Recognition of Hanacaraka Characters in Old Manuscripts Using Feed-Forward Networks and Elman Recurrent Networks

Gregorius Satia Budhi*[1], Hans Christian Indrayana, Liliana[2], Yulia[3], Rudy Adipranata[4]

Informatics Department, Petra Christian University, Surabaya, Indonesia
*Corresponding author, e-mail: greg@petra.ac.id[1], lilian@petra.ac.id[2], yulia@petra.ac.id[3], rudya@petra.ac.id[4]

Abstract

The Javanese language has a unique set of letters called Hanacaraka characters, which is different compared to the 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 an old manuscript or writing. For this system, we investigated and employed several methods of image processing, features extractions and machine learning for character recogniser. 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 recogniser. We also compare the results with the accuracies of the 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.

Keywords*: cultural preservation; Elman recurrent networks; feed-forward networks; Hanacaraka characters; machine learning*

1. 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 the 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 has been made by researchers and also the government of Indonesia. For example, the 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]. The problem we work here is how to preserve the Javanese language from extinction since this is one of Indonesia cultural wealth.

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 by developing a tool to recognise these characters automatically.

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].

In this study, we employ our expertise in image processing and machine learning, especially ANNs, to recognise Javanese characters automatically from the pictures of old manuscript pages, convert 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. Differs from other research, in this current research we use printed Hanacaraka characters that we extract directly from old manuscripts as training dataset and not the handwriting characters. 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 broaden the target of recognition from 20 aksara carakan to be 62 Hanacaraka characters. They are 20 aksara carakan (basic characters), 20 aksara pasangan (consonant at the end of the 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). For the pattern recogniser, we implement ANNs feed-forward networks and Elman recurrent networks architectures. ANNs is also used by several other studies such as classification of body posture [24], face and expression recognition [25], and also earthquake prediction [26].

The rest of paper is organized as follows. In Section 2, we propose a system to recognise Javanese characters based on ANNs. In Section 3, we discuss the research method which we used for this research. Experimental results and discussions are presented in Section 4. And finally, we present our conclusion and future research direction in Section 5.

2. The Proposed System

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 [27]. For Elman recurrent networks training process, we implement a training algorithm that is provided by the author [28]. 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 a data preparation system from our previous research. This preparation system is included several image processing that is image segmentation, histogram balance, thinning and skew correction [15].

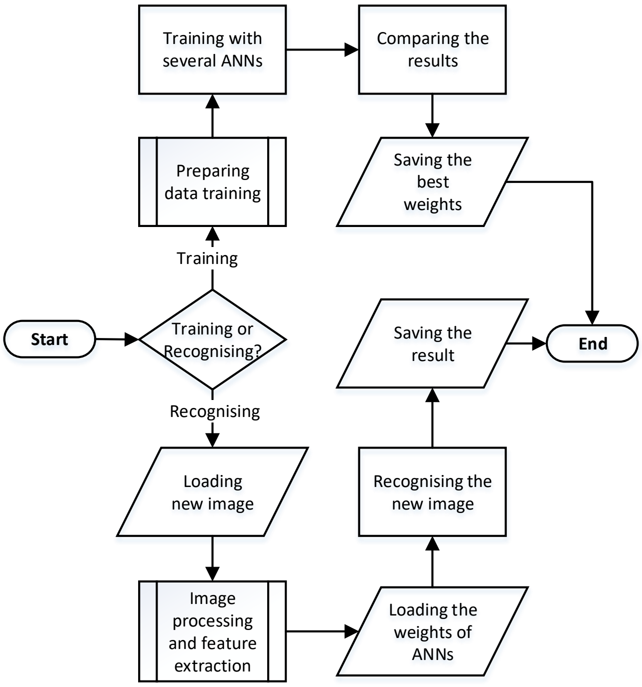

Figure 1. Overal system design

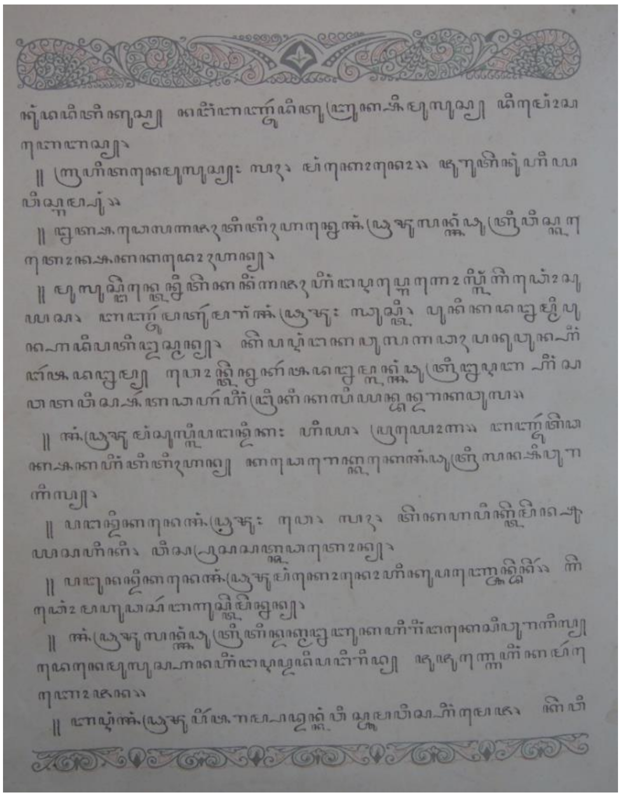

Figure 2. Old manuscript image

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.

3. Research Method

This research employs a quantitative research methodology [29] to tackle both research issues and questions. Panas & Pantouvakis [30] state that quantitative research methodology is associated with empirical research, simulation proposals formulation and performance of archival studies. Thus, the research methods that will be used in this project are simulation and experimentation. Simulation is chosen mainly due to its suitability for developing a new theory. Davis et al. [31] argue that simulation enables theoretical elaboration and exploration through computational experimentation. Additionally, Simulation attempts to derive semantic content from models which represent the actual systems [30]. The quantitative research methodology through simulation and experimentation will support our research.

Several steps have been undertaken to complete our research. First, the supporting theories of Javanese characters and the existing body of knowledge on machine learning and pattern recognition surveyed to gain a deeper understanding of the domain areas. Second, a model and prototype of Javanese characters recognition based on machine learning techniques were designed. Then experiments and simulations on these model would follow. In the experiments and simulations, we benchmarked the results with techniques we had found best in our previous research. Finally, the results of the experiments and simulations were documented and published.

4. Results and Discussion

We conducted the experiments for the purpose to search optimal parameters of the networks. In these experiments, we used the 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, it 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 the 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 of other algorithms both for data training and testing. Although for recognition of data training, the Induction Tree is perfect (accuracy = 100%), it tends to be overfitting. This is proved by its performance for data testing is the worst.

5. 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 using one hidden layer and eighty-eight hidden neurons outperforms the performance of other recogniser algorithms. Thus we will implement this ANNs method for our further research on the same topic. For the next research, 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.

Acknowledgement

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. 019/SP2H/LT/PTUPT/LPPM-UKP/III/2018, fiscal year 2018.

References

[1] Saddhono K. *Membangun Kearifan Lokal Melalui Karya Sastra dan Budaya Daerah [Jawa].* Seminar Nasional Bahasa Dan Budaya 2017. Denpasar. 2017; 8-17. ISBN: 978-602-50777-0-8.

[2] Marzuki. Tradisi dan Budaya Masyarakat Jawa dalam Perspektif Islam. *Kajian Masalah Pendidikan dan Ilmu Sosial "INFORMASI"*. 2006; 32(1): 1-13.

[3] Budhi G.S, Adipranata R. Java Characters Recognition Using Evolutionary Neural Network and Combination of Chi2 and Backpropagation Neural Network. *International Journal of Applied Engineering Research.* 2014; 9(22): 18025-18036.

[4] Setiawan A, Sulaiman A. M. Hanacaraka: Aksara Jawa dalam Karakter Font dan Aplikasinya sebagai Brand Image. *Ornamen.* 2015; 12(1): 33-47.

[5] Budhi G.S, Adipranata R. Handwritten Javanese Character Recognition Using Several Artificial Neural Network Methods. *Journal of ICT Research and Applications*. 2015; 8(3): 195-212.

[6] Gunarwan A., Kasus-Kasus Pergeseran Bahasa Daerah: Akibat Persaingan dengan Bahasa Indonesia?. *Linguistik Indonesia.* 2006; 24(1): 95-113.

[7] Ali L. Iktisar Sejarah Ejaan Bahasa Indonesia. Jakarta: Pusat Pembinaan dan Pengambangan Bahasa. 1998.

[8] Erawati N. K. R., Interpretasi Segmen Bunyi Bahasa Jawa Kuno: Analisis Speech Analyzer dan Fitur Distingtif. *Aksara.* 2017; 29(2): 225-238.

[9] Undang-Undang Republik Indonesia. Nomor 11 Tahun 2010. *Cagar Budaya.* M. o. L. a. H. Rights. 2010.

[10] Sularso P, Maria Y. Upaya Pelestarian Kearifan Lokal Melalui Ekstrakurikuler Karawitan di Smp Negeri 1 Jiwan Tahun 2016. *Citizenship Jurnal Pendidikan Pancasila dan Kewarganegaraan*. 2017; 5(1): 1-12.

[11] Suryono A, Pelestarian Aspek Kesemestaan dan Kesetempatan dalam Arsitektur Bangsal Sitihinggil di Kraton Yogyakarta. *RUAS*. 2016; 14(2): 1-10.

[12] Sunarni, Pelestarian Lingkungan Sosial Budaya Melalui Peningkatan Prestasi Belajar Bahasa Jawa dalam Materi Unggah-Ungguh. *GeoEco.* 2016; 2(1): 88-102.

[13] Budhi G.S, Adipranata R, Thedjakusuma R. *Java Characters Word Processing.* The 3rd International Conference on Soft Computing, Intelligent System and Information Technology Bali, Indonesia. 2012.

[14] Adipranata R, Budhi G. S, Liliana, Sebastian B. *Comparison Between Shape-Based And Area-Based Features Extraction for Java Character Recognition*. The 2nd Management and Innovation Technology International Conference. Bangkok, Thailand. 2015.

[15] Liliana, Soephomo S, Budhi G. S, Adipranata R. *Segmentation of Hanacaraka Character Using Double Projection Profile and Hough Transform*. The 8th EIA International Conference on Big Data Technologies and Applications. Gwangju, South Korea. 2017.

[16] Budhi G. S, Adipranata R. *Comparison of Bidirectional Associative Memory, Counterpropagation and Evolutionary Neural Network for Java Characters Recognition.* The 1st International Conference on Advanced Informatics: Concepts, Theory and Applications. Bandung, Indonesia. 2014.

[17] Budhi G. S, Adipranata R, Sebastian B, Liliana. *The use of Probabilistic Neural Network and ID3 Algorithm for Java Character Recognition.* The 2nd Management and Innovation Technology International Conference. Bangkok, Thailand. 2015.

[18] Astuty R. Y, Kusuma E. D, *Pengenalan Aksara Jawa Menggunakan Digital Image Processing.* The 2nd Informatics Conference. Jakarta, Indonesia. 2016; 32-35.

[19] Nurmila N, Sugiharto A, Sarwoko E. A. Backpropagation Neural Network Algorithm for Java Character Pattern Recognition. *Jurnal Masyarakat Informatika.* 2010; 1(1): 1-10.

[20] Rismiyati K, Nurhadiyatna A, *Deep Learning for Handwritten Javanese Character Recognition.* The 1st International Conference on Informatics and Computational Sciences. Semarang, Indonesia. 2017; 59-64.

[21] Wibowo M. A, Soleh M, Pradani W, Hidayanto A. N, Arymurthy A. M. *Handwritten Javanese Character Recognition Using Descriminative Deep Learning Technique*. The 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering. Yogyakarta, Indonesia. 2017; 325-330.

[22] Dewa C. K, Fadhilah A. L, Afiahayati A. Convolutional Neural Networks for Handwritten Javanese Character Recognition. *Indonesian Journal of Computing and Cybernetics Systems (IJCCS).* 2018; 12(1): 83-94.

[23] Arifianto T, Segmentasi Aksara pada Tulisan Aksara Jawa Menggunakan Adaptive Threshold. *Smatika.* 2017; 7(1): 1-5.

[24] Tripathy R.K, Acharya A, Choudhary S.K, Sahoo S.K. Artificial Neural Network based Body Posture Classification from EMG Signal Analysis, *Indonesian Journal of Electrical Engineering and Informatics.* 2013; 1(2): 59-63.
[25] Khandait S.P, Thool R. C., Khandait P.D. Hybrid Facial Geometry Algorithm for facial feature Extraction and Expression Recognition using ANFIS and BPNN. *Bulletin of Electrical Engineering and Informatics.* 2013; 2(1): 11-22.
[26] Bhandarkar T, Vardaan K, Satish N, Sridhar S, Sivakumar R, Ghosh S. Earthquake Trend Prediction using Long Short-Term Memory RNN. *International Journal of Electrical and Computer Engineering (IJECE).* 2019; 9(2): 1304-1312.
[27] Rumelhart D. E, Hinton G. E, Williams R. J. Learning Internal Representations by Error Propagation. James M.L. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press; 1986: 1: 318-362.
[28] Elman J. L, Finding Structure in Time. *Cognitive Science.*1990; 14: 179-211.
[29] Jongker J, Pennink B. W. Quantitative Research: Observing Through The Eyes Of The Researcher Using A Closed Research Question. The Essence of Research Methodology: A Concise Guide for Master and PhD Students in Management Science. Berlin Heidelberg: Springer-Verlag; 2010; 65-75.
[30] Panas A, Pantouvakis J. P. Evaluating Research Methodology In Construction Productivity Studies. The Built & Human Environment Review. 2010; 3(1): 63-85.
[31] Davis J. P, Eisenhardt K. M, Bingham C. B. Developing Theory Through Simulation Methods. Academy of Management Review. 2007; 32(2): 480-499.