1998 International Computer Symposium
Workshop on Image Processing and Character Recognition
December 17-19, 1998, N.C.K.U., Tainan, Taiwan, R.O.C.

PREPROCESSING AND SEGMENTATION OF ADDRESS CHARACTERS ON CHINESE ENVELOPES

Yih-Ming Su[1,2] *and Jhing-Fa Wang*[2,3]

[1]*Department of Electronic Engineering, I-Shou University,*
Kaohsiung County, Taiwan.

[2]*Department of Computer Science and Information Engineering,*
National Cheng Kung University, Tainan, Taiwan.

[3]*Department of Electrical Engineering, National Cheng Kung University,*
Tainan, Taiwan.

ABSTRACT

To cope with the requirements of a real-time mail-sorting system, some empirical methods are proposed in this paper for the preprocessing and segmentation of hand-written or machine-printed Chinese addresses on standard envelopes. The preprocessing includes displacement modification, skew adjustment and binarization of envelope image, and the location of address block. After preprocessing, the address block image is extracted from the envelope image. A set of isolated characters is then extracted from the address block by a proposed segmentation approach based on profile analysis and a curve-mapping function. The proposed segmentation approach contains three steps, they are component analysis, fine segmentation, and component combination. The experimental results show the effectiveness of the proposed methods.

Key words: Preprocessing, Address segmentation, Skew adjustment, and Address block location.

1. INTRODUCTION

A new mail-sorting system based on a real-time recognition of hand-written/machine-printed Chinese postal addresses has been developed for automatic mail handling in Taiwan. Besides, some mail-sorting systems [1-4] illustrate that the preprocessing and segmentation process is more important than recognition process because the former is a critical section to build up a stable and reliable mail-sorting system. Therefore, the goal of this study is to provide a set of isolated characters for address recognition and resolve some issues from illumination and mechanism of the system.

Some researches including the analysis of preprocessing task [5], address layout [6], and address block location [7], have been described to develop more efficient approaches for mail handling. In addition, some related segmentation methods of handwriting Chinese characters [8,9] have been reported. The method [8] consisting of connected component, feature-complexity, and spatial analysis is capable of processing the segmentation of mixed handprint Chinese/English characters, but the neighboring characters are too close to find a correct segmentation point. The other method [9] consisting of stroke bounding, knowledge-base merging and dynamic programming is used to segment handwritten Chinese characters, but the overlapping characters may be incorrectly segmented and there is a consuming time for segmentation process. Based on practical considerations, our methods are emphasized on the speed and effectiveness for a real-time mail-sorting system.

For envelope preprocessing, we develop here a simple and efficient method, which consists of several steps: displacement modification, skew adjustment, translation, binarization, and address block location. Displacement modification step is performed to evaluate the displacement between even and odd frame image captured from an interlaced CCD camera, and uses it to reconstruct a good quality overlapped static image. The skew adjustment and translation step is performed to adjust a slanted image and extract a limited region of the static image, to reduce processing time. Skew adjustment for this image ensures high consistency in processing different captured images, while the translation step enables future extraction of image data to remove the totally black background information. Due to variations in the intensity of illuminating light (connected to AC source), an adaptive thresholding technique is also employed during the binarization of gray level images. Finally, an address block location step is used to extract the region from an envelope image that contains the destination address. In order to achieve this, specific geometric features on a standard Chinese style envelope are first located before an address block is identified.

For address segmentation, a new address segmentation method based on profile analysis and curve-mapping function is broken down into several steps. Firstly, component analysis is used to analyze the address characters and process them by statistic estimation, which calculates character parameters including stroke width, character height, and its variance. Then, three character parameters are used to separate over long character blocks within connected components (i.e., touching/overlapped characters) for fine segmentation step, using a curve-mapping function to partition connected components. Finally, component combination based on the operation of a decision table is used to determine whether adjacent components should be merged or not. Some assumptions are described as follows. (a) Most Chinese characters have than those between radicals of a character.

In the rest of this paper, Section 2 describes a preprocessing task. Section 3 presents a new address segmentation method. In section 4, experimental results and discussions are reported. The concluding remarks for this paper are mentioned in the final section.

2. PREPROCESSING TASKS

2.1 Envelope Detection And Displacement Modification

An envelope to be sorted is first carried by a conveyor belt and pulled up by vacuum onto the mail-sorting machine. The mask of the envelope stimulates a photoelectric detector and an interrupt signal is sent to the personal computer, which then activates the camera to snap the image of the moving envelope, and the image data is stored in the memory of the frame grabber board. The image caught by the camera is composed of even frame and odd frame image, as shown Fig. 1. However, there is a displacement between even frame and odd frame image, caused by an interlaced CCD camera and motion of the envelope. In order to preserve the quality of the envelope image, via the formation of a static image, it is necessary to calculate the displacement between even frame and odd frame image. The concepts of gradient and least-mean-square-error (LMSE) fitting are adopted to evaluate the displacement. Firstly, the vertical boundary points in the even and odd image frame are detected by scanning horizontal direction from right to left and finding maximum difference in gray scale between two adjacent pixels. The LMSE fitting is respectively used to transform boundary points into a line for two image frames. The boundary line in the vertical direction of even and odd image fields are denoted respectively by $x = a_1 + b_1 y$ and $x = a_2 + b_2 y$. The displacement between two image fields is then denoted by the absolute difference between a_1 and a_2. The equation of the boundary line in horizontal direction is denoted by $y = c + dx$. Fig. 1 shows an original envelope image and the detection process of an envelope boundary.

2.2 Envelope Skew Adjustment And Translation

The envelope image shown in Fig. 1 is slanted. Such skew may be caused by many factors including the strength of vacuum pump, the moving speed of conveyor belt, and envelop quality. If the image is slanted, difficulties will arise for subsequent processes such as address block location and address segmentation. Therefore, the skew angle of an envelope should be evaluated in order to adjust a slanted image at this step. The skew angles in vertical and horizontal directions are respectively defined by ${}_H = \tan^{-1}(d)$ and ${}_V = \tan^{-1}((b_1 + b_2)/2)$, where b_1, b_2 and d are given in Equations above. If the difference of two skew angles from horizontal and vertical direction is more than 5 degrees, the processing of the envelope image is abandoned in order to avoid too much distortion in the adjusted image. The intersection of the vertical and horizontal boundary line is at the upper-right hand corner of the envelope. According to this intersection point and the vertical skew angle of the vertical boundary line, the slanted image can be translated and rotated to coordinate (511,0). The overlapped static image after skew adjustment and translation operation is shown in the right half side of Fig. 2.

2.3 Binarization And Address Block Location

An image captured by an interlaced CCD camera is a gray-scale image. For convenience of the further processing, this image should be transformed into a binary image. The histogram of the image pixel, as plotted in left hand part of Fig. 2, shows two peak values existed in the image. These two peak values represented the foreground and background of the gray-scale image. Using the following formulae, a binary image can be obtained.

$$Threshold = \begin{cases} P_x * 0.85, & H_{fp} \geq 150, \\ P_x * 0.81, & 100 \leq H_{fp} < 150, \\ P_x * 0.78, & H_{fp} < 100, \end{cases}$$

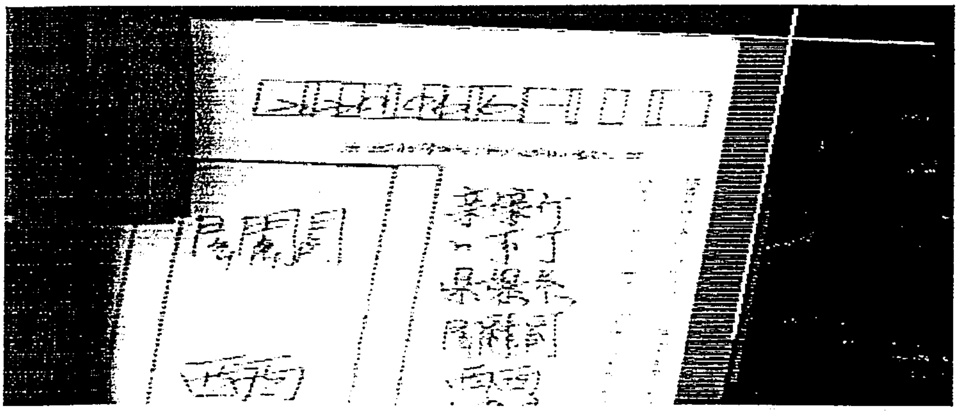


Fig. 1. An example of an original envelope image. The boundary line in the vertical direction: *x=431-0.08y* and in the horizontal line: *y=3.2+0.078x* obtained after the process of boundary detection.

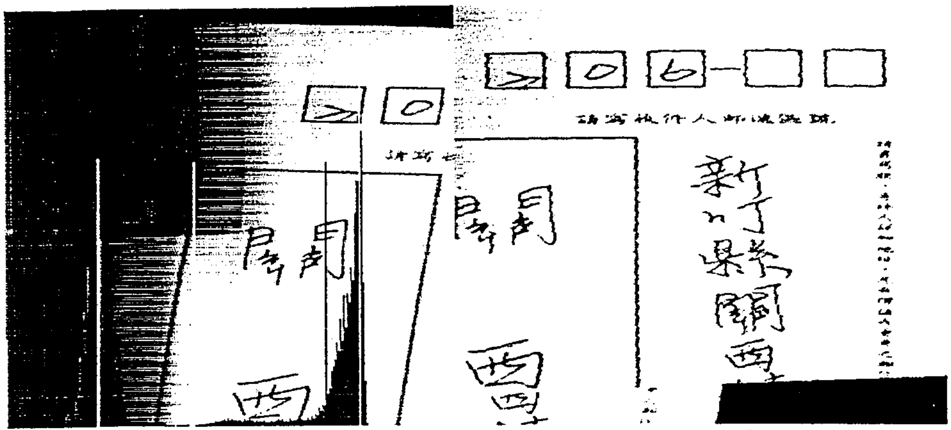


Fig. 2. An example of the modified image (right side) after the process of the skew adjustment and translation.

where Px represents the total pixel of the foreground image and H_{fp} is the peak value of the foreground image. The factors used in each formula are arbitrarily selected according to experimental results for optimum binarization. Since the shutter speed of the CCD camera is very fast, and the intensity of the illuminating light (as connected to AC source) changes continuously, the images caught by the camera will differ in gray levels. This will be true even for the image of the same envelope being captured at different time.

On standard Chinese envelopes, a number of simple geometrical features can assist in the identification of address block on the envelopes. The geometrical features include a bounding rectangle located on the central part of an envelope with sides parallel to the edges of the envelope and five small bounding rectangles on the top right hand corner of the envelope. The name of the consignee is filled within the central bounding rectangle, and the ZIP code filled within the first three of the five small rectangles. In addition, the address of the consignee is filled in the space at the right side of the central rectangle. The address block is identified by the following. (1) The right vertical frame line of the central rectangle is extracted by seeking maximum gradient in the vertical projection of the given bottom block as shown in Fig. 3. (2) The upper horizontal frame line of the central rectangle is extracted by seeking maximum gradient in the horizontal projection of the strip. (3) The upper-right corner of the central rectangle is extracted by calculating the intersection of the vertical and horizontal frame line of the central rectangle. (4) The upper horizontal starting position of address block is located by selecting a small block at the right side of the corner as shown in the top binarized block, on the right side of Fig. 3. The small printed character postal tag can be removed by identifying their properties such as their larger aspect ratio and more narrow width. In the Fig. 3, the address block is enclosed by a rectangle, the two frame lines of the central rectangle are denoted by two lines, and the upper-right corner of the central bounding rectangle is indicated by a spot.

3. ADDRESS SEGMENTATION

After the preprocessing task, the binarized image of the address block is partitioned into separated character blocks by scanning the vertical projection of the address block, in order to resolve two or more parallel character strings. The vertical projection of the address block is composed of simple running count of the black pixels in each column. Moreover; the projection indicates positions of connected strings in the address block. When the character strings are touched or overlapped, the projection often contains less value at the proper segmentation point. The block width of a vertical character string is calculated by counting successive columns with non-zero vertical projection value.

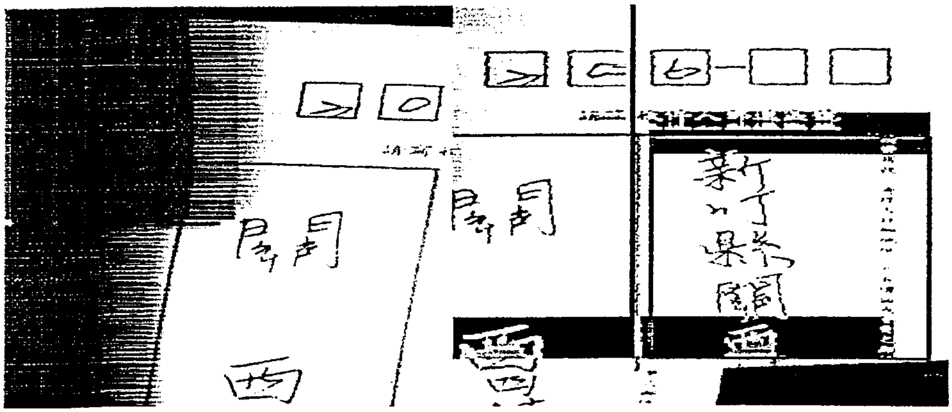


Fig. 3. Location of the address block (enclosed by a rectangle) and the upper-right corner of the central bounding rectangle indicated by a spot.

If the width of a block is too small, the block is then regarded as the small print postal tag on the envelope and discarded.

3.1 Component Analysis

The component analysis offers an efficient operation to obtain character parameters in a vertical character string including stroke width, character height, and its variance. The three parameters are used as a referred basis for fine segmentation and component combination. The more accurate estimation for the three character parameters is, the better performance for segmentation process will be.

The estimation of stroke-width (*SW*) is defined as

$$SW = \max(W_1, W_2, ..., W_i), \quad (3.1)$$

where W_i indicates the *i*th thickness value of stroke in pixels, through scanning from white-to-black transition and then from black-to-white transition at different positions of the horizontal projection of the character string. The stroke width is used as a criterion to partition a character string into character components during fine segmentation. The estimation of character width (*CW*) is defined as

$$CW = averag\!e(P_1, P_2, ..., P_i), \quad (3.2)$$

where P_i indicates the projection value located the maximum projection value in the *i*th section of horizontal projection. The horizontal projection of the character string consists of a simple running count of the black pixels in each row. The character width is used as a criterion to determine whether a connected component is too long, because the height of a character is proportional to its width. The estimation of the character height (*CH*) is defined as

$$CH = average(H_1, H_2), \quad (3.3)$$

where H_i indicates the height of the *i*th group. The collected data from the height of components (i.e., characters or radicals) is clustered into two groups by the K_means algorithm [10]. The character height is used as a criterion for fine segmentation and component combination. Finally, the variance of *CH* is calculated by statistic variance. Although these estimated parameters are not quite accurate, they are essential for fine segmentation and component combination.

3.2 Fine Segmentation

This step based on a curve-mapping function is performed to segment connected components. In operation of the curve-mapping function, there are two observations as follow. (1) The possible position of a segmentation point in the connected components is located at a less projection value, due to ligatures existed. (2) The possible position of a segmentation point is near the position corresponding to character height (*CH*).

A block of a character string is scanned to detect successive black pixels in the horizontal projection. If a black block is too long, it has to be further segmented into components by the curve-mapping function. As the block ranges from maximum height (*MaxH*) to minimum height (*MinH*), a segmentation point is normally located near the position corresponding to character height (*CH*). The Estimation of the *MaxH* and *MinH* can be obtained by:

$$\begin{aligned} MaxH &= CH + HV, \\ MinH &= CH - HV. \end{aligned} \quad (3.4)$$

The operation of the curve-mapping function is represented in Fig. 4 and described as follows. The horizontal projection is scanned in the range from *MaxH* to *MinH* to detect and the less projection values are detected by less than stroke width (*SW*). Then, the position of less projection is regarded as a cutoff point (i.e., a candidate of segmentation point). The successively cutoff points are grouped together, and a represented point is selected. Each represented point is then mapped to the triangular curve as shown in Fig. 4(d). The horizontal axis of the curve represents distance. The vertical axis of the curve represents a confidence degree (*CD*), which ranges from 0 to 1. A point with higher *CD* indicates a higher priority for being a segmentation point. *CD* of the curve-mapping function is defined by

$$CD(index) = \begin{cases} (index - MinH)/(CH - MinH) & \text{if}\,(index < CH), \\ (MaxH - index)/(MaxH - CH) & \text{if}\,(index \geq CH), \\ 0 \quad \text{otherwise}, \end{cases}$$

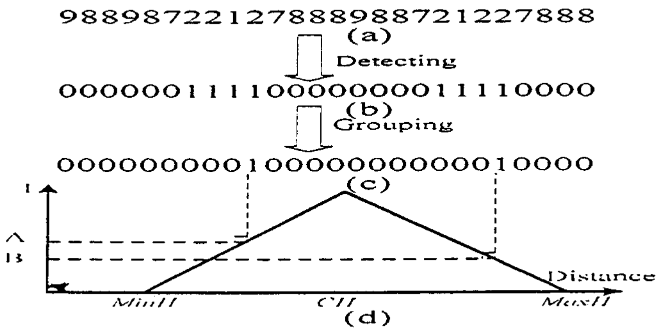


Fig. 4. (a) The projection part of the connected components. (b) The cutoff points are denoted by "1". (c) The represented points are denoted by "1". (d) The triangular curve and the segmentation point located in the mapping of the position 'A'.

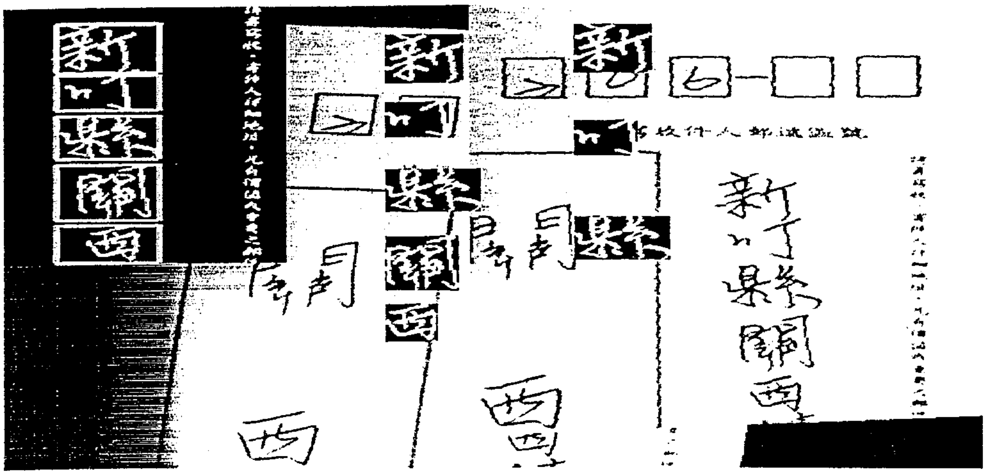


Fig. 5. An example of the extracted individual characters after address segmentation.

where *index* is denoted by the position of the represented points. For example, the projection value of the connected component is shown in Fig. 4(a). The cutoff points are represented by "1" in Fig. 4(b). The represented points are denoted by "1" in Fig. 4(c). The triangular curve is plotted as Fig. 4(d), in which the mapping value in position 'A' is greater than that in position 'B'. Therefore, position 'A' is regarded as the position of the segmentation point of the connected component.

3.3 Component Combination

The process of fine segmentation should be employed to segment the large connected components, to avoid existence of connected characters. However, some characters may be separated into two adjacent components in the process, and require combinations later. Component combination step is based on a decision table, which determines if two adjacent components should be merged or not. The decision table (Table 1) contains two variables and follows a set of rule. The two variables include character height (*CH*) and character gap (*CG*) between characters. The *CG* is defined as $CG = average(G_1, G_2)$, where G_i indicates the gap of the *i*th group. A set of collected data from the gap between two adjacent components is clustered into two groups by K-means algorithm. The character gap is used to check with Table 1, to decide for combination. Table 1 shows the different combinations for *CH* and *CG*, with "M" represents recommendation to merge and "NM" not to merge. The accompanied set of rules is:

If $(MH \le 0.8CH)$ and $(Gap \le 1.0CG)$ then do merging operation;

If $(MH \le 2.0CH)$ and $(Gap \le 0.5CG)$ then do merging operation;

If $(0.8CH < MH \le 1.4CH)$ and $(0.5CG < Gap \le 1.0CG)$ then do merging operation;

Otherwise, do not merge;

where *MH* indicates the merged height of two adjacent components, and *Gap* indicates the gap between two adjacent components. Fig. 5 shows an example for extracting individual characters after address segmentation. If the gap between adjacent characters is obvious, then characters can be segmented very accurately.

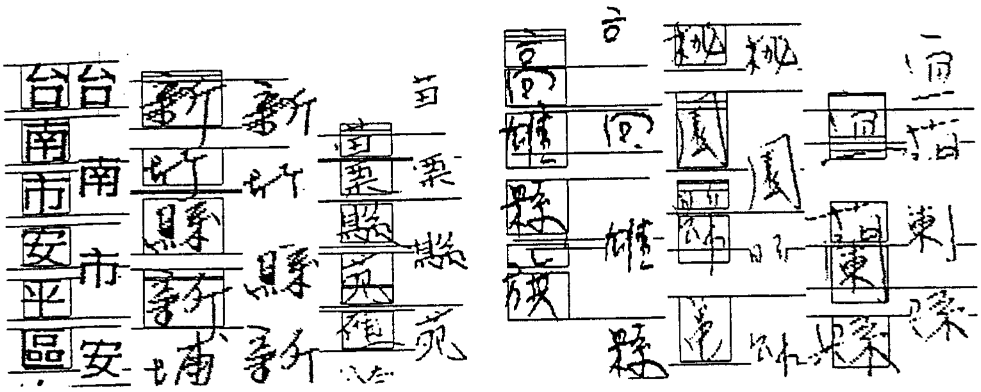
Table 1. The decision table of component combination

	0.5*CG*	1.0*CG*	1.5*CG*
0.8*CH*	M	M	NM
1.4*CH*	M	M	NM
2.0*CH*	M	NM	NM

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we analyzed the experimental results of the preprocessing and address segmentation for 1468 handwritten/machine-printed envelopes extracted from real mail pieces. The experimental results show only 0.3% of failure rate in the preprocessing stage and 9.5% of failure rate in the address segmentation stage. Examples of the successful and failure cases in address segmentation are given in Fig. 6a and 6b respectively, where the left part of each image string indicates the original address image and the right part is the segmented address.

There are several causes of failure. In the preprocessing stage, poor contrast caused by poor envelope quality results in the extraction of incomplete geometric features. The instability of camera's sensitivity to light changes results in undesirable changes in the gray level. In the address segmentation stage, stroke features of scripted characters are very clear which make the extraction of characters incomplete. When the address writing style is irregular, it will cause extraction of broken characters. The inherent segmentation ambiguity in the address characters will result in the selection of incorrect segmentation points.



(a) (b)

Figure 6: (a) successful cases and (b) failure cases.

5. CONCLUSIONS

In this paper, some empirical methods for preprocessing and segmentation have been described and implemented in a real-time mail-sorting system to process handwritten/machine-printed Chinese envelopes. Simple and efficient approaches, such as profile analysis and curve-mapping function are adopted to improve both the speed and effectiveness of the system. The experimental results obtained from a real-time operation show that the proposed methods are capable of coping with system requirements. The main contributions of this study are as follows. Firstly, a computer vision approach is applied to handle a large amount of standard Chinese mails. Secondly, a new segmentation technique based on profile analysis and curve-mapping function is efficiently used to extract a set of isolated characters from an address block. Finally, a preprocessing task is performed to resolve some issues caused by illumination, sensor sensitivity, and mechanism.

6. ACKNOWLEDGEMENTS

The authors sincerely acknowledge financial support for this study from the Ministry of Transportation and Communications and the envelope database provided by the Postal Pavilion.

7. REFERENCES

[1] S.N. Srihari, "High-performance Reading Machines," Proc. of the IEEE, vol. 80, No. 7, pp. 1120-1132, July 1992.

[2] V. Govindaraju, A. Shekhawat and S.N. Srihari, "Interpretation of Handwritten Addresses in US Mailstream," The Third Inter. Workshop on Frontiers in Handwriting Recognition, pp. 197-206, 1993.

[3] P.W. Palumbo and S. N. Srihari, "Postal Address Reading in Real Time," Inter. J. of Imaging Science and Technology, 1996.

[4] E. Cohen, J. J. Hull, S. N. Srihari, "Understanding Handwritten Text in a Structured Environment: Determining ZIP Codes from Addresses," Inter. J. of Pattern Recognition and Artificial Intelligence, Vol.5, No. 1&2, pp. 221-264, 1991.

[5] A.P. Whichello and Hong Yan, "Fast Location of Address Block and Postcodes in Mail-piece Images," Pattern Recognition Letters, Vol. 17, pp. 1199-1214, 1996.

[6] N. Nakajima, T. Tsuchiya, T. Kamimura, K. Yamada, "Analysis of Address Layout on Japanese Handwritten Mail- A Hierarchical Process of Hypothesis Verification," Proc. of the 13th Inter. Conf. on Pattern Recognition, pp. 726-731, 1996.

[7] S.W. Lee and K.C. Kim, "Address Block Location on Handwritten Korean Envelopes by The Merging and Splitting Method," Pattern Recognition, Vol. 27, No. 13, pp. 1641-16511, 1994.

[8] H. D. Kuo and J.F. Wang, " A New Method for the Segmentation of Mixed Handprinted Chinese/English Characters," Proc. ICDAR, pp. 810-813, 1993.

[9] L. Y. Tseng and R.C. Chen, "A New Method for Segmenting Handwritten Chinese Characters," Proc. ICDAR, pp. 568-571, 1997.

[10] J. T. Ton and R. C. Gonzalez, " Pattern Recognition Principles," Massachusetts, Addison-Wesley Publishing Co. 1974.