Framework Design of Information Retrieval System for Official Letter Using Extraction of Geometry Feature Method

L Fabrianto[1], Hendra[1], Surohman[1], Sfenrianto[2], NM Faizah[3]

[1]Master of Computer Science-Postgraduate Program, STMIK Nusa Mandiri, Jakarta, Indonesia
[2]Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia, 11480
[3]Information Systems Department, Universitas Tama Jagakarsa, Jakarta, Indonesia.

*1402269@nusamandiri.ac.id

Abstract. In a government institution many official letters are made daily, the letters have an official format in the *Tanggal, No. Surat, Lampiran dan Hal.* area, with an OCR (Optical Character Recognition) scanner an official letter can be digitized and processed to obtain characters that will be used as searches, but in fact in certain desired parts there are often handwriting, a method of pattern recognition is needed to recognize the handwriting. The process in this study including of normalization, segmentation and pattern recognition from labels in the *Tanggal, No. Surat, Lampiran dan Hal.* area,this system will generate output in text form, the output text is result that will use as a term in the process of the Information Retrieval System, so the search for an official letter will be easy to do. This study is only makes the framework design for the next research authors will explain the detail of process and results obtained.

1. Introduction

Official correspondence in matters of government and private offices has become a mandatory and very important matter; official letters are letters containing official matters or certain matters, such as examples: invitation letters, circulars, decrees, assignments, official notes, announcements and others.

In some cases there is an official letter that has an official format but is still filled with handwriting on certain small parts. Official letters, also can be written evidence that has the power of law, in an institution the documents that have been filed can also be a tool of historical evidence and a reminder of activities that have been carried out by the institution, official letters usually become unstructured information , so that the search for official documents is quite difficult because of the large number of letters in an institution, in this study which will be taken from an official letter only at the top of the letter such as *Tanggal, No. Surat, Lampiran dan Hal.*[1].

In the modern era, many scanners have been produced that can digitize official company documents such as the letters mentioned above. One tool that is quite popular is OCR (Optical Character Recognition) as an identifier of letters and numbers to be converted become text.

In this study an official letter will be scanned first and then recognized using artificial neural networks with geometric feature extraction methods, this method is used because many official letters still use handwriting in certain parts and as problem solving to recognize handwriting is very related to pattern recognition, geometry feature extraction method can be used to help computers recognize patterns or motives [3].

2. Pattern Recognition

2.1 Digital imagery

It can be defined as a function of two variables f (x, y), where x and y are spatial coordinates and the value of f (x, y) is the intensity of the image at that coordinate. An image is converted to digital form so that it can be stored in computer memory or other media [2].

2.2 RGB

RGB is an array of m x n multiplied by 3, because it contains 3 color definitions namely red, green and blue. Each pixel is a combination of red, green and blue. RGB is a 24-bit image where each of these colors is 8 bits so that it has a brightness of 256 levels (2^n). Calculation of RGB conversion to grayscale is formulated as follows: [4]

$$Grayscale = (0.299*R) + (0.587*G) + (0.114*B) \quad (1)$$

Where *R* as Red, *G* as Green, *B* as Blue

2.3 Threshold

Process of making grayscale images into images with binary values. Value 1 = white and value 0 = black. The following is an equation to determine the binary value based on the threshold value [2].

$$g(x,y)=\begin{cases}1, if\ f(x,y) \geq T\\ 0, if\ f\ (x,y) < T\end{cases} \quad (2)$$

g (x, y) is a binary image of a grayscale image f (x, y) while T is a threshold value. The quality of binary images is very dependent on the value of T.

2.4 The segmentation

This process is divided into 2, line segmentation and character segmentation. Line segmentation traces the image and cuts lines horizontally. Character segmentation is cuts characters from the results of line segmentation [5].

	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	1	1	1	1	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	1	1	0	0
6	0	1	1	1	1	0
7	0	0	0	0	0	0

Figure 1. Each line is traced and every pixel is added up, if the result is not 0 then there is a line. Character segmentation is to separate each character based on the line that has been cut, done by summing each column.

	2	3	4	5
2	1	1	1	1
3	0	0	0	0
4	0	0	0	0
5	0	1	1	0
6	1	1	1	1

Figure 2. Segmentation result

2.5 Edge detection

Detecting marks the part that becomes the image detail to correct the details of blurred images, which occur because of the effects of the image acquisition process. A point (x, y) is said to be the edge of an image if that point has a high difference with its neighboring pixels [10], therefore the effect of neighboring pixels will differ according to its location to the point where the gradient is calculated [6]. The method used in this research is Sobel, this method takes the principle of Laplace and Gaussian functions known as functions to generate HPF, and the advantage Sobel method is reducing noise before doing edge detection calculations [6].

$$M = \sqrt{Gx^2 + Gy^2} \quad (3)$$

M is the magnitude of the gradient.

2.6 Image Thickening

Dilation is the process of combining background points (0) into parts of objects (1), based on structuring element S that used [7].

Dilation steps, for each point on A, do the following:

A

$D(A,S) = A \oplus S$

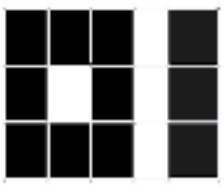
$A \oplus S$

Figure 3. 1st Place the axis point S at point A

Figure 4. 2nd Give the number 1 for all points (x, y) that are affected by the structure S in that position.

2.7 Closing

Dilation process that followed by erosion, the effect of filling a small hole in an object, combining objects that are close together and smoothing the boundary of a large object without significantly changing the object area [7].



$A \cdot S = (A \oplus S) \otimes S$

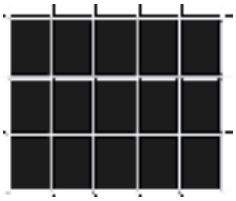


Figure 5. Before closing process

Figure 6. After closing process

2.8 Characteristics of geometry

Characteristic that based on the relationship between two points, lines, or fields in a digital image. Characteristics of geometry include distance and angle. The distance between two points (with pixel units) can be determined using the Euclidean equation [8].

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (4)$$

To distinguish the shape of an object, the metric parameter is used. Metric is the value of comparison between the area and circumference of an object. Metric has a range of values between 0 and 1. Objects that are in the form of elongated / approaching a straight line, the value of the metric is close to 0, while the object is round / circular, the value of the metric is close to 1[8].

$$M = \frac{4\pi A}{P^2} \quad (5)$$

A = Area (Number of pixels per row) and
P = Parameter (Number of pixels from the regional boundary).

2.9 ANN Perceptron

Perceptron changes the weight continuously until it finds a weight that is used to recognize patterns optimally. The perceptron activation function can be -1, 0 or 1, the input and target can be free and the threshold used is 0, with the determination of θ (threshold) as follows [9]:

$$f(net) = \begin{cases} 1, if\ net > \theta \\ 0, if\ -\theta \leq net \leq \theta \\ -1, if\ net < -\emptyset \end{cases} \quad (6)$$

3. Information Retrieval System

3.1 TF-IDF

Method Term Frequency Inverse Document Frequency is a way to give weight relationship of a term to a document. This method is combining two concepts for weight calculation, namely frequency of occurrence of a word in a document certain and inverse frequency documents containing words [10].

3.2 *Vector Space Model*

This method is an algebraic model that describes several text documents as vector identifiers. In this case it is used as information retrieval, indexing and giving the most relevant document ranking [11]. This method assumes that there are already index terms/words that represent documents and queries. TF value is the number of occurrences of a term/word in the document. While the IDF value is calculated by the formula [12]:

$$IDF = \log(\frac{D}{DF}) \quad (7)$$

D = Number of documents
DF = Document Frequency
The formula for weighting (*W*) is *W = tf x IDF*
Then calculate the distance of each document to the query (word).

$$\sqrt{Q} = \sqrt{\sum_{j=1}^{n} D_j^2} \quad (8)$$

Calculate dot product:

$$\sum Q * Di = \sum_{j=1}^{n} Q_j D_{i,j} \quad (9)$$

The next stage calculates similarities with the formula:

$$Cosine\ \theta\ D_i = \frac{Q*D}{\sqrt{Q}*\sqrt{D_i}} \quad (10)$$

The D value with the largest θ is the first rank and so on.

4. Framework Design

The design of this study consisted of 2 processes which were put together, beginning with preprocessing consisting of gray scaling, threshold, binarization, and proceeding with the normalization of the position and size of the image to be processed.

The first process is character recognition which includes: Segmentation, geometry extraction, and classification by ANN.

The next process is the information retrieval system which consists of: Indexing, retrieval and document ranking.

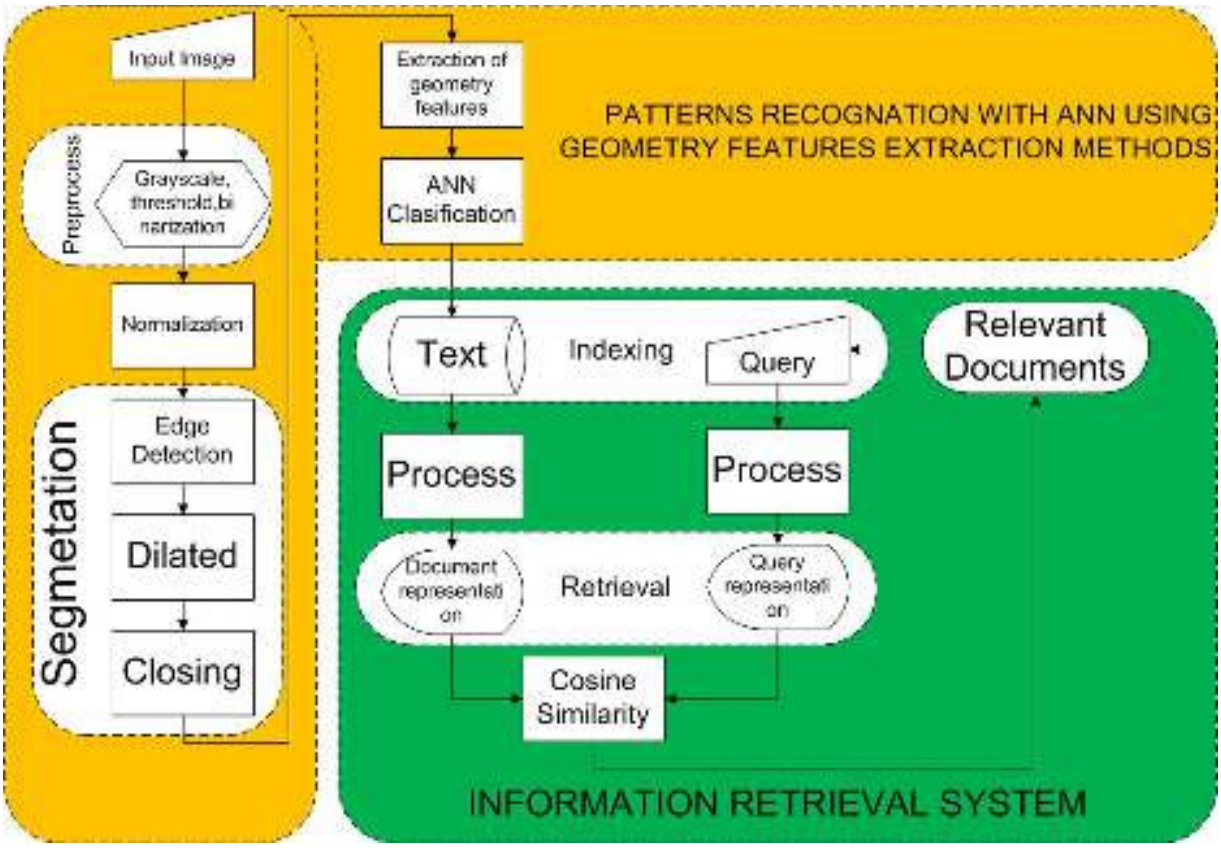


Figure 7. Framework design

The algorithms of the first system to be designed are as follows:

1. Input in the form of RGB official letter image
2. Preprocess, Image through Grayscale, Threshold, and Binaryization processes.
3. In the normalization stage, regional positions are determined *Tanggal, Nomor Surat, Lampiran* and *Hal.* In order for the letters to be read by ANN, the letters are normalized to become 10x12 pixel images.
4. The segmentation process of each object will go through the edge detection process, the dilation process is the thickening of the pixels so that the image becomes bigger than before, and the closing process is done which is filling in the blank space in detail in the image. The image generated at this stage is stored as a segmented file

5. The segmented file is extracted with geometry features and will generate numeric data that will be stored as an extracted file
6. Recognize the character pattern using ANN used data from the extracted file.
7. ANN will recognize the image according to matching training data from the segmented file to the extracted data file.
8. The output of the first system is text with regionalization based on position *Tanggal, Nomor Surat, Lampiran* and *Hal.*

The output from above algorithm is text that will use as an input to the information retrieval system (IRS) which will be used for the indexing process.

There are 2 processes that are run on IRS; indexing and retrieval.

Figure 8. Indexing Process

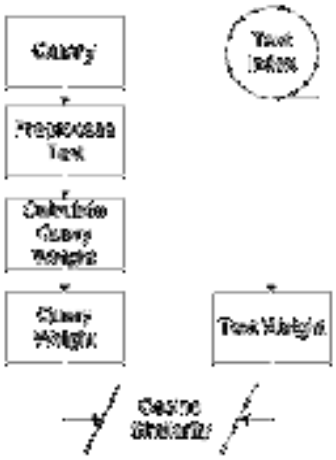


Figure 9. Retrieval Process

5. Discussion

The test results from the pattern recognition of handwriting done by Herviana Masrani [13] is quite good results, while the results of testing of letter/character recognition in the research conducted by Rizqia Lestika Atimi [1] is good enough results, and research on IRS conducted by D. Putung Karter [14] obtained good enough results too.

6. Conclusion and Recommendation

Framework design in this study is not tested yet, but the result of three studies before [13][1] [14] authors found high percentage with quite good results and as a suggestion for further research authors will develop this study to find the real result, otherwise authors create a new design that simpler to obtain percentage of success more higher.

References

[1] Atimi, Rizqia Lestika. "Pengenalan Karakter Pada Surat Masuk Menggunakan Neural Network Back propagation" *Jurnal Sistem dan Teknologi Informasi (JUSTIN)* 1.1 (2012): 1-6.

[2] Gonzalez, R., & Woods, R 2002 *Digital Image Processing (2nd ed.)* New Jersey: Prentice Hall

[3] Fanani, Aris, Anny Yuniarti, and Nanik Suciati. "Geometric Feature Extraction of Batik Image Using Cardinal Spline Curve Representation." *Telkomnika* 12.2 (2014): 397.

[4] Kadir, Abdul., dan Susanto, Andi. 2012 *Teori dan Aplikasi Pengolahan Citra* Yogyakarta: Andipublisher
[5] Gonzalez, R, & Wintz, P. 1987 *Digital Image Processing (2nd ed.)* Boston: Addision-Wesley
[6] Vilas H Gaidhane,et al. 2017 *An improved edge detection approach and its application in defect detection* in IOP Conf. Ser.: Mater. Sci. Eng. 244 012017
[7] Pengolahan Citra Digital: Morfologi Citra [online]. Available : http://staff.ui.ac.id/system/files/users/dodi.sudiana/material/kuliah08-morfologicitra.pdf [accessed on 10 July 2019]
[8] Ekstraksi Ciri Citra [online]. Available : https://pemrogramanmatlab.com/pengolahan-citra digital/ekstraksi-ciri-citra-digital/ [accessed on 9 July 2019]
[9] Siang, JJ. 2004 *Aplikasi Jaringan Syaraf Tiruan dan Pemrograman Menggunakan MATLAB* Yogyakarta: Andi
[10] Karmayasa, O. 2012 Implementasi Vector Space Model Metode Term Frequency Inverse Document Frequency (TF-IDF) Pada Sistem Temu Kembali Informasi in *Universitas Udayana Denpasar*
[11] Nadirman, F. 2006 Sistem Temu-kembali Informasi Dengan Metode Vector Space Model Pada Pencarian File Dokumen Berbasis Teks in *Skripsi Program Studi Ilmu Komputer Universitas Gadjah Mada, Yogyakarta*
[12] Manning, D. Christopher, Raghavan, P. & Schütze H. 2009 *An Introduction to Information Retrieval* Cambridge University Press
[13] Herviana Masrani, Ilhamsyah, Ikhwan Ruslianto 2018 Aplikasi Pengenalan Pola Pada Huruf Tulisan Tangan Menggunakan Jaringan Saraf Tiruan Dengan Metode Ekstraksi Fitur Geometri in *Jurnal Coding, Sistem Komputer Untan*
[14] Karter D. Putung, Arie Lumenta, Agustinus Jacobus 2016 Penerapan system Temu Kembali Informasi Pada Kumpulan Dokumen Skripsi in *E-journal Teknik Informatika*