Javanese Document Image Recognition Using Multiclass Support Vector Machine

Yuna Sugianela[1] and Nanik Suciati[2]
[1−2]Departemen Informatika, Fakultas Teknologi Informasi dan Komunikasi, Institut Teknologi Sepuluh Nopember Surabaya
Jawa Timur 60111, Indonesia
Email: [1]nelaneliyuna@gmail.com, [2]naniksuciati@gmail.com

***Abstract*—Some ancient documents in Indonesia are written in the Javanese script. Those documents contain the knowledge of history and culture of Indonesia, especially about Java. However, only a few people understand the Javanese script. Thus, the automation system is needed to translate the document written in the Javanese script. In this study, the researchers use the classification method to recognize the Javanese script written in the document. The method used is the Multiclass Support Vector Machine (SVM) using One Against One (OAO) strategy. The researchers use seven variations of Javanese script from the different document for this study. There are 31 classes and 182 data for training and testing data. The result shows good performance in the evaluation. The recognition system successfully resolves the problem of color variation from the dataset. The accuracy of the study is 81.3%.**

***Index Terms*—Javanese Script, Recognition, Classification, Multiclass Support Vector Machine, One Against One Strategy**

I. Introduction

THE Javanese script is often used in ancient Javanese documents that contain Javanese and Nusantara (Indonesian) culture [1]. However, only a few people understand the Javanese script [2, 3]. The automation system is needed to translate the book with the Javanese script. Thus, it will help people to understand the document written in the Javanese script and help people to study the knowledge of those documents.

The automatic translation of the Javanese script consists of four main stages. The first stage is the segmentation to get the Region of Interest (ROI) image, each character of Javanese script, or letter. The second step is feature extraction of each ROI. In the third step, each character is recognized as the alphabet using the classification method. The last step is combining the alphabet into meaningful words. The Javanese script uses the *Scriptio Continua* (writing continuously) model [4] or script that does not use spaces or other punctuation.

Received: Jan. 10, 2019; received in revised form: Apr. 04, 2019; accepted: Apr. 09, 2019; available online: Apr. 30, 2019.

Moreover, the alphabet of the Javanese script consists of 20 main characters which are syllabic. Javanese script also has *Sandhangan* character, *Pasangan* character, *Murda* characters, *Swara* character, punctuation marks, and numbers [5–7]. There are several studies on Javanese script and other ancient scripts in different culture and country. The topics of this research are about segmentation, feature extraction, recognition or classification, and transliteration. Reference [3] proposed the Hidden Markov Model (HMM) to recognize the character of the Javanese script. The study had good performance. The accuracy is 85.7%. In 2017, Ref. [4] improved the method on her research, but the accuracy was still in 85.7%. Reference [2] also proposed the new method in recognition of the Javanese script. He used the deep learning technique to classify the dataset. The datasets for his study were handwritten Javanese character. He showed a good result in performance with an accuracy of 94.57%.

The challenge in the study of Javanese script automatic translation system is to get a high accuracy in the recognition system and solve the various kind of script for the dataset. The variations are from the shape of the letter and color of the letter and background. In this study, the researchers propose a method in classification to recognize the Javanese script written in the document. The method used in the study is the Multiclass Support Vector Machine (SVM). The strategy of the multiclass classification is One Against One (OAO). This method has been proposed in the ancient script by Ref. [8]. In that study, they used English and Bengali script document for the dataset.

II. Literature Review

A. Characteristics of the Javanese Script

Javanese culture is one of popular culture that shows the identity of Nusantara or Indonesian. Many intel-

Fig. 1. The main character of Javanese script.

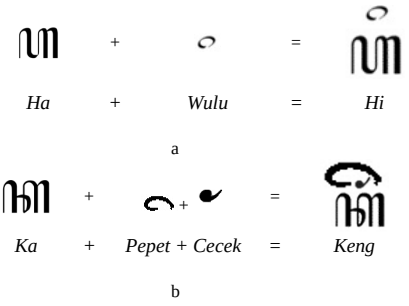

Fig. 2. The example combination main and *Sandhangan* letters of the Javanese script: (a) *Hi*, (b) *Keng*.

lectual properties about the culture are written in the ancient book using the Javanese script. The contents of ancient Javanese books are linguistics, myths, religion, philosophy, laws and norms, folklore, kingdoms, and histories.

The Javanese script consists of 20 main characters known as *Ha-Na-Ca-Ra-Ka*. The name of *Ha-Na-Ca-Ra-Ka* is from five first letters of Javanese script. Figure 1 is the list of the main character of the Javanese script.

The Javanese script also consists of *Sandhangan* letters. Those can change the pronounce of the main character. The main character can be added by one or more *Sandhangan* letter. For example, if there is the main character of *Ha* and *Sandhangan Wulu*, *Ha* is changed into *Hi*. Moreover, if there is the main character of *Ka* and *Sandhangan, Pepet* and *Cecek*, *Ka* is changed into *Keng*. The example is shown in Fig. 2.

Table I is an example of *Sandhangan* letters. In this study, the researchers only discuss the main and *Sandhangan* letters of the Javanese Script. Seven variations of Javanese script are used from a different document. One of the datasets is the printed document of the Javanese script. The document title is *Bloemlezing Uit Javaansche Werken (Proza)* published in 1942. Although the title is written in Dutch, the content of the document is written with the Javanese Language.

TABLE I
THE SANDHANGAN LETTER OF JAVANESE SCRIPT.

Sandhangan	Name	Meaning
	Adeg-adeg	Start of the sentence
	Cecek	-ng
	Koma	,
	Layar	-r
	Pepet	e
	Suku	u
	Taling	e'
	Wulu	i
2	*Taling-tarung*	o
	Wignya	-h
	Titik	.

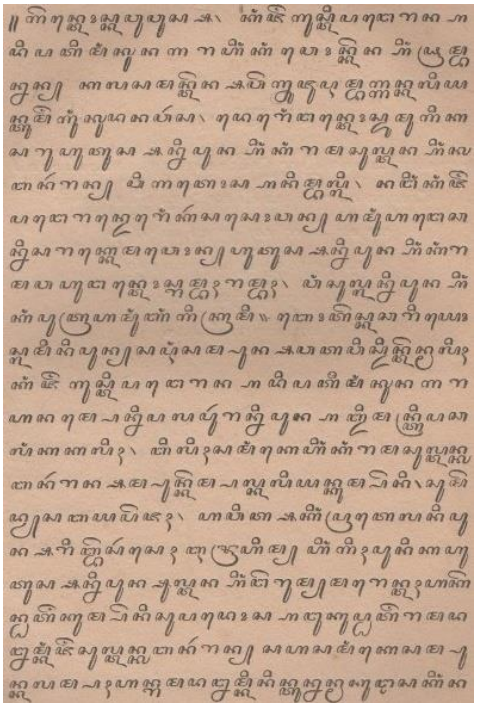

Fig. 3. Dataset from Javanese script.

Figure 3 is one of the documents used in this research.

The other six datasets are the image of the Javanese script from the website. The examples of the dataset are shown in Table II. Because of the good quality of the printed text, the distance between the lines is consistent, but there are some overlaps between characters. From Table II, it can be seen that there are variations of shape and color of letters from a different source.

TABLE II
THE EXAMPLE OF DATASET FROM DIFFERENT SOURCE.

Script	Character Label	Source 1	Source 2	Source 3
Ha	Main			
Da	Main			
Pa	Main			
Ma	Main			
Nya	Main			
Taling	*Sandhangan*			
Wignya	*Sandhangan*			
Pepet	*Sandhangan*			
Suku	*Sandhangan*			
Cecek	*Sandhangan*			

B. Support Vector Machine (SVM)

SVM was developed by Boser, Guyon, and Vapnik, and presented the first time in 1992 in the Annual Workshop on Computational Learning Theory [9]. The concept of SVM is the combination of computation theories that have existed many years before such as margin hyperplane [10]. The basic idea of SVM is a linear classifier, but in the next development, SVM can be used in the non-linear problem by using the kernel trick concept. SVM is the method of learning machine that has a basic idea of Structural Risk Minimization (SRM). SVM works by finding the best hyperplane in the input space. Hyperplane in the vector room in dimension d is $d-1$ dimension affine subspace that separates the vector room into two parts that each of them corresponds to different classes [11].

In general, problems in the real world are rarely linear separable. Most of those are non-linear. To solve this problem, SVM is modified using kernel functions. In this study, the researchers use the Radial Basis Function (RBF) kernel [12]. As mentioned in the Burges tutorial [13], the SVM method has been used for pattern recognition cases such as for isolated handwritten digit recognition [10] and text recognition [14]. In pattern recognition case, dataset input is classified into many classes. To solve the problem, the multiclass SVM classification is needed. One popular strategy in multiclass SVM is OAO. This strategy builds an SVM for each pair of classes, so it is also known as pairwise [15]. This method is like a knockout system in the football tournaments. For a problem with k classes, $k(k-1)/2$ SVMs are trained to separate the input dataset of a class from the dataset of another class [16].

III. RESEARCH METHOD

In this study, the researchers use a multiclass SVM using OAO to recognize the document of the Javanese script. Figure 4 shows the steps of the proposed method. First, it is getting ROI from the document image. Second, each ROI image will be processed in feature extraction using the Histogram of Oriented Gradient (HOG). Third, after getting the feature vector from the HOG process, getting the label of each ROI using multiclass SVM in the classification step will be done. The label of the document consists of name and kind (main or *Sandhangan*).

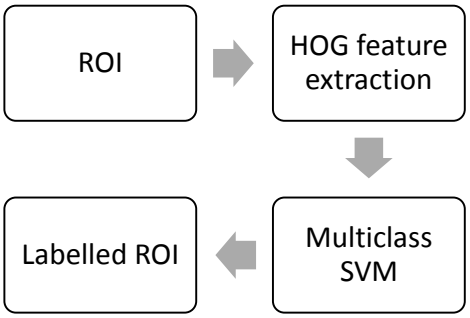

Fig. 4. The flow of the proposed method.

IV. RESULTS AND DISCUSSION

The dataset for this research is from the printed document and images from the website. The researchers select the main character and *Sandhangan* for ROI from seven Javanese script documents. There are many variations in the size of ROI, so the researchers normalize the size of ROI into 64×64 pixels. If the height and width of the original ROI are not the same and the width is less than height, the researchers will add pad in the left and right side. Moreover, the researchers will add pad in above and below side when the height is less than the width. Thus, the height and width will be the same.

Figure 5 is the detail of the number for each letter used for the dataset. The total of the dataset is 182 images of Javanese script letter. Moreover, the researchers set the color of the pad like the color of the background image of ROI. The examples of padding step are in Figs. 6b and 6e. The number of datasets for each letter is different.

The next step is extracting features using the HOG method. The inputs in this stage are grayscale ROI image. HOG is a method for discriminating features. This feature is a histogram description based on edges and orientations that apply to object recognition. This method is widely used in face recognition, animals, and detection of vehicle images [17]. HOG is also used to extract features in the multiclass classification of Batik [18]. Reference [19] show the experiment result that the HOG method applied to Javanese scripts image can achieve good accuracy for recognition.

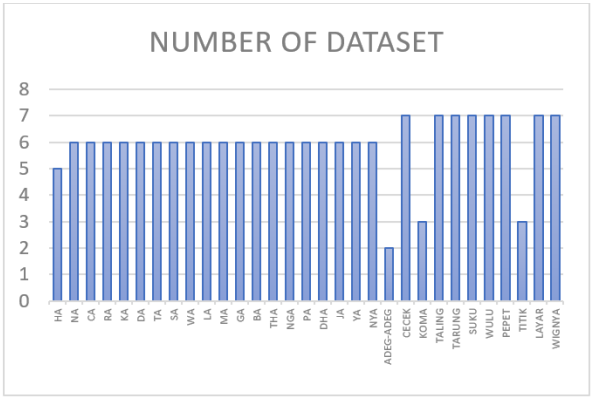

Fig. 5. The number of datasets for each letter.

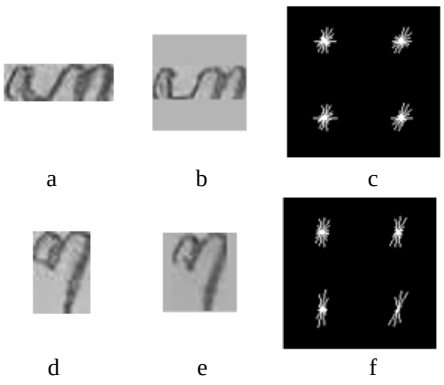

Fig. 6. (a) Original Region of Interest (ROI) *Ha*, (b) Normalized Region of Interest (ROI) *Ha*, (c) Illustration of gradient in Histogram of Oriented Gradient (HOG) *Ha*, (d) Original Region of Interest (ROI) *Taling*, (e) Normalized Region of Interest (ROI) *Taling*, (f) Illustration of gradient in Histogram of Oriented Gradient (HOG) *Taling*.

There are some steps in the HOG method. It is started by calculating the gradient of each pixel. After that, the ROI image is segmented into four cells (32×32 pixels) to calculate the histogram cell. Figures 6c and 6f are examples of gradient illustration in the HOG method.

Every pixel in the cell gives the vote for a histogram channel based on orientation. The researchers use nine bin orientations in voting to set the histogram value contribution. The number of voting is based on Eqs. (1)–(2) [19]. Histogram value in a cell is grouped with others to be normalized for reducing the difference of object brightness. The number of the feature vector that the researchers get in this step is four cells $\times$ 9 bins = 36 feature vectors. Figure 7 is an example of the HOG results of Fig. 6a.

0.04	0.03	0.05	0.17	0.28	0.25	0.21	0.07	0.07
0.04	0.03	0.05	0.18	0.28	0.26	0.21	0.07	0.07
0.03	0.03	0.05	0.16	0.28	0.28	0.25	0.07	0.04
0.03	0.03	0.06	0.15	0.28	0.28	0.27	0.08	0.04

Fig. 7. The example of Histogram of Oriented Gradient (HOG) feature vector.

		System	
		A class	Non-A class
Actual	A class	TP	FN
	Non-A class	FP	TN

Fig. 8. Illustration of the confusion matrix.

$$\nu_j = \mu \frac{c_{j=1} - \theta}{w} \text{ for bin to } j = \left(\frac{\theta}{w} - \frac{1}{2}\right) \text{ mod B}, \tag{1}$$

$$\nu_{j+1} = \mu \frac{\theta - c_j}{w} \text{ for bin to } j = j + 1 \text{ mod B}. \tag{2}$$

It shows that ν as the contribution of the histogram value, μ as the gradient on pixels, c as the middle angle value in bin, θ as the gradient orientation angle at pixel, w as the width of the middle angle value $\left(w = \frac{180}{B}\right)$, and B as bin width of the histogram used.

The next step in this research is classification using multiclass SVM with OAO strategy. Inputs for this stage are 36 feature vectors and a label for each letter. The researchers use all the dataset for training and testing data. The total class used is 31, and the total number of datasets is 182.

To evaluate the performance of our proposed method, the researchers test the accuracy. Accuracy is defined in Eq. (3).

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100. \tag{3}$$

In Confusion Matrix, there is True Positive (TP) statistics, which are the results of the correct system classification for class A. True Negative (TN) is the Non-A class. False Positive (FP) class is the Non-A class classified as an A class. Then, False Negative (FN) is A class which is classified as a Non-A class. Figure 8 is an illustration of the confusion matrix [20].

In this research, each class of letters has a different amount of training data. There are classes that have more than five training data such as *Ha, Na, Ca,*

TABLE III
THE EXAMPLES OF THE MISTAKES IN RECOGNITION SYSTEM.

Letter	Actual Class	Classification Result	
	Da		*Ya*
	Dha		*Ca*
	Ga		*Ka*
	La		*Ya*
	Nga		*Ta*
	Nya		*Ba*
	Taling		*Ga*
	Wa		*Pa*

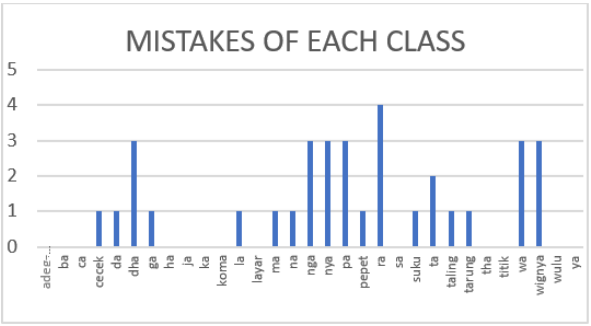

Fig. 9. The number of mistakes in each class.

Ra, Ka, and *Sandhangan Taling, Tarung,* and *Cecek*. However, there are classes less than four training data such as *Sandhangan Adeg-adeg, Titik,* and *Koma*. This does not affect the classification results because the ROI classifies the class consisting of only one training data and has the correct results.

The result shows good performance in the evaluation. However, there are some mistakes in the model classification. The factor that impacts in misclassification is a similarity in the shape of the letter that gives the effect in similar value in the HOG. Figure 9 shows the detail results for each letter. The recognition system successfully resolves the problem of color variation from the dataset. Most of the mistakes are from the letters that have italic shape. It is shown in Table III.

The researchers also compare the performance of Multiclass SVM with other popular classification methods. It is compared with Random Forest (RF), K-Nearest Neighbor (KNN), and Artificial Neural Network (ANN). Table IV shows the comparison of Multiclass SVM with other classification methods. It can be seen that Multiclass SVM has the best result. It reaches 81.3% of accuracy.

V. CONCLUSION

For this study, the researchers use seven different documents. There are 31 classes and 182 datasets for training and testing data. The result shows good performance in the evaluation. The accuracy of the evaluation is 81.3%. This is the best result compared to the other popular classification method (RF, KNN, and ANN).

TABLE IV
THE COMPARISON WITH OTHER CLASSIFICATION METHOD.

Classification Method	Accuracy (%)
RF	37.0
KNN	62.1
ANN	45.0
Multiclass SVM	81.3

For further research, the segmentation method is needed to translate the document of the Javanese script automatically. The evaluation for another kind of Javanese script is also required such as *Pasangan, Swara, Angka,* and *Murdha*. Moreover, the recognition system successfully resolves the problem of color variation from the datasets, but there are still many mistakes from the letters that have italic shape. It can be solved by the method of feature extraction. Moreover, this kind of study can be done in printed Sundanese and Balinese script because they have similar characteristics. The improvements are needed to recognize the ancient handwritten script.

REFERENCES

[1] A. R. Widiarti, A. Harjoko, and S. Hartati, "Preprocessing model of manuscripts in Javanese characters," *Journal of Signal and Information Processing*, vol. 5, no. 04, pp. 112–122, 2014.

[2] M. A. Wibowo, M. Soleh, W. Pradani, A. N. Hidayanto, and A. M. Arymurthy, "Handwritten Javanese character recognition using descriminative deep learning technique," in *2017 2^{nd} International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*. Yogyakarta, Indonesia: IEEE, Nov. 1–2 2017, pp. 325–330.

[3] A. R. Widiarti and P. N. Wastu, "Javanese character recognition using hidden Markov model," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 3, no. 9, pp. 2201–2204, 2009.

[4] A. R. Widiarti, A. Harjoko, Marsono, and S. Hartati, "The model and implementation of Javanese script image transliteration," in *2017 International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT)*. Denpasar, Indonesia: IEEE, Sept. 26–29 2017, pp. 51–57.

[5] Tim Ahli Bahasa Jawa, *Pedoman penulisan Aksara Jawa*. Yogyakarta: Yayasan Pustaka Nusatama, 2002.

[6] A. M. Sulaiman, "Hanacaraka: Aksara Jawa yang mulai ditinggalkan," Institut Seni Indonesia, Tech. Rep., 2011. [Online]. Available: https://bit.ly/2IpgEa0

[7] A. R. Widiarti, A. Harjoko, and S. Hartati, "Line segmentation of Javanese image of manuscripts in Javanese scripts," *International Journal of Engineering Innovations and Research (IJEIR)*, vol. 2, pp. 239–244, 2013.

[8] A. Tikader and N. Puhan, "Histogram of Oriented Gradients for English-Bengali script recognition," in *International Conference for Convergence for Technology-2014*. Pune, India: IEEE, April 6–8 2014, pp. 1–5.

[9] A. S. Nugroho, A. B. Witarto, and D. Handoko. (2003) Support Vector Machine – Teori dan aplikasinya dalam bioinformatika. [Online]. Available: http://www.asnugroho.net/papers/ikcsvm.pdf

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] T. Zhang, "An introduction to Support Vector Machines and other kernel-based learning methods," *AI Magazine*, vol. 22, no. 2, pp. 103–104, 2001.

[12] H. C. S. Ningrum, "Perbandingan metode Support Vector Machine (SVM) Linear, Radial Basis Function (RBF), dan Polinomial Kernel dalam klasifikasi bidang studi lanjut pilihan alumni UII," 2018. [Online]. Available: https://dspace.uii.ac.id/handle/123456789/7791

[13] C. J. Burges, "A tutorial on Support Vector Machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[14] T. Joachims, *Advances in kernel methods - Support vector learning*. Cambridge, Massachusetts: The MIT Press, 1998, ch. Making large-scale SVM learning practical.

[15] B. Aisen. (2006) A comparison of multiclass SVM methods. [Online]. Available: https://courses.media.mit.edu/2006fall/mas622j/Projects/aisen-project/

[16] J. Milgram, M. Cheriet, and R. Sabourin, ""One Against One" or "One Against All": Which one is better for handwriting recognition with SVMs?" in *Tenth International Workshop on Frontiers in Handwriting Recognition*. La Baule, France: Suvisoft, Oct. 23–26 2006.

[17] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for human detection," in *International Conference on Computer Vision & Pattern Recognition (CVPR'05)*, vol. 1. San Diego, CA, USA: IEEE Computer Society, June 20–25 2005, pp. 886–893.

[18] M. N. Fuad and N. Suciati, "Klasifikasi multilabel motif citra batik menggunakan boosted random ferns," *JUTI: Jurnal Ilmiah Teknologi Informasi*, vol. 16, no. 1, pp. 79–89, 2018.

[19] Y. Sugianela and N. Suciati, "Ekstraksi fitur pada pengenalan karakter Aksara Jawa berbasis Histogram of Oriented Gradient," *JUTI: Jurnal Ilmiah Teknologi Informasi*, vol. 17, no. 1, pp. 64–72, 2019.

[20] Y. Sugianela, Q. L. Sutino, and D. Herumurti, "EEG classification for epilepsy based on wavelet packet decomposition and random forest," *Jurnal Ilmu Komputer dan Informasi*, vol. 11, no. 1, pp. 27–33, 2018.