Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Automatic detection and recognition of Korean text in outdoor signboard images

Jonghyun Park [a,*], Gueesang Lee [a], Euichul Kim [a], Junsik Lim [a], Soohyung Kim [a], Hyungjeong Yang [a], Myunghun Lee [a], Seongtaek Hwang [b]

[a] *School of Electronics and Computer Engineering, Chonnam National University, Republic of Korea*
[b] *Multimedia Lab., Telecommunication R&D Center, Telecommunication Network Business, Samsung Electronics Co. Ltd., Republic of Korea*

ARTICLE INFO

Article history:
Received 27 November 2009
Available online 1 June 2010
Communicated by C.L. Tan

Keywords:
Text detection
Text extraction
Text recognition
Text translation
Signboard image

ABSTRACT

In this paper, an automatic translation system for Korean signboard images is described. The system includes detection and extraction of text for the recognition and translation of shop names into English. It deals with impediments caused by different font styles and font sizes, as well as illumination changes and noise effects. Firstly, the text region is extracted by an edge-histogram, and the text is binarized by clustering. Secondly, the extracted text is divided into individual characters, which are recognized by using a minimum distance classifier. A shape-based statistical feature is adopted, which is adequate for Korean character recognition, and candidates of the recognition results are generated for each character. The final translation step incorporates the database of shop names, to obtain the most probable result from the list of candidates. The system has been implemented in a mobile phone and is demonstrated to show acceptable performance.

1. Introduction

Text contained in an image usually provides semantic information, and is often the crucial key to understanding the image content. These days, digital image capturing devices, including multi-functional mobile devices, are prevalent, and interest in the extraction of text information from natural scene images is increasing (Obinata and Dutta, 2007; Jung et al., 2004). Although text in a regular font of fixed size, with uniform background, can be successfully recognized by existing methods, it is difficult to recognize variable text as used in the real world. Compared with other text detection algorithms developed for computer vision applications, difficulties are compounded in natural scene images due to numerous reasons. The primary problems lie in the variety of the text appearance as it can vary in font style, size, orientation and position. Also, the text images are often corrupted by shadows, reflection of lights, or uneven illumination. They can also be distorted by slants or tilts caused by the position of the camera. Although many commercial OCR systems have good recognition capabilities on high quality scanned documents under well controlled environments, much higher error rates are common for character recognition in real world situations, when the input images do not satisfy the enforced constraints.

This paper presents an automatic translation system for Korean signboard images, where the text usually represents a shop name. The system was originally designed for foreigner tourists with little, or no, knowledge of the Korean language. The system consists of text detection, binarization, recognition and translation. First, local clustering is used to effectively handle the luminance variations of the captured images. The bounding boxes of the individual characters are obtained by connected component analysis. Secondly, the text is recognized using direction features extracted from each character region using a non-linear mesh. Finally, a database of shop names has been used to generate a translation result from the list of recognition candidates. The novel idea of this system is in the translation scheme, in which a database of shop names has been incorporated to compensate for the possible incorrectness of the recognition result, and the most probable interpretation of the word is generated by referencing the database. The block diagram of the proposed system is shown in Fig. 1.

The rest of this paper is organized as follows: In Section 2, problems involved in automatic detection and recognition of text in natural scene images are introduced, with related works in this area. In Section 3, the new proposed method for text detection and binarization is described. In Section 4, a shape-based statistical feature is adopted, which is adequate for Korean character recognition, and the final translation step is described in Section 5. In Section 6, experimental results are presented and Section 7 concludes the paper.

* Corresponding author. Tel.: +82 62 530 0147.
E-mail addresses: jhpark@chonnam.ac.kr, jhpark@jnu.ac.kr (J. Park), gslee@jnu.ac.kr (G. Lee), eckim@jnu.ac.kr (E. Kim), jslim@jnu.ac.kr (J. Lim), shkim@jnu.ac.kr (S. Kim), hjyang@jnu.ac.kr (H. Yang), shwang@samsung.com (S. Hwang).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.05.024

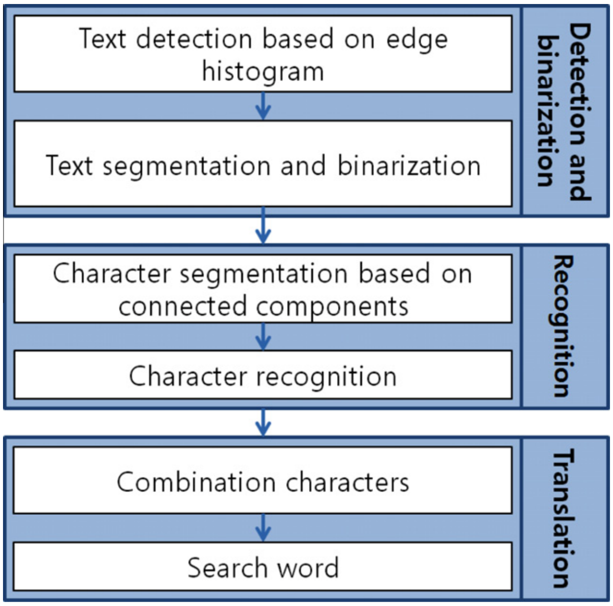

Text detection based on edge histogram
Text segmentation and binarization
Detection and binarization
Character segmentation based on connected components
Character recognition
Recognition
Combination characters
Search word
Translation

Fig. 1. The block diagram of the proposed system.

2. Related works

The interest of this paper is in the automatic detection and recognition/translation of texts in outdoor signboard images for mobile applications. The application scenario is as follows: a camera is used to capture an image, or a sequence of images containing the text of interest.

There have been techniques previously developed for the detection and recognition of written text, targeting applications such as document image processing (Nagy, 2000; Peng et al., 2003), content-based image/video indexing (Li et al., 2000; Xi et al., 2001), assistance for visually impaired persons (Ezaki et al., 2005), text restoration (Ye et al., 2007; Lim et al., 2007), and text recognition or sign translation in natural scene images (Zhang et al., 2002; Yang et al., 2001).

There are many relevant works, particularly, for the detection of a text region. Most of the text detection methods can be classified as either edge-based methods, connected component based methods or texture-based methods. These methods have their own advantages and disadvantages in terms of reliability, accuracy and computation complexity.

The edge-based methods focus on the high contrast between the text and the background (Gllavata et al., 2003; Wu et al., 2005). Once the edges of the text boundary are identified, heuristics are used to filter out the non-text regions. Generally, edges are very useful for image analysis, and they can be used to find the text area, or the bounding area around the text. Text is mainly composed of strokes in a certain direction, so the region with higher edge strengths in a specific direction has a high probability of containing text.

Text detection methods based on connected components use the spatial structure of the connected components, and these work well on text on book covers, news titles or video captions (Jain and Yu, 1998). The connected component based method is a bottom-up approach in the sense that small components are grouped into successively larger components until all relevant regions in the image have been identified. The geometric properties of the connected components are considered in order to filter out the non-textual components, and to mark the boundaries of the region containing the text of interest.

The texture-based method assumes that the text in an image has a distinct textural property which can be used to discriminate the text from the background, or from other non-textual regions (Fujii and Hoefer, 2001). The texture-based algorithms are more robust in dealing with complex backgrounds than the connected component based methods. Tang et al. (2002) has proposed a text detection algorithm that uses texture properties, but the algorithm fails if the texture information cannot be extracted, e.g., in the case of small sized text. Fujii and Hoefer (2001) has proposed a method using wavelet features obtained from the fixed sized blocks of the pixels, and the feature vectors are classified into textual or non-textual groups using neural networks. However, neural networks are not generally efficient in terms of their large computation cost in highly complex images.

For the extraction of a text area, there have been many approaches developed to deal with the possible variations in text orientation, text size, the language used, low image quality, and so on. Most of the existing approaches assume that the text strokes are either horizontal or vertical, and that the text font size is fixed, hence they are very restrictive in the text types they can process.

In order to obtain information in an arbitrary image or a scene captured using a camera, the recognition system automatically recognizes characters of various conditions in the scene, and then provides information about the location of text within the captured image. The text regions, detected in the previous step, can be fed into the recognition step for the classification. This work is related to existing research in the recognition of text on special objects such as car license plates (Mullot et al., 1991). While the early methods required manual selection of the text area (Watanabe et al., 1998; Yang et al., 1999), recent attempts have moved toward automatic detection and recognition of text in natural scenes, for mobile system applications (Zhang et al., 2002; Yang et al., 1999).

Aside from the challenges for the recognition of text in natural scenes, the restrictions on execution time or the overall complexity of the algorithm also need to be enforced in real-time environments, as in mobile systems or personal digital assistants (PDAs). In order to overcome these problems, more robust features are required for the recognition and translation. Efficient methods for text detection, recognition and translation, with their robustness features, are given in the following sections.

3. Text detection

Texts in outdoor signboard images are affected by changes of lights, orientation and the actual location of the signboards, where the orientation is decided by the viewing angle of the camera. In this paper, it is assumed that the text region is located around the center line. In this section, a hierarchical detection framework is presented, including the computation of the edge histogram and the text extraction using fuzzy clustering and connected component analysis.

3.1. Detection of the candidate text region

Although the intensity of pixels in a digital image are important features for text detection, it is not robust in dealing with the variations in the lighting. On the other hand, the edge component is less sensitive to light changes and therefore is more dependable for identifying the text detection. The canny edge detector is applied to the gray-scale image to obtain the edges of the input image. In order to detect the candidates for the region containing the text, the horizontal profile of the edges are computed.

In this paper, it is assumed that texts are aligned horizontally, but the proposed approach can also be extended to vertically aligned texts. Fig. 2 shows edge profiles in the horizontal and

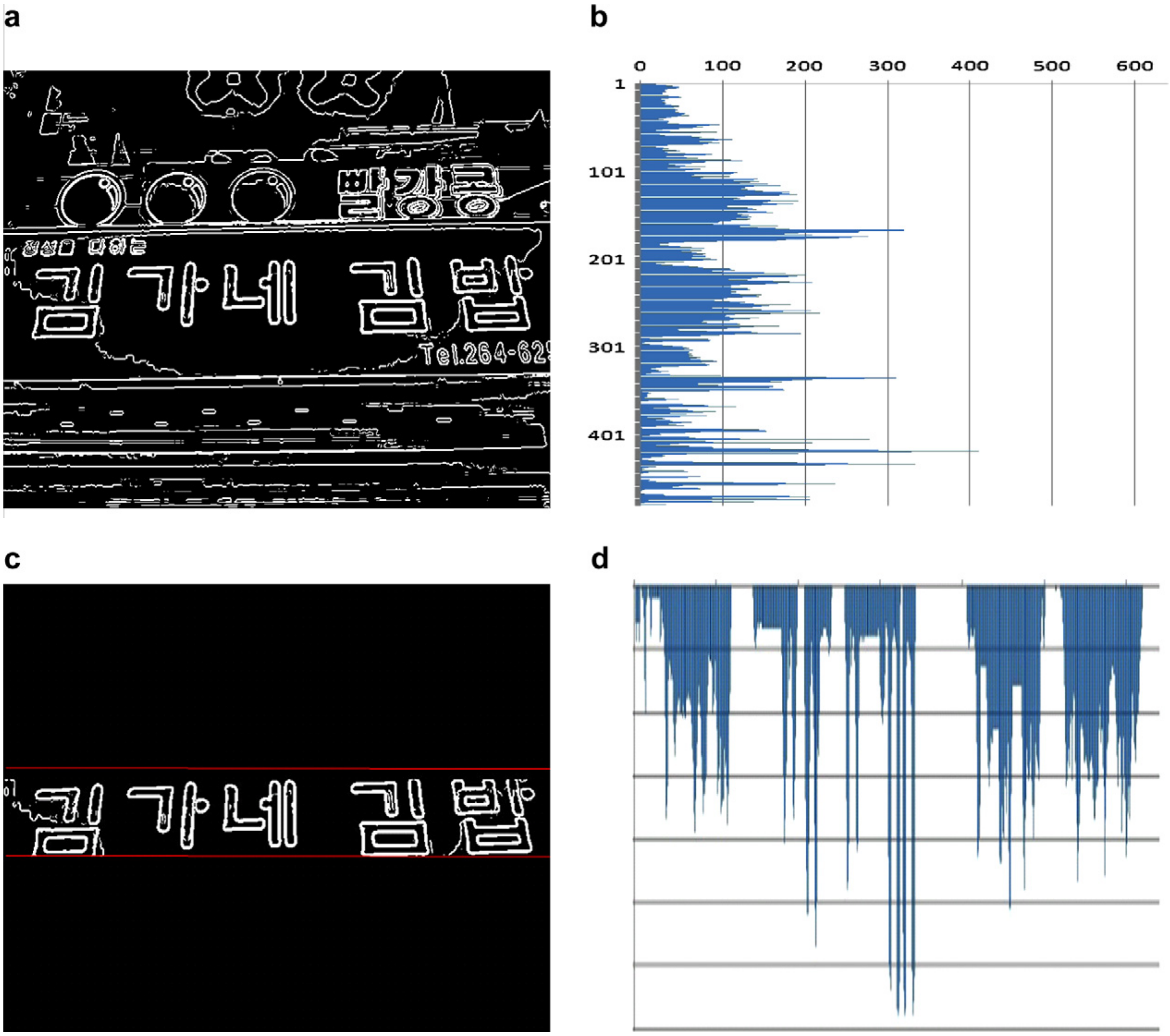


Fig. 2. Detection of the candidate text region with edge profiles: (a) detected edges (b) horizontal edge profile (c) text region obtained by from the horizontal profile (d) vertical edge profile.

vertical directions. By scanning vertically, to the upper or lower ends of the image, starting from the center line of the image, valleys in the histogram are detected, which lead to the identification of the candidate regions with the horizontal profile, as shown in Fig. 2b with a blue box. If the value in the horizontal profile is less than the threshold of HTR/C, it is considered as the valley dividing the candidate region and the background. Horizontal text region (HTR) is defined by the Eq. (1), and C is a constant. Likewise, VTR/C is used for the vertical detection of the candidate region, where vertical text region (VTR) is defined by Eq. (2). Only the region detected from the horizontal profile is used for the vertical detection of the candidate region, as shown in Fig. 2c. To compute the HTR and VTR areas, the summation of the edges in the horizontal or vertical direction is divided by the region size, which approximates the edge density along the scan line.

$$\mathrm{HTR} = \frac{\text{sum of horizontal edges}}{\text{selected region size}} \tag{1}$$

$$\mathrm{VTR} = \frac{\text{sum of vertical edges}}{\text{width of image}} \tag{2}$$

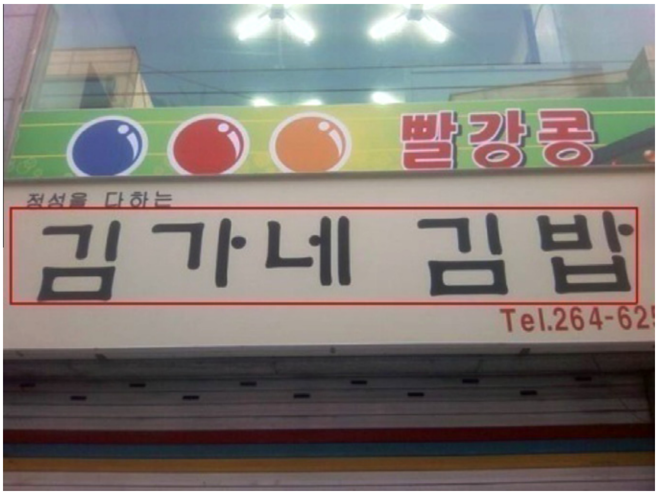


Fig. 3. Detected text region by scanning the edge profiles.

In the vertical profile, shown in Fig. 2d, the valleys detected around the left and right ends are considered as noise or non-text elements in the image. Fig. 3 shows the final bounding box, indicating the candidate text region as a result of the profile scanning.

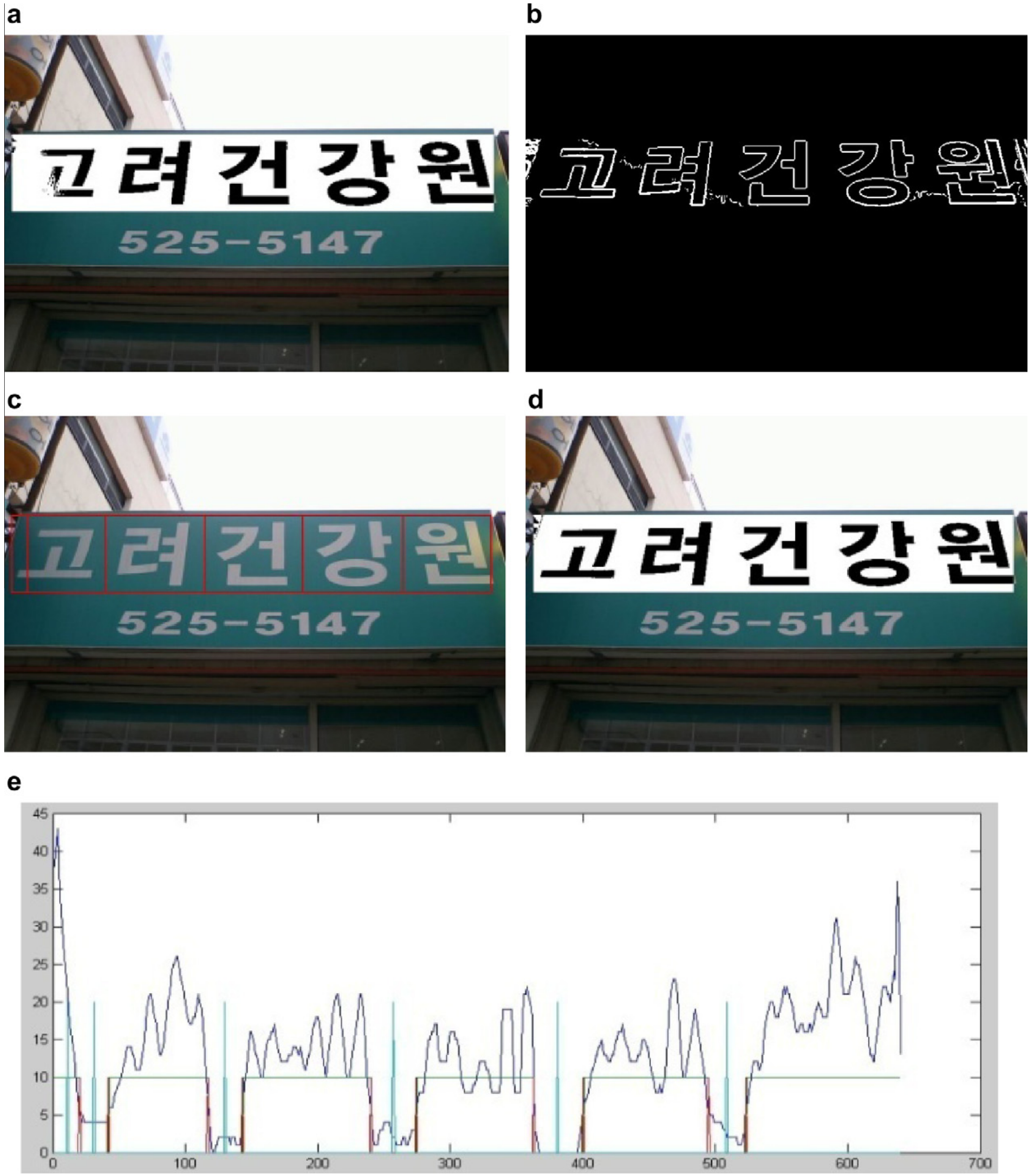


Fig. 4. Segmentation and binarization by the proposed method: (a) segmentation results, (b) edge image in the detected text region, (c) separated image by vertical profile, (d) segmentation results by local features, (e) vertical profile in the (b) image.

3.2. Segmentation and binarization

After the candidate text region is detected, the detected text is binarized. Although the candidate text region is detected, the text in the region may still contain light variations along the text itself. To deal with light changes in the image, fuzzy *c*-means (FCM) clustering is used for the text segmentation as it allows one piece of data to belong to two or more clusters and reduces the likelihood of missing correct matches due to the effects of the noise. It is an unsupervised approach based on the minimization of an objective function that has been used in the image segmentation (Lim and Lee, 1990). It assigns pixels to each cluster by using fuzzy membership. Let $Y = \{y_k | 1 \leqslant k \leqslant N, y_k \in R^d\}$ denote the input image of pixels y_k with d colors, where $d = 1$ for gray level images and N is a constant. The FCM objective function for partitioning a dataset X into c-clusters is given by the following equation:

$$J_m = \sum_{i=1}^{c} \sum_{k=1}^{N} \pi_{ik}^{m} ||y_k - v_i||^2 \tag{3}$$

where $\{v_i\}_{i=1}^{c}$ denotes centroids or prototypes of clusters, m is a real number greater than 1 denoting the amount of fuzziness of the resulting classification, π_{ik} is the degree of membership of y_k in the cluster i, $y_k \in R^d$ is the kth element of d-dimensional pixels,

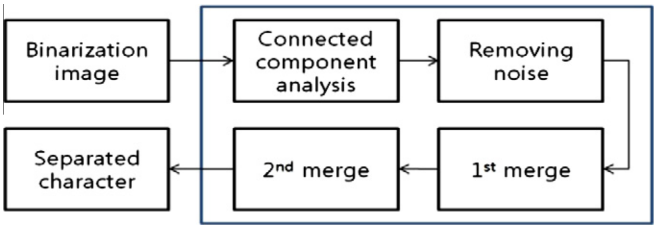


Fig. 5. Character segmentation procedure.

and $||^*||$ denotes any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function in Eq. (3), with the update of membership π_{ik} and the cluster centers $\{v_i\}_{i=1}^{c}$ by update functions:

$$\pi_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{||y_k - v_i||}{||y_k - v_j||}\right)^{\frac{2}{m-1}}}$$
$$v_i = \frac{\sum_{j=1}^{N}\pi_{jk}^{m}\cdot y_k}{\sum_{j=1}^{N}\pi_{jk}^{m}} \quad (4)$$

The iteration stops when the difference between two successive iterations becomes negligible. $\Lambda = [\pi_{ik}]$ is a fuzzy partition matrix satisfying:

$$\Lambda = \left\{\pi_{ik} \in [0,1] | \sum_{i=1}^{c}\pi_{ik} = 1, \forall k,\ N > 0, \forall i\right\} \quad (5)$$

Upon convergence of the clustering algorithm, two mean vectors are found for two dominant groups in a region of text, i.e., text and non-text so that the text region can be digitized.

Fig. 4 shows the binarized result for a sample of text. If text segmentation is performed for the entire candidate region, as shown in Fig. 4a, the digitized result for the first character is not good because of uneven illumination and usually segmentation in local regions can improve the results. Therefore, individual character regions are separated by detecting gaps in the vertical edge profiles of the candidate sub-regions, as shown in Fig. 4e. If a value in the vertical edge profile is less than mean/K for some constant K, we separate the candidate region into local regions for the individual characters. Then, each region enclosed by a red box as shown in Fig. 4c has FCM clustering applied to it for the text segmentation. The results are shown in Fig. 4d.

3.3. Character segmentation

The objective of the character segmentation with the text image is to bind up the connected components into a sequence of characters (Forsyth and Ponce, 2003; Negi et al., 2003). In other words, the text extracted from the previous step is to be divided into individual characters for character recognition. Every Korean character is bounded by a rectangle, in which 2–6 consonants and/or vowels are placed rectilinearly. There are 6 types of such placement, and each of 2350 character classes belong to one of them. Theoretically, the maximum number of Korean characters generated from stroke elements is 11,172, but the Korean standard (KS5601) uses only 2350 characters to form a 2-byte complete-type code. Therefore, we assume that 2350 characters are enough for signboards. Utilizing this fact, a bottom-up procedure is proposed for the character segmentation, as shown in Fig. 5.

Given a binarized image of the text region, the first step is to collect the information of every connected component, namely the width and height of its bounding rectangle, and number of constituent pixels, etc.

The second step is for noise removal; if a connected component C_i, whose width, height and number of pixels are w_i, h_i, and n_i, respectively, satisfies the following condition, it is regarded as noise and is removed from further processing:

$$\{w_i/h_i > 10\ \mathrm{OR} w_i/h_i < 1/10\}\ \mathrm{AND}\ \{n_i/(w_i^* h_i) < 0.5\}$$

The third step is the merging of vertically adjacent components. Given a set of connected components, and a center line which halves the text image vertically, find all the components whose center of mass lie under the center line – see the circled components in Fig. 6a. For each of these components identified, adjacent components which overlap vertically with that component are then merged as a character – see the three overlapping groups in Fig. 6b.

The fourth and final step is the merging of horizontally adjacent components. Given the components are not connected into characters during the first stage of merging we compute the average width a_w of all the remaining components. Then, each component is scanned one by one from left to right, and merged with the component with the next few components as a character if and only if the width of the merged character is no larger than 1.5 × a_w – see the second example in Fig. 7 which has been derived from Fig. 6b.

Fig. 7 shows some examples of character segmentation where the bounding rectangles represent single characters. These examples show that the proposed method can separate the given text image successfully even though there is significant noise and character degradation.

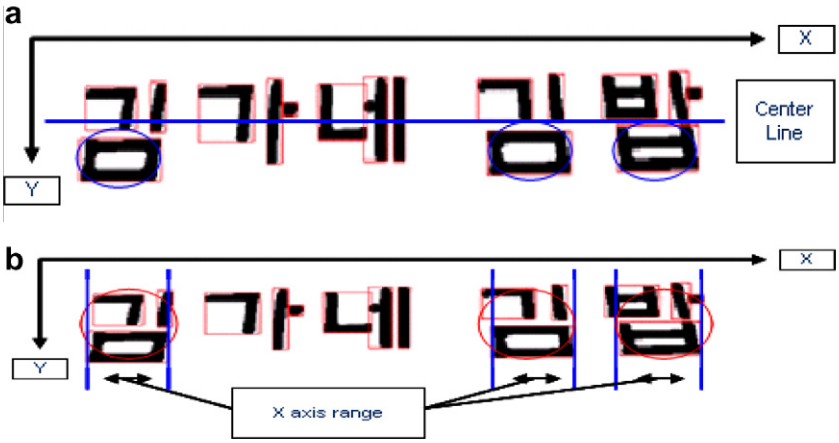


Fig. 6. First stage of merging: (a) Connected components whose center of mass lie below the center line are circled (b) Merging of vertically overlapping components.

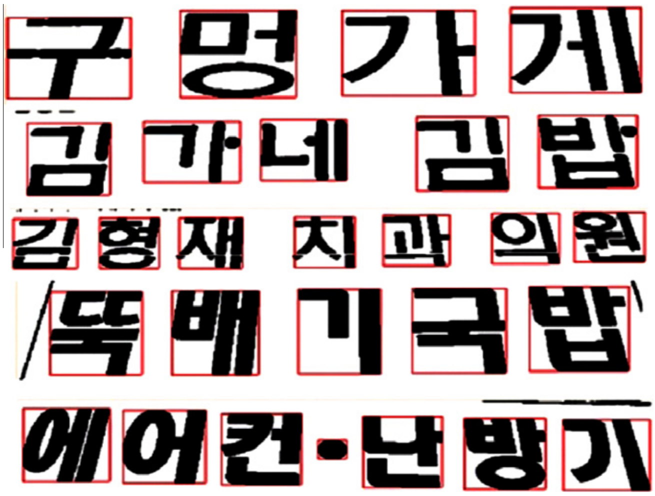


Fig. 7. Examples of character segmentation.

4. Feature extraction and character recognition

A recognition system for the characters captured from outdoor signboard images is more difficult than that needed for normal document analysis. For a traditional document analysis task, a scanner with a stable embedded lighting system is used to obtain high quality images. In the case of outdoor signboard images, however, the noise ratio is much higher because the images are captured by a camera under outdoor conditions with various lighting situations. In addition, there is a wide variety of font styles and font sizes.

For the recognition of Korean characters, two different types of approaches have been developed – structural and statistical. The structural approaches attempt to split the characters into their constituent consonants and vowels, and then recognize each component independently. This approach is intuitive and reduces the number of pattern classes drastically, but the problem of character decomposition becomes non-trivial when the character pattern is degraded, or the image contains some noise or character decorations. The statistical approaches, on the other hand, regard the whole character as a pattern and try to differentiate it from other characters using mesh-based shape information. This approach is simple to implement, but the recognition accuracy is not normally satisfactory because there are many confusing pairs having similar shapes in Korean characters.

Based on the these observations, a statistical approach is selected, which is felt to be more appropriate to the outdoor images containing lighting variations, noise and changes in the use of font styles and character decorations. In addition, this approach has an added advantage of producing a reasonable set of recognition candidates for the input characters by searching similar patterns from the entire character classes.

Once the text image is segmented into isolated characters, the new proposed system classifies the characters one by one using a minimum distance classifier (MDC). The MDC is a Bayesian classifier in which the samples in every class are assumed to be statistically independent and distributed normally with a common variance, and a-priori probabilities are the same for all classes. Given a set of K mean vectors (for K classes), $\mu_1, \mu_2, \ldots, \mu_K$, and an unknown pattern $\mathbf{x}$, the MDC measures the Euclidean distance $||\mathbf{x}-\mu_i||$ from $\mathbf{x}$ to the ith mean vector μ_i, $i = 1, \ldots, K$, and assigns $\mathbf{x}$ to the category of the nearest mean vector.

Since the assumption of normal distribution with a common variance is not generally valid, the classification accuracy of the MDC is not optimal. However, the simplicity of the algorithm makes it a popular method for a preliminary classification in large-set character recognition applications (Kim, 1997; Tsukumo, 1992; Tung et al., 1994; Yamashita et al., 1983). Here the MDC is adapted as a primary classifier for the recognition of Korean characters. There are 2350 character classes in Korean Standard, but the proposed MDC deals with only the 808 classes which are used most frequently in signboards and have 99% coverage of our signboard dictionary. 808 characters were selected in this research to cover 99% of all characters in the Korean signboard database. If the character set is increased, the coverage gets higher with the penalty of degraded performance in the recognition. This number has been chosen as the best candidate in the trade-off.

4.1. Feature extraction

A shape-based statistical feature is adopted, which is adequate for Korean character recognition, as shown in Fig. 8. Given a character pattern represented by a black and white image, an $N \times M$ non-linear mesh and four directional segments for horizontal (H), vertical (L), left-diagonal (L), and right-diagonal (R) directions are prepared, and then $4 \times N \times M$ feature values are computed using the mesh and the four directional segments.

Assuming that the character pattern is an $H \times W$ image $f(i, j)$, where (i, j) is the pixel at ith row and jth column ($1 \leqslant i \leqslant H$, $1 \leqslant j \leqslant W$), then the construction of an $N \times M$ mesh ($N \leqslant H$, $M \leqslant W$) can be described as follows. First, the horizontal and vertical projection profiles, $h(i)$ and $v(j)$, are computed using

$$h(i) = \sum_{k=1}^{W} f(i,k), \quad 1 \leqslant i \leqslant H, \tag{6}$$

$$v(j) = \sum_{k=1}^{H} f(k,j), \quad 1 \leqslant j \leqslant W. \tag{7}$$

and the rows of the input image, indexed from 1 to H, are partitioned into N consecutive strips, $[1, e_1]$, $[e_1 + 1, e_2], \ldots, [e_{N-1} + 1, H]$, in which the boundary indices $e_1, e_2, \ldots, e_{N-1}$ are determined so that the black pixel densities defined by the following equation are the same for all strips:

$$D(k) = \sum_{i=s_k}^{e_k} h(i), \tag{8}$$

where $s_1 = 1$, $s_{k+1} = e_k + 1$, $e_N = H$, and $1 \leqslant k \leqslant N$. Similarly, the columns of the input image, indexed from 1 to W, are partitioned into M strips with the vertical projection $v(j)$. Combining the boundary indices of the N horizontal strips and M vertical strips, an $N \times M$ mesh can be formed, as in Fig. 8a, where $N = 8$, $M = 8$. The non-linear partitioning of the input pattern is used to allow for the variation of different font styles and character decorations that are commonly used in Korean characters.

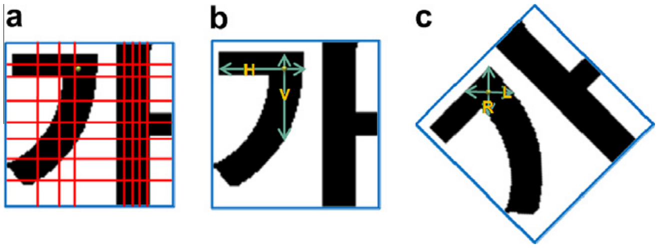


Fig. 8. An example of feature extraction: (a) 8 × 8 non-linear mesh, (b) horizontal (H) and vertical (V) segments from the input image, (c) left-diagonal (L) and right-diagonal (R) segments from 45-degree rotated input image.

Next, the four directional segments are calculated for every black pixel in the image, considering the fact that the structural characteristics of Korean characters are represented by horizontal, vertical and two main diagonal lines. Given a black pixel (x, y) in the input character image, the two run-lengths $RLH_{x,y}$ and $RLV_{x,y}$ are computed by measuring the length of horizontal and vertical runs passing through the pixel (x, y), as shown in Fig. 8b. Using these two values, compute the values of horizontal and vertical contributions $DCH_{x,y}$ and $DCV_{x,y}$ as follows:

$$DCH_{x,y} = RLH_{x,y}/(RLH_{x,y} + RLV_{x,y}) \tag{9}$$

$$DCV_{x,y} = RLV_{x,y}/(RLH_{x,y} + RLV_{x,y}) \tag{10}$$

The directional contribution values $DCH_{x,y}$ and $DCV_{x,y}$ are averaged over all the black pixels within each cell of the $N \times M$ mesh to get $2 \times N \times M$ feature values. The other $2 \times N \times M$ feature values for the left-diagonal and right-diagonal directions are computed by a similar computation with the 45-degree rotated image, as in Fig. 8c.

In the proposed system, a 9×7 mesh is selected, according to statistics of the aspect ratio of Korean characters. So, the character pattern in a digital image is transformed to a 252 $(9 \times 7 \times 4)$ dimensional feature vector.

4.2. MDC with prototype learning

In the minimum distance classifier (MDC), one prototype is used for each class. Since 808 classes of Korean characters have been chosen by analysis of the frequency of appearance in a dictionary of signboard names, 808 prototypes, $\mu_1, \mu_2, \ldots, \mu_{808}$, should be available for MDC.

For learning each prototype, 200 different font samples have been used. The training samples are obtained by using a computer typesetting system for changing the font typefaces, font styles, and font sizes to make its shape similar to that of outdoor signboards. The ith prototype μ_i is trained by averaging the feature vectors of all the samples belonging to the ith class:

$$\mu_i = \frac{1}{200}\sum_{j=1}^{200} F_j^i, \text{ where } F_j^i \text{ is the } j\text{-th sample of the } i\text{-th class.}$$

Given an unknown pattern $\mathbf{x}$, the MDC measures the Euclidean distance $||\mathbf{x} - \mu_i||$ from $\mathbf{x}$ to the ith prototype μ_i, $i = 1, \ldots, 808$, and produces a set of the 10 most probable classes, which are ordered according to the distance value. The smaller the distance between $\mathbf{x}$ and μ_i, the higher the probability of $\mathbf{x}$ belonging to the ith class.

4.3. Reordering of the recognition candidates

Given an ordered set of recognition candidates computed from the MDC, the reordering module rearranges them by the use of pairwise classifiers. The reordering starts with invoking a pairwise classifier for the pair of two candidates of higher distance (or smaller similarity), and then the winner of this classification and the next candidate are prepared for another reordering, and so on.

The possible number of character pairs composed of K different classes is $K(K-1)/2$ and as there are 808 character classes in the system, there are as many as 326,028 possible pairs. This large number of pairwise classifiers makes the pair reordering task impractical, but it has been observed that just 1.4% of the most frequent pairs cover more than 40% of the pairs in a real application. In addition, it has also been observed that approximately 14% of the most frequent pairs cover more than 90% of the entire pairs. Therefore the classifiers have been constructed only for the 14% of all possible character pairs.

A pairwise classifier for the ith class and jth class $(1 \leqslant i, j \leqslant 808)$ is another MDC, which uses a 32D feature. The 32D feature is constructed as a subset of the original 252D directional segment feature. The selection is based on the Fisher discriminant measure (Sugiyama, 2006, 2007) $F_{ij}(k)$ computed as

$$F_{ij}(k) = \frac{\sigma_{ij}(k)}{\sigma_i(k) + \sigma_j(k)} \tag{11}$$

where $\sigma_i(k)$ is the variance of the kth feature value of the samples belonging to the ith character class, $\sigma_j(k)$ is the variance of the kth feature value of the samples belonging to the jth character class, and $\sigma_{ij}(k)$ is the variance of the kth feature value of the samples belonging to both classes. A higher value of $F_{ij}(k)$ implies that the kth feature has a high discrimination power between the ith class and the jth class. We choose 32 features having the higher Fisher measure.

The 32D sub-vectors for the most confusing pairs are prepared during the training stage. Since the pairwise classifiers are prepared only for 14% of the entire number of possible pairs, the reordering is not performed for the pair whose classifier does not exist. An ordered list of 5 recognition candidates is generated as a final result of the character recognition.

5. Translation

The translation phase considers n candidates of characters from the recognition step to boost the accuracy of the translation. One of the combinations among the candidate characters is translated to a target word in English.

We define the input data $R = \{R_1, \cdots, R_m\}$ for translation, where m is the number of characters and $R_m = \{r_{m1}, \cdots, r_{mn}\}$ as a set of candidate characters where n is the number of candidates. For example, if $m = 4$ characters which should be recognized and $n = 5$ candidates are collected for each character, then the following matrices are evident:

$$R = \begin{bmatrix} r_{11} & \cdots & r_{15} \\ \vdots & \ddots & \vdots \\ r_{41} & \cdots & r_{45} \end{bmatrix} \tag{12}$$

Table 1 shows an input data set from the recognition result in which the correct word is “구멍가게”in Korean.

As seen in Table 1, only considering the first ranked characters does not guarantee an accurate result since the right combination is $[r_{14}, r_{22}, r_{31}, r_{41}]$ for “구멍가게”. Therefore, all combinations of the candidates should be considered, aligning the possible results with a word dictionary to ensure that a correct word has been formulated. If the top n candidates from the recognition step are considered to be combined, the number of word combinations is as follows:

$$C = n^i, \tag{13}$$

where C is the number of combinations, i is the number of letters, and n is the number of candidates.

However, aligning all the combinations takes a long time with respect to the number of characters in a word. Therefore, words having long length and which are not frequently used in signboards have been excluded to save processing time. To minimize

Table 1
Input data from the recognition result (corrected characters are in bold red-colored).

Character No.	Candidate characters				
	1st	2nd	3rd	4th	5th
1	긋	곳	꿋	**구**	근
2	멈	**멍**	엄	덩	엉
3	**가**	겨	거	기	게
4	**게**	계	개	자	지

unnecessary combinations, a range of combinations will be set up with the input letters. Fig. 9 shows the length of words used in the names of Korean signboards. As a result, 2-letter words occupy around 60%, 3-letter words occupy around 30%, and 4-letter words occupy less than 10%, respectively. Words with 5 or more letters words are not included and become meaningless words leading to segmentation error words as they are not frequently used words for signboards.

Therefore, in this paper, 4-letter words are considered to be the longest length for practical combinations, while 2-letter words are the minimum length to be translated. If there are more than two possible words from the combinations, the longest word is chosen since it is more common in the Korean language that two words are concatenated in a word.

In combining words from the possible candidates, a top-down method is used in which the range is reduced starting from the longest combination. Since the maximum and minimum length of words are set, when the length of a word is less than the maximum the whole length of the word is considered in order to produce the possible combinations. However, if the length of the text is more than the maximum, the index of the starting character of the combination is defined as follows:

$$S = (N + 1) - M, \qquad (14)$$

where S is a starting index of the combination, N is the number of characters, and M is the maximum length of the words; as N is the last index of the combination, the range of the possible combinations is from S to N. If there is no word which is able to be aligned, the range of the combination is reduced by increasing S so that the distance to N is decreased. When the search fails with the minimum possible combination, the first ranked characters are translated as it sounds. The above process is repeated with the remaining characters. In this process, however, there can be multiple results. In that case, the words with the highest matching rate and highest frequency are selected. Fig. 10 shows how the starting index is decided. The correct word is [대학생]+[선교회] in Korean.

The translation, by aligning with words in a dictionary, requires frequent access to a dictionary. To reduce the search time, the frequent words list and two-level index are used, as shown in Fig. 11. The words generated from the same combination range have the same word length. By using this same length property, the search space is reduced. The word is searched by using the initial consonant which is 14 in Korean and the words in a dictionary are classified into 14 consonants from ‘ㄱ’ to ‘ㅎ’.

The hash function h is given by Eq. (15) and M is a constant number to adjust the minimum length of the words and the number of initial characters in Korean is 14. The time of translation is reduced, saving the searching cost by using a binary searching algorithm

$$h = (L - M) \times 14 + I, \qquad (15)$$

where L is the length of a word, I is the order of initial characters, and M is the minimum length of the word.

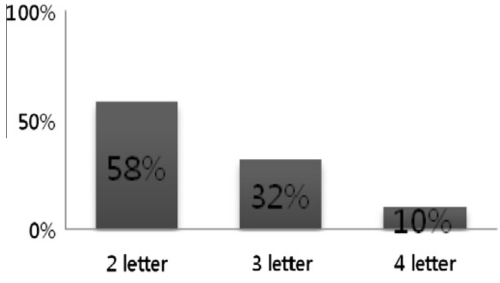
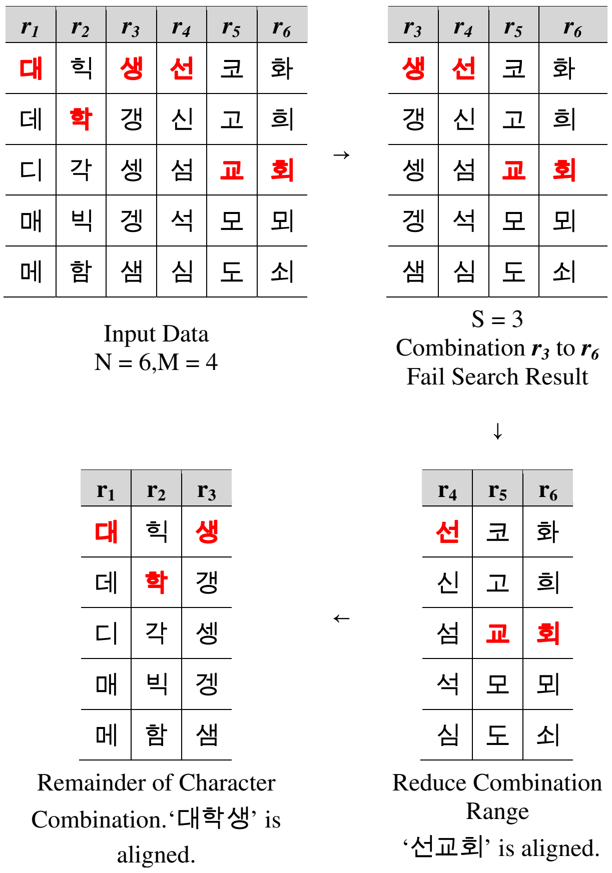


Fig. 9. Percentage of words used with their length.



r_1	r_2	r_3	r_4	r_5	r_6
대	힉	생	선	코	화
데	학	갱	신	고	희
디	각	셍	섬	교	회
매	빅	겡	석	모	뫼
메	함	샘	심	도	쇠

Input Data
N = 6,M = 4

→

r_3	r_4	r_5	r_6
생	선	코	화
갱	신	고	희
셍	섬	교	회
겡	석	모	뫼
샘	심	도	쇠

S = 3
Combination r_3 to r_6
Fail Search Result

↓

r_4	r_5	r_6
선	코	화
신	고	희
섬	교	회
석	모	뫼
심	도	쇠

Reduce Combination Range
‘선교회’ is aligned.

←

r_1	r_2	r_3
대	힉	생
데	학	갱
디	각	셍
매	빅	겡
메	함	샘

Remainder of Character Combination.‘대학생’ is aligned.

Fig. 10. Example of reducing the combination range.

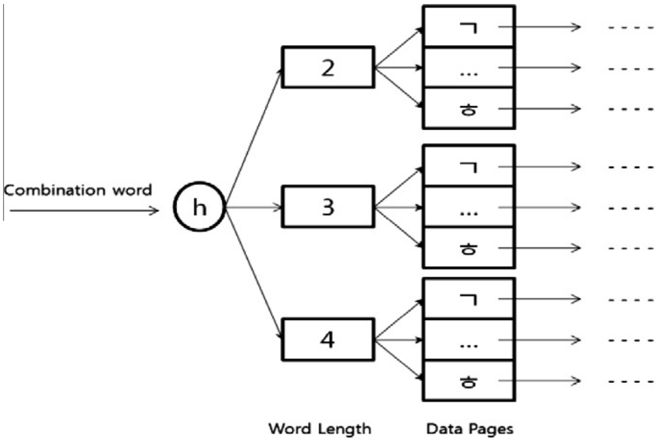


Fig. 11. Multiple index structure of the dictionary.

Words not recorded in a dictionary, such as proper nouns, will be translated to Roman alphabet phonetically. There are two ways to translate phonetically. The first way is to make the sounds with the Roman alphabet after dividing one character into three parts, such as the initial sound, the middle sound and final sound. This way does not generate the correct sound in Roman alphabets as it sounds in Korean, since if there are two consonants in a final part it does not sound both of them. Therefore, if the actual pronunciation of the alphabet sound is different to the other sound converted to the Roman alphabet, the output is not accurate. The second way is to build and use codes of the frequently used 808 words among the 2350 characters into it equivalent Roman translation. Even though this way needs more space, it will give more accurate results in the Roman notation than the first method.

6. Experimental results

Experiments have been conducted on signboard scene images captured in different light conditions. The text appearance varies with different colors, orientations and font sizes. Korean signboard images are captured by a Samsung Smart-phone, with an image resolution of 640 × 480. The mobile system is equipped with intel XScale PXA270 416 MHz and the operating system is MS window mobile 5.0 for Pocket PC.

6.1. Text region detection and binarization

Fig. 12 shows the results of the detection and binarization of texts, where the red rectangles denote detection results. Generally, outdoor signboard images might be taken under uneven illumination, such as gradation, highlights, shadows in the text region, which degrades the performance of the system. Fig. 13 shows the

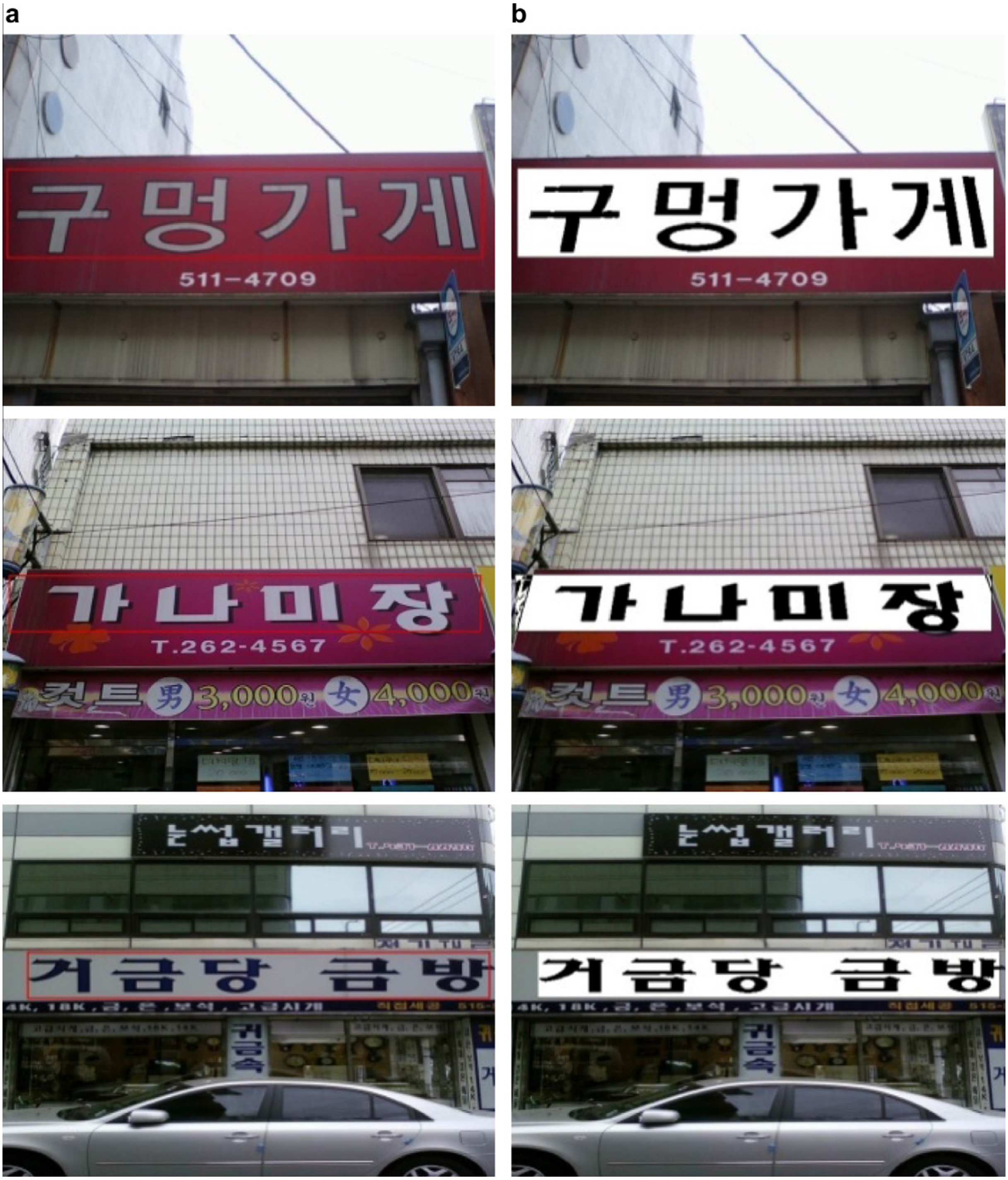


Fig. 12. Text detection and binarization results in Korea signboards: (a) detected regions, (b) digitized results

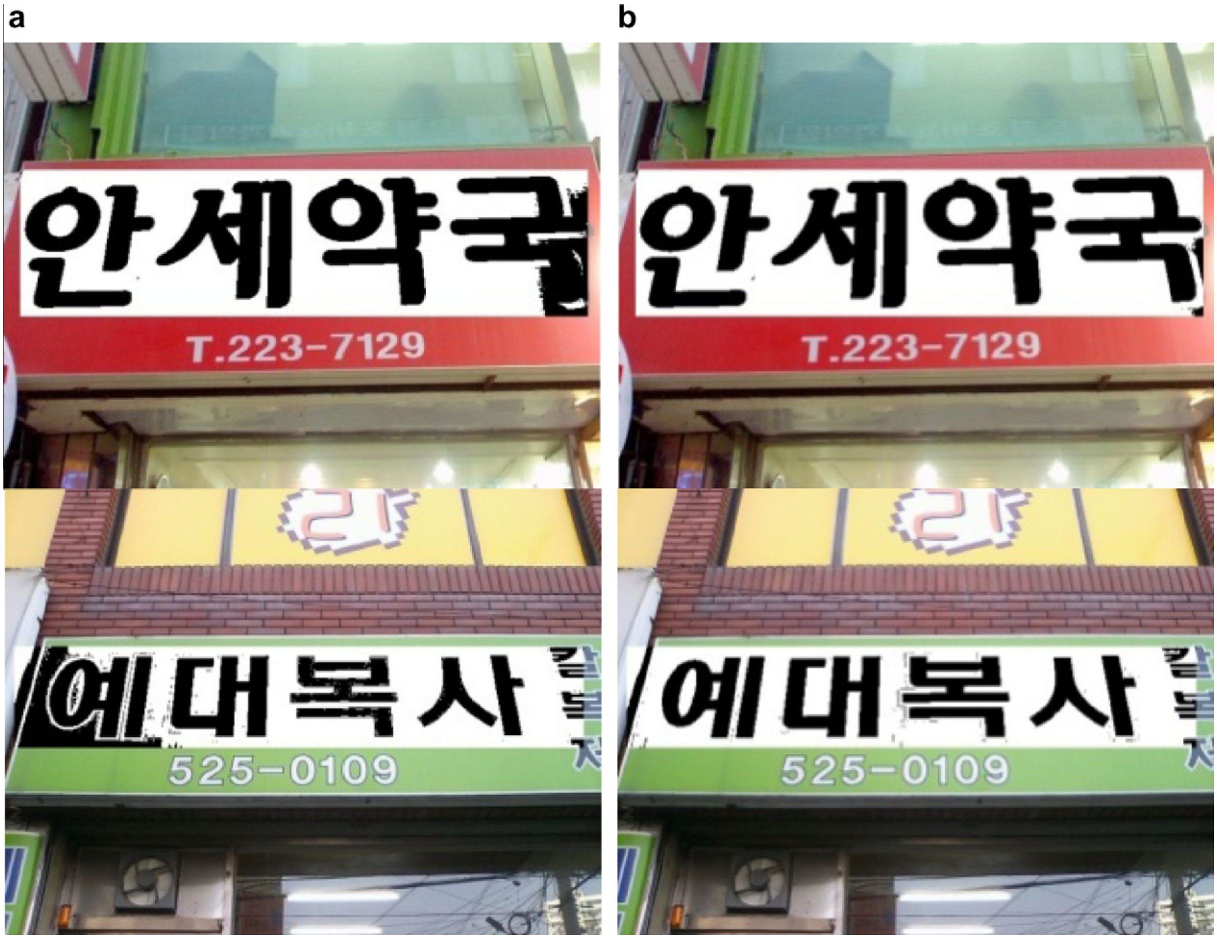


Fig. 13. Examples of binarization in signboard images with uneven illumination: (a) binarization results using global features, (b) binarization results using local features.

binarization results of images with uneven lights, where Fig. 13a shows the result of the segmentation using global features, and Fig. 13b shows the result using local features, indicating approaches with local features are generally better for dealing with variations in light changes with other noise or deformations.

The evaluation result of the test image set is summarized in Table 2. The experiments involved tests using 445 images in total. In order to evaluate the performance, the results are classified into three groups where the first group is *failure*, when the results cannot be used in the next step for text recognition; the second group is *acceptable*, in which case the main text and the background can be separated but includes isolated regions or noise; and the third group is *success*.

6.2. Recognition results

The proposed character recognition algorithm has been evaluated with 563 Korean characters extracted from outdoor signboard images. The recognition accuracy is summarized in Table 3 where it can be seen that the recognition accuracy of MDC is 60.92% for the first choice, and 84.01% for the top five choices. After the reordering module is performed, the recognition accuracy is increased to 73.18 and 85.97%, respectively.

In addition to the recognition accuracy for the first candidate, the top-five cumulative accuracy is also an important performance criterion for the proposed system because the dictionary-based post-processing is able to detect the correct answers if they are included in the candidate list.

Table 2
Text detection and binarization ratio in the dataset used.

	Num. of images	Detection and binarization (%)
1st group	57	13
2nd group	252	57
3rd group	136	30

6.3. Translation results

Fig. 14 shows the results of the translations where Fig. 14b shows that the candidate text region is detected and then divided

Table 3
Recognition accuracy for signboard characters.

Recognition order	The proposed approach			
	MDC with DSF		After reordering	
	Recognition accuracy (%)	Cumulative accuracy (%)	Recognition accuracy (%)	Cumulative accuracy (%)
1st	60.92	60.92	73.18	73.18
2nd	12.61	73.53	7.28	80.46
3rd	5.68	79.22	2.84	83.30
4th	3.20	82.42	1.60	84.90
5th	1.60	84.01	1.07	85.97
Failure	15.99	15.99	14.03	14.03

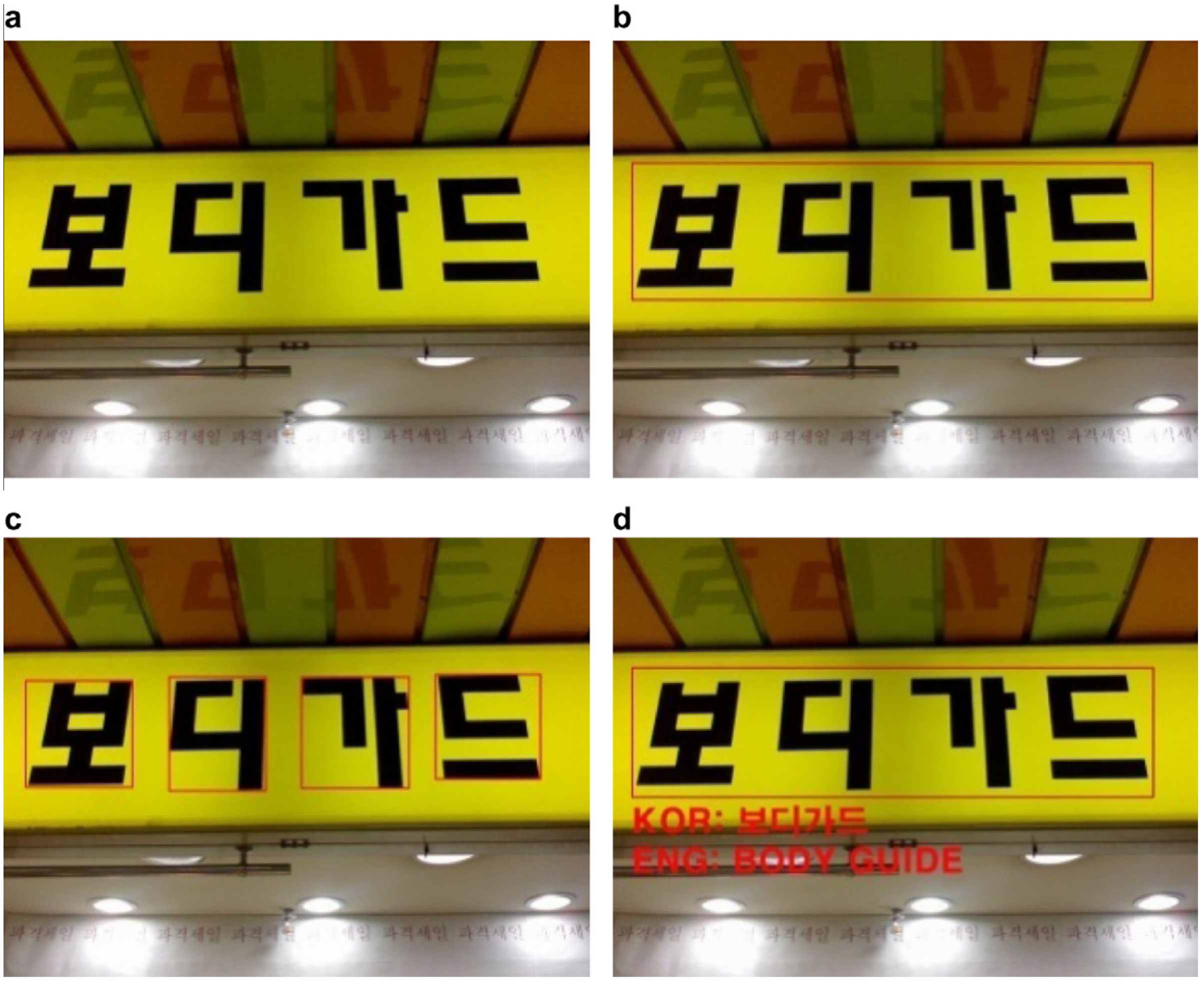


Fig. 14. The translation result of character extraction and recognition: (a) signboard image, (b) detected text region, (c) separated characters and (d) recognition and translation result.

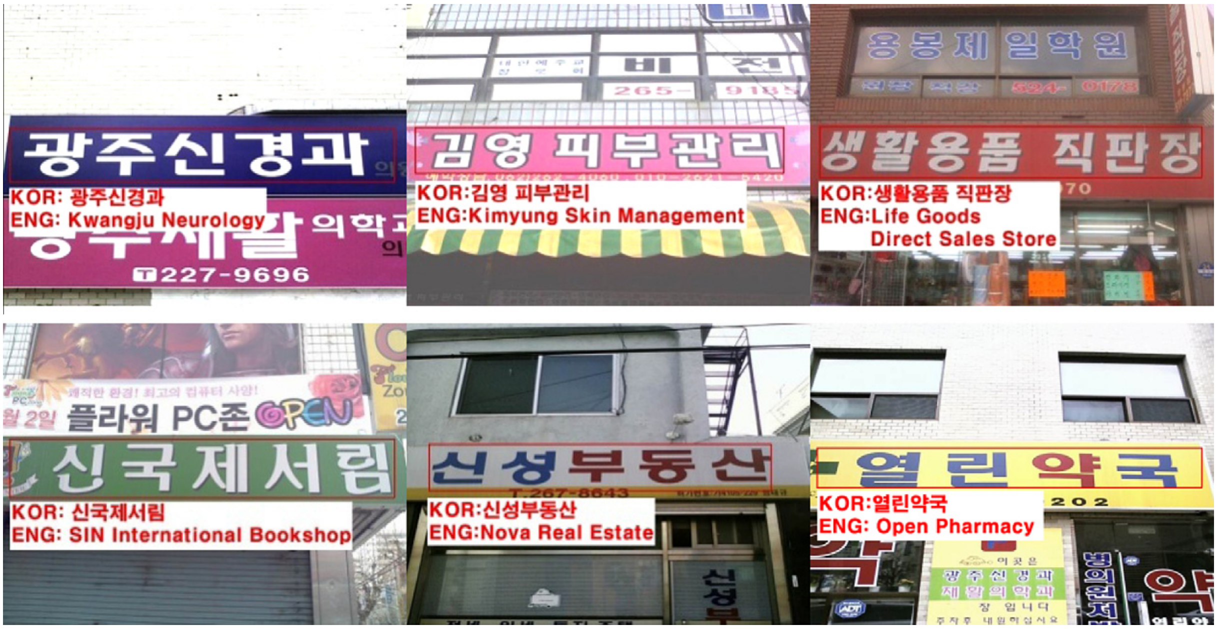


Fig. 15. Examples of automatic text detection, recognition and translation.

Table 4
Translation accuracy for recognized characters.

Success		Failure	Total
Success	Acceptable		
240	15	14	269
90%	5%	5%	–
95%		5%	–

into individual characters, as shown in Fig. 14c. The final result is given Fig. 14d. Fig. 15 shows the detection, recognition and translation results with other samples posters.

The performance of the translation is evaluated with 269 signs of which the recognizer generates the correct characters using the 5 candidates. The results are classified as success, acceptable and failure. "Success" is when the full-text translation is performed successfully, "acceptable" is when key-words are correctly translated (here the key-words are the main words which distinguish the kind of stores). As shown in Table 4, success and acceptable form 90% and 5% of the overall percentages of the results, respectively.

Since the system consists of cascading stages, the final result can be obtained when all steps must be performed successfully. The final success rate is the product of the success rates for each step, which is about 71%. The minimum success rate for the system has been set by industry experts, therefore we assume that the performance reaches the requirement of a commercial product. In reality, input images contain all kinds of exceptional inputs such as unusual alignments or graphic font styles, and frequently they are corrupted by noises such as uneven illumination, occlusions, color changes, and others. Therefore the success rate over 70% is not an easily attainable goal. Much higher success rate is expected if the user captures the input image with care excluding exceptional noises.

7. Conclusions

This paper has described a system for automatic detection, and recognition of Korean texts in signboard images captured by a mobile phone camera and finally translating the content into English. The captured images are taken in natural lighting environments, where noise and irregularities of fonts occur frequently. The proposed system can robustly detect and recognize texts from such images. The edge-based method is used for the detection of the candidate text region, and the text is digitized by using fuzzy c-means clustering in local regions of individual characters. For character recognition, a shape-based statistical feature extraction is employed and the translation is carried out by checking against a definitive list of recognized candidates and referencing the top shop names in a dictionary to find the most probable answer. The experimental results show that the proposed method has been successfully applied to recognize and translate Korean shop names into English with their outdoor signboard images.

Acknowledgement

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-C1090-1011-0008).

References

Ezaki, N., Kiyota, K., Minh, B.T., Bulacu, M., Schomaker, L., 2005. Improved text-detection methods for a camera-based text reading system for blind persons. International Conference on Document Analysis and Recognition, 257–261.

Forsyth, D.A., Ponce, J., 2003. Computer Vision A Modern Approach. Prentice-Hall.

Fujii, M., Hoefer, W.J.R., 2001. Filed-singularity correction in 2-D time-domain Haar-wavelet Modeling of waveguide components. IEEE Trans. Microwave Theory Technol. 49 (4), 685–691.

Gllavata, J., Ewerth, R., Freisleben, B., 2003. A robust algorithm for text detection in images. Int. Symp. Image Signal Process. 2, 611–616.

Jain, A.K., Yu, B., 1998. Automatic text location in image and video frames. Int. Conf. on Pattern Recognition 2, 1497–1499.

Jung, K.J., Kim, K.I., Jain, A.K., 2004. Text information extraction in images and video: a survey. Pattern Recognition 37, 977–997.

Kim, S.H., 1997. Performance improvement strategies on template matching for large–set character recognition. Int. Conf. Comput.Process. Oriental Lang., 250–253.

Li, H., Doermann, D., Kia, O., 2000. Automatic text detection and tracking in digital videos, IEEE Transactions on Image Processing 9(1), 147–156.

Lim, Y.W., Lee, S.U., 1990. On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques. Pattern Recognition 23 (9), 935–952.

Lim, J.G., Park, J.H., Medioni, G.G., 2007. Text segmentation in color images using tensor voting. Image Vision Comput. 25, 671–685.

Mullot, R., Olivier, C., Bourdon, J.L., Courtellemont, P., Labiche, J., Lecourtier, Y., 1991. Automatic extraction methods of container identity number and registration plates of cars. Int. Conf. Ind. Elect. Control Instrument. 2591, 1739–1744.

Nagy, G., 2000. Twenty years of document image analysis. IEEE Trans. Pattern Anal. Machine Intell. 22 (1), 38–62.

Negi, A., Shanker, K.N., Chereddi, C.K., 2003. Localization, extraction and recognition of text in Telugu document images. Int. Conf. Document Anal. Recognit., 1193–1197.

Obinata, G., Dutta, A., 2007. Vision Systems: Segmentation and Pattern Recognition. I-Tech.

Peng, H., Long, F., Chi, Z., 2003. Document image recognition based on template matching of component block projections. IEEE Trans. Pattern Anal. Machine Intell. 25 (9), 1188–1192.

Sugiyama, M., 2006. Local Fisher discriminant analysis for supervised dimensionality reduction. Int. Conf. Machine Learn., 905–912.

Sugiyama, M., 2007. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. J. Machine Res. 8, 1027–1061.

Tang, X., Gao, X., Liu, J., Zhang, H., 2002. A spatial-temporal approach for video caption detection and recognition. IEEE Trans. Neural Network 13 (4), 961–971.

Tsukumo, J., 1992. Handprinted Kanji character recognition based on flexible template matching. Int. Conf. Pattern Recognit., 483–486.

Tung, C.H., Lee, H.J., Tsai, J.Y., 1994. Multi-stage pre–candidate selection in handwritten Chinese character recognition systems. Pattern Recognition 27 (8), 1093–1102.

Watanabe, Y., Okada, Y., Kim, Y.B., Takeda, T., 1998. Translation camera. Int. Conf. Image Process., 613–617.

Wu, W., Chen, X., Yang, J., 2005. Detection of text on road signs from video. IEEE Trans. Intell. Transport. Syst. 6 (4), 378–390.

Xi, J., Hua, X., Wenyin, L., Zhang, H.J., 2001. A video text detection and recognition system. International Conference on Multimedia and Expo, 873–876.

Yamashita, Y., Higuchi, K., Yamada, Y., Haga, Y., 1983. Classification of hand-printed Kanji characters by the structured segment matching method. Pattern Recognition Letter 1, 475–479.

Yang, J., Yang, W., Denecke, M., Waibel, A., 1999. Smart Sight: A Tourist Assistant System. Int. Symp. Wearable Comput., 73–78.

Yang, J., Gao, J., Zhang, Y., Waibel, A., 2001. Toward automatic sign translation. Human Language Technology, 269–274.

Ye, Q., Jiao, J., Huang, J., Yu, H., 2007. Text detection and restoration in natural scene images. J. Vis. Commun. Image Represent. 18, 504–513.

Zhang, J., Chen, X., Hanneman, A., Yang, J., Waibel, A., 2002. A robust approach for recognition of text embedded in natural scenes. Int. Conf. on Pattern Recognition 3, 204–207.