Detecting and Segmenting Text from Images for a Mobile Translator System

Thanarat H. Chalidabhongse, and Poonsak Jeeraboon

Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
(Tel : +66-2-737-2551; E-mail: thanarat@it.kmitl.ac.th)

Abstract: Researching in text detection and segmentation has been done for a long period in the OCR area. However, there is some other area that the text detection and segmentation from images can be very useful. In this report, we first propose the design of a mobile translator system which helps non-native speakers to understand the foreign language using ubiquitous mobile network and camera mobile phones. The main focus of the paper will be the algorithm in detecting and segmenting texts embedded in the natural scenes from taken images. The image, which is captured by a camera mobile phone, is transmitted to a translator server. It is initially passed through some preprocessing processes to smooth the image as well as suppress noises. A threshold is applied to binarize the image. Afterward, an edge detection algorithm and connected component analysis are performed on the filtered image to find edges and segment the components in the image. Finally, the pre-defined layout relation constraints are utilized in order to decide which components likely to be texts in the image. A preliminary experiment was done and the system yielded a recognition rate of 94.44% on a set of 36 various natural scene images that contain texts.

Keywords: Text detection, text segmentation, text translation, mobile imaging

1. INTRODUCTION

Presently, Multimedia Message Service (MMS) is a key application in the wireless messaging business and one of the enablers of the Mobile Information Society. Multimedia Messaging brings richer content to mobile communication, and a wide range of value-added services currently plays a dominant role in the Mobile Information Society, in which mobile users are able to easily access a variety of information and services for their specific personal needs. According to BBC news, more mobile phones with built-in camera were sold last year than digital cameras worldwide. From 2002, sales went up almost five-fold to 84 millions. The booming popularity of camera phones, which can take and instantly send photos and short video clips, motivates us to design an imaging system that can assist international tourists and business people to overcome language barriers in Thailand. Millions of foreigners come to Thailand every year either for business or leisure. We believe most of them might have trouble understanding the language when looking for directions, reading menu, etc.

The combination of mobile and ubiquitous computing is emerging as a promising new paradigm with the goal to provide computing and communication services everywhere, at any time, transparently to users. This is the underlying motivation of our work. We have designed a system that detects, recognizes, and translates Thai texts into English texts. The texts embedded in the natural scenes are detected from the images taken by digital mobile cameras. The detected text regions are then fed to character recognition to recognize Thai words. The interpreted words are then translated to English and sent back to the user.

1.1 Related works

Yang et al. [1] developed a tourist assistant system at CMU. The system was equipped with combination of wearable hardware and software such as computer, GPS, camera, head-mounted display, audio in-out, speech recognition, machine translation, OCR, and multimodal interfaces. Recently, they have been developing technologies for automatically translate sign from images and video on handheld PDA device [2] [4]. Their approach employs multi-resolution, adaptive search in a hierarchical framework for text detection, and intensity-based OCR for recognition. The system translates Chinese sign to English text. Haritaoglu [3] and his colleague at IBM Almaden Research Center developed a prototype system of IBM InfoScope. The system ran on a Pocket PC handheld device with an attached digital camera. The prototype can translate Chinese, French, German, and Italian into English. They were also working on Japanese and some other language. Isotani et al. [5] proposed an automatic Japanese-English bi-directional speech-to-speech translation system on PDA that helps oral communication between Japanese and English speakers.

This report addresses only the first key challenge of detecting texts from the natural scene images. The other challenges -Thai OCR and Thai-English machine translation- are still under developed. The paper is organized as follows: the next section describes an overview of the mobile translator system architecture. The details of our method in detecting texts from the natural scene images and some reviews are presented in Section 3. Section 4 presents and discusses the experiments and the results. The conclusion of the paper is in Section 5.

2. SYSTEM ARCHITECTURE

The system architecture of our under-developed mobile Thai-English translator is shown in Fig.1.

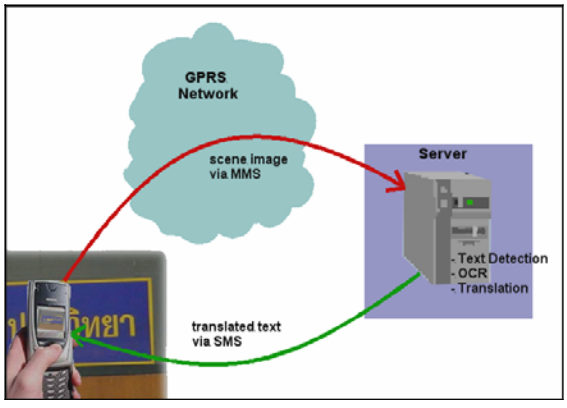

Fig. 1 The mobile Thai-English translator system architecture

To use the service, the user first captures scene image containing texts he wants to translate using his mobile with

built-in camera. The image is then sent to the wireless operator's server via MMS. At the server, a system for text detection, recognition, and translation is running. The following section describes in detail our algorithm used in text detection from natural scene image. The result from this module is detected text regions and segmented components. The text segmentations will be further processed by a Thai character recognition module. Afterward, the recognition result will be fed to a machine translation to translate the recognized Thai words to English. Finally, the translated message is sent back to the user's handheld device via the SMS.

3. SCENE TEXT DETECTION

The text detection can be grouped into two general approaches; area analysis and edge analysis. Wu et al. [6] detected and extracted text in images using multi-scale texture segmentation and spatial cohesion constraints. Gao and Yang [2] proposed an adaptive color modeling and searching algorithm on hierarchical structure approach to discriminate text/non-text regions. Jung et al. [7] introduced an approach of using multiple-CAMShift algorithm on a text probability image produced by a multi-layer perceptron (MLP). Tang et al. [8] extracted caption text from video sequences using temporal information and gray-level vector tracing method. Recently, Zhang and Chang [9] proposed a statistical method to detect text on planar or non-planar with limited angles in natural 3D scene. Their parts-based approach uses a MRF model with higher-order potential and incorporate intra-part relational features at the clique level.

The area-based methods are sensitive to light and scale changes. The edge-based methods are more robust and more suitable for scene text detection [4]. Our approach uses the component relationship analysis on the edge image to detect the text from the natural scene image. The algorithm is shown in Fig.2.

3.1 Preprocessing

First, the input image (I), that is received from the mobile device, is preprocessed to suppress some noise. A discrete smoothing kernel,

$$G_{i,j} = (2\pi\sigma^2)^{-1} \exp(- [(i-k-1)^2 + (j-k-1)^2] / 2\sigma^2), \quad (1)$$

is then convolved with I to yield R:

$$R_{i,j} = \Sigma_{u,v} G_{i-u,j-v} I_{u,v}. \quad (2)$$

where σ is the standard deviation of the Gaussian kernel,
k is the radius of the kernel, and
i,j define pixel's position.

After smoothing the image by the linear filtering, the thresholding is then performed to binarize the image (see Fig.3).

3.2 Edge detection and connected component analysis

The Laplacian edge detection is performed on the binary image. Afterward, the connected component analysis is applied on the edged image (see Fig.4).

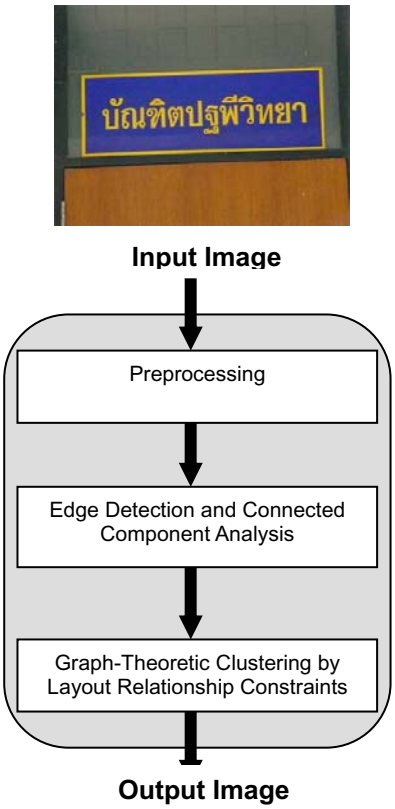

Fig. 2 The proposed scene text detection algorithm

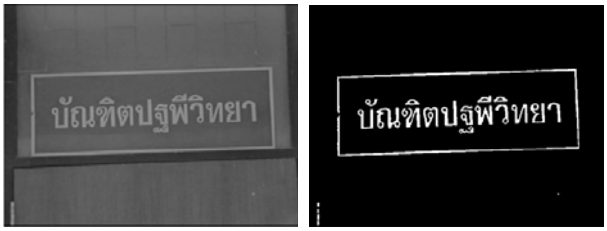

Fig. 3 Filtered image (left) and the result binary image (right)

Fig. 4 Edged image (left) and the labeled connect component image (right)

3.3 Graph-Theoretic Clustering by Layout Relationship Constraints

Based on an observation, we found:

- Characters in the same text have small proximity to each other.
- Characters in the same text have the same alignment.
- Characters in the same text have almost the same compactness.

Then we can formulate the text region detection problem as a graph-theoretic clustering [10] using the above layout relationship constraints. Before clustering the text region, we

first perform a size filtering to eliminate too small components. Afterward, we build a weighted graph by associating each component with a vertex in the graph, and assigning weights on the edges between elements with affinity measures between them. We define three affinity measures based on the three constraints as follows:

Affinity by proximity is defined as

$$Aff_p(x,y) = \exp(-[(x-y)^t (x-y) / 2\sigma_p^2]), \quad (3)$$

affinity by alignment is defined as

$$Aff_a(x,y) = \exp(-[(\boldsymbol{a}(x) - \boldsymbol{a}(y))^t (\boldsymbol{a}(x) - \boldsymbol{a}(y)) / 2\sigma_a^2]), \quad (4)$$

and the affinity by compactness is defined as

$$Aff_c(x,y) = \exp(-[(\boldsymbol{c}(x) - \boldsymbol{c}(y))^t (\boldsymbol{c}(x) - \boldsymbol{c}(y)) / 2\sigma_c^2]). \quad (5)$$

where $\boldsymbol{a}(x)$ is angle between the principle axis of the element x and the vertical line,

$\boldsymbol{c}(x)$ is the compactness value of the element x, and

σ_p, σ_a, and σ_c are user defined parameters used to specify the scale of the above affinity affects.

After constructing the graph, our goal is then to cut the graph into segments with relatively large interior weight or high cohesion. Each segment of the graph represents each text region in the scene image (see Fig.5). The detected text is then forwarded to the recognition module which is beyond this paper.

Fig. 5 The text detection result

4. EXPERIMENTS AND RESULTS

We evaluated our scene text detection algorithm on the database contains 36 color images, shown in Fig. 6. The images were taken from various natural scenes containing texts; images of buildings, airplane, T-shirt, signs, product logo, etc, using a Nokia 6600 camera phone and a 2Mpixel Fujifilm 2600 digital camera. The resolution of the images is 640x480 pixels.

Some detection results are given in Fig. 7. The rectangles indicate the detected text regions. The overall performance of the proposed algorithm on the dataset is given in Table 1. The output produced by the detection algorithm is compared against the ground truth from operator's visual inspection. The text detection is considered to be correct if the rectangles completely cover the whole text. Otherwise, it is considered incorrectly detecting. Fig. 8~9 show some fault detections.

Table 1 Detection rate varied by the size of the Gaussian kernel.

k	Accuracy
1	94.44%
2	86.11%
3	83.33%

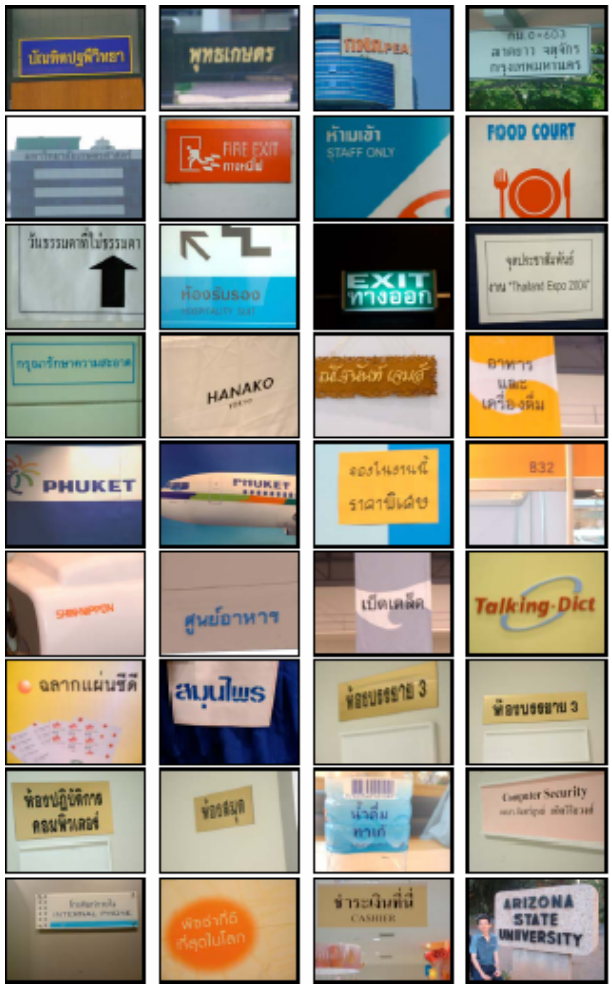

Fig. 6 Dataset used in evaluating our text detection algorithm.

Fig. 7 Some detection results. The rectangles indicate the detected text regions.

Results show that size of the kernel used in image filtering apparently effects the accuracy of text detection. The smoothing kernel forms a weighted average that weighted pixels at its center much more strongly than boundary. When the kernel is large, the neighboring pixels will have larger weights in the weighted average. The noise will largely suppress at the cost of some blurring. In some cases, the characters are blurred and merged to each other and that effects the edge detection and connected component analysis.

Fig. 8 A fault detection results due to using inappropriate kernel size (left) comparing to the one using the right kernel size (right).

Fig. 9 (left) shows another limitation of the proposed algorithm when facing the cases that there are other objects with similar size as text and located close to the text region.

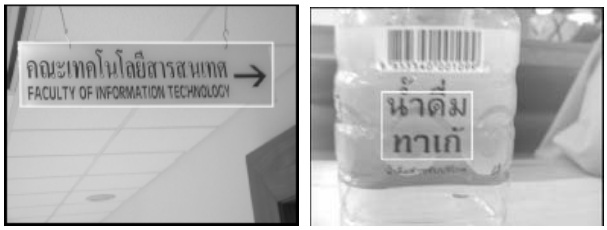

Fig. 9 Other fault detections due to text-like object (left) and the complex structure of written Thai language (right)

Another problem we faced is the fundamental structure of Thai language that is more complex comparing to some other languages such as English. Written Thai language composes of 4 lines of characters: above upper, upper, middle, and lower (see Fig.10). The detection problem arises when there are characters in lower and above upper because most of those characters have smaller size comparing to others and sometimes they were filtered out.

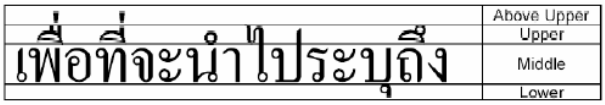

Fig. 10 Written Thai language composed of 4 lines of characters.

5. CONCLUSION AND FUTURE WORK

We present an algorithm for automatic text detection from natural scenes. The work is a part of an on-going development of a mobile Thai-English translator system aims to assist international tourists and business people overcome language barriers. To use the service, the users first capture scene images containing texts they are interested in by cameras equipped with mobile phones. The image is then sent to the service provider's server. At the server, the image is initially filtered to suppress noise and smooth image. Thresholding technique is then applied to binarize the image. Afterward, the Laplacian edge detection algorithm is performed before connected component labeling. Finally, the text region is defined using the graph-theoretic clustering under the layout relationship constraints. The result shows the algorithm can detect the text in various types of scene images with 94.44% accuracy. However, there are still some limitations with the current algorithm especially for complex structure of written Thai language. We are continuing improve the detection algorithm as well as developing the recognition and translation parts to complete the system.

REFERENCES

[1] J. Yang, W. Yang, M. Denecke, and A. Waibel, "Smart sight: a tourist assistant system," *Proc. of 3rd Int'l Symposium on Warable Computers* , pp. 73-78, 1999.

[2] J. Gao, and J. Yang, "An adaptive algorithm for text detection from natural scenes," *Proc. of Computer Vision and Pattern Recognition*, 2001.

[3] I. Haritapglu, "Scene text extraction and translation for handheld devices, " *Proc. of Computer Vision and Pattern Recognition*, Vol. 2, pp.408, 2001.

[4] J. Zhang, X. Xhen, J. Yang, and A. Waibel, "A PDA-based sign translator," *Proc. of the 4th IEEE Int'l Conf. on Multimodal Interfaces*, 2002.

[5] R. Isotani, K. Yamabana, S. Ando, K. Habazawa, S. Ishikawa, T. Emori, K. Iso, H. Hattori, A. Okumura, and T. Watanabe, "An automatic speech translation system on PDAs for travel conversation," *Proc. of the 4th IEEE Int'l Conf. on Multimodal Interfaces*, 2002.

[6] V. Wu, R. Manmatha, and E.M.Riseman, "TextFinder: an automatic system to detect and recognize text in images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 11, pp. 1224 -1229, 1999.

[7] K. Jung, K.I. Kim, T. Kurata, K. Kourogi, and J.H. Han,"Text scanner with text detection technology on image sequence," *Proc. of IEEE Int'l Conf. on Pattern Recognition*, Vol.3, pp.473-476, 2002.

[8] X. Tang, B. Luo, X. Gao, E. Pissaloux, and H. Zhang, "Video text extraction using temporal feature vectors," *Proc. of IEEE Int'l Conf. on Multimedia and Expo*, 2002.

[9] D.Q. Zhang, and S.F. Chang, "Learning to detect scene text using a higher-order MRF with belief propagation," *Proc. of IEEE Workshop on Learning in Computer Vision and Pattern Recognition, in conjunction with CVPR*, 2004.

[10] D.A. Forsyth, and J. Ponce, *Computer Vision, A Modern Approach,* Prentice Hall, 2003.