Myanmar Text Areas Identification from Video Scenes

Thuzar Tint, and Dr.Nyein Aye

***Abstract*—**Text embedded in image and video provides short-lived and important information to describe video content. Text detection and extraction from image and video plays a vital role in digital image processing to support optical character recognition. This paper presents the design and implementation of a system that detects and localizes Myanmar text areas in video scenes. First of all, Gaussian filter is applied the output of video analysis to remove noise. As there exits contrast between text and background in video, second order edge detection operator (Laplacian operator) is employed to find areas of rapid change (edges) in images. And second order edge maps are analyzed to reduce processing time. Then, candidate regions are identified by using connected component labeling approach and geometric properties such as aspect ratio. To eliminate false positives and to verify Myanmar text area, a correlation model is constructed in frequency domain. Experimental results show that the proposed system is efficiently able to localize Myanmar text areas from video clips and the system can applied in real time.

***Keywords*—**Myanmar Text Area, Gaussian filter, Second order edge detection operator, 2D Fast Fourier Transformation (fft2).

I. Introduction

NOWADAYS, video is the most popular media type delivered via TV broadcasting, Internet and wireless network. Video is one of the sources for presenting the valuable information since it contains sequence of video images, audio and text information. The text embedded in video provides clearer and more obvious information about the content of specific media. It is a powerful source of information about the content of an image or video. It is necessary automatically to detect and extract the text information from these video frames before feeding to OCR engine. Nowadays commercial OCR engines cannot detect and recognize text embedded in complex background directly, e.g. a video image. Consequently, preprocessing, i.e., detecting accurate bounding box of text and simplifying the background, is very important [1].

Text detection is a preliminary step in automatic text recognition. It needs to be fast, efficient and robust in order to feed an OCR classifier with the correct input. Therefore, text detection step plays a vital role in text recognition process. A large number of techniques have been proposed in the process of retrieving text from images and video clips. Edge-based text extraction algorithm is a general-purpose method, which can quickly and effectively localize and extract the text from both document and indoor/ outdoor images. 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 [2]. Edge-based method performs fast text detection but also results in high false detections. Another popular method is region based approaches. Region based methods detect characters as monochrome regions satisfying certain heuristic constraints. The pixels of each character are assumed to have similar color and can be segmented from background by image segmentation or color clustering [3] preprocess. The resulting monochrome regions are selected as characters under some simple heuristic constraints, such as the size, the height/ width ratio of the region or baselines. Region-based methods not only identify the embedded text regions but also segment characters from background. However, the monochrome constraints are not always be satisfied and, therefore, the methods are not robust to complex background and compressed video [4]. Texture-based methods use the observation that texts in images have distinct textural properties that distinguish them from the background. The techniques based on Gabor filters, Wavelet, FFT, spatial variance, etc. can be used to detect the textural properties of a text region in an image [5]. Although Texture-based method is able to detect text in complex background, it is very time consuming and cannot always perform accurate localization.

In [6] and [7], vertical edges are first detected and connected into text clusters by using a smoothing filter. As with region and texture-based methods, the text clusters are then selected by using heuristic constraints. Jain and Yu [8] first employ color reduction by bit dropping and color clustering quantization, and afterwards a multi-value image decomposition algorithm is applied to decompose the input image into multiple foreground and background images. Then, CC analysis is performed on each of them to localize text candidates.

Xin Zhang [9] used the color and edge features to extract the text from the video frame. In this work, two methods are combined, called color-edge combined algorithm, to remove text background. One of the combined methods 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. After removing complex background, text location is determined using the vertical and horizontal projection

Thu Zar Tint is with the University of Technology(Yantanarpon Cyber City), Pyin Oo Lwin, Myanmar (corresponding author's phone: +95 9420703151 ; e-mail: thuzartint1984@gmail.com).

Dr. Nyein Aye is now with the Department of Hardware, University of Computer Studies, Mandalay, Myanmar (e-mail: nyeinaye@gmail.com).

http://dx.doi.org/10.15242/IIE.E0314187

method. 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. Sin et al. [10] use frequency features such as the number of edge pixels in horizontal and vertical directions and Fourier spectrum to detect text regions in real scene images. Based on the assumption that many text regions are on a rectangular background, rectangle search is then performed by detecting edges, followed by the Hough transform. However, it is not clear how these three stages are merged to generate the final result.

The rest of the paper is organized as follows. The proposed method of Myanmar text areas detection step is explained in Section II. And Section III shows how Myanmar text areas localize. The experimental results on various videos which contain Myanmar texts are shown in Section IV, followed by conclusion in Section V.

II. Proposed Method

The proposed method is based on the observation of contrast between text and its adjacent background. Therefore, Gaussian filter is firstly applied frames of input video to reduce noise from each frame and get smoothing images (frames). Then, Laplacian zero crossings are computed at each frame to produces a set of zero crossings edge maps. After generating a set of zero crossings edge maps from each frame, frame analysis is performed by using XOR operation to remove redundancy frames. And we carry out region filling processes and find candidate regions from each edge map by using connected component labeling method and some heuristic rules. Since the next step is to verify whether Myanmar text or non-text in each candidate region, similarity measure of each candidate region is calculated cross-correlation as a multiplication in the frequency domain. After detecting Myanmar text areas from each frame, localization process is applied to show the Myanmar text regions with bounded rectangles. The overall procedure of Myanmar text area identification steps is shown in Fig. 1.

A. Zero Crossings Edge Map Generation

An edge can generally be defined as a boundary or contour that separates adjacent image regions having relatively distinct characteristics according to some feature of interest. In many images and videos, bright pixels correspond to text pixels while dark pixels correspond to non-text pixels. Since texts in images and videos have the greater intensity discontinuity between the text and its adjacent background, a set of edge maps is generated based on this property. First of all, each frame of the input video is filtered with Gaussian operator because the Gaussian smoothing operation serves to band-limit the image to a small range of frequencies, reducing the noise sensitivity problem when detecting zero crossings.

When the input frame is this image is convolved by a Gaussian kernel

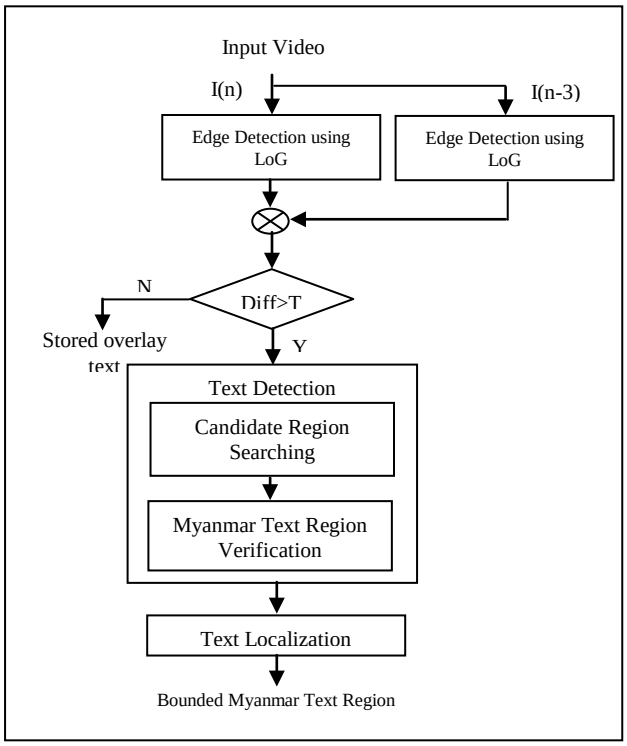


Fig. 1 Overall procedure of the proposed system

$$g(x,y,t) = \frac{1}{2\pi t^2} e^{-\frac{x^2+y^2}{2t^2}} \quad (1)$$

at a certain scale t to give a scale space representation

$$L(x,y;t) = g(x,y,t) * f(x,y) \quad (2)$$

And then, Laplacian zero crossings are computed at each as as it can highlight intensity discontinuities in an image and deemphasizes regions with slowly varying intensity levels. The mathematical equations of Laplacian *L(x,y)* of a frame with pixel intensity values *I(x,y)* is given by:

$$L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \quad (3)$$

To include a smoothing Gaussian filter, combine the Laplacian and Gaussian functions to obtain a single equation:

$$LoG(x,y) = -\frac{1}{\pi\sigma^4}[1 - \frac{x^2+y^2}{2\sigma^2}]e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (4)$$

In the proposed system, 5 x 5 kernel map is used and σ is set to 0.6. The LoG response will be positive just to one side of the edge , negative just to the other side of the edge, and zero at the uniform intensity. In this way, LoG second derivative edge detector can produces a set of edge maps from not only high contrast frames not only low contrast frames. The end result of this step plays a vital role in the proposed method as the subsequent steps will not be able to recover those lines if it misses low contrast text lines. The edge map of a sample frame is shown in Fig 2(b).

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

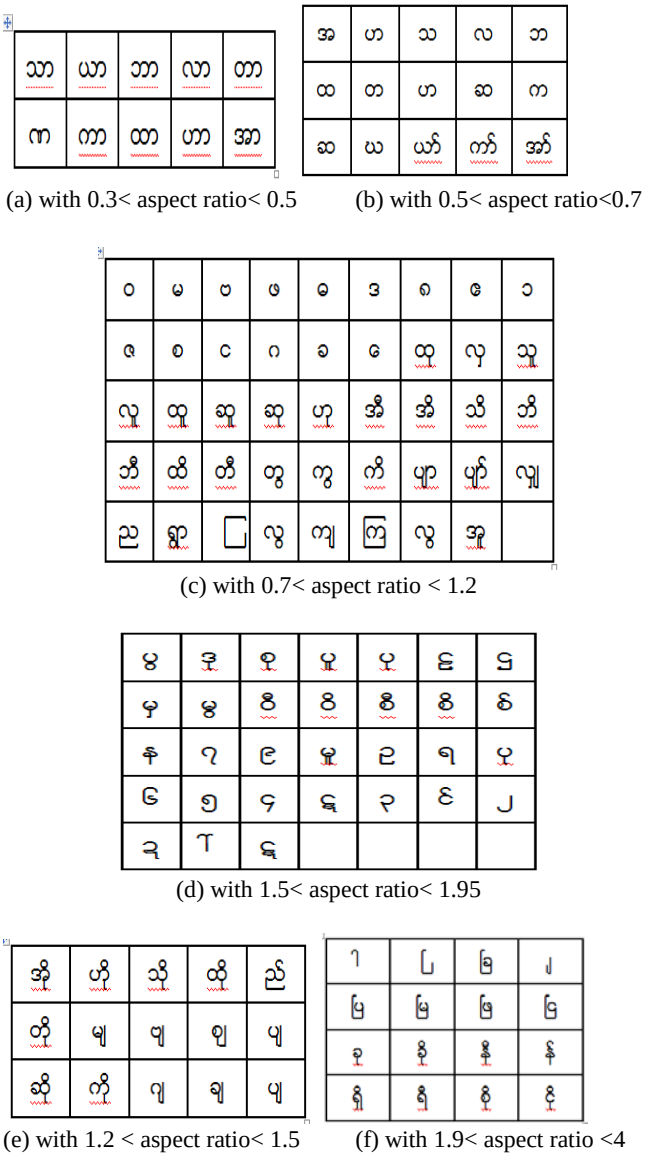
Fig.2 (a) Input Frame (b) Zero crossing edge map for the input frame (c) Link map for the input frame, and (d) Candidate region for the input frame

B. Edge Maps Analysis

In order to reduce processing time, it is necessary to analyze the edge map of the input frame whether it is similar edge map of frames or not. Therefore, we find dissimilarity measure between the current edge map and the previous third edge map by using XOR operation. If the different is higher than threshold value that is seek in advanced, the current frame is assumed to be a new edge map and then performs the remaining steps. Otherwise, the current frame is assumed to be the same with the previous third frame.

C. Candidate Region Extraction

The second step is to find the candidate text regions in frames, which is based on zero crossings edge map from the previous step. There involve many processes in the second step. Firstly, images containing zero crossing edge maps are converted into binary image. Then, filling process is done within gaps of consecutive transition pixels. In this process, they are filled with 1s if a gap of consecutive pixels between two non-zero points in the same row is shorter than 10% of the frame width. And connected components are labelled and if they are smaller than the threshold value, they are removed. The threshold value is empirically selected by observing the minimum size of text region. Since it is reasonable to assume that the text regions are generally in rectangular shapes, a rectangular bounding box is generated by linking four points, which correspond to(min_x,min_y), (max_x,min_y), (min_x,max_y), (max_x,max_y) taken from the link map shown in Fig. 2(c). To remove small parts from link maps, height and width of every link map are counted for getting aspect ratio of every link map. If the aspect ratio of a component is smaller than the threshold, that link map is considered as a false positive and removed; if not, it is accepted as a candidate region. Threshold is selected based on the Myanmar character rule. In Myanmar word, at least three characters are combined to get a meaningful word. Therefore, the width of a meaningful word is larger than its height. The refined candidate regions are shown in Fig 2(d).



(a) with 0.3< aspect ratio< 0.5 (b) with 0.5< aspect ratio<0.7

(c) with 0.7< aspect ratio < 1.2

(d) with 1.5< aspect ratio< 1.95

(e) with 1.2 < aspect ratio< 1.5 (f) with 1.9< aspect ratio <4

Fig. 3 (a) (b) (c) (d) (e) (f) groups of trained templates by their shapes and aspect ratios

D. Myanmar Text Region Verification

The previous state has a high detection rate but there exists relatively low precision. This step is to verify that the smoothed boundary candidate regions are Myanmar text region or not because these smoothed boundary candidate regions can contain not only real Myanmar text but also other non-text. Since it is necessary to calculate the similarity probability of the smoothed boundary candidate regions, we use a template matching approach. This approach involves :

(1) Myanmar Character Templates Creation

(2) Finding Myanmar Characters in each candidate region.

1) Myanmar Character Templates Creation: For any template based approach, it is very much necessary to obtain a template which is a good representative of the data. As Myanmar characters have significant features and they are similar in shape with one another, these templates are learned into six groups by their shapes and aspect ratios which are

shown in Fig. 3. In the pre-processing state, each grey scale Myanmar character template is converted into binary image by using otsu's threshold method. And then six average templates are found from similar shape Myanmar character templates. The average template can be formally defined as

$$T(i,j) = \frac{1}{N}\sum_{K}^{N} B(i,j) \quad (5)$$

where N is the number of similar shape Myanmar character template used for average template creation and $B(i, j)$ and $T(i, j)$ represent the pixel values of the (i, j)th pixel of B and T respectively.

2) Finding Myanmar Character in Each Candidate Region: To search Myanmar Character in each candidate region, we can compute the cross-correlation as a multiplication in the frequency domain. In the frequency domain, convolution corresponds to multiplication can be described by

$$I * T' = \mathfrak{F}^{-1}(\mathfrak{F}(I)\mathfrak{F}(T')) \quad (6)$$

Where, $\mathfrak{F}$ denotes Fourier transformation and $\mathfrak{F}^{-1}$ denotes the inverse FFT. The implementation takes each candidate region and each average template. The template is zero-padded because the template must have the same size as each candidate region and then transforms are evaluated. The required convolution is obtained by multiplying the transforms and then applying the inverse. The resulting candidate region is the magnitude of the inverse transform. Each of the templates of differing size is then matched by frequency domain multiplication. The maximum frequency domain value, for all sizes of template, indicates the position of the template and gives a value for its size [11].

The similarity values for corresponding candidate regions are obtained from above step. If similarity value in the candidate region is larger than a predefined threshold value, the component is assumed Myanmar character and match count is increased by 1. After all components of a candidate region have been matched against templates, the matching probability is computed and compared with predefined threshold. Then, if the matching probability is greater than the threshold value, the corresponding candidate region is finally determined as Myanmar text area. The threshold value is empirically set to 0.6 based on various experimental results.

III. Myanmar Text Area Localization

After Myanmar text region verification step, it is necessary to refine the detected Myanmar text object region for better accurate text extraction because these detected Myanmar text object region can include single text line or multiple text lines. In this localization step, all the edge pixels of each row of the detected text areas are firstly counted in order to form a histogram of the number of zero crossing edge pixels. Then, the null points, which denote the pixel row without transition pixels, are removed and separated regions are relabeled. The projection is conducted vertically and null points are removed once again. At the end of this step, we can bound corresponding a single text line from each detected Myanmar text regions with yellow color lines.

IV. Experimental Result

In this section, the proposed system is implemented and evaluated to show the efficiency and robustness of the proposed method. Since there is no standard benchmarking dataset available, we have tested a various kind of video such as news programming videos, movie clips, sport videos which involve not only graphical but also scene text written with Myanmar language.

A. Performance measure for text region detection

For performance measure for text region detection, the following categories are used:

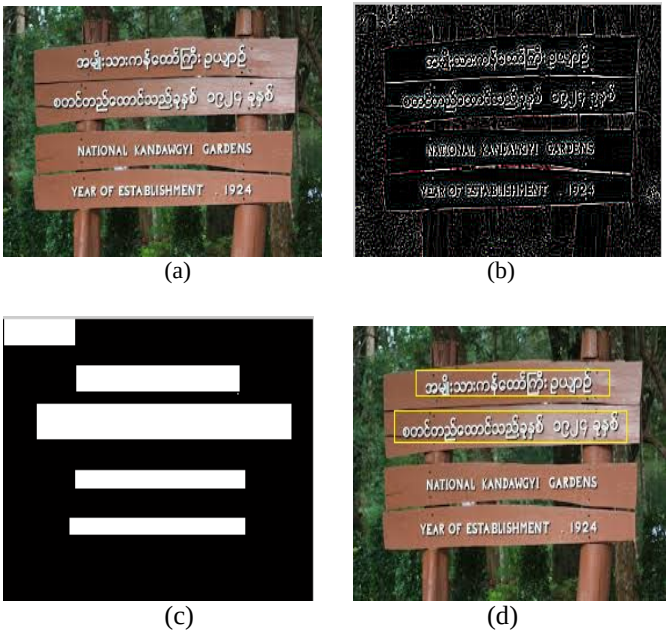
- Recall(R)=TDB / ATB
- Precision(P)=TDB / (TDB+FDB)
- F-measure(F)=2 _ P _ R / (P + R)
- Misdetection Rate(MDR)=MDB / TDB

Truly Detected Block (TDB): A detected block that contains at least one true overlay Myanmar character. Thus, TDB may or may not fully enclose a text line.

Falsely Detected Block (FDB): A detected block that does not contain any text.

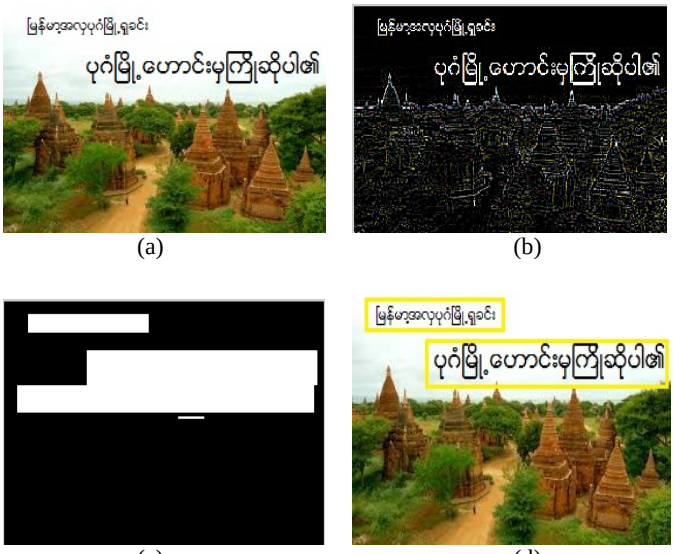
Text Block with Missing Data (MDB): A detected block that misses more than 20% of the characters of a text line. For each image frame in the experiment, we also manually count the number of *Actual Text Blocks (ATB)*, i.e, the true overlay text blocks.

We use MDR for the partial detection of Myanmar text regions to ensure a fair performance analysis. Detection Rate



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

Fig. 4 Myanmar Text Area Identification process (a) Sample Video frame (b) Zero crossing edge map for the sample frame (c) Candidate region for the sample frame (d) bounded Myanmar text regions for the sample frame

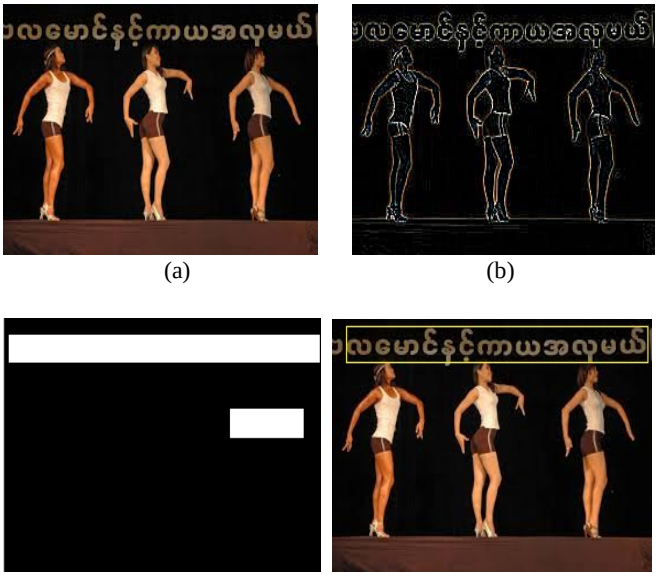


(a) (b) (c) (d)

Fig. 5 Myanmar Text Area Identification process (a) Sample Video frame (b) Zero crossing edge map for the sample frame (c) Candidate region for the sample frame (d) bounded Myanmar text regions for the sample frame

and False Positive Rate can also be converted to Recall and Precision: Recall= Detection Rate and Precision = 1-False Positive Rate. So only the above four performance measures are used for evaluation.

In this experiment, our approach handles well the Myanmar texts that are independent of contents types, such as movie, drama, animation, and so on. Although we have tested our proposed system with a large number of videos which contain Myanmar texts, we will describe three short sample videos in this paper. One is a documentary video clip for national Kan Daw Gyi garden located in Pyin Oo Lwin. This video contains more



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

Fig. 6 Myanmar Text Area Identification process (a) Sample Video frame (b) Zero crossing edge map for the sample frame (c) Candidate region for the sample frame (d) bounded Myanmar text regions for the sample frame

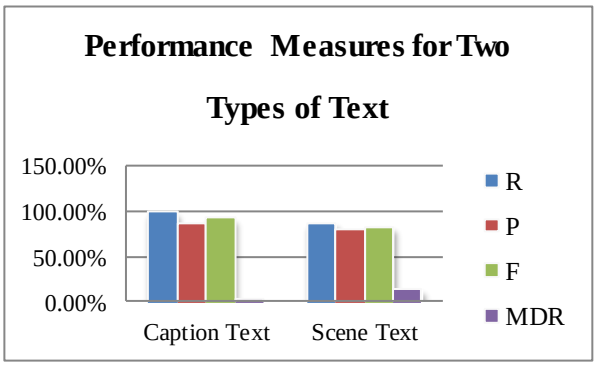


Fig. 7 Performance Measure of the Proposed System

than 500 scene texts which are not only Myanmar language but also English language. This sample video frame is shown in Fig (4). Our approach intends only to support Myanmar OCR. Therefore, candidate regions extraction step include both Myanmar and English scene texts, but only Myanmar scene text can correctly be localized after verification step. The next sample video, an advertisement short video about Myanmar for tourist attraction, has Myanmar caption texts with different font size shows in Fig (5). This sample video frame shows that our approach can easily be localized for different font size. The last one, Fig (6) is a video clip of competition of Miss Myanmar which contains a large number of both caption texts and scene texts written with Myanmar language. Myanmar scene texts included in some video frame can be verified that the robustness of the proposed approach. The proposed system successfully detected Myanmar text regions which are not only captions but also scenes regardless of color, size, style, and contrast. The accuracy of the proposed approach for detection of both caption texts and scene texts is shown in Fig (7).

B. Performance Measure for total processing time

In the proposed system, the incoming video frames are encoded with size of 240 x 320. And edge maps analysis is performed to reduce processing time. Moreover, a cross correlation approach in frequency domain is exploited in verification step for faster computation. The processing time depended on the number of detected candidate region. Since post processing after zero-crossing edge map generation is performed in the proposed system, the total processing time can be reduced. It makes our method much faster than other previous methods. It makes possible real time implementation. The experiments were performed on the low-end PC (core i5 2.4 GHz) with 2GB memory. To check the efficiency of edge maps analysis, we measure both the total processing time without update algorithm, and with update algorithm separately using video. The comparison of total processing time is shown in Fig (8).

Although the experimental results in this section showed the efficiency and robustness of the proposed method, the limitations of the proposed method is that the scene text lines

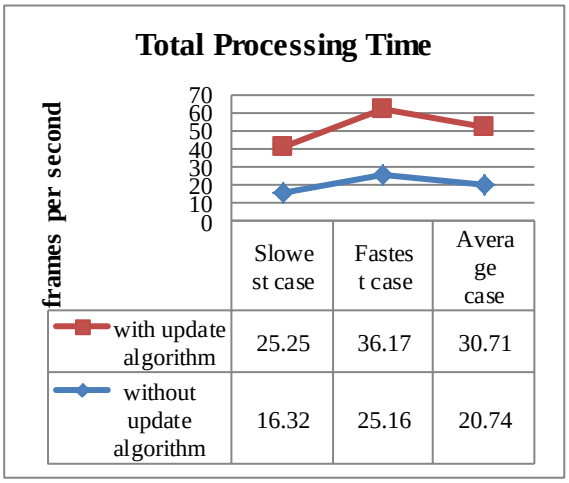


Fig. 8 Performance Evolution of Processing Time

in the experiment are more difficult to detect due to arbitrary orientation.

V. Conclusion

This paper intends to present an improved approach for the identification of Myanmar text areas in digital video considering for both graphical and scene text. Text area identification is critical step to support the final recognition result. Our detection method is based on contrast between text and its adjacent background. Although it is oriented to Myanmar language, text region detection via zero crossing edge maps is well suite for other languages. In verification step, a correlation model in frequency domain is constructed using Myanmar character templates. This is computationally faster than its direct implementation, given the speed advantage of the FFT (Fast Fourier Transform). Various videos have been tested to validate the performance of our detection and localization method. Our approach is able to handle not only caption text but also scene text. Experimental results show that our approach achieves high performance according to the effective and efficient results evaluated by the precision rate and the recall rate.

Acknowledgment

We would like to thank the anonymous reviewers for their constructive comments which helped to improve the paper.

References

[1] M. Cai, J. Song, Michael R.Lyu, “A New Approach For Video Text Detection,” *IEEE ICIP*, 2002, pp. 117–120.

[2] C.P. SUMATHI, T. SANTHANAM, N. PRIYA, "Techniques and Challenges of Automatic Text Extraction in Complex Images: A Survey", Journal of theoretical and Applied Information Technology, 31st January 2012, vol.35, No.2, www.jatit.org.

[3] K. Sobottka, H. Bunke, H. Kronenberg, "Identification of text on colored book and journal covers" , ICDAR, pp: 57-63, 1999.

[4] D. Chen, H. Bourlard, J.P. Thiran, “Text Identification in Complex Background Using SVM” , *IEEE*, 0-7695-1272-0/01 $ 10.00 © 2001.

[5] K. Jung, K.I. Kim and A.K. Jain, “Text information extraction in images and video: a survey,” *Pattern Recognition*, 37, 2004, pp.977-997. http://dx.doi.org/10.1016/j.patcog.2003.10.012

[6] T. Sato, T. Kanade, E.K. Hughes, and M. A. Smith, “Video ocr for digital news archives", In *IEEE* Workshop on Content Based Access of Image and Video Databases, Bombay, January 1998.

[7] M. A.Smith and T. Kanade, "Video skimming for quick browsing based on audio and image characterization", Carnegie Mellon University, Technical Report CMU_CS_95_186, July 1995.

[8] A. K. Jain, and B. Yu, "Automatic Text Location in Images and Video Frames”, Pattern Recognition 31(12), 1998, pp. 2055-2076. http://dx.doi.org/10.1016/S0031-3203(98)00067-3

[9] X. Zhang, L. Gu, "A Combined Algorithm for Video Text Extraction", *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, 2010.

[10] B. Sin, S. Kim, and B. Cho, "Locating Characters in Scene Image using Frequency Features, Proc. of International Conference on Pattern Recognition, 2002, Vol. 3, pp. 489-492

[11] Mark S. Nixon, Albert S. Aguado, *Feature Extraction & Image Processing*, 2nd ed, pp. 193-195

Thuzar Tint received the B.E degree and M.E degree in Information Technology from the Technological University, Mandalay, Myanmar in 2006 and 2008, respectively. She is currently pursuing the Ph.D. degree in Information Technology at University of Technology (Yatanapon Cyber City).Her current research interests are digital image processing, pattern recognition and object segmentation.