Japanese Sign Language Recognition Based on Three Elements of Sign Using Kinect v2 Sensor

Shohei Awata(✉), Shinji Sako, and Tadashi Kitamura

Graduate School of Engineering, Nagoya Institute of Technology, Aichi, Japan
{awata, sako, kitamura}@mmsp.nitech.ac.jp

Abstract. The visual feature of Japanese sign language is divided into two of manual signals and non-manual signals. Manual signals are represented by the shape and motion of the hands, and convey mainly the meaning of sign language words. In terms of phonology, sign language words consist of three elements: hand's motion, position, and shape. We have developed a recognition system for Japanese sign language (JSL) with abstraction of manual signals based on these three elements. The abstraction of manual signals is performed based on Japanese sign language words dictionary. Features like coordinates of hands and depth images are extracted from manual signals using the depth sensor, Kinect v2. This system recognizes three elements independently and the final result is obtained under the comprehensive judgment from the results of three elements recognition. In this paper, we used two methods for recognition of hand shape, a contour-based method suggested by Keogh and template matching of depth image. The recognition methods of other elements were hidden Markov model for recognition of motion and the normal distribution learned by maximum likelihood estimation for recognition of position, as a same manner of our previous research. Based on our proposal method, we prepared recognition methods of each element and conducted an experiment of 400 sign language words recognition based on a sign language words dictionary.

Keywords: Sign language recognition · Kinect · Hand pose · Contour · Template matching

1 Introduction

In general, sign is represented by combinations of posture or movement of the hands and facial expressions such as eyes or month. These visual features of sign are happened both sequentially and simultaneously. Communication between the hearing people and the deaf can be difficult, because the most of hearing people do not understand sign language. To resolve a communication problem between hearing people and deaf, projects for automatic sign language recognition (ASLR) system is still under way.

One of major problem of current ASLR system is performing small vocabulary. Corresponding to the unknown vocabulary is also important from the view of practical aspect. It is said that the number of JSL vocabulary is over 3,000. In addition, a new

© Springer International Publishing AG 2017
C. Stephanidis (Ed.): HCII Posters 2017, Part I, CCIS 713, pp. 95–102, 2017.
DOI: 10.1007/978-3-319-58750-9_13

sign is introduced to adjust the situation. Obviously, it is inefficient to perform the recognition on individual sign units.

From the point of view, we employ a JSL dictionary and notation system proposed by Kimura et al. [1]. Our system is based on three elements of sign language: hand motion, position, and pose.

This study considers a hand pose recognition using depth image obtained from a single depth camera. We apply the contour-based method proposed by Keogh et al. [2] to hand pose recognition and evaluated by comparison of typical template matching method. The contour-method recognizes a contour by means of classifiers trained from several hand shape contours.

To recognize hand motion and position, we adopted statistical models such as Hidden Markov models (HMMs) and Gaussian mixture models (GMMs). To address the problem of lack of training data, our method utilizes the pseudo motion and hand shape data. We conduct experiments to recognize 400 JSL sign targeted professional sign language interpreters.

2 Overview of the System

An overview of our proposed system is shown in Fig. 1. The features of sign motion are captured by using Microsoft Kinect v2 sensor [3]. At first, time series of hand position is split into moving segment. Second, the three phonological elements are recognized individually by using hand position and hand depth image. Finally, the recognition result is determined by the weighted sum of each score of three elements. The recognition process of the hand pose and other two components employs depth data of the hand region and coordinates of joints, respectively.

We used JSL dictionary proposed by Kimura et al. [1]. In this dictionary, hand poses are classified by several element as shown in Table 1. These elements are also illustrated in Fig. 2. Currently, the vocabulary of this dictionary is approximately 2,600.

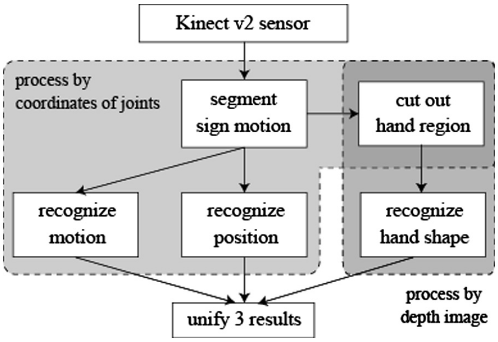

Fig. 1. Flowchart of the entire system.

Table 1. Portion of the database in the dictionary.

Word	SL type	Hand type	Palm direction	Position	Motion
Love	3	B	Down	NS	Circle
Between	4	B	Side	NS	Down
Blue	1	B	Back	Lower face	Back
Red	f	1	Back	Lower face	Right
Baby	4	B	Back	Whole face	Front back

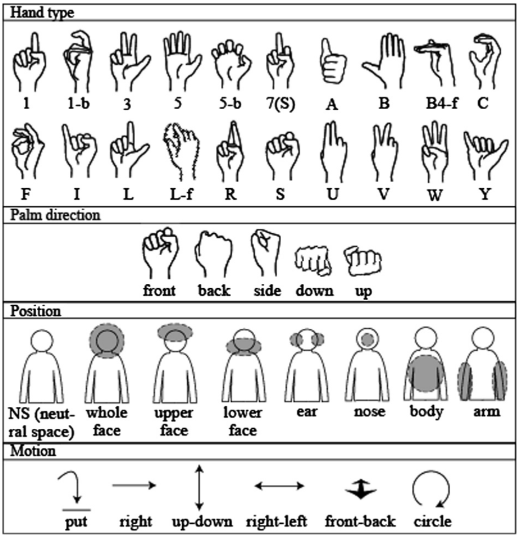

Fig. 2. Elements in sign language dictionary.

3 Hand Pose Recognition

Several study on hand pose recognition using a technique of estimating the finger joints has been proposed [4, 5]. However, these methods still have difficulties when some fingers are invisible due to the complex hand shapes of sign language. From the point of view, we adopt the contour-based technique proposed by Keogh et al. [2] to recognize hand pose. This technique is considered to be robust even when the finger is partially occluded. The details of the method are described below.

3.1 Feature Extraction

Hand shapes can be converted to distance vectors to form one-dimensional sequence. Figure 3 shows the procedure to extract a distance vector from a hand image. At first, the center point of the hand region is determined by distance transform. Distance transform convert one-pixel value of the binary image with the distance between the

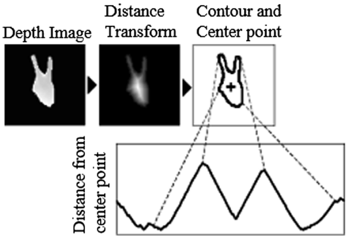

Fig. 3. Feature extraction from an image of hand region

nearest zero value pixel. Next, each distance from the center point to every pixel on the contour is calculated. The distance vector represents a series of these distances.

3.2 Calculation of Distance

A distance D between two distance vectors $P = \{p_0, p_1, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$ is calculated according to the followings.

$$D(P,Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{1}$$

If the length of two distance vectors is different, some normalization process should be required such as dynamic time warping (DTW). To simplify, we adjust length of vector to be same in advance for low computation cost reason.

It can be compared contours by calculating their distances or using classifiers generated from contours. These classifiers are called wedges. Wedges have set of maximum and minimum values at each point. If a contour is located inside a wedge, the distance is zero. The distance D between a wedge W ($U = \{u_o, u_1, \ldots, u_n\}$ means its top, $L = \{l_0, l_1, \ldots, l_n\}$ means its bottom) and a contour $P = \{p_0, p_1, \ldots, p_n\}$ can be calculated by following equation.

$$D(W,P) = \sqrt{\sum_{i=1}^{n} \begin{cases} (p_i - u_i)^2 & (p_i > u_i) \\ (p_i - l_i)^2 & (p_i < l_i) \\ 0 & (otherwise) \end{cases}} \tag{2}$$

3.3 Generate Wedges

Wedges are produced according to the following steps.

1. Extract features from hand images
2. Calculate distances of all contours
3. Combine two contours in ascending order of distances. Wedge is represented by set of maximum and minimum values of merged contours.

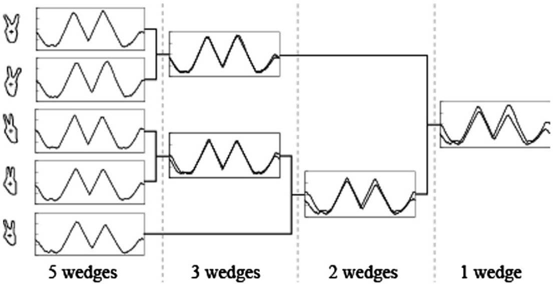

Fig. 4. Making wedges from five contours

Repeat Step 3 until the pre-determined number of wedges. The step of generating wedges is also illustrated in Fig. 4. We prepare various wedges to recognizing each hand type.

4 Sign Movement and Position Recognition

In this paper, HMMs are utilized to recognized hand movement using the feature parameter of hand position provided by the Kinect sensor. 3-dimensional hand position and its speed are used as feature parameter of HMMs. HMMs corresponding to the typical movement of sign are constructed from pseudo-training data. It can be omitted the cost of collecting the sign data. The definition of the hand position is ambiguous in JSL. It is necessary to consider for the hand position recognition. In this paper, the particular position of the hand in sign is modeled by GMMs. 3-dimensional hand position is used as feature parameter of GMMs. GMMs corresponding to the typical position of sign are also trained from pseudo-training data.

5 Experiments

We conduct JSL words recognition experiments by recognizing three elements independently. In order to recognize the hand shape, we used a contour-based method and template matching.

5.1 Experimental Condition

We use 400 JSL words commonly used in the social life for the test data. To recognize this 400 words requires to distinguish 24 hand poses defined by hand types and palm directions. Because hand shapes transform with motions, each hand type is not separated even if the palm direction is different. However, there are a few exceptions to distinguish sign language words which have same motion, position, and hand types, but only palm direction is different.

To simplify the collection of data in our experiments, we used depth images of stationary hand instead of hand images obtained during natural sign motion. Table 2

Table 2. Condition of shape recognition

	Contour-based method	Template matching
Number of test data	223 words × 2 speakers × 2 trials = 892 data	188 words × 2 speakers × 2 trials = 752 data
Image size of test data	120 × 120 pixel	120 × 120 pixel
Image for recognition	120 × 120 pixel 150 images per person, hand type	90 × 90 pixel 12 × 18 images per person and hand type (Rotate by 20°)
Hand type	24 types	

Table 3. Condition of position and motion recognition

Position type	8 types
Features	3-dimensional hand coordinates
Number of training data	6 hand coordinates per a position
Number of GMM mixes	1–6
Motion type	40 types
Features	10 pseudo hand movements per a motion
Number of training data	3-dimensional direction vector + Four-dimensional information on speed
Number of states of HMM	5–18
Number of GMM mixes	1

shows the condition of shape recognition by contour-based method and template matching. The similarity used in template matching is calculated by a method incorporating normalization by luminance. 12 template images were selected from each of the belts when the number of the belt was 12. The target image is the frame with the slowest speed in the sign language movement.

Table 3 shows the condition of position and motion recognition. For the parameters required for position recognition, at each operating position of sign language, draw a circle by hand and use the coordinates of the hand obtained at that time. In the training of HMM, we performed motions that reproduced the movement pattern of dictionary data 10 times and trained the parameters from the obtained feature values.

After recognizing hand shape, position, and motion for the test data, sign language word can be determined by weighted sum of each score of three elements.

5.2 Results

Table 4 shows the results of JSL words recognition experiments. The scores of the three elements are weighted after performing normalization so that the maximum

Table 4. Word recognition rate (%)

	Shape	Position	Motion	Word
Contour-based (Weight)	28.7 (0.1)	78.3 (0.4)	60.0 (0.5)	33.8
Template matching (Weight)	32.7 (0.1)	73.7 (0.7)	51.0 (0.2)	28.1

values are equal. In recognition of hand shape, the recognition rate of template matching was 32.7%, which was better than the contour-based method. Word recognition rate by contour-based method was 33.8%, and word recognition rate by template matching was 28.1%. In either method, the recognition accuracy of the hand shape was the lowest among the recognition of the three elements. One of the main causes of misrecognition is difficulty in recognizing the hand shape during sign language motion using a single camera image. It is assumed that hand shape weight was suppressed to the minimum because hand shape recognition accuracy was low.

6 Conclusion

In this research, we proposed a method to recognize sign language words by constructing recognition models corresponding to hand shapes, hand positions and movements, which are three elements of sign language, based on the notation method of Japanese sign language/Japanese dictionary system. In sign language recognition research, it is difficult to obtain a sign language database currently. As in this research, the method of introducing the sign language academic knowledge and determining the constituent elements of sign language by top down has the advantage that a small number of learning data is enough. Therefore, our method can be said to be suitable for sign language recognition research. Furthermore, by using a sign language word dictionary with a large number of recorded words, we can expect to develop into large vocabulary recognition in the future.

We also conducted sign language word recognition experiments on Japanese sign language words. In this research, pseudo data corresponding to each element of sign language was used as learning data, and recognition was attempted for actual sign language motion. In recognition of hand shape, the recognition rate of template matching was 32.7%, which was better than the contour-based method. Word recognition rate by contour-based method was 33.8%, and rate by template matching was 28.1%.

Improvement of hand shape recognition method and improvement of learning data are future issues.

Acknowledgment. This research was partially supported by JSPS KAKENHI Fostering Joint International Research (15KK0008).

References

1. Kimura, T., Hara, D., Kanda, K., Morimoto, K.: Expansion of the system of JSL-Japanese electronic dictionary: an evaluation for the compound research system. In: Kurosu, M. (ed.) HCD 2011. LNCS, vol. 6776, pp. 407–416. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21753-1_46
2. Keogh, E., Wei, L., Xi, X., Lee, S.-H., Vlachos, M.: LB Keogh supports exact indexing of shapes under rotation invariance with arbitrary representations and distance measures. In: 32nd International Conference on Very Large Data Bases (VLDB2006), pp. 882–893 (2006)
3. Kinect for Windows. http://kinectforwindows.org
4. Liang, H., Yuan, J., Thalmann, D.: Parsing the hand in depth images. IEEE Trans. Multimedia **16**(5), 1241–1253 (2014)
5. Tang, D., Yu, T.-H., Kim, T.-K.: Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In: Proceedings of the 2013 IEEE International Conference on Computer Vision, ICCV 2013, pp. 3224–3231 (2013)