Indian Journal of Science and Technology, Vol 10(12), DOI: 10.17485/ijst/2017/v10i12/113012, March 2017
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Online Application of Printed Jawi Character Recognition

Sayed Muchallil[1]* and Nazaruddin[2]

[1]Department of Electrical and Computer Engineering, Faculty of Engineering, Syiah Kuala University, Indonesia; sayed.muchallil@unsyiah.ac.id
[2]Department of Informatics, Faculty of Mathematic and Natural Science, Syiah Kuala University, Indonesia; anzaro@unsyiah.ac.id

Abstract

Objectives: In this paper, a system of extracting moment feature based on online system for printed Jawi character recognition was presented. This application was built using PHP programming language. **Methods/Analysis:** We tested this application by transforming the character into condition: rotating and scaling. We rotated the image character by using 45, 90, 180, and 270 of degree and we scaled by using 2, 3, 4, and 5 scaling factor. **Findings:** Generally, the online application was able to extract moment invariant feature from a character. This system has around 93.24% sucessful rate of scaling character and 92.98% of rotating character.**Novelty/Improvement:** This research is the first research of Jawi character recognition for online application.

Keywords: Jawi Character Recognition, Moment Feature, Moment Invariant, Online OCR

1. Introduction

Researchers have carried on experiments in Optical Character Recognition (OCR) since 1970s. Since then many applications have been developed for OCR system. However, standalone character is much easier to be recognized compare to cursive characters. On the other hand, the OCR application for Jawi Language has not been built and developed. An OCR system is constructed by four main stages namely pre-processing[1,2,3] such as binarization and denoising, Segmentation[4,5], feature extraction[6,7], and pattern classification.

The researchers have realized the important of Jawi character recognition application because this language was used by many countries in South East Asia region in their ancient documents and artifacts. There are almost 15.000 historical manuscripts using Jawi characters in Indonesia only. These manuscripts contain religion related text, stories and some historical events. Many of these documents were kept in the museum in Dutch, British, Malaysia, Indonesia and other countries[8].

Researcher in computer vision and image processing are trying to develop some OCR application for Chinese, India and Arabic characters. Compare to Jawi Characters, the algorithm for recognition the characters mentioned before has been well-developed. In the last few years, the Farsi and Urdu character recognition also developed. In the International Conference for Document Analysis and Recognition (ICDAR)[9], there is no Jawi related articles.

In earlier of 2000s, the recognition of Jawi character has been started. In[10,11] developed Jawi character recognition system by using neural network system. Zaidi started on segmentation focused[4,5] while [6,7]focused on feature extraction.

[12]Experimented and developed an Optical Character Recognition (OCR) for Chinese Character in 2001. There were three main blocks for this application. Feature Extracted is the first step of these main blocks. There was

*Author for correspondence

Clustering Phase for the second blocks. The last block was recognizing stage. This research tried to recognize 13.053 Chinese Characters. The result of this study is the about 97.4 % Chinese Characters can be recognized.

[13]Studied how to create an OCR that is simple and efficient for basic words for Kannada which is one of India written text. For this research, they used typed characters, not hand-written, including vocal and consonant. This research used Zernike moment and Hu Moment Invariant. In addition, the method also adds some classification using Radial Basis Function (RBF). All of these method is compared for their performance in recognizing Kannada Characters. Zernike Moment performance could recognize 96.8 % compare to Hu Moment Invariant that can only know 82% of the characters.

[14]Studied hand-written recognition using Moment Invariant in 2006. The objects for this research is Latin characters. The characters written cursively. Methods that were implemented for this research are Hu moment Invariant, Zernike, and Affine were the method that were used for this experiment. Classification also used Redial Basis Function (RBF). There was pre-processing step for calculating each moment for those three methods. The result of this study showed that the process using RBF was better.

[15]Proposed a new algorithm for Character Recognition that named as Radial Sector Coding (RSC). The differences between RSC and Invariant Character Recognition (IRC) that is the proposed method did not used complex computation. The objects for this study is 26 Latin alphabet. There were four experiments, the first was 40px x 40px Arial Font. The rotation from 0 to 90-degree angle for training and 0 - 350-degree angle for recognition process. The second experiments still used the same size ad font. The only difference is the angle for training 0 - 135 and for testing 0 - 355. The third experiment used a bigger size 50px x 50px with the same font and the same angle for training and testing as the first experiment. The last experiment used a smaller size 30px x 30px with the same angle for training and testing as the first experiment.

[16]Researched how to recognize hand-written Arabic characters using mapping invariant. Invariant Mapping consist of four processes. The process was moment invariant technique, Fourier descriptors, boundary-based techniques and some other techniques like vector analysis. This experiment clustered characters into 18 categories based on the character form. For example, ba and ta will be in the same class since they have the same form. The only difference they have is the number of dots. Feature can be calculated into two parts, the first was invariant mapping and the latter was combination. The accuracy for this study is about 92.75 %.

All this research focused on offline OCR. Most of this study used MATLAB to calculate the moment invariant. This paper presented how to calculated Hu Moment Invariant using PHP that can be accessed via web server. Internet connection and browser which is already installed by default for many Operating System (OS) are the only thing required for running this application. The aim of using PHP that this application can be run anywhere without any software installation. This research focus on calculated seven moment invariants from a digital version of character image.

2. Experiment Details

2.1 Template Creation

DPJ database[17] was used as dataset for experiment. The dataset used in this study is printed Jawi character image. Total number of data that were used in this experiment is 1524 characters. 10% for every font style from the dataset were taken as sample data in the experiment. The samples which are taken was considered to represent every shape that are different from each character. This is to ensure that every different shape of each character drawn as samples in this experiment. Figure 1 is examples of the experiment's character.

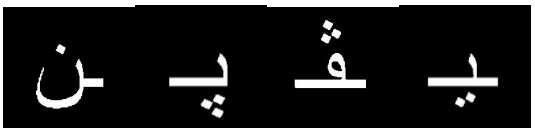

Figure 1. Example of experiment characters.

2.2 Feature Extraction

Feature extraction is a step to obtain a spesific characteristic from an image. This characteristic uses as feature to recognize Jawi character. There is two kinds of feature: sintaxial based and statistical based. Sintaxial feature is based on feature shape representation and statistical feature based on distribution of the image pixel.

Hu moment invariant is one of extraction feature method that often using to get image feature which is based on statistical feature[18]. Hu extracted the feature

from an image using moment method. Hu moment invariant also known as geometrical moment invariant. Hu moment is defined as[19]:

$$\phi_1 = \eta_{20} + \eta_{02}$$
$$\phi_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2$$
$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$
$$\phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2$$
$$\phi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$
$$+(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$
$$\phi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + 3\eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$$
$$\phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$$
$$+(3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$$
$$+(3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \quad (6)$$

Every moment invariant value using as feature for an image.

2.3 System Development

On this phase, the application development process started on the calculation of moment invariant of every character. The programming language that are used for the development is PHP. Formula and calculation from Digital Image Processing book converted into the PHP language[20].

The application in his book used MATLAB, on contrary our application used PHP so it can be accessed via internet. However, the procedures and stages for both are the same. There are five phases for recognition process. (i) Reading image. (ii) Convert image to be two-dimension image. (iii) Calculate central moment. (iv) Normalize central moment (v) calculated seven moment invariant according to equation 6.

On the fourth stage, normalized central moment, the result is in decimal format. This is why the moment invariant resulted from this calculation is in decimal form. However, the computation process need to be accelerated, so the moment invariant will be converted into integer.

The final result of this process is the value of seven moment invariants. These values were kept in a database and would be used for comparison in next step.

2.4 Experimental Testing

Testing stage for the proposed application used the same image with different size uses rotating and scaling transform. For the rotating character, we used 45, 90, 180, and 270 angle of degree. The non-transformed character was scaled by using two, three, four, and five of scaling factor. The testing was applied to assest that rotating and scaling invariant of Hu moment on handwritten Jawi character and the online system implementation works appropriately.

To get comparison of the transformed character, we compare the scaling and the rotating character with the non-transformed character by using equation 7 as define follow:

$$Rr = \frac{sm}{tme} \quad (7)$$

Rr is the accuracy of the feature extraction by using Hu moment while Sm is the similar moment to the basic character that accomplished by using the online application. Tme is total of moment feature that extracted using Hu method.

3. Result and Discussion

3.1 Basic Character

We selected 28 characters from total 125 as experiment data. This 28 characters are representing all of character shape. The characters which isused in the experimentare : isolated alif, isolated ba, ta begin-form, jim begin-form, ca isolated, da isolated, za isolated, sin end-form, sya middle-form, shad middle-form, dhad end-form, tha end-form, ain begin-form, ng end-form, fa isolated, p middle-form, qaf end-form, Kaf begin-form, g begin-form, lam isolated, mim middle-form, nun end-form, ny begin-form, waw end-form, haa middle-form, lamalif end-form, hamzah, ya middle-form.

From the overall character of the sample that was calculated value of its moment feature, we found no characters that have the same seventh grades moment. Table I showed the value of Hu moment invariant feature from eachnon-transformed chacracter.

3.2 Scaling Transforms Testing Characters

Total moments that are extracted for each of character is 112 moment feature. We used four scaling factors for every character. The factors are 2, 3, 4 and 5 factor.

The recognation rate of the images that transform by using factor scaling of 2 is 91.84%. Most of the moment that unable to be recognize is the 7th moment of the

characters. Alif isolated, ta begin-form, ca isolated, sin end-form, shad middle-form and ain begin-form are the characters which have all moment extracted correctly.

The character that scaled by using factor scaling of 3 and 5 had accuracy of 93,88%. These recognition rate are the highest rate of the experiment. Most of the moment in the case are failed to be recognize in the 6th moment. Ng end-form, ny begin-form, ya middle-form, jim middle-form, da isolated-form are the examples of the character that failed in extracting the 6th moment value.

In the experiment that uses scaling factor of 4, we achieved 93.37% of accuracy. Most of the 7th moment value in this experiment step had different moment value from basic character. Figure 2 show the recognition rate of scaling transform character.

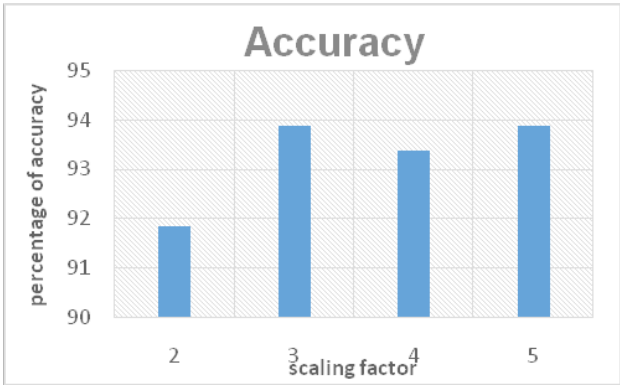

Figure 2. Recognition rate of scaling transform character.

3.3 Rotating Transforms Testing Characters

In rotating transform character, we used four angle of rotation. The angle degree are 45, 90, 180, and 270 degree. Total moment which is extracted in this testing are 112 moment values.

The recognation rate of the images that transform by 45 degree is 88.27%. Most of the moment that unable to be recognize in this condition is the 7th moment of the characters. Alif isolated, ba begin-form, ca isolated, sin end-form, ya middle-form and ghain begin-form are the characters which have all moment extracted correctly.

In the experiment that uses angle of 90 degree and 270 degree, we achieved 93.37% of accuracy. Most of the 7th moment value in this experiment condition had different moment value from basic character. The example of the character that have correct all of the moment are hamzah, waw end-form, nun end-form, mim middle-form, g begin-form, qaf end-form, isolated fa, and pa middle form.

The character that scaled by 180 degree is 96.94%. This recognition rate are the highest rate of the experiment. Most of the moment in the case are failed to be recognize in the 6th moment. Ca end-form, ny begin-form, ya middle-form, ba middle-form, da end-form are the examples of the character that failed in extracting the 6th moment value. Figure 3 showed the accuracy of moment feature.

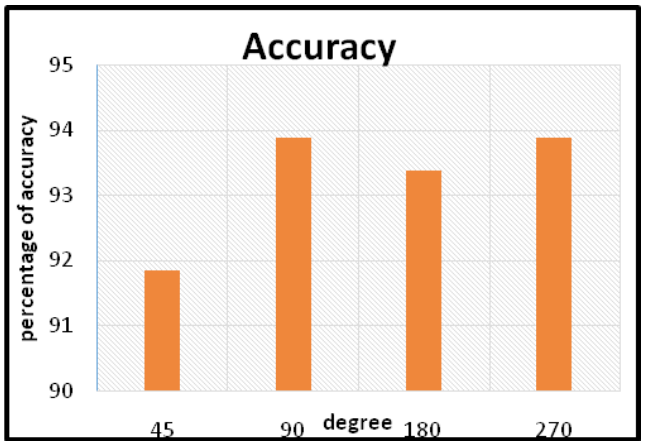

Figure 3. Recognition rate of rotating transform character.

4. Conclusion

In this paper, we introduced the online OCR system for extracting Hu moment from a Jawi character. The online OCR system was developed on web server based programming. In the experiment we used two conditions: scaling character transformation and rotating character transformation.

The result showed that the scaling character obtained the highest recognition rate for the character that scaling by using 3 and 5 scaling factor. The accuracy of ths condition is 93,88%. Furthermore, The highest recognition rate of the rotating transformation condition is the character that rotating by angle of 180 degree. The accuracy of this condition is 96.94%.

5. References

1. Arnia F, Munadi K, Fardian F, Muchallil S. Improvement of binarization performance by applying dct as preprocessing procedure. IEEE Press: Proceedings of 2014 6th International Symposium on Communications, Control, and Signal Processing, ICCSP 2014. 2014; p. 128–32.
2. Muchallil S, Arnia F, Munadi K, Fardian. Performance comparison of denoising methods for historical documents. Jurnal Teknologi. 2015; 77(22):137–43.

3. Fardian F, Arnia F, Muchallil S, Munadi K. Identification of most suitable binarisation methods for acehnese ancient manuscripts restoration software user guide. Jurnal Teknologi. 2015; 77(22):95–102. Crossref.
4. Razak Z, Rosli S, Mashkuri Y. Hardware Design of On-Line Jawi Character Recognition Chip using Discrete Wavelet Transform. Proceedings 2005 Eighth International Conference on Document Analysis and Recognition. 2005; p. 91–5.
5. Razak Z, Zulkiflee K, Noor NM, Salleh R, Yaacob M. Off-line handwritten Jawi character segmentation using histogram normalization and sliding window approach for hardware implementation. Malaysian Journal of Computer Science. 2009; 22(1):34–43.
6. Nasrudin MF, Petrou M, Kotoulas L. Jawi character recognition using the trace transform. 2010 Seventh International Conference on Computer Graphics, Imaging and Visualization (CGIV). 2010; p. 151–6. Crossref.
7. Nasrudin MF, Petrou M. Offline handwritten Jawi recognition using the trace transform. 2011 International Conference on Pattern Analysis and Intelligent Robotics (ICPAIR). 2011; p. 87–91. Crossref.
8. Moain AJ, Pustaka DB. Dewan Bahasa dan Pustaka: Perancangan bahasa: sejarah aksara Jawi. 1996.
9. Nasrudin MF, Omar K, Zakaria MS, Yeun LC. Handwritten cursive Jawi character recognition: a survey. 2008 CGIV'08 Fifth International Conference on Computer Graphics, Imaging and Visualisation. 2008; p. 247–56. Crossref
10. Omar K. Universiti Putra Malaysia: Jawi handwritten text recognition using multi-level classifier. PhD Thesis. 2000.
11. Manaf M. Universiti Kebangsaan Malaysia: Jawi handwritten text recognition using recurrent bama neural networks. PhD Thesis. 2002.
12. Yang T-N, Wang S-D. A rotation invariant printed Chinese character recognition system. Pattern Recognition Letters. 2001; 22(2):85–95. Crossref.
13. Kunte RS, Samuel RDS. A simple and efficient optical character recognition system for basic symbols in printed Kannada text. Sadhana. 2007; 32(5):521–33. Crossref.
14. Lacrama DL, Snep I. The use of invariant moments in hand-written character recognition. 2009. arXiv Prepr arXiv09043650.
15. Iqbal A, Musa ABM, Tahsin A, Sattar MA, Islam MM, Murase K. A Novel Algorithm for Translation, Rotation and Scale Invariant Character Recognition. SCIS & ISIS. 2008; p. 1367–72.
16. Kharma NN, Ward RK. A novel invariant mapping applied to hand-written Arabic character recognition. Pattern Recognition. 2001; 34(11):2115–20. Crossref.
17. Saddami K, Munadi K, Arnia F. A database of printed Jawi character image. IEEE Press: Proceedings of 2015 3rd International Conference on Image Information Processing, ICIIP 2015. 2015; p. 56–9. Crossref.
18. Shih FY. New Jersey: John Wiley and Sons, Inc.: Image Processing and Pattern Recognition: Fundamentals and Techniques. 2010. Crossref.
19. Hu M-K. Visual pattern recognition by moment invariants. IRE Transactions on Information Theory. 1962; 8(2):179–87.
20. Gonzalez RC, Woods RE, Eddins SL. New Jersey: Pearson Prantice Hall: Digital Image Processing Using Matlab. 2004.