Master's Thesis
석사 학위논문

A Group Based Personalized Approach to Efficient Hand Gesture Recognition Using Sensor Fusion

Seongjoo Shin(신 성 주 辛 晟 柱)

Department of

Information and Communication Engineering

DGIST

2018

Master's Thesis
석사 학위논문

A Group Based Personalized Approach to Efficient Hand Gesture Recognition Using Sensor Fusion

Seongjoo Shin(신 성 주 辛 晟 柱)

Department of
Information and Communication Engineering

DGIST

2018

A Group Based Personalized Approach to Efficient Hand Gesture Recognition Using Sensor Fusion

Advisor: Professor Yongsoon Eun
Co-advisor: Professor Sang Hyuk Son
Co-advisor: Professor Youngmi Baek
by

Seongjoo Shin
Department of Information and Communication Engineering
DGIST

A thesis submitted to the faculty of DGIST in partial fulfillment of the requirements for the degree of Master of Science in the Department of Information and Communication Engineering. The study was conducted in accordance with Code of Research Ethics[1].

12. 21. 2017

Approved by

Professor Yongsoon Eun (signature)
(Advisor)

Professor Sang Hyuk Son (signature)
(Co-Advisor)

Professor Youngmi Baek (signature)
(Co-Advisor)

[1] Declaration of Ethical Conduct in Research: I, as a graduate student of DGIST, hereby declare that I have not committed any acts that may damage the credibility of my research. These include, but are not limited to: falsification, thesis written by someone else, distortion of research findings or plagiarism. I affirm that my thesis contains honest conclusions based on my own careful research under the guidance of my thesis advisor.

A Group Based Personalized Approach to Efficient Hand Gesture Recognition Using Sensor Fusion

Seongjoo Shin

Accepted in partial fulfillment of the requirements for the degree of Master of Science.

11. 21. 2017

Head of Committee ______________(signature)___
Prof. Yongsoon Eun

Committee Member ______________(signature)___
Prof. Sang Hyuk Son

Committee Member ______________(signature)___
Prof. Kyoung-Dae Kim

Committee Member ______________(signature)___
Prof. Youngmi Baek

MS/IC
201622009

신 성 주. Seongjoo Shin. A Group Personalized Approach to Efficient Hand Gesture Recognition Using Sensor Fusion. Department of Information and Communication Engineering. 2018. 29p. Advisors Prof. Yongsoon Eun, Co-Advisors Prof. Sang Hyuk Son, Co-Advisors Prof. Youngmi Baek

Abstract

Multimodal interface keeps evolving in order to better represent people’s intention. A gesture as a type of the multimodal interface is one of the effective ways for people to express their intention. Specially, hand gesture recognition provides an eidetic and convenient way of human-machine interaction (HMI).

In this thesis, we investigate the problems of dynamic hand gesture recognition and develop a Korean sign language (KSL) recognition system which can help many hearing and speech-impaired people communicate with the public.

To recognize sign language, the system should first determine the shape of the hand and the movement of the arm. Since sign language consists of a sequence of movements, it is difficult to distinguish a certain gesture from gestures (movements). To address this problem, the recognition system has to know the beginning and end of the gesture. To get the starting and ending points, we have defined the basic posture. The sign language also has various lengths of gestures. It is effective to make the fixed length input data (gestures) rather than predefine the length of each gesture for recognition.

Many attempts to study the hand gesture recognition commonly use various types of sensors such as cameras, electromyograms (EMG), glove sensors, and inertial measurement units (IMU). Inconvenience caused by their weight, the shapes uncomfortable to wear, and cumbersome calibration processes might decrease the usability of them. Wearable devices like smart watches and armbands can solve this problem. Furthermore, in order to improve recognition accuracy, the effective way is to exploit multiple heterogeneous sensors (both an EMG sensor and an IMU sensor) which can produce the redundant information to the same physical variable. It is necessary to pre-process before classification since it is important to classify the gesture using the values extracted from the sensor. We evaluated the performance of two different methods, min-max and z-score normalization.

Specially, we focus on the fact that EMG signals depends on physical features of people because the amount of muscle and the thickness of the fat layer are different for each person. Unfortunately, in the traditional recognition technique not to consider human physical features, since a single model is applied to all users, it does not guarantee the performance in terms of accuracy. To address these issues, we create group-dependent Neural Network (NN) models based on a sensor fusion technology. Our approach on group-dependent NN models is to separate the models so that people can use different models. People are experimentally divided into several groups according to persons’ data with similarity in body features after learning. We proved that the physical similarity exists in our created models.

Finally, We compare our model with models of Artificial neural networks (ANNs) including convolution neural networks (CNNs) and long short-term memory (LSTM) since the performance of those is high in the classification. The experimental results show that the proposed method has high accuracy (99.13% of CNN without dropout and 98.1% of CNN with dropout).

Keywords: Korean sign language; Electromyography; Hand gesture; Sensor fusion; Artificial Neural Network

List of Contents

List of Figures

List of Tables

I. INTRODUCTION

There are many traditional input/output (I/O) devices through which computers/machines easily communicate with human beings. With the increasing interest in augmented reality (AR) applications, this human-computer interaction technologies are evolving and new types of I/O devices are being developed to improve interaction, such as gloves, haptic devices, and head mounted devices (HMD). Specially, HMD called an eyewear display is a device which can display a sequence of images on the screen of human eye level. It is useful tool for the user to experience the real world. Since HMD generally uses a wireless controller to obtain human intention, including a clickable trackpad, a trigger, some buttons, and an IMU sensor, applications using HMD can only execute the preprogrammed functions corresponding to several buttons. In addition, users familiarize themselves with all operation prior to using such devices. If gestures as a non-verbal communication method are used for the intention input function instead of such controller, we including speech-impaired persons and deaf persons better enjoy the comfort and convenience of this interaction method. For example, we are already used to finger-based gestures for smartphones and tablets without any button. Especially, hand gestures, which is one of the most common types of the multimodal input, also allows humans to communicate naturally and intuitively. In terms of effective human-computer interaction, hand gestures is capable of easily getting much more information from users in diverse situations without any limitation of the input device.

The structural characteristics of hand gestures with a sequence of the hand movements causes two problems having a serious effect on the performance of hand gestures recognition. First, while we move our arms or hands to perform the next gesture after performing one gesture, it is difficult to distinguish between these two gestures (movements). To address this problem, we define a basic posture to better understand the start and end points of the given gesture. This defined basic posture represents no movement, and should be performed between the two different hand gestures in order to recognition them accurately and effectively. This is an effective way to distinguish between two gestures. Second, there are hand gestures of various lengths depending on what we want to express. We used an interpolation filter to solve this problem, which makes the gestures equal in length.

To recognize the hand gestures, various multiple sensors is commonly exploited. Here is another challenge we might face. In daily life, the body of a human being generates various bio-signals such as electrocardiography (ECG), electrooculography (EOG), electroencephalography (EEG) and electromyography (EEG) [1]. Specially, EMG is generated while a person uses a muscle. Since this measured signal can determine the intensity of muscle movement, the EMG sensors are used to recognize hand movements to make a shape. The hardware redundancy of the EMG sensors enables the recognition system to improve the reliability and the accuracy. An inertial measurement unit (IMU) is an electronic device which mainly has a gyroscope, and an accelerometer. It is commonly used to measure the movement and direction of an object in several wearable devices [2], [3]. Therefore, the IMU sensor can also be used to measure the movement of the arm to make a gesture. Sensor fusion technology focuses on combining various sensory data obtained from heterogeneous sensors. Such technology is used to drive towards convergence in various fields in order to satisfy the non-functional requirements of systems. Therefore, using sensor fusion is more effective way to improve accuracy, reliability and robustness of systems of recognition systems than exploiting the redundancy of homogeneous sensors. By using both the EMG sensor and the IMU sensor, while we take advantage of the sensor fusion technology, unpaired data from the differences of sampling frequencies of sensors disturbs the accurate fusion due to multi-view data. To address this problem, we select under-sampling being not complex process within the range with no distortion of sensory data, considering the real-time processing, among under and oversampling approaches [25]. Therefore, in order to equalize the sampling frequency, we down-sampled one sensor to match the sampling frequency of both sensors.

Finally, the hand recognition model using bio-signals measured by sensors should designed for the diversity of human physical features. It directly affects the performance of the designed recognition model. Since individual data obtained from people has different features, the accuracy of recognition might be low if one learning model is applied to all users. The reason is that: (1) The EMG signal can be measured differently depending on the arm's muscle and subcutaneous fat layer thickness. (2) The IMU value can have different values depending on the height of the person and the length of the arm. For this reason, we create several models that can be applied to individuals. This model, called GDM (group-dependent model), learns only the data of people in each group.

We focus on how to recognize Korean sign language (KSL) in real-time by addressing those technical challenges as described above. There are 2.5 million people with disabilities in Korea. 10% of them are hearing-disabled and speaking-disabled persons [4]. People with hearing impairments or speech impairments use sign language to communicate with others. Since people still lack knowledge of sign language, it is difficult for them to communicate with others. A system that recognizes and translates KSL is helpful for communicating with others. The system recognizes sign language, converts it into a natural language, and output it by using a speaker.

In this paper, we propose a method to accurately recognize hand gestures related to KSL, which is based on the fusion of EMG and IMU sensors and leverage artificial neural networks (ANNs) as a learning model. The main objective is to design a learning model and provide an optimized learning model for each individual group which consists of persons with similarity in body features. For that, the cross validation method is performed by testing one person’s data on another’s learning model in order to find people with similar data, and then they are grouped together.

The first technical contribution of this thesis is the development of the transformation method from numerical sensor data to image data in order to use the designed learning model with the CNN architecture since CNN achieves good performance in the 2D image recognition. In addition, after the transformation, we assess the two normalization methods to find out the effective pre-processing method.

The second contribution is the development of a novel group-dependent learning model. People are experimentally divided into some groups according to persons’ data with similarity in physical characteristics. It turns out that the experimental results of the models divided by the group are better, and we proved that the characteristics of the group’s people are similar, comparing with the real data of people.

The third contribution is that we provide the CPS (cyber-physical systems) design for tight interaction between sensors of the physical world and classification of the cyber world through networks. The CPS system is implemented as a prototype for the KSL translation services and is available anywhere and anytime.

The remainder of this thesis is organized as follows. In Section II, we describe the existing hand gesture recognition methods and the challenges of those. We clarify the aim of our proposed method and then describe how to design it to recognize the hand gestures with a high level of accuracy in detail in Section III. In Section IV, the validity of our implemented the Korean sign language system is verified by showing experimental results for real data obtained from ten human participants. Finally, the thesis is concluded with future work in Section V.

II. BACKGROUND

2.1 Hand Gesture Recognition

There are some prior research to recognize gestures in relation to Human-machine interaction (HMI) [1, 2, 5]. They focus on how to recognize the given gestures by using sensory data and understand their meanings in order to control the devices by performing certain functions for applications.

To control electronic devices used in a smart home, Costanza *et al.* have developed a hand gesture technique [7]. They not only provide recognition without any user calibration, but also reduce computational complexity. It can only recognize pre-defined gestures using muscle signals. Their method do not consider movements at all. A recognition method is developed to raise the level of human acceptance in AR, it does not also allow the movement of the arms and hands [6].

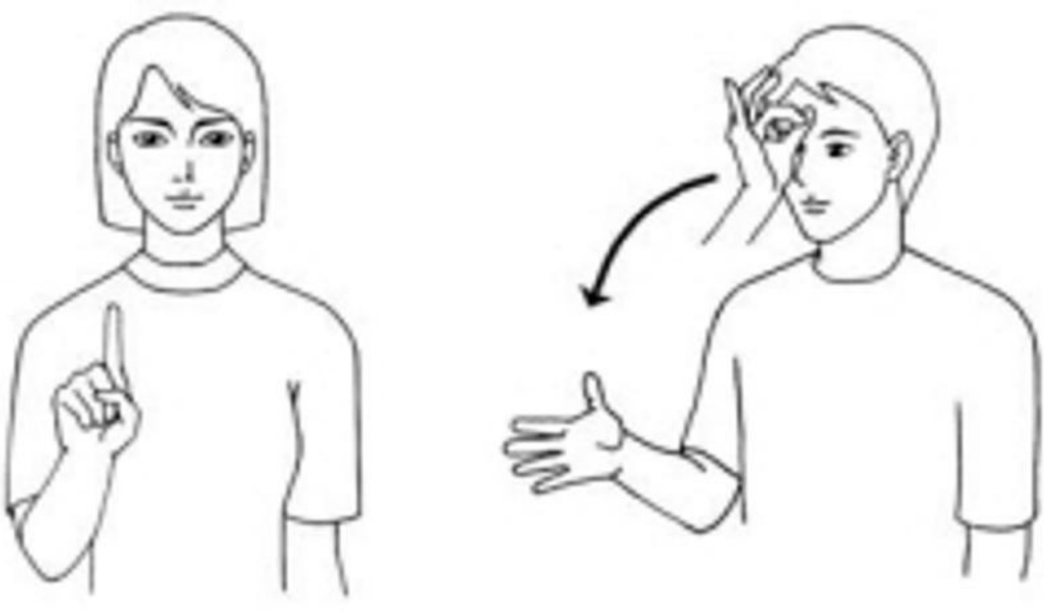
Rahman *et al.* have proposed a method to recognize the sequence of movements [5]. Their approach is limited in providing high mobility because the equipment required for the recognition must be installed in one place. Radkovski and Strizke have proposed a method of hands-free interaction between a person and a device in AR [8]. This requires a lot of sensors for the landmark and provides low portability.

Recently, machine learning algorithms such as support vector machine (SVM) [9], hidden Markov model (HMM) [10], [11], and Artificial neural networks (ANNs) [12] have been studied to recognize gestures. In particular, ANNs, including Convolutional neural networks (CNNs) and Long-short term memory (LSTM), are one of the most popular classifiers. Architecture of CNNs is usually designed to recognize 2D input data, and many researchers use it to classify camera data in order to detect pedestrians, objects, and drones [13, 26, 27].

These techniques mainly create a single model for a specific application by using a large amount of input data. If the designed model based on those is applied to hand gesture recognition, it might have low performance. This is because it does not consider the diversity of human physical characteristics. For example, the signal of EMG depends on physical characteristics and behavior of individuals such as the amount of the user’s muscle and fat, and how the person is moving. To address this problem, when learning the user model, there is an attempt to use both the sensor values and estimated physical characteristics [14]. They try to estimate this user-dependent factors by observing one motion. In this case, although it shows good performance, it fails to identify some signs. The estimation error under the uncertainty and the noise, which might affect recognition performance, are inherent in the method.

2.2 Sensors for Hand Gestures

There are two kinds of hand gestures: static gestures and dynamic gestures [10] as shown in Figure 1 [15].



(a) “One” sign

(b) “Sorry” sign

Figure 1. Examples of static and dynamic gestures

When performing static gestures, users do not move hands and arms and hold a posture of the hand to be used in the communication as show in Figure 1(a). Each static gesture is distinguished only by the shape of the hand representing different meaning.

Since dynamic gestures have a sequence of movements and postures of both the hand and arm, dynamic gesture recognition is the same as sequence recognition. Many studies have used cameras, EMG sensors, globe sensor with IMU, and flex sensors to recognize dynamic hand gestures. The camera sensor is used to extract and recognize the features of arms and hands [16], [17]. The globe sensor has a flex sensor and an IMU sensor. It recognizes the shape of the hand using the flex sensor and recognizes the position of the hand using IMU. Both sensors have a high recognition rate. But it has disadvantages and inconveniences. Camera sensors are difficult to use outdoors because they have to be fixed in one place. In the case of glove sensors, people must wear uncomfortable gloves. In the case of the globe sensor, it is difficult for people to use their hands on other things. Because they have to wear the cumbersome gloves. Wearable devices such as smart watches and armbands are not restricted by location and are not inconvenient to use hands [18], [19], [20].

III. METHODOLOGY

In this section, we introduce the method to recognize hand gestures. In order to create an optimized learning model for many people, we perform a series of operations: (1) feature extraction, (2) pre-processing, (3) generation of architectures for the ANN models, (4) learning and testing the architectures by using cross-validation.

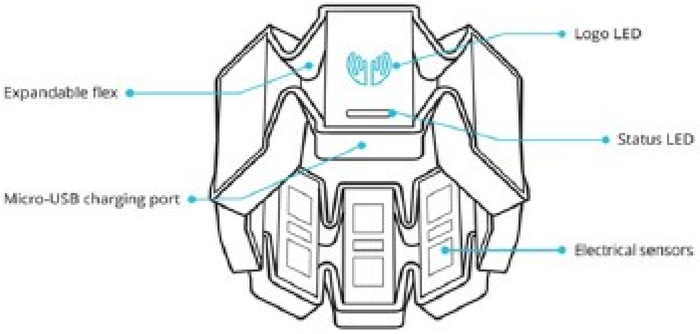
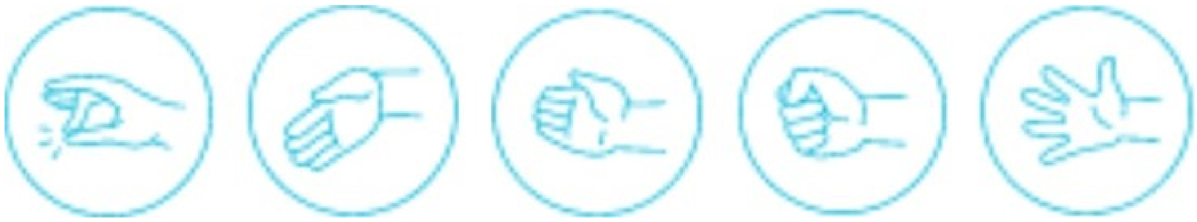


Figure 2. An armband called MYO

3.1 Feature extraction

We use an armband, called MYO, to extract the data as shown in Figure 2. This armband is worn on the forearm and has the function to recognize five static gestures as shown in Figure 3 [21]. This armband's SDK (Software Development Kits) allows us to get the raw data from the sensors. Therefore we can use this data to create our recognition model without using the static gestures provided by the armband.



Double Tap Wave Left Wave Right Make Fist Spread Fingers

Figure 3. Five static gestures recognized by MYO

This arm band contains eight EMG sensors and one IMU sensors. The EMG sensors are non-permeable stainless steel. The IMU includes a three-axis accelerometer and a three-axis gyroscope. This armband transmits data from the sensor via Bluetooth.

It is difficult to distinguish between two gestures in the case of a dynamic gesture with multiple motions. These dynamic gestures are difficult to find starting and ending points because they have different lengths. We define a basic posture to effectively recognize dynamic gestures that have a sequence of motions. This basic posture has no hand and arm movement and is located between the two gestures. Every single gesture consists of three step: basic posture (start), gesture, basic posture (end). The start basic posture and the end basic posture are the same. One gesture recognized by this system must start with the basic posture and return to the basic posture. Figure 4 shows the sequence of postures of a single gesture recognized by the system.

The accelerometer of the IMU recognizes the starting and ending point of the gesture. When the hand or arm moves in the basic posture, it is recognized as the starting point, and when it returns to the basic motion after the gesture, it is recognized as the ending point. To reduce the effect of noise in the accelerometer, we use the *n* average values recently obtained

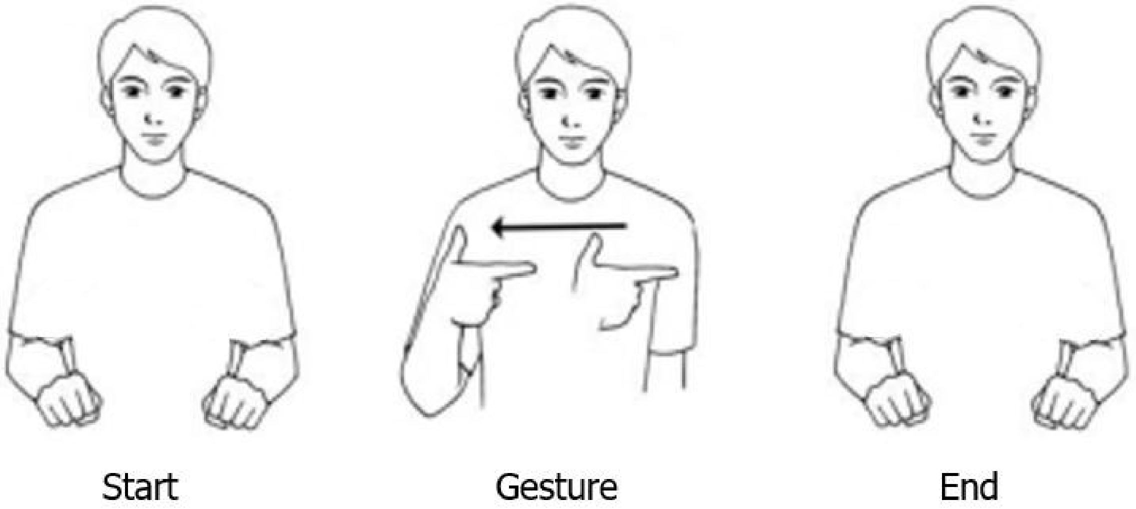


Figure 4. A sequence (one unit) defined for different signs to be distinguished

on each axis. The average value of the accelerometer, $\overline{ACC}_{\psi}(m)$, is shown in Equation (1).

$$\overline{ACC}_{\psi}(m) = \frac{1}{n}\sum_{i=0}^{n-1} V_{\psi}(m-i) \tag{1}$$

In Equation (1), $\psi \in \{x, y, z\}$ means each axis of the accelerometer and m is the average value obtained on the axis ψ, and $V_{\psi}(j)$ is the value obtained on the j-th axis ψ. We set the size of the n to 10, taking into account the effect of gesture length and noise. When the value of $\overline{ACC}_{\psi}(m)$ is greater than the upper bound (a certain positive value) or less than the lower bound (a certain negative value), the system recognizes that the gesture has started. When all values of $\overline{ACC}_{\psi}(m)$ all ψ are kept within between upper bound and lower bound for ten consecutive times, the system recognizes this as an ending point. Because the value of $\overline{ACC}_{\psi}(m)$ obtained in one part of the gesture can be near the boundaries. Figure 5 shows the starting and ending point in $\overline{ACC}_{\psi}(m)$ graph of one gesture.

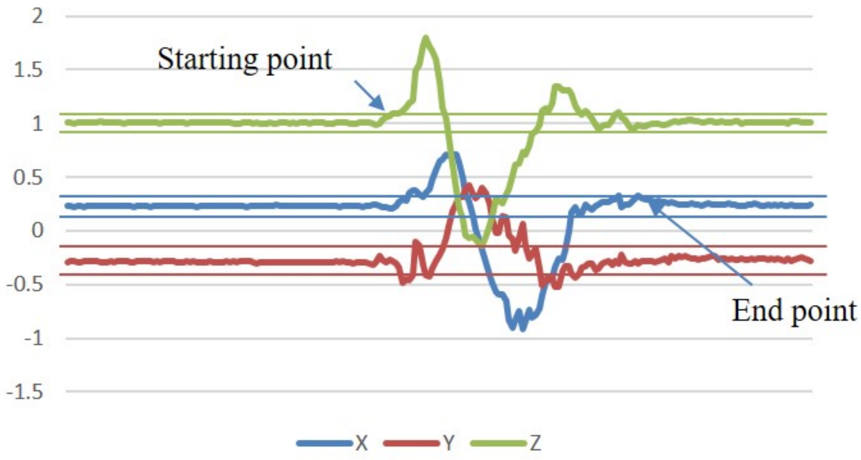


Figure 5. An example of the starting and ending points in the accelerometer’s values

3.2 Preprocessing and Acquisition

The raw data of the EMG sensors included in the MYO arm band have a value of -128 to 128. We converted this value to an absolute value. This is because the value of the EMG signal is the magnitude of the force. This can be expressed as $E_{c,i} = |E_c^i|$, and E_c^i is the i-th data obtained from the channel c ($c \leq 8$) of the EMG sensors.

We have two problems because we used EMG sensor and IMU sensor together. First, the sampling frequencies of the two sensors are different. The EMG sensor of MYO has a sampling frequency of 200 Hz and the IMU sensor has a sampling frequency of 50 Hz. When two different types of sensors are used together, the sensor frequency with a low sampling frequency is selected for sensor synchronization. Therefore, the sampling frequency of the EMG sensor are down-sampled to 50 Hz. Secondly, the range of raw data from two sensors is different. We try two method of min-max and z-score normalization to find a normalization method that can improve recognition performance. The raw data is converted to a z-score and is represented by Equation (2).

$$\begin{gathered} z_{c,i} = \frac{E_{c,i} - \mu}{\sigma}, and \\ z'_{\psi,i} = \frac{M_{\psi,i} - \mu'}{\sigma'} \end{gathered} \tag{2}$$

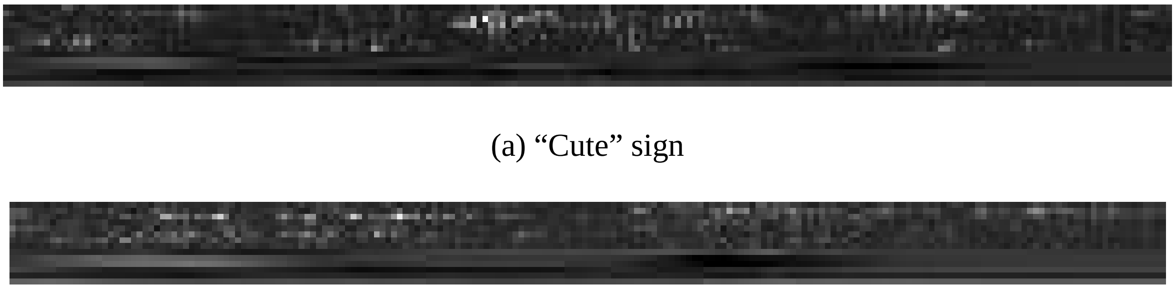
where $E_{c,i}$ and $M_{\psi,i}$ are the i-th values of each channel ($c \leq 8$ and $\psi \leq 6$) of the EMG and IMU sensors respectively. In the case of the EMG sensor, μ and σ are the mean and standard deviation of the values obtained from all channels of the EMG sensors. In the case of the IMU sensor, μ' and σ' are the mean values and standard deviations obtained from accelerometer, gyroscope of IMU, respectively.

The min-max normalization is a method using a maximum value and a minimum value different from the z-score. The raw data of the EMG and the accelerometer, gyroscope of the IMU are transformed into a common range by using Equation (3), respectively.

$$\begin{gathered}\tilde{d}_{c,i} = \frac{E_{c,i} - S_{min}}{S_{max} - S_{min}}, and \\ \tilde{d}'_{\psi,i} = \frac{M_{\psi,i} - S_{min}}{S_{max} - S_{min}}\end{gathered} \tag{3}$$

where S_{min} is 0 and S_{max} is 255.

Many studies use zero padding to make the input data the same size. These studies are image processing, signal processing, and research using a neural network. Zero padding is a method of filling the necessary parts with zeros [22, 23, 24]. In our work, interpolation filter is used to equalize the length of gestures because the lengths of the gestures are different and the lengths of each sample of one gesture are different. As people become accustomed to gestures, the speed of gestures is faster, so the length of the sample in the same gesture can be shorter than before. Interpolation is the estimation of a value located between fixed values. Interpolation filters are used in up-sampling, linear and spline, and we use the spline method [28]. The longest sample of all gestures is 192. We set the size of the training data size for learning to 200. Because it is longer than the longest length of the sample and to perform max-pooling. In



(a) “Cute” sign

(b) “Wait” sign

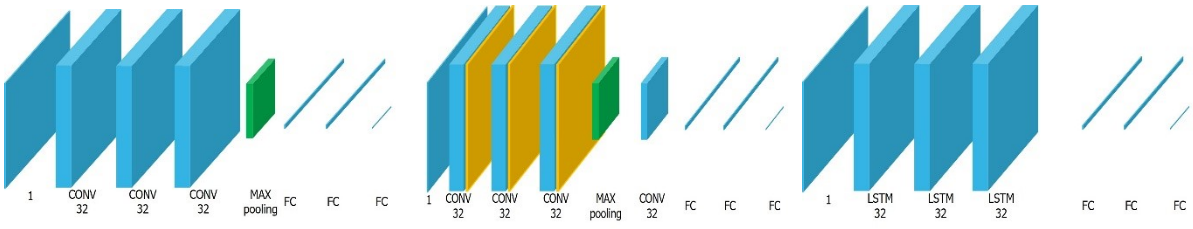
Figure 6. Preprocessed input data of gestures as an image

Figure 6, the preprocessed data of two types of gestures (word “Cute” and “Wait”) are represented by two-dimension image. One of the gesture data consists of 8 channels of EMG sensor and 6 channels of IMU sensor (three channels of accelerometer, three channels of gyroscope). Therefore, the sample of gesture is transformed into a fixed size (the x-axis is 200 and the y-axis 14).

3.3 Creation of Architectures Using Neural Networks

We designed the NN architecture for hand gesture recognition systems. CNN is a specialized network for image classification. To take advantage of this, we transformed the sensor data as an image and used it as learning data. LSTM is often used for data classification such as speech recognition with variable length. We create CNN and LSTM architectures and compare their performance.

Figure 7 shows our three architectures: (a) CNN architecture without dropout, (b) CNN architecture with dropout and (c) LSTM architecture. Dropout is a way of changing the value of one neuron to the next layer to zero at a certain probability. This has the same effect as operating except for that neuron [25]. Dropout is usually used to avoid overfitting. The CNN architecture without dropout consists of three convolution layers, one max-pooling, and three full connected layers. The CNN architecture with dropout has the same structure as (a) architecture and has dropout applied. The hyper-parameters used in the three architectures are: the learning rate is 0.001, the batch size is 10, the activation function is Relu, the optimizer is SGD and the epoch is 50. We used Tensorflow and Keras together. Tensorflow is an open source library for machine learning created by Google. Keras is a neural network API written in Python.

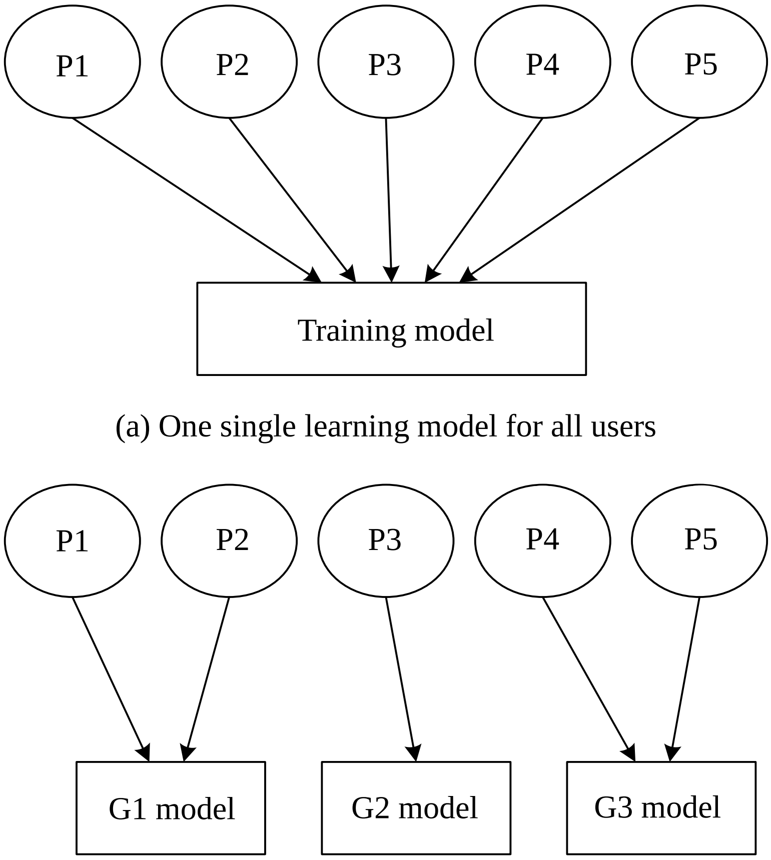


(a) (b) (c)

Figure 7. Designed NN architectures (a) CNN architecture without dropout (b) CNN architecture with dropout (c) LSTM architecture

3.4 Group-Dependent NN Models

We first create models in which all the people's data are learned. These models are made up of a combination of two normalizations (z-score and min-max) and three learning



(a) One single learning model for all users

(b) Group-dependent models

Figure 8. Comparison between a traditional approach and our novel approach

architectures as shown in Figure 7. A single model l that has been learned from the data of all people has a low recognition rate because people’s physical characteristics are different. We create a learned model with only one person’s data and test it with data from another person and found a similar type of person. We create a group of similar types of people and a learned model from the data of the people in the group. We call this as GDM and Figure 8 shows this. In Figure 8, the label P is data of the person and the label G is the group.

IV. KSL RECOGNITION SYSTEM

4.1 Data Acquisition

The data for the hand gesture recognition system are extracted from 10 men (5 people are 20s and 5 people are 30s) without muscle disease. They wear a MYO armband on their right arm and perform prescribed gestures as follows: They watch the video of the gestures two or three times, and then perform it without watching the video. They perform 30 classes of KSL each 70 times. These data are preprocessed and normalized in two ways (z-score and min-max).

4.2 Generation of Group-Dependent NN models

We create six models by combining two ways of normalization, each of architecture based on CNN and LSTM, and the dropout solution. After the designed CNN architecture without dropout is learned by using training data which are pre-processed by z-score, a single model for all people is created. We also design CNN architecture with dropout and LSTM architecture. Each of single models with the design architecture is trained with 15,000 training data and is tested with 6,000 test data.

Table I. The recognition results of a single model for all data

Number	Pre-Processing	NN model	Accuracy (%)
1	Z-score	CNN without dropout	94.2
2	Min-max	CNN without dropout	94.6
3	Z-score	CNN with dropout	93.7
4	Min-max	CNN with dropout	91.1
5	Z-score	LSTM	91.1
6	Min-max	LSTM	3.3

As shown in Table I, the single model using the CNN architecture without dropout have the best performance when it is learned with the normalized data by the z-score method. Every model using data normalized by the z-score method achieves the higher recognition rate (the rate = 80.1, 79.5, and 69.5%) than models (the rate = 71.6, 70.6, and 16%) using min-max normalization. Also, all models using CNN architecture have got better accuracy than two single models using LSTM architecture. Note that the results of all models are not high in terms of the recognition accuracy. It turns out that any model cannot respond well to data for all of the involved people with different physical characteristics.

Therefore, we hypothesize that people with similar characteristics would have a good result on the same learning model. First, a model with CNN architecture is learned only by the data of one person is defined as *a user model*. Second, data of the rest except him are tested on *a user model* learned only by one person's data in order to further analyze the performance (accuracy) of individual user models learned from using data of ten subjects which participate in training.

Experimental results for all user models are presented in Table II, Table III, and Table IV, and the result is highlighted in bold if it is greater than median value of accuracy. User models of Table II(a) and (b) are based on CNN without dropout. CNN with dropout is used for user models of Table III(a) and (b). User models of Table IV(a) and (b) have LSTM architecture. Table II(a), Table III(a), and Table IV(a) show the results of user models to which z-score normalization is applied respectively. The results of user models performing min-max normalization are in Table II(b), Table III(b), and Table IV(b), respectively.

From Table II and Table III, we find that the combination of the z-score normalized method and the CNN architecture has many cells with values higher than median value rather than those of other models.

In Table II(a), when the first person's data (Label Person1) are tested on the ninth person's model (Label User Model 9), the result is 89.7%. Also, when the ninth person's data (Label Person 9) are tested on the tenth person's model (Label User Model 10), the result is 95.2%.

Table II. Recognition accuracy (%) of the user model based on CNN without dropout for each person

(a) Z-score

	User Model	Person									
		1	2	3	4	5	6	7	8	9	10
CNN without dropout	1	**99.7**	43.5	51.2	40.8	38.2	48.0	**80.2**	**60.7**	**71.0**	**67.2**
	2	**58.7**	**98.8**	47.7	26.2	43.8	42.8	**56.8**	44.8	**55.7**	**54.5**
	3	**58.5**	40.3	**98.8**	42.5	33.8	**75.2**	**57.5**	**57.0**	**56.2**	38.7
	4	47.5	32.2	49.8	**98.3**	26.5	**66.2**	43.2	37.5	31.8	27.2
	5	**55.2**	42.7	38.8	28.0	**99.7**	34.3	**58.5**	51.0	41.5	46.5
	6	**54.3**	42.0	70.8	**52.7**	30.5	**98.3**	**58.2**	48.3	49.3	39.7
	7	**84.2**	**60.7**	51.5	37.3	**57.0**	48.8	**99.8**	**57.8**	**83.7**	**82.3**
	8	**54.8**	**51.8**	45.8	31.5	43.8	37.5	**53.5**	**99.7**	51.5	**56.8**
	9	**89.7**	**52.0**	49.8	31.8	46.3	42.8	**92.5**	**59.0**	**99.8**	**95.2**
	10	**81.5**	**59.5**	43.5	33.5	50.2	40.0	**94.0**	**56.8**	**97.7**	**99.8**

(b) Min-max

CNN without dropout	1	**99.8**	46.7	43.7	36.3	42.5	39.5	**78.2**	**63.5**	**68.0**	**61.8**
	2	**60.5**	**98.7**	43.7	34.0	49.0	47.3	**74.3**	**55.5**	**51.2**	**56.3**
	3	**56.3**	38.8	**99.3**	34.2	42.2	**72.5**	**56.0**	**54.8**	43.7	36.3
	4	45.0	29.3	**51.3**	**97.8**	33.7	**66.5**	**52.3**	34.2	35.8	25.2
	5	48.3	36.5	33.2	25.5	**99.7**	29.0	**54.3**	46.3	40.5	45.3
	6	**59.3**	46.2	**67.5**	**51.5**	37.0	**98.5**	**63.2**	49.0	**50.3**	46.0
	7	**81.3**	**52.7**	48.7	29.7	**53.7**	43.3	**99.8**	**54.5**	**79.8**	**76.2**
	8	**51.3**	42.7	44.5	26.8	40.2	35.7	**52.3**	**99.0**	47.2	**53.5**
	9	**86.8**	48.3	**52.2**	34.3	47.7	46.8	**92.7**	**60.2**	**99.7**	**88.0**
	10	**74.8**	**58.0**	47.0	30.3	49.5	43.2	**91.8**	**60.5**	**92.0**	**99.7**

In Table III(a), when the first person's data (Label Person1) are tested on the ninth person's model (Label User Model 9), the result is 85.0%. Also, when the ninth person's data (Label Person 9) are tested on the tenth person's model (Label User Model 10), the result is 97.2%. Depending on the person and the model with the high results, we make a group.

Table III. Recognition accuracy (%) of the user model based on CNN with dropout for each person

(a) Z-score

	User Model	Person									
		1	2	3	4	5	6	7	8	9	10
CNN with dropout	1	**99.7**	44.8	45.3	38.3	41.3	42.5	**78.2**	**55.5**	**66.5**	**62.0**
	2	**58.3**	**98.7**	47.8	34.3	42.0	46.3	**59.0**	42.5	**52.3**	**51.7**
	3	**56.3**	38.7	**99.2**	39.2	33.3	**74.0**	**56.5**	**54.7**	48.8	40.3
	4	48.8	40.2	**54.2**	**98.3**	37.3	**71.5**	48.3	43.3	38.7	30.0
	5	**52.7**	40.3	34.7	27.3	**99.8**	30.0	**57.8**	45.3	40.7	43.8
	6	**52.8**	39.3	**69.3**	**54.8**	35.7	**98.5**	**57.0**	48.8	47.7	38.0
	7	**83.5**	**58.7**	47.7	42.7	**52.7**	**49.5**	**99.8**	**53.5**	**80.0**	**80.2**
	8	**54.0**	42.2	36.5	23.0	39.7	28.5	48.7	**99.5**	44.3	**52.8**
	9	**85.0**	**50.3**	**50.3**	34.2	44.5	45.2	**90.5**	**52.7**	**99.8**	**94.5**
	10	**81.7**	**57.3**	41.8	34.3	46.7	40.3	**92.2**	**56.2**	**97.2**	**99.8**

(b) Min-max

CNN with dropout	1	**99.8**	37.2	46.3	35.7	35.5	41.8	**75.8**	**60.3**	**67.5**	**61.7**
	2	**60.7**	**98.7**	48.2	32.2	48.5	46.7	**66.0**	**51.7**	**51.0**	**54.2**
	3	**59.5**	39.7	**99.5**	41.5	42.7	**76.0**	**51.7**	**54.7**	46.0	33.7
	4	**53.2**	34.0	45.0	**98.0**	27.0	**64.8**	**54.3**	37.3	38.7	29.0
	5	47.2	38.3	31.7	27.8	**99.5**	29.0	**53.3**	47.0	39.0	39.0
	6	**57.0**	38.0	**65.5**	**52.7**	27.8	**98.7**	**61.8**	**49.7**	47.7	36.8
	7	**76.3**	**53.2**	**53.5**	31.2	**54.0**	46.8	**99.8**	**54.7**	**78.7**	**78.5**
	8	**55.0**	44.0	38.3	29.7	40.3	35.5	**50.0**	**99.2**	44.3	50.0
	9	**85.5**	45.3	**48.8**	29.8	43.0	42.0	**95.8**	**50.3**	**99.8**	**92.5**
	10	**75.2**	**55.5**	42.3	27.7	48.2	34.7	**88.7**	**55.3**	**94.7**	**99.8**

Table IV. Recognition accuracy (%) of the user model based on LSTM for each person

(a) Z-score

	User Model	Person									
		1	2	3	4	5	6	7	8	9	10
LSTM	1	**99.7**	39.3	38.7	39.3	31.3	43.0	**73.2**	**44.2**	**67.8**	**61.5**
	2	**47.3**	**98.7**	41.7	25.5	31.5	37.0	**48.2**	**46.8**	43.0	42.3
	3	**44.5**	34.2	**99.3**	29.0	23.3	**65.5**	37.5	**43.8**	35.3	35.3
	4	**45.5**	26.3	**44.2**	**98.0**	20.8	**63.0**	**45.8**	32.8	30.7	26.0
	5	35.3	39.0	29.8	25.3	**99.7**	24.5	**52.3**	33.8	32.3	30.8
	6	**45.7**	33.3	**68.3**	**50.8**	16.5	**98.5**	**44.2**	**46.3**	39.2	32.8
	7	**72.7**	41.8	**44.8**	29.2	41.7	**44.3**	**99.8**	**44.8**	**72.2**	**71.5**
	8	**48.7**	34.5	33.7	20.5	20.7	31.8	**50.8**	**99.3**	35.8	41.2
	9	**72.5**	41.8	**45.8**	35.5	30.7	**44.3**	**81.5**	**45.0**	**99.8**	**88.5**
	10	**68.3**	**55.7**	40.2	32.2	39.2	36.3	**79.0**	**49.2**	**94.7**	**99.5**

(b) Min-max

LSTM	1	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	2	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	3	3.5	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	4	2.5	5.5	2.3	1.7	4.7	1.7	3.8	4.0	3.8	2.2
	5	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	6	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	7	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	8	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	9	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3	3.3
	10	3.5	3.3	3.3	3.3	3.2	3.3	3.3	3.2	3.3	3.3

Consequently, according to the results in Table II(a) and Table III(a), people are divided into four groups (Group A: Label Person1, 6, 7, 9, and 10, Group B: Label Person 2, 4 and 8, Group C: Label Person 3, Group D: Label Person 5). Each person was divided into groups according to the K-means algorithm, and the results of the table were used as the input data. The K-means algorithm is an algorithm for grouping given data into k clusters. Based on

Table V. Recognition results for GDMs

(a) CNN without dropout

Group	A	B	C	D
Accuracy (%)	96.3	94.7	98.8	99.7

(b) CNN with dropout

Group	A	B	D	D
Accuracy (%)	96.1	94.0	99.2	99.8

data of individual groups, we create a group-dependent model and perform the test again.

Table V(a) is the result of the CNN model without dropout and (b) is the result of the CNN model with dropout. The model without dropout and the model with dropout have average 97.4% and 97.2% recognition results, respectively. In addition, we test learning models with data from two experimental participants. The data of these experimental participants are not used in the learning data. The CNN without dropout model (number 1 of Table II) and the CNN with dropout model (number 3 of Table II) have 81.0% and 79.3%, respectively.

Table VI shows the results of testing the two data in a GDMs. In Table VI (a) The first person (Label Person 10)'s data has the highest accuracy (85.1%) when tested on the model of group B. The second person (Label Person 11)'s data has the highest accuracy (80.5%) when tested on the model of group A. Label person 10 and 11 set group B and A as their own learning models, respectively. When a new person not included in the learning data uses GDM, the model of the group with the highest accuracy is set as the learning model.

Note that the recognition result of the model learned by the data of the group is higher than a single model learned by all the data as shown in Table I. It turns out that the physical characteristics of the same group's people are similar, comparing with the real data of people as shown in Table VII. It shows the body information of the people in the group.

Table VI. Recognition results (%) for group-dependent models

with data of not included in learning data

(a) CNN without dropout

Person	A	B	C	D
10	78.3	85.1	49.6	42.6
11	80.5	70.5	37.3	38.8

(b) CNN with dropout

Person	A	B	D	D
10	74.3	84.0	50.1	35.1
11	81.5	73.3	35.5	38.5

Table VII. Body information of people in each group

Group	Person	Height (cm)	Weight (kg)	Fore-arm (cm)	Upper-arm (cm)	Arm circumference (cm)
A	1	172	65	25	30	26
	6	178	72	27	34	27
	7	173	77	28	27	27
	9	172	68	26	30	27
	10	172	69	27	30	27
B	2	177	75	29	33	27
	4	178	69	28	30	26
	8	178	75	29	33	27
C	3	181	85	30	33	29
D	5	174	68	28	28	27
	10	176	71	27	30	26
	11	171	67	27	30	26

4.3 System Implementation

The group-dependent learning models, which show a high recognition performance (more than 99%) as mentioned above, are used to implement the KSL recognition system. The overall system design is illustrated in Figure 9. After the user wears the armband and performs the KSL, the data is transmitted to the mobile PC via Bluetooth. The mobile PC sends this data to the classifier server with the learning model. The server pc pre-processes the data and applies it to the learning model. The server sends the text result on the mobile PC. The mobile PC sends the text results to the Google TTS server and receives the mp4 result. The mobile PC outputs the result on the speaker.

The implemented system's mobile pc is a laptop with windows 7. This PC communicates with the arm band via Bluetooth. The PC also communicates with a classifier server and a google TTS server using a wireless network such as WiFi. The pc will be replaced by a mobile device.

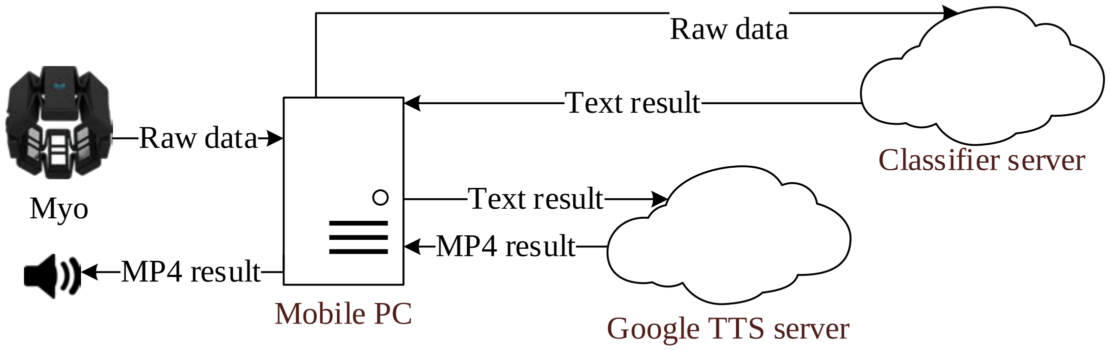


Figure 9. A diagram of our KSL recognition system

V. CONCULSION

In this thesis, we propose the effective method to Korean sign language recognition using EMG and IMU sensors. This method focus on developing the group-dependent models based on deep learning. This specified model comes from the fact that the physical characteristics of human being are absolutely different but similarity among them may exist. The deep learning approaches such as CNN and LSTM are exploited to create the group-dependent models. To achieve more high performance, the raw data in the numeric range is transformed to the 2D image as an input of group-dependent models. In order to recognize sign language accurately and efficiently, we investigate several problems to be addressed. First, it is difficult to distinguish between given gestures when we communicate with sign language using dynamic hand gestures. The second problem is that the length of the dynamic hand gestures is different. To address these problems, we define the basic posture and apply interpolation filter to initial data of the gestures in order to fix the length. Although using both the EMG sensor and the IMU sensor is helpful for improving the reliability and accuracy for recognition, unpaired data from the differences of sampling frequencies of sensors disturbs the accurate fusion due to multi-view data. We select under-sampling being not complex process within the range with no distortion of sensory data, considering the real-time processing. In addition, since the range of variables obtained from the two sensors is quite different, two normalization methods such as z-sore and min-max are applied to process the raw sensor variables and we evaluate their performance.

When the user models based on the same architecture are used, our results indicate that the performance (in terms of the accuracy) of them with z-score is higher than min-max normalization of them. All user models based on the CNN architecture show better performance than those based on the LSTM architecture for Korean sign language recognition. We think that the preprocessing method contributes to CNN much rather than LSTM. In addition, we see that the user model corresponding to each group are more accurate than one single model for all people data. In this regard, we suggest that to create the individual user model for each of groups with similarity is more effective than that of one single model learnt from data of all available people.

We create a KSL recognition system. This system helps the hearing-impaired and the language impaired to communicate with the public. This system can be used in any place. The system recognizes the KSL that the person performs and outputs it to the speaker.

We plan to create a GDM with better performance by collecting more data since the number of participants in the current work is too small and to compare it with other experimental results. It is important to determine the number of groups if many people will use the group-dependent model. The elbow method [29] is a method of determining the number of clusters, which is a method of increasing the number of clusters. This method is a method of setting the number of clusters to n if the performance of n+1 cluster models is not better than that of n cluster models.

We will implement and optimize it to be used outdoors in real time so that it can work on smart devices with limited processing capacities, such as smart phones. It is worthwhile to provide our KSL recognition system for the people who want to communicate with or help the speech-impaired people but they have never learned how to speak sign language.

References

[1] R. B. Reilly and T. C. Lee. “Electrograms (ecg, eeg, emg, eog),” *Technology and Health Care,* pp. 443-458, 2010.

[2] S. Wan and E. Foxlin. "Improved pedestrian navigation based on drift-reduced MEMS IMU chip," in *Proceedings of the 2010 International Technical Meeting of the The Institute of Navigation, San Diego, CA, USA*, 2010, pp. 220-229.

[3] J. Wu, L. Sun and R. Jafari. “A Wearable System for Recognizing American Sign Language in Real-Time Using IMU and Surface EMG Sensors,” *IEEE Journal of Biomedical and Health Informatics*, pp. 1281-1290, 2016.

[4] "2016 Disablility Statistics." Korea Employment Agency for the Disabled, [Online]. Available: https://www.kead.or.kr/webzine/ibook/ttsbook/WEB/KEAD.html. [Accessed sept 2017].

[5] A. M. Rahman, M. A. Hossain and J. Parra. “Motion-path based gesture interaction with smart home services,” *ACM international conference on Multimedia,* pp. 761-764, 2009.

[6] S. Reifinger, F. Wallhoff, M. Ablassmeier, T. Poitschke and G. Rigoll. “Static and Dynamic Hand-Gesture Recognition for Augmented Reality Applications,” *International Conference on Human-Computer Interaction*, pp. 728-737, 2007.

[7] E. Costanza, S. A. Inverso and R. Allen. “Toward subtle intimate interfaces for mobile devices using an EMG controller,” *SIGCHI conference on Human factors in computing systems,* pp. 481-489, 2005.

[8] R. Radkowski and C. Stritzke. “Interactive Hand Gesture-based Assembly for Augmented Reality Applications,” 2012.

[9] M. Yoshikawa, M. Mikawa and K. Tanaka. "Real-time hand motion estimation using EMG signals with support vector machines," in *SICE-ICASE, 2006. International Joint Conference*, IEEE, 2006, pp. 593-598.

[10] X. Zhang, X. Chen, Y. Li, V. Lantz, K. Wang and J. Yang. "A framework for hand gesture recognition based on accelerometer and EMG sensors," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* vol. 41, no. 6, pp. 1064-1076, 2011.

[11] Z. Yang, Y. a. C. W. Li and Y. Zheng. "Dynamic hand gesture recognition using hidden Markov models," in *Computer Science & Education (ICCSE), 2012 7th International Conference on,* IEEE, 2012, pp. 360-365.

[12] M. R. Ahsan, M. I. Ibrahimy and O. O. Khalifa. "Electromygraphy (EMG) signal based hand gesture recognition using artificial neural network (ANN)," in *Mechatronics (ICOM), 2011 4th International Conference On,* IEEE, 2011, pp. 1-6.

[13] Y. Tian, P. Luo, X. Wang and X. Tang. "Deep learning strong parts for pedestrian detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1904-1912.

[14] T. Matsubara and J. Morimoto. "Bilinear modeling of EMG signals to extract user-independent features for multiuser myoelectric interface," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2205-2213, 2013.

[15] "Korean numeral words dictionary," National Institue of Korean Language, [Online]. Available: http://sldict.korean.go.kr/signhand/hand/main.do. [Accessed sept 2017].

[16] R. Hartanto, A. Susanto and P. I. Santosa. "Real time hand gesture movements tracking and recognizing system," in *Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS), 2014*, IEEE, 2014, pp. 137-141.

[17] Z. Ren, J. Yuan and Z. Zhang. “Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera,” *19th ACM international conference on Multimedia*, 2011.

[18] C. Xu, P. H. Pathak and P. Mohapatra. "Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch," in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, ACM, 2015, pp. 9-14.

[19] Z. Lu, X. Chen, Q. Li, X. Zhang and P. Zhou. "A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices," *IEEE transactions on human-machine systems*, vol. 44, no. 2, pp. 293-299, 2014.

[20] M. Sathiyanarayanan and S. Rajan. "MYO Armband for physiotherapy healthcare: A case study using gesture recognition application," in *Communication Systems and Networks (COMSNETS), 2016 8th International Conference on*, IEEE, 2016, pp. 1-6.

[21] “Myo,” [Online]. Available: https://www.myo.com/.

[22] D. Wang, N. Canagarajah, D. Redmill and D. Bull. “Multiple description video coding based on zero padding. In Circuits and Systems,” *IEEE International Symposium on Circuits and Systems*, 2004.

[23] J. Borkowski and J. Mroczka. “LIDFT method with classic data windows and zero padding in multifrequency signal analysis,” *Measurement*, pp. 1595-1602, 2010.

[24] B. Hu, Z. Lu, H. Li and Q. Chen. “Convolutional neural network architectures for matching natural language sentences,” *Advances in neural information processing systems*, pp. 2042-2050, 2014.

[25] H. Wu and X. Gu. "Towards dropout training for convolutional neural networks," *Neural Networks*, vol. 71, pp. 1-10, 2015.

[25] Dubey, Rashmi, et al. "Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study." NeuroImage 87 (2014): 220-241.

[26] Ren, Shaoqing, et al. "Faster R-CNN: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[27] Aker, Cemal, and Sinan Kalkan. "Using deep networks for drone detection." Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on. IEEE, 2017.

[28] Dyer, Stephen A, and Justin S. Dyer. "Cubic-spline interpolation. 1." IEEE Instrumentation &

Measurement Magazine 4.1 (2001): 44-46.

[29] Bholowalia, Purnima, and Arvind Kumar. "EBK-means: A clustering technique based on elbow method and k-means in WSN." International Journal of Computer Applications 105.9 (2014).

요 약 문

센서 퓨전 기반의 핸드 제스처 인식을 위한 그룹 기반의 개인 맞춤 접근

입력 장치 발전으로 사람들의 의사를 좀더 표현 할 수 있는 장치들이 개발 되고 있다. 이에 따라 증강 현실 또는 실생활에서 사용 할 수 있는 입력 장치가 필요하다. 핸드 제스처는 사람의 의사를 가장 직관적으로 표현할 수 있는 방법이며 이 것을 기계가 인식한다면 컴퓨터와 사람간의 소통이 더욱 좋아질 것이다. 핸드 제스처는 움직임이 있는 다이나믹 제스처와 움직임이 없는 스태틱 제스처로 나뉜다. 스태틱 제스처의 경우 손의 모양에 따라 사용되는 근육이 다르기 때문에 측정 할 수 있는 근전도 센서를 이용하여 손의 모양을 추측 할 수 있다. 다이나믹 제스처의 경우 IMU를 통하여 측정 할 수 있다. 이 두 센서를 퓨전 하여 다이나믹 핸드 제스처를 인식하고자 한다. 다이나믹 제스처는 움직임의 길이가 다르기 때문에 시작 지점과 끝 지점을 찾기 어렵다. 따라서 기본 동작을 정의함으로써 시작 지점과 끝 지점을 구분 할 수 있도록 한다. 기본 동작에서 시작 지점과 끝 지점을 인식하는 방법은 IMU의 가속도계를 사용하며, 노이즈의 영향을 줄이기 위해 최근 10개의 데이터의 평균값을 사용한다. 이 평균값이 일정 범위 이상 벗어나면 시작 지점이라 하며, 일정 범위 안에 10번 연속으로 들어오면 끝 지점이라 한다. 제스처의 한 동작이 범위 근처에서 동작 할 수 있기 때문이다. 다이나믹 제스처의 경우, 한 제스처의 여러 샘플의 길이가 다르기 때문에 zero padding을 적용하여 샘플들의 길이를 동일하게 한다. EMG 센서와 IMU 센서를 같이 사용 했기 때문에, 샘플링 주기가 높은 EMG 센서를 다운 샘플링 했다. 또한 센서들의 출력되는 값의 범위가 다르기 때문에 정규화를 수행하며 우리는 2가지 정규화 (z-score, min-max)를 수행하여 성능을 비교한다. 우리는 CNN과 LSTM을 구조로 가지는 모델들을 만들었으며, CNN의 경우 dropout을 적용한 모델과 적용하지 않은 모델을 만들었다. CNN모델을 사용하기 위해 전처리 된 센서 값을 이미지처럼 표현했다. 사람들은 각자 다른 신체적 특성을 가지고 있기 때문에 센서로부터 측정된 값이 다를 수 있다. 이는 분류 모델의 성능 저하를 일으킬 수 있는 원인이 된다. 따라서 비슷한 사람마다 그룹을 지어 각각 다른 모델을 사용하는 방법으로 인식률을 높이고자 한다. 모든 데이터로 학습된 하나의 학습모델보다 비슷한 특성을 가지는 그룹의 데이터로 학습된 모델들 중 자신에게 맞는 모델을 사용했을 때 성능이 크게 향상 됨을 확인했다. 한국에 있는 수화를 사용하고 있는 장애인들은 수화를 모르는 일반인들과의 의사소통에 어려움을 겪는다. 따라서 수화를 인식해서 실시간으로 번역할 수 있는 시스템을 만든다. 이 시스템은 수화를 인식하고 뜻을 스피커를 이용해 알려줌으로써 일반인이 수화를 모르더라도 그 뜻을 알 수 있게 한다.

핵심어: 한국 수화, 근전도, 핸드 제스처, 센서 퓨전, 인공 신경망