Regular paper

Journal of the KIECS. pp. 291-298, vol. 15, no. 2, Apr. 30. 2020, t. 100, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2020.15.2.291

딥러닝 방식의 웨어러블 센서를 사용한 미국식 수화 인식 시스템

정택위[*] · 김범준[**]

American Sign Language Recognition System Using Wearable Sensors with Deep Learning Approach

Teak-Wei Chong[*] · Beom-Joon Kim[**]

요 약

수화는 청각 장애인이 다른 사람들과 의사소통할 수 있도록 설계된 것이다. 그러나 수화는 충분히 대중화되어 있지 않기 때문에 청각 장애인이 수화를 통해서 일반 사람들과 원활하게 의사소통하는 것은 쉽지 않은 문제이다. 이러한 문제점에 착안하여 본 논문에서는 웨어러블 컴퓨팅 및 딥러닝 기반 미국식 수화인식 시스템을 설계하고 구현하였다. 이를 위해서 본 연구에서는 손등과 손가락에 장착되는 총 6개의 IMUs(Inertial Measurement Unit) 센서로 구성된 시스템을 구현하고 이를 이용한 실험을 수행하여 156개 특징이 수집된 데이터 추출을 통해서 총 28개 단어에 대한 미국식 수화 인식 방법을 제안하였다. 특히 LSTM (Long Short-Term Memory) 알고리즘을 사용하여 최대 99.89%의 정확도를 달성할 수 있었고 향후 청각 장애인들의 의사소통에 큰 도움이 될 것으로 예상된다.

ABSTRACT

Sign language was designed for the deaf and dumb people to allow them to communicate with others and connect to the society. However, sign language is uncommon to the rest of the society. The unresolved communication barrier had eventually isolated deaf and dumb people from the society. Hence, this study focused on design and implementation of a wearable sign language interpreter. 6 inertial measurement unit (IMU) were placed on back of hand palm and each fingertips to capture hand and finger movements and orientations. Total of 28 proposed word-based American Sign Language were collected during the experiment, while 156 features were extracted from the collected data for classification. With the used of the long short-term memory (LSTM) algorithm, this system achieved up to 99.89% of accuracy. The high accuracy system performance indicated that this proposed system has a great potential to serve the deaf and dumb communities and resolve the communication gap.

키워드

American Sign Language, Deep Learning, Human-Computer Interaction, Sign Language, Wearable Sensors
미국식 수화, 딥러닝, 인간-컴퓨터 상호 작용, 수화, 웨어러블 센서

[*] 계명대학교 전자전기공학과 (chongteakwei@gmail.com)
[**] 교신저자 : 계명대학교 전자전기공학과

・접 수 일 : 2020. 02. 12
・수정완료일 : 2020. 03. 14
・게재확정일 : 2020. 04. 15

・Received : Feb. 12, 2020, Revised : Mar. 14, 2020, Accepted : Apr. 15, 2020

・Corresponding Author : Beom-Joon Kim
Dept. of Electronic Engineering, Keimyung University,
Email : bkim@kmu.ac.kr

Ⅰ. Introduction

Communication is a tool to allow people to convey message or share their self-inner feeling with each other. In the society, people communicate with each other verbally and non-verbally. However, deaf and dumb communities are incapable of communicating verbally. Thus, sign language was designed to serve deaf and dumb communities. Sign language is formed by a series of hand and body gestures and facial express [1]. Similar to spoken language, sign language is not standardized globally. It is classified by many different dialects regionally, for instance, Chinese Sign Language (CSL), Arabic Sign Language (ArSL), French Sign Language (LSF), and etc. [2]. Each types of sign language has their own unique expressions, some required both hands to perform but some are only one-handed.

In fact, this study focused on American Sign Language (ASL). ASL is one of the most common sign language worldwide and it was reported as the most targeted sign language among all published papers during the last decade [3]. In general, ASL was classified into two, namely fingerspelling and word-based ASL. Fingerspelling consists of 36 signs with designated handshapes that represent 26 letters (A-Z) and 10 digits (0-9) respectively. Most of the fingerspelling signs have no movement involved, except for the letter "J" and letter "Z". Conversely, word-based ASL consists hand and fingers movement to represent English words and phrases. Even so, ASL is not coded English in signs representation [4]. ASL has different grammars and sentence structure with English language [5]. According to William Stokoe's terminology, word-based ASL were formed by five components, i.e. (1) handshape, (2) movement, (3) orientation, (4) location of articulation, and (5) facial and body expression [6].

Unfortunately, most of people in the society have no knowledge and experience on sign language [3]. Therefore, it has become a communication barrier in our current society. Deaf and dumb communities were often found isolated from the society. As such, sign language recognition systems were proposed to aid deaf and dumb communities and vanish the unresolved society gap.

The objective of this study targeted to design and implement a system with smart wearable sensors that able to recognized dynamic word-based ASL signs which involved movements. The system aimed to serve the deaf and dumb communities as a sign language interpreter that able decode sign language into text or audio to mass public.

Ⅱ. Related Works

In the field study of sign language recognition, two approaches was often found for hand sign and gesture recognition, namely vision-based and sensor-based approach. For vision-based approach, Haria et al. [7] proposed a low cost marker-less hand gesture recognition system using webcam. This study utilized contours, convexity defects and Haar cascade to performance hand detection, and reported that complex background was an impactful factor to cause low detection accuracy. Then, Elmezain et al. [8] adopted a stereo camera for hand gesture recognition. This study achieved 98.6% of accuracy for 10 Arabic numbers recognition by skin segmented technique using Gaussian Mixture Model (GMM). The technique successfully overcame the difficulties of handling occlusion and overlapping regions between hands and faces. Moreover, Molchanov et al. [9] integrated color camera, depth camera, and short range radar for 10 dynamic gesture recognition using 3D-based convolutional neural network (CNN). The study revealed the highest accuracy rate of 94.1% by including all features extracted from three cameras, while the features that extracted from color camera delivered the lowest accuracy of 60.1% only.

For sensor-based approach, several sensors were widely adopted by researchers. Preetham et al. [10] utilized 10 flex sensors to develop a single-handed data glove for hand gesture recognition. These flex sensors were placed on each of two joints over every fingers. on top of that, Patil et al. [11] proposed another data glove that only utilizing 5 flex sensors to put on each fingers to reduce bulkiness of the wearable glove. However, flex sensors only measured fingers flexion but not capable to obtain finger and hand movement and orientation. Thus, Wang et al. [12] proposed a gesture recognition system using data glove that utilized 5 flex sensors and one 3-axis accelerometer. The accelerometer was placed at the center back of hand palm. The system recognition 50 CSL gestures in real-time with accuracy rate over 91%. For further improvement, Lee et al. [13] added two pressure sensors which were mounted on the top and the left of middle fingertip. The proposed study recognized 26 ASL fingerspelling letters using support vector machine (SVM) classifier had achieved the highest accuracy of 98.2%. Other than flex sensors, Mummadi et al. [14] developed a data glove that consisted of 5 inertial measurement unit (IMU) sensors that placed on each of fingertips to recognize 24 static ASL fingerspelling letters. The system achieved 92.95% of accuracy by random forest classifier (RF). Unfortunately, this proposed system incapable to recognize dynamic ASL signs with movement.

In short, vision-based has a limited range of detection and suffering from the background noise, while sensors-based approach allow user to perform the sign in most kind of circumstances but the bulkiness of the wearable glove must put into consideration.

Ⅲ. System Design

3.1 Sensing Module

Figure 1 depicted the proposed custom-made smart wearable sensors. This wearable sensors consisting 6 BNO055 IMU sensors, Bluetooth module HC-06, 3.7V Lithium-Ion battery, Teensy 3.2 microcontroller (MCU), and I2C multiplexer TCA9548A. Those IMUs were placed on each of fingertips and one on back of hand palm to capture fingers and hand movement data. All the sensors readings were initially received and process by Teensy 3.2 MCU. However, Teensy 3.2 MCU had not enough I2C to support 6 IMUs concurrently. To solve the problem, an I2C multiplexer was included as a medium between MCU and sensors. Besides, a bluetooth module HC-06 was included to transmit collected data to computer for further data analysis and processing. All of these components were powered by a 3.7V Lithium-Ion battery.

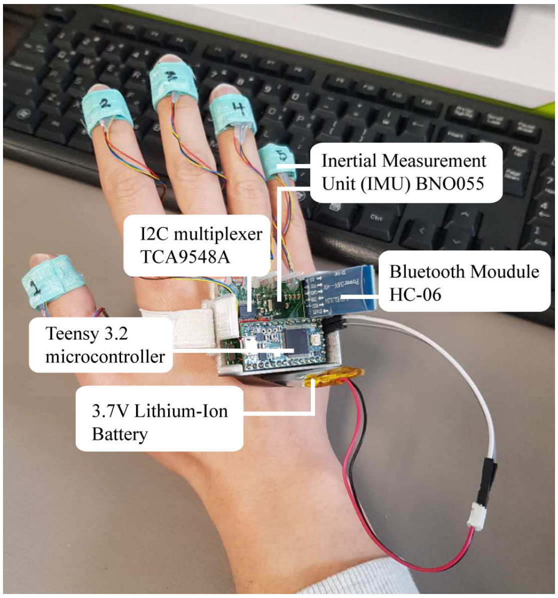

Fig. 1 Proposed smart wearable sensors

3.2 Processing Module

The Teensy 3.2 MCU was programmed by Arduino programming languages to acquire data from sensors. Due to Teensy 3.2 MCU is not native Arduino components [15], Teensyduino was utilized to upload Arduino sketches to the MCU. It was a third-party software add-on for Arduino that convert and compile Arduino code to Teensy MCU boards instantly, no extra coding task was needed.

However, the main data analysis and classification tasks were performed and programmed in Python language [16, 17]. It was chosen due it was highly compatible with most of the deep learning frameworks and libraries such as. Scikit-learn and Keras library.

Ⅳ. Experiment and Results

4.1 Data Collection

There were 12 subjects performed data collection in this experiment. The experiments were under supervision of researcher(s) in order to make the data collection was obtained correctly and safely. 12 data collection sections were conducted seperately. Each of them were requested to performed the given ASL gestures that showed on monitor which placed in front of the participants. There were 28 gestures proposed in this study. 27 word-based ASL signs and one non-ASL gesture. The 27 word-based ASL gestures were showed as Table 1, all of these gestures were selected based on the ASL components by William Stokoe's terminology. For instance, gesture “good” and gesture “happy” had same handshape but different in gesture movement, hand orientation and location of articulation. Exceptionally, gesture “Ok” and “rice” were two gesture that formed by a series of fingerspelling letters. “o”-“k” and “r”-“i”-“c”-“e” respectively. While the non-ASL is a relax gesture that had all fingers fully opened that allow participants have some rest during the experiment.

In this study, 4 types of raw data were collected from the sensors during the experiments, namely accelerometer (A), gyroscope (G), orientation (O), and quaternion (Q). Accelerometer and gyroscope data were acquired from the IMU BNO055 sensors directly, however, orientation and quaternion data were calculated by utilizing sensor fusion computation. Then all these data were further reorganized and grouped into 15 sensor data categories by adopting leave-one-out techniques. Table 2 showed the list of sensor data categories.

Table 1. Proposed list of ASL words

No.	Gesture	No.	Gesture
1	Good	15	Yes
2	Happy	16	Please
3	Sorry	17	Drink
4	Hungry	18	Eat
5	Understand	19	Look
6	Pretty	20	Sleep
7	Smell	21	Hearing
8	There	22	Water
9	You	23	Rice
10	Me/ I	24	Search
11	Ok	25	Onion
12	Hello	26	Apple
13	Bye	27	Vegetable
14	Thank you	28	None(Relax gesture)

Table 2. List of sensor data categories

No.	Gesture
1	Quaternion (Q)
2	Orientation (O)
3	Gyroscope (G)
4	Accelerometer (A)
5	Q + O + G + A
6	Q + O + G
7	Q + O + A
8	Q + G + A
9	O + G + A
10	Q + O
11	Q + G
12	Q + A
13	O + G
14	O + A
15	G + A

4.2 Data Preprocessing

Before data analysis and classification. all of the collected data were being preprocessed into two modality, i.e. standard deviation (STD) and mean (MEAN). STD was measured to observe degree of movement, and MEAN was calculated to observe the average or arithmetic mean value of one window size. Then, all these features were further normalized between a range of 0 to 1 in order to reduce the processing time and data complexity. Lastly, all these data were segmented into a window size of value 20 which equaled to the number of instance for a seconds of collected data.

4.3 Classification Model

This study proposed the state-of-art time series data classification model, long short-term memory (LSTM). In comparison of typical recurrent neural network, LSTM introduced functional gates to overcome the vanishing and exploding gradient for a long sequence, namely forget gate (f), input gate (i), and output gate (o). All of these function utilized sigmoid function to decide what information to pass through the cell. Sigmoid function had output value between 0 and 1, where 0 means let nothing to pass through and 1 means let everything to pass through.

Meanwhile, the LSTM network model was computed as showed in Fig. 2. There were six hidden layers consisting one LSTM layer, two dropout layers, and 3 dense layers. Dropout layers were included to prevent overfitting the it selectively remove number of neurons in a rate of 0.2% and 0.1% respectively. While, dense layers were included to fully connected all neurons of the network with rectified linear unit (ReLU) activation and Softmax activation function. Lastly, the model was compiled with adaptive moment estimation (Adam) optimizer and learning rate of 0.001.

Moreover, validation was designed to gauge generalizability of a model. Specifically, hold-out validation was introduced during the model training and testing phrase. This validation algorithm split the entire dataset into training and testing sets in a portion of 90% and 10% respectively.

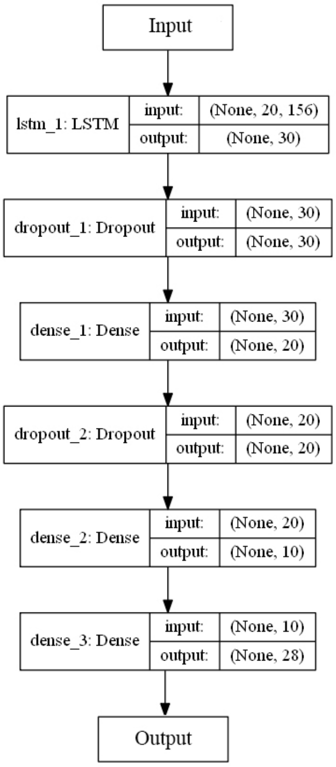

Fig. 2 Proposed LSTM model

4.4 Results

A system performance evaluation metrics named accuracy (ACC) was computed as below:

$$ACC = TP+TN \,/\, (TP+TN+FP+FN) \qquad (1)$$

where TP and TN were denoted as true positive and true negative respectively. While, FP and FN were denoted as false positive and false negative respectively. Figure 3. depicted the experiment results. The experiment results indicated that the integrated of four sensors data had the highest mean accuracy of 99.89% with STD and MEAN

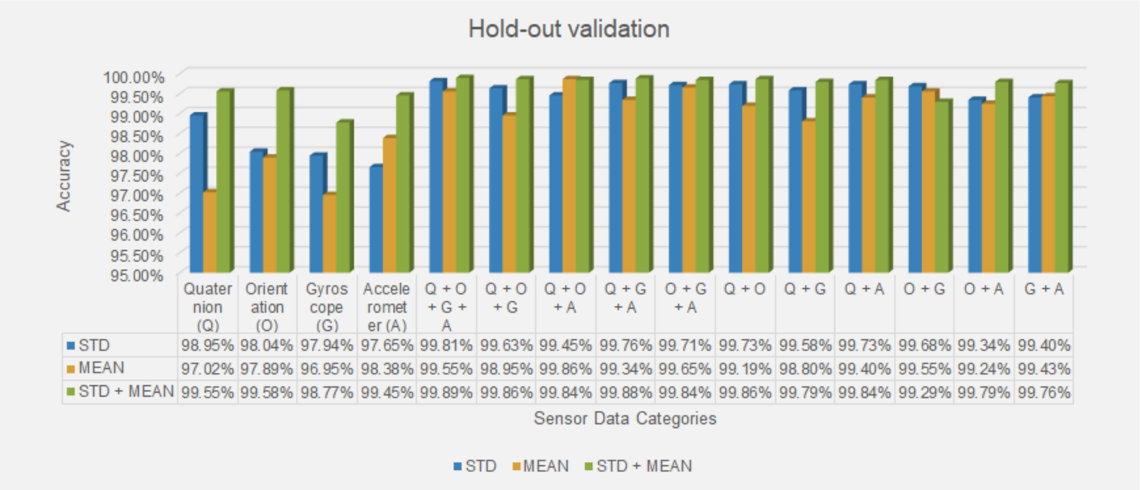

Fig. 3 Experiments results

features. While the single sensor data categories, Q, O, G, and A had similar lowest accuracy compare to the rest of sensor data categories, especially the G data with MEAN features, 96.95% only. On the hand, standard deviation (STD) showed better performance than mean (MEAN) feature, 99.23% and 98.88% respectively. While, the combination of both features had even higher mean accuracy of 99.67% among all sensor data category.

Ⅴ. Conclusion

Conclusively, the experiment results indicated that integration of four sensor data had the best system performance. However, among all proposed sensor data categories, gyroscope was found had the least modality to represent hand and finger movement. While, the number of sensor data integration was found has positive correlation to the system performance. The higher number of integrated sensor data the higher the system performance accuracy. Besides that, the combination of standard deviation and mean features were recommended for word-based ASL recognition system.

As the results, this proposed system successfully translated sign language into text and audio with high accuracies. Meanwhile, it also revealed great potential to vanish the current communication barriers in the society.

References

[1] T. Chong and B. Lee, “American Sign Language Recognition Using Leap Motion Controller with Machine Learning Approach,” *Sensors,* vol. 18, no. 10, 2018, pp. 35-54.

[2] M. J. Cheok, “A review of hand gesture and sign language recognition techniques,” *Int. J. Mach. Learn. Cybern.*, vol. 10, issue 1, 2019, pp. 131-153.

[3] M. A. Ahmed, B. B. Zaidan, and A. A. Zaidan, “A Review on Systems-Based Sensory Gloves for Sign Language Recognition State of the Art between 2007 and 2017,” *Sensors,* vol. 18, no. 7, 2018, pp. 1-44.

[4] I. Infantino, R. Rizzo, and S. Gaglio, “A Framework for Sign Language Sentence Recognition by Commonsense Context,” *IEEE Trans. Syst. Man, Cybern.* vol. 37, no. 5, 2007,

pp. 1034-1039.

[5] L. Ding and A. M. Martinez, “Modelling and Recognition of the Linguistic Components in American Sign Language Title,” *Image Vis. Comput.*, vol. 27, no. 12, 2009, pp. 1826-1844.

[6] H. Lane and F. Grosjean, *Recent perspective on American Sign Language.* New York, Pyschology Press, 2017.

[7] A. Haria, A. Subramanian, N. Asokkumar, and S. Poddar, “ScienceDirect ScienceDirect Hand Gesture Recognition for Human Computer Interaction,” *Procedia Comput. Sci.*, vol. 115, 2017, pp. 367-374.

[8] M. Elmezain and A. Al-hamadi, “A Hidden Markov Model-Based Isolated and Meaningful Hand Gesture Recognition,” *Int. J. Electr. Comput. Syst. Eng.*, vol. 3, no. 3, 2009, pp. 156-163.

[9] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, “Multi-sensor System for Driver's Hand-Gesture Recognition,” In Proc. *2015 11th IEEE Int. Conf. Work. Autom. Face Gesture Recognit.*, vol. 1, 2015, pp. 1-8.

[10] C. Preetham, G. Ramakrishnan, S. Kumar, and A. Tamse, “Hand Talk- Implementation of a Gesture Recognizing Glove,” In Proc. *2013 Texas Instruments India Educators' Conference*, Bangalore, India, 2013, pp. 328 - 331.

[11] K. Patil, G. Pendharkar, and P. G. N. Gaikwad, “American Sign Language Detection,” *Int. J. Sci. Res. Publ.*, vol. 4, no. 11, 2014, pp. 4-9.

[12] X. Wang, M. Xia, H. Cai, Y. Gao, and C. Cattani, “Hidden-Markov-Models-Based Dynamic Hand Gesture Recognition,” *Math. Probl. Eng.*, vol. 2012, 2012, p. 11.

[13] B. G. Lee and S. M. Lee, “Smart Wearable Hand Device for Sign Language Interpretation System With Sensors Fusion,” *IEEE Sens. J.*, vol. 18, no. 3, 2018, pp. 1224-1232.

[14] C. K. Mummadi, F. Leo, K. Verma, S. Kasireddy, P. Scholl, J. Kempfle, and K. Laehoven, “Real-Time and Embedded Detection of Hand Gestures with an IMU-Based Glove,” *Informatics,* vol. 5, no. 2, 2018, pp. 1-18.

[15] J. Kim and W. Lee, “An User-aware System using Visible Light Communication,” *J. of the Korea Institute of Electronic Communication Sciences,* vol. 14, no. 4, 2019, pp. 715-722.

[16] J. Jo, “Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance,” *J. of the Korea Institute of Electronic Communication Sciences,* vol. 14, no. 3, 2019, pp. 547-552.

[17] J. Jo, “Performance Comparison Analysis of AI Supervised Learning Methods of Tensorflow and Scikit-Learn in the Writing Digit Data,” *J. of the Korea Institute of Electronic Communication Sciences,* vol. 14, no. 4, 2019, pp. 701-706.

Authors

정택위(Teak-Wei Chong)

Teak-Wei Chong received B.M.M degree in 2015 from Multimedia University, Cyberjaya, Malaysia.

He is a final year master's degree student in Dept. Electronic Eng. at Keimyung University, Daegu, South Korea.

김범준(Beom-Joon Kim)

Beomjoon Kim received his B.E. degree in electronic engineering, M.S. and Ph.D. degrees in electronic engineering from Yonsei University in 1996, 1998, and 2003, respectively. Currently, he is a professor in Dept. Electronic Eng., Keimyung University, Daegu, South Korea, since 2006.