Basic Investigation for Sign Language Sentence Interpretation Using Acceleration Sensor Information

Hiroshi Tanaka
Dept. of Information and Computer Sciences
Kanagawa Institute of Technology
Atsugi, Kanagawa, Japan
email: h_tanaka@ic.kanagawa-it.ac.jp

Yoshimori Umeda
Course of Information and Computer Sciences
Graduate School of Kanagawa Institute of Technology
Atsugi, Kanagawa, Japan
email: s2385004@cco.kanagawa-it.ac.jp

Yuusuke Kawakita
Dept. of Information and Computer Sciences
Kanagawa Institute of Technology
Atsugi, Kanagawa, Japan
email: kwkt@ic.kanagawa-it.ac.jp

Hiromitsu Nishimura
Dept. of Information Media
Kanagawa Institute of Technology
Atsugi, Kanagawa, Japan
email: nisimura@ic.kanagawa-it.ac.jp

Jin Mitsugi
SFC Laboratory
Keio University
Fujisawa, Kanagawa, Japan
email: mitsugi@keio.jp

***Abstract*— This paper presents a method for segmenting a sign language sentence consisting of multiple words into individual word motions, which is an elemental technique for achieving the final goal of interpreting sign language sentences. We propose a segmentation method based on the similarity of motions, focusing on the fact that the word motion is included in the sentence motion. We selected 22 frequently occurring sign words and created 5 short sentences using them and acquired word and short sentence motion data. The results of the segmentation method using these data are presented. In addition, we show the results of word classification and confirm the feasibility of the proposed method for sentence interpretation.**

Keywords- Sign language; Acceleration sensor; Segmentation; LSTM; SVM; Motion classification.

I. INTRODUCTION

The hearing of more than 430 million people worldwide is impaired [1]. Traditionally, communication with normal-hearing people using written or text input tools has been the norm, but with the use of automatic transcription through speech recognition [2], the barriers to communication are gradually being lowered. The use of automatic interpretation has been increasing for verbal communication between different languages, and automatic interpreters have already been commercialized in Japan [3]. If automatic interpretation for sign language, which is considered to be an extension of these technologies, comes into practical use, the communication barrier between people with hearing disabilities and people with normal hearing will be eliminated.

Research on sign language interpretation has so far focused on word-level recognition of sign language motions. A camera, an accelerometer, and a data glove with a built-in strain gauge have been proposed as sensors for detecting sign language motions [4][5]. In recent years, there have been studies using OpenPose, MediaPipe, and other applications that can extract skeletal nodal information from camera images [6], and multimodal use of multiple sensors for higher accuracy [7]. These have ensured a certain level of recognition accuracy in word count limitations.

Since sign language is composed of multiple words, similar to a normal conversational dialogue, research is currently developing away from a focus on recognition at the word level to a deeper recognition of entire sign language sentences [8]. In this context, there are initiatives to recognize signed sentences using Transformer and Conformer technologies, which have been increasingly used in the field of natural language processing and speech recognition in recent years [9][10]. These are approaches that recognize sign language sentences without splitting them into the individual words that make up the sentence. However, learning data for sign language sentences is required, and a huge amount of learning data is needed to make these approaches practically workable. Acquisition of sign language motion data also imposes a significant burden.

In contrast, we propose a method to interpret sign language sentences from word motion data, considering the situation where a database of sign word motions is provided to make the proposed method feasible [11][12]. While some papers have proposed a method for determining the segmentation point of each word in a signed sentence by the motion speed, etc. [13], we focus on the similarity between the motions of each word in sentences and the hand motions of individual words and propose a method for interpreting by segmenting the sentence into words. The proposed method is

ISBN: 978-1-68558-185-5

made feasible by limiting the target domain for sign language interpretation, meaning that the number of words required can be reduced, and existing sign language word data can be used.

In Section 2, we present the final sign language interpretation sequence and the research target of this study; in Section 3, we present the method used to segment a sign language sentence into its component words and the results of our experiments; in Section 4, we present the classification results of the segmented motion data and its evaluation; and in Section 5, we discuss the results of this paper and future work.

II. Sequence of Interpretation and Investigation Target

A. *Sequence of sign language interpretation*

The sequence for interpretation of sign language sentences is shown in Figure 1. We focus on the fact that the motions of each word that makes up a sentence are contained in the sentence. We divide the sentence into words using a segmentation process, which will be described in Section 3. We already have the sign words' motion data and classify the segmented words using a learning model for the sign word motions.

This paper clarifies this segmentation method and attempts to classify each word based on segmentation results. The research target is the region shown by the blue rectangle in Figure 1. Modification of the classification results and composition into sentences based on a sign language linguistic model remains a future work.

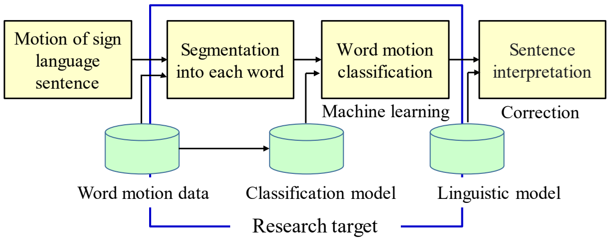

Figure 1. Sequence of sign language sentence interpretation.

TABLE I. Target sign language words and short sentences

Sign language words					
1. new	2. system	3. create	4. human	5. animal	6. difference
7. driving	8. license	9. update	10. family	11. put	12. work
13. prioritize	14. ordinary	15. people	16. familiar	17. shop	18. basic
19. power	20. public	21. election	22. law		

Short sign language sentences	
1. new/system/create (Create a new system.)	2. human/animal/difference (Humans and animals are different.)
3. driving/license/new/update (Update a driving license.)	4. family/put/work/prioritize (Prioritize work over family.)
5. ordinary/people/familiar/shop (A shop is familiar to ordinary people.)	

B. *Target sign language motions*

From approximately 10,000 sign words in the "New Japanese-Sign Language Dictionary" [14], we selected 22 sign words from those with the highest number of references in the dictionary. Short sentences combining these words were created by a sign language instructor, and these sentences were used for segmentation and word classification. Table I shows the 22 sign words and 5 short sentences that are combinations of these words.

C. *Acquisition of sign language motion data*

It is not necessary to attach a sensor to the person signing when using a camera, which may be advantageous from the standpoint of real use. There are applications that output body node information from camera images, such as OpenPose and Media Pipe, but they are limited to detecting motion in a plane. An acceleration sensor can measure the motions of sign language in 3D and has a higher sampling rate than a normal video acquisition camera, with each of the methods having their own advantages and disadvantages. We acquired sign language motion data using both an acceleration sensor and a camera. Figure 2 shows the data acquisition setup.

As the purpose of this investigation is to confirm the possible practical application of the proposed method, detailed finger motions were excluded. In order to acquire 3D motions including depth movements, an acceleration sensor (model: Analog Devices, ADXL362) was used to acquire motion data. The sampling rate was 10 ms, with a maximum measurement of 8 G. The sensors were attached at four locations on the elbows and wrists of both hands; sensor data from the four locations were synchronized for data reception using a backscatter communication system [15].

Acceleration data were acquired for the motions of individual words and sentences. Each word was acquired by repeating the sign language motion from a starting position in which the signer was standing still with both hands down by the sides of the body. The beginning and end of the short sentences were the same as for word acquisition. The data set used in this study is shown in Table II: 15 samples were acquired per word, for a total of 330 samples, and 5 different of sentences, 3 samples per sentence, for a total of 15 samples per short sentence. Since the purpose of this study was to verify the feasibility of the method, it was decided to use data from one signer (the aforementioned sign language instructor),

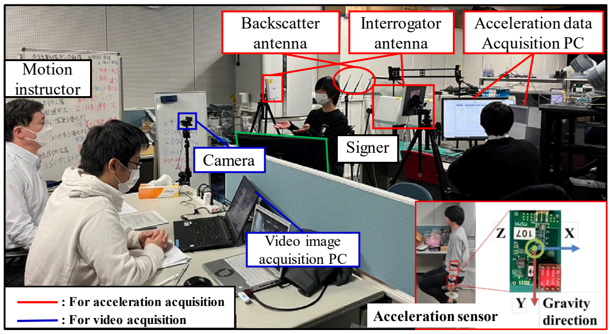

Figure 2. Data acquisition configuration. [16]

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

TABLE II. TARGET SIGN LANGUAGE WORDS AND SHORT SENTENCES

	No. of signers	No. of motions	No. of samples / motion	No. of total samples
Words	1	22	15	330
Sentences	1	5	3	15

whose signing motions are correct and stable. This was done to minimize the influence of differences in the motions of individual signers.

III. SEGMENTATION METHOD AND RESULTS

In this section, we discuss the concept underlying the sentence segmentation method and the results of our experiments.

A. *Fundamental concept*

Figure 3 shows an example of acceleration data during the motions for a word and a sign language sentence (data in the x-axis direction for the left wrist). This is for the word “driving” and the sentence “driving/license/new/update” which means "I renew a driver's license". From this graph, we can see that parts of the word motion data are similar to what is seen in the sentence motion data.

Based on this concept, the segmentation method involves the extraction of similar portions of word motions from the sentence motion data. In a sentence, the interval between word actions includes a transition section, which is neither of the two motions (shown in Figure 4). This transition section is considered to be shorter than the duration of the sign word motion and was assumed to be included as part of the sign word in this investigation.

There are two methods for detecting similar parts of actions: the Dynamic Time Warping (DTW) method [17] and the method based on likelihood information using a Long Short Term Memory (LSTM) model [18]. When using an LSTM model, it is possible to quickly evaluate similar parts taking into account all word data by creating the LSTM model in advance, so in this study, we decided to use an LSTM model that had been trained by word motions.

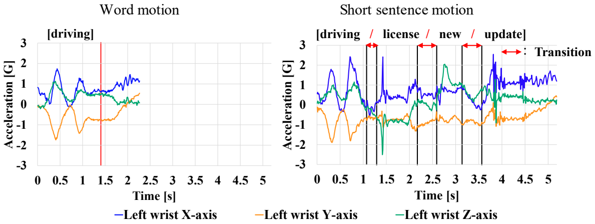

Figure 3. Example of acceleration data.

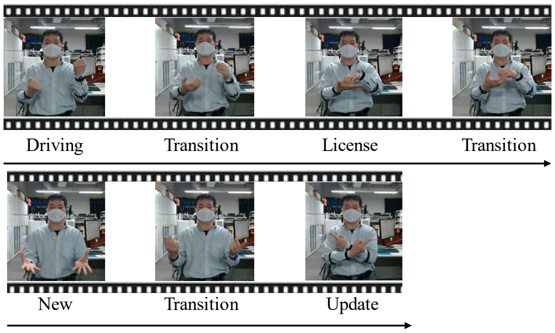

Figure 4. Sentence motions and transition section.

Learning parameters	
Hidden unit	500
Optimizer	adam
Maximum epoch	100
Batch size	128
Learning rate	0.001

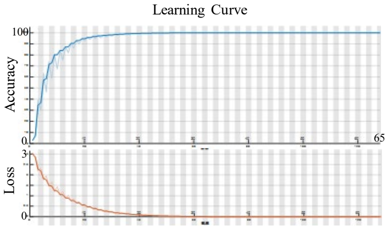

Figure 5. Parameters and learning curves for LSTM model creation.

B. *Word learning model and likelihood output*

To classify the 22 words, the LSTM model was created using 15 samples of each word, with motion data acquired for a total of 330 words as described above. The parameters used to create the model and the learning curve are shown in Figure 5. The convergence of the curve indicates that sufficient learning was achieved.

Figure 6 shows the likelihood output from the LSTM model when the short sentence “driving/license/new/update” is input. Here, the likelihood is the values obtained from the softmax layer of the model (a probability value ranging from 0 to 1, the sum of the elements (22 in this model) is 1). It was confirmed that the likelihoods of the words driving, license, new, and update, which make up the short sentence, are output with the highest values, and that the order of these words is also output correctly in this example.

As another example, Figure 7 shows the results when "family/put/work/prioritize" is input to the LSTM model. Unlike Figure 6, the likelihood of each word is not stable and its variation is large. In the case of Figure 6, it is considered relatively easy to divide the words that make up the sentence, but in the case of Figure 7, division into individual words is difficult and some sort of division criteria must be set.

The following two conditions were imposed for the segmentation of each word using this likelihood information. Here, we also added the condition that the segmentation time should not be less than 0.3 seconds, considering the minimum time required to make a sign language word motion.

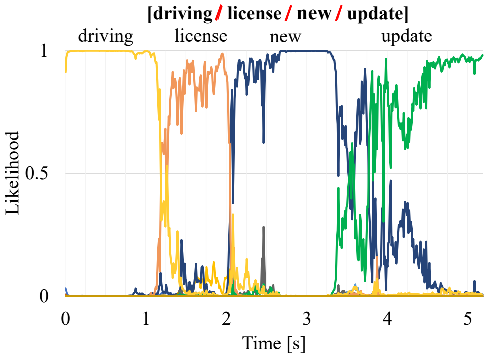

Figure 6. Example of likelihood (Case 1).

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-185-5

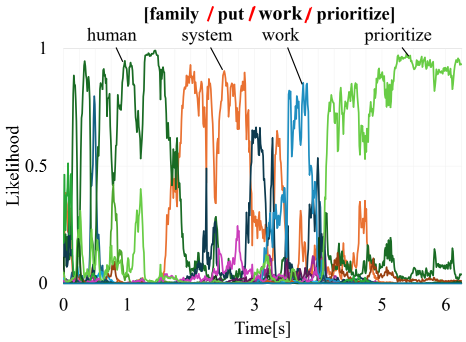

Figure 7. Example of likelihood (Case 2).

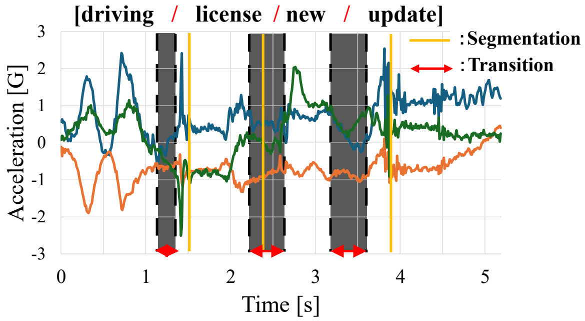

Figure 8. Example of a short sentence segmentation result.

(a) Selection of the word that has the maximum value of the 22-word likelihood integration value

(b) Saturation of the integrated value of the likelihood of that word

C. *Segmentation results*

An example of the segmentation results from the likelihood output in Figure 6 is shown in Figure 8. The number of words composing the sentence is 4, and the number of segments is 4. The position of the segmentation as well as the number of segments is important, and in this study, the segmentation results were evaluated from the following perspective. The segmentation position was evaluated based on whether it was in the transition section or not. Since a larger number of segments generates a higher number of segmentation positions included in the transition section, the segmentation position index was calculated according to the number of segmentation positions in the transition section and the number of segments. Here, the transition section was determined visually by a person familiar with the word motions.

(a) Number of segments

(b) Segment position index: Number of segmentation positions in the transition section / number of segments

The result for 5 short sentences, 3 sentences each, for a total of 15 sentences is shown in Table III. The number of segments tends to be larger than the actual number of constituent words. This indicates that the risk of missing a constituent word is small, and from the perspective of sign language sentence interpretation, it tends to be better than under-division. The segment position index, which indicates the probability that a segment position falls within the transition section, was 0.31 on average. This leads to a decrease in word classification accuracy, and more correct segmentation remains an issue.

IV. Word Classification Using Segmentation Results

The segmentation results were used to classify word motions in that segment. We compared the classification accuracy between LSTM and Support Vector Machine (SVM) in a preliminary study and found that SVM performed better with the current number of data for training. Therefore, we used SVM, which has proven performance as an accurate classifier, taking into account the small number of samples for learning. Based on the results of our previous studies [19], the acceleration data for each segmentation section was divided into 10 parts, and the mean value and standard deviation in this region were used as the feature values. Then, a normalization parameter of 10 as the SVM parameter and RBF as the kernel were set for the classification model by SVM. Table IV shows the results of each word classification for each of the 4 segments in Figure 8. The top three classification results are shown in this table. Here, the number of words to be classified is 22.

TABLE III. Evaluation Result for Segmentation

Short sentences	No. of words	No. of Segment			Segment position index		
1. new/system/create	3	4	3	3	0.25	0.33	0.67
2. human/animal/difference	3	3	3	4	0.00	0.33	0.25
3. driving/license/new/update	4	4	4	4	0.25	0.50	0.50
4. family/put/work/prioritize	4	4	3	7	0.00	0.33	0.43
5. ordinary/people/familiar/shop	4	4	5	5	0.25	0.20	0.40

(three sample sentences for each short sentence)

TABLE IV. Word Classification Result Base on Segmentation

Sentence	1st seg.	2nd seg.	3rd seg.	4th seg.
driving/license/new/update	driving	familiar	new	system
	people	license	people	update
	law	law	driving	driving

TABLE V. Word Classification Results for Segmented Sections

Short sentences	First place only			Up to 3rd place		
1. new/system/create	0.50	0.33	0.66	0.50	0.66	1.00
2. human/animal/difference	0.33	0.66	0.25	0.33	0.66	0.75
3. driving/license/new/update	0.50	0.75	0.75	1.00	1.00	1.00
4. family/put/work/prioritize	0.50	0.66	0.28	1.00	1.00	0.42
5. ordinary/people/familiar/shop	0.25	0.80	0.40	0.75	0.80	0.60

(three sample sentences for each short sentence)

Word classification of the segmented sections was performed using the segmentation results for five different sentences, a total of 15 sentences. The results are shown in

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-185-5

Table V. As a measure of the classification performance of words in a segmented section, the Evaluation Index (*EI*) was defined as expression (1). The order of words was not considered, and multiple occurrences of a correct word were counted as one.

$$EI = \alpha / \beta \qquad (1)$$

where α is the number of words correctly classified, and β is the number of segments.

The classification performance was evaluated for two cases by assuming that (a) only the first place out of 22 words was correct, and that the classification was correct, and if (b) it was included in the third place. Although there were cases in which words not included in the sentence were classified, basically the words that constituted the sentence were reliably classified. It was confirmed that the proposed segmentation method enables the classification of words that make up sentences, although this is partly because the number of words in our experience was 22.

V. Conclusion

This paper proposed a method for segmenting short sentences into their component words and evaluated its performance. The proposed method is based on the likelihood information obtained from LSTM models learned on word motions. Twenty-two words and 5 different of sentences consisting of those words were created, and sign language data were obtained to evaluate the proposed method. The word classification rate after segmentation was approximately 50% for the first-place criterion and approximately 76% when the third-place criterion was applied, confirming the feasibility of the method.

In future work, it will be necessary to examine methods for eliminating the effect of transition sections to achieve highly accurate word classification, create word learning models for more accurate segmentation, and evaluate the results using sign language data from a large number of sign language users. In parallel, we will collect linguistic information, such as word frequency and word order of a signed language restricted to a specific field, and investigate more accurate methods of interpretation that incorporate such information in the segmentation and word classification results.

Acknowledgment

This work includes the results of the research and development for expanding radio resources (JPJ000254) of the Ministry of Internal Affairs and Communications. It is also supported by "The Telecommunication Advancement Foundation", Japan.

References

[1] Deafness and hearing loss. Available from: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss (accessed June 06, 2024).

[2] A. Radford et al., "Robust Speech Recognition via Large-Scale Weak Supervision," Proceedings of the 40th Int. Conf. on Machine Learning, PMLR 202:28492-28518, pp. 1-27, 2023.

[3] PokeTalk. Available from: https://pocketalk.jp/(accessed June 06, 2024).

[4] M. J. Cheok, M. Jin, Z. Omar, and M. H. Jaward, "A review of hand gesture and sign language recognition techniques," Int. Journal of Machine Learning and Cybernetics 10 pp. 131-153, 2019.

[5] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues," IEEE Access 9, 126917-126951, 2021.

[6] S. Qiao, Y. Wang, and J. Li, "Real-time human gesture grading based on OpenPose," 10th Int. Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1-6, 2017.

[7] Q. Miao et al., "Multimodal gesture recognition based on the resc3d network," In Proceedings of the IEEE int. conf. on computer vision workshops, pp. 3047-3055. 2017.

[8] S. B. Abdullahi, S. Bala, and K. Chamnongthai, "American sign language words recognition using spatio-temporal prosodic and angle features: A sequential learning approach," IEEE Access 10: 15911-15923, 2022.

[9] P. Villegas and L. Francisco, "Sign Language Segmentation Using a Transformer-based Approach.". Available from: https://hdl.handle.net/20.500.14468/14662, 21 pages, 2022.

[10] J Huang et al., "Video-based sign language recognition without temporal segmentation." Proc. of the AAAI Conf. on Artificial Intelligence. Vol. 32. No. 1. pp. 2257-2264, 2018.

[11] SmartDeaf -- Video dictionary for learners of sign language. Available from: https://www.smartdeaf.com/.(accessed June 06, 2024)(in Japanese).

[12] Sign language shower. Available from: https://www.nhk.or.jp/school/tokkatsu/syuwashower/.(accessed June 06, 2024)(in Japanese).

[13] K. Renz, N. C. Stache, S. Albanie, and G. Varol, "Sign Language Segmentation with Temporal Convolutional Networks," IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 2135-2139, 2021.

[14] A. Yonekawa, "New Japanese Sign Language Dictionary," Japan Federation of the Deaf and Dumb, 2011.

[15] J. Mitsugi, Y. Kawakita, K. Egawa, and H. Ichikawa, "Perfectly Synchronized Streaming from Multiple Digitally Modulated Backscatter Sensor Tags, " IEEE J.RFID, vol. 3, no. 3, pp. 149-156, 2019.

[16] Y. Umeda et al., "Proposal and Evaluation of Sign Language Sentence Segmentation Method based on Acceleration Information," RISP Int. Workshop on Nonlinear Circuits, Communications and Signal Processing, pp. 360-363, 2024.

[17] W. Li, Z. Luo, and X. Xi, "Movement trajectory recognition of sign language based on optimized dynamic time warping," Electronics 9, no. 9, 1400, 15 pages, 2020.

[18] Sequence-to-Sequence Classification Using Deep Learning. Available from: https://www.mathworks.com/help/deeplearning/ug/sequence-to-sequence-classification-using-deep-learning.html (accessed June 06, 2024).

[19] T.Wakao et al., "Application of multimodal methods to sign language motion classification and its effectiveness evaluation," RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, pp. 269-272, 2023.