Improvement of Services through Digital Documents Filing with The Template Matching Correlation Method

Teddy Setiady*, Nasril, Kuswandi, Setiawan, Hesti Rian

{Teddysetiady007@gmail.com}

Polytechnic LP3I Jakarta, Jakarta, Indonesia

Abstract. Scanning printed documents into digital documents has become a necessity in universities. Each document has a unique number as an identity with a certain code structure. The purpose of this research was to determine the effectiveness of document scanning results by utilizing an optical character recognition technique known as OCR (Optical Character Recognition) on this unique number. The method used in this optical character recognition process was to use template matching correlation,a technique to get the highest value in the comparison of characters in the input image. The image with the highest value is determined as the image that best fits the template. The technology developed in this study was Template Matching Correlation so that it can produce applications using MATLAB that can detect unique numbers in documents as a reference in file filing in the form of naming new files and placing the files in the desired folder.

Keywords: Document Imaging; Optical Character Recognition; Matching Correlation

1. Introduction

Scanning printed documents into digital documents has become a necessity for many institutions and companies and the Jakarta LP3I Polytechnic. One of the documents that are often transferred to the media is a diploma. With an average number of graduates reaching 1,600 per year, of course, many important documents must be transferred and filed properly, but research is still rare that discusses the automatic filing of digital files scanned by these documents.

Optical Character Recognition (OCR) using template matching correlation to recognize letters and numbers from an image is a method that is quite simple and easy to use. OCR research using the template matching correlation algorithm has been carried out to obtain a high success accuracy or about 99% for the vehicle number plate and can be achieved by assuming the condition of the image captured from a fixed distance, center position, angle parallel to the horizontal line, year of the vehicle plate. the same, as well as for certain motorized vehicles. (Rathore & Kumari,014:43-53).

BIC 2021, October 11-12, Batusangkar, Indonesia

DOI 10.4108/eai.11-10-2021.2319533

2. Methods

2.1 Analysis of Font Data Needs

The first step that must be prepared was a printed document containing an arrangement of uppercase, lowercase letters, and numbers. The font types used in this study were Arial, Times New Roman, Calibri, Tahoma, and Book Antiqua. The font shape was described below:

Figure 1. Needs Analysis of Character Templates

2.2 Hardware Requirements Analysis

Computer

To conduct this research, a computer that is compatible with MATLAB software and the scanner used was used.

Scanner

To perform a scan, a good scanner was needed so that it canproduce quality digital images according to the desired resolution.

2.3 Software Requirements Analysis

Scanner Software

This software was the default software from the scanner hardware that was used to scan paper containing letters into a digital image from 50 dpi to 12,800 dpi for both colors, gray and black and white types.

MATLAB R2017b

The software used in this thesis was MATLAB R2017b. This software can be utilized in terms of MATLAB's ability to quickly find solutions to various numerical problems.

2.4 Analysis Techniques

Correlation Analysis Techniques

How to calculate the correlation value using MATLAB using the Corr2(A, B) syntax, where A is the matrix of the template image and B is the matrix of the input image. The image with the highest correlation value is determined as the image that best fits the template.

3. Research Result

3.1 Hardware Selection

The printer needed was a DeskJet printer that can print in large quantities but with maintained quality, for that, the researchers chose the EPSON L565 printer. While the scanner needed was a scanner with an ADF type (Auto Document Feeder) to get a stable scan resolution and a faster process.

3.2 Software Selection

The software for scanning documents used the default hardware, namely the Kodak Scanmate i900 Series Scanner. As for the OCR process using MATLAB version 2017b x64.

3.3 Character Template Creation Process

Following the object of research, namely the diploma number or transcript using Arial letter numbers, the character template created was in the form of numbers as shown below.

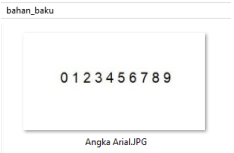

Figure 2. Number Character Template

Then using MATLAB, it was extracted into a binary image with the following working order:

a. Read image
b. Convert to gray image
c. Convert to binary image
d. Extract per letter to be like the image below:

Figure 4. Extract per letter

Next, convert the image into an array of matrix cells with the name templates. mat.

3.4 Application Program Algorithm Development

The working order of the applications made is as follows:

a. Reading the folder to be OCR
b. Command tool to run the OCR function in a loop according to the number of *.jpg files in the folder.

3.5 Document Print Process

The printed documents were transcripts of LP3I Jakarta Polytechnic students who graduated in 2019 from 4 Study, selected proportionally according to the number of graduates of each Study Program. The process for printing the document was as follows: a) Specify the bold font format for the diploma number on the Smart Campus application to print transcripts where the diploma number was the same as the transcript number without the prefix "T."; and b) Select Transcripts per Study Program proportionally from 1,487 transcripts to a total of 100 transcripts and save them in a PDF file, namely: 1) Business Administration as many as 53 students (AB-Arial-10-bold.pdf); 2) Computerized Accounting for 24 students (KA-Arial-10-bold.pdf); 3) Information Management as many as 21 students (MI-Arial-10-bold.pdf); and 4) Public Relations of 2 students (HM-Arial-10-bold.pdf); c) Prepare 70-gram HVS A4 paper and print only the first page using the EPSON L565 Printer; d) Repeat from points 1 to 3 with a regular font size of 10 on the diploma number and 12 on the transcript number with the name: (1) AB-Arial-10-12-Regular.pdf; (2) KA-Arial-10-12-Regular.pdf; (3) MI-Arial-10-12-Regular.pdf; (4) HM-Arial-10-12-Regular.pdf; e) Repeat from points 1 to 3 with bold font size 12 on the transcript number with the name: (1) AB-Arial-12-bold.pdf; (2) KA-Arial-12-bold.pdf; (3) MI-Arial-12-bold.pdf; (4) HM-Arial-12-bold.pdf

3.6 Document Scan Process

The document scanning process was carried out as follows:

a. Setting the Kodak ScanMate i940 Scanner with the following conditions:
   1) Prefix ar10b for Arial Font, Size 10, Type Bold
   2) File Type: JPEG
   3) Scan as Color Perfect Document
b. 5x ADF scanning per 20 sheets for a total of 100 sheets
c. Repeat from points 1 and 2 for Arial Font, Size 10, Regular Type (incl. size 12)
d. Repeat from points 1 and 2 for Arial Font, Size 12, Type Bold.

3.7 OCR (Optical Character Recognition) Process

It was necessary to test the accuracy of the coordinates of the location of the diploma number, if the coordinates are wrong, then correct it first by looking at it with the imshow () command. If true it will produce text like in the image below:

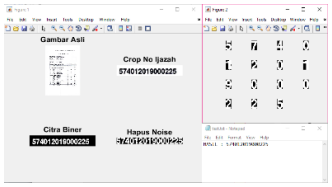

Figure 5. Results of the OCR Process Text

3.8 OCR Process Success Rate Tabulation

Through an application designed using MATLAB 2017b x64 with the template matching correlation method, all transcript files were subjected to an OCR (Optical Character Recognition) process for the diploma number (size 10) and a fragment of the transcript number (size 12) whose number was the same as the diploma number. For the research to focus more on the part being studied, the template used only numbers from the Arial font. The captured part of the transcript was with the following conditions: (1) Font Arial, Size 10, Type Bold; (2) Font Arial, Size 12Type Bold; (3) Font Arial, Size 10, Type Regular; (4) Font Arial, Size 12, Type Regular. The recognition results from the scanned digital files were arranged in the form of tabulations with the following results:

Table 1. Recognition results from digital files

Huruf Arial, Size 10, Type Bold

PROGRAM STUDI	JML FILE	BENAR	SALAH	%
Administrasi Bisnis	53	47	6	89%
Komputerisasi Akuntansi	24	22	2	92%
Manajemen Informatika	21	18	3	86%
Hubungan Masyarakat	2	2	0	100%
	100	89	11	**89%**

One of the failures in the OCR process with Arial, Size 10, and Type Bold letters was that there were double numbers that were read as one number, for example in the transcript of the Business Administration Study Program with diploma number 634112019000131 the OCR Data shows that the number 900 was read as number 1 so that the diploma number carried out by the OCR was considered as 6341120110131 so that with a trial of 100 files, the success of 89% was achieved.

Table 2. Test results 1

Huruf Arial, Size 12, Type Bold

PROGRAM STUDI	JML FILE	BENAR	SALAH	%
Administrasi Bisnis	53	34	19	64%
Komputerisasi Akuntansi	24	22	2	92%
Manajemen Informatika	21	20	1	95%
Hubungan Masyarakat	2	1	1	50%
	100	77	23	**77%**

Similar to the previous test, one of the causes of failure with Arial, Size 12, Type Bold letters were that there were double numbers that were read as one number, 48 was considered as number 4 so that with a trial of 100 files, the success of 77% was achieved. An explanation of this can be seen in the OCR process below:

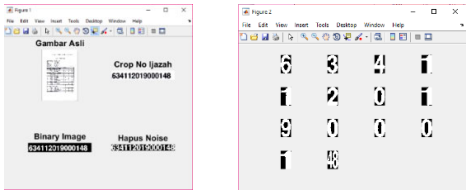

Figure 6. OCR Process 2

Table 3. Test results 2

Huruf Arial, Size 10, Type Reguler

PROGRAM STUDI	JML FILE	BENAR	SALAH	%
Administrasi Bisnis	53	53	0	100%
Komputerisasi Akuntansi	24	24	0	100%
Manajemen Informatika	21	21	0	100%
Hubungan Masyarakat	2	2	0	100%
	100	100	0	**100%**

By using Arial, Size 10, Regular Type, there was an increase in success with the same file, which is 100%.

Table 4. Test results 3

Huruf Arial, Size 12, Type Reguler

PROGRAM STUDI	JML FILE	BENAR	SALAH	%
Administrasi Bisnis	53	53	0	100%
Komputerisasi Akuntansi	24	24	0	100%
Manajemen Informatika	21	21	0	100%
Hubungan Masyarakat	2	2	0	100%
	100	100	0	**100%**

Likewise, by using Arial, Size 12, Regular Type letters, it achieved 100% success.

4. Conclusion

The result of this research is that an application prototype using MATLAB can detect a unique number in the document as a reference in file filing in the form of naming a new file and placing the file in the desired folder. Based on the results of this study, it is known that the effectiveness of digital document filing can be increased by using a computer application that uses the template matching correlation method with an average accuracy of 91.5%. File accuracy can be investigated further by using different templates or other methods.

References

[1]. Adhvaryu, Rachit Virendra. “Optical Character Recognition Using Template Matching (Alphabet& Numbers).” International Journal of Computer Science Engineering and InformationTechnology Research (IJCSEITR) 3, no. 4 (2013): 227-232.

[2]. Chandarana, Jagruti, dan Mayank Kapadia. “Optical Character Recognition.” InternationalJournal of Emerging Technology and Advanced Engineering (UKA TARSADIA University) 4,no. 5 (May 2014): 219-223.

[3]. Patil, Jatin M, dan Ashok P Mane. “Multi Font And Size Optical Character Recognition UsingTemplate Matching.” International Journal of Emerging Technology and Advanced Engineering3, no. 1 (2013): 504-506.

[4]. Rathore, Manisha, dan Saroj Kumari. "Tracking Number Plate From Vehicle Using MATLAB.”International Journal in Foundation of Computer Science & Technology (IJFCST) 4, no. 3 (May2014): 43-53.

[5]. Rustanto, A.E. Analysis of Perceived Benefits, Convenience, and Risk to the Effectiveness of Non-Cash Payments. *IJISRT*. 682-886

[6]. Setiady, Teddy. “Analisa Batas Sudut Kemiringan Hasil Pemindaian Dokumen MenggunakanTemplate Matching Correlation.” Jurnal Lentera ICT Vol 3 No.1 (Mei 2016): 112-130.