Robust Segmentation for Automatic Data Extraction from Passport

Mya Mya Thinn, Myint Myint Sein †
Mandalay Technological University, Mandalay,
† University of Computer Studies, Yangon.
ngenge17@gmail.com, chuchu0218@gmail.com

Abstract

The objective of this paper is to develop an automatic data entry system of passport for security system. A passport contains the important personal information of holder such as photo, name, date of birth and place, nationality, date of issue, date of expiry, authority and so on. The proposed system is provided to prevent mistakes of writing in personal information and to obtain the true information from the passport in short time. A new segmentation process applied to unconstrained handwritten character. The novelty of the proposed algorithmis based on the combination of two types of structural features in order to provide the best segmentation path between connected entities. This method was developed to be applied in a segmentation-based recognition system. The handwritten character recognition algorithm is extended by Gaussian elimination method. The handwritten character A to Z and digits 0 to 9 are convert to the printed form. Once the passport is scanned, the proposed system is executed automatically the data extraction, converting to printed form and entry the data and facts to worksheet. Pre-limitary investigation is performed to confirm the effectiveness of the proposed approach.

Keywords: Automatic Data Entry System, Character Recognition, Gaussian elimination method, data extraction

1. Introduction

The term biometrics refers to the science of measuring identifying features or attributes of human beings. We distinguish two approaches: passive and active schemes. Example of passive biometric is face recognition. Handwritten recognition is one of an example of the behavioral biometrics. By using biometric data, officials can determine whether someone making a new passport has ever been issued a passport under another identity. A typical identity certification such as a driver's licenses, passport, or visa, consists of a personal portrait photo, an arbitrary message, and one or more feature whose purpose is to guarantee authenticity. To achieve the security Criterion of non-transferability, most documents contain a photograph and an image of the handwritten signature of the legitimate holder and some cases some information on the holder's appearance, like date of birth, eye color, nationality.

We describe a system to recognize characters such as name, date of birth, Nationality, passport number, date of issue, place of issue, date of expiry, etc. In this paper, automatic handwriting recognition and data entry system of passport is proposed. Stefan Hellkvist [1] studied the system designed for recognizing the single characters that are written on either a touch pad or a touch screen. The elastic matching method is proposed for recognition. *Fadi Biadsy* et al. [2] was presented the algorithm of Handwriting recognition of Arabic script. Hidden Markov Model (HMM) is applied for most of the difficulties inherent in recognizing Arabic script including: letter connectivity, position-dependent letter shaping, and delayed strokes.

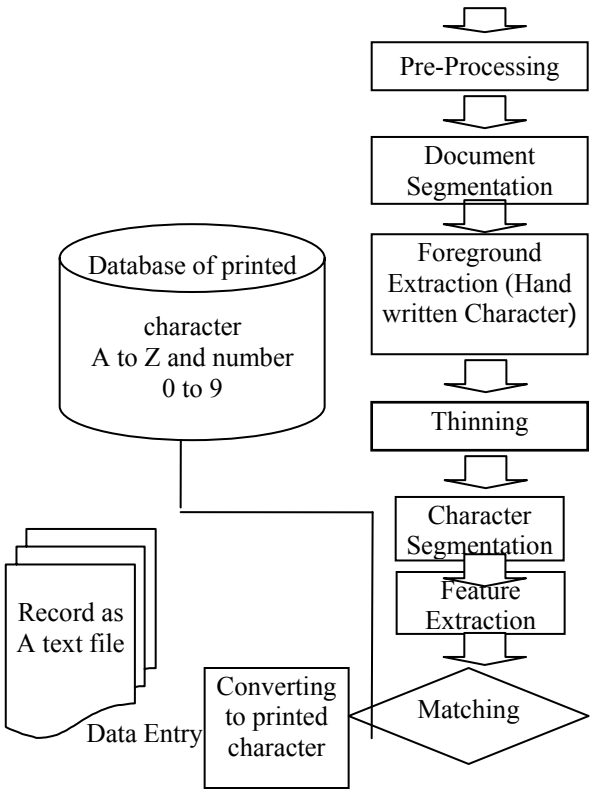


Figure. 1. Block diagram of the proposed system

In our proposed system, the handwritten character recognition algorithm is developed based on the Gaussian elimination approach. Skewing, resizing, filtering, gray scale converting, normalization step are done in Pre-processing steps. Foreground extraction is performed to extract the handwritten characters in the passport. For word and character segmentation, region-based segmentation method is used. Thinning and skeleton methods are extended for good feature extraction. The connections of the neighborhood pixels have been detected to convert the handwritten character A to Z and digits 0 to 9 are to the printed form. Recognizing which is included matching with the database information. Finally, data is automatically entered to worksheet. The block diagram of the proposed system is shown in figure (1).

2. Preprocessing

2.1. Skewing

In pre-processing step, skewing angle which may be appeared due to the user scanning is performed. Each document is moved from an automatic feeder into a scanner and angle of skew is sometimes introduced. So that the input image is tested for skewed angle detection algorithm has been developed using the equation.

$$P = x \cos \alpha + y \sin \alpha$$

2.2. Resizing

The size of input image is rather so that it may cause the performance of the system slow down and sometimes may cause the system hang. Therefore the image is resized. Resizing is to reduce the processing time and to convenience for thinning method. Noise filtering step are included in the pre-processing step.

The process of scanning is not exact. Even bi-level documents that are scanned, when digitized as grayscale or color, show more then two levels of darkness. This may be due to variances in the painting or reflectance of the document, or imprecision in the scanner itself. Because of factors such as these, it is difficult to pick a good threshold. The extraction of the foreground textual matter by removing such as textured background, salt and pepper noise and interfering strokes. The effect of setting the threshold too low is the presence of spurious black pixels throughout the image, often known as *speckle* noise. This noise should be eliminated from the image as much as possible so as not to cause confusion recognition.

3. Segmentation and Extraction

3.1. Document Segmentation

To extract the personalization data from a passport, words and documents segmentation process are performed by the mean positions of training sets. Figure (3) shows the segmented document regions of hand written characters sets. Mean position of each word or document can be derived from the following relation:

$$\text{Mean Registered Position} = \frac{\text{Total of register position}}{\text{Number of test images}}$$

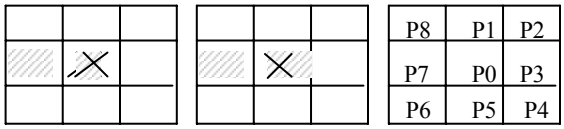
Fourteen documents are extracted from a passport image separately. Results of registered positions are depend on the types of scanner and scanned position. Some experimental results are shown in table1.

Table. 1

Types of scanner	Scanned Position	Scanned Rate
Same	**fix**	**97.97%**
Different	**anywhere**	**90.33%**

3.2. Thinning

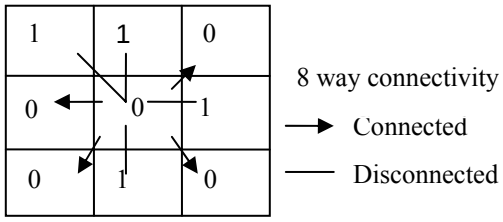
Thinning algorithms are iterative algorithms that is pixels at 0 ⟶ 1 transitions in a binary image. Connectivity is an important property that must be preserved in the thinned object. Border pixels are deleted in such a way that object connectivity is maintained. Some thinning parts are illustrated in figure (4).



P8	P1	P2
P7	P0	P3
P6	P5	P4

3.3. Character Segmentation

Image segmentation is one of the most important steps leading to the analysis of processed image. Segmentation methods can be divided into three groups, these are global knowledge about, edge based segmentation and region based segmentation. Region based segmentation method is easy to construct regions from the borders, and it is easy to detect borders of photo from the passport existing regions. Individual region of the character are estimated segmenting the extracted region of the image.
The individual words and characters are extracted using the neighborhood pixel connection approach. Translation of the pixel intensity is considered vertically. Then the characters are segmented using the recursive algorithm.



A connected component is defined as a set of black pixels where each pixel is a direct neighbor of at least one other black pixel in the component. The pixel connectivity to split or extract each line and each digit from input document. Value '0' refers to black pixel and '1' refers to white pixel.

3.4. Feature Extraction

After data segmentation is performed, foreground words are extracted from the segmented parts. The foreground is obtained by eliminating the background based on the color range. RGB color segmentation approach is applied to get the word and character written by black color. Convert the gray scale and the extract the black pixel. Figure (5) shows a segmented part and its extracted feature points, respectively.
i. e. For any image,

If the R, G, B color values of a pixel are less than or equal 100 then each color component value is set to 0. And, the R, G, B color values of a pixel are greater than 100, then each color component value is set to 255. Then, convert the gray scale and the extract the black pixel. Figure (5) shows a segmented part and its extracted feature points, respectively.

4. Transforming Printed Character

To convert the handwritten charter to standard printed character, the pattern matching is required between two patterns. That can be considered as the off-line character images. Then the corresponding regions of each character pattern are detected by the Gaussian elimination method. Let ***I*** and ***J*** be two images, containing *m* features *Ii*(*i*=1….*m*) and *n* features *Jj* (*j*=1….*n*),respectively.

Firstly, a proximity matrix ***G*** is build by the two sets of feature points and computed the Gaussain weighted distance between two features *Ii* and *Jj*

$$G_{ij} = e^{-r_{ij}^2 / 2\sigma^2} \quad i=1...m, j=.1\ldots,n \quad (1)$$

where $r_{ij} = \|I_i - J_j\|$ is their Euclidean distance if we regard them as lying on the same planes.And, decompose the matrix into the multiple of orthogonal matrices *T, U* and diagonal matrix *D*.

$$[TDU^T] = svd(G) \quad (2)$$

Its diagonal elements D_{ii} in descending numericalorder. Then, new matrix *E can be* obtained by replacing every Diagonal element D_{ii} with ***1*** and then computes the product.

$$P = TEU^T \quad (3)$$

Rogue points cause lots of ambiguous, equally good matching possibilities *I* the space of paring *s*, and the sole proximity used to build G in Equation (1) does not have enough "character" to discriminate amongst them.

If we represent two W* W areas centered on features Ii and Jj as two W*W arrays of pixel intensities A and B, respectively, the normalized correlation is defined as

$$C_{ij} = \frac{\sum_{u=1}^{w}\sum_{y=1}^{w}(A_{uv} - \overline{A}).(B_{uv} - \overline{B})}{W^2.\sigma(A).\sigma(B)} \quad (4)$$

where $\overline{A}(\overline{B})$ is the average and σ(A)(σ(B)) the standard deviation of all the elements of *A* (*B*) C_{ij} varies from -1 for completely uncorrelated patches to 1 for identical patches.

One way of including this correlation information into the proximity matrix is to transform the elements *G* as follows:

$$G_{ij} = \frac{(C_{ij} + 1)}{2} e^{-r_{ij}^2 / 2\sigma^2} \quad (5)$$

Figure (6) show the matching and translation pattern between two character patterns, respectively.

5. Experimental Results

We first created a large database which includes classified printed characters. Using this database, we have conducted several experiments to access the performance under know variations of lighting, scale and orientation. There are fifty Myanmar passports are used in this experiment. .Region based segmentation and color segmentation algorithm are developed for words characters segmentation and extraction. Transformed printed characters are matched with the A to Z characters and 0 to 9 digits into the database. If they are same, they will convert to text file by using automatic data entry system. The extracted characters are recognized with high data accuracy rate.

Table. 2.

Input	Recognition Rate	Error rate
Digits	99.2%	0.8%
Character	95.6%	4.4%

Figure. 2. Corrected skewing angle and resized image

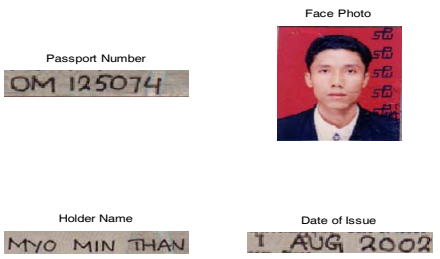


Figure. 3. Segmented Documents Data

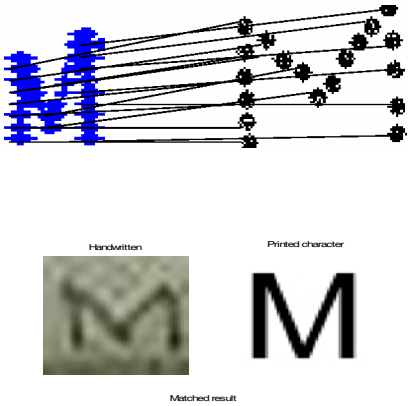


Figure. 6. Transforming of Printed character patterns

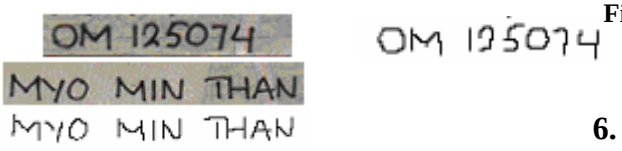


Figure. 4. Results of Thinning Parts

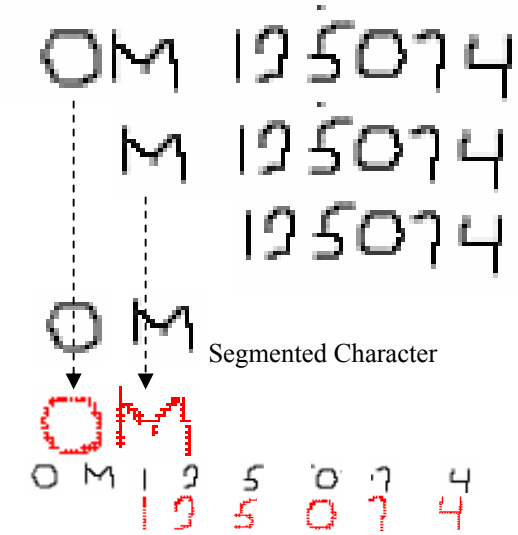


Figure. 5. Extracted digits and characters

6. Conclusion and Future Work

We have proposed a system of automatic data entry of passport and a handwritten character segmentation system to register the passport's holder data from the diplomatic passport. The personal information of passport holder described as a text file. In future work, personalization will perform by extending the face and signature recognition system. The personal identification can do from every where thought over the network connecting to database. We've tried our best to obtain the system with less error and more accuracy recognition.

References

[1] Stefan Hellkvist, (1999)," On-line character recognition on small hand-held terminals using elastic matching", Royal Institute of Technology, Department of Numerical Analysis and Computing Science

[2] Biadsy, Jihad El-Sana and Nizar Habash "Online Arabic Handwriting Recognition Using Hidden Markov Models" Columbia University Department of Computer Science, New York

[3] Q. Tian, P.Zhang, T.Alexander Y. Kim. Survey: "Omnifont Printed Character Recognition", *Visual Communication and Image Processing 91: Image Processing*, pp 260-268, 1991.

[4] Z .ShiandV. Govindaraju. "Segmentation and recognition of connected handwritten numeralstrings. In Progress in Handwriting

Recognition", *World Scientific*, pp. 515-518, 1996.

[5]. Dr.A.Zidouri, K.Fand, "Automatic Text Recognition: A need in Arabization" URL: Electrical Engineering Department, University of Petroleum & Minerals, Kingdom of Saudi Arabia.

[6] M.M.Thinn and M.M.Sein, "Handwritten Recognition System for Automatic Data Entry of Passport", *Proceedings of the ACRS organizing committee*, pp 240-241, Nov 2007.