Verification of the Document Components from Dual Extraction of MRTD Information

Young-Bin Kwon and Jeong-Hoon Kim

Departmen Computer Science and Engineering, Chung-Ang University,
221 Heuksukdong, Dongjakku, Seoul, Korea
ybkwon@cau.ac.kr, jhkim@cvlab.cau.ac.kr

Abstract. This paper proposes a method for character region extraction and picture separation in a passport by adopting a preprocessing phase for the passport recognition system. Character regions required for the recognition make black pixel and remainder of the passport regions makes white pixel in the detected character spaces. This method uses the MRZ sub-region in order to automatically calculate the threshold value of the binary image which is also applied to the other character regions. This method also applies horizontal/ vertical histogram projection in order to remove the picture region from the binary image. After region detection of picture area, the image part of the passport is stored in the database of face images. The remainder of the passport is composed of character part. Recognition on the extracted character is performed on the various different passports. From the obtained information, auto-correlation of extracted characters within a given passport are accomplished after character recognition. A cross-check process of MRZ information and field information of passport on similarity is implemented. For this purpose, this paper uses the auto-correlation between the binarization method based on the color information and an extracted image from a passport to propose a characteristic extraction method to prevent passport forgery.

Keywords: passport, verification, MRTD(Machine Readable Travel Document), cross-check, auto-correlation.

1 Introduction

Keys used for the personal identification method, which is generally used in the airport, is the passport number and personal records including picture and character information. The number of passport issued by each country is increasing. With the increase in the number of people travelling, the number of passports processed by the immigration is also increased. Immigration control is done to control immigration, find forged passports, emigration and immigration forbidder, wanted criminal, and emigration and immigration ineligibility persons such as alien. An efficient and accurate passport recognition system is required for the immigration control judgment. Most passport recognition system only processes the MRZ (Machine Readable Zone) code and the picture on the passport. MRZ code refers to the recognition code used to express the information on the passport using 44 OCR-B

W. Liu, J. Lladós, and J.-M. Ogier (Eds.): GREC 2007, LNCS 5046, pp. 235–244, 2008.

font characters per line for passport recognition. By comparing the data on the passport with the information on the MRZ code, we can improve the forged passport detection rate through the recognition and verification method.

The existing methods only extracted the MRZ code and picture from the passport for the recognition [1, 2]. These methods are not able to utilize the other information on the passport and are vulnerable to passport forgery if the MRZ code is altered with. In this paper, we propose a method that extracts the printed personal information as well as the MRZ code to cross-check the information for a more thorough recognition. The use of the MRZ proves effective in passport processing at the immigration. However, the simple structure of the MRZ is vulnerable to forgery and the auto-recognition process makes it even easier to pass-by with forged passports. Checking for the validity of a passport is limited with the MRZ auto-recognition method. In this paper, the redundant information appearing on both the MRZ area and the printed personal information area are extracted for comparison to check for the validity of the passport.

Extent recognition system does not require any special binarization method [5] because it is only used for the MRZ code recognition. However, a binarization method that separates the characters in the passport from the background is required to for the recognition of the data on the passport. This paper proposes a binarization method which uses the character's RGB property and histogram of the MRZ region. Extraction of picture region is proposed by a method which executes horizontal/vertical histogram projection and analyzes the result value. After extraction of each character, a recognition method based on template matching and feature comparison is implemented. A self-checking of the traced characters samples is performed in order to compare the similarity of extracted characters. This method is simple and easy to compare all recognized characters within a same passport. If the same characters are not found in the same passport, we utilize the standard OCR-B font to compare the similarity. A cross-check for the same field information from MRZ and upper personal information area is also performed.

2 Passport Data Extraction

2.1 Passport Data Extraction Process

In order to extract the data part from passport, it is necessary to divide the regions into the picture area, the MRZ area, and the upper printed personal information area. The MRZ area is easily detected through the horizontal projection method because they are composed of two lines at the bottom of the passport and each line contains 44 characters. During the projection, threshold value of 180 on each RGB component is used to detect the gray scale value of printed characters. The separation of the picture area and the upper printed personal information area is accomplished by detecting the boundary line between two regions. As shown in Fig. 2, a vertical projection of the passport except the MRZ area can detect the separation line. There is no passport which uses more than half of its area for the picture area. Thus, only the left half part of the passport is used for vertical projection. To decide the line boundary, the right

Fig. 1. Detection of MRZ area

Fig. 2. Detection of Picture & Personal Information areas

most point of the zero values is selected. If no zero value is found after projection, we may adjust the threshold value to create the zero values for the boundary.

After detection of each area, the picture area is saved as an image. Binarization is only performed on the MRZ area and the upper personal information area. Passports with hologram backgrounds are more difficult to process. However, the knowledge of the printed characters gray value information which has same value when they are issued helps us to detect the printed value correctly. Character recognition based on template matching using OCR-B font and feature comparison is performed after binarization. The location information on each character filed is already standardized by ICAO helps us to improve the recognition ratio for the numbers and alphabets [6, 7]. After recognition of printed characters, cross-check verification on each region is also performed in order to detect the possible fraud passport. But some countries use different fonts between two areas. Thus, a development of a self-correlation method which calculates a correlation between extracted characters is needed. We call it extracted auto-correlation character analysis.

2.2 Colour Image Binarization

The photo area is stored separately for the face recognition and the information from the MRZ and personal information area is each extracted. The image is binarized to distinguish the character part from the background. Most documents are black and white or have a simple background color which makes it easier to perform the binarization using a fixed threshold value and a method such as the Trier, Text [4] or iterative binarization method [5]. However, passports usually use various colors for the background with various features. Some passports even use holograms for the background which makes the binarization even more important in the process of accurately extracting the passport information.

The MRZ area of the passport has to be checked by binarization in order to extract the passport information. In previous studies [1], the average value of the background and character pixels was calculated from the initial threshold value for the binarization. In other words, the initial threshold value is set by dividing the sum of all the RGB values by the number of pixels, before comparing it against the current threshold value. If this number is bigger than the threshold value, the RGB value is added to the object. If the value is smaller, the process of adding the object and black and dividing it by 2 is repeatedly applied to get the value of R, G and B before the binarization. The threshold value for the Red, Green and Blue were acquired for binarization. The value is put through the median filter to get the final result. The background of the personal information area is more complex than the MRZ area which makes it difficult to get good results by simply applying binarization. The RGB elements of most printed letters on the MRZ and personal information area are similar in most cases. In the personal information area, an improved binarization method of laying over is used which consists of two stages. In the first stage, the RGB value of the original image acquired using the above method is extracted and stored as the threshold value. In the second stage, the RGB element, which is the RGB value, is acquired. The RGB element of the pixels in the personal information area is compared against the RGB value of the threshold value. After acquiring the R:G:B ration of the personal information area, this value is compared with the R:G:B ratio of the MRZ area to add or subtract the range value. After this process, the same binarization method applied to the MRZ area is applied repeatedly. This method has an advantage of more clearly expressing the contrast for a more thorough binarization result.

2.3 Extracted Character Analysis by Automata

The MRZ information stores the information based on the information of the location of each field. MRZ has a certain pattern with the '<' as the identifier. The result of template matching using the '<' identifier is used to design the automata for the segmentation in each field. The starting point of each field is decided by the analysis of each MRZ line and the information is extracted using the automata at each location. The extracted information is stored along with its number of characters for cross-checking with the personal information, as shown in Fig. 3.

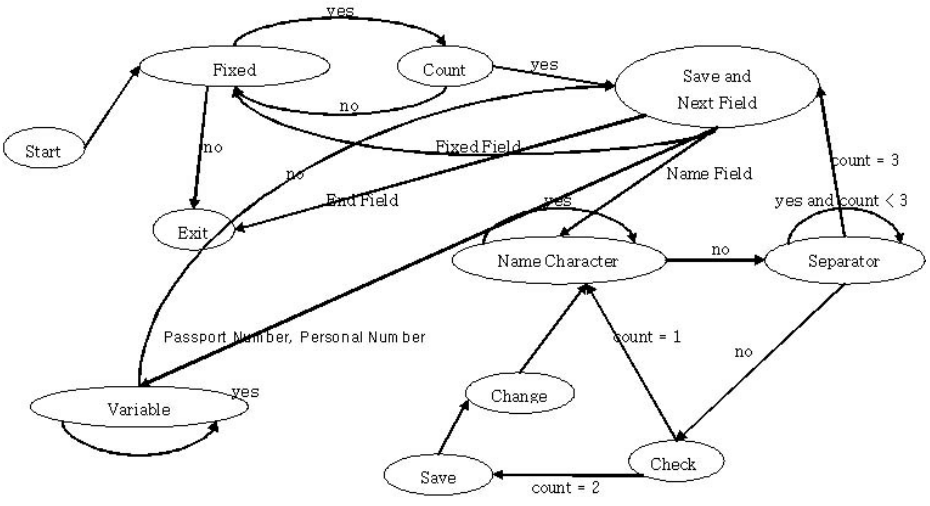

Fig. 3. Automata for the Segmentation of MRZ

2.4 Character Extraction and Auto-Correlation Analysis by Alphabet

The field can easily be divided in the MRZ area based on the regular order and the identifier, '<', but there is a few more steps to process it in the personal information area. Each extracted character has to be expressed as a set of characters for the information extraction, as shown in Fig. 4 and characters either too small or too big are filtered.

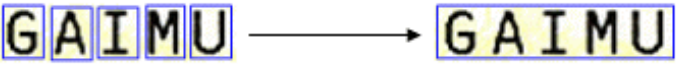

Fig. 4. Character set

The personal information area is regrouped into a few sets after this process. Each set is recognized using the method shown in Figure 4 and this result is string-matched against the field extracted form the MRZ to recognize the set most similar to the MRZ. A different algorithm has to be applied for the field matching of the date of birth and date of expiry because of the different format used to express dates by personal information area and the MRZ area. The date on the MRZ area is expressed using 6 digits whereas the expression used in the personal information area is different in various formats. Fig. 5 shows a few examples.

Although there are many different formats used to express the date in the personal information area, the order of appearance, from day to month to year, is the same as in the MRZ area. The matching algorithm can be produced using this pattern. The year is first located before locating the day field after which the month field can easily be

Fig. 5. Examples of Data Expression

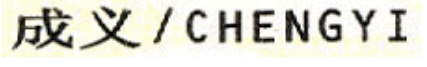

Fig. 6. '/' Identifier Example

found by finding the number in-between. This algorithm is applied differently by the expression of the year. The year data is first extracted from the MRZ area, and if the first digit is a 0 or 1, the digit 20 is added on to express the years after the year 2000, and 19 is added otherwise. To express the years after 1900, the same set of digits is looked up for in the personal information area, and the same digits used to express the day in the MRZ area is looked up for to find the day. Finally, if the height of the day and year area is the same, the set between the two is set as the month. The same algorithm is applied to the personal information area where the year is expressed with only two digits. Finally, a few exceptions handling process is required to divide the field within the personal information area. Chinese passports divide the English and Chinese name using the '/' identifier as shown in Fig. 6. The '/' identifier is used to divide the Chinese and English area into two different sets.

A different algorithm is required to divide the information in the personal information area. If the value of each set is smaller than the height of each character multiplied by 1.5, the two characters are recognized to be a single set of characters. The extracted MRZ and personal information can be used for cross-checking since they contain the same information.

A single character from 0~9 and 'A' to 'Z', extracted from the passport is used for the comparison. For example, if you chose 'A', both 'A' from the MRZ and upper personal information area, is printed on the screen. The result of match comparing the 'A' in the MRZ and the upper personal information area is then printed. The average shape of the 'A' in the MRZ and the upper personal information area are compared for the difference in their font shape. The shape that is most irrelevant from the others can be displayed. The matching score for each character is acquired through dividing the compared score by n. The comparison with the extracted characters becomes possible using the auto-correlation. The type and number of characters extracted from the information on each passport is different due to the different names and nationalities. The information on the MRZ and the personal information area share at least two same alphabets since they contain the same information. Therefore, it is possible to

look for the cross-characteristic based on this fact. The relation between the two areas is very high since the information on a passport is printed at the same time. Various auto-correlation can be found from the result for result analysis.

3 Results

Character extraction experiment is performed using 50 different passport images. The passports used for the experiment are from 13 Asian/Oceania countries, 5 European countries, 2 American countries and 2 African countries. Fig. 7 shows the improved result of the binarization method applied in this paper.

The image on the left shows the result of applying the iterative binarization method on the image and the image on the right shows the result of applying the improved binarization method explained in section 2.2.. The E and K on the name field and the E on the date of expiry are expressed beyond recognition on the left side image whereas they are expressed properly on the right side image. Field extraction result demonstrates perfect extraction results. The recognition of characters using template matching and feature comparison is implemented. The recognition rate of character is 99.9% and personal information extraction is 99.8% with an average speed of 2.3 second required for the recognition of the whole page. Fig. 8 shows the result of the cross-checking.

Fig. 9 and Fig. 10 show the result of single character analysis. Each figure illustrates the extracted data from each region and the rate of auto-correlation between extracted characters.

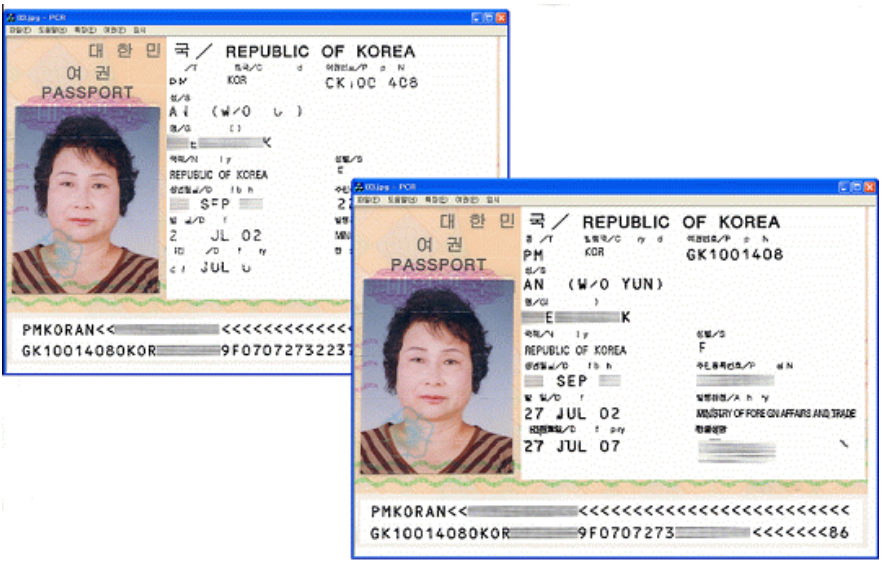

Fig. 7. Iterative Binarization (left) and Proposed Colour Binarization

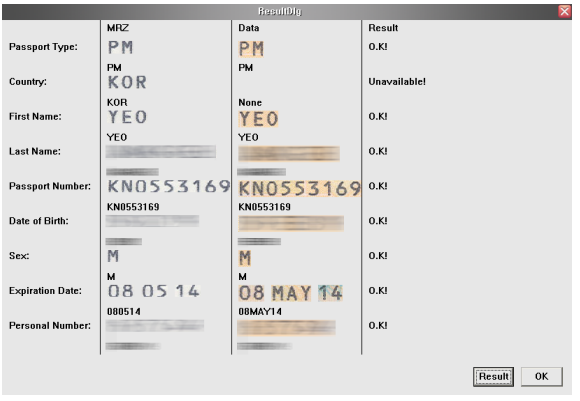

Fig. 8. Result of Cross Check

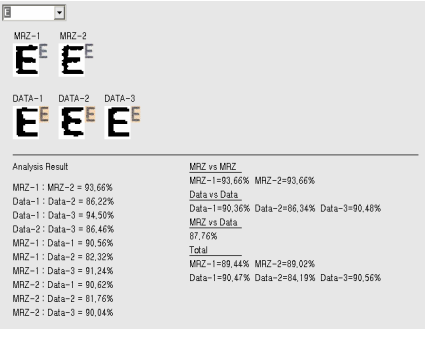

Fig. 9. Analysis Result of 'E'

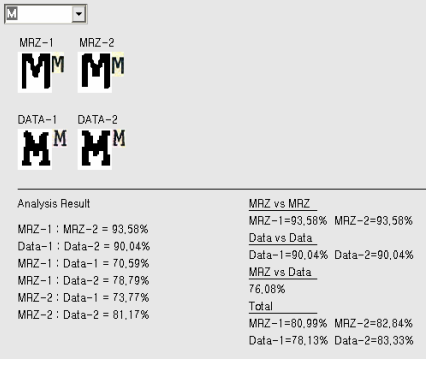

Fig. 10. Analysis Result of 'M'

In the result of Fig. 9, we can see that the similarity is higher when comparing the same area type than comparing the MRZ from the Personal Information area. The second 'E' on the Personal Information area, DATA-2, shows 84.19% similarity level which means that it is the most heterogeneous one.

Fig. 10 shows the result of analyzing the character 'M' on a European passport. The similarity level is at a low level of 76% between the MRZ and Personal Information area. This means that the British passport uses a different font for the MRZ and Personal Information area. We can construct a knowledge database of the passports for the date of issue and the country based on the detected information. Forgery detection may be performed using this knowledge base.

4 Conclusion

In this paper, we have proposed a method for extracting and utilizing the information on the MRZ area and personal information area through the structure analysis which showed a high recognition rate through experiments. This method also applies an automata for the sorting of the MRZ information into different fields for a more effective classification. The concept of cross-checking can be used to check for fraud information using redundant information on the passport for more accuracy. This will contribute to the identification of forged passports by comparing the information on the passport image itself by using the characteristic of each image. Experiments have been performed by using 50 different passports from 22 different countries. The characters on the MRZ area of these passports consist of the same font type and have a simple background image for easier recognition. However, the background image is complex in the personal information area which made the recognition of only 22 passports possible through this method. Therefore, a method for removing the hologram from the passport and a method for extracting the characters from a background with the same color is required for a more precise recognition.

In this paper, we have implemented and experimented with the concept of passport structure analysis and character recognition module. A method for removing the hologram and a module for extracting the characters from the background has to be developed for a more precise passport recognition system.

Acknowledgments. This work is supported by Seoul R&BD program (10544).

References

1. Kim, T.J., Kwon, Y.B.: Crosscheck of Passport Information for Personal Identification. In: GREC 2005, pp. 162–172 (2005)
2. Kim, K.B., Kim, Y.-J., Oh, A.-s.: An Intelligent System for Passport Recognition Using Enhanced RBF Network. In: Zhang, J., He, J.-H., Fu, Y. (eds.) CIS 2004. LNCS, vol. 3314, pp. 762–767. Springer, Heidelberg (2004)
3. Trier, P.D., Taxt, T.: Evaluation of Binarization Methods for Document Images. IEEE Trans. On PAMI 17(3), 312–315 (1995)

4. Dawoud, A., Kamel, M.S.: Iterative Multimodel Subimage Binarization for Handwritten Character Segmentation. IEEE Trans. On IP 13(9), 1223–1230 (2004)
5. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Addison Wesley Longman (1992)
6. Tan, H.L.: Hybrid feature-based and template matching optical character recognition system, United States Patent 5077805 (1991)
7. ICAO, Document 9303, http://mrtd.icao.int/content/view/33/202/