Journal of Korea Multimedia Society Vol. 25, No. 9, September 2022(pp. 1257-1265)
https://doi.org/10.9717/kmms.2022.25.9.1257

Fast Convergence GRU Model for Sign Language Recognition

Barathi Subramanian[†], Bekhzod Olimov[††], Jeonghong Kim[†††]

ABSTRACT

Recognition of sign language is challenging due to the occlusion of hands, accuracy of hand gestures, and high computational costs. In recent years, deep learning techniques have made significant advances in this field. Although these methods are larger and more complex, they cannot manage long-term sequential data and lack the ability to capture useful information through efficient information processing with faster convergence. In order to overcome these challenges, we propose a word-level sign language recognition (SLR) system that combines a real-time human pose detection library with the minimized version of the gated recurrent unit (GRU) model. Each gate unit is optimized by discarding the depth-weighted reset gate in GRU cells and considering only current input. Furthermore, we use sigmoid rather than hyperbolic tangent activation in standard GRUs due to performance loss associated with the former in deeper networks. Experimental results demonstrate that our pose-based optimized GRU (Pose-OGRU) outperforms the standard GRU model in terms of prediction accuracy, convergency, and information processing capability.

Key words: Deep Learning, Gesture Recognition, Human Pose Detection, OpenPose, Sign Language

1. INTRODUCTION

A sign language is an interactive, vision-based language with unique and complex linguistic rules. People with hearing impairments use various parts of their bodies as ways of communicating and exchanging feelings, ideas, and thoughts [1]. According to its geographic location, sign language differs linguistically from one region to another due to its unique linguistic structure [2]. Every country has developed its own sign language for communication between its deaf and hard-of-hearing populations [3]. Sign languages that are widely used include American sign language (ASL) in the US [4], British Sign Language (BSL) in the UK [5], and Korean sign language (KSL) in Korea [6]. In recent years, it has become increasingly important to deal with communication obstacles encountered by people with normal hearing and people with hard hearing [7]. According to a World Health Organization report, about 500 million people worldwide suffer from hearing loss [8].

SLR effectively bridges the communication gap between hearing and non-hearing communities, and it provides a new path for applications using human-computer interaction (HCI) [9]. Although,

※ Corresponding Author : Jeonghong Kim, Address: (41566) Daehag-ro 80, Buk-gu, Daegu, Korea, TEL : +82-53-950-5551, FAX : +82-53-950-5551, E-mail : jhk@knu.ac.kr
Receipt date : Jul. 18, 2022, Approval date : Aug. 19, 2022
[†] Dept. of Computer Science and Engineering, Graduate School, Kyungpook National University (E-mail : achu_samriti@yahoo.com)
[††] Dept. of Computer Science and Engineering, Graduate School, Kyungpook National University (E-mail : bekhzod.olimov@gmail.com)
[†††] Dept. of Computer Science and Engineering, Graduate School, Kyungpook National University
※ This study was supported by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (4199990214394). Also this study was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1I1A3043970).

it is still difficult to track [10] hand movements, obstruct hand movements [11], select features and learn the information in the given input data (image/text/audio) [12], despite different techniques. To address these drawbacks, we proposed a Pose-OGRU SLR system that discards the deep weighted reset gate from GRU cells and neglects the importance of past input integrated with a real-time human pose detection library called OpenPose. Moreover, we developed an SLR system that is simpler, faster, and more cost-effective by changing the activation function of candidate memory state of each GRU cell in the output layer. In summary, this study makes the following contributions:

• In this paper, we propose a novel and optimized GRU with a minimized gated cell architecture, by eliminating the reset gate of the standard GRU model. As a result, significant past content can be preserved without favoring current input exclusively and thereby accelerating the convergence rate, eliminating the gradient depletion problem, and improving the learning efficiency.

• In the candidate memory state, we replaced the hyperbolic tangent (Tanh) activation function with a sigmoid activation function in order to overcome the performance loss issue. Thus, Pose-OGRU requires less training time and performs better than other variants

The rest of this paper is organized as follows: Section 2 introduces existing SLR methods along with their limitations. Section 3 introduces and describes the proposed methodology. Section 4 presents the experimental settings, results and comparison of our method with other methods, and analyzes the model performance. Finally, Section 5 presents our contributions and outlines the future work.

2. RELATED WORKS

In recent years, artificial intelligence algorithms have been used to predict sign language gestures to help people with hearing impairments. The hard-of-hearing community requires more research to address limitations and improvements, despite the amount of research that has been conducted for SLR [13]. In this section, we review recent literature on deep learning SLR methods.

2.1 Visual and Pose Based Methods

To create holistic representations of all video frames utilizing raw RGB data as input representations, Convolutional Neural Networks (CNNs) were employed in the early research [14-16]. Then, temporal information has been encoded using recurrent neural networks such as the long short-term memory (LSTM) network , bidirectional LSTM network, gated recurrent unit network and transformers. For this task, 3D CNNs have also been utilized, as they can learn a combination of spatiotemporal and frame representations. Generally, action recognition appears to be easier when human poses are extracted from images or videos. In their first study, Yan et al. [17] proposed a spatio-temporal graph convolutional network for the recognition of action dynamics. A graph convolutional neural network was used to extract a sequence of body poses from individual video frames for SLR, as well as a 3D CNN in the latter work. These two studies have demonstrated that appearance-based representations can deliver comparable results.

2.2 OpenPose

OpenPose is a real-time multiple-person detection library that jointly detects the human body, face, and foot key points (in total 135 key points). As shown in Fig. 1, an OpenPose model estimates the pose of multiple people in real-time on a 2D image. OpenPose pipeline consist of two parts:

• Two tensors are inferred from a Neural Network: Part Confidence heatmaps and their pairwise relations (PAF, Part Affinity Field).

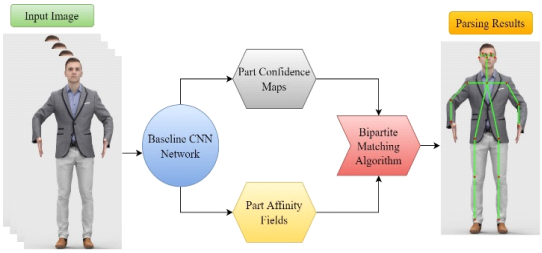

Fig. 1. Overview of OpenPose pipeline.

• Person instances are grouped by keypoints. A key point is extracted from the heatmap peaks, and the instances are grouped accordingly.

2.3 Standard GRU

It is common for computer vision problems to involve short and long-term sequence modeling and handling temporal dependencies among inputs. This type of sequential data can be efficiently managed and processed by recurrent neural networks (RNNs). The problem of vanishing and exploding gradients makes training RNNs difficult. GRUs have been used to solve gradients that vanishes and explodes in convention RNNs. GRU constructs candidate memory by using current inputs and previous memory state by using only two gates instead of three gates in LSTM. Thus, it is significantly less parameterized and minimizes the training time while maintaining dependency on information. The general structure of the standard GRU cell is shown in Fig. 2.

From Fig. 2, at each time step t, a GRU cell takes the contents of previous hidden state and present input operates them through reset and update gates, and passes the computed current state to the next time step. The general formulae of a standard GRU cell are as follows:

$$R_t = sig(W_t X + U_r H_{t-1} + B_r) \quad (1)$$

$$Z_t = sig(W_z X_t + U_z H_{t-1} + B_z) \quad (2)$$

$$N_t = \tanh(W_h X_t + (U_h H_{t-1} \oplus R_t)) + B_h \quad (3)$$

$$H_t = Z_t \oplus H_{t-1} + (1 - Z_t) \otimes N_t \quad (4)$$

An update gate is represented by the vector Z_t, a reset gate is represented by the vector R_t, and the current candidate memory is represented by H_t, which is computed as a linear interpolation between H_{t-1} and the previous candidate memory N_t. To limit the values between 0 and 1, Sigmoid(sig) activation function is used in both reset and update gates. N_t is computed from a hyperbolic tangent activation (tanh), $\otimes$ denotes the element-wise multiplication (Hadamard product). The current input into the network is X_t, and the trainable weights of feed-forward connections are W_z, W_r and W_h, whereas the recurrent connections weights are W_z, W_r, W_h and B_r, B_z, B_h are the bias vectors.

However, their gated structures lead to the omission of crucial content in a long sequence, even though standard GRUs handle long-term sequential data effectively and solve the vanishing gradient problem of RNNs. A novel GRU model that can sustain crucial content in long-term sequential data is presented in this paper in order to alleviate this problem.

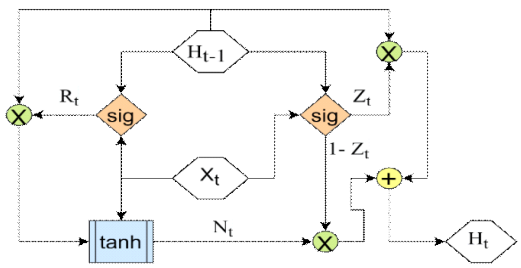

Fig. 2. General architecture of standard GRU cell.

3. PROPOSED METHODOLOGY

A three-stage methodology is proposed to accurately recognize sign gestures and translate them into text: data augmentation, feature extraction, and gesture recognition. General overview of the proposed SLR system architecture is shown in Fig. 3 respectively. The following sections provide a detailed explanation of the three stages of the proposed methodology.

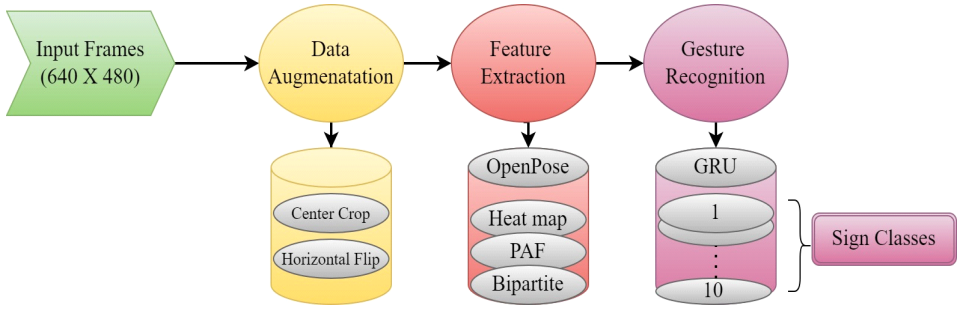

Fig. 3. General overview of our proposed Pose-OGRU.

3.1 Data Augmentation

An augmenting process creates variants of input data while maintaining naturalness, so that trained models generalize more effectively. Our video input is augmented as follows:

• Cropping Center - Video can only be cropped to the center using the Center Crop command. The frame was cropped by 40% based on experiments.

• Flipping horizontal - Rotating the frame of the video horizontally, so that a signer signing with the right hand appears to be signing with the left hand after flipping.

As a result, different versions of the same video can be obtained by passing it through these augmentations.

3.2 Feature Extraction

In general, a video is a high dimensional data. In order to extract the high dimensional features from the video, a pretrained OpenPose model is used first to extract lower-dimensional features. In this process, the augmented input video is passed frame wise and as a result, three sets of features are computed: *Key point vector*: A vector of x- and y- coordinates for each frame. *Pose video:* A frame-wise pixel map of all limbs. *PAF video:* A frame-wise pixel map with Part Affinity Fields (PAF) aggregations. As a result, a greedy bipartite matching algorithm is used to extract the poses for each person in an image as shown in Fig. 1.

3.3 Gesture Recognition

The extracted features are then passed through the GRU network for training and recognizing the sign gestures. As a major modification to conventional GRUs, we eliminate the reset gate and use sig activation function as a substitute for tanh activation function in equation 3.

3.3.1 Eliminating the reset gate and featuring sigmoid activation

In GRUs, R_t in equation 1 is used to determine whether the previous memory state H_{t-1} is relevant in calculating the candidate memory state H_t. Depending on the need, the network can either forward information from the previous state or exit completely. Especially when the previous memory state affects the computation of the candidate memory state, the reset gate can be fully opened to attain values closer to zero. In order to generate a candidate memory state for signing gesture recognition, each frame sequence in a video must maintain its memory state. Maintaining smooth continuity among frames in this way facilitates prolonging the retention of temporal information due to the fact that the previous memory state H_{t-1} is not sacrificed for the current input X_t. Removing the reset gate can result in a poorly computed candidate state, which in turn may affect the accuracy of the current memory state H_t. Since gesture recognition problems generally treat information from previous and current frames reasonably and un-

biasedly, the reset gate is unnecessary for our work. Discarding the reset gate from a GRU cell is achieved by removing equation 1 in GRU formulation and modifying equation 3 to the following form:

$$N_t = \tanh(W_h X_t + U_h H_{t-1}) + B_h \quad (5)$$

As a result, the GRU model structure are computationally lightweight. Fig. 4(a) shows the modification to be carried on the standard GRU cell. As seen in the diagram below, the blue-dashed box represents R_t which has been removed from standard GRU, while the light, pink-dashed box shows sig activation instead of tanh activation function.

Featuring sigmoid activation: Another modification we made to the standard GRU was to replace tanh activation function with sigmoid activation function as shown in Fig. 4(b). This is highlighted with a pink dashed box and thus the candidate memory N_t needs to be calculated as follows:

$$N_t = sig(W_h X_t + U_h H_{t-1}) + B_h \quad (6)$$

Tanh activation functions are ineffective when training feed-forward connections, especially in standard GRU, the exponential operation of this algorithm results in a vanishing gradient and also is computationally expensive. Thus, the candidate memory states are computed using sigmoid activation. Models trained with sigmoid activation are found to be superior for sign recognition in our experiments.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Dataset

Using a web camera in different directions and different lighting conditions, we created a real-time dataset of 30 videos with 45 frames each for 10 different sign gestures (10 classes). Each video frame of size 640 by 480 pixels respectively. Our training and testing datasets are formed by dividing the collected dataset in the ratio 80:20.

4.2 Experimental Setup

Simulations were conducted using python 3.7 version on a desktop computer equipped with 32GB RAM and an i7 core processor running on windows 10 with a 64-bit operating system. A web camera was used to capture the input images which had a resolution of 720 pixels/30 frames per second and was captured using RGB color format.

4.3 Learning Curve Analysis

Data-driven algorithms use learning curves to diagnose problems that occur over time. In this study different models are evaluated against the testing accuracy and loss and the performance results are plotted as shown in Fig. 5 and Fig. 6 respectively. On the basis of the learning pattern of each curve in the learning graph, we are able to observe that the convergence rate of RNN,

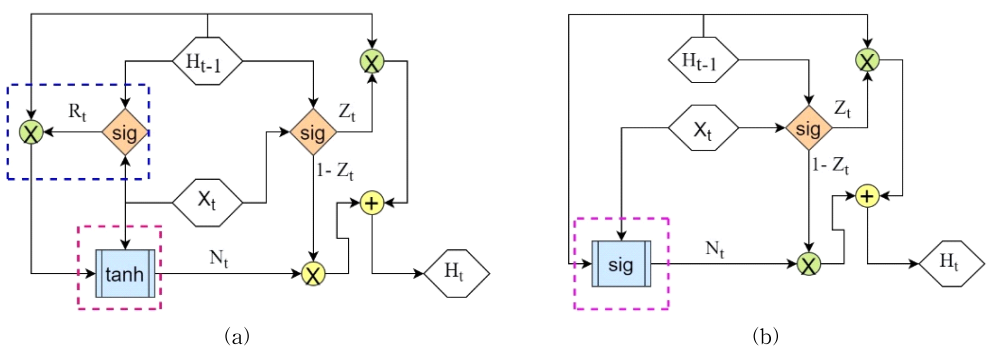

Fig. 4. Modified structure of GRU cell. (a) Modifications on the standard GRU cell and (b) Proposed structure of GRU cell.

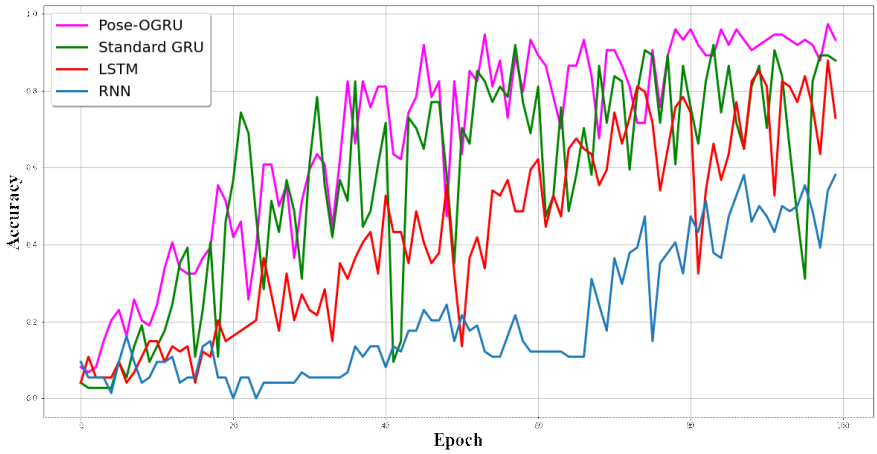

Fig. 5. Learning graph of different models with their test accuracy for 100 epochs.

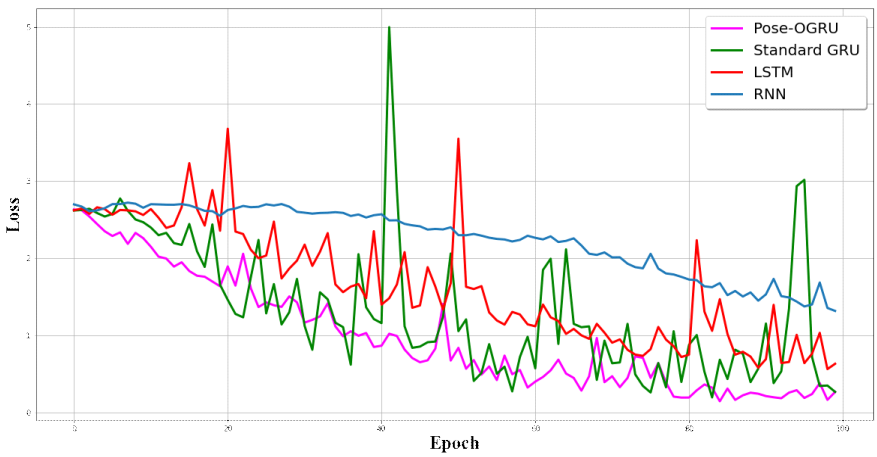

Fig. 6. Learning graph of different models with their test loss for 100 epochs.

LSTM, and standard GRU is slower than our proposed Pose-OGRU. Approximately 50 epochs after the training our Pose-OGRU model converged with high accuracy without any over-fitting or under-fitting issues. As with Fig. 6, comparing the loss curve of Pose-OGRU to other models reveals that the Pose-OGRU model converges rapidly with the minimum loss occurring around epoch 70.

Additionally the results of our comparison of models with respect to training and test accuracy show that our proposed Pose-OGRU model achieves the highest training and test accuracy of 99.4% and 95.1% as sown in Table 1 respectively.

4.4 Performance Evaluation Metrics

We evaluated the performance of the model us-

Table 1. Different model comparison in terms of training and test accuracy.

Model		RNN	LSTM	Standard GRU	Pose-OGRU
Accuracy(%)	Training	90.0	93.4	94	99.4
	Testing	89.9	85.2	90.0	95.1

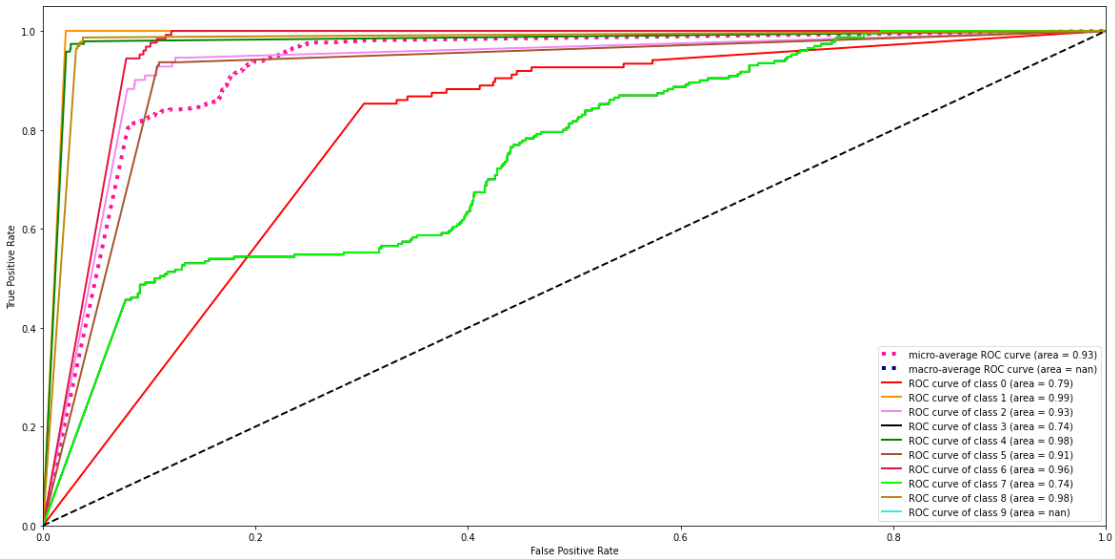

Fig. 7. General ROC curve of our Pose-OGRU model.

ing a Receiver Operator Characteristic (ROC) curve. Based on the ratio of true positive rate (TPR) to false positive rate (FPR), a ROC curve can be constructed. TPR is calculated by dividing all positive observations by the number of observations which were correctly predicted as positive as shown in equation 7. In the same way, FPR of observations are calculated as the proportion of observations incorrectly predicted as positive as shown in equation 8.

$$FPR = FalsePositive / TrueNative + FalsePositive \quad (7)$$

$$TPR = TruePositive / Truepositive + FalseNative \quad (8)$$

Generally, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test. From the ROC graph plot shown in Fig. 7, we can observe that the model trained well and learned the information efficiently except for two classes.

5. CONCLUSION

In this study, we proposed a system for word-level SLR system combined with a real-time human pose detection library and a minimized version of the GRU model. Each gate unit is optimized by discarding the depth-weighted reset gate in GRU cells and considering only current input. In addition, we used sigmoid activation rather than tanh activation function in standard GRUs due to the performance loss associated with the former in deeper networks. Our experimental results demonstrated that our proposed pose-OGRU outperforms the standard GRU model in terms of prediction accuracy, convergency, and information processing capability. Although our proposed model outperformed RNN and its two basic variants it still lacks in competing with the latest sequential deep learning models with limited dataset. So our future goal is to expand this work by collecting more data and develop an efficient SLR system that combines our proposed model with the latest existing sequential model.

REFERENCE

[1] N. Aloysius, M. Geetha, and P. Nedungadi, "Incorporating Relative Position Information in Transformer-Based Sign Language Recognition and Translation," *IEEE Access*,

Vol. 9, pp. 145929-145942, 2021.

[2] D. Li, C.R. Opazo, X. Yu, and H. Li, "Word-Level Deep Sign Language Recognition from Video: A New Large-Scale Dataset and Methods Comparison," *IEEE Winter Conference on Applications of Computer Vision, WACV 2020*. pp. 1448-1458, 2020.

[3] R.A. Kadhim and M. Khamees, "A Real-Time American Sign Language Recognition System Using Convolutional Neural Network for Real Datasets," *TEM Journal*, Vol. 9, No. 3, pp. 937-943, 2020.

[4] A. Wadhawan and P. Kumar, "Deep Learning-Based Sign Language Recognition System for Static Signs," *Neural Computing Applications*, Vol. 32, No. 12, pp. 7957-7968, 2020.

[5] P. Kumar, H. Gauba, P.P. Roy, and D.P. Dogra, "Coupled HMM-Based Multi-Sensor Data Fusion for Sign Language Recognition," *Pattern Recognition Letters*, Vol. 86, pp. 1-8, 2017.

[6] R. Elakkiya and K. Selvamani, "Subunit Sign Modeling Framework for Continuous Sign Language Recognition," *Computers and Electrical Engineering*, Vol. 74, pp. 379-390, 2019.

[7] O. Koller, "Quantitative Survey of the State of the Art in Sign Language Recognition," *arXiv Preprint*, arXiv:2008.09918, 2020.

[8] T.R. Gadekallu, M. Alazab, R. Kaluri, P.K.R. Maddikunta, S. Bhattacharya, K. Lakshmanna, and M. Parimala, "Hand Gesture Classification Using A Novel CNN-Crow Search Algorithm," *Complex & Intelligent Systems*, Vol. 7, No. 4. pp. 1855-1868, 2021.

[9] B. Kanisha, V. Mahalakshmi, M. Baskar, K. Vijaya, and P. Kalyanasundaram, "Smart Communication Using Tri-Spectral Sign Recognition for Hearing-Impaired People," *Journal of Supercomputing*, Vol. 78, No. 2. pp. 2651-2664, 2022.

[10] Patil, A., Kulkarni, A., Yesane, H., Sadani, M., Satav, P., "Literature Survey: Sign Language Recognition Using Gesture Recognition and Natural Language Processing," *In: Sharma, N., Chakrabarti, A., Balas, V.E., Bruckstein, A.M. (eds) Data Management, Analytics and Innovation. Lecture Notes on Data Engineering and Communications Technologies*, Vol. 70, 2021.

[11] R. Rastgoo, K. Kiani, and S. Escalera, "Hand Sign Language Recognition Using Multi-View Hand Skeleton," *Expert Systems with Applications*, Vol. 150, p. 113336, 2020.

[12] R.C. Chen, C. Dewi, S.W. Huang, and R.E. Caraka, "Selecting Critical Features for Data Classification Based on Machine Learning Methods," *Journal of Big Data*, Vol. 7, No. 1, Article number 52, 2020.

[13] D. Naglot and M. Kulkarni, "Real Time Sign Language Recognition Using the Leap Motion Controller," *International Conference on Inventive Computation Technologies (ICICT)*, Vol. 2016, pp. 1-5, 2016.

[14] N.C. Camgöz, O. Koller, S. Hadfield, and R. Bowden, "Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020 c*, pp. 10020 - 10030, 2020.

[15] O. Koller, N.C. Camgoz, H. Ney, and R. Bowden, "Weakly Supervised Learning with Multi-Stream CNN-LSTM-HMMs to Discover Sequential Parallelism in Sign Language Videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 9, pp. 2306-2320, 2020.

[16] B. Saunders, N.C. Camgoz, and R. Bowden, "Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks," *International Journal of Computer Vision*, Vol. 129, No. 7. pp. 2113-2135, 2021.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D : Generic Features for

Video Analysis C3D," *arXiv Preprint*, arXiv: 1412.0767v1, 2015.

Barathi Subramanian

received her Bachelor's and Master's degree in mathematics with computer application obtained from Bharathiar University, Coimabatore, India, in 2013 and 2015, respectively. She is currently pursuing her Ph.D degree in computer science and engineering from Kyungpook National University, South Korea. Her research interest includes sign language recognition and computer vision using machine learning and deep learning techniques.

Bekhzod Olimov

received his B.S. degree in economics from Fergana Polytechnic Institute, Uzbekistan and obtained M.S. degree from Yeungnam University, South Korea in 2014 and 2018, respectively. He received the PhD degree from computer science and engineering department of Kyungpook National University, South Korea. His research interests are computer vision and pattern recognition using deep learning techniques.

Jeonghong Kim

received the B.S. and M.S degrees from Kyungpook National University, in 1986 and the Ph. D. degree from Chungnam National University, Daejeon, South Korea in 2001. He worked as a senior researcher at the Electronics and Telecommunications Research Institute from 1988 to 1996. He is currently working as a professor in School of Computer Science and Engineering, Kyungpook National University, South Korea. His research areas are biosignal processing and pattern recognition using deep learning techniques.