The Generalized Learning Vector Quantization Model to Recognize Indonesian Sign Language (BISINDO)

Tri Handhika
Computational Mathematics Study Center
Gunadarma University
Depok, Indonesia
trihandika@staff.gunadarma.ac.id

Ilmiyati Sari
Computational Mathematics Study Center
Gunadarma University
Depok, Indonesia
ilmiyati@staff.gunadarma.ac.id

Revaldo Ilfestra Metzi Zen
Metra Digital Media
Telekomunikasi Indonesia
Jakarta, Indonesia
revaldo.ilfestra@mdmedia.co.id

Dewi Putrie Lestari
Computational Mathematics Study Center
Gunadarma University
Depok, Indonesia
dewi_putrie@staff.gunadarma.ac.id

Murni
Computational Mathematics Study Center
Gunadarma University
Depok, Indonesia
murnipskm@staff.gunadarma.ac.id

***Abstract*—There is a fundamental difference between image and gesture recognition where image recognition only works against one frame while gesture recognition works on a sequence of frames. It means that the accuracy formulas implemented on each issue are different. The accuracy of the image recognition is calculated based on the prediction accuracy of each frame, while gesture recognition is based on each sequence of frames. The incompatibility of using these accuracy formulas generate the misleading outputs and interpretation. Thus, the classification model used also needs to be adjusted with this problem. In this paper, we use GLVQ model as a classification algorithm based on machine learning approach to recognize the gestures of Indonesian sign language (BISINDO). However, this algorithm is used to classify every single frame so it needs to be modified by adding a new function for a sequence of frames, e.g. mode. In addition, there is a parameter known as the number of prototypes that affects the accuracy of the model. Based on the results of this research, GLVQ model with mode function has a higher degree of accuracy when compared with Hidden-Markov Model (HMM) in recognizing BISINDO. However, it is necessary to specify a more appropriate function instead of mode which is not give uniquely results. We also know that the increasing number of prototypes does not increase the accuracy significantly. In fact, the increasing number of prototypes used can increase the computational time.**

***Index Terms*—GLVQ, gesture recognition, BISINDO, machine learning, accuracy, prototype**

I. Introduction

Sign language is a language in which communication between people are made by visually transmitting the sign patterns to express the meaning [1]. It has its own vocabulary and syntax which is purely different from spoken/written languages [2]. Sign language is more to combine the different gesture, shape and movement of hand, body and facial expression where each of them has special assigned meaning [3].

Sign language is used by deaf and hard-hearing people for effective communication tool between their own community and with other people. In different part of the world, the different sign languages are used. It depends on the spoken language and culture of that particular place [4]. For example in USA, American Sign Language (ASL) is used while in England, the deaf use British Sign Language (BSL). Similarly, Indian Sign Language (ISL), Japanese Sign Language (JSL), and French Sign Language (FSL) [3].

Currently, there are two models of sign language used in Indonesia namely *Sistem Isyarat Bahasa Indonesia* (SIBI) and *Bahasa Isyarat Indonesia* (BISINDO). SIBI is more impractical and unnatural for the deaf because it follows Indonesian spoken language grammar structure. On the other hand, BISINDO uses some expressions for translating a word from Indonesian spoken language to represent its context [6]. For those reasons, in this paper we choose BISINDO to our research. In fact, there is still a difficulty in communication between the deaf and ordinary people who does not know about sign language. Therefore, the ordinary people needs a software application that can translate sign language into a spoken/written languages.

Researches in sign language recognition become more and more popular during the past decades. In the last several years, there has been an increased interest among the researchers in the field of sign language recognition to introduce means of interaction from human–human to human–computer interaction [5]. Several studies have been conducted to build an automatic sign language translator through computer vision technology based on gesture recognition. Sign language recognition has emerged as one of the important area of research in gesture

This research was supported by PUPT 2018 Research Grant, Ministry of Research, Technology, and Higher Education, Indonesia to the Gunadarma University No. 014.18/LP/UG/V/2018.

recognition. We used Microsoft Kinect XBox as a recording device [6]–[14] to recognize gestures. It has various sensor features that can receive multi-modal gesture inputs such as face, fingers, hands, forearms, upper arms, and shoulders [6].

In the introductory phase, different methods were used, e.g. Hidden-Markov Model (HMM), Dynamic Time Wrapping (DTW), Support Vector Machine (SVM), or k-Nearest Neighbor (kNN) [7]. Most of researchers used HMM for sign language recognition. Bhoir et. al. [8] showed that HMM is the most frequent tool for sign language recognition through hand gesture based on the shape parameters. It is a statistical model that has been successfully applied for spatial-temporal processes with finite number of states. Ghotkar et. al. [7] proposed hand gesture recognition for few subset of ISL. They used ten state HMM based on the skeleton joint information obtained by Kinect sensor. This algorithm had been tested on the four persons who performed 20 words of ISL for total 800 training set with an average of accuracy is 89.25%. Parcheta et. al. [9] also proposed a Spanish Sign Language recognition system. This work extends previous works by augmenting the data size and work on phrase rather than just a word of Spanish sign language. They also used HMM for recognizing this sign language and then compared with other classification techniques.

Rakun et. al. [10] proposed the first part of the automatic Indonesian Sign Language (SIBI) into text translation system. They combined a Kinect camera, Discrete Cosine Transform, Cross Correlation Function, and classification algorithm called Generalized Learning Vector Quantization (GLVQ). They obtained a high degree of accuracy in their experiment to create a simple system for recognizing alphabets A to Z and numbers 1 to 10 in SIBI. On another research, Rakun et. al. [7] proposed a model to recognize SIBI. They used GLVQ combined with WEKA data mining tools for implementing Random Forest training algorithm. The highest accuracy of their experiment results is 96,67%.

There is a difference between the accuracy calculations performed by Handhika et. al. [6] and Rakun et. al. [10], [11]. The accuracy in Rakun et. al. [10], [11] is calculated based on the suitability of predicted label for each frame on all gestures. Meanwhile, the accuracy in Handhika et. al. [6] is calculated based on the suitability of predicted label on each gestures which is a sequence of frames. In this paper, we follow the accuracy calculations used by Handhika et. al. [6] to recognize the gestures of BISINDO by considering that the words in BISINDO can be recognized in the form of a sequence of frames instead of just one frame.

Handhika et. al. [6] develop a translator model of BISINDO through computer vision technology, i.e. Microsoft Kinect XBox, and translation machine using HMM with optimal number of hidden states. They used skeleton data from Kinect sensor for feature extraction such as movement of the shoulders, upper arms, forearms, and hands. They got accuracy around 60% for their experiment to recognize the gesture of BISINDO. The average of accuracy is relatively small under the sequence of frames framework. Therefore, we tried to use another method to increase the average of accuracy for BISINDO's translation machine. Fig. 1 shows the technology framework proposed to help communication between the deaf and ordinary people by using an automatic BISINDO translator machine using the data provided by the Microsoft Kinect XBOX. The deaf's gestures is recorded by the Kinect camera and after the raw image processing, the automatic BISINDO translator provides the corresponding words in the spoken languages as an output. We use GLVQ model as a classification algorithm based on machine learning approach to recognize the gestures of BISINDO.

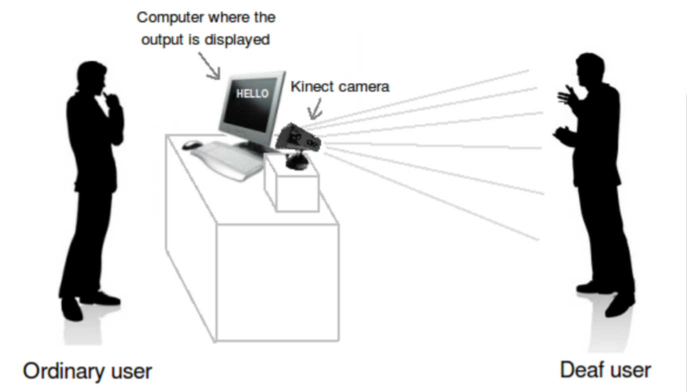


Fig. 1. Technology framework of automatic BISINDO translator [12].

II. Literature Review

The main concept in this paper is to implement the GLVQ classification algorithm into the problem of gesture recognition. The following is the fundamental theory of GLVQ which is the generalized form of Learning Vector Quantization (LVQ). LVQ is a family of classification algorithms which learning prototypes representing class regions which defined by hyperplanes for some of prototypes [15].

Let $\boldsymbol{\theta}_k = \left[\theta_{1_k}, \theta_{2_k}, \cdots, \theta_{L_k}\right]^T \in R^L$, are L-dimensional input sample of size N with $k = 1, 2, \cdots, N$ and superscript "T" referring to transpose. Given a training data set, $C_i, i = 1, 2, \cdots, K$, where K is the number of class labels. $\mathbf{w}^j = \left[w^{j^1}, w^{j^2}, \cdots, w^{j^L}\right]^T$ is the weight vector of the j-th class label C_j where the output class label which corresponds to a winner is assigned. It is determined the closest weight vector, $\mathbf{w}^*$, as the winning output vector , i.e. $\mathbf{w}^* = \arg d\left(\boldsymbol{\theta}_k, \mathbf{w}^j\right)$, where $d\left(\boldsymbol{\theta}_k, \mathbf{w}^j\right)$ is a distance measure between $\boldsymbol{\theta}_k$ and $\mathbf{w}^j$.

The sample margin of LVQ is hard to compute and numerically unstable [16]. Therefore, GLVQ solves this limitation by proposing a cost function to maximize the margin as follow [17]:

$$E = \Sigma_{i=1}^{N} \phi\left(\mu\right), \tag{1}$$

where $\phi\left(\cdot\right)$ is a logistic sigmoid function and μ is the relative distance difference

$$\mu\left(\boldsymbol{\theta}_i\right) = \frac{d^+ - d^-}{d^+ + d^-}. \tag{2}$$

where $d^+ = d(\theta_i, w^+)$ and $d^- = d(\theta_i, w^-)$ are the Euclidean distance of data point θ_i from its closest prototype w^+ having the same class label and w^- having a different class label, respectively.

III. Research Methodology

A. Skeleton Features

We use Microsoft Kinect XBox for collecting skeleton data for various features such as hands, forearms, upper arms, and shoulders. The skeleton data are obtained by previous research consists of 25 root words of BISINDO recorded five times each [6]. This recording involves two deaf people (male and female) performance from Pusat Layanan Juru Bahasa Isyarat Indonesia, Jakarta. To standardize the experiment, we transformed the extracted skeleton data into angles between shoulder-center and each hands, wrists, elbows, and shoulders [6]. These eight shoulder-center joint angles (hand-right, hand-left, wrist-right, wrist-left, elbow-right, elbow-left, shoulder-right, and shoulder-left) are processed separately using formula (3) and (4) such that there will be 16 angles as skeleton features for each frame, eight angles for each X-axis and Z-axis.

$$\theta_1 = \tan^{-1}\left(\frac{z_1 - z_2}{x_1 - x_2}\right) \tag{3}$$

$$\theta_2 = \tan^{-1}\left(\frac{y_1 - y_2}{z_1 - z_2}\right) \tag{4}$$

where θ_1 and θ_2 are the angles to the X-axis and Z-axis, respectively [11].

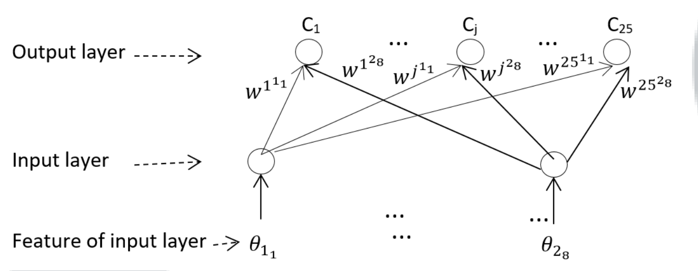


Fig. 2. GLVQ scheme for this research.

B. GLVQ Classification Algorithm

For a sample 16-dimensional input column vector $\boldsymbol{\theta} = [\theta_{1_1}, \theta_{1_2}, \cdots, \theta_{1_8}, \theta_{2_1}, \theta_{2_2}, \cdots, \theta_{2_8}]^T$ and given 25 class labels (words) $C_{i=1,2,\cdots,25}$, where the weight column vector of the j-th word C_j is $\mathbf{w}^j = \left[w^{j^{1_1}}, w^{j^{1_2}}, \cdots, w^{j^{1_8}}, w^{j^{2_1}}, w^{j^{2_2}}, \cdots, w^{j^{2_8}}\right]^T$, the predicted word which corresponds to a winner (the one which corresponds with input vector in the best way) is assigned as shown in Fig. 2. The winning output vector is determined as $\mathbf{w}^* = \arg d\left(\boldsymbol{\theta}_k, \mathbf{w}^j\right)$, where $d\left(\boldsymbol{\theta}_k, \mathbf{w}^j\right)$ is the squared Euclidean distance between $\boldsymbol{\theta}_k$ and $\mathbf{w}^j$. It is adjusted by equation (5) as follow

$$\mathbf{w}^*_{k+1} = \mathbf{w}^*_k \pm \xi\left(\boldsymbol{\theta}_k - \mathbf{w}^*_k\right), \tag{5}$$

for some convergence conditions, e.g. $\mathbf{w}^{*j}_k = \mathbf{w}^{*j}_{k-1}$ for $j = 1, 2, \cdots, 25$ and $k = 1, 2, \cdots, N_{train}$ where N_{train} is the sample size per epoch. The parameter ξ represents the learning rate decreasing with the number of iterations/epochs of training [18]. The sign "±" is taken "+" when $\boldsymbol{\theta}_k$ has been correctly classified, otherwise "−" such that the winning weight vector is driven toward the data when class label is correctly identified, or vice versa.

Training of GLVQ classifiers using equation (1) involves optimizing a cost function given on equation (6) which relates correctly classified samples to particular class weight vectors [18].

$$E = \frac{1}{2}\Sigma_{\forall k} f\left(\mu_k\right), \tag{6}$$

where $f(u) = \frac{1}{1+\exp^{-u}}$ for a measure of proximity $\mu_k = \mu\left(\boldsymbol{\theta}_k = \frac{d_k^+ - d_k^-}{d_k^+ + d_k^-}\right)$ as mentioned on equation (2). Dissimilarity measures $d_k^+ = d\left(\boldsymbol{\theta}_k, \mathbf{w}^*_k = \mathbf{w}^+\right) = \|\boldsymbol{\theta}_k - \mathbf{w}^+\|^2$ and $d_k^- = d\left(\boldsymbol{\theta}_k, \mathbf{w}^*_k = \mathbf{w}^-\right) = \|\boldsymbol{\theta}_k - \mathbf{w}^-\|^2$ are the squared distances of $\boldsymbol{\theta}_k$ to the closest prototype $\mathbf{w}^+$ and $\mathbf{w}^-$, respectively. Therefore, the weight update is then implemented as

$$\mathbf{w}^*_{k+1} \leftarrow \mathbf{w}^*_k \pm \xi\left(\mu_k\right)\left(\boldsymbol{\theta}_k - \mathbf{w}^*_k\right). \tag{7}$$

We can see that GLVQ algorithms on equation (7) adopt varying $\xi\left(\mu_k\right) = \frac{\partial f}{\partial \mu_k}\frac{d^{\pm}}{\left(d_k^+ + d_k^-\right)^2}$ for improving the accuracy with respect to sample $\boldsymbol{\theta}_k$ [19].

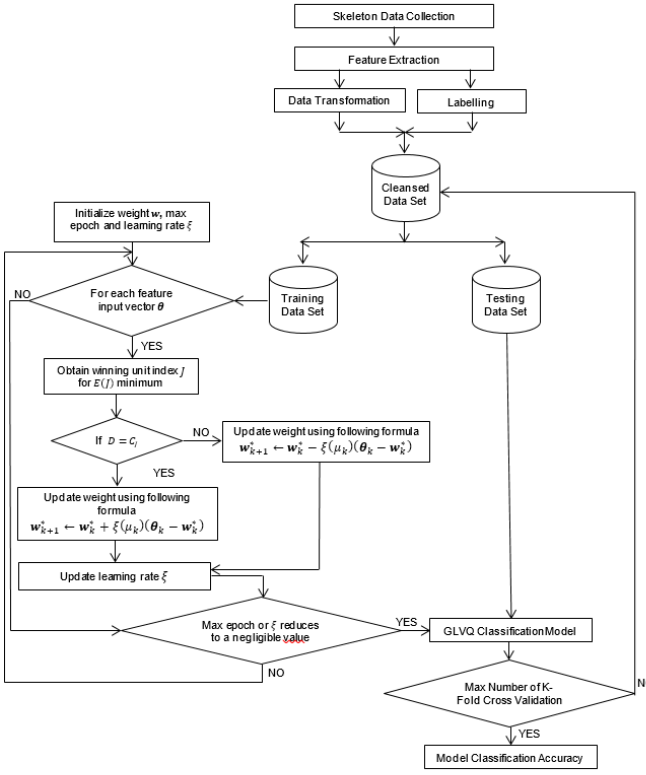
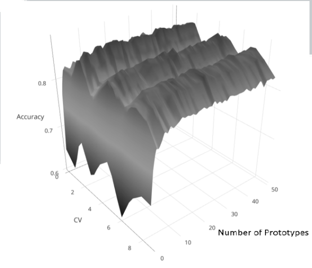
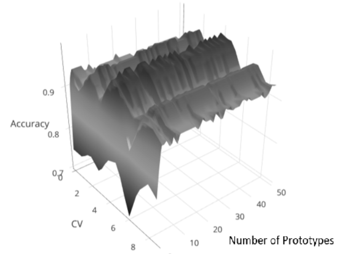
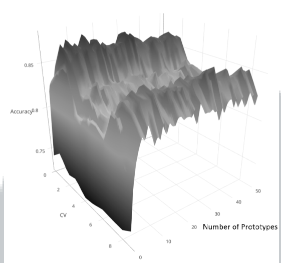
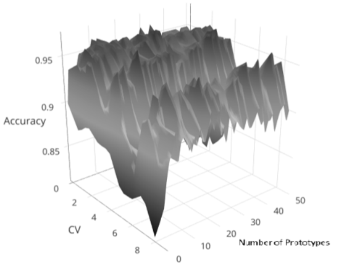
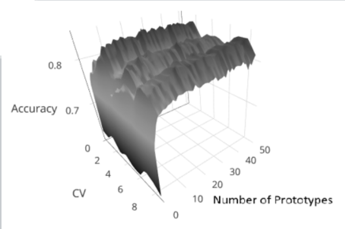
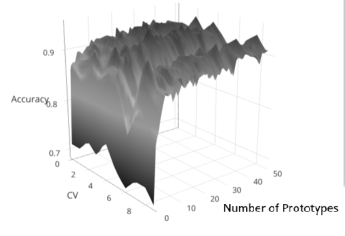

Skeleton Data Collection
Feature Extraction
Data Transformation
Labelling
Cleansed Data Set
Initialize weight w, max epoch and learning rate ξ
Training Data Set
Testing Data Set
NO
For each feature input vector θ
YES
Obtain winning unit index J for $E(J)$ minimum
If $D = C_j$
NO
Update weight using following formula $w^*_{k+1} \leftarrow w^*_k - \xi(\mu_k)(\theta_k - w^*_k)$
YES
Update weight using following formula $w^*_{k+1} \leftarrow w^*_k + \xi(\mu_k)(\theta_k - w^*_k)$
Update learning rate ξ
Max epoch or ξ reduces to a negligible value
YES
NO
GLVQ Classification Model
Max Number of K-Fold Cross Validation
NO
YES
Model Classification Accuracy

Fig. 3. Flowchart of this research.

TABLE I
THE ACCURACY COMPARISON OF BISINDO RECOGNITION: GLVQ MODEL VS (GLVQ + MODE) MODEL

Sex	Accuracy per K-Fold CV per Number of Prototypes	
	GLVQ Model (per Frame)	**(GLVQ + Mode) Model (per Sequence of Frames)**
Male		
Female		
Mixed		

C. GLVQ for Gesture Recognition

In this research, we use GLVQ algorithm to classify 25 sample words of BISINDO based on machine learning approach. Therefore, we divide the cleansed extracted skeleton features data and its labels into training and testing data sets. Training process using GLVQ classification algorithm (7) produce a prototype for each word. We can also determine the number of prototypes used which gives the highest degree of accuracy in our experiment. Finally, we use squared Euclidean distance formula for testing process. Fig. 3 shows the flowchart of this research. To evaluate the model, we use K-fold cross-validation (CV) [20]. We also calculate the average of each accuracy of K models to determine the accuracy of the model. Note that a word in BISINDO is represented by a sequence of frames, while GLVQ works for every single frame. It allows multiple frames of a word in BISINDO to have different predicted results. In this research, mode is used to generate word prediction from the sequence of frames.

IV. RESULTS AND DISCUSSION

We repeat the procedure on Fig. 3 three times of experiments, i.e. male, female, and mixed. We run GLVQ model with maximum number of prototypes is 50 for each experiment for selecting the best model via CV procedure. Table I shows the accuracy comparison of BISINDO recognition between GLVQ model (per frame) and (GLVQ + mode) model (per sequence of frames). We can see that most of the accuracy obtained by (GLVQ + mode) model is higher than GLVQ model for each experiment. Table II is derived by Table I. It shows the average of accuracy comparison of BISINDO recognition between GLVQ model and (GLVQ + mode) model. Both models shows that male performer has the highest degree of accuracy than other two experiments as summarized on Table III. However, it appears that the increasing number of prototypes does not increase the accuracy significantly. In fact, the increasing number of prototypes used can increase the computational time. The GLVQ model with mode function has a higher degree of accuracy when compared with HMM in recognizing BISINDO as obtained by Handhika et. al. [6]. Fig. 4 shows the accuracy of each word in the dataset using (GLVQ + mode) model. Most of the words in the dataset can be well predicted using the methods proposed in this research. The word “makan” (eat) is a rather difficult word to predict in all experiment. The word “apa” (what) and “gemuk” (fat) are also difficult to be recognized by this model for both female and mixed experiments.

TABLE II
AVERAGE OF ACCURACY COMPARISON OF BISINDO RECOGNITION: GLVQ MODEL VS (GLVQ + MODE) MODEL

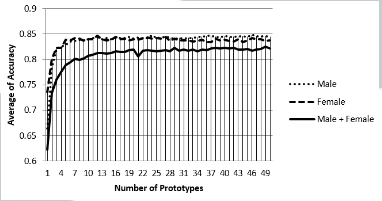
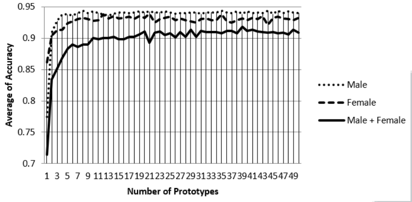
GLVQ Model (per Frame)	**(GLVQ + Mode) Model (per Sequence of Frames)**
	

TABLE III
THE HIGHEST DEGREE OF ACCURACY SUMMARY FOR EACH EXPERIMENT IN THIS RESEARCH

Sex	GLVQ Model		(GLVQ + Mode) Model	
	Number of Prototypes	**Accuracy**	**Number of Prototypes**	**Accuracy**
Male	45	84.829%	8	94.375%
Female	12	84.6903%	22	93.975%
Mixed	49	82.5575%	40	91.8125%

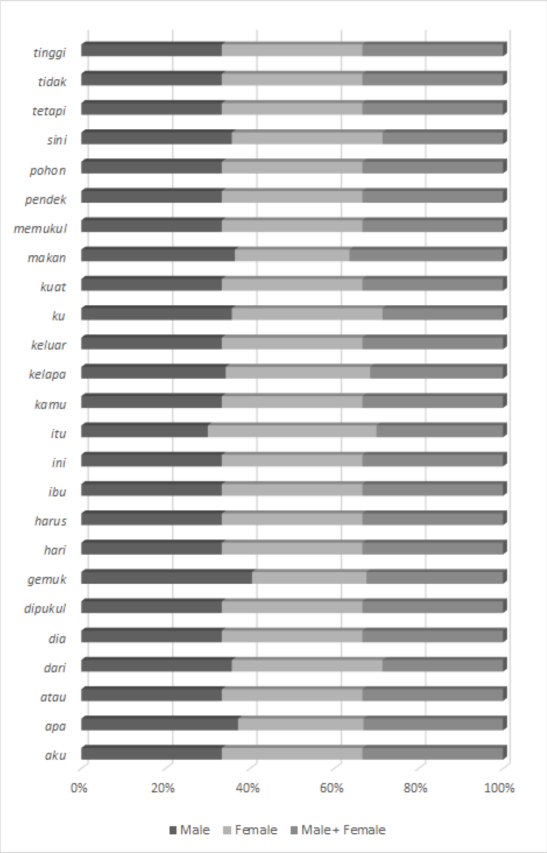


Fig. 4. The accuracy of (GLVQ + mode) model for each word for each experiment.

V. CONCLUSION AND FUTURE WORKS

This paper has shown that there are differences in accuracy that can be misleading in terms of interpretation. Based on the results of this research, GLVQ model with mode function has a higher degree of accuracy when compared with HMM in recognizing BISINDO. However, it is necessary to specify a more appropriate function instead of mode which is not give uniquely results. Based on the results, we know that the increasing number of prototypes does not increase the accuracy significantly. In fact, the increasing number of prototypes used can increase the computational time. In addition, the optimal number of prototypes used on the GLVQ model needs to be determined in recognizing BISINDO.

ACKNOWLEDGMENT

This research was supported by PUPT 2018 Research Grant, Ministry of Research, Technology, and Higher Education, Indonesia to the Gunadarma University No. 014.18/LP/UG/V/2018.

REFERENCES

[1] J. Singha and K. Das, "Recognition of Indian sign language in live video," International Journal of Computer Applications, vol. 70, pp. 17–22, May 2013.

[2] A. K. Sahoo, G. S. Mishra, and K. K. Ravulakollu, "Sign language recognition: State of the art," ARPN Journal of Engineering and Applied Sciences, vol. 9, pp. 116–134, February 2014.

[3] P. Pandey and V. Jain, "Hand gesture recognition for sign language recognition: A review," International Journal of Science, Engineering and Technology Research (IJSETR), vol. 4, pp. 464–470, March 2015.

[4] V. Adithya, P. R. Vinod, and U. Gopalakrishnan, "Artificial neural network based method for Indian sign language recognition," IEEE Conference on Information and Communication Technologies (ICT), pp. 1080–1085, April 2013.

[5] P. S. Rajam and G. Balakrishnan, "Real time Indian sign language recognition system to aid deaf and dumb people," 13th International Conference on Communication Technology (ICCT), pp. 737–742, September 2011.

[6] T. Handhika, R. I. M. Zen, Murni, D. P. Lestari, and I. Sari, "Gesture recognition for Indonesia Sign Language (BISINDO)," Journal of Physics: Conf. Series, vol. 1028, pp. 1-8, June 2018.

[7] A. Ghotkar, P. Vidap, and K. Deo, “Dynamic hand gesture recognition using hidden Markov model by Microsoft Kinect sensor,” International Journal of Computer Applications, vol. 150, issue 5, pp. 5–9, September 2016.

[8] P.P. Bhoir, A.V. Nandyhyhh, D.S. Bormane, and R.R. Itkarkar, “Sign language recognition using hidden Markov model,” IOSR Journal of Electronics and Communication Engineering (IOSR-JECE), vol. 1, issue 18, pp. 88–92, 2015.

[9] Z. Parcheta and C.D.M. Hinarejos, “Sign language gesture recognition using HMM,” Pattern Recognition and Image Analysis IbPRIA, vol. 10255, pp. 419-426, May 2017.

[10] E. Rakun, M. F. Rachmadi, Andros, and K. Danniswara, “Spectral domain cross correlation function and generalized learning vector quantization for recognizing and classifying Indonesia sign language,” Proceeding of IEEE Advanced Computer Science and Information Systems (ICACSIS), pp. 213–218, December 2012.

[11] E. Rakun, M. Andriani, I.W. Wiprayoga, K. Danniswara, and A. Tjandra, “Combining depth image and skeleton data from kinect for recognizing words in the sign system for Indonesian language (SIBI [Sistem Isyarat Bahasa Indonesia]),” Proceeding of IEEE Advanced Computer Science and Information Systems (ICACSIS), pp. 387–392, September 2013.

[12] D. M. Capilla, Sign language translator using Microsoft Kinect XBOX 360TM, MSc Thesis VIBOT. Knoxville-USA, Department of Electrical Engineering and Computer Science: University of Tennessee, 2012.

[13] H.V. Verma, E. Aggarwal, and S. Chandra, “Gesture recognition using Kinect for sign language translation,” Proceeding of the IEEE Second International Conference on Image Information Processing (ICIIP), India, Waknaghat, pp. 96–100, December 2013.

[14] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, and M. Zhou, “Sign language recognition and translation with Kinect,” The 10th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR), China, Shanghai, 2013.

[15] T. Kohonen, “Self–Organizing Maps,” Springer-Verlag, New York, 1997.

[16] K. Crammer, R. Gilad-Bachrach, A. Navot, and A. Tishby, “Margin analysis of the LVQ algorithm,” Advances in Neural Information Processing Systems, vol. 15, pp. 462-469, 2002.

[17] A. Sato and K. Yamada, “Generalized Learning Vector Quantization,” Advances in Neural Information Processing Systems, vol. 8, pp. 423-429, 1996.

[18] T. Temel, “A New Classification Algorithm: Optimally Generalized Learning Vector Quantization (OGLVQ),” Neural Network World, vol. 27, issue 6, pp. 569-576, December 2017.

[19] M. Kaden, M. Lange, D. Nebel, M. Riedel, T. Geweniger, T. Villmann, “Aspects in Classification Learning-Review of Recent, Developments in Learning Vector Quantization,” Foundations of Computing and Decision Sciences, vol. 39, issue 1, pp. 79–105, May 2014.

[20] Y. Jung and J. Hu, “A K-fold averaging cross-validation procedure,” Journal of Nonparametric Statistics, vol. 27, issue 2, pp. 167–179, February 2015.