Investigation of Feature Elements and Performance Improvement for Sign Language Recognition by Hidden Markov Model

Tatsunori Ozawa, Hirotoshi Shibata, Hiromitsu Nishimura, and Hiroshi Tanaka(✉)

Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi-Shi, Kanagawa, Japan
{sl32l059, sl585009}@cce.kanagawa-it.ac.jp,
{nisimura, h_tanaka}@ic.kanagawa-it.ac.jp

Abstract. Sign language is commonly used as one means of communication for hearing-impaired or speech-impaired people. However, there are many difficulties in learning sign language. If automatic translation for sign language can be realized, it would be extremely valuable and helpful not just to those who are physically impaired but to unimpaired people as well. The cause of the difficulty in automatic translation is that there are many kinds of specific hand motions and shapes, which make it difficult to discriminate each motion. Consequently, this has a negative impact on accurate recognition. This paper presents a recognition method that is able to maintain accurate recognition of different signs that encompass a multitude hand motions and shapes. The main feature of our approach is the use of colored gloves to detect hand motions and shapes. For our investigation, a recognition scheme using HMM (Hidden Markov Model) has been introduced to enhance recognition performance. In this scheme, performance depends on the feature elements extracted from each sign language motion. Feature elements of sign language motions and their unification are investigated, and the recognition performance is clarified using these feature elements and compared with each result. Although the percentage of recognition successes for each feature element is low, from 21.7% to 42.7%, it was shown that recognition success for the combined element results increased from 55.2% to 61.9% for 25 different sign language motions. In addition, the removal of candidates was also examined to enhance performance as a form of preprocessing using a threshold obtained from DP matching. It is also confirmed through experiments that the recognition success rate increased by a few percentage.

Keywords: Sign language · Color gloves · Optical camera · Hidden Markov Model · Feature elements · DP matching

1 Introduction

Sign language is a widely used communication method for hearing or speech impaired people. It is quite difficult to learn sign language. If automatic translation for sign language could be realized, it would become very meaningful and valuable to both

M. Antona and C. Stephanidis (Eds.): UAHCI 2017, Part II, LNCS 10278, pp. 76–88, 2017.
DOI: 10.1007/978-3-319-58703-5_6

impaired people and physically unimpaired people. Although interpretation from sign language to speech has also been studied for many years, the technologies have not yet matured to the level at which they can be put into practical use. Specifically, some methods that use a special sensor or device [1, 2] are associated with high introduction costs or sensors must be attached to the body. The detection target is mainly limited to hand motions and hand shape, and finger motions are not included. For this reason, achieving highly accurate recognition is difficult and the number of words that can be recognized using these methods is limited [3, 4]. Thus, the results that can be obtained with existing technologies are insufficient in terms of developing a system that can be put into practical use. In addition, because the scenario for usage has necessarily been limited in experiments, in previous studies, researchers have not even attempted to put them to practical use.

The authors have been investigating a method of sign language recognition using the optical camera and colored gloves shown in Fig. 1 [5, 6]. Since an optical camera is implemented in smartphones, this configuration can be used anywhere, although the colored gloves are also necessary. By using an optical camera and colored gloves, each finger can be discriminated by color and therefore hand shape can be correctly detected.

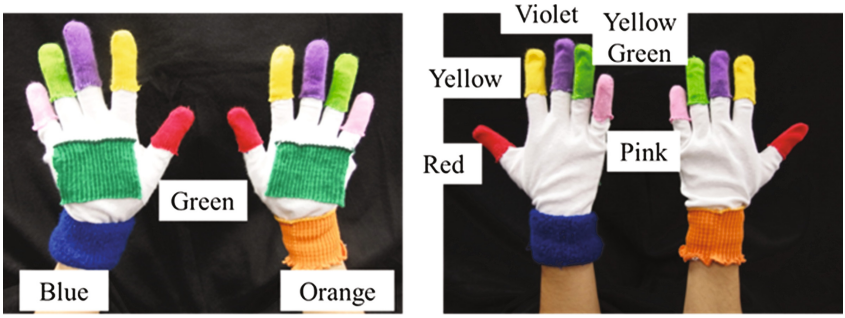

Fig. 1. Proposed colored gloves (Color figure online)

Automatic translation is our final goal, but remains too difficult to realize with current technology. Therefore, the authors are now trying to produce a kind of learning tool for sign language. Video data relevant to sign language can be obtained from the web site. Figure 2 shows an image from an instruction video demonstrating sign language motion. A learner memorizes the motions of each sign from this video. However, it is quite difficult not only to memorize the motion but also to confirm the validity of the motion memorized from the video. A tool for checking the learned motion is essential. An example of the use of our current research technology is shown in Fig. 3. After memorizing the sign language motion, the learner displays the same motion in front of a web camera connected to a PC. If the PC recognizes his/her motion, the positive result shows on the display. The learner uses this system as a review tool for sign language.

Fig. 2. Instructional video for sign language motions

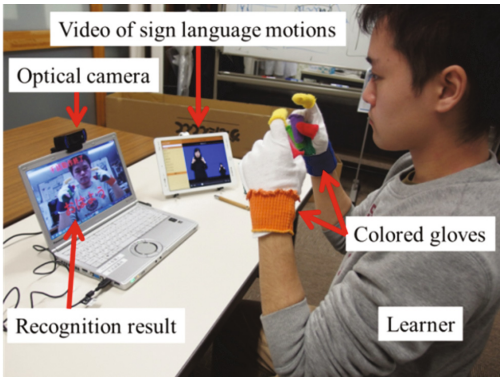

Fig. 3. Application of current investigation

In this paper, we describe recognition schemes for sign language motions, and experimental results. The hidden Markov model (HMM), which is used for recognition in several fields such as voice recognition, was used in this investigation. Since this method can easily be adapted to the recognition of a wide variety of feature elements of motions, excellent recognition performance can be expected.

2 Preparation of Learning Models for Recognition

The authors used DP matching [7] for sign language recognition in an earlier investigation [5, 6]. Hand motions were detected as the motion of the center of gravity of the colored region on the wrists. We decided to use the HMM recognition method, as this model can include many kinds of feature elements from video image data representing sign language motion, which can be useful in the instructional process. Improved recognition can be expected when the appropriate feature elements of sign language motions are used.

The instructional video data for sign language words included in Smart Deaf [8] is divided into 35 categories based on the usage areas of each sign. Each category includes roughly 50 to 100 words. In this study, recognition targets in sign language motions are selected according to the words included in each category. Demand for words relevant to medical and health issues is high, so we selected 25 words from this category (Table 1).

Table 1. Target sign language words

1.アトピー 1. Atopic	2.おしっこ 2. Urinary	3.ガン 3. Cancer	4.コンタクト 4. Contact lens	5.喘息 5. Asthma
6.体調 6. Physical condition	7.ハゲ 7. Bald	8.発熱 8. Fever	9.病気 9. Sickness	10.盲腸 10. Cecum
11.顔が赤い 11. Blushing	12.カテーテル 12. Catheter	13.禁煙 13. No smoking	14.喫煙 14. Smoking	15.薬を飲む 15. Take medicine
16.呼吸 16. Breath	17.耳鼻科 17. Otorhinology	18.頭痛 18. Headache	19.摘出 19. Remove	20.糖尿病 20. Diabetes
21.脳卒中 21. Stroke	22.吐き気 22. Nausea	23.鼻水 23. Runny nose	24.昼寝 24. Nap	25.虫歯 25. Tooth decay

One of the most important tasks in a recognition investigation is to compose the dataset for a recognition experiment and its evaluation. It is especially important to have sufficient motion data for each sign language word in order to create the learning model for HMM. Multiple motion data sets from multiple operators are required for the learning process. The recognition process uses HMM and calculates the likelihood values that the input motion data represents each possible sign language motion. The word with the highest likelihood value is selected as the recognition result.

Clearly, it is quite important to gather correct motion data. Therefore, the authors asked for the cooperation of the person in charge of making the motion video of Smart Deaf to compose the set of motion data used in this investigation. Figure 4 shows the signer directing and checking the sign language motions of the experimenter. The motion data for 25 words were recorded for learning and evaluation. The number of persons used for collecting motion data, and the number of samples for each motion are indicated in Table 2. The total number of collected motion data samples was 2250 ($60 \times 25 + 30 \times 25$). The number of experimenters and the number of samples for each word were based on the pre-study as the optimum for applying the learning HMM and recognition performance. The conditions under which the motion data were captured are as follows.

(i) Camera image resolution of 800×600 pixels is selected.
(ii) Illumination is set at about 200 lx for both the camera side and singer side.
(iii) Frame rate is 30 fps (frames per second). This is the maximum rate for a standard Web camera and smartphone.
(iv) The distance between the camera and signer is one mater, as this distance is considered to coincide with a real situation.

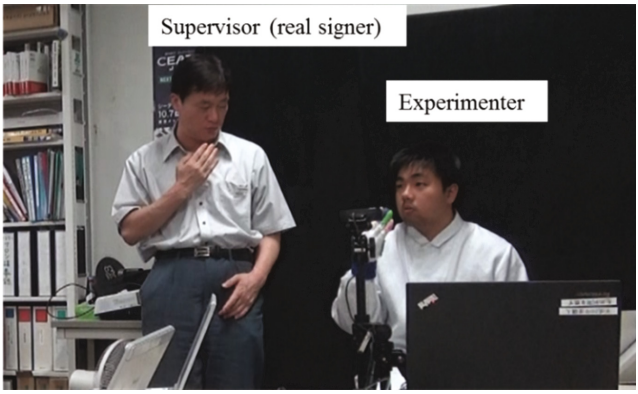

Fig. 4. Scene of sign language instruction

Table 2. Collected motion data for learning and evaluation

	Singers	Sample/Signer	Total
Data for learning	A, B, C	20	60
Data for evaluation	D, E, F	10	30

(v) The color of the signer's clothes and the background wall is black to facilitate easy detection of the colored region of colored gloves.
(vi) The height in the field of view of the camera is set at a position that prevents the wrists of the signer being detected when he/she lowers his/her arm in order to make clear the beginning and the end of a sign language motion.

We are now trying to compose sign language motion data as a corpus for analysis. Therefore, these conditions will be included the next time we collect motion data.

3 Feature Elements

DP matching was used for the recognition of hand motions in the previous investigation. However, hand speed and hand position are also important features of sign language motions. The following features were extracted from the motion data for each sign.

(i) Shape of hand motion

The sequence of positions recorded as x and y coordinates indicates the shape of the hand motion. Since the size of the motion differs for each user, a normalization scheme for these data was applied, based on the normalization process represented by expression (1). Figure 5 shows the shape of a hand motion, i.e. the progressive positions of the center of the gravity of the wrist. The figure on the left is from the raw data before normalization, and the one on the right is the result of normalization. In addition, linear interpolation was applied when there was no color region detection due to occlusion or no color detection.

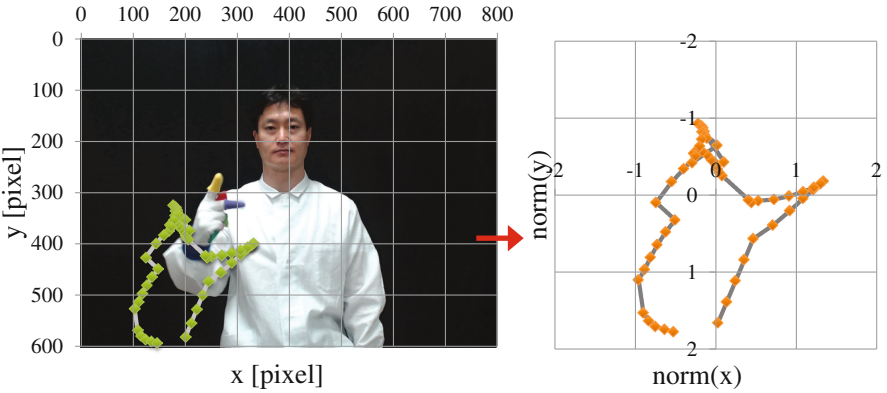

Fig. 5. Shape of hand motion before and after normalization

$$\left.\begin{aligned} norm(x_i) &= \frac{(x_i-\bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((x_i-\bar{x})^2+(y_i-\bar{y})^2\right)}} \\ norm(y_i) &= \frac{(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left((x_i-\bar{x})^2+(y_i-\bar{y})^2\right)}} \end{aligned}\right\} \quad (1)$$

Here,
i: i^{th} frame of motion data
n: the total number of frames of sign language motion data

(ii) Speed of hand motion

The speed of motion sometimes indicates meaning. The hand-position difference in each successive video frame can be regarded as a measure of the speed of motion, and

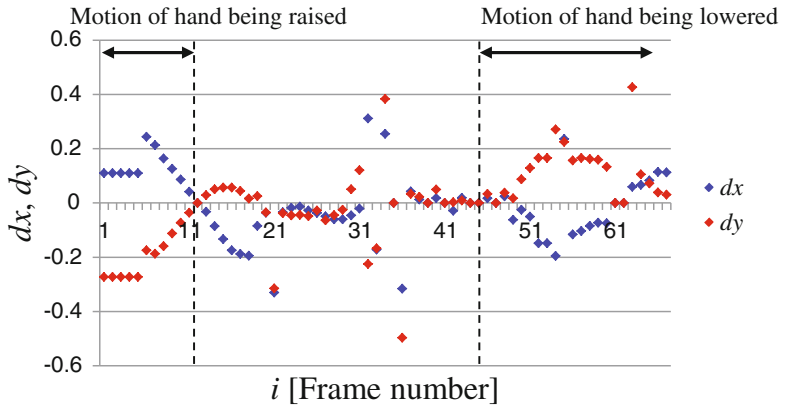

Fig. 6. Hand motion speed

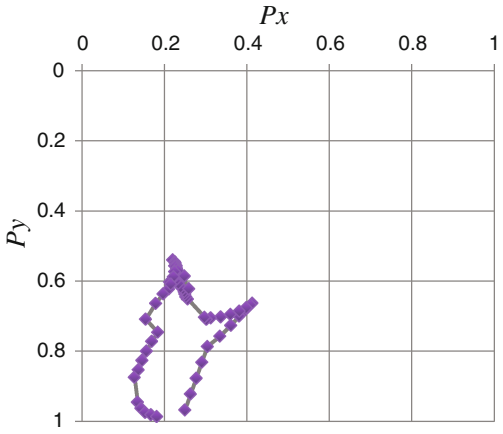

Fig. 7. Hand position

is used as such for this feature element. Hand speed is defined by the following expression. An example of hand motion speed is shown in Fig. 6.

$$\left.\begin{aligned} dx_i &= x_i - x_{i-1} \\ dy_i &= y_i - y_{i-1} \end{aligned}\right\} \tag{2}$$

Here,
i: i^{th} frame of motion data

(iii) Position of hand

The position of the hand includes meaning, so it should be included as a feature element, identified by its position in the image frame. Of course, the division by frame size (x:800, y:600) was used in the normalization process as represented by the following expression. An example of hand position is shown in Fig. 7.

$$\left.\begin{aligned} Px_i &= x_i / 800 \\ Py_i &= y_i / 600 \end{aligned}\right\} \tag{3}$$

Here,

i: i^{th} frame of motion data

(iv) Hand shape by distance

The shape of the hand itself, including finger positions and shapes, is also used as a feature of sign language words. The distance between the center of gravity of the wrist and the center of gravity of each finger-tip was used in the earlier investigation [6]. The distance is obtained by expression (4).

$$d_i = \sqrt{(fx_i - w_x)^2 + (fy_i - w_y)^2} \tag{4}$$

Here,
d: distance of center of gravity of wrist and each finger tip
i: Each finger
(1: Thumb, 2: Forefinger, 3: Middle finger, 4: Ring finger, 5: Little finger)
(f_x, f_y): Center of gravity of colored region of each finger tip
(w_x, w_y): Center of gravity of colored region of wrist (Fig. 8)

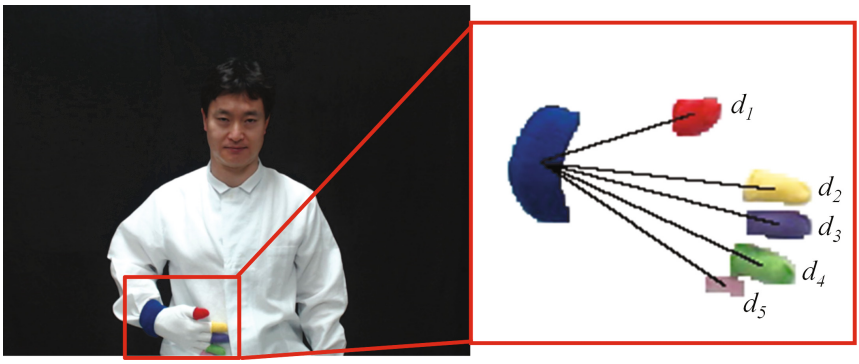

Fig. 8. Hand shape by distance

(v) Hand shape by number of pixels

When a finger is hidden, the distance represented by expression (4) is undefined. Nevertheless, the value 0 is set in method (iv). In this investigation, the number of visible pixels of each finger is also used as feature element for comparison. Of course, each value is divided by each maximum value as normalization. Either of (iv) or (v) is used in the recognition experiment.

4 Recognition Experiment Results

4.1 Recognition Experiment Result for Each Feature

All feature elements were used for the HMM learning process. The HMM learning and likelihood calculations were carried out with HTK. The dimensions of the state and initial values for the HMM, and the number of experimenters and samples used for HMM learning were determined in the pre-investigation. The data set shown in Table 2 was used for HMM learning and the recognition performance evaluation experiment.

The recognition success percentage for each feature element is shown in Table 3. The dimensions of each feature element and the number of HMM states are also shown in this table. The success ratios of the first through third ranks are shown in this table. The hand shape, i.e. the feature based on the number of pixels in each finger-tip image brought the best result.

4.2 Recognition Experiment Using Combined Feature

This experiment was carried out using HTK (Hidden Markov model toolkit) [9]. Table 3 shows the recognition success percentage as well as the learning value of each feature. It is expected that recognition performance can be raised by combining features. Two methods were used to examine the potential for improving the recognition performance. One involved combining three features (shape, speed and position) for HMM learning, and the other was to sum the rankings of the recognition results obtained with each feature. The ranking was in the order of likelihood, that is, 1, 2, 3 ···, 25, where 25 was the total number of words that might be recognized. Shape, speed and position are the features of hand motions. Therefore, these features were combined and the HMM was trained by using these features. Of course, both methods are considered, and the hand shape result and their ranking were added to determine the total ranking, i.e. recognition result. The results are shown in Tables 4 and 5.

Table 3. Recognition success ratio by each feature element

Feature value	Feature element dimension	Number of states	Recognition success ratio	
			1st	lst-3rd
Motion shape	2	4	21.7%	41.2%
Motion speed	2	4	28.5%	51.9%
Motion position	*2*	4	34.0%	69.6%
Hand shape (Distance)	*5*	3	39.7%	70.4%
Hand shape (Number of pixels)	*5*	3	42.7%	68.1%

Table 4. Recognition results by combining feature elements

Feature element			Success ratio	
			1st	lst-3rd
A	{Shape, Speed, Position}	Hand shape (Distance)	61.2%	84.1%
B	{Shape, Speed, Position}	Hand shape (Number of pixels)	55.2%	80.3%

Table 5. Recognition results by sum of ranking of each result

Feature element					Success ratio	
					1st	lst-3rd
A	Shape	Speed	Position	I land shape (distance)	61.9%	82.1%
B	Shape	Speed	Position	Hand shape (Number of pixels)	58.8%	81.1%

Similar results were obtained by both methods. The success percentage rose to around 60%, and increased to over 80% when 3rd rankings were included. It seems possible to use this method as a review tool for confirming the correctness of a learner's sign language motions.

5 Enhancement of Recognition Performance

The recognition results shown in Tables 4 and 5 are not sufficient for practical use. This section shows the methods we used to enhance recognition performance.

5.1 Selection of Candidates

As the number of words to be recognized increases, the more difficult it becomes to maintain recognition performance. The basic idea is that some of the recognition candidates should be eliminated before the HMM recognition process by using a number of relevant criteria. DP matching is used as the criterion in this investigation. This method has been widely used in voice and motion recognition [10]. There were experiences in the past investigation [5] where recognition using DP matching includes correct result when we selected the half of the recognition target words. This means that although the performance of DP matching cannot necessarily achieve the high recognition results obtainable using the HMM schemes, the possibility that correct results will be achieved is high if we take multiple candidates. Therefore, the authors used the results of DP matching to select the candidates of the recognition results before applying the HMM recognition scheme. DP matching calculates the distance of two vector elements, which is the motion data shown in Fig. 9. This is obtained from the movement of the center of gravity of the colored region of the wrist. This is an example of the motion of representing the word of "Diabetes". These motion data in a time sequence are used for DP matching. The more similar two motions are the smaller the distance for two motions by DP matching.

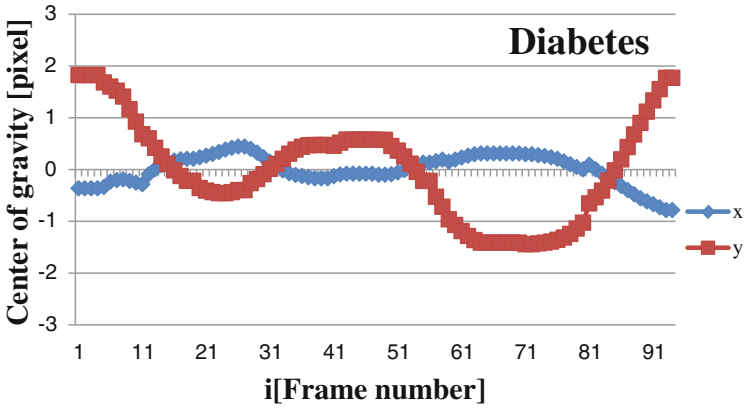

Fig. 9. Example of motion data for DP matching

The proposed recognition process sequence is shown in Fig. 10. Before the initiating the recognition process using HMM, the threshold values are decided for each word to be recognized as a pre-process, and these values are used for selection of

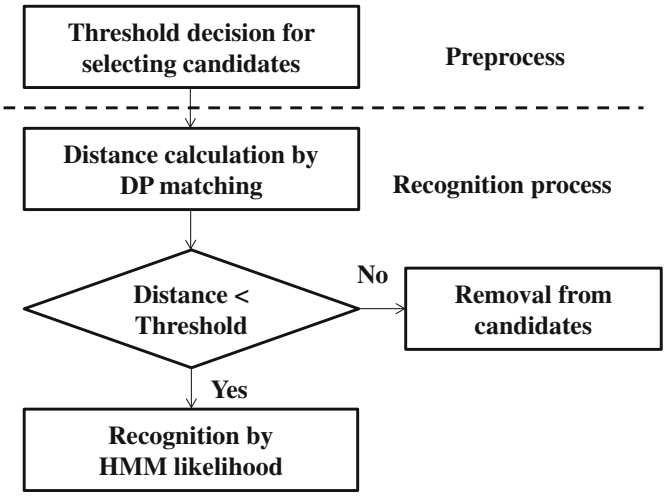

Fig. 10. Performance enhancement by introduction of threshold

correct candidates by sigh language recognition using HMM. The authors add a data set for this purpose (Table 6) as data for the threshold. Since the similarity of two motions is obtained as the distance of two motions, the distance of each word was used as the criterion for motion data to be recognized. We obtained 1800 (60×30) pieces of distance data from 60 samples for learning and 30 samples for deciding the threshold. The threshold values were determined from these distance values. If the distance of data for learning and the data for evaluation exceeds this threshold, these words are removed from the candidate of correct recognition results because there is no possibility that they will qualify as recognition results due to distance. Since the number of candidates for the recognition target can be reduced by this scheme, this leads to enhancement of the recognition performance using HMM method described in Sects. 3 and 4.

5.2 Results and Evaluation

The results obtained by the proposed method are shown in Table 7. Two cases, that is, combining feature ({Shape, Speed, Position} and hand shape by distance) and each feature (shape, speed, position and hand shape by number of pixels) are used for evaluation. The results without a threshold and some with thresholds are investigated for comparison in this experiment. The threshold values are determined from the average value and the standard deviation of distance values shown in this table. It is verified that the recognition performance can be enhanced, from 61.2% to 63.7% (from 81.6% to 85.6%) and 55.8% to 61.3% (from 81.1% to 84.5%), by setting appropriate threshold values in the experiment.

We confirmed the validity of this removal method. While the correct candidates must be retained, any word that has no possibility of a correct answer should be

Table 6. Dataset for experiments for proposed method

	Signers	Sample/Signer	Total
Data for learning	A, B, C	20	60
Data for evaluation	D, E, F	10	30
Data for threshold	G, H, I	10	30

Table 7. Recognition results by threshold

Threshold	Recognition success rate			
	Combining feature		Each feature	
	1st	1st-3rd	1st	1st-3rd
None	61.2%	81.6%	58.8%	81.1%
$\mu - \sigma$	55.9%	68.7%	53.2%	68.4%
μ	63.7%	85.6%	61.3%	84.5%
$\mu + \sigma$	63.5%	85.1%	60.8%	83.6%

Table 8. Evaluation of eliminated candidates

Number of samples (incorrect)			Number of samples (correct)		
Total	Removed	Percent	Total	Removed	Percent
18,000	8,733	48.5%	750	17	2.3%

removed by the proposed method. Among 25 words to be recognized, the number of correct samples is 750(= 1 × 25 × 30), and the number to be eliminated is 18,000 (= 24 × 25 × 30). We checked the eliminated samples using the proposed methods in each feature case. The results are shown in Table 8. The table on the left shows the removed candidates from incorrect candidates and the table on the right shows the results from correct samples. Almost half of the incorrect candidates and about 2% of the correct candidates can be eliminated. This shows the validity of the proposed method.

6 Conclusion

HMM learning and likelihood techniques were used to recognize sign language motions. The authors composed a sign language motion data set for model learning and performance evaluation together with supervision from a human signer. Each feature element that could be used for recognition was investigated and extracted from the motion data generated from video recordings of sign language. The percentage success of the two proposed methods was around 60% based on the 1st rank rating, and over 80% based on the 1st through 3rd ranks for 25 words. In addition, the selection of

candidates of correct answer was introduced to enhance the recognition percentage by using the DP matching results. The success percentage was increased from 61.2% to 63.7% (from 81.6% to 85.6%) and 55.8% to 61.3% (from 81.1% to 84.5%) by setting an appropriate threshold value.

It seems quite possible to use the proposed method as a learners' reviewing tool. However, it is necessary to enhance the performance by adding features, for example, hand direction, the visibility of the palm and other features. It is considered that the final ranking should be arrived at by considering the sum of the reliabilities, i.e. a weighted summation of the recognition success percentage for each word. Rejection criteria for a recognition result should be introduced to enhance the reliability of recognition methods. These investigations will be undertaken in future studies.

References

1. Baatar, B., Tanaka, J.: Comparing sensor based and vision based techniques for dynamic gesture recognition. In: The 10th Asia Pacific Conference on Computer Human Interaction (APCHI), Poster 2P-21 (2012)
2. Matsuda, Y., Sakuma, I., Jimbo, Y., Kobayashi, E., Arafune, T., Isomura, T.: Development of finger braille recognition system. J. Biometrical Sci. Eng. **5**(1), 54–65 (2010)
3. Humphries, T., Padden, C., O'Rourke, T.: Basic Course in American Sign Language. T. J. Pub., Inc., Silver Spring (1994)
4. Murakami, K., Taguchi, H.: Gesture recognition using recurrent natural networks. In: CHI 1991 Conference Proceedings, pp. 237–242 (1991)
5. Sugaya, T., Tsuchiya, H., Iwasawa, H., Nishimura, H., Tanaka, H.: Fundamental study on sign language recognition using color detection with an optical camera. In: International Conference on Imaging and Printing Technologies(ICIPT), Bangkok, Thailand, pp. 8–13 (2014)
6. Shibata, H., Nishimura, H., Tanaka, H.: Basic investigation for improvement of sign language recognition using classification scheme. In: Yamamoto, S. (ed.) HIMI 2016. LNCS, vol. 9734, pp. 563–574. Springer, Cham (2016). doi:10.1007/978-3-319-40349-6_55
7. KCC Corporation, Smart Deaf. http://www.smartdeaf.com/
8. HTK version 3.4.1. http://htk.eng.cam.ac.uk/
9. Sakoe, H.: Two-level DP-matching algorithm – a dynamic programming based pattern matching algorithm for continuous speech recognition. IEEE Trans. Acoust. Speech Sig. Process. **27**(6), 588–595 (1979)
10. Tanaka, H., Kimura, R., Ioroi, S.: Equipment operation by motion recognition with wearable wireless acceleration sensor. In: Next Generation Mobile and Services and Technologies, Cardiff, Wales, United Kingdom, pp. 114–118 (2008)