AI RECOGNITION OF JAPANESE SIGN LANGUAGE AND ITS APPLICATION

Tsutomu Kimura[1]*, Teppei Miura[1], Kazuyuki Kanda[2]

[1]National Institute of Technology, Toyota College, 2-1-1 Eisei-cho, Toyota City, Aichi, 471-8525, Japan

[2]Testing Organization for Sign Language Proficiency,6-13 Nihonbashi Kobunacho, Chuoku, Tokyo, 103-0024, Japan

Abstract

We are researching Machine Recognition of Japanese Sign Language using deep learning technology. JSL consists of the main parameters i.e., handshape, location, and movement of hands, otherwise non-manual markers such as facial expressions, posture, and others. JSL has unique morphemes called CL=classifiers, and some signs using fingerspelling. Understanding JSL needs recognizing all these elements, which makes the research entangled. We limit the first target as a simple sign and a simple sentence.

In our research, we use Conformer (a combination of CNN and Transformer) and CTC, which are used generally in speech recognition. The system learned 170 types of signs with each 100 tokens. The correct answer rate (Accuracy) was about 95%.

For sign language sentences, unlike a sign word recognition, the problem is the transition between words which make the sign recognition harder. Our system recognized over 58% for the sign sentences with 96 types of signs.

Keywords: *sign language, AI recognition, sign language engineering, sign language dictionary*

1. INTRODUCTION

In considering sign language, we would like to summarize the characteristics of the sign language, the basic knowledge of the morphology of spoken language, and the basic grammatical phenomena of Japanese sign language that are currently proposed.

1.1. Characteristics of Visual Language

In spoken language, the serial combination of sounds constitutes morphemes, words, and sentences, so the sequence order itself is a rule. Since sign language is a visual language, it is characterized by the fact that the constituents are arranged in space and often arranged at the same time. Moreover, since an ordinal array is also used, the composition becomes fused and complex. For example, the word "FATHER" is composed of three morphemes, "#KINSHIP" "#MAN" and "#ELDER" , but there is an order in which "#MAN" "#ELDER" are combined at the same time, and "#KINSHIP" always precedes this combination. When simultaneous joins are represented by / and ordered joins are represented by +, they can be expressed as follows.

<Father>=<#KINSHHIP>+<#MAN>/<#ELDER>*capitalized letters in < > show sign language

In spoken language, there are elements that span larger units, such as prosody, but in sign language, there are elements called NMM (non-manual markers) that function over words and phrases, such as facial expressions and postures. In addition, spatial perception shared by the speaker and the other party is used.

We consider that the grammar of sign language is often composed at the morphological level compared to spoken languages, and that there are fewer grammatical phenomena at the syntactic level than in spoken languages, but it should be noted in advance that this is not the consensus of all sign language researchers worldwide.

1.2. Word Order in Sign Language

Regarding the word order in sign language, the following four types are typical of the position of words in sentences or phrases.

1.2.1. Sentence Ending Interrogative

In Japanese Sign Language, interrogative words are arranged at the end of sentences. It is similar to the way interrogatives are placed at the beginning of sentences in English. In the so-called "Signed Japanese", which is a hybrid variant of Japanese and Japanese Sign Language, the order of Japanese may be followed instead of this rule.

<#YOU> / <#GO> + <WHERE> (accompanied by a questioning expression)

1.2.2. Modifier Word Order

In Japan Sign Language, the order is basically modified noun + modifier(NA). For example, Japanese "seven o'clock" is arranged to the order of <TIME>and <SEVEN>. A similar trend can be seen in adjectives and nouns, and it is difficult to distinguish between "the white flower" and "The flower is white." because they are both <FLOWER> and <WHITE>.

1.2.3. Special Syntax

When the interrogative word "WHAT" appears in middle of a sentence, it becomes a topicalization and it is equivalent to the Japanese copula construction "A is B". For example, "My father's job is a bank clerk" becomes <FATHER><JOB><WHAT><BANK> <WORKER>. This <WHAT>is distinguished from interrogative words that are accompanied by a facial expression at the end of the sentence, by not having those phenomena.

1.2.4. Sandwich Structure

In Japan Sign Language, there are often pointing at the beginning or middle of a sentence and at the end of a sentence. This phenomenon is called “a sandwich structure”. The person of the pointing often coincides with the other, but sometimes the movement of the pointing at the end of the sentence expresses emphasis, or the positional change expresses aspectual meaning.

1.2.5. Subject, Predicate and Object

In spoken languages, the sentence structure of SVO or SOV is controversial, but in sign language, the subject and object are internalized in the verb and are often expressed simultaneously, so the discussion of word order of SVO or SOV is meaningless.

1.2.5.1 Argument Structure

For example, the verb <MEET>is expressed by the motion of raising the index fingers of both hands and approaching mutually, but the grammatical person changes depending on the position. The index fingers of both hands are the meaning of a person, and the approaching action means the action of meeting. The index fingers of both hands are the arguments (of the predicate). Each hand can be made by two fingers, the index finger and the middle finger, to mean two people, or all five fingers can be opened to mean a large number of people. It is possible to change the number of person. Bringing open hands close together, the meaning of "MANY and MANY, MEET" is translated into Japanese as "gathering". In Japanese, "meet" and "gather" are different words, but in English it is the same “meet”, thus the semantic range of this sign language word is close to English.

1.2.5.2 Person of Argument

If you move your index finger forward from the position of your mouth, it becomes <SAY>but more precisely, it means "I say," and the subject is intrinsic. In general, verbs in sign language are consistent with the person by position or location as described below, but the direction of movement of the verb indicates a person change. In that case, the hand is a single argument. As for the verb type, there are two-type, one-type, and zero argument, and two-types are subject and object, one type is subject or

object, and zero type is intransitive. However, in the two-types, the distinction between which is the subject, or the object is not clear, as indicated by <MEET>.

1.2.5.3 Verb Inflection

The argument of the sign language verb changes the direction of movement according to the person that is localized in space. In addition, an adverbial meaning is added by moving manner.

1.2.5.3.1 Inflection by Person

In sign language, the person is localized by the line of eyesight. The signer is the first person, and it is in front of eyesight. The receiver is the second person, and it is located toward the eyesight. The person outside the line of eyesight is in the third person. This personal positions are shared by the signer and the receiver as a premise. So, for example, in the verb <MEET>, the right hand is placed in front of the signer's chest and the left hand is placed in the direction of the receiver, and then the hands are brought close together, resulting in a sentence that says, <I- MEET-YOU>. The signer intended <I- MEET-YOU> but the receiver understands it as <YOU-MEET-ME>. In either case, the actual meaning is the same. If you place your left hand outside of your eyesight line and then bring your hands closer, it becomes <I-MEET-HIM/HER>. At first glance, the change in the direction of the approach movement seems to be a personal change, but if the signer changes the way of his or her gaze to the left person and keeps the positions of <I- MEET-YOU>, it changes the person of the dialogue. That means by<I-MEET-HIM/HER> to the previous signer. In this case, for the signer, "you" is the current dialogue partner, and the previous dialogue partner is already in the third person. The signer has changed the receiver. In sign language, space reflects reality, so the gaze localizes the position of the person, and the direction of movement matches it. The sign verbs with arguments and moving direction behave in the same way. This type of verbs are classified as “directional verbs”. A sign <RAIN> has movement with no argument and does not inflect by person, therefor it is not classified as directional verb. In English, non-personal pronoun is presented as “It rains.”, but in sign language it is not presented.

1.2.5.3.2 Adverbial Inflection

Since a verb includes an adverbial meaning added to them by the way they are moved, this is called adverbial inflection. If you move the verb <WALK> slowly, it means <WALK-SLOWLY>, and if you move it quickly, it means <RUN>, and if you move it in a zigzag motion, it means <STAGGER>. The content of the adverbial meaning is diverse, and phase changes such as repetition and continuation are possible. <COMMUTE> can added the meaning of "going often" or "barhopping". <WALK> can be arranged by adverbial inflection, such as <GOING-UP-STAIRS>, <GOING-UP-SPIRAL-STAIRS>, <GOING-UP-SLOPE> etc., which is equivalent the objects in Japanese. Sign language utilizes a function that can be displayed visually by using space.

1.2.5.4. Number of Pronouns

In sign language, the hand-shape is equivalent to pronoun. Pointing behavior may have the function of a pronoun or clitic as some researchers think and may also be a personal pronoun in conjunction with a personal position in space. Instead of pointing, there is also a chin pointing (chin pointing) as a trace.

1.2.5.4.1. Singular, Dual and Plural

One index finger indicates singularity, index and middle fingers indicate duality, and five fingers indicate a group. The number can also be indicated by the index finger of either hand. In the case of two hands, it is used to indicate that the two are together, for example, <WE-TWO>, and in the case of one hand, it indicates that they are separate. For example, it is used as <WE-TOGETHER>. If each hand has <TWO> hands, it means <2-ON-2>. It's complicated, but it can also show <ONE-ON-ONE>. Since there are many five fingers, it is also possible to indicate <ONE-TO-MANY>.

1.2.5.4.2. Gender

The erect thumb represents <MAN> and the upright little finger represents <WOMAN>. <DOCTOR> is distinguish to a male and a female doctor. When representing the mayor or chairman, if the gender is unknown, it is expressed with a thumb due to "male priority rule", but if it is clear that it is a woman, it

is represented by a little finger. Gender indication occurs when this rule of noun formation makes the thumb and little finger function as pronouns indicating <HE> and <SHE> with thumb and little finger, respectively. If you don't want to show the gender, you can use your index finger or can change to pointing. However, if it is used as an argument such as <MEET>, it cannot be changed to a thumb or little finger.

1.2.5.4.3. Classifiers

A characteristic feature of sign language is a morpheme called CL. In sign language, classifiers are called CLs. Basically, it is a well-established theory that a specific hand shape forms the stem of the verb, but there is another theory that CL expands the interpretation to include movement and the body. The definition of a classifier in spoken language is ambiguous and is defined as "representing a size, shape, material, etc.", and in Japanese it is combined with a numeral to become "one-mai, two-hiki", etc. The discussion is omitted here.

A typical CL is to indicate "human behavior" when the index and middle fingers are downward, for example. Moving the hand becomes verbs such as <WALK>, <RUN>, <SLEEP>, or <SURPRISED>. In addition, a certain hand shape represents <AIRPLANE> or <AUTOMOBILE>. From the standpoint of interpreting CL in an expansive way, there is CL called semantic CLs that display geometric figures such as points, lines, and figures in the form of fingers and hands, and the gesture of handling some tool is also considered handle CLs. When describing an animal such as a dog, the body is likened to that animal, but some people call them body CL.

1.2.5.5. Syntactic Structure by NMM

Non-Manual Markers, as facial expressions and postures that perform a prosodic function like phonemes in spoken language. Changes in NMM can also have syntactic or adverbial functions.

1.2.5.5.1. Modals

Questioning, negation and imperative have their own unique facial expressions. Facial expressions may also be obligatorily accompanied to vocabulary. For example, <SAD> or <HAPPY> is expressed by a facial expression that correspond to each emotion. If you don't show the modal NMM in a sentence, you can't convey the meaning, but even if you don't show the NMM to a vocabulary, you will be felt uncomfortable.

1.2.5.5.2. Adverbial Function

The emphasis on <THIN> is accompanied by a peculiar NMS: squinting eyes, thinning lips and exhaling. When emphasizing <BIG>, it is accompanied by an NMS that puffs up the cheeks. It has been pointed out that there is an NMS associated with an aspectual change, but it has not been confirmed due to objections and it is under discussion.

1.2.5.6. Complex Sentence Structure

When connecting sentence and sentence, there are several signs, but the conjunction is not clear. Some studies have shown that the end of a sentence is displayed by an NMM called a nod, but it is thought that this function is similar to the descending intonation at the end of affirmative sentences in spoken language. There is a way to change the topic, such as the gesture of <BY-THE-WAY> but it is not clear whether it is a sentence or a conjunction in itself.

1.2.5.6.1. Relative Structure

If you specify a specific place in space, you can specify the relative structure by pointing or chin-pointing from the next discourse. The word <PLACE> is often used, but <HE> or <SHE>is often designated with the thumb or little finger. It has also been pointed out that there is an NMS in which the posture is fixed during the presentation of the relevant clauses.

1.2.5.6.2. Compound Sentence (Roll Shift)

It is an universal phenomenon in sign languages. In the scene of a dialogue, the posture changes from side to side, like the change of speakers. This is called a roll shift. The posture is often symmetrical, but

it can also show a positional hierarchical relationship, such as that of a parent and child. It has not been concluded in Japan how many people can be shifted.

1.2.5.6.3. Speech (Scale)

Some signs are shown in full size and others are shown in small scale. Volleyball, for example, is presented by the gesture of tossing the ball, which is full-size. However, <WALK> is a small scale shown by a person's legs with two fingers. In general, full-scale indicates that you are acting, and a small scale indicates that you are observing the situation. It is considered that the function of direct speech in English is shown by the full size, and the indirect speech is equivalent to the small scale. Since there are not all words with expressions that change the scale, the scenes are limited, but there are some deaf people who use full-scale expressions to create a more realistic scene, or to make full use of small-scale expressions to maintain objectivity.

1.2.5.7. Tense

Verb inflections and auxiliary verbs that present tenses are not found in sign language. The horizontal position of the side of the face is used, and the front indicates the future and the back indicates the past, so if you move the index finger forward next to the face, it will be <TOMORROW>, if you move backward, it will be <YESTERDAY>. When you use two fingers, and if you will move forward, it is <THE-DAY-AFTER-TOMORROW>. When you use <SEVEN>, it will be <NEXT-WEEK> and <LAST-WEEK>. If it is a flat hand, it will be <FUTURE> and <PAST>. When the vocabulary indicating these times is displayed as a sentence adverb, the context is formed as a tense, and it is considered that inflection of verb and auxiliary verb is unnecessary.

There are also examples of signs with the meaning of <FINISH> and the sign <END> used as if they were completion or past auxiliary verbs. However, the detailed distribution of whether it is a verb combination or an auxiliary verb is not known, so it is out of discussion here.

1.2.5.8. Word Formation by Body Parts and Directions

When a body part is used as a position, the implied connotational meaning of that part is represented. For example, the eye points to <EYE> . <MOUTH> draws a circle around the mouth with the index finger. <SPEAK> is done at the position of the mouth, <SEE> is done at the eyes, and <SMELL> is done at the nose.

In sign language, position is considered to be phonological constituent (strictly speaking, a parameter), so the phoneme that contains meaning may be contradictory in definition. However, since the meaning of a body part becomes meaningful only when it is combined with the certain shape of a hand, with a certain motion such as pointing, there is no contradiction if we consider the combination as a morpheme. Direction also has a metaphorical meaning. <HEAVEN> and <UP> are indicated by the upward direction. <RECEIVE> and <RAISE> also use the upward direction, and <GIVE> use the downward direction. <SMELL> also expresses the image of smell rising from the bottom up to the nose. It is not a metaphor when indicating the actual direction of the line of sight, such as <LOOK-UP> or <READ>.

1.2.5.9. Differentiation from Gestures

Sign language is a language, and although there is a distinction with gestures, it is difficult to distinguish them clearly, as many words originate from gestures and sometimes use NMM. We will examine the structure of syntax and word formation and distinguish between verbal and non-verbal phenomena in the future. However, there are many universal phenomena with spoken language at the base.

2. SIGN LANGUAGE ENGINEERING

According to the linguistic analysis of the above sign language, the sign language is different from the spoken language, and the syntax does not form a grammar by words or phrase, but the morphological combination mainly forms grammar. Therefore, Tsutomu Kimura Laboratory, National Institute of Technology (KOSEN), Toyota College, believes that analysis of sign language word (sign below) separately will analyze syntax at the same time. We extract signs by image processing, and we decided

to develop a sign dictionary and a machine translation system. The following is an outline of engineering and technical explanation.

2.1. Single Sign Recognition System Using Deep Learning

Kimura Laboratory has developed a system that learned signs using deep learning and created a learned model. It did not learn the signs of the video learn as they are, but it estimated the posture using the MediaPipe [2] developed by Google and learned the key points such as joints. We prepared 100 video data per a single sign type, and the system get coordinates of the key points for each frame. This was created as learning data for the basic 150 signs. This was applied to a combination of Transformer [3] used for natural language processing and Conformer [4] in which CNN used for image recognition, etc. and it generated a learned model. In this model, the correct answer rate (Accuracy) was about 95%.

2.2. Sign Dictionary System

We developed a sign language dictionary system by applying the sign language recognition model above. We are developing a system being realized online, operating in a web browser, and usable on a PC or smartphone equipped with a camera (Figure 1). A user expresses a sign language in front of the camera and send the video to the server. The server uses a sign language recognition engine to recognize it and presents the top 10 matching words. Clicking on the presented word plays a sample sign language video (Figure 2). We are currently making final adjustments and preparing for public releases.

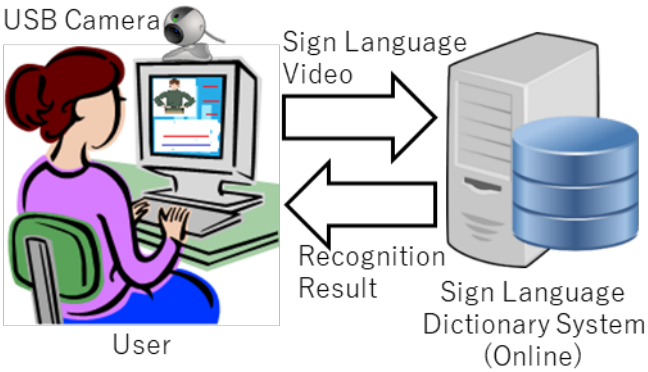

Figure 1. Online Dictionary System

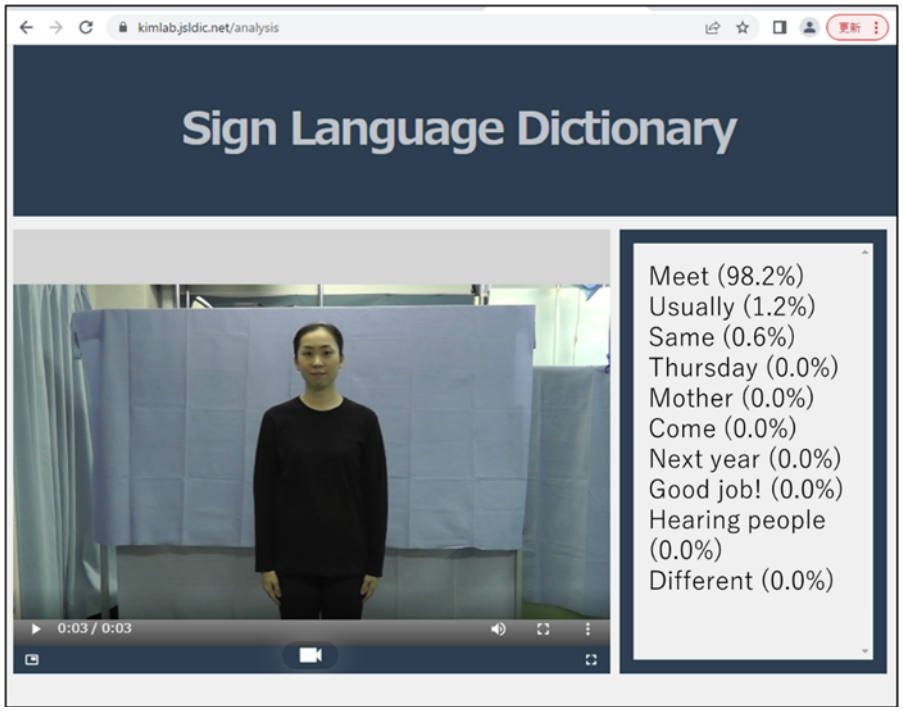

Figure 2. Sample Sign Video

2.3. Sign Language Sentence Recognition

Previous research has been word-by-word recognition as shown in Fig. 3. This makes it impossible to develop a sign language translation system because it is not possible to recognize sign language as a sentence. Therefore, we started from recognizing words used in the sentences. However, it is more difficult to recognize sign language sentences than single signs. This is because "transitions" between the signs which are unconscious movement occur. After completing the operation of one sign expression, an interpolation action is performed to perform the next sign action without returning to Home Position. (Fig. 4). Transitions are ambiguous meaningless movements, and the behavior fluctuates depending on the combination of signs before and after. The transitions have made recognition more difficult than recognition of single signs.

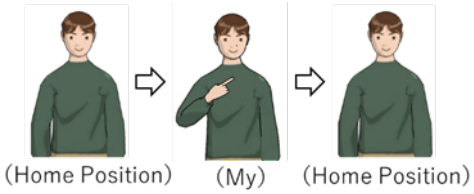

Figure 3. Word-by-word Recognition

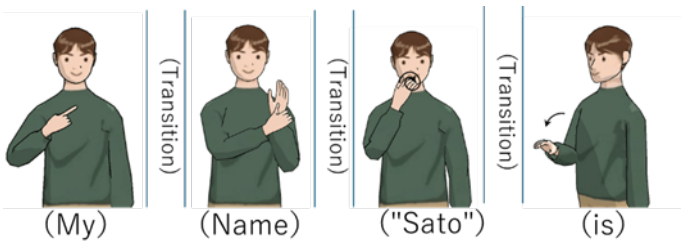

Figure 4. Real Sign Action

The process of sign language sentence recognition is similar to speech recognition, in which features are extracted from speech data and then learned. In this study, we implemented the Conformer method because we thought that Conformer, which combines a Transformer suitable for speech recognition and a CNN used for image recognition, would be highly compatible with sign language sentence recognition. However, the exact agreement rate (the percentage of signs and the order of occurrence) was as low as 1.3%. Therefore, CTC (Connectionist Temporal Classification), which is used in speech recognition, is used for model output.

Furthermore, for the learning of recognition models, more sign language data is needed. However, it is not easy to create a large amount of annotated sign language video data. Therefore, by introducing two-word sentences inspired by 2-gram in natural language processing, efficient transition learning and increased data volume are achieved. As a result of training with these methods, a correct answer rate (i.e., accuracy) was up to approximately 79%.

2.4. Sign Visualizing Tool

In the above-mentioned sign language sentence recognition, it is now possible to recognize not only signs but also the time at which they are expressed. We have developed a system that can be used to analyze sign language expressions by linking video and recognition results so that the recognition results can be dynamically understood. Figure 5 shows the flow and Figure 6 shows GUI of the system. It can, for example, analyze the motion of transition between the signs.

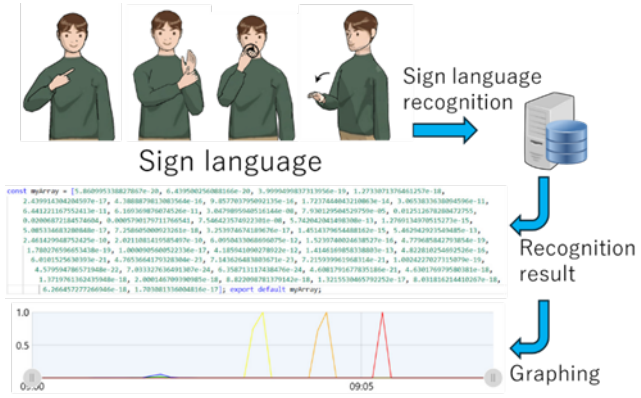

Figure 5. System Flow

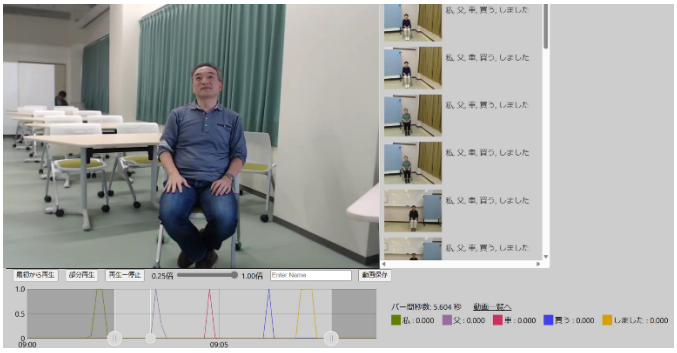

Figure 6. GUI of the System

The system uses a scroll bar to check which and when the sign is being represented in conjunction with the video.

2.5. Synthetic Sign Generation Tool

We have developed a synthetic sign language generation tool to address data shortages in the machine learning domain. This tool generates various types of synthetic sign data from KoSign, a large-scale JSL database [5]. Specifically, our tool obtains 3D motion capture data from KoSign and produces synthetic data including realistic images, 2D/3D joint keypoints, depth maps, body-part segmentations, optical flows, and surface normals (Fig. 7). The availability of diverse and large-scale synthetic data is expected to contribute advancements in sign language research.

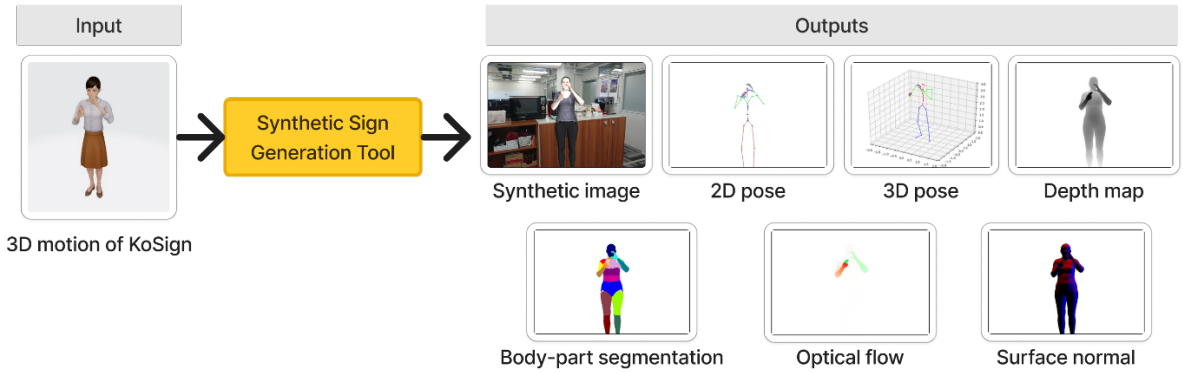

Figure 7. Overview of Synthetic Sign Generation Tool

3. DISCUSSION

As shown in Section 1.1 above, a feature of sign language is that the constituents are presented simultaneously in space. In spoken languages, the constituents are arranged consecutively along the time axis. In human recognition, phonemes are produced in order and received by the ear in order, but since the constituents of sign language are received three-dimensionally with the eyes, and when recognized by a machine, the characteristics are imitated. Although there is a difference in the structure of word formation in spoken languages that has a syntagmatic structure and in sign languages that has a paradigmatic structure, the basic semantic units are the words. Thus, we decided to start with word recognition.

The sign language recognition system has achieved a high rate of correct recognition of words, and has been applied to a sign language dictionary system. However, the vocabulary for word recognition from signed sentences is limited at this stage. This is due to the overwhelming lack of training data that has already been annotated. Annotation is a costly process. Therefore, research is being conducted using a method based on self-supervised learning, in which learning is conducted using unannotated data. The sign language visualization tool is based on this sign language sentence recognition system, and the number of vocabulary words that can be recognized should be increased for use in sign language studies.

4. CONCLUSIONS

In this paper, we explain the linguistic characteristics of Japanese Sign Language, and as part of sign linguistic engineering, development of a sign language recognition system and its application examples are described. Sign language dictionaries are currently being brushed up for publication. Lexical level recognition from sign language sentences with a small vocabulary has a high accuracy rate, and the number of recognizable vocabularies will be increased in the future. We believe that the sign language visualization tool will be a useful tool in sign language research, and we are considering making it available to researchers.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP22H00661, JP23K17511, JP23K21933.

REFERENCES

1. T. Kimura and K. Kanda, 2019, 'Sign Language Recognition through Machine Learning by a New Linguistic Framework', *Association for the Advancement of Assistive Technology in Europe 2019*, Proceedings, Bologna, Italia, S144-S145.
2. Google inc. 2021, *MediaPipe*, viewed 19 April, <https://developers.google.com/mediapipe>
3. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin 2017 'Attention Is All You Need', *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, pp5998-6008
4. Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang 2020, 'Conformer: Convolution-augmented Transformer for Speech Recognition', *INTERSPEECH 2020,* no.10. 21437, Shanghai, China, pp.5036-5040.
5. Y. Nagashima, 2021, 'Kogakuin University Japanese Sign Language Multi-Dimensional Database (KoSign)', *Informatics Research Data Repository, National Institute of Informatics, 2021.* DOI: 10.32130/rdata.5.1