*ISSN : 1978-774X*

# AN EVALUATION OF FEATURE EXTRACTION ALGORITHMS FOR AUTOMATIC LANGUAGE TRANSCRIPTION SYSTEM FOR ANCIENT HANDWRITING JAVANESE MANUSCRIPTS

Brian Karundeng[1] , Kho I Eng[1],
Anto Satriyo Nugroho[2].

[1] Faculty of Information Technology, Swiss German University, German Centre Building 6F Bumi Serpong Damai – INDONESIA 15321
brian.karundeng@student.sgu.ac.id
ie.kho@sgu.ac.id

[2] Center for Information & Communicaton Technology, Center for the Assessment & Application of Technology (PTIK-BPPT) Jalan MH Thamrin 8 BPPT 2nd bld. 4F, Jakarta INDONESIA 10340
asnugroho@inn.bppt.go.id

***ABSTRACT***

*In order to preserve and extract the implicit knowledge of ancient Javanese manuscript, a system is required to scan and translate these manuscripts. The problem faced is the fact that most of these manuscripts are written over a very brittle medium. The best way to solve the problem is by digitalizing them to digital images, then being processed to extract the content. Optical Character Recognition is an effort to extract the content of such invaluable documents that will be followed by language translation process. Optical Character Recognition System consists of preprocessing, feature extraction, character recognition and post processing. Feature extraction works to extract distinctive features of the character that will be fed to the next step: character recognition. In this study, we evaluated several feature extraction algorithms including Local Line Direction, mesh and other different approaches in term of classification rate obtained by Support Vector Machine (SVM).*

*Key words: Javanese Manuscripts, Optical Character Recognition, Local Line Direction, Feature Extraction.*

## 1. INTRODUCTION

Indonesia is a country consists of the mixture of great cultures. One of them is the Javanese culture that still exists today. Some of these writings, or so-called manuscripts are very old and written on top of palm leaves, some are carved on to tablets. These manuscripts are very delicate, they ware off over the years of preservation.

The manual system of preservation is to copy these manuscripts to papers. Not so many left in our generation that could actually understand, and nevertheless write in the ancient language. A system of preservation is to be created in order to preserve this valuable knowledge of history. The main idea of the system is to convert the manuscripts into digital media, and preserve them for longer decades.

In our knowledge, there are very limited studies in the development of ancient handwriting recognition system. In particular for Javanese manuscripts, Harjoko and Widiarti (2006) has developed a system to recognize ancient Javanese manuscript, they have successfully recognized the printed manuscript using an automated system compared using manual recognition,

ISSN : 1978-774X

nevertheless the study was conducted using printed documents.

The complexity of this study lies on the variety of the handwritten characters on different individuals, where as the complexity to recognize printed documents is dependent on what font used in the document.

**2. REVIEW OF RELATED STUDY**

The study in this research is for off-line handwriting recognition otherwise known as OCR. The system work as simple as scanned the document and the data is then captured and recognized by the system. Which can solve the preservation problem. As the development of computers advanced, the demand of more accurate OCR also increased as a lot of data such as forms, mails, bank slips, and especially historical data were handwritten.

There are 4 main stages in the system, the first one is the pre-processing stage, feature extraction stage, classification stage, and post-processing stage. This study is concentrated on the feature extraction stage to find the optimal method out of the introduced method in order to maximize the next stage, that is the classification stage.

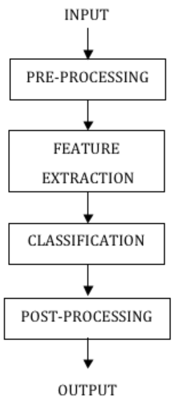



*Figure 1. Simple OCR diagram for Javanese manuscript.*

To extract the alphabet an image pre-processing step, the first stage is where the enhancements such as binarization occur to optimize the image for further process. As a result, a normalized image was created. The equation to produce the binary image is given Gonzales & Woods (1992):

$$g(x,y)=\begin{cases}1 & if \quad f(x,y)>k\\ 0 & if \quad f(x,y)\le k\end{cases} \tag{1}$$

With $g(x,y)$ being the binary image from black and white image $f(x,y)$, and *k* is the value of threshold.

The first step for normalization is to define the initial frame that fit the image with $xNorm_o \times yNorm_o$ then rescale the initial image to $xNorm \times yNorm$ pixels which is the maximum size according to $xNorm_o \times yNorm_o$, as follows:

$$\begin{aligned} xNorm &= \max(xNorm_0, yNorm_0)\\ yNorm &= \max(xNorm_0, yNorm_0)\end{aligned} \tag{2}$$

The next stage is the feature extraction process, in this research several scenarios were used and the result is compared to show the prominent method. The first scenario is the mesh (pixel by pixel) extracted feature, the second is by using Local Line Direction (LLD) method, and the final one is the combination of both scenarios.

The work of Fujisawa & Liu (2003) has shown the advances of the directional features for handwriting recognition features. The work of Suwa et al. (1991) and Iwata et al. (1991) also has shown a good result for LLD in handwriting character recognition.

The result from the previous stage contains valuable information differentiate each alphabet to have unique values that can maximize the performance towards the next and also the last stage that is to be covered in this research that is classification stage. Since the main interest of this research is on the feature extraction stage, and not in the differences between types of classifiers only both *k*-Nearest Neighbor (*k*-NN) and Support Vector Machine (SVM) were used as the classifiers.

All of the experiments in this study are using the 10-folds cross validation scheme, the reason is because of the

Javanese character is not used in the daily life, samples we gathered to build the data set is insufficient to use the hold-out method. Where holding a certain amount of data set and use for training does the holdout procedure. But in order to obtain maximum output, the data is use as much as possible for testing.

## 3. DATA PREPARATION

Before the character recognition can be conducted, a database of alphabet was built, and named Javanese Alphabet Database (JADB v.1). The process of building the database is performed by using a form consisting of each character of the Javanese letters, and then handed out for people to write the alphabet below the sample. The first row of the form is the sample of each printed Javanese alphabet, i.e.: *ha, na, ca, ra, ka, da,ta,sa, wa, la, pa, dha, ja, ya, nya, ma, ga, ba, tha, nga*. And below this row, boxes are available to be written on as samples of handwritten Javanese alphabet.

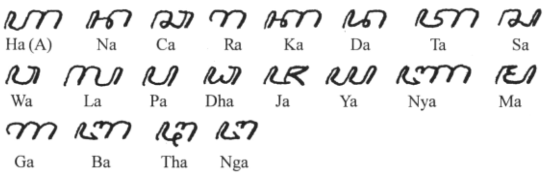



*Figure 2. Ancient Javanese Alphabet.*

This form was distributed to the total of 100 persons to fill out the forms, and the collected forms are manually sorted to find an appropriate set of alphabets that are qualified. As a result a collection of 3379 alphabet (samples) were acquired for the whole 20 classes that can be use as dataset.

## 4. PROPOSED METHOD

The experiments conducted in this research were divided into 3 different experiments to be compared in order to find the most prominent feature extraction method. The first experiment was using the mesh method, followed by the developed feature extraction method, LLD and the last experiment was by combining methods, mesh and LLD.

The mesh feature is basically converting the alphabet scanned from the form, and standardized them pixel-by-pixel through binarization and normalization process. The normalize dimension in this research is set to 20 by 20. This dimension is achieved to through a series of experiments, with the result showing that the dimension of the normalize image is proportional to the image quality, hence will affect the recognition process.

The feature extraction method developed, Local Line Direction is the second method used in the experiment. The general steps after the normalizing the character into standard size is to generate specific directional planes for each local stroke. Each of the directional planes created as masks is divided into 9 by 9 blocks which will create the total of 81 vectors, and for each mask as shown in Fig. 3 the values of black box = -1/36, white box = 1/45 for vertical and horizontal filters respectively, and the value of black box = -1/42, and white box = 1/39 for both diagonal filters. Each mask records the pixels on the image according to a particular stroke orientation. To obtain the values, each direction plane is partitioned into an equal sized number of zones and takes the sum of pixel values in each zone. Fig. 4 show the process.

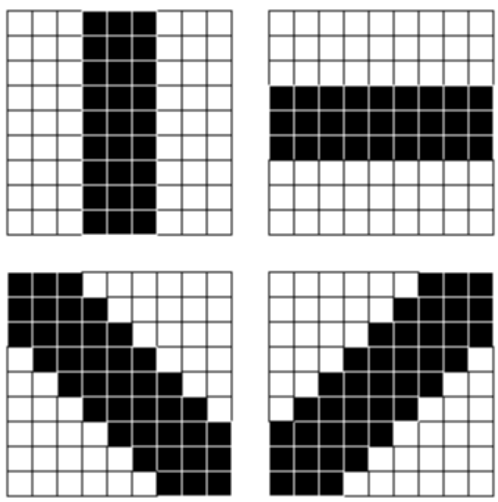


*Figure 3. Mask of four directions.*

ISSN : 1978-774X

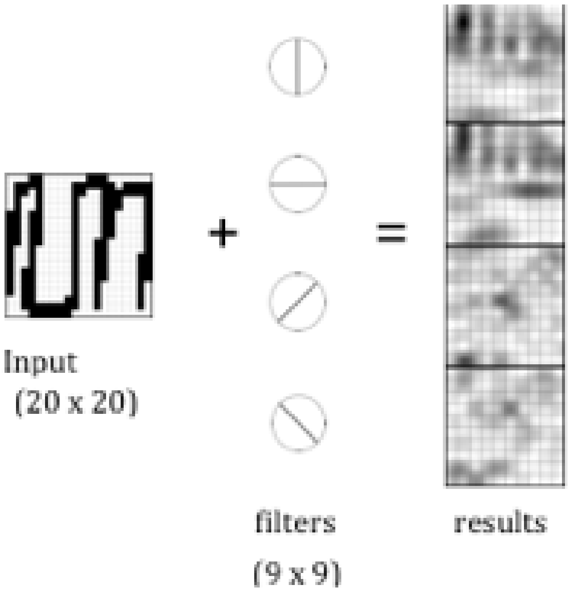



*Figure 4. LLD process.*

## 5. RESULT AND DISCUSSION

To find the recognition rate, the extracted features from each character are fed to the classification process. All of the experiments are using the classifiers *k*-NN with the *k* value set to 5 and SVM with the complexity parameter of *C* = 1.0, round of error *epsilon* = 1.0E-12, using the training data for logistic models, and with the kernel of *PolyKernel*. The *PolyKernel* parameter of *exponent* value set to 1.

The sample used were 3379 samples in a total of all 20 classes with each of them consisting of 401 attributes. The normalized image size of 20 x 20 creates a vector 20 x 20 which means that the values are 400, plus one attribute for the class (*ha,na,..,nga*).

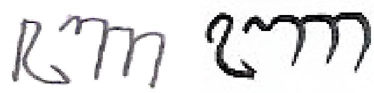
Table 1. The classifier comparison.

| Features Methods | Recognition Rate (%) | |
|---|---|---|
| | ***k*-NN** | **SVM** |
| Mesh | 77.1 | 81.1 |
| LLD | 82.7 | 87.8 |
| Mesh + LLD | 79.4 | 86.9 |

The result of experiment using Mesh Features showed the recognition rate of 77.1%% with *k*-NN classifier and 81.1% with SVM classifer. The result of experiment using LLD features showed the recognition rate of 82.7% with *k*-NN classifiers and a high recognition rate of 87.8% with SVM classifier. The experiment using combination of both methods show 79.4% recognition rate with *k*-NN classifier and 86.9% with SVM classifier.

An experiment to show the model performance for different variety of added classes to the model is conducted. For every class that added to the experiment, the recognition rate is then evaluated.

The samples used were selected alphabet that is rated as “good”. The definition of “good” is manually rated according to the similarity between the printed sample and the handwritten sample. Fig. 5 shows the example of the rated sample. The rated “good” samples used were 1751 samples.



(a) (b)

*Figure 5. (a) Low rated sample, (b) High (“good”) rated sample.*

Table 2 shows that the performance of LLD feature is better than the others. At the end of the experiment, where the full classes of the model has been added to the model shows a number of 84.30% recognition rate.

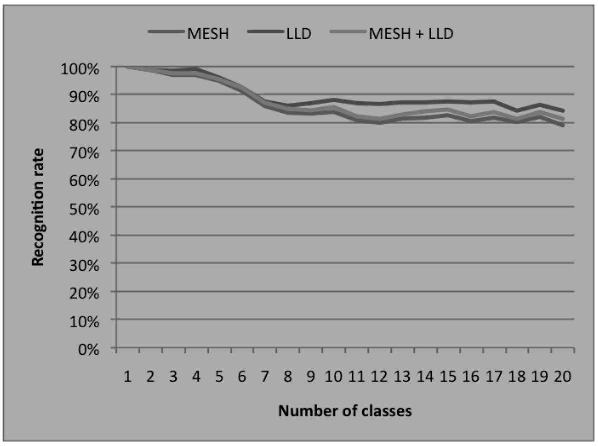



*Figure 6. Evaluation Graph*

Table 2. The model performance for each added class (in percentage).

| Class | MESH | LLD | MESH + LLD |
|---|---|---|---|
| 1 | 100 | 100 | 100 |
| 2 | 98.80 | 98.80 | 98.80 |
| 3 | 96.90 | 98.40 | 97.60 |
| 4 | 96.80 | 99.10 | 97.40 |
| 5 | 94.90 | 96.00 | 95.30 |
| 6 | 91.30 | 92.50 | 92.50 |
| 7 | 85.80 | 87.30 | 87.10 |
| 8 | 83.40 | 85.80 | 84.80 |
| 9 | 83.20 | 86.60 | 84.40 |
| 10 | 83.60 | 87.80 | 85.40 |
| 11 | 80.60 | 86.80 | 82.20 |
| 12 | 79.80 | 86.40 | 81.70 |
| 13 | 81.40 | 87.10 | 82.80 |
| 14 | 81.70 | 87.10 | 83.90 |
| 15 | 82.40 | 87.40 | 84.70 |
| 16 | 80.50 | 86.90 | 82.30 |
| 17 | 81.50 | 87.20 | 83.60 |
| 18 | 80.00 | 84.70 | 81.40 |
| 19 | 81.90 | 86.00 | 83.70 |
| 20 | 79.10 | 84.30 | 81.40 |

Curious with the misclassified class, an experiment was conducted to show the percentage of recognition for each of the classes, first the quantity sample for each class has to be measured. The next step is to know the number of misclassified count for each class. Then the number of samples is subtracted by the number of error count, and multiplied with a hundred percent will give the percentage of the recognized for each of the classes.

Table 3. Evaluation on each classes.

| No. | Class | Samples | Error Count | Percentage (%) | |
|---|---|---|---|---|---|
| | | | | Recog. | Error |
| 1 | ha | 171 | 30 | 82.46 | 17.54 |
| 2 | na | 171 | 23 | 86.55 | 13.45 |
| 3 | ca | 171 | 22 | 87.13 | 12.87 |
| 4 | ra | 171 | 5 | 97.08 | 2.92 |
| 5 | ka | 171 | 18 | 89.47 | 10.53 |
| 6 | da | 170 | 29 | 82.94 | 17.06 |
| 7 | ta | 169 | 42 | 75.15 | 24.85 |
| 8 | sa | 169 | 17 | 89.94 | 10.06 |
| 9 | wa | 169 | 23 | 86.39 | 13.61 |
| 10 | la | 169 | 8 | 95.27 | 4.73 |
| 11 | pa | 167 | 18 | 89.22 | 10.78 |
| 12 | dha | 167 | 17 | 89.82 | 10.18 |
| 13 | ja | 167 | 12 | 92.81 | 7.19 |
| 14 | ya | 167 | 17 | 89.82 | 10.18 |
| 15 | nya | 167 | 28 | 83.23 | 16.77 |
| 16 | ma | 168 | 20 | 88.10 | 11.90 |
| 17 | ga | 168 | 9 | 94.64 | 5.36 |
| 18 | ba | 169 | 34 | 79.88 | 20.12 |
| 19 | tha | 169 | 18 | 89.35 | 10.65 |

ISSN : 1978-774X

| | | | | | |
|---|---|---|---|---|---|
| 20 | nga | 169 | 22 | 86.98 | 13.02 |
| Total | | 3379 | 412 | | |

Table 3 shows the recognition rate for each of the classes, meaning the percentage of the recognized alphabet out of the total sample. For example the recognized percentage of the alphabet *ta* is 75.15% shows the lowest recognition rate, and the highest error rate of 24.85%. This percentage concludes that most of the sample from the class *ta* is misclassified as class *ha*. The reason of the misclassified result is because of the stroke similarity between both alphabet *ta* and *ha* as shown in Fig. 7.

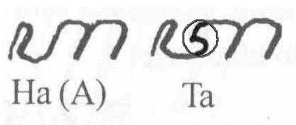



*Figure 7.* The main difference of alphabet *ha* and *ta.*

There are many similar characteristics in the Javanese alphabet. The similarity between the Javanese alphabet causes the confusion for the classifier to misclassified classes. Some of these alphabet differ only in a small stroke such as those in Fig 8.

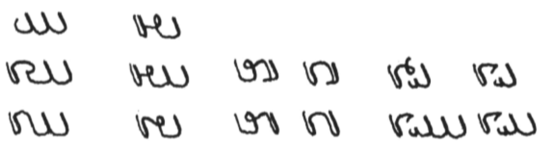


*Figure 8. Group of similar Javanese Alphabets*

## 6. CONCLUSION

The complexity of the research is caused by the unfamiliar alphabet for general public, causing a difficulty for the process of building the database. Due to the similarities of shape between different alphabets, it is challenging to achieve high recognition accuracy, especially for handwritten recognition. The classification experiments were using the cross-validation scheme due to the scarce number of data to maximize the quality of the model.

This study shows that LLD method performs better out of the other methods introduced in this research paper. For the classification stage experiments, SVM shows a better performance than *k*-NN. The recognition rate has shown a relatively high number of accuracy of 87.8% for the LLD method using SVM as the classifier.

## 7. FUTURE STUDIES

Recommendation for future studies includes the proposed method for the classification stage for Javanese handwriting recognition system, and other development includes:

- Segmentation process development.
- Recognition using real ancient Javanese manuscripts.
- Development of post-processing stage for character recognition.
- Further application for developed system to a similar set of alphabet. For example: Balinese alphabet


*Figure 9. Balinese alphabet.*

## 7.REFERENCE

(a) Cheriet, M., Kharma, N., Liu, C., Suen, C.Y. (2007). *Character Recognition Systems: A guide for Students and Practioners*, New Jerset: John Wiley & Sons, Inc.

(b) Fujisawa, H. & Liu, C.-L., “Directional Pattern Matching for Character Recognition Revisited”, in Proceedings of the Seventh International Conference on Document Analysis and Recognition, IEEE Computer

*ISSN : 1978-774X*

Society, Washington DC, vol.2, 2003, p. 794.

(c) Gonzales, R. C. & Woods, R. E., (1992). *Digital Image Processing*, USA: Addison-Wesley Publishing Company, Inc.

(d) Harjoko, A. & Widiarti, A. R., “Document Processing System for Javanese Manuscripts”, in Proceedings of Signal and Image Processing 2006, p.534

(e) Iwata, A., Kawajiri, H., and Suzumura, N., “Classification of handwritten digits by a large scale neural network ‘CombNET-II’.”, In Proceedings of IEEE & INNS International Joint Conference on Neural Networks, Singapore, 1991, pp.1021–1026.

(f) Suwa, Y., Kawajiri, H., Iwata, A., and Suzumura, N., ’0 numerals recognition by CombNET-II’. *IEICE Autumn National Conference Rec.* D-19. 1991.