Open Access Journal of Information Systems (OAJIS)

Available at http://is.its.ac.id/pubs/oajis/index.php

The Third Information Systems International Conference

The Implementation of Shape Based Feature Extraction and Similarity Measures to Prevent Falsification of Handwritten Document

Dian Pratiwi[a], Syaifudin[a]

[a]*Trisakti University, Jl.Kyai Tapa No.1, Jakarta-Indonesia, 15000*

Abstract

Document security in present is one of important thing because many crimes against the falsification of documents is growing anyway. This research was conducted with the aim of designing a software system that can recognize handwriting of the owner of the document through the characteristics of shape, so that the falsification of documents can be prevented. The method applied in this study consisted of writing stage analog to digital conversion, pre-processing, segmentation into the size of 500x200 pixels and 10 grids, feature extraction stage, and the percentage calculation of similarity through the similarity measures method. From the research that has been conducted on 20 documents of 4 owners handwriting, 11 documents identified the owner managed appropriately through matching the features of the 12 words in each document, namely "The", "You", "Will", " To", "He", "And", "It", "Is", "Are", "His", "Have", "For". So that the percentage of success obtained against the document security system that is equal to 55%. The percentage is still possible can be increased by adding more number of document and mapping features with better methods again

Keyword : Handwriting, Document Security, Similarity Measures, Feature Extraction, Segmentation

1. Introduction

Currently, the development of various types of computer technology has developed rapidly following the various needs that exist. Computer technology is no longer foreigners and almost every individual is able to use it in everyday, such as in the use of notebook or laptop. These advances also have an impact on developments in software device such as the handwriting recognition device. Handwriting recognition tool is pretty much up to now has been applied in a handheld device or mobile phone based touchscreen. In function, the device is generally used only as a reader of results of user's handwriting on the screen to be known the meaning of the sentence. Handwriting recognition function then can be developed further to be known owner of the handwriting of the results of the analysis of shapes or patterns that exist in it. Handwriting patterns can be known through a series of image processing and feature extraction of texts

that will produce features or special characteristics that can be used to identify each owner handwriting tested. With this handwriting recognition software, in the future is not just to be able to identify the owner of the handwriting pattern, but can also be used as a protection system such as in terms of securing important documents from counterfeiting writings and the system is maintained through the input of pattern article.

2. Theoretical

2.1. Pre-processing

Pre-processing is an early stage that needs to be done to get the post data in digital form with the same size of the pixel and the same greylevel of a set of analog handwriting that has been digitized by means of a digitizer or scanner. This stage consists of RGB color conversion into grayscale and thresholding.

- RGB to greyscale color conversion
  RGB to greyscale color conversion is a stage to change the color value of 24 bits to 8 bits, so the size of the resulting color will be smaller with the interval between 0 to 255 [1].
- Thresholding
  Thresholding is a process to saparate the object region (foreground) to the background area through a certain threshold value [2]. In this study, threshold value is also determined by trial and error

2.2. ROI Formation (Gridding)

ROI (Region of Interest) formation is a technique that is commonly performed to assist the analysis of the object to be observed, such as fMRI image analysis conducted by researchers from the UCLA - Los Angeles, Russel A Poldrack in 2007. This technique can improve the success of the introduction phase, due to the information of ROI, feature extraction process to be performed is limited to a specific region or area that has been restricted [3].

2.3. Feature Extraction

Feature extraction is an important stage of the pattern recognition application. This stage will give results in the form of values of the feature to be measured or recognized as a pattern. With feature or traits extraction, important information of data (which in this study is the form of image data) will be taken and stored in the feature vector [2]. Features that can be extracted in the form of image data including color features, shapes, and textures. And in this study, which will be extracted feature is based on the representation of handwritten form. The values of the handwriting feature extraction forms-based will be binary values (worth "0" and "1") for each grid for each image, where the value of "0" will be given if the representation of the grid is background object, and the value "1" if the grid is representation of foreground objects with a minimum of 15% of total pixels of each grid is a foreground object [4].

2.4. Pattern Recognition

Pattern recognition is one of the artificial intelligence techniques that aims to recognize the features or specific characteristics of data set (both text and image document) and classify [5]. Pattern recognition can be done in several ways, one of which is by using the method of *similarity measures*.

Similarity measures is a method that can be used to look for similarities from one object to another, by calculating the distance of which [6].

As in the research conducted Anna Huang in 2008, this study also used the technique of Similarity Measures to recognizing handwriting patterns by calculating the distance between the patterns by using the Euclidean Distance formula [6] :

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (qi - pi)2}$$

Where d is the distance between the total value of the handwriting image pixel with each other, q and p are the pixel image.

2.5. Calculation of Accuracy, Precision and Recall Percentage

This step is the final stage, where each image handwriting tested will be the level of success of its introduction. The formula used is :

$$Accuracy = \frac{Total\ of\ Image\ Handwriting\ Successfully\ Identified}{Overall\ Count\ of\ Handwriting\ Tested}\ x\ 100\%$$

$$Precision = \frac{Total\ of\ relevant\ images \cap total\ of\ retrieved\ images}{Total\ of\ retrieved\ images}\ x\ 100\%$$

$$Recall = \frac{Total\ of\ relevant\ images \cap Total\ of\ retrived\ images}{Total\ of\ relevant\ images}\ x\ 100\%$$

3. Procedure and Implementation

In this study, the procedure is carried out consists of several stages.

3.1. Collecting Data

Collecting data in this study is done through direct searches of a number of resources (randomly close person with researcher) by a certain time interval. They are recorded in a book in white background, with intervals of three days in a row one times, and intervals of one week later one times, and intervals of one month later one times. This is done to see in future studies if there's any change in the shape of a person's handwriting in a certain period. So that the number of documents collected for each resource amount to five documents.

Data in the form of handwritten documents obtained by asking authors to write a sentence like the following :

"*The man who loves you more will allow you to grow as a person without taking space. He will be patient and kind. It is because you are his priority. He will always have a reason for seeing you*."

These sentences chosen by the researchers because it has a number of conjuction and common words in the English language each article, such as "The", "Will", "And", and so on. In this study, the handwritten data are tested using a sentence written with the English language, but can also use the handwritten with other languages. The entire document then will be scanned and stored as digital data.

3.2. Design and Testing

After the document that containing the digitized handwriting, the next step is pre-processing, where each article/handwritten document will be converted into a greyscale color that has a little bit more color. After that, every file or document will be taken 12 words in it through assimilation and cropping stage by using this system and stored in the folder of trial data. The word assimilation is done in accordance with

the words to be taken as “The”,”You”,”Will”,”To”,”He”,”And”,”It”,”Is”,”Are”, “His”,”Have”, “For”. Cropping phase will produce an image with a size of 500x200 pixels automatically, and each document will produce 12 pictures. From 12 pictures, each will be extracted feature shape after forming grids sized 100x100 pixels and the values of the features that contain a 10-digits binary value is then stored in the form of a text (.txt) via the application.

Overall, the data collected by researchers amounted to 20 documents from 4 sources (we call : A, B, C, D). All these documents resulted in a total of 240 words in common, characteristics are then calculated through the method of similarity measures. The results of these calculations will show the distance value of each author/writer/source, where the owner of the real documents will be selected based on the value of the smallest distance generated. Because the smaller the distance writings produced, will be more similar to the tested article.

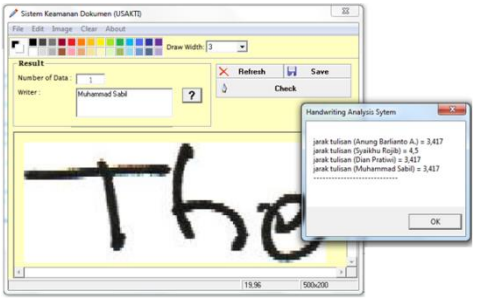

Fig. 1. Distance Calculation of Handwritten between each Writer

The following are summary tables the results of trials that have been done, where the implementation program of this study using Visual Basic 6.0 :

Table 1. The Test Result of First Day Document

Words / Author	1	2	3	4	5	6	7	8	9	10	11	12	True	False	Author's Prediction
A	D	B	A	A	D	A	D	D	D	C	C	D	3	9	D
B	D	B	D	B	D	D	B	B	D	D	D	B	5	7	D
C	D	B	D	C	D	D	C	C	D	D	C	B	4	8	D
D	D	D	D	D	D	D	D	D	D	D	C	B	10	2	D
Sum	1	2	2	4	1	2	3	3	1	1	1	1	22	26	
Words Accuracy													**45.8 %**		

Note : 1 = The, 2 = You, 3 = Will, 4 = To, 5 = He, 6 = And, 7 = It, 8 = Is, 9 = Are, 10 = His, 11 = Have, 12 = For

Table 2. The Test Result of Second Day Document

Words / Author	1	2	3	4	5	6	7	8	9	10	11	12	True	False	Author's Prediction
A	A	B	A	C	D	A	D	B	D	A	C	C	4	8	A
B	D	B	D	D	D	A	D	B	D	D	D	B	3	9	D
C	D	D	D	C	D	D	D	D	D	A	C	C	3	9	D
D	A	B	A	D	D	D	D	D	D	A	C	B	6	6	D
Sum	1	1	1	2	1	2	1	2	1	1	1	2	16	32	
Words Accuracy													**33.3%**		

Note : 1 = The, 2 = You, 3 = Will, 4 = To, 5 = He, 6 = And, 7 = It, 8 = Is, 9 = Are, 10 = His, 11 = Have, 12 = For

Table 3. The Test Result of Third Day Document

Words / Author	1	2	3	4	5	6	7	8	9	10	11	12	True	False	Author's Prediction
A	C	B	A	D	D	A	B	A	D	C	A	C	4	8	A
B	D	B	D	D	D	A	B	B	D	B	B	B	6	6	B
C	D	B	D	C	D	D	D	A	D	C	C	C	4	8	D
D	D	D	A	D	D	D	D	D	D	D	D	B	10	2	D
Sum	1	2	1	2	1	2	2	3	1	3	4	2	24	24	
Words Accuracy													**50%**		

Note : 1 = The, 2 = You, 3 = Will, 4 = To, 5 = He, 6 = And, 7 = It, 8 = Is, 9 = Are, 10 = His, 11 = Have, 12 = For

Table 4. The Test Result of First Week Document

Words / Author	1	2	3	4	5	6	7	8	9	10	11	12	True	False	Author's Prediction
A	A	B	A	C	D	A	D	B	D	C	A	C	4	8	A
B	B	B	A	D	D	D	D	B	D	D	D	B	4	8	D
C	D	B	C	C	D	A	D	B	D	C	C	C	5	7	C
D	D	D	A	C	D	C	B	B	D	D	C	B	5	7	D
Sum	3	2	2	1	1	1	0	1	1	2	2	2	18	30	
Words Accuracy													**37.5%**		

Note : 1 = The, 2 = You, 3 = Will, 4 = To, 5 = He, 6 = And, 7 = It, 8 = Is, 9 = Are, 10 = His, 11 = Have, 12 = For

Table 5. The Test Result of First Month Document

Words / Author	1	2	3	4	5	6	7	8	9	10	11	12	True	False	Author's Prediction
A	D	B	A	C	D	C	D	D	D	C	C	C	1	11	C / D
B	D	B	A	C	D	D	B	B	D	B	D	B	5	7	B
C	D	B	A	D	D	D	D	C	D	C	C	B	3	9	D
D	D	B	A	D	D	A	D	B	D	D	C	B	6	6	D
Sum	1	1	1	1	1	0	2	2	1	3	1	1	15	33	
Words Accuracy													**31.2%**		

Note : 1 = The, 2 = You, 3 = Will, 4 = To, 5 = He, 6 = And, 7 = It, 8 = Is, 9 = Are, 10 = His, 11 = Have, 12 = For

Table 6. Percentage of Precision and Recall

Author's Prediction by The Words				Author's Handwritten	Precision %	Recall %	Accuracy %
A	**B**	**C**	**D**				
16	8	16	20	**A**	51.6	26.7	39.5
4	23	1	32	**B**	47.9	38.3	
4	7	19	30	**C**	45.2	31.6	
7	10	6	37	**D**	31	61.67	

From the test results that shown in the tables above (table 1-5), the number of documents which successfully identified the owner of handwriting that is a total of 11 documents, of which as many as 3 document author A, B as much as 2 documents, 1 document C, and D as much as 5 documents. This indicates that the application still contained an error in recognizing handwriting characters of the author. Even in Table 5, the writing of the A author tends to be recognized as an author writing C or D. This could be due to the value of features used is relatively small, at only 10 binary digits so that the possibility of the value of the same features in different words still quite high because the method of similarity

measures can produce the same distance value in use amount of features slightly. And this may affect the prediction results of handwriting owners. This can also be seen from the calculation of precision and recall (table 6) the fairly low in value, where the precision obtained between 31 to 51.6 % and recall between 26.7 to 61.67 %. However, although there are errors, the results are still better than the rate of failure. Accuracy is achieved is still quite large, namely 55%. This level of accuracy can still be developed if the amount of data being tested and features that are used more. This is because the value of the features contained in the handwriting will be more varied, so the error factor in handwritten matching can be reduced.

4. Conclusion

Based on the results of the data and document security applications, the overall researchers can provide the following conclusions :

1. The percentage of success rate on the document security system in this study reached 55%, where 11 of 20 documents successfully distinguished handwriting correctly through the features form. It can be quite good considering the amount of test data that is used is still very small, so vulnerable to errors due handwritten documents matching feature data less well mapped.
2. The number of features that are used relatively little that amounted to only 10 feature values, causing some word used in writing the feature vector can have the same value even if the words a different.
3. Words that have a high accuracy is “To”,”Have”, ”The”, “His”, “It”, “Is”, where the results of the test is able to identify the owner of his writings better than other word said.
4. This research still requires further development in term of handwriting data used with more number or varied as well as the addition of more complex methods in order to improve the accuracy.

References

[1] Pratiwi, D. 2012. The Use of Self Organizing Map Method and Feature Selection in Image Database Classification System. *International Journal of Computer Science Issues (IJCSI)*, Vol.9, Issue 3 No.2 ISSN : 1694-0814

[2] Pratiwi, D. Santika, D.D, and Pardamean, B. 2011. An Application of Backpropagation Artificial Neural Network Method for Measuring The Severity of Osteoarthritis. *International Journal of Engineering & Technology (IJET-IJENS)*. Vol.11, No.3, ISSN: 117303-8585

[3] Poldrak, R.A. 2007. Region of Interest Analysis for fMRI. *Oxford Journal*. Vol.2 Issue 1, pp: 67-70. Los Angeles-USA.

[4] Lu, G. 1999. Multimedia Database Management Systems. Artech House Inc.

[5] Absultanny, Y.A. 2003. Pattern Recognition using Multilayer Neural Genetic Algorithm. *Neurocomputing*. Pp.237-247. Elseiver Science.

[6] Huang, A. 2008. Similarity Measures for Text Document Clustering. *New Zealand Computer Science Research Student Conference*. Christchurch. New Zealand