Jurnal Elektronik Ilmu Komputer Udayana
Volume 8, No 2. November 2019

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Segmentation of Certificate with Connected Component Labeling Method

Cokorda Gde Teresna Jaya[a1], I Gede Arta Wibawa[a2]

[a]Informatics Department, Mathematics and Science Faculty, Udayana University
Jalan Raya Kampus Unud, Bukit, Jimbaran, Bali, Indonesia
[1]cokordagedetresna@gmail.com
[2]gede.artha@unud.ac.id

Abstract

Certificate is one of the documents that can be used as evidence of ownership or an event. For example, when certificate used as requirement to participate in an event. If a document is made as a requirement, of course the file verification process will be done. Seeing the time optimization problem when verifying the file, the authors carry out research by segmenting important data contained in a certificate as an initial step in the development of an automatic document verification system. The segmentation process carried out in this study uses the Connected Component Labeling method in determining the area to be segmented and Automatic Cropping to cut the results of the segmentation process. By using these two methods obtained an accuracy of 60% with a total of 15 pieces of test data

Keywords: *Certificate, Document, Segmentation, Connected Component Labeling, Automatic Cropping*

1. Introduction

Documents are something that can never be separated from human life. Documents can be used as evidence to support information to be more convincing. One of the documents that can be used as evidence is a certificate.

Certificate is a written or printed statement or certificate from an authorized person that can be used as evidence of ownership or an event [1]. This document can be used as a condition for participating in an activity or event. the case that the authors found is that the certificate itself is used as a condition in the registration of an activity at Educational Institution such as University. Sample case that the authors found were the function of certificate documents as a requirement in registering the Final Project Workshop in Department of Computer Science, Faculty of Mathematics and Natural Sciences, Udayana University.

In practice, the certificate documents are collected and then manually checked again to ensure the requirements collected are correct. But the authors see a problem in the process of checking that is less efficient and the potential for human error, so that a more optimal way of checking is needed by retrieving data from the certificate and verifying it automatically.

In the process of making this automatic checking system, it takes several stages. One of them is the regional segmentation stage of important data in the certificate such as the data of the name of the activity and the data of the name of the certificate owner so that the next step can be a recognition to character from the results of the segmentation. Segmentation stage is an important area of this certificate which will be carried out by the authors in this study.

There are several studies that have been made previously related to segmentation. One of them is the recognition of handwritten Arabic numerals (Indian). To recognize the characters, the Connected Component Labeling is used in the segmentation process of the handwriting so that the letters contained in the handwriting can be separated into non-connected characters. Then the recognition of numbers is done by using the K-Nearest Neighbors method by finding out how big is the match between 100 test images and training images based on class to the nearest neighbor. The results of this study have an accuracy of 86% when the value of k = 1, 84% when the value of k = 3, and 83%

when the value of k = 5 [2]. In addition there are also studies that apply Optical Character Recognition (OCR) for PLN electricity meter readings. The segmentation process in the study used Connected Component Labeling and Template Matching for character recognition [3].

In this research, there are several steps that are carried out to do segmentation. The steps taken begin with the process of data acquisition, pre-processing, certificate region detection, and certificate region segmentation.

At the data acquisition stage digitization is performed on the certificate data which is then continued with the pre-processing stage which consists of grayscaling and binarization processes. After changing to binary image, the certificate area is detected using the *Connected Component Labeling* method by drawing a marginal line and determining the area to be segmented. And the last process that is done is cropping the area to be segmented

2. Reseach Methods

2.1 Data Acquisition

The certificate image was obtained from Computer Science students at Udayana University. The certificate has a predetermined format and was acquired using the scanner. The results of the acquisition have image specifications of RGB with 600 x 800 px resolution and * .png file format. The amount of data tested in the following studies is 15 certificates.

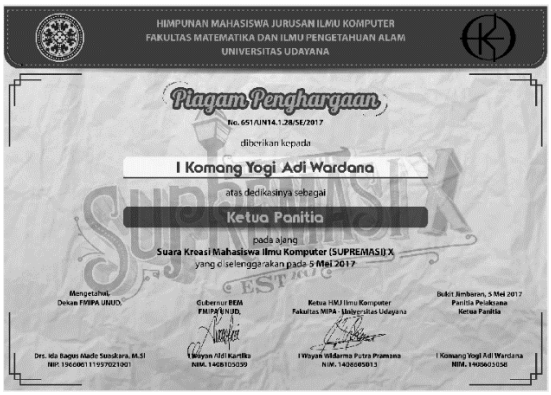
Figure 1. RGB image

2.2 Pre-Processing

2.2.1 Grayscaling

The image produced after data acquisition is in the form of a color image. So the grayscale process is needed to change the color image into an image with a gray level. The function of the grayscaling process is to change the image that originally had 3 layers to 1 layer so that the image will be simpler [4] . In the grayscaling process the following equation is used:

$$Gray = \frac{(\mathrm{Re}\, d + Green + Blue)}{3} \quad (1)$$



p-ISSN: 2301-5373
e-ISSN: 2654-5101

Figure 2. Grayscale Image

2.2.2 Binarization

After the image is converted into an image with a gray level then proceed with the process of binarization by changing the grayscale image into a binary image.

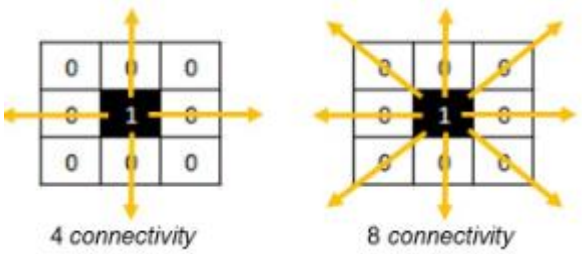
Figure 3. Binary Image

2.3 Certificate Region Detection

At this stage a section is detected in the area containing important data on a certificate. There are 2 stages in the process of detecting parts, namely the process of making marginal lines and the process of determining the part to be segmented.

In the process of making marginal lines, labeling using the Connected Component Labeling method. The process of Connected Component Labeling is done by scanning per line [5]. Here are the steps:

- Create a variable with an initial value of 0
- Search per pixel line to determine whether a pixel is 0 (white pixels) or 1 (black pixels)
- If the pixel is 0, then mark the pixel with a value of 0.
- If found a pixel with value 1 that is not connected with its neighbors (using 4 connectivity or 8 connectivity). Then increment the variable value and mark the pixel according to the variable value
- If found a pixel of value 1 that is connected with its neighbors (Using 4 connectivity or 8 connectivity). Then increment the value of the variable and mark the pixel according to the value of the smallest neighbor's sign
- Sometimes in scanning a conflict occurs when a pixel has a neighbor with two different signs. Then mark the pixel with the smallest sign value and record the relation data equivalent
- After all pixel rows have been scanned and all pixel conflicts found, re-mark them by combining the marks according to the equivalent relation data.



For marginal lines to form, the y-axis of each first and last pixel labeled on each character will be recorded. Then proceed with the stage of determining the part to be segmented by means of the y coordinate of the selected marginal line the x-axis value will be checked at each end of the sentence

Figure 4. Illustration of Marginal Lines

2.4 Certificate Area Segmentation

At this stage of segmentation, an auto cropping process is carried out which takes 2 coordinates, namely the initial coordinates which are the initial coordinates for the cut image and the final coordinate which is the final coordinate point of the cut image. So it forms a rectangular shape where each pixel is in a certain coordinate area and stored in a new image

Figure 5. Segmentation Results

3. Result and Discussion

In this study, several test results were obtained from 15 test data including the following:

Table 1. Result

No Data Sertifikat	Data ke – n	Status
1	Data 1	Success
2	Data 2	Failed
3	Data 3	Success
4	Data 4	Failed
5	Data 5	Success
6	Data 6	Success
7	Data 7	Success
8	Data 8	Failed
9	Data 9	Failed
10	Data 10	Success
11	Data 11	Success
12	Data 12	Failed
13	Data 13	Failed
14	Data 14	Success
15	Data 15	Success

Then testing using the blackbox testing technique with accuracy calculation using the following formula:

$$Accuracy = \frac{TotalSuccess\ \mathrm{Re}\ sult}{TotalData} x100\% \quad (2)$$

From the above test results obtained as many as 9 certificates correctly so that an accuracy of 60% is obtained.

4. Conclusion

From the results of the study found that the segmentation of important areas on the certificate can be done with the condition that the certificate data has similarities with the format that has been determined. The similarity here is seen from the marginal lines because the sequence of marginal lines determines which part of the cropping will be done.

References

[1] Tim Penyusun, K. B. B. I. (2008). Kamus Besar Bahasa Indonesia. *Balai Pustaka: Jakarta*.
[2] Gunawan, R., Suwarno, S., & Hapsari, W. (2014). Penerapan Optical Character Recognition (Ocr) Untuk Pembacaan Meteran Listrik Pln. Informatika: Jurnal Teknologi Komputer dan Informatika, 10(2).
[3] Akbar, R., & Sarwoko, E. A. (2016). Studi Analisis Pengenalan Pola Tulisan Tangan Angka Arabic (Indian) menggunakan Metode K-Nearest Neighbors dan Connected Component Labeling. Dinamika Rekayasa, 12(2), 45-51.
[4] Isnanto, R. R., & Zahra, A. A. (2014). Pengenalan Plat Kendaraan Secara Waktu Nyata Menggunakan Framework Aforge. Net. TRANSIENT, 3(2), 262-269.
[5] Bollman, J. E., Rao, R. L., Venable, D. L., & Eschbach, R. (1999). *U.S. Patent No. 5,978,519*. Washington, DC: U.S. Patent and Trademark Office.
[6] Putra, E. D., & Santosa, S. (2017). Optimasi Kemampuan Segmentasi Otsu Pada Identifikasi Plat Nomor Kendaraan Indonesia Menggunakan Metode Gaussian. *Pseudocode*, *4*(1), 47-60.