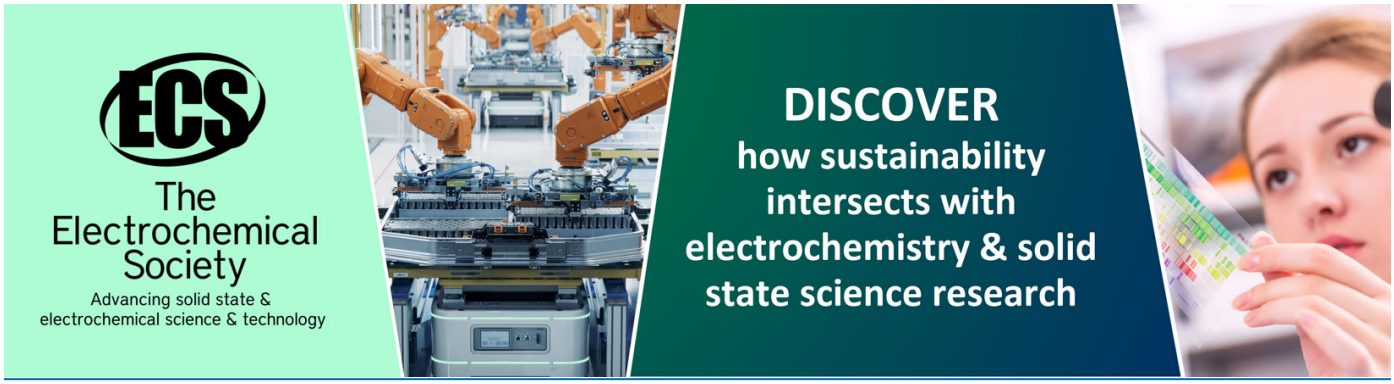
Gesture recognition for Indonesian Sign Language (BISINDO)



IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012173 doi:10.1088/1742-6596/1028/1/012173

Gesture recognition for Indonesian Sign Language (BISINDO)

T Handhika[1*], R I M Zen[2], Murni[1], D P Lestari[1] and I Sari[1]

[1]Computational Mathematics Study Center, Gunadarma University, Depok, 16424, Indonesia
[2]Metra Digital Media, Telkom Indonesia, Jakarta, 12780, Indonesia

*trihandika@staff.gunadarma.ac.id

Abstract. Sign language is different from spoken language that emphasizes both the audio and verbal aspects. There are two sign languages applicable in Indonesia, namely Indonesian Signal System (SIBI) and Indonesian Sign Language (BISINDO). SIBI converts spoken Indonesian language into sign language and follows the Indonesian spoken language's grammatical structure complete with prefix and suffix. In contrast to SIBI, BISINDO translates one word from the Indonesian spoken language in accordance with its context followed by an expression showing the ongoing events. We choose BISINDO rather than SIBI in line with the deaf' people's suggestions and requests out there to make BISINDO as an official Indonesian sign language to replace SIBI. This research aims to develop a translator model of BISINDO through computer vision technology such as Microsoft Kinect XBox and machine translation using Hidden-Markov Model (HMM) with optimal number of hidden states. We utilize skeleton data from Kinect sensor for feature extraction. There are four kinds of skeleton features used in this study consisting of the movement of the shoulders, upper arms, forearms, and hands. The experiment results by using this methodology successfully recognize the gesture of BISINDO with an accuracy is around 60%.

1. Introduction

Sign language is different from spoken language that emphasizes both the audio and verbal aspects. The deaf people only use visual aspects of sign language as a medium in communicating using gestures such as hands, shoulders, eyes, eyebrows, and other facial expressions. The differences between these two languages make it difficult for the deaf to blend in the society since their average literacy skills are equivalent to a 10-year-old kid [1]. Although the deaf still have the visual ability, but this ability cannot be compared to people who have the hearing ability that learn written language as the visual representation of spoken language [2].

Sign language had a complex system. A term could have different meaning in sign language depending on several aspects, such as the shape of the hand, movement of the hand and arms, also the part of body where the gestures are articulated (parallel to the eye or parallel to the chin) [3, 4]. Other aspects that can cause the difference of meaning in sign language are body posture, facial expression also rhythm and the speed of hand movement. There is no sign language that has been applied internationally. Each country has their own sign language [5, 6], e.g. Great Britain has British Sign Language while United States of America has American Sign Language.

Currently, there is a dualism in sign language in Indonesia. There are two sign languages used in Indonesia, i.e. *Sistem Isyarat Bahasa Indonesia* (SIBI) and *Bahasa Isyarat Indonesia* (BISINDO). Although SIBI is used as the official sign language at school, but it is not commonly used by the deaf on their daily life. SIBI turns the Indonesian spoken language into sign language and follows its grammar structure along with prefix and suffix. Those make SIBI more impractical and unnatural for the deaf. On the other hand, BISINDO translates a word from Indonesian spoken language followed by an expression that represents its context. We choose BISINDO rather than SIBI in line with the deaf' people's suggestions and requests out there to make BISINDO as an official Indonesian sign language to replace SIBI. This research aims to develop a translator model of BISINDO through computer vision technology such as Microsoft Kinect Xbox [7, 8, 9].

Microsoft Kinect Xbox is equipped with various sensor features that can receive multi-modal inputs such as gestures of shoulders, upper arms, forearms, hands, fingers, and face. Therefore, it can be used to recognize sign language. Hidden-Markov Model (HMM) is one of the most popular model represented the gesture recognition problem [10, 11, 12]. However, we cannot directly implement HMM into the problem without combining the model with other algorithm such as clustering or distance sequence learning for classification purposes.

2. Research methodology

We collect the gestures data performed by two deaf people (male and female) from *Pusat Layanan Juru Bahasa Isyarat Indonesia*, Jakarta, who used BISINDO in their daily life. They tried to demonstrate 25 root words of BISINDO, recorded five times each (see Figure 1 and Figure 2 for the sample data). We used Microsoft Kinect XBox for collecting skeleton data for various features such as shoulders, upper arms, forearms, and hands from Kinect sensor. The extracted skeleton data then transformed into angles between shoulder-center and each shoulders, elbows, wrists, and hands. The following are the formula to compute each angle to the X plane (θ_1) and Z plane (θ_2), respectively [12]:

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

$$\theta_2 = tan^{-1}\left(\frac{y_1 - y_2}{z_1 - z_2}\right) \quad (2)$$

We also need to label every frame recorded to perform the accuracy analysis of the model.

Figure 1. Sample data of root word *hari* (day) by male performer.

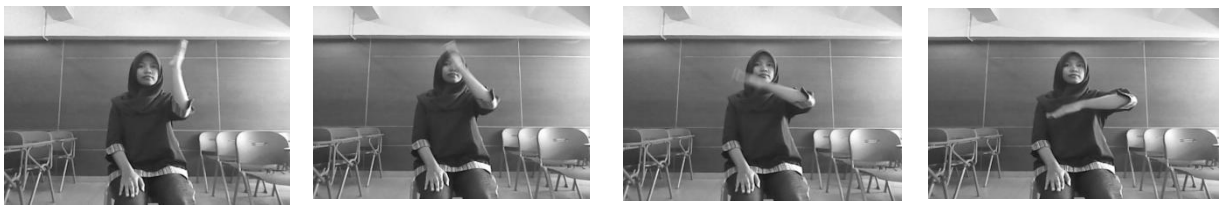


Figure 2. Sample data of root word *hari* (day) by female performer.

Figure 3 shows the flowchart of experiment conducted in this study. The cleansed data set is divided into two groups, i.e. training and testing data sets, after data transformation and labeling had been performed simultaneously. Furthermore, all frames in the training data set that have been labeled are classified into several hidden states tested using the Hidden-Markov Model (HMM) [13, 14, 15] with Gaussian densities [16]. The number of optimal hidden states is obtained by maximizing the Bayesian Inference Criterion (BIC) [17] for a maximum trial number of hidden states (N). Next, The HMM with optimal number of hidden states then implemented for the testing data set. We use K-fold cross-validation [18] as a method for model evaluation. The accuracy of the model is obtained by taking the average of each accuracy of K models. It was the rate of correct (incorrect) predictions made by each model over their own testing data set [19]. Moreover, the prediction is obtained by determining the root words in training data set that has the shortest modified Levenshtein distance [20] with the root words in testing data set. We repeated this procedure for three times of experiments to the performers who is: (i) male; (ii) female; and (iii) combination of male and female.

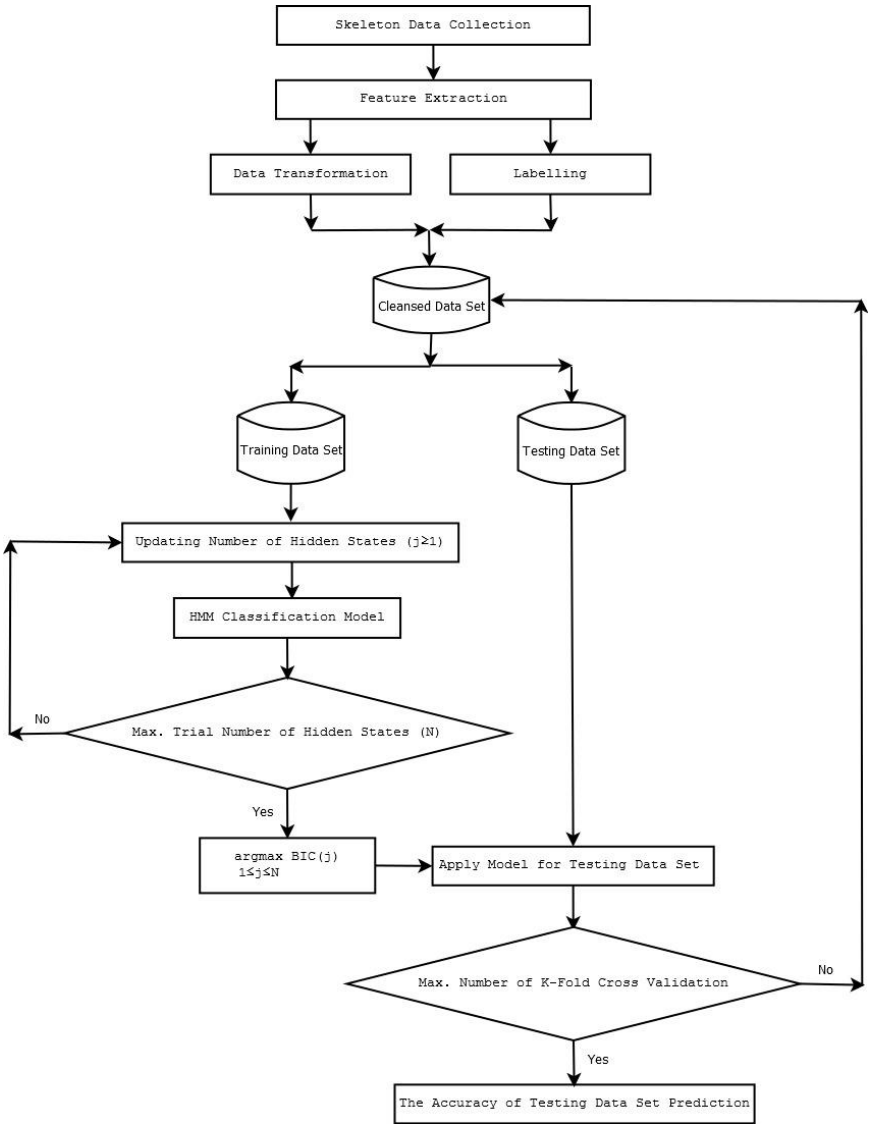


Figure 3. Flowchart of this research

3. Results and discussion

In this research, we tried to take the maximum trial number of hidden states (N = 50). First, log-probability of the model for each number of hidden states is estimated. Figure 4 shows the sample of log-probability estimation for Experiment (iii) when we trained the first three recording for each performer. The maximum BIC for this sample is reached when the number of hidden states is 25. By using 25 hidden states, then HMM is applied to the testing data set for this sample. The result shows that the accuracy of the model is 65%.

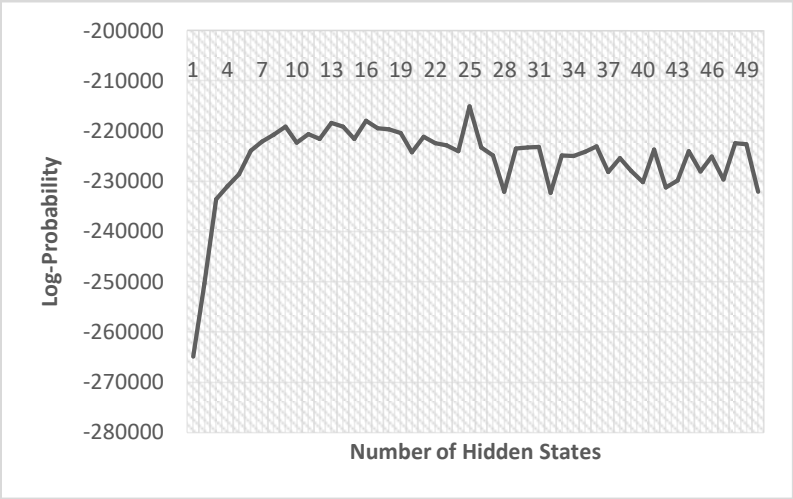


Figure 4. Sample of log-probability for experiment (iii)

The cross-validation procedure is then applied to each experiment for model selection. We run the experiment by training the combination of three records randomly for each performer (ten times per experiment or $K = 10$). The new findings in this study is about the distribution of the optimal number of hidden states to perform BISINDO modelling as shown in Table 1. This finding indirectly disputes the methodology of the previous research which determining the number of both root and inflectional words as the number of hidden states used in gesture recognition [12]. However, the average of accuracy does not seem depend on the optimal number of hidden states being used. This may be due to the methodology for measuring the distance between the hidden states sequences of words in testing data set with training data set.

Table 1. The frequency of optimal number of hidden states

Experiment	Optimal Number of Hidden States	Frequency	Average of Accuracy (%)
(i)	8	4	60.9375
	9	3	63.75
	13	2	75
	14	1	73.75
(ii)	6	1	56.25
	8	1	60
	9	3	57.9167
	10	1	58.75
	11	1	67.5
	13	2	59.375
	15	1	56.25

Table 2. The frequency of optimal number of hidden states (continued).

Experiment	Optimal Number of Hidden States	Frequency	Average of Accuracy (%)
(iii)	10	1	60
	23	1	63.75
	24	1	60.625
	25	1	65
	29	2	66.5625
	30	1	76.25
	31	1	71.875
	32	1	66.875
	35	1	70

Based on the experiment results as shown in Table 2, we know that the Experiment (iii) that combine both of male and female performers has the best performance. Although the average of error for Experiment (i) is smaller than Experiment (ii), but it has the largest standard deviation. The BISINDO performed by the female performer has the highest error rate but it is relatively constant in each recording. It can be caused by the speed of movement affecting the sequence of hidden states. Meanwhile, the BISINDO performed by the male performer has better performance than female, but there is some inconsistency on his gestures that cause the standard deviation become large. Therefore, combining both of male and female data sets can improve the accuracy of Hidden-Markov Model.

Table 3. Performance of each experiment

Experiment	(Average $\pm$ Standard Deviation) of Error (%)
(i)	34.125 ± 7.3112
(ii)	40.875 ± 3.3564
(iii)	33.25 ± 4.7335

The accuracy of each word in each experiment can be further analysed as shown in Figure 5. All of words tested can be recognized by Experiment (iii). Conversely, we can see that there are two words tested cannot recognized by both Experiment (i) and (ii), i.e. *pendek* (short) and *makan* (eat). Moreover, the Experiment (ii) cannot recognize the other six words tested such as *sini* (here), *keluar* (out), *kamu* (you), *ini* (this), *harus* (must) and *aku* (I). Meanwhile, another word tested that is not recognized by Experiment (i) is *dipukul* (hitted). The other ten words tested recognized by all of experiments with relatively same accuracy, they are *tinggi* (tall), *pohon* (tree), *kuat* (strong), *ku* (my), *kelapa* (coconut), *ibu* (mother), *hari* (day), *gemuk* (fat), *dari* (from) and *apa* (what).

4. Conclusion and future works

Gesture recognition for Indonesian sign language (BISINDO) conducted in this study has been performed systematically through *K*-fold cross-validation by combining: (i) computer vision technology, i.e. Microsoft Kinect Xbox; (ii) Hidden-Markov Model (HMM) with Gaussian densities and the optimal number of hidden states obtained by maximizing the Bayesian Inference Criterion (BIC); and (iii) the shortest modified Levenshtein distance criteria. We repeated this procedure for three times of experiments to the performers who is: (i) male; (ii) female; and (iii) combination of male and female. The results show that it is not appropriate to take the number of words tested as the number of hidden states used for implementing HMM to the gesture recognition problem. The experiment results by using this methodology successfully recognize the gesture of BISINDO with an accuracy is around 60%-70%.

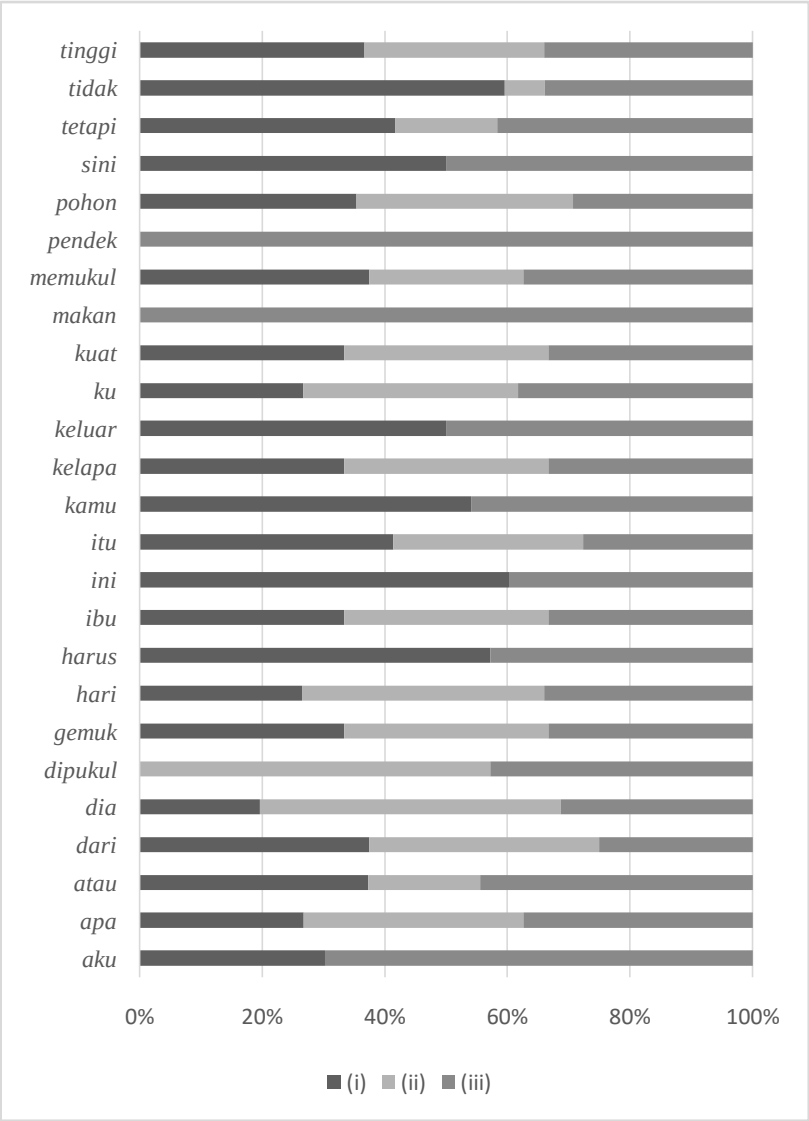


Figure 5. The accuracy comparison of each word for each experiment

This research should be developed to recognize not only the root words of BISINDO but also a series of words in the formation of a sentence. This is necessary because it allows for the changes in the gesture depending on the position of the word in the sentence and its context. The gesture recognition for a sentence of BISINDO can then be implemented into an automatic translator application. It is expected to decrease the dependency of the deaf to the sign language interpreter in bridging the communication between the deaf and those who have the ability for hearing. Moreover, there is a face recognition problem for the additional features of facial expressions in representing the context of a sentence. In addition, we can add the other skeleton features, i.e. fingers, to improve the accuracy of the model. The feature selection techniques might be also applied for solving this accuracy issue. Finally, the average of accuracy in this research does not seem depend on the optimal number of hidden states being used. It can be improved by using the appropriate methodology for measuring the distance between the hidden states sequences of words in testing data set with training data set.

References

[1] Holt, J.A. 1993. Stanford achievement test – 8th Edition: Reading comprehension subgroup results. *American Annals of the Deaf*, 138(2), 172-175.
[2] Agris, U.v., Zieren, J., Canzler, U., Bauer, B., & Kraiss, K.F. 2008. Recent developments in visual sign language recognition. *Journal of Universal Access in the Information Society*, 6(4), 323-362.
[3] Stokoe, W.C. 2001. The study and use of sign language. *Sign Language Studies*, 1(4), 369-406.
[4] Stokoe, W.C. 2005. Sign language structure: An outline of the visual communication systems of the American deaf. *Journal of Deaf Studies and Deaf Education*, 10(1), 3-37.
[5] Satryawan, I., Wijaya, L.L., Isma, S.T.P., Suwiryo, A.I., Woodward, J., Sze, F., & Lee, J. 2014. *Bahasa Isyarat Jakarta Buku Pedoman Siswa 1 Tingkat 1*. Indonesia, Jakarta: Fakultas Ilmu Pengetahuan Budaya Universitas Indonesia.
[6] Satryawan, I., Wijaya, L.L., Isma, S.T.P., Suwiryo, A.I., Woodward, J., Sze, F., & Lee, J. 2014. *Bahasa Isyarat Jakarta Kamus Pendamping untuk Buku Pedoman Siswa 1 Tingkat 1.* Indonesia, Jakarta: Fakultas Ilmu Pengetahuan Budaya Universitas Indonesia.
[7] Verma, H.V., Aggarwal, E., & Chandra, S. 2013. Gesture recognition using Kinect for sign language translation. *Proceeding of the IEEE Second International Conference on Image Information Processing (ICIIP), India, Waknaghat*, 96-100.
[8] Chai, X., Li, G., Lin, Y., Xu, Z., Tang, Y., Chen, X., & Zhou, M. 2013. *Sign language recognition and translation with Kinect, The 10th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR)*. China, Shanghai.
[9] Rakun, E., Andriani, M., Wiprayoga, I.W., Danniswara, K., & Tjandra, A. 2013. 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, Indonesia, Bali: IEEE*.
[10] Starner, T., & Pentland, A. 1995. Visual recognition of American sign language using hidden-Markov models. *Proceeding of the International Workshop on Automatic Face and Gesture Recognition*.
[11] Alexandre, L.A., Sánchez, J.S., & Rodrigues, J.M.F. 2017. Sign language gesture recognition using HMM. *Proceeding of the 8th Iberian Conference of Pattern Recognition and Image Analysis (IbPRIA), Portugal, Faro*, 419-426.
[12] Rakun, E., Fanany, M.I., Wisesa, I.W.W., & Tjandra, A. 2015 A heuristic hidden markov model to recognize inflectional words in sign system for Indonesian language known as SIBI (Sistem Isyarat Bahasa Indonesia). *Proceeding of International Conference on Technology, Informatics, Management, Engineering & Environment (TIME-E), Indonesia, North Sumatra: IEEE*.
[13] Rabiner, L.R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceeding of the IEEE,* 77(2), 257-285.
[14] Dymarski, P. 2005. *Hidden Markov Models Theory and Applications*. Croatia, Rijeka: Intech,
[15] Cappé, O., Moulines, E., & Rydén, T. 2005. *Inference in Hidden Markov Models.* USA, New York: Springer.
[16] Euler, S. 1992. Clustering of Gaussian densities in hidden Markov models. *Speech Recognition and Understanding: Recent Advances, Trends and Applications, NATO ASI Series*, 75, 83-88.
[17] Jeebun, S., Ballgobin, R.R., & Al-ani, T. 2015. Optimal number of states in Hidden-Markov models and its application to the detection of human movement. *University of Mauritius Research Journal,* 21, 438-469.
[18] Jung, Y., & Hu, J. 2015. A K-fold averaging cross-validation procedure. *Journal of Nonparametric Statistics*, 1-13.
[19] Kohavi, R., & Provost, F. 1998. Glossary of terms: special issue on applications of machine

learning and the knowledge discovery process. *Machine Learning*, 30, 271-274.

[20] Nyirarugira, C., Choi, H.R., Kim, J.Y., Hayes, M., & Kim, T.Y. 2013. Modified Levenshtein distance for real-time gesture recognition. The 6th International Congress on Image and Signal Processing (CISP), China, Hangzhou: IEEE, 974-979.