Machine Translated by Google

こちらは英語原著論文の機械翻訳版です。
次ページが原著論文で、翻訳版と交互に展開されます。機械翻訳のため、誤字や誤訳、翻訳が未反映の部分が含まれている可能性があります。

Paper

Evaluation of Sign Language Recognition with Higher Resolution Face Images

Takeshi Kajiyama †, †† (member) and Yoshiaki Shishikui †† (member)

Abstract Sign language recognition technology is expected to be realized as a technology to support communication among deaf people whose first language is sign language. In addition to the hands and arms, facial expressions are also important for understanding sign language, and there are words and expressions that require the shape of the mouth, eyes, and eyebrows to be distinguished. In sign language recognition, the upper body image including all these body parts is analyzed, but in sign language recognition using deep learning, the computational and memory amounts are limited and it is not possible to handle high-resolution upper body images. In this paper, we analyze high-resolution images only for the mouth, eyes, and eyebrows, and aim to improve recognition performance while suppressing the increase in computational and memory amounts. Experiments have confirmed that by combining the analysis of a low-resolution upper body image as a reference and the analysis of a mouth image with double resolution, recognition of words that require the shape of the mouth to be distinguished was improved. On the other hand, no improvement was confirmed for the eyes and eyebrows. As a result of statistically evaluating the recognition performance for all words, it was confirmed that analyzing the mouth image with double resolution is effective in improving recognition performance.

Keywords:Sign language recognition, Image recognition, Facial expression recognition, Deep learning, Japanese Sign Language, Non manual signals

1. Introduction

There are approximately 70 million deaf people in the world whose first language is sign language1). For deaf people, it is important to have a means of communication with hearing people, but there is a shortage of sign language interpreters, which is the main means of communication2). To make up for the shortage of sign language interpreters, there are hopes for the realization of technology that converts speech and text into sign language and technology that recognizes sign language. In Japan, there are 1000 sign language interpreters that convert Japanese text into sign language. Research is underway into converting computer-generated sign language expressions into computer-generated sign language3). Research into the recognition of sign language , focusing on Japanese sign language4)-6), has also been reported.

As a technology for recognizing sign language, an approach that uses image analysis to read words and sentences in sign language is gaining attention. Sign language uses not only hands and arms but also facial expressions to express words and sentences7)-10). For example, the shape of the mouth is used to distinguish between words with different meanings7) that have the same finger shapes (hereafter referred to as "same-finger-opposition words"); the shape of the eyes and eyebrows can be used to indicate doubt or negation of the content expressed by the hands and fingers. They may add flavor8)9) or modify the degree of content10) In sign language recognition research, analysis of the fingers and arms has been the main focus, but there have also been examples of attempts to classify synonyms by analyzing the mouth11). Furthermore, in sign language, these functions do not depend solely on facial expressions, but arm movements can also perform similar functions12)13).

In recent years, the use of deep learning for sign language analysis from images has become mainstream14)-16).

The system takes an image as input and recognizes sign language words through an integrated analysis of the fingers, arms, and face contained in the image. However, in sign language recognition using deep learning, the amount of calculation and memory required to learn a time series of several hundred upper body images is a constraint, so a high-end GPU (Graphics Processing Unit) is required.

Even when using a 3D image processing unit, it is difficult to handle high-resolution images. Therefore, low-resolution images are used, which are created by reducing the resolution of the captured video. For this reason, it is possible that some parts of the body cannot be adequately identified due to insufficient resolution. In research on facial expression recognition, it has been shown that the resolution of facial images affects the recognition accuracy. There have been reports that this has been the case17). Therefore, we consider improving recognition accuracy while suppressing the increase in memory size by increasing the image resolution only for specific parts. In this paper, we consider the mouth, eyes, and eyebrows as candidates for specific parts, and perform analysis of these parts using higher resolution images of the upper body image.

In verifying the proposed method, we first verify the effect of increasing the resolution by two-class classification under limited conditions. That is, when high-resolution mouth images are used in combination, we verify the accuracy of distinguishing between synonyms of the same hand. In addition, when high-resolution eye and eyebrow images are used in combination, we verify the accuracy of distinguishing between the presence or absence of modifiers. Next, in multi-class classification with several thousand candidate words, we verify whether there is an improvement in the recognition accuracy of words for which two-class classification experiments were effective. Finally, we conduct recognition experiments on all words excluding those for which two-class classification was effective, and verify whether there is any adverse effect on the recognition of other words.

2023年5月31日受付，2023年8月25日再受付，2023年9月12日採録
† NHK財団
（〒157-0073 世田谷区砧1-10-11，TEL 03-5494-2400）
†† 明治大学
（〒164-8525 中野区中野4-21-1，TEL 03-5343-8040）

【機械翻訳コンテンツの著作権について】
当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
詳細は映像情報メディア学会著作権規定をご覧ください。

論　文

顔表情を解析する手話認識における画像解像度と認識精度の評価

Evaluation of Sign Language Recognition with Higher Resolution Face Images

正会員　梶山岳士[†,††]，正会員　鹿喰善明[††]
Takeshi Kajiyama[†,††] and Yoshiaki Shishikui[††]

あらまし　手話認識技術は，手話を母語とするろう者のコミュニケーションを支援する技術として実現が期待されている．手話の理解には手指と腕だけでなく顔の表情も重要であり，判別に口や目眉の形が必要な単語や表現が存在する．手話認識ではこれらすべての身体部位を含む上半身画像を解析対象とするが，深層学習を用いる手話認識では計算量・メモリー量が制約となり解像度の高い上半身画像を扱えない．本論文では，口および目眉に限り解像度の高い画像を解析し，計算量・メモリー量の増加を抑えながら認識性能の向上を図る．実験により，基準となる低解像度な上半身画像の解析と2倍解像度の口画像の解析を併用することで，判別に口の形が必要な単語について認識の改善が確認された．一方，目眉については改善効果が確認されなかった．すべての単語を対象とした統計的な認識性能を評価した結果，2倍解像度の口画像の解析が認識性能向上に有効であることが確認された．

キーワード：手話認識，画像認識，顔表情認識，深層学習，日本手話，非手指動作

1．ま　え　が　き

世界には手話を母語とするろう者が約7,000万人いる[1]．ろう者にとって，聴者とのコミュニケーション手段を確保することは重要であるが，主要な手段である手話通訳士の数は不足している[2]．手話通訳士の不足を補うため，音声やテキストを手話に変換する技術や，手話を認識する技術の実現が期待されている．国内では，日本語テキストをCGによる手話表現に変換する研究が進められている[3]．手話の認識についても日本の手話を対象とした研究[4)～6)]が報告されている．

手話を認識する技術として，画像解析により手話の単語や文を読み取るアプローチが注目されている．手話は手指と腕だけでなく，顔の表情も使って単語や文を表現する[7)～10)]．例えば，指の形が同じでありながら意味の異なる単語[7)]（以下，同手指異議語と呼ぶ）の判別に口の形を用いる，また，目と眉の形により手指で表現される内容に疑問や否定の意味を加える[8) 9)]，あるいは内容に対し程度を修飾することがある[10)]．手話認識研究では，主として手指と腕の解析が行われていたが，口の解析によって同手指異義語の分類を試みる例[11)]も見られた．なお，手話ではこれらの機能を顔の表情のみに依存するのではなく，腕の動きで類似の機能を果たすこともある[12) 13)]．

近年はこのような画像からの手話解析に深層学習を活用することが主流となっている[14)～16)]．深層学習では上半身画像を入力とし，その中に含まれる手指，腕，顔を統合的に分析することを通じて手話単語の認識を行っている．ところが，深層学習による手話認識では，数百枚からなる時系列の上半身画像の学習に要する計算量とメモリー量が制約となるため，ハイエンドなGPU（Graphics Processing Unit）を用いた場合も高解像度な画像を扱うことが難しい．そのため，撮影された映像から解像度を落として作った低解像度画像を使用している．このため部位によっては解像度不足で充分な判別ができない可能性もある．

顔表情認識の研究で顔画像の解像度が認識精度に影響があったとの報告もある[17)]．そこで特定の部位のみ画像解像度をあげることにより，メモリー量の増加を抑えつつ，認識精度を向上させることを考える．本論文においては，口部分，および目眉部分を特定部位の候補とし，これらの部位についてより高解像度の画像による解析を上半身画像の解析に併用する．

提案手法の検証にあたり，まず限定された条件における2クラス分類により高解像度化の効果を確認する．すなわち，高解像度口画像の併用については同手指異義語の判別精度の検証を行う．また，高解像度目眉画像の併用については修飾語の有無の判別精度の検証を行う．次に数千種類の単語を候補とする多クラス分類において，2クラス分類実験で効果が認められる単語の認識精度の向上があるかを検証する．最後に，2クラス分類で効果が認められた単語を除いた全単語の認識実験を行い，他の単語の認識への悪影響がないかを確認する．

2023年5月31日受付，2023年8月25日再受付，2023年9月12日採録
†NHK財団
（〒157-0073 世田谷区砧1-10-11，TEL 03-5494-2400）
††明治大学
（〒164-8525 中野区中野4-21-1，TEL 03-5343-8040）

Machine Translated by Google

こちらは英語原著論文の機械翻訳版です。
次ページが原著論文で、翻訳版と交互に展開されます。機械翻訳のため、誤字や誤訳、翻訳が未反映の部分が含まれている可能性があります。

As a result of the experiment, it was found that the accuracy of two-class classification was improved by using high-resolution images for the mouth images, but no improvement was observed for the eye and eyebrow images.

In terms of usage, improvements were also observed in multi-class classification, demonstrating the effectiveness of the proposed method.

In the following, Chapter 2 introduces related research, and in Chapter 3, we conduct a verification experiment of the effect of increasing the resolution of the mouth and eyebrows using two-class classification. In Chapter 4, we conduct a word recognition experiment to verify the effect of increasing the resolution of the mouth. The effectiveness of using the same image is verified in Chapter 5. To summarize.

2. Related research

In sign language, expressions using parts of the body other than the hands and arms, such as facial expressions, head movements, and body orientation, are called non-manual actions (or non-manual signals)8) 9). Among non-manual actions, facial expressions are important, as they play a role in distinguishing homophonic words7) 12) whose meaning cannot be determined by the shape of the hands and fingers alone, using the shape of the mouth, adding meanings of doubt or negation8) 9), and acting as modifiers expressing degree10) 13).

Early research into sign language recognition only dealt with the analysis of hands and arms. Analysis of hands and arms began with the use of gloves equipped with sensors to obtain information18)-20). One issue with gloves is that they restrict the movement of the speaker, so research has been reported on image analysis to identify words from the position and contour shape of the fingers as a method that does not restrict the speaker's movement21)-22). In addition, there have been reports of a case in which the shape of the mouth is analyzed using feature points extracted from a facial image to identify synonyms11), and a case in which the shape of the eyes and eyebrows is analyzed to identify questions and negative expressions23)-25).

In recent years, research has been reported on applying deep learning, which has a proven track record in recognizing body movements and facial expressions, to images of the speaker's upper body and comprehensively analyzing the shape and positional relationship of each body part14)-16). Deep learning involves a convolutional neural network (CNN) that analyzes images of each frame and a time series of
Recurrent Neural Network (RNN) that analyzes CNN output
In deep learning, the computational complexity and memory size
required to handle hundreds of time-series images are a constraint, so in previous studies, images of the upper body with reduced resolution were used from the captured images, as shown in Table 1.
Depending on the body part, the resolution of these cameras may be insufficient to allow for sufficient discrimination, and the resolution is limited to the area around the fingers.

Table 1. CNN input image sizes in previous studies of sign language recognition

文献番号	手話映像データセット 名称	画像サイズ [画素]	CNN入力 画像サイズ [画素]
14	PHOENIX-2014-T [31]	210×260	224×224
	CSL [32]	1920×1080	
15	PHOENIX-2014-T [31]	210×260	224×224
	CSL [32]	1920×1080	
16	CSL-BS [16]	非公開	224×224
	CSL [33]	640×480	

Research has also been reported on improving recognition performance by analyzing high-quality images6).

In the field of facial expression recognition research, differences in the resolution of facial images have been recognized. It has been reported that this affects the recognition accuracy. 17) To the best of the author's knowledge , there has been no verification as to whether the current resolution of face images is sufficient for sign language recognition applications.

In this study, we analyzed high-resolution images of the mouth, eyes, and eyebrows to distinguish between synonyms and modifiers.
By combining this with the analysis of upper body images, the accuracy of classification can be improved. Plan.

3. Effect of high resolution on two-class classification experiment verification

In this chapter, we confirm the improvement effect of using high-resolution mouth and eye and eyebrow images under limited conditions. For the mouth, we use pairs of synonymous words with the same hand and finger as the target, and conduct a two-class classification experiment of word videos to evaluate the discrimination accuracy. Since eyes and eyebrows have meaning as modifiers due to their co-occurrence with modified words, we use images of modified words as the target and conduct an experiment to classify them into two classes based on whether or not they have a modifier.

3.1 Experimental Method

The experiment uses footage of a news program in Japanese sign language (Table 2). There are two types of sign language used in Japan: "Japanese Sign Language," which has its own grammar and vocabulary, and "Japanese Sign Language," which is based on Japanese grammar and vocabulary. However, the program footage shown in Table 2 does not distinguish between Japanese Sign Language and Japanese Sign Language, and is designed to be understood by as many people as possible. Therefore, this footage can be considered to have the characteristics of both Japanese Sign Language and Japanese Sign Language.

From the videos in Table 2, the start and end frames of the word expressions to be classified into two classes were manually identified, and word videos were extracted. For the two-class classification of word videos, a network combining CNN and RNN , which are used for body movement recognition27) and sign language word classification28), is used. An overview of the classification network is shown in Figure 1, and the network specifications are shown in Table 3. The network shown in Figure 1 sequentially analyzes the images of each frame using CNN, and the output is input to a bidirectional RNN. The RNN summarizes the input data using time series analysis, and after integrating the summary results in the forward and backward directions in the fully connected layer, the classifier obtains the discrimination result. The upper body image is used in combination with the mouth or eyebrow images.

In this case, the CNN output of each image is concatenated and input to the RNN. For example, if the CNN output of the upper body image is an N-dimensional vector and the CNN output of the mouth image is an M-dimensional vector,

Table 2. Specifications of the sign language video used in the experiment

番組名	NHK手話ニュース NHK手話ニュース845 NHK週間手話ニュース
手話文の数	35,000
単語の種類数	3,726
画像サイズ(話者上半身)	800×800 画素
フレームレート	29.97枚/秒
話者数	17名

【機械翻訳コンテンツの著作権について】
当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
詳細は映像情報メディア学会著作権規定をご覧ください。

実験の結果，口画像については，高解像度画像の併用により2クラス分類の精度の向上が認められたが，目眉画像については向上が認められなかった．高解像度口画像の併用については多クラス分類においても改善効果が認められ，提案手法の有効性が示された．

以下，2章においては関連研究について紹介し，3章において口と目眉の部位の高解像度化の効果の検証実験を2クラス分類により行う．4章で単語認識実験により口の高解像度画像の併用の効果検証を行う．5章で考察を述べ，6章でまとめる．

2．関連研究

手話における顔の表情，頭部の動き，体の向きなどの手指と腕以外の部位による表現は非手指動作（または非手指信号）と呼ばれる[8)9)]．非手指動作の中でも顔の表情は重要であり，手指の形だけでは意味が定まらない同手指異義語[7)12)]を口の形で判別する役割や，目と眉の形により疑問，否定の意味をつけ加える役割[8)9)]，および程度を表す修飾語の役割を持つ[10)13)]．

初期の手話認識研究では手指と腕の解析のみを扱うにとどまっていた．手指と腕の解析はセンサ付きグローブで情報を取得する手法から検討が始まった[18)～20)]．グローブは話者の動きを制限するという課題があり，話者の動きを制限しない手法として画像解析により，手指の位置や輪郭形状から単語を判別する研究が報告されている[21)22)]．また，顔画像から抽出した特徴点を利用して口の形を解析し同手指異義語を判別する事例[11)]や，目と眉の形の解析により疑問や否定の表現を判別する事例が報告されている[23)～25)]．

近年では，身体の動きや顔表情の認識において実績のある深層学習を話者の上半身画像に適用し，各身体部位の形と，各部位の位置関係を統合的に分析する研究が報告されている[14)～16)]．深層学習では，各フレームの画像を解析するCNN（Convolutional Neural Network）と，時系列のCNN出力を解析するRNN（Recurrent Neural Network）やTransformer[26)]を同時に学習させている．

深層学習では数百枚の時系列画像を扱う計算量とメモリー量が制約となるため，先行研究では**表1**に示すような撮影された画像から解像度を落とした上半身画像を使用している．身体部位によってはこれらでは解像度不足で充分な判別ができない可能性があり，手指周辺に限り解像度の高い画像を解析して認識性能を改善する研究も報告されている[6)]．

表情認識の研究分野において顔画像の解像度の違いが認識精度に影響を与えることが報告されている[17)]．筆者の知る限り，手話認識の応用において顔画像の解像度が現状のもので充分かどうかの検証はされていない．

本研究では同手指異義語および修飾語の判別について，口部分および目と眉部分に限り解像度の高い画像を解析し，上半身画像の解析と併用することで判別精度の改善を図る．

表1　手話認識の先行研究におけるCNN入力画像サイズ

文献番号	手話映像データセット 名称	手話映像データセット 画像サイズ [画素]	CNN入力画像サイズ [画素]
14	PHOENIX-2014-T [30)]	210×260	224×224
	CSL [29)]	1920×1080	
15	PHOENIX-2014-T [31)]	210×260	224×224
	CSL [32)]	1920×1080	
16	CSL-BS [16)]	非公開	224×224
	CSL [33)]	640×480	

3．2クラス分類実験による高解像度化の効果検証

本章では解像度の高い口画像および目眉画像による改善効果を，限定された条件により確認する．口については，対象を同手指異義語のペアとし，単語映像の2クラス分類実験を行い判別精度を評価する．目眉は被修飾語との共起により修飾語として意味を持つため，被修飾語となる単語映像を対象し，修飾語の有無を2クラス分類する実験を行う．

3.1　実験方法

実験には日本の手話によるニュース番組の映像を使用する（**表2**）．国内で使われる手話には，独自の文法と語彙をもつ「日本手話」と，日本語の文法と語彙をベースとする「日本語対応手話」があるが，表2に示す番組映像は日本手話と日本語対応手話を区別せず，より多くの人に伝わるように表現が考えられている．そのため，この映像は日本手話と日本語対応手話の両方の性質を併せ持つと考えられる．

表2の映像から，2クラス分類の対象とする単語表現について開始フレームから終了フレームまでを手作業により特定し，単語映像を抽出した．単語映像の2クラス分類には，身体動作の認識[27)]や手話単語のクラス分類[28)]に用いられるCNNとRNNを組み合わせたネットワークを使用する．分類ネットワークの概要を**図1**に，ネットワークの諸元を**表3**に示す．図1に示すネットワークは各フレームの画像をCNNで順次解析し，その出力を双方向のRNNに入力する．RNNは時系列解析により入力データを要約し，順方向と逆方向の要約結果を全結合層で統合した後，分類器により判別結果を得る．上半身画像に口または目眉画像を併用する場合は，それぞれの画像のCNN出力を連結してRNNに入力する．例えば，上半身画像のCNN出力がN次元ベクトル，口画像のCNN出力がM次元ベクトルの場合，

表2　実験に使用する手話映像の諸元

番組名	NHK手話ニュース NHK手話ニュース845 NHK週間手話ニュース
手話文の数	35,000
単語の種類数	3,726
画像サイズ(話者上半身)	800×800 画素
フレームレート	29.97枚／秒
話者数	17名

Machine Translated by Google

こちらは英語原著論文の機械翻訳版です。
次ページが原著論文で、翻訳版と交互に展開されます。機械翻訳のため、誤字や誤訳、翻訳が未反映の部分が含まれている可能性があります。

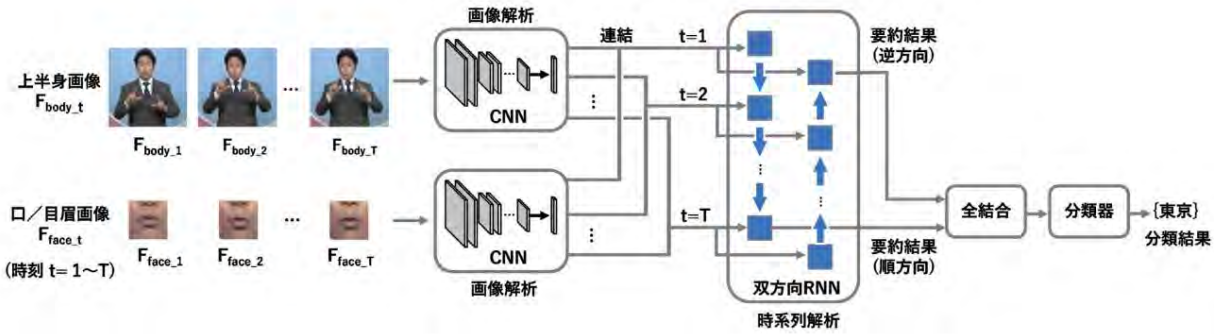

Figure 1 Word video classification network

Table 3. Classification network specifications

画像解析（CNN）	ResNet18
時系列解析（双方向RNN）	Bidirectional LSTM
損失関数	Cross Entropy Loss
バッチサイズ	32
学習率	0.0001
最適化アルゴリズム	ADAM
深層学習フレームワーク	PyTorch

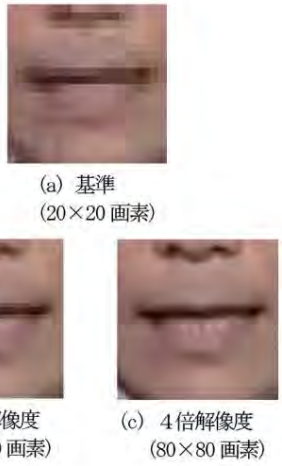

Figure 2. Differences in mouth image resolution ((a)

and (b) are enlarged to 80 x 80 pixels using nearest neighbor interpolation)

Figure 3 Differences depending on the resolution of eye and eyebrow images

((a) and (b) are enlarged to 140 x 90 pixels using nearest neighbor interpolation.)

Table 4 Total number of pixels in the input image

入力画像サイズ［画素］(解像度倍率) 上半身	口	目眉	総画素数 (対基準比)
200×200 (基準)	—	—	40,000 (100%)
	40×40 (2倍)	—	41,600 (104%)
	80×80 (4倍)	—	46,400 (116%)
	—	70×45 (2倍)	43,150 (108%)
	—	140×90 (4倍)	52,600 (132%)

The input of an RNN is an N+M dimensional vector.

The upper body image input to the CNN was reduced in both horizontal and vertical resolution to 1/4 of the original image (800 x 800 pixels), making it the same size as in previous studies (200 x 200 pixels). This was used as the standard. The mouth image and eye and eyebrow images were cut out from the original image and the original image was cut in half the number of pixels horizontally and vertically.

In other words, the resolution of these images is four times higher in the horizontal and vertical directions than that of the reference image. The difference in the resolution of the mouth image due to the difference is shown in Figure 2, and the difference in the resolution of the eye and eyebrow images is shown in Figure 3.

The amount of calculation and memory required for image analysis is proportional to the number of pixels in the input image. Table 4 shows the increase in the number of pixels when a mouth image or eye and eyebrow images are used in combination. If an upper body image (four times the resolution of the original image) is used, the number of pixels will be 1600%, whereas the use of a mouth image with four times the resolution will increase the number of pixels by 16%, and the use of a double resolution image will increase the number of pixels by 4%. Similarly, the use of an eye and eyebrow image with four times the resolution will increase the number of pixels by 1600%. Using the same images increases the number of pixels by 32%, while using double resolution

increases it by 8%. 3.2 Using high

resolution mouth images together We evaluate three pairs of homographs that appear frequently in the sign language videos in Table 2: {East}/{Tokyo}, {West}/{Kyoto}, and {Military}/{Hyogo}*1 . Homographs with the same hand and finger shapes have different mouth shapes, as shown in the image in Figure 4. For each word, 100 samples are extracted and used for evaluation. All of the extracted samples have different mouth shapes, as in the example images in Figure 4. Note that for {Tokyo} and {Kyoto}, not only the mouth shapes but also the arms are different.

*1 The notation of enclosing Japanese characters in brackets { } is used to identify sign language words9).

【機械翻訳コンテンツの著作権について】
当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
詳細は映像情報メディア学会著作権規定をご覧ください。

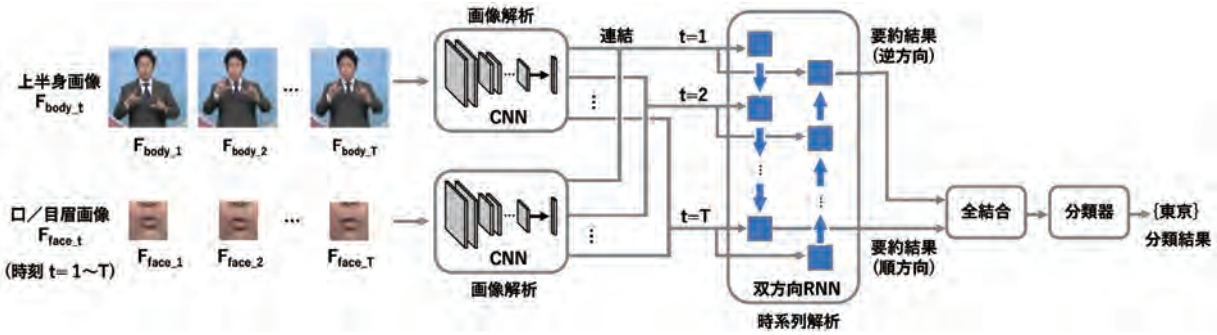

図1　単語映像のクラス分類ネットワーク

表3　クラス分類ネットワークの諸元

画像解析（CNN）	ResNet18
時系列解析（双方向RNN）	Bidirectional LSTM
損失関数	Cross Entropy Loss
バッチサイズ	32
学習率	0.0001
最適化アルゴリズム	ADAM
深層学習フレームワーク	PyTorch

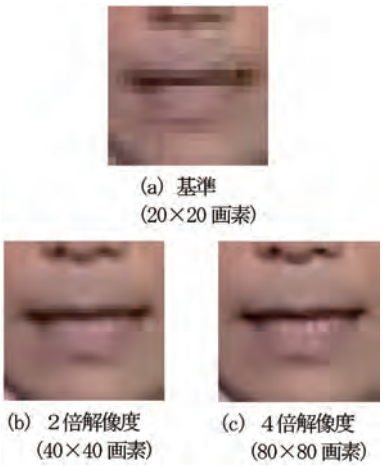

図2　口画像の解像度による違い
((a) (b) は最近傍補間により 80 × 80 画素に拡大表示)

図3　目眉画像の解像度による違い
((a) (b) は最近傍補間により 140 × 90 画素に拡大表示)

表4　入力画像の総画素数

入力画像サイズ［画素］（解像度倍率） 上半身	口	目眉	総画素数（対基準比）
200×200（基準）	—	—	40,000（100%）
	40×40（2倍）	—	41,600（104%）
	80×80（4倍）	—	46,400（116%）
	—	70×45（2倍）	43,150（108%）
	—	140×90（4倍）	52,600（132%）

RNNの入力はN + M次元のベクトルとなる．

CNNに入力する上半身画像は元画像（800 × 800画素）の水平・垂直方向の解像度を各々1/4に縮小し，先行研究と同程度のサイズ（200 × 200画素）とした．これを基準とする．口画像および目眉画像は，元画像から当該部位を切り出した画像，および元画像を水平・垂直各々1/2の画素数に縮小した画像から切り出した画像を使用する．すなわち，これらの画像の解像度は基準画像に比べて，水平垂直の解像度が各々4倍，2倍になる．口画像の解像度による違いを図2に，目眉画像の違いを図3に示す．

画像解析に係る計算量とメモリー量は入力画像の画素数に比例する．口画像または目眉画像の併用による画素数増加について表4に示す．仮に元画像（基準に対して4倍解像度）の上半身画像を使用する場合，画素数は1600%となるのに対して，4倍解像度の口画像の併用では16%の増加，2倍解像度では4%の増加となる．同様に，4倍解像度の目眉画像の併用では画素数は32%の増加，2倍解像度では8%の増加となる．

3.2　高解像度口画像の併用

表2の手話映像において出現回数の多い同手指異義語{東}/{東京}，{西}/{京都}，{軍}/{兵庫} *1の3ペアについて評価する．手指の形が同じ同手指異義語は，図4の画像に示すように口の形に違いがある．各単語について，100個のサンプルを抽出し，評価に用いる．抽出したすべてのサンプルは図4の画像例と同様に口の形に違いを持つ．なお，{東京} および {京都} については口の形だけでなく，腕を

＊1　日本語の文字を｛｝で括る表記は手話単語の識別に用いられる[9]．

Machine Translated by Google

こちらは英語原著論文の機械翻訳版です。
次ページが原著論文で、翻訳版と交互に展開されます。機械翻訳のため、誤字や誤訳、翻訳が未反映の部分が含まれている可能性があります。

（a）左：{東}，右：{東京}　（b）左：{西}，右：{京都}

（c）左：{軍}，右：{兵庫}

Figure 4 Examples of synonyms for hands and mouths

Table 5. Two-class classification results for the same hand-fingered opposing words

入力画像サイズ［画素］(解像度倍率) 上半身	口	判別精度［%］ 平均	{東} {東京}	{西} {京都}	{軍} {兵庫}
200×200 (基準)	—	83.5	82.1	89.6	78.8
	40×40（2倍）	92.7	93.3	95.6	89.1
	80×80（4倍）	92.5	93.6	95.6	88.2

In some cases, {East} and {West} can be distinguished by shaking the arm up and down twice. {Tokyo} has 8 out of 100 samples with two arm swings up and down, and {Kyoto} has 18 out of 100 samples with two arm swings up and down.

Only upper body images and a combination of mouth and upper body images were input to the network to compare the discrimination accuracy. Cross-validation was performed using one of the 10 speakers included in all word video samples for evaluation and the remaining nine for training, and the average discrimination accuracy calculated for all speakers is shown in Table 5. The use of high-resolution mouth images in combination improved the discrimination accuracy by 9 points for both double and quadruple resolution.

In order to confirm in detail the effect of using high-resolution mouth images in combination with samples that can be distinguished even from the arm, we compare the number of samples that failed to be recognized. For {Tokyo}, the number of recognition failure samples was 5 when only the upper body image was analyzed, but when the mouth was also used, this reduced to 2 at 2x resolution and 3 at 4x resolution. For {Kyoto}, the number of recognition failure samples was 3 when only the upper body image was used, but this reduced to 2 each at 2x and 4x resolution when the mouth was also used. 3.3 Using High-Resolution Eye and Eyebrow Images In the modifier discrimination experiment, we

targeted the modified word {rain}, which appears

frequently in the videos in Table 2, and {very heavy rain}, which is accompanied by the modifier "very heavy," and selected 100 word videos for each {rain}/{very heavy rain} pair to evaluate the discrimination accuracy.

Figure 5. Examples of the modified word "hands" and the modifier "eyes"

Table 6. Two-class classification results for modifiers

入力画像サイズ［画素］(解像度倍率) 上半身	目眉	判別精度［%］ {雨} {非常に激しい雨}
200×200 (基準)	—	93.4
	70×45（2倍）	90.6
	140×90（4倍）	92.4

As shown in the example image of {rain}/{very heavy rain} in Figure 5, there is a difference in the shape of the eyes and eyebrows. All of the extracted samples have different eye and eyebrow shapes, just like the example image in Figure 5. However, in addition to the eyes and eyebrows, there is also an expression that involves the movement of the arms up and down quickly and widely to add the modifier "very heavy" to {rain}. 78 out of 100 samples include this kind of movement.

We input only upper body images and images of the eyes and eyebrows together with the upper body into the network and compared the discrimination accuracy. As in the experiment on the same hand and finger synonyms, cross-validation was performed for each speaker, and the average discrimination accuracy calculated for all speakers is shown in Table 6. When upper body images and eyes and eyebrow images were used together, the accuracy was 90.6% at double resolution and 92.4% at four times resolution, both of which were lower than the 93.4% achieved by the upper body image alone.

In order to confirm the effect of using high-resolution eyebrow and eyebrow images in detail, we compare the number of recognition failure samples for modifier samples that are distinguishable even with arms.

The number of eyes and eyebrows was nine in both cases (2x and 4x resolution), and there was no increase or decrease due to the use of eye and eyebrow images in combination.

From these results, it was not confirmed that the use of high-resolution eye and eyebrow images improved the discrimination of the {rain}/{very heavy rain} pair.

4. Verification of the effect of high resolution on sign language recognition

We verified the improvement of the recognition of sign language sentence videos for the discrimination of synonyms with the same hand, which was shown to be improved in a two-class classification experiment using high-resolution images.

Since many words are used, the classification becomes a multi-class classification problem in which one word is identified from multiple word candidates. In this section, we first classify all words from the video of the sign language sentence, and evaluate the classification accuracy of the same hand synonyms included in the classification results by F-measure. Next, we examine the effect of word errors on the recognition accuracy for all words.

The evaluation is based on the rate29).

【機械翻訳コンテンツの著作権について】
当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
詳細は映像情報メディア学会著作権規定をご覧ください。

（a）左：{東}，右：{東京}　（b）左：{西}，右：{京都}

（c）左：{軍}，右：{兵庫}

図4　同手指異義語の手指と口の例

表5　同手指異議語の2クラス分類結果

入力画像サイズ［画素］（解像度倍率） 上半身	口	判別精度［%］ 平均	{東} {東京}	{西} {京都}	{軍} {兵庫}
200×200（基準）	—	83.5	82.1	89.6	78.8
	40×40（2倍）	92.7	93.3	95.6	89.1
	80×80（4倍）	92.5	93.6	95.6	88.2

上下に2回振ることで {東} および {西} と区別される場合もある．{東京} は腕を上下に2回振るサンプルが100個中8個，{京都} は100個中18個含まれる．

上半身画像のみ，および口と上半身画像の併用をそれぞれネットワークに入力し，判別精度を比較する．単語映像の全サンプルに含まれる10名の話者のうち1名を評価用，残りの9名を学習用とする交差検証を行い，全話者の判別精度から算出した平均値を表5に示す．高解像度口画像の併用により2倍解像度および4倍解像度のいずれも9ポイント改善している．

腕でも判別可能なサンプルについて，高解像度口画像の併用効果を詳細に確認するため，認識に失敗したサンプルの個数を比較する．認識失敗サンプルは，{東京} については上半身画像のみ解析する場合が5個に対して，口併用は2倍解像度が2個，4倍解像度が3個に減少した．{京都} については上半身画像のみが3個に対して，口併用は2倍・4倍解像度で各々2個に減少した．

3.3　高解像度目眉画像の併用

修飾語の判別実験では，表2の映像において出現回数の多い被修飾語 {雨} と，「非常に激しい」という修飾を伴う {非常に激しい雨} を対象とし，{雨} / {非常に激しい雨} のペアについて各100単語映像を選び判別精度を評価する．

図5　被修飾語の手指と修飾語の目眉の例

表6　修飾語の2クラス分類結果

入力画像サイズ［画素］（解像度倍率） 上半身	目眉	判別精度［%］ {雨} {非常に激しい雨}
200×200（基準）	—	93.4
	70×45（2倍）	90.6
	140×90（4倍）	92.4

図5の {雨} / {非常に激しい雨} の画像例のように，両者には目眉の形に違いがある．抽出したすべてのサンプルは図5の画像例と同様に目眉の形に違いを持つ．ただし，{雨} に「非常に激しい」という修飾を加える方法としては，目眉以外にも，腕を素早く大きく上下させる動作が伴う表現がある．このような動作を伴うサンプルは100個中78個含まれる．

上半身画像のみ，および目眉と上半身画像の併用をそれぞれネットワークに入力し，判別精度を比較する．同手指異義語の実験と同様に話者別の交差検証を行い，全話者の判別精度から算出した平均値を表6に示す．上半身画像と目眉画像を併用する場合，2倍解像度では90.6%，4倍解像度では92.4%となり，いずれも上半身画像のみの93.4%を下回った．

腕でも判別可能性な修飾語のサンプルについて，高解像度目眉像の併用効果を詳細に確認するため，認識失敗サンプルの個数を比較する．上半身画像のみ解析する場合および目眉併用（2倍・4倍解像度）いずれにおいても9個となり，目眉画像の併用による増減は無かった．

これらの結果から，{雨} / {非常に激しい雨} のペアの判別については，解像度の高い目眉画像の併用による改善効果は確認されなかった．

4．手話文の認識による高解像度化の効果検証

解像度の高い画像を用いた2クラス分類実験で改善効果が示された同手指異義語の判別について，手話文映像の認識での改善効果を検証する．実際の手話文では多くの種類の単語が使われるため，判別は複数の単語候補から一つの単語を特定する多クラス分類問題となる．本章ではまず手話文の映像からすべての単語を判別し，判別結果に含まれる同手指異義語について判別精度をF値により評価する．次にすべての単語についての認識精度への影響を単語誤り率[29]により評価する．

Machine Translated by Google

こちらは英語原著論文の機械翻訳版です。
次ページが原著論文で、翻訳版と交互に展開されます。機械翻訳のため、誤字や誤訳、翻訳が未反映の部分が含まれている可能性があります。

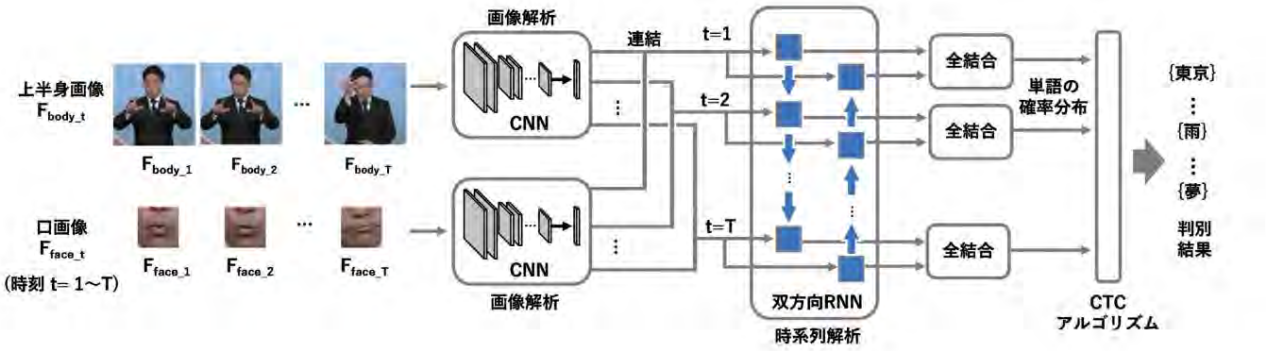

Figure 6 Sign language sentence recognition network

4.1 Experimental conditions Classification of sign language words with one-to-one correspondence between images and words In contrast, the mainstream method for recognizing sign language sentences is to distinguish the time sequence of words from video in which the boundaries between words are unclear14)-16). For the videos shown in Table 2, the video was manually extracted from the start frame to the end frame of the sign language sentence, and time sequence word information was added to the video, which was then used for training and evaluation.

To identify time-sequence words from video, we used CNN and bidirectional The network output of the combined RNN is CTC (Connectionist We apply the Temporal Classification (TCM) algorithm30) to the dataset (Figure 6). CTC is an algorithm that determines the time-series word that maximizes the likelihood from the probability distribution of the words output by the network in a time series.

The input images to the network are either upper body images alone, or a combination of high-resolution mouth images and upper body images. For the mouth images, images with 2x and 4x resolution are used. The specifications of the video used in the experiment are shown in Table 7, and the specifications of the sign language recognition network are shown in Table 8.

4.2 Accuracy of homograph discrimination We extract homograph results from all words discriminated from the sign language video and calculate the discrimination accuracy of them.

Table 7. Image specifications used in sign language sentence recognition

	学習用	評価用
手話文の数 (内，同手指異義語を含む文)	30,000 (360)	5,000 (240)
総語数 (内，同手指異義語数)	335,154 (360)	56,253 (240)
単語の種類数	3,726	1,776
文あたりの平均語数	11語	
話者数 (内，同手指異義語話者)	17名 (10名)	

Table 8. Specifications of the sign language sentence recognition network

画像解析（CNN）	ResNet18
時系列解析（双方向RNN）	Bidirectional LSTM
損失関数	CTC Loss
バッチサイズ	256
学習率	0.0001
最適化アルゴリズム	ADAM
深層学習フレームワーク	PyTorch

is the harmonic mean (F-value) of the proportion of correctly identified words (recall) and the proportion of correctly identified words (precision).

In order to evaluate the precision rate of erroneous detections from sign language sentences that do not contain homosynonyms, Table 9 shows the results of identifying homosynonyms from the evaluation video of 5,000 sentences (240 of which contain homosynonyms) shown in Table 7. The average F-measure was 90.1% for upper body images alone, and 95.5% for the combined use of double-resolution mouth images, an improvement of about 5 points, and 94.0% for the quadruple resolution, an improvement of about 4 points.

4.3 Recognition accuracy of all words

Sign language sentences contain many words that are unrelated to the shape of the mouth, and we verified whether the analysis of the added mouth image would hinder the discrimination of these words. Word discrimination accuracy (F-value) is calculated by comparing the detected word sequence (discrimination result) with the correct word sequence, and identifying the number of successful detections, false positives, and missed detections for each word. However, if the detected word order differs from the correct word order, it is not possible to uniquely determine successful detection, false positives, or missed detections*2. In the detection of synonymous words with the same hand shown in Table 9, the F-value was calculated after confirming that there were no cases that could not be determined, but if there are cases that could not be determined, measures such as excluding them will be necessary. A total of more than 50,000 sign language sentences were included in the evaluation. For words (Table 7), we use the word error rate, which is generally used as an index to evaluate the recognition accuracy of sign language sentences and measures the degree of error on a sentence-by-sentence basis. The word error rate treats the number of word operations (insertion, deletion, and substitution) required to convert the detected word sequence to the correct word sequence as the degree of error, and evaluates the difference between the detected word order and the correct word order as the number of operations required for the conversion.

The 5,000 sentences used for evaluation in Table 7 were excluded from the 240 sentences containing homophonic digits, leaving 4,760 sentences (total number of words: 52,668). The average word error rate calculated under the above experimental conditions is shown in Table 10. When a mouth image with double resolution was used in combination, the error rate was 0.319, and when it was 4 times resolution, the error rate was 0.327, both of which were lower than the 0.335 obtained when only an upper body image was used. For comparison, the average word error rate for the 240 sentences containing homophonic digits (total number of words: 3,585) is also shown.

*2 For example, if the correct word sequence is {A}{B}{C} and the detected word sequence is {B}{A}{C}, the first {B} is a false positive, {A} is a successful detection (it should come next).
It is possible that {B} is a false negative, {A} (which is originally at the beginning) is a false negative, {B} is a successful detection, and {A} is a false positive.

【機械翻訳コンテンツの著作権について】
当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
詳細は映像情報メディア学会著作権規定をご覧ください。

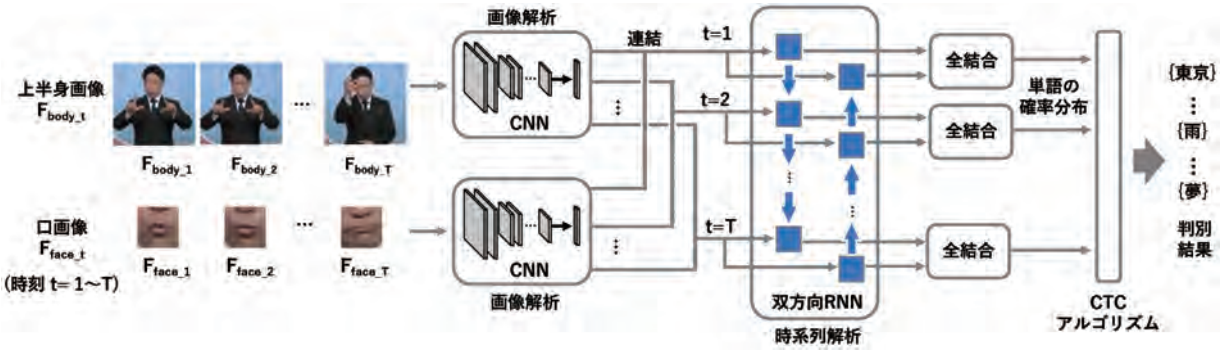

図6　手話文の認識ネットワーク

4.1　実験条件

映像と単語が1対1で対応する手話単語のクラス分類と異なり，手話文の認識では単語間の境界が不明な映像から時系列の単語を判別する方法が主流となっている[14)～16)]．表2に示す映像について，手話文の開始フレームから終了フレームまでを手作業により抽出した映像に時系列の単語情報を付与したデータを学習と評価に使用する．

映像から時系列の単語を判別するため，CNNと双方向RNNを組み合わせたネットワーク出力にCTC (Connectionist Temporal Classification) アルゴリズム[30)]を適用する (図6)．CTCは，ネットワークが時系列に出力する単語の確率分布から，尤度が最大となる時系列の単語を決定するアルゴリズムである．

ネットワークへの入力画像は，上半身画像のみ，あるいは解像度の高い口画像と上半身画像の組み合わせである．口画像として，2倍および4倍解像度の画像を使用する．実験に使用する映像の諸元を表7に，手話文の認識ネットワークの諸元を表8に示す．

4.2　同手指異義語の判別精度

手話文映像から判別したすべての単語の中から同手指異義語の結果を抽出し，それらの判別精度を求める．評価には出現した単語を正しく判別した割合 (再現率) と，判別した単語が正解であった割合 (適合率) の調和平均 (F値) を用いる．

適合率については同手指異義語を含まない手話文からの誤検出も評価するため，表7に示す5,000文 (内，240文が同手指異義語を含む) の評価用映像を対象として同手指異義語を判別した結果を表9に示す．F値の平均値は，上半身画像のみの90.1%に対して，2倍解像度の口画像の併用が95.5%となり約5ポイント，4倍解像度が94.0%となり約4ポイントの改善があった．

表7　手話文の認識に用いる映像の諸元

	学習用	評価用
手話文の数 (内，同手指異義語を含む文)	30,000 (360)	5,000 (240)
総語数 (内，同手指異義語数)	335,154 (360)	56,253 (240)
単語の種類数	3,726	1,776
文あたりの平均語数	11語	
話者数 (内，同手指異義語話者)	17名 (10名)	

表8　手話文の認識ネットワークの諸元

画像解析 (CNN)	ResNet18
時系列解析 (双方向RNN)	Bidirectional LSTM
損失関数	CTC Loss
バッチサイズ	256
学習率	0.0001
最適化アルゴリズム	ADAM
深層学習フレームワーク	PyTorch

4.3　全単語の認識精度

手話文には口の形状に無関係な単語が多数存在するが，追加した口画像の解析がこれらの単語の判別を阻害しないかを検証する．単語の判別精度 (F値) は，検出語列 (判別結果) と正解語列を比較し，単語ごとに，検出成功，誤検出，検出漏れの回数を特定して算出するが，検出した語順が正解の語順と異なる場合，一意に検出成功・誤検出・検出漏れを判断できない*2．表9で示した同手指異義語の検出では判断不能な事例がないことを確認しF値を計算したが，判断不能な事例が含まれる場合は除外するなどの対応が必要となる．評価用の手話文に含まれる延べ5万以上の単語 (表7) については，手話文の認識精度を評価する指標として一般に用いられ，文単位で誤りの程度を測る単語誤り率を用いる．単語誤り率は，検出語列から正解語列への変換に必要な単語の操作回数 (挿入，削除，置換) を誤りの程度として扱い，検出した語順と正解の語順の違いもその変換に必要な操作回数として評価される．

表7に示す評価用の5,000文から同手指異義語を含む240文を除いた4,760文 (総語数52,668) を対象とし，先の実験条件において算出した単語誤り率の平均値を表10に示す．2倍解像度の口画像の併用では0.319，4倍解像度の場合は0.327となり，いずれも上半身画像のみの0.335と比較して低い誤り率となった．比較のため，同手指異義語を含む240文 (総語数3,585) を対象とした単語誤り率の平均値を表

*2　例として，正解語列が {A} {B} {C}，検出語列が {B} {A} {C} であるとき，最初の {B} は誤検出・{A} は検出成功・(本来次にあるべき) {B} は検出漏れとなる場合，(本来冒頭の) {A} の検出漏れ・{B} は検出成功・{A} は誤検出となる場合とが考えられる．

Machine Translated by Google

こちらは英語原著論文の機械翻訳版です。
次ページが原著論文で、翻訳版と交互に展開されます。機械翻訳のため、誤字や誤訳、翻訳が未反映の部分が含まれている可能性があります。

Table 9. Recall, precision, and F-score for same-sign object words in sign language sentence recognition

	入力画像サイズ [画素] (解像度倍率) 上半身	口	平均	{東}	{東京}	{西}	{京都}	{軍}	{兵庫}
再現率 [%]	200×200 (基準)	—	93.3	88.9	91.2	85.5	93.9	100.0	100.0
		40×40 (2倍)	**96.8**	**97.8**	91.2	**94.6**	**97.0**	100.0	100.0
		80×80 (4倍)	93.9	93.3	91.2	90.9	90.9	96.8	100.0
適合率 [%]	200×200 (基準)	—	87.5	83.3	75.6	88.7	93.9	88.6	94.9
		40×40 (2倍)	**94.5**	91.7	**91.2**	83.9	**100.0**	**100.0**	**100.0**
		80×80 (4倍)	94.2	**97.7**	83.8	**98.0**	88.2	**100.0**	97.4
F値 [%]	200×200 (基準)	—	90.1	86.0	82.6	87.0	93.9	93.9	97.4
		40×40 (2倍)	95.5	94.6	**91.2**	88.9	**98.5**	**100.0**	**100.0**
		80×80 (4倍)	94.0	**95.5**	87.3	**94.3**	89.6	98.4	98.7

Table 10. Mean word error rate

入力画像サイズ [画素] (解像度倍率) 上半身	口	単語誤り率 平均値 同手指異義語を含まない文	同手指異義語を含む文
200×200 (基準)	—	0.335	0.221
	40×40 (2倍)	0.319	0.162
	80×80 (4倍)	0.327	0.202

The error rate for the two-times resolution mouth image was 0.162, and for the four-times resolution it was 0.202, both of which were lower than the error rate for the upper body image alone (0.221).

5. Observations

The use of high-resolution mouth images in combination with the 3D-based image processing has been shown to improve the two-class classification of word images in the discrimination of synonyms with the same hand, and has also been shown to improve the multi-class classification of sign language sentence recognition. In this study, the difference in the improvement effect between 2x and 4x resolution was small, and 2x resolution is considered sufficient. Considering that 4x resolution would increase the amount of calculation and memory, it seems appropriate to use 2x resolution in combination. For the {Tokyo} {Kyoto} sample, which can be distinguished by arm movements, the number of failures was reduced by analyzing high-resolution mouth images. It is unclear to what extent the arms and mouth each contribute to the distinction, but it is considered that the use of high-resolution mouth images is effective even in samples that can be distinguished by arms. In addition, the analysis results of only upper body images showed a high F-value for multi-class classification of sign language sentence recognition compared to the two-class classification accuracy of word images. This is thought to be due to the word prediction effect based on the context of the sign language sentence. For example, when expressing "Hyogo Prefecture," {Ministry}*3 co- occurs immediately after {Hyogo}*4, so it is thought that it is easy to distinguish it from {Military} from the context.

When using high-resolution eye and eyebrow images to distinguish between modifiers, no improvement was observed in the two-class classification of whether modifiers were present or not. Since the current images of the eyes and eyebrows are part of the upper body images, it may be sufficient to use images of the eyes and eyebrows as part of the upper body images. Even when focusing on samples that can be distinguished by the arms, no improvement was observed because there was no increase or decrease in the number of failures when using high-resolution eyes and eyebrows in combination with the analysis of only the low-resolution upper body. Since there are many samples in which arm movements are meaningful, learning of the arms progresses and the contribution of the arms to discrimination increases, which may result in a relative decrease in the contribution of the eyes and eyebrows to discrimination, and the effect of increasing the resolution of the eyes and eyebrows may not be as pronounced. In this experiment, only one type of modified word was targeted, but it is thought that it is necessary to collect evaluation videos for other modified words and conduct separate verification. It is also necessary to verify video discrimination dealing with questions The word error rate of all the remaining sign language sentences, except for those with homophonic digits, did not worsen with the addition of mouth image analysis, and instead showed a tendency to improve.

This suggests that there is no or very little adverse effect on other word discrimination. The cause of this tendency to improve is the accuracy of discrimination for each word, which needs to be examined in detail. The word error rate was also evaluated for 240 sentences containing six types of homophonic digits, and the improvement was confirmed by the analysis of mouth images. From these results, it is believed that the performance of sign language recognition can be improved while suppressing the increase in the amount of calculation and memory by combining the analysis of upper body images and the analysis of mouth images with double resolution. In this paper, a recognition network with the most basic structure was used to verify the improvement effect of high-resolution face images. To realize more practical sign language recognition, it is necessary to consider the network structure and the method of analyzing input images.

6. Conclusion

In this paper, we verified the effect of analyzing high-resolution face images in improving sign language recognition. For homophonic words in which the hand and finger shapes are the same but the mouth shapes express different meanings, we confirmed that the accuracy of discrimination was improved by combining analysis of mouth images with twice the resolution compared to the standard upper body image. Although the amount of calculation and memory required increased by 4% by adding analysis of mouth images, the accuracy improved by 9 points in two-class classification and 5 points in multi-class classification.

*3 The sign language word {Sho} corresponds to the Japanese word "ken."

*4 As a similar example, to express "Kyoto Prefecture," the sign language word {Fu} is written immediately after {Kyoto}. co-occur.

【機械翻訳コンテンツの著作権について】
当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
詳細は映像情報メディア学会著作権規定をご覧ください。

表9　手話文の認識における同手指異議語の再現率，適合率，F値

	入力画像サイズ [画素]（解像度倍率） 上半身	口	平均	{東}	{東京}	{西}	{京都}	{軍}	{兵庫}
再現率 [%]	200×200（基準）	—	93.3	88.9	91.2	85.5	93.9	100.0	100.0
		40×40（2倍）	**96.8**	**97.8**	91.2	**94.6**	**97.0**	100.0	100.0
		80×80（4倍）	93.9	93.3	91.2	90.9	90.9	96.8	100.0
適合率 [%]	200×200（基準）	—	87.5	83.3	75.6	88.7	93.9	88.6	94.9
		40×40（2倍）	**94.5**	91.7	**91.2**	83.9	**100.0**	**100.0**	**100.0**
		80×80（4倍）	94.2	**97.7**	83.8	**98.0**	88.2	**100.0**	97.4
F値 [%]	200×200（基準）	—	90.1	86.0	82.6	87.0	93.9	93.9	97.4
		40×40（2倍）	95.5	94.6	**91.2**	88.9	**98.5**	**100.0**	**100.0**
		80×80（4倍）	94.0	**95.5**	87.3	**94.3**	89.6	98.4	98.7

表10　単語誤り率の平均値

入力画像サイズ [画素]（解像度倍率） 上半身	口	単語誤り率 平均値 同手指異義語を含まない文	同手指異義語を含む文
200×200（基準）	—	0.335	0.221
	40×40（2倍）	0.319	0.162
	80×80（4倍）	0.327	0.202

10に示す．2倍解像度の口画像の併用では0.162，4倍解像度では0.202となり，いずれも上半身画像のみの0.221と比較して低い誤り率となった．

5．考　察

高解像度口画像の併用により同手指異義語の判別において，単語映像の2クラス分類で改善効果が示された上，手話文認識の多クラス分類でも改善が確認された．両方の評価において，改善効果は2倍解像度と4倍解像度での差は小さく，この結果からは2倍の解像度で充分と考えられる．4倍解像度を用いると計算量とメモリー量がさらに増加することを考慮すると2倍解像度の併用が適切と思われる．腕の動きでも判別可能な {東京} {京都} のサンプルについて，高解像度な口画像の解析により失敗例が減少した．腕，口が各々どの程度判別に寄与しているかは不明であるが，腕でも判別可能なサンプルにおいても口の高解像度画像併用は効果があると考えられる．なお，上半身画像のみの解析結果は，単語映像の2クラス分類精度と比較して，手話文認識の多クラス分類のF値が高い水準となっているが，この要因として手話文の文脈に基づく単語予測効果が考えられる．例えば「兵庫県」を表す場合は {兵庫} の直後に {省} *3が共起するため*4，文脈から {軍} との区別が容易であると考えられる．

高解像度目眉画像の併用による修飾語の判別では，修飾語の有無の2クラス分類において改善効果は確認できなかった．低解像度の上半身画像でも93%の判別精度が得られているので，目眉については現在の上半身画像の一部としての画像で充分であるかもしれない．腕で判別可能なサンプルに注目した場合についても，低解像度な上半身のみの解析に対して，高解像度な目眉併用による失敗例の増減がないことから，改善効果は認められなかった．修飾語は腕の動きが意味を持つサンプル数が多いことから，腕の学習が進み，腕による判別の貢献が高まった結果，相対的に目眉による判別の貢献が下がり，さらに目眉の高解像度化の効果が顕著にならなかったかもしれない．今回の実験では一種類の被修飾語のみを対象としたが，他の被修飾語ついては評価用映像を収集のうえ，別途検証が必要と考えられる．さらに疑問・否定を扱った映像判別も検証する必要がある．

評価用の手話文から，同手指異義語を伴う文を除いた残りすべての手話文の単語誤り率は，口画像の解析の追加による悪化はなくむしろ改善傾向がみられたことから，他の単語判別への悪影響はない，もしくは非常に小さいと考えられる．この改善傾向の要因については，各々の単語の判別精度を詳細に検証する必要がある．6種類の同手指異義語を含む240文を対象とした単語誤り率についても評価した結果，口画像の解析により改善が確認された．以上の結果から，上半身画像の解析と2倍解像度の口画像の解析を併用することで，計算量とメモリー量の増加を抑えながら，手話認識の性能向上が可能と考えられる．

本論文では解像度の高い顔画像による改善効果を検証するため，最も基本的な構造の認識ネットワークを使用した．より実用的な手話認識実現に向けては，ネットワーク構造や入力画像の解析方法の検討が必要である．

6．む　す　び

本論文では，解像度の高い顔画像の解析による手話認識の改善効果を検証した．手指の形が同じでありながら口の形で異なる意味を表す同手指異義語については，基準となる上半身画像と比較して2倍解像度の口画像の解析を併用することで判別精度の改善が確認された．口画像の解析の追加により計算量・メモリー量は4%増加するが，2クラス分類で9ポイント，多クラス分類で5ポイントの精度の向

＊3　手話単語 {省} は日本語の「県」に相当する．
＊4　類似の例として「京都府」を表す場合は {京都} の直後に手話単語 {フ} が共起する．

こちらは英語原著論文の機械翻訳版です。
次ページが原著論文で、翻訳版と交互に展開されます。機械翻訳のため、誤字や誤訳、翻訳が未反映の部分が含まれている可能性があります。

上が示された．一方，目眉の形により表現される修飾語の判別については，解像度の高い目眉画像による改善効果は確認されなかった．手話文認識の実験により，口画像の解析の追加は同手指異義語以外の単語を対象とした判別精度についても悪影響を与えていない．これらの結果から，上半身画像と併せて解像度の高い口画像を解析することで，手話認識の性能向上が可能であることが明らかにされた．

今後，深層学習を用いた認識処理技術自体の改善と平行して，適正な認識のための各部位の時空間解像度といった入力画像のあり方についてより検討を行う必要がある．

手話映像の使用に協力頂いたNHK放送技術研究所に感謝する．

〔文　献〕

1) World Federation of the Deaf, https://wfdeaf.org
2) 日本手話通訳士協会, http://www.jasli.jp
3) N. Hiruma, M. Azuma, T. Uchida, S. Umeda, T. Miyazaki, N. Kato and S. Inoue: "Automatic Generation System of Japanese Sign Language (JSL) with CG Animation of Fixed Pattern Weather Information", ABU Technical Journal **264**, pp.2-5 (2015)
4) B. Kwolek, W. Baczynski and S. Sako: "Recognition of JSL Fingerspelling Using Deep Convolutional Neural Networks", Neurocomputing (June 2021)
5) N. Takayama, G. Benitez-garcia and H. Takahashi: "Skeleton-based Online Sign Language Recognition Using Monotonic Attention", Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp.601-608 (2022)
6) 梶山岳士，遠藤伶，加藤直人，河合吉彦，金子浩之：“深層学習を用いた日本手話認識の評価実験”，2019年映情学年次大，11B-2（2019）
7) 加藤直人，宮崎太郎，井上誠喜，内田翼，東真希子，梅田修一，比留間伸行，長嶋祐二：“手話CGアニメーションの口型作成システム”，HCGシンポジウム，HCG2015-A-3-2，pp.111-116（2015）
8) R. Pfau and J. Quer: "Nonmanuals: Their Grammatical and Prosodic Roles", In D. Brentari (Ed.), Sign Languages, Cambridge Language Surveys, pp.381-402, Cambridge University Press (2010)
9) 神田和幸：“手話の言語的特性に関する研究”，福村出版（2010）
10) 小薗江聡，木村晴美，市田泰弘：“日本手話における視線の文法化―目の開き方と眉の動きについて―”，日本手話学会第28回大会，pp.15-16（2002）
11) C. Schmidt, O. Koller, H. Ney, T. Hoyoux and J. Piater: "Using Viseme Recognition to Improve a Sign Language Translation System", Proceedings of the 10th International Workshop on Spoken Language Translation (2013)
12) 全国手話研修センター日本手話研究所：“新日本語-手話辞典”，中央法規出版（2011）
13) NPO手話技能検定協会：“一目でわかる実用手話辞典”，新星出版社（2007）
14) H. Zhou, W. Zhou, Y. Zhou and H. Li: "Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation", IEEE Trans. on Multimedia, **24**, pp.768-779 (2022)
15) R. Zuo and B. Mak: "C2SLR: Consistency-enhanced Continuous Sign Language Recognition", IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5121-5130 (2022)
16) W. Qin, X. Mei, Y. Chen, Q. Zhang, Y. Yao and S. Hu: "Sign Language Recognition and Translation Method based on VTN", International Conference on Digital Society and Intelligent Systems, pp.111-115 (2021)
17) Y. Tian: "Evaluation of Face Resolution for Expression Analysis", Conference on Computer Vision and Pattern Recognition Workshop, pp.82-82 (2004)
18) K. Murakami and H. Taguchi: "Gesture Recognition using Recurrent Neural Networks", in proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.237-242 (1991)
19) J. Kim, W. Jang and Z. Bien: "A Dynamic Gesture Recognition System for the Korean Sign Language (KSL) ", IEEE Trans. Syst., Man, & Cybernetics, Part B (Cybernetics), **26**, 2, pp.354-359 (Apr. 1996)
20) M.B. Waldron and S. Kim: "Isolated ASL Sign Recognition System for Deaf Persons", IEEE Transactions on Rehabilitation Engineering, **3**, 3, pp.261-271 (Sep. 1995)
21) J. Zieren and K. Kraiss: "Non-intrusive Sign Language Recognition for Human Computer Interaction", in proceedings of IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design and Evaluation of Human Machine Systems (2004)
22) S. Liwicki and M. Everingham: "Automatic Recognition of Fingerspelled Words in British Sign Language", IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp.50-57 (2009)
23) N. Michael, C. Neidle and D. Metaxas: "Computer-based Recognition of Facial Expressions in ASL: from Face Tracking to Linguistic Interpretation", Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, pp.164-167 (2010)
24) M. Mukushev, A. Sabyrov, A. Imashev, K. Koishybay, V. Kimmelman and A. Sandygulova: "Evaluation of Manual and Non-manual Components for Sign Language Recognition", International Conference on Language Resources and Evaluation (2020)
25) C. Zhang, Y. Tian and M. Huenerfauth: "Multi-modality American Sign Language Recognition", IEEE International Conference on Image Processing, pp.2881-2885 (2016)
26) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin: "Attention is All You Need", arXiv preprint (2017)
27) J. Donahue et al: "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description", IEEE Transactions on Pattern Analysis and Machine Intelligence, **39**, 4, pp.677-691 (Apr. 1, 2017)
28) H. Reza, V. Joze and O. Koller, MS-ASL: "A Large-Scale Data Set and Benchmark for Understanding American Sign Language", arXiv preprint (2018)
29) Ye-Yi Wang, A. Acero and C. Chelba: "Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy", IEEE Workshop on Automatic Speech Recognition and Understanding, pp.577-582 (2003)
30) A. Graves, S. Fernández, F. Gomez and J. Schmidhuber: "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks", Proceedings of the 23rd International Conference on Machine Learning, pp.369-376 (2006)
31) N.C. Camgoz, S. Hadfield, O. Koller, H. Ney and R. Bowden: "Neural Sign Language Translation", IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.7784-7793 (2018)
32) J. Huang, W. Zhou, Q. Zhang, H. Li and W. Li: "Video-based Sign Language Recognition without Temporal Segmentation", Proceedings of AAAI Conference on Artificial Intelligence, pp.2257-2264 (2018)
33) H. Wang, X. Chai, X. Hong, G. Zhao and X. Chen: "Isolated Sign Language Recognition with Grassmann Covariance Matrices", ACM Transactions on Accessible Computing, **8**, 4, pp.1-21 (2016)

梶山（かじやま）　岳士（たけし）　2003年，電気通信大学大学院電気通信学研究科修了．同年，NHKに入局．札幌放送局，NHK放送技術研究所を経て，2022年より，NHK財団に勤務．薄型光ディスク，スーパーハイビジョン記録装置，手話CGの研究に従事．正会員．

鹿喰（ししくい）　善明（よしあき）　1983年，東京大学大学院工学系研究科修了．同年，NHK入局．1986年より，同放送技術研究所にて，ディジタル信号処理，ハイビジョンの伝送方式，映像圧縮符号化，スーパーハイビジョン放送，IP放送，新サービスに関する研究に従事．2014年より，明治大学総合数理学部教授．博士（工学）．当会フェロー認定会員．

[機械翻訳コンテンツの著作権について]
当サイトに掲載されている原著論文の著作権は映像情報メディア学会に帰属します。原著論文および翻訳論文の無断使用は禁止します。
詳細は映像情報メディア学会著作権規定をご覧ください。

Machine Translated by Google

The results showed that the accuracy of the analysis of modifiers expressed by the shape of the eyes and eyebrows was improved by using high-resolution images. Experiments on sign language sentence recognition showed that the addition of mouth image analysis did not adversely affect the accuracy of the analysis of words other than homoglyphs. These results demonstrate that the performance of sign language recognition can be improved by analyzing high-resolution mouth images in conjunction with upper body images.

In the future, in parallel with improvements to the recognition processing technology itself using deep learning, it will be necessary to further consider the nature of the input image, such as the spatiotemporal resolution of each part for proper recognition.

We would like to thank the NHK Science and Technology Research Laboratories for their cooperation in allowing us to use the sign language video.

[References]

1) World Federation of the Deaf, https://wfdeaf.org
2) 日本手話通訳士協会, http://www.jasli.jp
3) N. Hiruma, M. Azuma, T. Uchida, S. Umeda, T. Miyazaki, N. Kato and S. Inoue: "Automatic Generation System of Japanese Sign Language (JSL) with CG Animation of Fixed Pattern Weather Information", ABU Technical Journal **264**, pp.2-5 (2015)
4) B. Kwolek, W. Baczynski and S. Sako: "Recognition of JSL Fingerspelling Using Deep Convolutional Neural Networks", Neurocomputing (June 2021)
5) N. Takayama, G. Benitez-garcia and H. Takahashi: "Skeleton-based Online Sign Language Recognition Using Monotonic Attention", Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp.601-608 (2022)
6) 梶山岳士，遠藤伶，加藤直人，河合吉彦，金子浩之：“深層学習を用いた日本手話認識の評価実験”，2019年映情学年次大，11B-2（2019）
7) 加藤直人，宮崎太郎，井上誠喜，内田翼，東真希子，梅田修一，比留間伸行，長嶋祐二：“手話CGアニメーションの口型作成システム”，HCGシンポジウム，HCG2015-A-3-2，pp.111-116（2015）
8) R. Pfau and J. Quer: "Nonmanuals: Their Grammatical and Prosodic Roles", In D. Brentari (Ed.), Sign Languages, Cambridge Language Surveys, pp.381-402, Cambridge University Press (2010)
9) 神田和幸：“手話の言語的特性に関する研究”，福村出版（2010）
10) 小薗江聡，木村晴美，市田泰弘：“日本手話における視線の文法化―目の開き方と眉の動きについて―”，日本手話学会第28回大会，pp.15-16（2002）
11) C. Schmidt, O. Koller, H. Ney, T. Hoyoux and J. Piater: "Using Viseme Recognition to Improve a Sign Language Translation System", Proceedings of the 10th International Workshop on Spoken Language Translation (2013)
12) 全国手話研修センター日本手話研究所：“新日本語-手話辞典”，中央法規出版（2011）
13) NPO手話技能検定協会：“一目でわかる実用手話辞典”，新星出版社（2007）
14) H. Zhou, W. Zhou, Y. Zhou and H. Li: "Spatial-Temporal Multi-Cue Network for Sign Language Recognition and Translation", IEEE Trans. on Multimedia, **24**, pp.768-779 (2022)
15) R. Zuo and B. Mak: "C2SLR: Consistency-enhanced Continuous Sign Language Recognition", IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.5121-5130 (2022)
16) W. Qin, X. Mei, Y. Chen, Q. Zhang, Y. Yao and S. Hu: "Sign Language Recognition and Translation Method based on VTN", International Conference on Digital Society and Intelligent Systems, pp.111-115 (2021)
17) Y. Tian: "Evaluation of Face Resolution for Expression Analysis", Conference on Computer Vision and Pattern Recognition Workshop, pp.82-82 (2004)
18) K. Murakami and H. Taguchi: "Gesture Recognition using Recurrent Neural Networks", in proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.237-242 (1991)
19) J. Kim, W. Jang and Z. Bien: "A Dynamic Gesture Recognition System for the Korean Sign Language (KSL) ", IEEE Trans. Syst., Man, & Cybernetics, Part B (Cybernetics), **26**, 2, pp.354-359 (Apr. 1996)
20) M.B. Waldron and S. Kim: "Isolated ASL Sign Recognition System for Deaf Persons", IEEE Transactions on Rehabilitation Engineering, **3**, 3, pp.261-271 (Sep. 1995)
21) J. Zieren and K. Kraiss: "Non-intrusive Sign Language Recognition for Human Computer Interaction", in proceedings of IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design and Evaluation of Human Machine Systems (2004)
22) S. Liwicki and M. Everingham: "Automatic Recognition of Fingerspelled Words in British Sign Language", IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp.50-57 (2009)
23) N. Michael, C. Neidle and D. Metaxas: "Computer-based Recognition of Facial Expressions in ASL: from Face Tracking to Linguistic Interpretation", Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, pp.164-167 (2010)
24) M. Mukushev, A. Sabyrov, A. Imashev, K. Koishybay, V. Kimmelman and A. Sandygulova: "Evaluation of Manual and Non-manual Components for Sign Language Recognition", International Conference on Language Resources and Evaluation (2020)
25) C. Zhang, Y. Tian and M. Huenerfauth: "Multi-modality American Sign Language Recognition", IEEE International Conference on Image Processing, pp.2881-2885 (2016)
26) A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin: "Attention is All You Need", arXiv preprint (2017)
27) J. Donahue et al: "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description", IEEE Transactions on Pattern Analysis and Machine Intelligence, **39**, 4, pp.677-691 (Apr. 1, 2017)
28) H. Reza, V. Joze and O. Koller, MS-ASL: "A Large-Scale Data Set and Benchmark for Understanding American Sign Language", arXiv preprint (2018)
29) Ye-Yi Wang, A. Acero and C. Chelba: "Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy", IEEE Workshop on Automatic Speech Recognition and Understanding, pp.577-582 (2003)
30) A. Graves, S. Fernández, F. Gomez and J. Schmidhuber: "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks", Proceedings of the 23rd International Conference on Machine Learning, pp.369-376 (2006)
31) N.C. Camgoz, S. Hadfield, O. Koller, H. Ney and R. Bowden: "Neural Sign Language Translation", IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.7784-7793 (2018)
32) J. Huang, W. Zhou, Q. Zhang, H. Li and W. Li: "Video-based Sign Language Recognition without Temporal Segmentation", Proceedings of AAAI Conference on Artificial Intelligence, pp.2257-2264 (2018)
33) H. Wang, X. Chai, X. Hong, G. Zhao and X. Chen: "Isolated Sign Language Recognition with Grassmann Covariance Matrices", ACM Transactions on Accessible Computing, **8**, 4, pp.1-21 (2016)

梶山　岳士（かじやま　たけし）　2003年，電気通信大学大学院電気通信学研究科修了．同年，NHKに入局．札幌放送局，NHK放送技術研究所を経て，2022年より，NHK財団に勤務．薄型光ディスク，スーパーハイビジョン記録装置，手話CGの研究に従事．正会員．

鹿喰　善明（ししくい　よしあき）　1983年，東京大学大学院工学系研究科修了．同年，NHK入局．1986年より，同放送技術研究所にて，ディジタル信号処理，ハイビジョンの伝送方式，映像圧縮符号化，スーパーハイビジョン放送，IP放送，新サービスに関する研究に従事．2014年より，明治大学総合数理学部教授．博士（工学）．当会フェロー認定会員．