Preservation of Hanacaraka Characters in Old Manuscripts Using Machine Learning Approach

Gregorius Satia Budhi, Hans Christian Indrayana, Liliana, Yulia, and Rudy Adipranata

Informatics Department, Petra Christian University
greg@petra.ac.id

***Abstract*—The Javanese language has a unique set of the letters called Hanacaraka characters, which is different compared to Latin alphabet. Since modern Javanese ethnics of Indonesia don't use it anymore for formal conversation and education, this language, especially its Hanacaraka characters begins to extinct. For the preservation purpose of old manuscripts in Hanacaraka characters, we create a system that can recognise Javanese characters automatically from old manuscript or writing. For this system, we investigated and employed several methods of image processing, features extractions and machine learning for character recognizer. In this paper, we present the result of our investigation of traditional feed-forward neural networks and Elman recurrent networks and comparing their accuracies to obtain the best recognizer. We also compare the results with the accuracies of probabilistic neural network and induction tree from our previous experiments. From the comparison we found that Elman recurrent network outperforms the performance of other algorithms, with accuracy more than 97% for data training and 85% for data testing.**

***Index Terms*—Cultural preservation; Elman recurrent networks; Feed-forward networks; Hanacaraka characters; Machine learning.**

I. INTRODUCTION

Local wisdom in Javanese literature and culture are usually interpreted as a Javanese way of life to solve life problematic based on traditions. In literature, they are presented in the form of Javanese language words, symbols and picture [1, 2]. The Javanese language owns a different set of letterforms and shapes compared to Latin alphabet, namely Hanacaraka [3, 4]. This difference creates difficulty for people to read or write Javanese language literature or script [5]. This condition is compounded by the ordination of Melayu language to be the formal language of the Republic of Indonesia in 1945, called Bahasa Indonesia (Bahasa), and is used as a language in formal education, news and other electronic or printed media [6, 7]. The implication of Indonesian as a national language is that people start using Bahasa as a colloquially to replace the Indonesians local ethnic language, like Javanese [6]. This fact slowly erodes the existence of Javanese language and Hanacaraka since people don't use it anymore in daily life [5, 8]. For the preservation of the cultural wealth of Indonesia from extinction, some attempt have been made by researchers and also the government of Indonesia. For example, Indonesian government release a law to preserve cultural heritage in the form of objects, buildings, structures and sites [9]. Implementing the law and also from personal awareness, some Indonesian researchers conduct research to preserve Indonesian culture heritage using their knowledge and expertise [1, 4, 8, 10-12].

This study is the continuity of our attempt to preserve Indonesian cultural heritage that is begun in 2012 [13]. We especially focus on the preservation of Hanacaraka characters and literature that use these characters. In this study, we employ our expertise in image processing and machine learning, especially Artificial Neural Network, to recognise Javanese characters automatically from the pictures of old manuscript pages, covert them to Hanacaraka font, and then save it into a document. To build a sophisticated and reliable system, we have investigated several image processing and feature extraction methods [14, 15], and also artificial neural networks algorithms [3, 5, 16, 17]. In this paper, we present the results of our current experiments compared to our previous works. We built the training dataset manually from several pictures of old manuscript pages that we gathered from several places like Sultan's Kraton (palace) in Jogjakarta and libraries in several cities of Java island. We use 62 different kind of Javanese characters for the target of recognition. They are 20 aksara carakan (basic characters), 20 aksara pasangan (consonant in the end of word), 10 aksara wilangan (numbers), and 12 sandhangan (signs to make changes to the sound of aksara carakan, i.e. Ha to He or Ho).

The rest of paper is organized as follows. In Section II, we review several works related to this research. In Section III, we present the methods used, which include the system design, data and targets, as well as artificial neural networks algorithms to be used. Experimental results and discussions are presented in Section IV. And finally, we present our

conclusion and future research direction in Section V.

II. Related Works

Some attempts have been made by researchers, academicians and other Indonesian people to preserve Indonesia's cultural heritage, especially Javanese cultural heritage. Setiawan and Sulaiman, in 2015, proposed the preservation of Hanacaraka characters by using them as a brand image on the blackboard, t-shirt, street signs, etc. [4]. Sunarni in 2016 proposed the preservation of Java language, by implementing contextual learning method to Unggah-Ungguh course material in Javanese language course for junior high school level [12]. The preservation of Javanese local wisdom in the form of traditional Javanese music, called Karawitan, was implemented as an extracurricular course for junior high school level by Sularso and Maria [10]. Suryono made an action to preserve the traditional architecture and interior layout of Bangsal Sitihinggil, a building inside Sultan's palace in Jogjakarta [11]. In 2017, Saddhono proposed an idea to preserve Javanese culture and local wisdom by implementing them in the form of literary works and writing. In 2017, Erawati employed her research to interpret the sound segment of old Javanese language using speech analyser and distinctive features analysis [8]. Few researchers from computer studies have taken part in Javanese cultural heritage preservation. They are usually focusing their research to the handwritten carakan (20 basic Hanacaraka characters) using some image processing techniques [14, 18], artificial neural networks (ANNs) [3, 5, 16, 19] and deep neural networks [20-22] to recognise these characters automatically. While other researchers used printed Hanacaraka characters from old manuscripts as their source of inputs [14, 15, 17, 23].

Differs from other research, in our current research we use printed Hanacaraka characters that we extract directly from old manuscripts as training dataset and not the handwriting characters. We broaden the target output from 20 aksara carakan to be 62 Hanacaraka characters as we mentioned in section 1. For the pattern recogniser, we implement ANNs feed forward networks and Elman recurrent networks architectures.

III. Design

In Fig. 1, can be seen the overall design of our system. Our system design is straightforward. We use backpropagation training algorithm to train our feed-forward networks [24]. For Elman recurrent networks training process, we implement training algorithm that is provided by the author [25]. This system produces a dataset of Hanacaraka characters from some old Javanese characters manuscripts (see Fig. 2 for the example). To provide data for the ANNs training process, we implement data preparation system from our previous research. This preparation system is included several image processing that are image segmentation, histogram balance, thinning and skew correction [15].

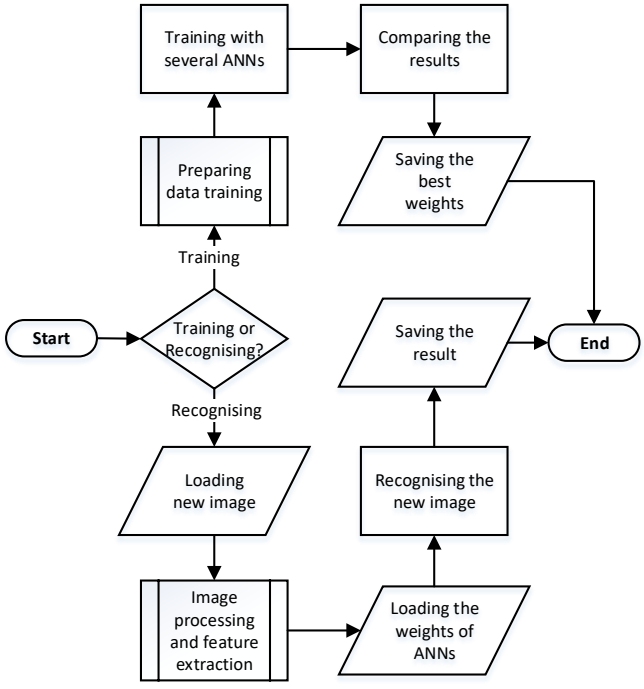

Figure 1: Overal system design

We implement area-based feature extraction from our previous research [14] to provide data input for both training and testing processes. For comparison purpose between the results of the current research to our previous research [17], we use data training from the same source and using same fashion to extract the Hanacaraka characters from old manuscripts images.

The first step in the recognising system is uploading an image of a page of the old manuscript. Then we process it using the same fashion in data training preparation [14, 15]. After choosing and loading the ANNs weights from training, we use it to recognise Hanacaraka characters within the picture, then save the results of recognition to a text document.

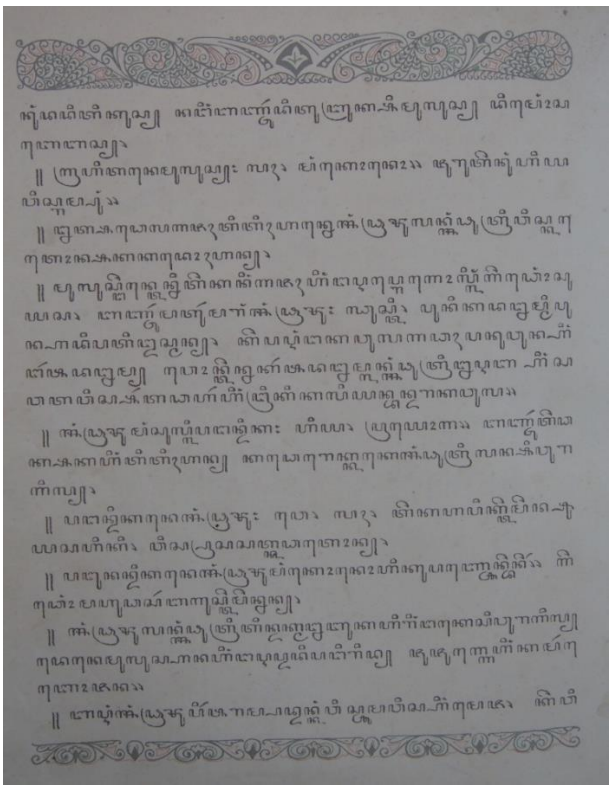

Figure 2: Old manuscript image

IV. RESULT AND DISCUSSION

We conducted the experiments for the purpose to search optimal parameters of the networks. In these experiments, we used learning rate = 0.05 and error threshold = 0.01, which we got from simple experiments before. We employed the experiments by splitting sample datasets to be three parts, used two parts as data training and one part for testing. In Table 1, can be seen the results of experiments for the optimal amount of hidden neurons. For these experiments, we used one hidden layer for each network architectures. From the results, we found that the best accuracy is achieved by 50 hidden neurons for feed-forward networks and 88 hidden neurons for Elman recurrent networks. Another set of experiments would be done for the amount of hidden layer (see Table 2). For these experiments, we used the best hidden neurons amount we found that are 50 neurons for feed-forward networks and 88 neurons for Elman recurrent networks. The results say that one hidden layer provides best accuracies for both feed-forward and Elman recurrent networks.

Table 1
Results of Hidden Neurons Amount Experiments

Networks Type	Hidden Neurons amount	Average accuracy (%)	
		Training Data	Testing data
Feed-forward networks	25	97.02	81.41
	50	96.56	83.48
	88	97.48	83.37
	100	96.1	82.20
Elman recurrent networks	25	78.25	65.74
	50	94.5	80.61
	88	97.71	85.16
	100	97.48	85.02

Table 2
Results of Hidden Layer Amount Experiments

Networks Type	Hidden layer amount	Average accuracy (%)	
		Training Data	Testing data
Feed-forward networks	1	96.56	83.48
	2	96.1	80.25
	3	96.56	81.71
	4	96.56	78.57
Elman recurrent networks	1	97.71	85.16
	2	95.64	81.20
	3	95.18	82.09
	4	96.1	81.72

Table 3
Comparison Between Algorithms for Hanacaraka Characters Recognition

Type of algorithm for recognition	Parameters	Average accuracy (%)	
		Data training	Data testing
Feed-forward networks	α = 0.05; error threshold = 0.01; 1 hidden layer; 50 neurons; activation = sigmoid	96.56	83.48
Elman recurrent networks	α = 0.05; error threshold = 0.01; 1 hidden layer; 88 neurons; activation = sigmoid	97.71	85.16
Probabilistic neural networks(*)	Activation: radial basis function	92.35	61.08
Induction Tree(*)	Entrophy; multi-class splitting method	100	15.57

(*) From previous research [17]

The comparison between the result of our current research and previous research can be seen in Table 3 [17]. After observing the comparison of accuracies in Table 3, we have the conclusion that Elman recurrent networks outperform the performance other algorithms both for data training and testing. Although for recognition of data training the Induction Tree is perfect (accuracy = 100%), but it has a tendency to be overfitting, proved by its performance for data testing is the worst.

V. Conclusion

Preservation of cultural wealthiness and local wisdom of Indonesia is not only the job of the government but shall be the duty of all Indonesian citizens. Using this research, we take part in this preservation attempt by making a system that can recognise Hanacaraka characters automatically from old manuscripts and writing. From the observation of experimental results, we found that Elman recurrent networks outperform the performance of other recogniser algorithms. Thus we will implement this ANNs method for our further research on the same topic. We plan to broaden out our system with the capability to convert Hanacaraka characters to alphabet style words, voicing them out and translate the meaning of the produced texts to Bahasa or English.

Acknowledgment

This research was funded by DIPA Directorate General of Research and Development Reinforcement (Direktorat Jenderal Penguatan Riset dan Pengembangan Kementrian Riset, Teknologi, dan Pendidikan TInggi) no. 120/SP2H/LT/DRPM/2018, fiscal year 2018.

References

[1] K. Saddhono, "Membangun kearifan lokal melalui karya sastra dan budaya daerah [Jawa]," in Seminar Nasional Bahasa Dan Budaya 2017, Denpasar, 2017.
[2] Marzuki, "Tradisi dan budaya masyarakat Jawa dalam perspektif Islam," *Kajian Masalah Pendidikan dan Ilmu Sosial "INFORMASI"*, vol. 32, no. 1, pp. 1-13, 2006.
[3] G. S. Budhi and R. Adipranata, "Java characters recognition using evolutionary neural network and combination of Chi2 and backpropagation neural network," *International Journal of Applied Engineering Research*, vol. 9, no. 22, pp. 18025-18036, 2014.
[4] A. Setiawan and A. M. Sulaiman, "Hanacaraka: Aksara Jawa dalam karakter font dan aplikasinya sebagai brand image," *Ornamen*, vol. 12, no. 1, pp. 33-47, 2015.
[5] G. S. Budhi and R. Adipranata, "Handwritten Javanese character recognition using several artificial neural network methods," *Journal of ICT Research and Applications*, vol. 8, no. 3, pp. 195-212, 2015.
[6] A. Gunarwan, "Kasus-kasus pergeseran bahasa daerah: Akibat persaingan dengan bahasa Indonesia?," *Linguistik Indonesia*, vol. 24, no. 1, pp. 95-113, 2006.
[7] L. Ali, *Iktisar sejarah ejaan bahasa Indonesia*. Jakarta: Pusat Pembinaan dan Pengambangan Bahasa, 1998.
[8] N. K. R. Erawati, "Interpretasi segmen bunyi bahasa jawa kuno: Analisis speech analyzer dan fitur distingtif," *Aksara*, vol. 29, no. 2, pp. 225-238, 2017.
[9] *Undang-Undang Republik Indonesia Nomor 11 Tahun 2010 Tentang Cagar Budaya*, M. o. L. a. H. Rights, 2010.
[10] P. Sularso and Y. Maria, "Upaya pelestarian kearifan lokal melalui ekstrakurikuler karawitan di smp negeri 1 Jiwan tahun 2016," *Citizenship Jurnal Pendidikan Pancasila dan Kewarganegaraan*, vol. 5, no. 1, pp. 1-12, 2017.
[11] A. Suryono, "Pelestarian aspek kesemestaan dan kesetempatan dalam arsitektur Bangsal Sitihinggil di kraton Yogyakarta," *RUAS*, vol. 14, no. 2, pp. 1-10, 2016.
[12] Sunarni, "Pelestarian lingkungan sosial budaya melalui peningkatan prestasi belajar bahasa Jawa dalam materi unggah-ungguh," *GeoEco*, vol. 2, no. 1, pp. 88-102, 2016.
[13] R. Adipranata, G. S. Budhi, and R. Thedjakusuma, "Java Characters Word Processing," in *proceedings of The 3rd International Conference on Soft Computing. Intelligent System and Information Technology*, Bali, Indonesia, 2012.
[14] R. Adipranata, G. S. Budhi, Liliana, and B. Sebastian, "Comparison between Shape-Based and Area-Based features extraction for Java character recognition," in *proceedings of the 2nd Management and Innovation Technology International Conference*, Bangkok, Thailand, 2015.

[15] Liliana, S. Soephomo, G. S. Budhi, and R. Adipranata, "Segmentation of Hanacaraka character using double projection profile and Hough transform," in *proceedings of the 8th EIA International Conference on Big Data Technologies and Applications*, Gwangju, South Korea, 2017.
[16] G. S. Budhi and R. Adipranata, "Comparison of Bidirectional Associative Memory, Counterpropagation and Evolutionary Neural Network for Java Characters Recognition," in *Proceedings of The 1st International Conference on Advanced Informatics: Concepts, Theory and Applications*, Bandung, Indonesia, 2014.
[17] G. S. Budhi, R. Adipranata, B. Sebastian, and Liliana, "The use of Probabilistic Neural Network and ID3 algorithm for Java character recognition," in *proceedings of the 2nd Management and Innovation Technology International Conference*, Bangkok, Thailand, 2015.
[18] R. Y. Astuty and E. D. Kusuma, "Pengenalan aksara Jawa menggunakan digital image processing," in *proceedings of the 2nd Informatics Conference*, Jakarta, Indonesia, 2016, pp. 32-35.
[19] N. Nurmila, A. Sugiharto, and E. A. Sarwoko, "Backpropagation neural network algorithm for Java character pattern recognition," *Jurnal Masyarakat Informatika*, vol. 1, no. 1, pp. 1-10, 2010.
[20] K. Rismiyati and A. Nurhadiyatna, "Deep learning for handwritten javanese character recognition," in *proceedings of the 1st International Conference on Informatics and Computational Sciences*, Semarang, Indonesia, 2017, pp. 59-64.
[21] M. A. Wibowo, M. Soleh, W. Pradani, A. N. Hidayanto, and A. M. Arymurthy, "Handwritten Javanese character recognition using descriminative deep learning technique," in *proceedings of the 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering*, Yogyakarta, Indonesia, 2017, pp. 325-330.
[22] C. K. Dewa, A. L. Fadhilah, and A. Afiahayati, "Convolutional neural networks for handwritten Javanese character recognition," *Indonesian Journal of Computing and Cybernetics Systems (IJCCS)*, vol. 12, no. 1, pp. 83-94, 2018.
[23] T. Arifianto, "Segmentasi aksara pada tulisan aksara Jawa menggunakan adaptive threshold," *Smatika*, vol. 7, no. 1, pp. 1-5, 2017.
[24] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1: MIT Press, 1986, pp. 318-362.
[25] J. L. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, pp. 179-211, 1990.