Recognition of Natural Hand Gestures Using Bidirectional Long Short-Term Memory Model

A-Ram Kim¹ and Sang-Yong Rhee²

¹Department of IT Convergence Engineering, Kyungnam University, Changwon, Korea

²Department of Computer Engineering, Kyungnam University, Changwon, Korea

Ijfis

Abstract

Recent studies have been trying to use voices and body gestures similar to human communication methods to control intelligent devices such as robots instead of giving commands using a keyboard or a mouse. Human being obtains about 80% of information using vision, and about 55% of all meaning of communication is related to vision. In addition, hand gestures are most frequently used in non-verbal communication. Therefore, there have been many researches to give commands to a robot. Existing studies were limited to recognition of fixed hand gestures and shapes, requiring users to be educated on motions that can be used to communicate with robots. To solve this problem, we use fuzzy inference to select meaningful gesture in continuous gesture at first. Hand position is interpolated using the Lagrangian method. Also, we use the Kalman filter for object-occlusion or self-occlusion. After generating a sequence of continuously received hand positions using a chain code, the fuzzy theory is used to select meaningful motions among various hand gestures. Eventually, we perform a method of recognizing meaningful hand gestures using a recurrent neural network with bidirectional long short-term memory (LSTM) architecture. Even though selecting the meaningful motion was hard, if it was selected correctly, the ratio of hand gesture recognition was very high in experiment results.

Keywords: Bidirectional LSTM, Hand gesture recognition, Natural hand motion, Recognition of direction command

1. Introduction

Body language is one of the non-verbal communication methods and refers to conveying the intention by using the body. Since gesture language is not only able to convey a relatively wide variety of meanings, but also because the language is intuitive enough to allow communication between human beings who use different languages, much researches have been done on gesture language. In general, gesture language is used at the same time as voice language, and hand gesture is the most frequently used gesture language. Therefore a lot of researches have been trying to give instructions to robots or intelligent devices using hand gestures.

Shin et al. [1] conducted a study of detecting each hand using skin color. To change shape, position and posture of a virtual object, hand gesture was used according to predefined motions. Malima et al. [2] proposed a system that after counting the number of fingers to recognize the hand, commands were given to the robot according to the number of fingers. A research using

Received: Aug. 2, 2018
 Revised : Dec. 17, 2018
 Accepted: Dec. 21, 2018

Correspondence to: Sang-Yong Rhee
 (syrhee@kyungnam.ac.kr)
 ©The Korean Institute of Intelligent Systems

©This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

natural hand motions has been also conducted. They used the chain code to represent the hand movements, recognition process was performed using a hidden Markov model (HMM), and gave the motion command to a robot [3, 4].

Pugeault and Bowden [5] created a data set by labeling ASL (American Sign Language) finger spelling using three-dimensional (3D) depth images and recognized the hand shape. This study had the disadvantage of recognizing only the still hand shape. Nishida and Nakayama [6] constructed a multi-stream recurrent neural network composed of several stages of long short-term memory (LSTM) and recognized hand movements. Though it showed 97.8% recognition ratio, it had serious limitation. Because the dataset was collected with only a hand and forearm visible while the background was not complicate. It is different from the actual situation in which the whole body of the user is viewed. Jiang et al. [7] created a mask that distinguished a user from the background using a 3D depth image, traced the hand using the particle filter, and recognized the motion using the condensation algorithm.

Neverova et al. [8] won with a recognition rate of about 85% by using deep neural network (DNN) at ChaLearn challenge in 2014. However, when they used the Kinect sensor data in actual situation, recognition rate declined to 74%. Chai et al. [9] used natural continuous hand motion. They obtained an region of interest (ROI) of a hand using faster region-based convolutional neural network (R-CNN) and tried to recognize the hand motion command in the ROI by using two streams recurrent neural network (2S-RNN) that two LSTMs were connected in parallel. The recognition rate was as low as 28.6%. It was shown that it was not easy to extract only command motion in continuous one. Marin et al. [10] used multi-class support vector machine (SVM) classifier to recognized hand gesture with leap motion and Kinect devices.

The crucial problem of most existing researches is that they used predefined instructions. If the instructions differ from device to device, a great deal of confusion will occur. Therefore, it is most reasonable to use the natural motion that human being commonly use on giving commands to an intelligent device using a hand gesture. However, there was a challenge to overcome in this method. Even though human being may move their hands for other purposes or unconsciously, a robot may mistakenly assume that human being give a command to it. In the case unexpected results might happen. In order to prevent such case, a research has been conducted using start and completion hand motion commands [4].

We thought the study suggested a limited solution using the

Figure 1. System flow.

specified hand gestures. The most convenient method is to distinguish hand commands from unconscious or meaningless gesture in continuous hand movement just like people without using the specified actions. A purpose of this study is to separate the command motion in continuous hand movement and recognize it using bidirectional LSTM.

The rest of this paper is composed as follows: in Section 2, we describe the proposed system. Experiments and analysis are described in Section 3. Section 4 provides some concluding remarks

2. System for Separating and Recognition of Command Motion

The overall of proposed system flow in this study is shown in Figure 1.

In order to recognize the command motions using the hand gesture, the joint and the hand position data are acquired from the color image, the 3D depth image, and the skeleton data of a human body from the Kinect Sensor V2. The hand position is continuously tracked using the Kalman filter. They are interpolated by using the Lagrangian method. Then these are expressed by the chain code to generate a sequence represented by the time. Fuzzy inference is used to extract the command motion and the moving speed. The moving speed refers to expressing ‘fast’, ‘slow’, ‘careful’, etc. together while presenting the direction of movement and determines the speed factor of the command to be transmitted to the intelligent device [11]. Finally, we use bidirectional LSTM to recognize the hand gestures that extracted by using the inference.

2.1 Acquisition of Hand Position Data

The Kinect Sensor V2 can recognize multiple dynamic objects at the same time, and the recognized dynamic objects are performed skeletonization to distinguish persons from the other dynamic objects. There may be several persons in front of the sensor. In this study, only the hand movements of the nearest person are recognized when the system is in operation.

After the user is determined, the position data of the user's hand is obtained. However, since the Kinect Sensor V2 skeletonizes dynamic objects, they try to recognize a similar object as a human if it's nearby. If the user is holding an object, the object also can be skeletonized. Therefore the position data of the joint cannot be correctly calculated. To prevent the situation, the hand position is confirmed using the skin color [6].

Hand position data acquired from the Kinect Sensor V2 isn't accurate due to errors frequently generated by noise. Also, since the hand cannot be traced if it is occluded by the other objects, the hand position data cannot be obtained normally. In this study, we partially solve this problem by using Kalman filter [11].

Since the distance between the camera and the user changes, the magnitude of the motion also changes depending on the position of the person. In addition, since each person does not have the same height, arm length, hand moving speed, and hand moving distance, we use the Lagrangian method to interpolate each data.

The recommended distance for the Kinect Sensor V2 is 2-3 m. Therefore, the X and the Y coordinate values are interpolated by determining the reference value of 3 m. N-dimensional polynomial using Lagrange interpolation is as shown in Eq. (1).

$$f_n(x) = \sum_{i=0}^n L_i(x) f(x_i). \quad (1)$$

The polynomial $f(x)$ passing through all $n + 1$ points (x_i, y_i) is obtained as shown in Eq. (2). Since $f(x)$ have to pass through all points, Eq. (3) holds.

$$f(x) = a_0 + a_1x + a_2x^2 + \dots + a_nx^n, \quad (2)$$

$$f(x_i) = y_i (i = 0, 1, \dots, n). \quad (3)$$

$L(x)$ is a weighting function and is given by the following Eq. (4).

$$L(x) = \prod_{i=0j \neq i}^n \frac{x - x_j}{x_i - x_j}. \quad (4)$$

In this study, a third order polynomial is obtained by applying a minimum distance of 1.5 m, a recommended distance of 2 m and 3 m, and a maximum user distance of 4 m. Because the data is changed depending on the user's x-coordinate value and the length of the user's arm, the interpolation is also performed.

Figure 2. Configuration of 4-way chain code.

2.2 Convert to Chain Code

In this study, we use two dimensional 4-way chain code and 26-way 3D chain code to recognize the hand movement using hand position data obtained at 0.1 second interval. A 4-way chain code using 2D data is expressed in the following manner. The difference between the previous horizontal position x_1 and the current horizontal position x_2 of the hand is used to indicate which direction the user has moved in the east or west direction and the difference between the previous vertical axis position y_1 and the current vertical axis position y_2 of the hand is used to express which direction the south or north direction has been moved.

Using the Δx and Δy obtained from the difference between the previous and current hand position, calculate the angle of the hand movement as shown in Eq. (5). We construct a 2D chain code using the obtained angle and direction (Figure 2).

$$\begin{aligned} &\text{if } x_1 - x_2 > 0 \text{ then} \\ &\quad \text{Direction} = \text{Left} \\ &\text{else Direction} = \text{Right} \\ &\text{if } y_1 - y_2 > 0 \text{ then} \\ &\quad \text{Direction} = \text{South} \\ &\text{else Direction} = \text{North} \end{aligned} \quad (5)$$

$$\theta = |\Delta y| \sqrt{\Delta x^2 + \Delta y^2},$$

$$\angle a = \sin^{-1}(\theta).$$

Also, the 3D direction can be discriminated by using the difference Δz on the Z axis obtained from the position of the previous and current hand. The 3D chain code representing the 26 ways based on the centroid of the cube is shown in Figure 3.

Figure 3. Composition of 3D chain code.

2.3 Hand Motion Recognition Using the LSTM

In this study, when a directional hand gesture is given as a sequence of 3D chain code, the fuzzy rule is used to judge whether or not the gesture is a command gesture. If it is judged to be instructive, it is inferred what kind of the command is performed using the LSTM.

As mentioned earlier, we try to recognize the gesture on the behaviour that people commonly use. However, in order to recognize a person's general actions, the following problems arise. Apart from communicating with robots, people move their hands unconsciously, habitually, and reflexively in the situation such as office or home, staying for a long time and live naturally.

Then, when it is necessary to give a command the robot, we raise up the hand as shown in Figure 4(b) and give a command (Figure 4(c)) and then lay down the hand as shown in Figure 4(d). The motion from Figure 4(a) to 4(b) is the one in order to generate the motion from Figure 4(b) to 4(c). The motion from Figure 4(c) to 4(d) is the action of lay down the arm because it is not necessary to hold the arm after giving the command to the robot. In this process, human being recognizes only the command motion without paying attention to movements that raise or lower the arm. Someone may repeat the same motion several times to emphasize the command.

In this series of motions, we select only the one that is determined to be the command [11], and recognize by the LSTM which command this motion is. The LSTM structure is used for hand gesture recognition is shown in Figure 5. The number of units of the input and the output layer is set to 26, which is the number of 3D chain codes, and when each sequence of motions is sequentially given, it is represented as one-hot encoding. The embedding layer maps the value received from the input layer to an embedding space of a specific size and represents each

Figure 4. Distinguishing behaviour of hands.

Figure 5. Structure of LSTM model.

motion sequence as a vector.

Since the constructed neural network model is bi-directional, two pairs of 256 LSTM units are used in the hidden layer. The output layer hands out a hand gesture having the highest probability value among all defined ones.

The probability value of each hand command is calculated with respect to the 3D chain code which is given at regular intervals, and the final output result is judged as the command.

3. Experiments

The environment for testing the proposed method is as follows: the operating system is Windows 10, and 1920×1080 color images and a 512×424 distance image at 30 frames per second are given from a Kinect Sensor V2. Visual Studio 2015 processes images using OpenCV 3.2.0 and the Kinect for Windows SDK 2.0 library from Microsoft. After a person object is extracted from the depth image, skeleton information can be obtained, and display it to the monitor screen. Since the detection range of the Kinect Sensor V2 is limited to 0.5-4.5 m, and the minimum distance for recognizing the upper body of

Figure 6. Defined hand commands. (a) Forward, (b) backward, (c) left, (d) right, (e) turn left, and (f) turn right.

Table 1. Neural network structure used in experiments

Model	LSTM	BLSTM	RNN	BRNN
Input layer	26	26	26	26
Embedding layer	200	200	200	200
Hidden layer	512	256×2	512	256×2
Output layer	6/12	6/12	6/12	6/12

the user is 1.5 m, the distance between the user and the camera is set to 1.5 m–4 m. The framework used for deep learning is Microsoft Cognitive Toolkit (CNTK) 2.0 Release Candidate 2 and uses Python 3.6.0.

In this study, the six hand motions defined in the output layer are the natural hand motions used by Koreans [4]. Six commands are ‘forward’, ‘backward’, ‘left shift’, ‘right shift’, ‘clockwise rotation’, and ‘counter-clockwise rotation’ as shown in Figure 6. We believe all of human being can be understood them.

In this study, the data used for training and test consisted of six motions that were performed by five people, and the total number of data was 24,000, consisting of sequences of 3DI chain codes. The whole data was divided into training set 90% and test set 10%.

We wanted to use the best model among the four deep-learning models of LSTM, bidirectional LSTM (BLSTM), RNN,

Table 2. Comparison of the proposed method and the results of existing studies

	HMM (%)	Our system (%)
Forward	95	100
Backward	100	100
To leftside	95	100
To rightside	100	100
Turn left	85	100
Turn right	95	100

and bidirectional RNN (BRNN) as shown in Table 1. First, experiments were conducted to find an effective learning method in the current data set. Experiments were conducted using the general LSTM structure shown in Table 1, and the Adam, FSAdaGrad, RMSprop, Momentum-SGD, and SGD were used as the training method. We used 0.002 as the learning rate and 50 as the epoch. We got the result that Adam method is most effective by varying the learning rate according to epoch, and the ReLU function showed the best result for the sigmoid, tanh, ReLU and softsign activation function tests.

Using the selected Adam and ReLU, the best model with the lowest error rate among the four models was selected and the final experiment was conducted. The structure of the neural network used in the experiment is shown in Table 1. Why the output layer is divided by 6 and 12 was to compare the existing paper that used HMM to recognize the hand motion. It was divided into four sub-motions per rotation direction to recognize hand rotation [8]. We also experimented with two kinds of labels, one of which is divided into four motions as the paper (4 units) and the other of which is composed of one label (1 unit). The training time of the RNN was more than two times faster than that of other models, but the error rate of classification of the BLSTM was the lowest. When the rotation motion was used as 4 units, the error rate is 0.23% in the validation dataset. It was much lower than 1 units. Eventually we decided to use the 4 units for the main experiment.

In order to know the recognition rate of each command, the experiment using BLSTM was performed in the same way as the conventional HMM. In order to meet the same experimental conditions, the start point and the end point were determined based on the assumption that the meaningful hand motion was accurately extracted in the above experiment. The number of experiments was 50 consecutive times for six operations.

Experimental results showed that the recognition rate of the

Table 3. Comparison with other methods

	[6]	[7]	Proposed method for extracted commands
Recognition rate (%)	97.8	95.9	100

rotation commands was not high when the HMM was used (Table 2), but when the BLSTM was used, it was confirmed that the recognition rate was 100% in all the command as well as the rotation, so we could know the BLSTM model constructed in this study was superior.

Comparisons with other studies are as follows: first, it is about research to select command motions. Chai et al. [9] was performed using the 2S-RNN model with the LSTM. In this study, the selection rate was 70% [12] and in the study of Chai et al. [9], was 28.6%.

Table 3 compares the results with similar studies that recognized hand gestures. Since the environments in which the experiments were conducted are not the same, it is not possible to conclude of which the research results are good using only the recognition rate. Nishida and Nakayama [6] showed a very high recognition rate of 97.8% when only the hand was photographed by using the Kinect sensor fixed. Jiang et al. [7] reported recognition rate of 95.9% on recognizing hand gesture of a sitting person.

4. Conclusions

In this study, we have studied communication between people and intelligent devices using the hand gestures that people use in real life, without artificially pre-determined hand gestures or the specific action to start hand gesture. We obtained 3D depth images and color images using Kinect Sensor V2, acquired human joint position data from the depth image using the Microsoft's Kinect API, and verified hand position using the color model. The Lagrangian method was used to interpolate the movement of the hand according to the distance from the camera and to solve the occlusion caused by the other objects and self-occlusion problem when the hand was tracked by using the Kalman filter.

If it was judged that the movement of the hand is a command operation, the hand gesture was inferred using the BLSTM, and the recognition ratio was very good. In the future, the research will be combined with recognition of user's eye focus or voice to further improve communication capability with robots.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgements

This paper is an excerpt from a part of a Ph. D. thesis.

References

- [1] M. C. Shin, L. V. Tsap, and D. B. Goldgof, "Gesture recognition using Bezier curves for visualization navigation from registered 3-D data," *Pattern Recognition*, vol. 37, no. 5, pp. 1011-1024, 2004. <https://doi.org/10.1016/j.patcog.2003.11.007>
- [2] A. K. Malima, E. Ozgur, and M. Cetin, "A fast algorithm for vision-based hand gesture recognition for robot control," in *Proceeding of the IEEE Conference on Signal Processing and Communications Applications*, Antalya, Turkey, 2006, pp. 17-19. <https://doi.org/10.1109/SIU.2006.1659822>
- [3] H. K. Lee and J. H. Kim, "An HMM-based threshold model approach for gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 961-973, 1999. <https://doi.org/10.1109/34.799904>
- [4] A. R. Kim and S. Y. Rhee, "Recognition of natural hand gesture using HMM," *Journal of Korean Institute of Intelligent Systems*, vol. 22, no. 5, pp. 639-645, 2012. <https://doi.org/10.5391/JKIS.2012.22.5.639>
- [5] N. Pugeault and R. Bowden, "Spelling it out: real-time ASL fingerspelling recognition," in *Proceedings of 2011 IEEE International Conference on Computer Vision Workshops*, Barcelona, Spain, 2011, pp. 1114-1119. <https://doi.org/10.1109/ICCVW.2011.6130290>
- [6] N. Nishida and H. Nakayama, "Multimodal gesture recognition using multi-stream recurrent neural network," in *Image and Video Technology*. Cham: Springer, 2015, pp. 682-694. https://doi.org/10.1007/978-3-319-29451-3_54
- [7] H. Jiang, B. S. Duerstock, and J. P. Wachs, "A machine vision-based gestural interface for people with upper extremity physical impairments," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, No. 5,

- pp. 630-641. 2014. <https://doi.org/10.1109/TSMC.2013.2270226>
- [8] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Computer Vision - ECCV 2014 Workshops*. Cham: Springer, 2014, pp. 474-490. https://doi.org/10.1007/978-3-319-16178-5_33
- [9] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen, "Two streams recurrent neural networks for large-scale continuous gesture recognition," in *Proceedings of 23rd International Conference on Pattern Recognition*, Cancun, Mexico, 2016, pp. 31-36. <https://doi.org/10.1109/ICPR.2016.7899603>
- [10] G. Marin, F. Dominio, and P. Zanuttigh, "Hand gesture recognition with leap motion and Kinect devices," in *Proceedings of 2014 IEEE International Conference on Image Processing*, Paris, France, 2014, pp. 1565-1569. <https://doi.org/10.1109/ICIP.2014.7025313>
- [11] A. R. Kim and S. Y. Rhee, "Mobile robot control using natural hand motion," *Journal of Korean Institute of Intelligent Systems*, vol. 24, no. 1, pp. 64-70, 2014. <https://doi.org/10.5391/JKIIS.2014.24.1.064>
- [12] A. R. Kim, "Recognition of Meaningful hand gesture using fuzzy inference and recurrent neural network with bidirectional long short-term memory," Ph.D. dissertation, Kyungnam University, Changwon, Korea, 2017.

A-Ram Kim received his B.S. and M.S. degrees in Computer Engineering and Advanced Engineering from Kyungnam University, Changwon, Korea, in 2011 and 2013, respectively, and is received the Ph.D. degree in Advanced Engineering at Kyungnam University, Changwon, Korea. His present interests include intelligent robot, image processing and pattern recognition.
E-mail : han0440@naver.com

Sang-Yong Rhee received his B.S. and M.S. degrees in Industrial Engineering from Korea University, Seoul, Korea, in 1982 and 1984, respectively, and his Ph.D. degree in Industrial Engineering at Pohang University, Pohang, Korea. He is currently a professor at the Computer Engineering, Kyungnam University, Changwon, Korea. His research interests include computer vision, augmented reality, neuro-fuzzy and human-robot interface.
E-mail : syrhee@kyungnam.ac.kr