CHARACTER RECOGNITION AND INFORMATION EXTRACTION ON IMAGE OF ABSTRAK SKRIPSI USING SUPPORT VECTOR MACHINE AND RULES BASED SYSTEM

Muhammad Fajry Hamzah[1], Galih Hermawan[2]

[1,2] Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur No 112-116 Bandung, 40132
E-mail : fajryhamzah@gmail.com[1], galih.hermawan@email.unikom.ac.id[2]

ABSTRACT

Technological advances have influenced various aspects of human life, for example, the management of information. Most people nowadays prefer to store their information digitally because it is provide convenience in accessing and manipulating the information.

One example that needed to be retrieved is a skripsi report. Complete information about skripsi can actually be obtained by the abstract page only. But sometimes, the skripsi abstract that is available only has a hard copy, so the librarian must enter the identity of a document by filling in the necessary data into the system. The Support Vector Machine method can be used to convert the text in the images into digital character so that later it can be recognized by each section in the thesis abstract by using Rule Based method.

The result of this study obtained an accuracy of 5,47% for the text recognition on the skripsi abstract images. As for the accuracy of the character recognition using SVM itself, it reaches 54.30%. The low result of recognition accuracy in abstract images is influenced by the segmentation process that is less able to solve existing problems in character segmentation. As for categorizing information using Rule Based it reaches an accuracy of 99%.

Keywords : Support Vector Machine, Rule Based, Skripsi abstract, Text recognition on images, Information Extraction

1. INTRODUCTION

Technological advances have influenced various aspects of human life, for example, the management of information. Most people nowadays prefer to store their information digitally because it is provide convenience in accessing and manipulating the information. One example of document that needed to be retrieved is a skripsi report. Complete information about skripsi can actually be obtained by the abstract page only, for example, title of the report, author name, author NIM number, the content, and the keywords of the report, But sometimes, for the old skripsi documents, the skripsi abstract that is available only has a hard copy, so the librarian must enter the identity of a document by filling in the necessary data into the system [1]. This will become a problem when the document that needed to be converted comes in large quantity so that it will take time and effort to convert the document. In addition, human error can occur due to the process of converting and categorizing information. These problems can be handled by changing the information form from hard copy to soft copy in the form of digital text and obtaining the information in the document automatically so that the problem can be minimized.

Research about text recognition has been done a lot, but there is rarely research using a full document as an input of the research. There are several methods that can be used to recognize written characters in an image [2], such as Artificial Neural Network (ANN) [3], K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) [4]. Previous studies tried to compare the performance of ANN and SVM [5] with the best accuracy obtained from the SVM method with a result of 94.43%, other studies also tried to compare the KNN and SVM method [6] with SVM is getting the best accuracy of 93.13%. So it can be concluded that the Support Vector Machine method classifies characters better so that makes the reason why this study will use the SVM method to recognize the written characters in the abstract image of the skripsi. In addition, research on the extraction of information in skripsi reports, in particular the skripsi report of University of Computer Indonesia has been done, with the Rules Based method the results obtained are very good [1] so the same method will be used in this study with slightly different rules due to the difference in input used in the research. The rules that are used in this study will try to minimize the information extraction errors due to the poor results of the text recognition.

2. RESEARCH CONTENT

2.1 Research Method

This research has 8 stages of research workflow, The stages are identification of problems, determining objectives and boundaries, collecting data, analyzing the system to be built, implementing

the system, testing and finally making conclusions and suggestions for the next research. The following image below is the flow of the stages of the research.

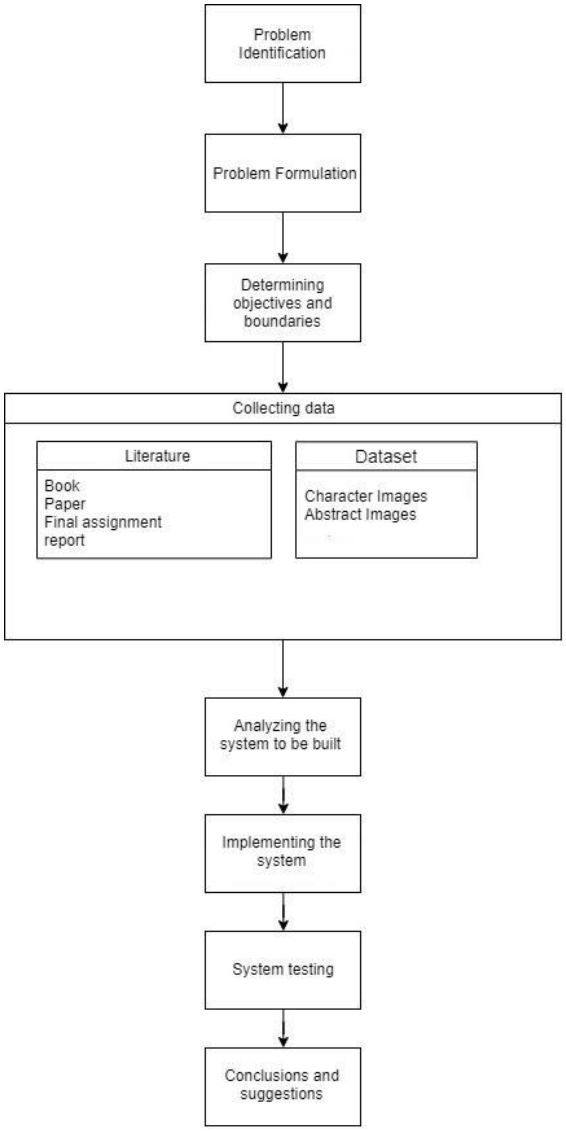


Figure 1. Research Workflow

2.2 System Overview

The system built is divided into two sections, The training section and the testing section. The figure below is an overview of the system:

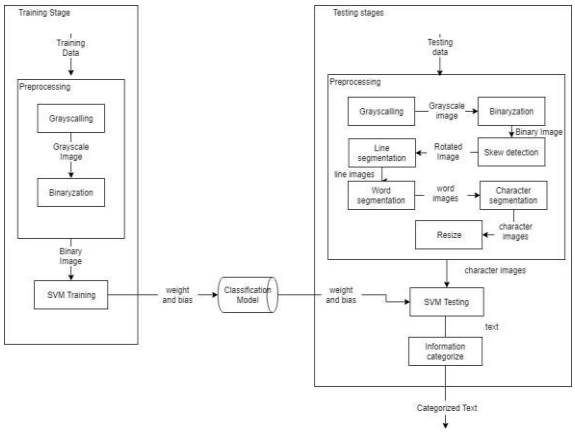


Figure 2. System Overview

These two sections have different processes. In the training section, training data is an image of each alphabet character, number and symbol character that often appears on the skripsi abstract page. There is also a preprocessing process that contains grayscalling and binarization, followed by the SVM model training which later the result will be stored in the storage media. The second section is the testing phase, the test data is the image of the skripsi abstract page that can be obtained by scanning or photos. There are several processes in this section, including the preprocessing process that contains the process of grayscalling, binarization, skew correction, line segmentation, word segmentation, character segmentation and resizing. After the preprocessing process is done, the characters that have been segmented and resized will become an input for the SVM testing using the SVM model that has been trained at the training stage. After all characters are recognized in the SVM testing process, the next process is extracting the information into predetermined categories, such as the title of the skripsi, the author's name, the author's NIM, abstract content and keywords in the abstract.

2.3 Training Data

The training data used in this study are the image of characters with Times New Roman fonts in the form of alphabets, numbers and some symbol characters that often appear on the abstract pages with variations such as rotation and blur so that the total data used in the training is 300 data. The following figure is an example of the training data used in this study:

Figure 3. Training Data

2.4 Grayscale

The grayscale process changes the color mode of the image into grayscale mode. The following figure is the formula used in the grayscale process.

$$Y' = 0{,}299R' + 0{,}587G' + 0{,}114B' \quad (1)$$

All pixels in the image will be calculated using the formula above, so that the pixel value that used to have more than 1 color channel changes to 1 color channel with values in the range of 2^8.

2.5 Binaryzation

Binaryzation method used in this research is Bradley-Roth Adaptive Thresholding [7]. The Bradley-Roth method uses an integral image. Integral image is the result of the sum of the specified area, this technique is able to speed up in getting the results of the calculation of the threshold. Below is a block diagram of the binaryzation process:

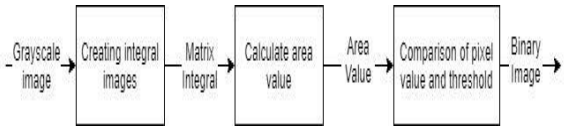


Figure 4. Block Diagram of Binaryzation

Process of making an integral images is done by the following equation:

$$I(x,y) = \sum_{\substack{x' \le x \\ y' \le y}} i(x',j') \quad (2)$$

Then the area value will be calculated from the region as figure and equation follows.

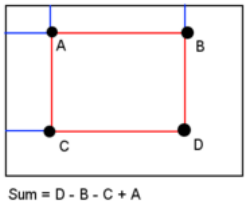


Figure 5. Reference Area

$$i(x,y) = I(D) - I(B) - I(C) + I(A) \quad (3)$$

So that later the value of the area will be compared with the pixel value with a predetermined percentage, if the calculation meets the requirements then the pixel will be changed to black pixels. The following comparison equation between pixel values and color intensity in the area.

$$(pixel_{value} * pixel_{inthearea}) < \left(sum * \frac{100-15}{100}\right) \quad (4)$$

2.6 SVM Training

Support Vector Machine (SVM) is a learning system that uses a hypothetical space in the form of linear functions in a high-dimensional feature [8] and trained using learning algorithms that are based on optimization theory. The SVM training process is done to find the vector α, w value and constant b to get the best hyperplane. In the training process, a set of input-output data is needed or in this case an image of the written character and the name of the character is needed. Below following the process flow in the SVM training section.

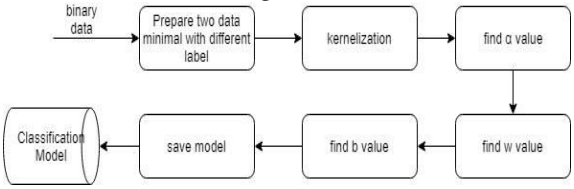


Figure 6. Flow of SVM Training

The binaryized image will then be transformed into a vector shape which will later be used in SVM training or testing. The format used is as follows:

[pixel_value_1,pixel_value_2,...,pixel_value_n]

SVM is one of the algorithms in supervised learning, therefore we need one more vector containing the class name of the vector to be trained. The sum of class vectors (y) is = N of the training data used. SVM training requires a minimum of 2 training data with different classes. The first step in training is kernel input vectorization. The kernel used can be various kinds, for example the Linear kernel. The following below is the equation of the linear kernel:.

$$K(x_i, x) = x_i^T x \quad (5)$$

The next step is to look for the values of w and b but first the alpha value need to be found first. The value of the alpha that is converted using the Lagrange Multiplier can be solved using Quadratic Programming by solving the equation in the dual form. Solving Quadratic Programming problems can be done using optimization algorithms such as Sequential Minimal Optimization (SMO) [9]. The alpha value can be found using the equation below:

$$a_2^{new} = a_2 + \frac{y_2(E_1 - E_2)}{n} \quad (6)$$

Where the value of n and E can be found using below equation:

$$n = K(x1,x1) + K(x2,x2) - 2K(x1,x2) \quad (7)$$
$$E_i = w.X + b = -1 \quad (8)$$
$$E_2 = w.X + b = 1 \quad (9)$$

In SMO, The completion performed on two alpha in one iteration and thus require a limit of boundary. If y1 is not the same as y2 then the boundary equation is as follows:

$$L = \max(0, a_2 - a_1) \quad (10)$$
$$H = \min(C, C + a_2 - a_1) \quad (11)$$

And if y1 = y2, then the boundary equation will be:

$$L = \max(0, a_2 + a_1 - C) \quad (12)$$
$$H = \min(C, a_2 + a_1) \quad (13)$$

After the selected alpha has been found, then the value needs to be checked to make sure that the alpha is still in the boundary. Below is the equation for checking the alpha.

$$\alpha_2^{\text{new,clipped}} = \begin{cases} H & \text{if} \quad \alpha_2^{\text{new}} \geq H; \\ \alpha_2^{\text{new}} & \text{if} \quad L < \alpha_2^{\text{new}} < H; \\ \text{L} & \text{if} \quad \alpha_2^{\text{new}} \leq L. \end{cases}$$

Then the second alpha can be found using the equation below:

$$a_1^{new} = a_1 + s(a_2 - a^{newclipped}{}_2) \quad (14)$$

After all the value of support vector (alpha) optimized, then the value of w can be obtained using the equation below.

$$w = \sum_{i=1}^{n} a_i y_i x_i \quad (15)$$

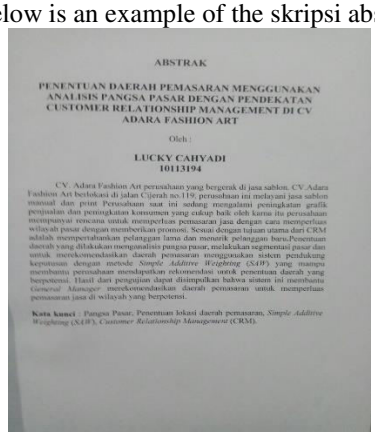
And the value of b is obtained by substituting w, x and y in the following equation.

$$b = y - w.x \quad (16)$$

The above process will be carried out on all training characters using the One vs One approach so that in this study, the number of models that will be created is 75 (75 −1) / 2 = 2775 models.

2.7 Testing Data

On this stage, skripsi abstract will be used. The figure below is an example of the skripsi abstract.



ABSTRAK

PENENTUAN DAERAH PEMASARAN MENGGUNAKAN ANALISIS PANGSA PASAR DENGAN PENDEKATAN CUSTOMER RELATIONSHIP MANAGEMENT DI CV ADARA FASHION ART

Oleh :

LUCKY CAHYADI
10113194

CV. Adara Fashion Art perusahaan yang bergerak di jasa sablon. CV.Adara Fashion Art berlokasi di jalan Cijerah no.119, perusahaan ini melayani jasa sablon manual dan print Perusahaan saat ini sedang mengalami peningkatan grafik penjualan dan peningkatan konsumen yang cukup baik oleh karna itu perusahaan mempunyai rencana untuk memperluas pemasaran jasa dengan cara memperluas wilayah pasar dengan memberikan promosi. Sesuai dengan tujuan utama dari CRM adalah mempertahankan pelanggan lama dan menarik pelanggan baru,Penentuan daerah yang dilakukan menganalisis pangsa pasar, melakukan segmentasi pasar dan untuk merekomendasikan daerah pemasaran menggunakan sistem pendukung keputusan dengan metode *Simple Additive Weighting (SAW)* yang mampu membantu perusahaan mendapatkan rekomendasi untuk penentuan daerah yang berpotensi. Hasil dari pengujian dapat disimpulkan bahwa sistem ini membantu *General Manager* merekomendasikan daerah pemasaran untuk memperluas pemasaran jasa di wilayah yang berpotensi.

Kata kunci : Pangsa Pasar, Penentuan lokasi daerah pemasaran, *Simple Additive Weighting (SAW), Customer Relationship Management* (CRM).

Figure 7. Abstract Image

Abstract images will be converted to grayscale and binarized using the same method at the training stage.

2.8 Skew Correction

Skew correction is a process whereby changing the image with rotation to improve the type of image that has a tilt. The slope of the image can result in the detection of writing on the image being difficult to recognize. Detection and correction of the image tilt level using the horizontal profile projection method where correction will be tried with various angles and angles with the maximum amount of energy function that will be used as the latest image angle [10]. The following figure is the flow that used in this process:

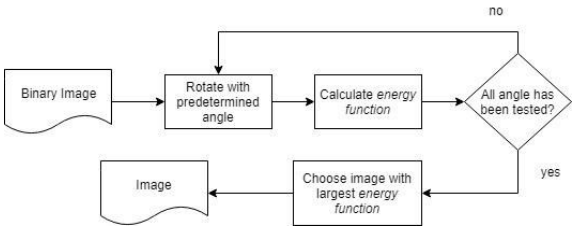


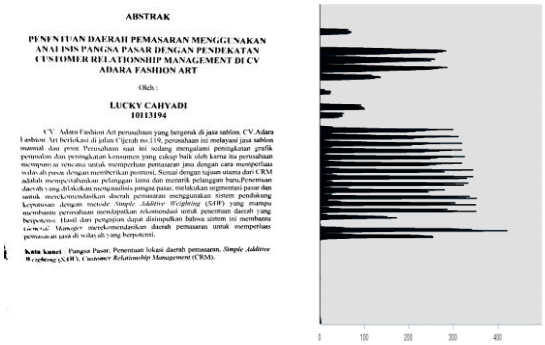
Figure 8. flow of Skew Correction

The following is the equation to calculate the energy function value:

$$h = \left(\text{area}_{\text{top}} - \text{area}_{\text{bottom}}\right)^2 \quad (17)$$

2.9 Line Segmentation

The next step is to do the Line Segmentation process which is used to separate the text of each line. The method used in Line Segmentation is the Horizontal Profile Projection method that works by counting the number of black pixels in an image horizontally (x axis) [11]. Below is example figure of pixel projection on abstract image:



ABSTRAK

PENENTUAN DAERAH PEMASARAN MENGGUNAKAN ANALISIS PANGSA PASAR DENGAN PENDEKATAN CUSTOMER RELATIONSHIP MANAGEMENT DI CV ADARA FASHION ART

Oleh :

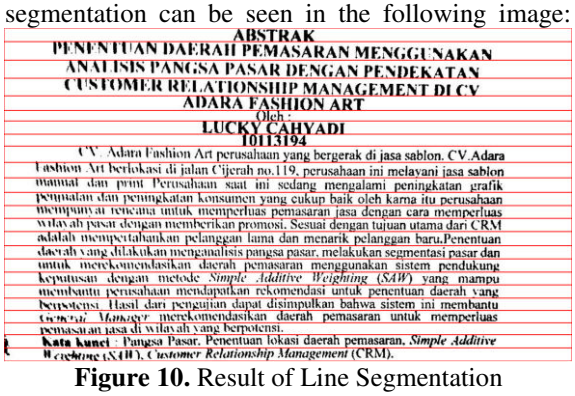
LUCKY CAHYADI
10113194

CV. Adara Fashion Art perusahaan yang bergerak di jasa sablon. CV.Adara Fashion Art berlokasi di jalan Cijerah no.119, perusahaan ini melayani jasa sablon manual dan print Perusahaan saat ini sedang mengalami peningkatan grafik penjualan dan peningkatan konsumen yang cukup baik oleh karna itu perusahaan mempunyai rencana untuk memperluas pemasaran jasa dengan cara memperluas wilayah pasar dengan memberikan promosi. Sesuai dengan tujuan utama dari CRM adalah mempertahankan pelanggan lama dan menarik pelanggan baru,Penentuan daerah yang dilakukan menganalisis pangsa pasar, melakukan segmentasi pasar dan untuk merekomendasikan daerah pemasaran menggunakan sistem pendukung keputusan dengan metode *Simple Additive Weighting (SAW)* yang mampu membantu perusahaan mendapatkan rekomendasi untuk penentuan daerah yang berpotensi. Hasil dari pengujian dapat disimpulkan bahwa sistem ini membantu *General Manager* merekomendasikan daerah pemasaran untuk memperluas pemasaran jasa di wilayah yang berpotensi.

Kata kunci : Pangsa Pasar, Penentuan lokasi daerah pemasaran, *Simple Additive Weighting (SAW), Customer Relationship Management* (CRM).

Figure 9. Projection of Black Pixel Horizontally

Cutting line image is done by looking at the empty gaps in each line so that the results of line segmentation can be seen in the following image:



ABSTRAK

PENENTUAN DAERAH PEMASARAN MENGGUNAKAN ANALISIS PANGSA PASAR DENGAN PENDEKATAN CUSTOMER RELATIONSHIP MANAGEMENT DI CV ADARA FASHION ART

Oleh :

LUCKY CAHYADI
10113194

CV. Adara Fashion Art perusahaan yang bergerak di jasa sablon. CV.Adara Fashion Art berlokasi di jalan Cijerah no.119, perusahaan ini melayani jasa sablon manual dan print Perusahaan saat ini sedang mengalami peningkatan grafik penjualan dan peningkatan konsumen yang cukup baik oleh karna itu perusahaan mempunyai rencana untuk memperluas pemasaran jasa dengan cara memperluas wilayah pasar dengan memberikan promosi. Sesuai dengan tujuan utama dari CRM adalah mempertahankan pelanggan lama dan menarik pelanggan baru,Penentuan daerah yang dilakukan menganalisis pangsa pasar, melakukan segmentasi pasar dan untuk merekomendasikan daerah pemasaran menggunakan sistem pendukung keputusan dengan metode *Simple Additive Weighting (SAW)* yang mampu membantu perusahaan mendapatkan rekomendasi untuk penentuan daerah yang berpotensi. Hasil dari pengujian dapat disimpulkan bahwa sistem ini membantu *General Manager* merekomendasikan daerah pemasaran untuk memperluas pemasaran jasa di wilayah yang berpotensi.

Kata kunci : Pangsa Pasar, Penentuan lokasi daerah pemasaran, *Simple Additive Weighting (SAW), Customer Relationship Management* (CRM).

Figure 10. Result of Line Segmentation

2.10 Word Segmentation

After each line has been segmented, the next step is to do the word segmentation to separate each word in each line that has been found. The method used to separate each word is to use the Vertical Profile Projection method. The method used is almost the same as the method for finding the line

except that the vertical profile projection of the number of pixels is done vertically or in the y-axis. The following are examples of projections on word segmentation:

Figure 11. Projection of Word Segmentation

The threshold range used for column separators is calculated from 1/3 of the height of the image. The value is obtained from previous studies [12] but the value will not be used if the value exceeds the maximum spaced value in the image, if that is so, the value used is 80% of the maximum spaced value in the line image. Here are the results of the word segmentation process:

Figure 12. Result of Word Segmentation

2.11 Character Segmentation

Character segmentation is done to separate each character in the image after word segmentation. The segmentation performed for each character is carried out using the same method as word segmentation, which counts the number of black pixels on the y axis (vertical). Separation for each character, done by utilizing the gap in each word. Following are the results of the character segmentation process:

Figure 13. Result of Character Segmentation

The result of segmented character will be represented as an matrix as follows:

1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1
1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1
1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1
1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1
1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1
1	1	1	1	1	0	0	0	1	0	0	0	0	1	1	1
1	1	1	1	1	0	0	1	1	0	0	0	0	0	1	1
1	1	1	1	1	0	0	1	1	1	0	0	0	0	1	1
1	1	1	1	0	0	1	1	1	1	0	0	0	0	0	1
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
1	1	1	0	0	1	1	0	1	1	1	0	0	0	0	1
1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0
1	1	0	0	1	1	1	1	1	1	1	1	0	0	0	0
1	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0

Figure 14. Character 'A' Matrix

2.12 Resize

The results of character segmentation will produced in different sizes. Therefore, we need a method to change the size of the segmented image into the same size. Resizing is done to adjust the image size used for the classifier training. The method used for resizing is to use the Nearest Neighbor method. Nearest Neighbor is the simplest and fastest method of interpolation by moving empty space with adjacent pixels (the nearest neighboring pixel) when reducing or enlarging the image scale [13]. Following figure is an example of resizing using 15x15px as the size.

1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
1	1	1	1	1	1	1	1	0	0	0	1	1	1	1
1	1	1	1	1	1	1	0	0	0	0	1	1	1	1
1	1	1	1	1	1	0	0	0	0	0	1	1	1	1
1	1	1	1	1	1	0	0	0	0	0	1	1	1	1
1	1	1	1	1	0	0	0	0	0	0	1	1	1	1
1	1	1	1	1	0	0	0	0	0	0	0	1	1	1
1	1	1	1	0	0	1	1	0	0	0	0	0	1	1
1	1	1	1	0	0	1	1	1	0	0	0	0	1	1
1	1	1	1	0	1	1	1	1	0	0	0	0	0	1
1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
1	1	1	0	0	0	0	0	0	0	0	0	0	0	1
1	1	1	0	1	1	0	1	1	1	0	0	0	0	1
1	1	0	0	1	1	1	1	1	1	1	0	0	0	0
1	0	0	0	1	1	1	1	1	1	1	0	0	0	0

Figure 15. Matrix Result of Resize

2.13 SVM Testing

After the characters have been segmented, the next step is the classification stage. Classification is done to recognize what character is in the image. The value of the segmented image will be converted into an input vector format to the SVM (X) model that has been trained. The vector format used consists of NxN depending on the size of the image when the model is trained and the value taken from each pixel image that has been segmented previously. Following is the flow at the SVM testing stage.

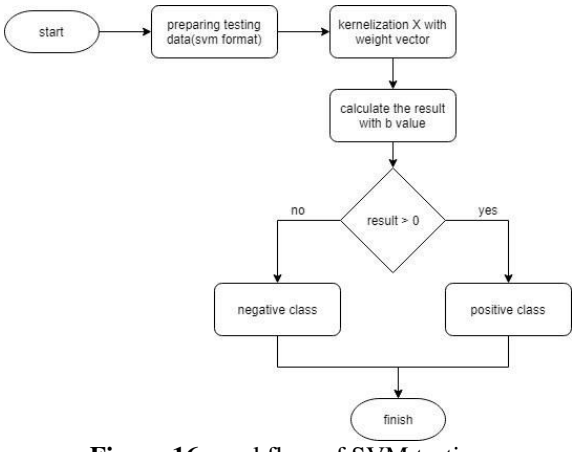


Figure 16. workflow of SVM testing

The input vector will be kernelization with the same kernel function at the training stage, then the values of w and b that already obtained during training will be substituted with the following equation.

$$w.x_i + b \quad (18)$$

A vector will be decided to be in a positive class if the predicted results of the equation (18)> 0. The above calculation will be done on all SVM models that have been trained and the final decision will be taken from the maximum vote SVM model that successfully classifies the input as a class.

2.14 Information Categorizing

The last step is taking the pieces of information contained in the results of the classification using Rules Based. The information that will be taken is the title, the name of the author, the author's NIM, the contents of the abstract and the keywords. To retrieve the information, a rule set will be made to obtain information on the results of the text recognition. The rules used are keywords and row positions based. The following are the rules used to identify information in the skripsi abstract.

Table 1. Rules Table

Component	Rules set
Title	1. Set the starting line to the second line 2. Set the end line to the line before the line that only has the word "oleh" or "by" or the line that has 2 words and the second word has only 1 character 3. If until line 6 the keyword is not found, then the final line is set to the line before the line with the least number of characters 4. Take writing from the beginning to the end of the line that has been set
Name	1. Set theline to two lines after the end of the title line 2. Take the writing from the line
NIM	1. Set the NIM line to the line after the author's name 2. Take the writing from the line
Keywords	1. Check the text from the very last line 2. If you find the word "kata kunci" or "keyword", take the text from the last line 3. If not, check the last two lines 4. If not found, the number of characters in the two lines will be counted and compared 5. If the last number of rows < second to last row, then the initial limit is set to the last second row 6. If not, then the end of line = the last line
Content	1. Set the initial line to the line after NIM 2. Set the ending line to the line before the initial line of the keyword section 3. Take the text from the initial line to the end line that has been set

After all the component have been identified, then the system is done.

2.15 Testing Result

Tests will be accomplished using 40 skripsi abstract images obtained from photos and scans. Tests are first performed on 6 data samples to find the best SVM parameters for later it will be used in overall testing. The following parameters are used.

Table 2. SVM Parameter Table

Kernel	Parameter	
	C	Gamma (γ)
Linear	1	-
	10	-
	100	-
RBF	1	Scale
	100	Scale
	1	1
	100	1

The test was also carried out with a model that was trained in 3 different sizes, 10x10,15x15,20x20px. In addition, testing will also be carried out only on the SVM model to get accuracy from the actual model without being interrupted by other processes using 225 character images. And in the final test, the categorization of information will be done to get the accuracy of information extraction using Rules Based.

2.15.1 Testing Sample Result

From the results of sample testing, it was found that models that were trained using 20x20 images exceeded the capabilities of models trained with other size images. Following are the test results from the model.

Table 3. Testing Sample Table

Model		Akurasi	
Kernel	Parameter	Case Sensitive	Case Insensitive
Linear	C=1	**16.4%**	**17.61%**
	C=10	**16.4%**	**17.61%**
	C=100	**16.4%**	**17.61%**
RBF	C=1,Gamma=scale	15.86%	17%
	C=100,Gamma=scale	16.05%	17.22%

	C=1,Gamma=1	0.2%	1.96%
	C=100,Gamma=1	0.37%	2.12%

From the test results, the parameters that will be used in the overall test data are the model that is trained with an image size of 20x20px and a linear kernel parameter with a value of C = 100.

2.15.2 Testing Whole Data Result

By using the best model and parameter on the sample test results, Below are the accuracy values of the entire test data.

$$Accuracy_{(case\ sensitive)} = 5{,}02\%$$
$$Accuracy_{(case\ insensitive)} = 5{,}47\%$$

2.15.3 Testing SVM Model Result

To get the accuracy of the SVM models to be more accurate, the SVM model testing carried out separately. From the results of sample testing, it was found that models that were trained using 20x20 images exceeded the capabilities of models trained with other size images. Following are the test results from the model.

Table 4. SVM Testing Table

Model		Accuracy	
Kernel	Parameter	Case Sensitive	Case Insensitive
Linear	C=1	**46.61%**	**54.30%**
	C=10	**46.61%**	**54.30%**
	C=100	**46.61%**	**54.30%**
RBF	C=1,Gamma=scale	46.15%	52.94%
	C=100,Gamma=scale	45.25%	51.58%
	C=1,Gamma=1	1.36%	2.71%
	C=100,Gamma=1	1.36%	2.71%

Then the best results obtained from the test is 46.61% for case sensitive and 54.30% for case insensitive.

2.15.4 Testing Information Extraction Result

The test is done by comparing the category section in the actual image with the category section found by the system. The following are the results of categorizing information testing on abstract images.

Table 5. Information Extraction Testing Table

Abstract Code	Detected accordingly?				
	Title	Name	Nim	Content	Keyword
F1	yes	yes	yes	yes	yes
F2	yes	yes	yes	yes	yes
F3	yes	yes	yes	no	no
F4	yes	yes	yes	yes	yes
F5	yes	yes	yes	yes	yes
F6	yes	yes	yes	yes	yes
S1	yes	yes	yes	yes	yes
S2	yes	yes	yes	yes	yes
S3	yes	yes	yes	yes	yes
S4	yes	yes	yes	yes	yes
S5	yes	yes	yes	yes	yes
S6	yes	yes	yes	yes	yes
S7	yes	yes	yes	yes	yes
S8	yes	yes	yes	yes	yes
S9	yes	yes	yes	yes	yes
S10	yes	yes	yes	yes	yes
S11	yes	yes	yes	yes	yes
S12	yes	yes	yes	yes	yes
S13	yes	yes	yes	yes	yes
S14	yes	yes	yes	yes	yes
S15	yes	yes	yes	yes	yes
S16	yes	yes	yes	yes	yes
S17	yes	yes	yes	yes	yes
S18	yes	yes	yes	yes	yes
S19	yes	yes	yes	yes	yes
S20	yes	yes	yes	yes	yes
S21	yes	yes	yes	yes	yes
S22	yes	yes	yes	yes	yes
S23	yes	yes	yes	yes	yes
S24	yes	yes	yes	yes	yes
S25	yes	yes	yes	yes	yes
S26	yes	yes	yes	yes	yes
S27	yes	yes	yes	yes	yes
S28	yes	yes	yes	yes	yes
S29	yes	yes	yes	yes	yes
S30	yes	yes	yes	yes	yes
S31	yes	yes	yes	yes	yes
S32	yes	yes	yes	yes	yes
S33	yes	yes	yes	yes	yes
S34	yes	yes	yes	yes	yes

So we get the following accuracy values:

$$\text{Accuracy} = \frac{\text{Correct result}}{\text{n data}} x\ 100\% = 99\ \%$$

3 CONCLUSION

The conclusion that can be drawn from the research on text recognition in the skripsi abstract document with the Support Vector Machine method are the best accuracy is obtained from overall system is 5.02% for case sensitive and 5.47% for case insensitive with the best SVM model with linear kernel with value parameter C = 100 and trained with an image size of 20x20px, while the accuracy

for character recognition using SVM itself reaches 46.61% for case sensitive and 54.30% for case insensitive. The poor result of overall system is influenced by the segmentation process that is less able to solve existing problems in character separation. Whereas for categorizing information using the Rule Based System has an accuracy of 99%.

Suggestions for next research in order to be better are as follows:

1. Text image segmentation method is needed for cases of broken characters and touching characters.
2. Using other recognition algorithms such as Artificial Neural Network or other algorithms as a comparison.
3. Using other methods at the preprocessing stage in order to separate the uppercase and lower case characters more better.

REFERENCES

[1] D. Mustaqwa and I. Nelly, "Implementasi Ekstraksi Informasi Pada Dokumen Teks Skripsi Menggunakan Metode Rule Based," Universitas Komputer Indonesia, 2017.

[2] I. M. Raden Sofian Bahri, "Perbandingan Algoritma Template Matching dan Feature Extraction pada Optical Character Recognition," *J. Komput. dan Inform.*, vol. 1, no. 1, p. 29, 2012.

[3] R. Anugrah and K. B. Y. Bintoro, "Latin Letters Recognition Using Optical Character Recognition to Convert Printed Media Into Digital Format," *J. Elektron. dan Telekomun.*, vol. 17, no. 2, p. 56, 2017.

[4] P. A, "A Comparative Study of Optical Character Recognition for Printed and Handwritten Tamil Text," *Int. J. Eng. Res. Appl.*, vol. 7, no. 8, pp. 56–60, 2017.

[5] M. R. Phangtriastu, J. Harefa, and D. F. Tanoto, "Comparison between Neural Network and Support Vector Machine in Optical Character Recognition," *Procedia Comput. Sci.*, vol. 116, pp. 351–357, 2017.

[6] B. El Kessab, C. Daoui, B. Bouikhalene, and R. Salouan, "A Comparative Study between the Support Vectors Machines and the K-Nearest Neighbors in the Handwritten Latin Numerals Recognition," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 8, no. 2, pp. 325–336, 2015.

[7] T. Kalaiselvi, "A Comparative Study On Thresholding Techinques For Gray Image Binarization," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 7, pp. 1168–1172, Aug. 2017.

[8] M. F. Soleh and K. K. Purnamasari, "Implementasi Metode Support Vector Machine (Svm) Dan Zoning Untuk Pengenalan Tulisan Tangan Pada Kasus Pengecekan Jawaban Ujian," Universitas Komputer Indonesia, 2018.

[9] J. C. Platt, "Sequential Minimal Optimization : A Fast Algorithm for Training Support Vector Sequential Minimal Optimization : A Fast Algorithm for Training Support Vector Machines," *Microsoft Res.*, vol. 98, no. Technical Report, p. 14, 2013.

[10] B. Jain and M. Borah, "A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical," *Int. J. Sci. Res. Publ.*, vol. 4, no. 6, pp. 4–7, 2014.

[11] A. Septiarini, "Segmentasi Karakter Menggunakan Profil Proyeksi," *Inform. Mulawarman J. Ilm. Ilmu Komput.*, vol. 7, no. 2, pp. 66–69, 2012.

[12] N. Priyanka, S. Pal, and R. Mandal, "Line and Word Segmentation Approach for Printed Documents," *Int. J. Comput. Appl. IJCA ,Special Issue RTIPPR*, no. 1, pp. 30–36, 2010.

[13] S. Safinaz, "An Efficient Algorithm for Image Scaling with High Boost Filtering," *Int. J. Sci. Res. Publ.*, vol. 4, no. 5, pp. 1–9, 2014.