontal projections, the image is binarized according to the threshold limitation. Then, text region are continuously filled by morphology processes (dilation and erosion) and then the text region on the input image can be cropped as output image result by using region labeling and regional properties.

In text extraction, the input colour images are received from the output key frames of the images. In the key frames, employee cards were detected and cropped.

B. Card Identification by Color

After extracting and cropping employee card, the card is needed to identify guest card or employee.

As Red card is used for guest and Blue card is for employee in our proposed system, it needs to define which color of cropped employee card image is. To define the color of the cropped card image, the following color defining algorithm is used.

TABLE I
COLOR IDENTIFICATION ALGORITHM

```
Begin
Read     cropped RGB image
Estimate   mean value of Red Channel (meanR)
Estimate   mean value of Blue Channel (meanB)
Estimate   mean value of Green Channel (meanG)
Find     MaxMeanValue among meanR, meanB and meanG.
If (MaxMeanValue == meanR)
        display (" The Card is GUEST")
elseif(MaxMeanValue == meanB)
        display (" The Card is EMPLOYEE")
else
        display (" The Card is UNDEFINE")
End
```

By the above algorithm, the cropped card image can be defined Red card or Blue and distinguished between guest and our employee. The performance and accuracy measured result of the proposed method can be seen in next section.

IV. RESULTS AND DISCUSSION

The results of the proposed system are described in this section. Over 100 video clips are now being tested for the system. To be obviously described, two implemented results of the system, which are obtained from two different video files, are discussed and demonstrated to identify the card whether it is employee card or guest card. The video clip, "use1.avi" has the total frames of 3808 and each frame size is resolution of 288x352. More than 300 key frames can be generated for this video clip depending on threshold value. Similarly, video clip, "use2.mp4" has total frames of 184 and each is resolution of 720x1280. There are 37 key frames for this video clip. "use3.mp4" has 170 total frames, 720x1280 resolutions, and received 35 key frames.

The cropped images from the input video frames are described in Fig. 3(a) and (d). The output key frames from text detection and extraction are shown in Fig. 3(b) and (e). Finally, from the card identification, the proposed system can identify the card whether it is employee card or guest card as shown respectively in Fig. 3(c) and (f).

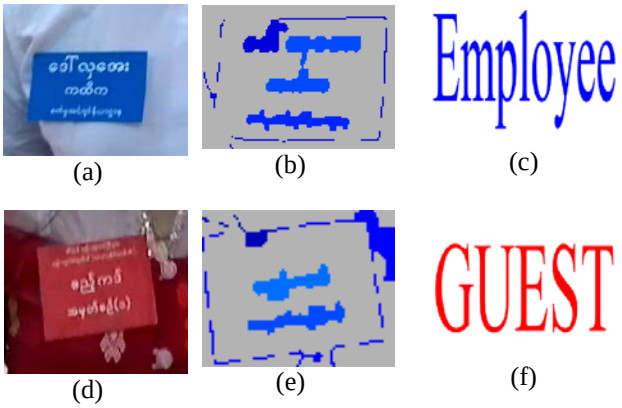

Fig.3. Experimental Results of the Proposed System

Table II shows the performance comparison of the proposed method with several existing methods, where the proposed method shows a clear improvement over existing methods. In this table, the performance statistics of other methods are cited from published work. The performance of the previous work is similar with each other but less than the performance of the proposed system.

The overall average computation time for 100 test images (with 720 × 1280 resolution) using unoptimized Matlab codes on a personal laptop with Intel Pentium(R) 1.8GHZ processor and 1.0G RAM is 35.15 seconds (standard deviation =0.156), which includes entire execution time including video file reading, key frame generation, card extraction, computation as well as image display.

TABLE I
PERFORMANCE COMPARISONS

Method	Precision Rate (%)	Recall Rate (%)
Proposed Method	93.5	97.6
Wolf et al.[6]	-	93.5
Xi et al.[7]	88.5	94.7
Gllavata et al. [8]	83.9	88.7

V. Conclusion

This system can successfully detect and extract text. Moreover, this system can identify the card whether it is employee or guest. Video processing and analysis is one of the interesting technologies. This system aims to be used in controlling the security among the employees and guests in the government offices. According to experimental results, the proposed method can generate the accuracy via than the previous works. As the future works, the system will be continued to be more precise and accurate in card extraction and retrieval of text.

References

[1] Xiaoqing Liu and Jagath Samarabandu, "MULTISCALE EDGE-BASED TEXT EXTRACTION FROM COMPLEX IMAGES", University of Western Ontario, Department of Electrical & Computer Engineering, London, Ontario, N6A 5B9, Canada.

[2] Keechul Junga;∗, Kwang In Kimb, Anil K. Jainc pattern recognition," Text information extraction in images and video: a survey", School of Media, College of Information, Soongsil Unversity, SangDo-Dong, DongJak-Gu, 1, Seoul 156-743, South Korea, Pattern Recognition 37 (2004) 977 – 997.

[3] S.Antani, D.Crandall and R.Kasturi "Robust Extraction of Text in Video", *IEEE Transactions of Image and Video processing*, vol 3,pp78-92, April 2000.

[4] KongqiaoWang and Jari A. Kangas, "Character location in scene images from digital camera," Pattern Recognition, vol. 36, no. 10, pp. 2287 2299, 2003.

[5] K. C. Kim, H. R. Byun, Y. J. Song, Y. M. Choi, S. Y. Chi, K. K. Kim, and Y. K. Chung, "Scene text extraction in natural scene images using hierarchical feature combining and verification," in Pattern Recognition, 2004, Aug. 2004, vol. 2 of ICPR 2004. Proceedings of the 17th International Conference on, pp. 679–682.

[6] C. Wolf, J. M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents," in Pattern Recognition, 2002, Aug. 2002, vol. 2 of Proceedings. 16th International Conference on, pp.1037–1040.

[7] Jie Xi, Xian Sheng Hua, Xiang Rong Chen, LiuWenyin, and Hong Jiang Zhang, "A video text detection and recognition system," in Multimedia and Expo, 2001. ICME 2001, 2001, IEEE International Conference on, pp. 873–876.

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

[9] Jiang Gao and Jie Yang, "an adaptive algorithm fot text detection from natural scenes," in Computer Vision and Pattern Recognition, 2001. CVPR 2001, 2001, Proceedings of the 2001 IEEE Computer Society Conference on, pp. II–84–II–89.

Khin Thandar Tint was born on 5th October, 1984 in Mandalay Township, Myanmar. The author received the BE degree in information technology from Technological University, Mandalay Division (Myanmar), in 2007, and the ME degree in information technology from Technological University, Mandalay Division, Myanmar, in 2010. She is an Assistant Lecturer at Government Technological University (Myingyan) Mandalay Division, Myanmar since 2010. Now she is studying for PH.D degree of Information and Communication Technology Engineering at University of Technology, Yatanarpon Cyber City, Pyin Oo Lwin, Mandalay Division, Myanmar.

The second author Kyi Soe is supervisor of first author. He is a Lecture from Faculty of Information and Communication Technology, University of Technology, Yatanarpon Cyber City, Pyin Oo Lwin, Mandalay Division, Myanmar.