Online Feature Extraction Technique for Optical Character Recognition System

Khairun Saddami[1,2], Khairul Munadi[1,2,3] and Fitri Arnia[1,2,3,*]

[1]Doctoral Program of Engineering, Faculty of Engineering, Syiah Kuala University, Darussalam Banda Aceh 23111, Indonesia
[2]Graduate Program of Electrical Engineering, Faculty of Engineering, University of Syiah Kuala Darussalam, Banda Aceh 23111, Indonesia
[3]Department of Electrical and Computer Engineering, Faculty of Engineering, Syiah Kuala University Darussalam, Banda Aceh 23111, Indonesia
*Corresponding Author: f.arnia@unsyiah.ac.id

Abstract

In this paper, we evaluated the performance of existing online Character Rcognition System (CSR) in extracting the feature from a standard testing image and a Roman character image. The rotating angle that use in the experiment are 90 and 180 of degree, and the scaling factors that use in the experiment are 0.5 and 2. Then, we compared the original image feature with the feature of scaled and rotated image. The result of the image feature was compared to the rotating and scaling image. We concluded that the application is succeeded to recognize with accuracy up to 95% in average.

Keywords:feature extraction, online OCR, character recognition system, Hu moment invariant

Introduction

Research in image processing field increases significantly. Now,communication is done not only through the text but also through an image. Letters, Checks, or other documents are transmitted by using computer applications. One of the ways to facilitate the storage and to distribute the documents is by digitalizing the document by photographing or scanning. Getting back or changing the image into the text format requires special technique. The technique should be capable to recover the text from the image. A system that can convert an image into a text is character recognition.

Character recognition is an interesting field in image processing research area. Character recognition has become popular research because increasing variation of characters. The main purpose of character recognition is to convert a character from digital image format become digital text format. The system that was developed for character called as Optical Character Recognition (OCR). To create an OCR application, at least we need four stages: pre–processing such as denoising (Arnia, *et al*., 2014), segmentation, feature extraction and pattern classification.

Feature extraction is an important step in the character recognition system. Feature extraction is performed for generating some features from an abjad or an alphabet. Each alphabet required having different and unique features. This feature will be assigned to pattern classifier for learning and classifying the feature into a particular character. Increasing the number of features that assign to a classifier, the pattern classification will get a better result.

Apart of feature extraction, to develop a character recognition system, we need a pre–processing procedure such as binarization (Fardian, *et al*., 2015), denoising (Muchallil, *et al*., 2015) and skew correction. Furthermore, segmentation and pattern classifier are also important procedure. Previous OCR techniques is developed based on an offline system. If we use the offline system, we require the reinstallation process, and it cost more times. On the other hands, the development of science and technology, many applications are developed and connected to the internet. This application was adequated to use without reinstallation process. All script commands and code will be processed by the browser through an application called a web server. This web server which will compile all of the client commands that are desired by the user then sends its result to the client. All offline application can be convertedto the online version including character recognition system.

In this paper, we evaluated an online feature extraction by using moment invariant techniques for Roman character and standard testing image. This online OCR system was proposed by Saddami for Jawi character (Saddami *et al.* , 2016). We implemented this system on local Apache server or known as localhost.

The rest of this paper is literature review on the second section. In the third section, we illustrate experiment method while in section fourth we present result and discussion. Finally, we give the conclusion on the fifth section.

Literature Review

Feature Extraction

An image feature is a value which represents identity or attribute from the image. The image feature should be extracted to get the attribute of the image, and we donate that the attribute should be different for each image object. There are two types of feature: syntaxial and statistical. Syntaxial is a feature that represents boundary or shape of the image while statistical feature is the feature that based on the image statistic distribution. One of the benchmark statistical features is Hu moment or called as geometric moment invariant (GMI) (Hu, *et al.*,1962). GMI have been widely used for extracting an image feature. Noh proposed palmprint recognizing by using Hu moment invariant (Noh, *et al.*, 2005). Then, Munadi et al., used GMI to extract image features for secure online trading (Munadi, *et al.*, 2013). In 2015, Sun extracted speech emotion feature using Hu moment invariant (Sun, *et al.*,2015).

Hu Moment Invariant

Geometric moment invariant (GMI) was proposed by Hu in 1962. Hu assumed that the statistical distribution of an image could be seen as a set of statistical distribution. The moments are defined as follow:

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)\,dx\,dy \qquad p, q = 0,1,2,\ldots \tag{1}$$

for discrete image, the moment value are defined as:

$$m_{pq} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} x^p y^q f(x,y) \tag{2}$$

the central moment of the image is

$$\mu_{pq} = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} (x-\bar{x})^p (y-\bar{y})^q f(x,y) \tag{3}$$

while the $\bar{x}$ and $\bar{y}$ is evaluated as follow:

$$\bar{x} = \frac{m_{10}}{m_{00}}, \bar{y} = \frac{m_{01}}{m_{00}} \tag{4}$$

the normalize central moment is defined as:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu^{\gamma}{}_{00}} \quad \text{where } \gamma = \frac{p+q}{2} + 1 \tag{5}$$

then Hu moment value was camputed as:

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

Experimental Methods

In this section, we describe the experimental method. The experimental method consists of three sub–section: Dataset creation template, creating the online application and assessing the application.

Dataset Creation Template

In this experiment, character images and standard testing images was tested on online OCR. We use lena, cameraman, coins, rice and moon image for the standard image digital. The standard image digital was obtained from Signal Image Processing Institue (SIPI) (SIPI, 2005). Examples of the standard digital image showed in Fig. 1 and Fig. 2.
The character image was created by following steps:
1. Write all Roman alphabets by using font size 60.
2. Print the characters.
3. Scan the character document to change it into image format
4. Split each alphabet into a single image file.
5. Change the size of image character into 256x256.

Total Roman characters that we used in the experiment were 27 characters. Fig. 3 showed examples of Roman alphabets in image format.

Figure 1. Lena image

Figure 2. Cameraman image

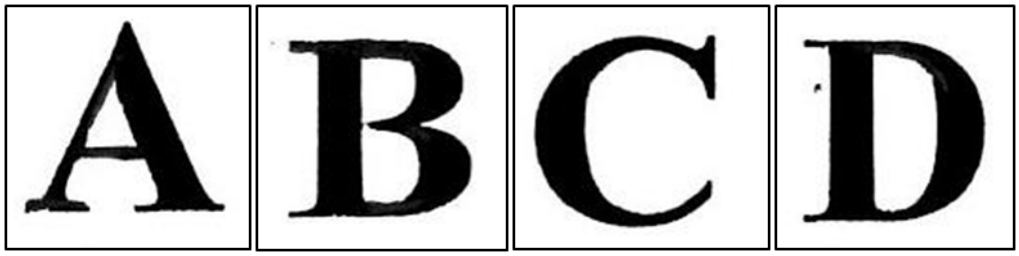

Figure 3. Roman character image

Creating The Online Application

The online OCR for character recognition was developed using PHP programming. In this paper, we rebuilt and evaluated the online OCR system for binarizing and feature extraction stages. The online application was accomplished by performing following step:
A. Binarization Phase
The binarization process was run by applying these steps:
1. Read the image
2. Determine the image thresholding.
3. Convert the image pixel value lower than the threshold into zero and higher than the threshold into one.

B. Feature Extraction Phase
The feature was extracted by using equation 2 to equation 6. The steps below are the process of extracting an image feature (Gonzalez *et al*., 2004).

1. Compute m_{pq} value of the object by using eq. 2.
2. Determine the central moment from the picture by using eq. 3.
3. Calculate the $\bar{x}$ and $\bar{y}$.
4. Determine the normalized central moment
5. Establish the moment value based on eq. 6.

Assessing The Application

To ensure the application performance in extract moment feature, we assess the application by performing on transformation image and character. The transformations that applied to the character are scaling and rotation. The image and the character were scaled by a half and two–factor scaling. The rotation degrees were 90^0 and 180^0. The recognition rate is used to approach the the result of the testing method to the untransformed character. The recognition rate is as described below (Saddami, *et al*., 2016).

$$Rr = \frac{sm}{tme} \qquad (2)$$

while Rr is recognition rate, sm is similarity moment with a basic image, and tme is the total moment that was extracted. The basic image is the image that not transform by using scaling nor rotating.

Results and Discussion

In this section, we described the result of the experiment. The result of experiment was analyzed using recognition rate equation that noticed in eq.3.

Transformed Image and Character

The result of transformed image and character was presented in Table 1 and Table 2. Table 1 demonstrates the result of seven moments feature from standard digital image The standard image was transformed by using two scaling factors: 0.5 and 2, and rotated by using two angles: 90 and 180 degrees. Table 2 showed the outcome of Hu moment features from character image. The character image was also transformed by using the scaling factor and angle of degree as in the standard image.

In the image which was scaled by using 0.5 scaling factor, we found that only one moment from 35 extracted moment which has different from the basic image. Furthermore, the image which was scaling by using 2 scaling factor had two moments that differ to the basic image. The result showed that the image which was scaled by 0.5 scaling factor had 97.1% of recognition rate, while the 2 factor scaling image had the accuracy of 94.3%. We established that eight images had the dissimilar moment from the basic image for the image that we rotated by using angle of 90 degrees. In the image that rotated by using angle of 180 degrees, we obtained five moments that differ to the basic image moment. Based on Table 1, the image that rotated by angle of 90 degrees had 77.1% of accuracy and the image that rotated by angle of 180 degrees had recognition rate 90.0%.

In the image that we transformed by using angle of 90 degree, we found that 19 moments from 756 extracted moment which has different from the basic image character. Furthermore, the image which was rotated by using angle of 180 degrees had 14 moments that differ to the basic image characters. The result showed that the image that scaled by angle of 90 degrees had 97.5% of recognition rate, while the accuracy of the character image with angle of 180 degrees increase to 98.1%. We established that eighty images had dissimilar moment from the basic image character for the image that we transformed by using 0.5 scaling factor. In the image that rotated by using 2 scaling factor, we obtained 66 moments that differed to the basic image character moment. Based on Table 2, the image that rotated by 0.5 scaling factor had 89.4% of accuracy and the image that rotated by 2 scaling factor had recognition rate 91.3%.

The similarity level of the standard image is less than thatthe character image. It is caused by difference number of object in an image. In the standard image there are many objects in an image frame but in the character image, there is only one object in one image frame. To increase the recognition rate in extracting Hu moment feature, we suggest to segment objects in an image to be a separate object.

Table 1. Result of scaling and rotating image feature extraction

No	Scaling					Rotation				
	Scaling factor	Similar moment	Dissimilar moment	Total	Accuracy	Angle	Similar moment	Dissimilar moment	Total	Accuracy
1	0.5	34	1	35	97.1%	90	27	8	35	77.1%
2	2	33	2	35	94.3%	180	30	5	35	85.7%
Average					95.7%	Average				90.0%

Table 2. Result of scaling and rotating character feature extraction

No	Scaling					Rotation				
	Scaling factor	Similar moment	Dissimilar moment	Total	Accuracy	Angle	Similar moment	Dissimilar moment	Total	Accuracy
1	0.5	737	19	756	97.5%	90	676	80	756	89.4%
2	2	742	14	756	98.1%	180	690	66	756	91.3%
Average					97.8%	Average				90.4%

Conclusions

In this paper, we evaluated online OCR system for Roman character. This system uses Hu moment as feature extractor. This online OCR tested for rotating and scaling image. The rotating angle that use in the experiment are 90 and 180 of degree, and the scaling factors that use in the experiment are 0.5 and 2. The result of this research showed that the feature extraction process of Hu moment by using the online application was successfully applied. The value of Hu moment in variant that was extracted after scaling and rotating had recognition rate reached 60.8% of scaling and91.75% of rotating respectively. The result onthe character image is better than the standard images because in the standard image there are many objects in one image meanwhile in character image there is only one object.

Acknowledgements

This research was supported by *Ministry of Research, Technology and Higher Education* under a scheme that called “PMDSU.”

References

Arnia, F. Munadi, K. Fardian, and Muchallil, S. (2014). Improvement of Binarization Performance by Applying DCT as Pre–Processing Procedure. In *Proceeding of International Symposium on Communications, Control, and Signal Processing*, pp. 128–132.

Fardian, Arnia. F., Muchallil. S. and Munadi, K. (2015). Identification of Most Suitable Binarization Method for Acehnese Ancient Manuscripts Restoration Software User Guide. *Jurnal Teknologi*, 95–102.

Gonzalez, W., Woods, R. E. and Eddins, S. L. (2004). *Digital Image Processing Using MATLAB*. Third New Jersey: Prentice Hall.

Hu., M. K. (1962) Visual pattern recognition using by moment invariant. *IRE Transactions on Information Theory*, 179–187.

Muchallil, S. Arnia, F. Munadi, K. and Fardian. (2015). Performance Comparison of Denoising Methods for Historical Documents. *Jurnal Teknologi*, 143–137.

Munadi, K., Syaryadhi, M., Arnia, F., Fujiyoshi, M. and Kiya, H. (2013). Secure online image trading scheme using DCT coefficients and moment invariants feature. In Proceeding of *2013 IEEE International Symposium on Consumer Electronics (ISCE)*, 291–292.

Noh, J. S. and Rhee, K. H. (2005). Palmprint identification algorithm using Hu invariant moments and Otsu binarization. In Proceeding of *Fourth Annual ACIS International Conference on Computer and Information Science (ICIS'05)*,94–99.

Saddami, K. Arnia. F. and Muchallil, S. (2016).Online Application of Handwritten Jawi Character Recognition. *International Conference on Engineering, Science, and Technology* [Accepted].

SIPI, U. (2005).*The USC–SIPI image database*. http://sipi.usc.edu/database/(July 15, 2016)

Sun, Y., Wen, G. and Wang, J. (2015). Weighted spectral features based on local Hu moments for speech emotion recognition. *Biomedical Signal Processing and Control*, *18*: 80–90.