一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2023-52(2024-03)

未知手話単語獲得に向けた手話動作特徴量空間の検討

山田大記 [a)†]　井上勝文 [b)†]　Partha Pratim Roy [c)††]　岩村雅一 [b)†]　吉岡理文 [b)†]

† 大阪公立大学大学院情報学研究科，〒 599-8531 堺市中区学園町 1-1
†† Department of Computer Science and Engineering, Indian Institute of Technology Roorkee
Roorkee-247667 Uttarakhand, India
E-mail: [a)]sb22664w@st.omu.ac.jp,
[b)]{inoue, masa.i, ysok}@omu.ac.jp, [c)]proy.fcs@iitr.ac.in

あらまし　近年，様々な手法が提案されている手話認識研究において大きな課題となるのが，データセット内の手話動画の語彙の不足である．この課題に対し，特に未学習の手話単語 (未知単語) のデータに対して自動で学習動画の収集とラベル付けする手法が求められている．幸い Youtube 等に多数字幕情報のついた手話動画が存在するため，これらの内，特定の未知単語が含まれる複数の動画より類似する動作を抽出することで，未知単語を自動獲得できるのではないかと期待されている．これを実現するために必要となりえる要素技術は，(1) 類似の手話動作を精確に表現できる動作特徴量と，(2) その特徴量に基づいて，手話動画から目的の単語の部分を精確に切り出す処理である．本研究では，これの実現の第一歩として，手話単語の細かな動作を一つ一つ記述可能な 3DCNN である I3D を用いて手話単語を表現する特徴量について検討する．本稿では，この特徴量表現の良し悪しを評価するテストヘッドとして，既知単語と未知単語を判別する問題を考え，特徴量の識別性能について評価した結果について述べる．

キーワード　手話単語認識，3DCNN，I3D，クラスタリング，WLASL，時空間特徴量

1. はじめに

手話とは，聴覚障害者にとって，手や指の動きや顔の表情を組み合わせることによって物事の意味や内容を伝え，他者とのコミュニケーションを取るための重要な手段の一つである．耳が聞こえる健聴者が聴覚障害者とコミュニケーションをとるためには，手話の意味を理解する必要がある．しかし，手話は語彙数が多く，表現が複雑であることから，その習得には多くの時間と労力が必要である．

手話が理解できない健聴者が聴覚障害者の手話を介してコミュニケーションをとる手段の 1 つとして，手話認識が研究されている．このような手話認識の実現において課題となるのが，データセットに含まれる手話動画の語彙の不足である．代表的な大規模手話単語データセットである WLASL [1] には，2000 単語の手話動画が含まれる．一方で，日常的に使用される手話単語は数千単語規模と言われており [1]，このような既存のデータセットを用いるだけでは不十分である．さらに，手話は一種の言語であり，日々新たな手話単語が生み出されている．未知単語の動画の収集とラベル付けを手動で行う方法には自ずと限界があるため，その自動化が望まれる．

ラベル付けの自動化に関連した従来手法として，手話ニュースのように連続的に手話を行う動画から，学習済みの手話単語 (既知単語) の位置を特定する手法（既知単語の Spotting）[2–4] がある．これらの手法は既知単語と特徴量が合致する箇所を動画中から探し出すので，既知単語のデータを増やすことはできるが，未知単語のデータを増やすことは難しい．しかし，動画中から既知単語を探し出すのでは無く，未知単語の可能性がある箇所を探し出すようにすれば，未知単語が同定できるのでは無いか．つまり，特徴量が類似する手話を動画中から集める，いわばクラスタリングのような処理で未知単語を同定できるのでは無いか，というのが本研究の基本的なアイディアである．しかし，この方法では類似した手話を集める事は出来ても，その手話が何を表す単語なのかというラベルを付与出来ない．

そこで本研究では字幕情報が付いた手話動画や手話ニュースの利用を念頭に置き，未知単語の自動ラベル付けを目指す．一般に字幕情報は単語単位ではなく，文単位あるいは文の一部（複数単語）毎に付けられているため，未知の手話動作と字幕中の単語（未知単語）は一意に対応付かない．しかし，複数の動画の字幕情報を勘案することで，未知単語の意味を特定し，未知単語のデータを増やすことを期待する．

本稿では，手話の未知単語の自動獲得を目指す上で必要となり得る要素技術の確立を目指し，その上で手話の未知単語の自動獲得の実現可能性や課題を検討する．前述の目標を達成するには，以下の要素技術が重要となる．すなわち，(1) 類似の手話動作を抜き出すために，動作を精確に表現できる特徴量と，(2) その特徴量に基づいて，手話動画から目的の単語の部分を精確に切り出す処理である．まず前記 (1) に関して，手話単語は類似する細かな動作の組み合わせで表現されるため，一つ一つの動作の情報を特徴量として記述できる 3DCNN である I3D [5] を用いる．本稿では，この特徴表現の善し悪しを評価す

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2024 by IEICE

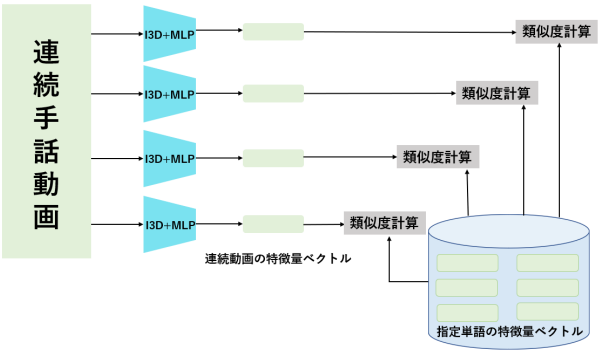

図 1: Momeni らの Spotting 手法の概要 [4]

るテストベッドとして，既知単語と未知単語を判別する問題を考える．これらが精確に区別できるならば，既知単語と未知単語だけでなく，未知単語同士も十分に区別出来ると考えられる．次に前記 (2) に関して，類似の手話動作を同定できる処理として，手話動作の特徴量のクラスタリングを検討する．クラスタリングの結果，同一の意味を表す手話動作がどの程度集められるかを調査することで，このアプローチの可能性を検討する．

本稿で行う実験には，手話単語動画データセットである WLASL [1] を用いる．これは単語毎に切り出された手話動作の動画が複数用意されているタイプのデータセットであり，前述の手話ニュースのような，複数単語の手話動作が連続しているものでも，字幕情報が付いたものでも無い．したがって，本稿はあくまでも未知単語の自動獲得を目指して，現在位置を確認するものであることに留意いただきたい．

2. 関 連 研 究

本節では，まず，本研究と似た問題設定である手話の Spotting 手法について述べる．CV 分野の手話に関連する研究は，手話認識 [6–9]・手話翻訳 [10–12]・Spotting [2–4, 13] の 3 つの分野に大きく分けられる．手話認識とは，手話動画を入力として，コンピュータが理解できる形への手話動作の変換を指す．手話翻訳とは，手話動画を入力として，手話から口語言語への翻訳を指す．Spotting は，特定の単語の手話 (既知単語のみ) が動画中のどこに位置するかを推定するタスクである．本研究では，Spotting に近い内容を扱うため，Spotting に関する研究を概観する．それに加えて，本研究で用いるモデルの元手法である，multi-stream neural networks を活用した高精度な手話単語認識手法 [8] と，従来手法である軌跡を用いた既知単語と未知単語の判定手法 [14] について述べる．

2.1 手話の Spotting 手法

連続手話動画の Spotting に関する 3 つの研究を取り上げる．

1 つ目は，手話動作の距離関数を導入した Buehler らの Spotting 手法 [2] である．この研究では，手話単語を表現する新たな距離関数を提案しつつ，字幕のようなノイズの多い教師情報に対応するために，Multiple Instance Learning を実施することで手話の Spotting を実現している．距離関数では，手の位置の距離や手の形状の距離，手の向きの差異を数式化する．これらの距離は手話動作の特徴を表現しており，各距離情報を重み付け加算することで 2 つの手話動画の距離を測っている．

また，あるターゲットの単語に対して，字幕にターゲットの単語が存在する動画を Positive bag，存在しない動画を Negative bag に分け，これらを用いて Multiple Instance Learning することで，ターゲットとなる単語手話の特徴を学習している．以上の 2 つの手法により，ターゲットの単語位置を特定している．

2 つ目は，手話の複合的な特徴量と SVM を用いた Multiple Instance Learning による Pfister らの Spotting 手法 [3] である．Pfister らは，手話において重要な部位である顔や手の位置情報と，口領域から抽出した SIFT 特徴量 [15] をフレーム毎に抽出し，手話の複合的な特徴量としている．また，MI-SVM [16] とよばれる Multiple Instance Learning の思想をもとにした SVM によって，手話の特徴を学習する．これにより，口や顔，手の情報の特徴量が近い手話動作の部分を見つけ，手話位置を特定する．この研究と Buehler らの研究 [2] は，使用する特徴量が異なるが，Multiple Instance Learning によって共通する単語の手話動作位置の特定を目指している点で共通している．

3 つ目は，Momeni らの Spotting 手法 [4] である．この研究では，BSLDict と呼ばれるイギリス手話単語動画データセットと BSL-1K と呼ばれる連続手話動画データセット用いて Multiple Instance Learning することで，連続手話動画に対する Spotting を可能にする手法が提案されている．この手法では，BSLDict と BSL-1K に存在する単語を 2 つの Positive bag と Negative bag に分け，これらを用いて Multiple Instance Learning することで手話単語の動作特徴量を学習している．そして，図 1 に示すように，この動作特徴量とある連続手話動画から抽出された動作特徴量とのコサイン類似度を測ることで，最も類似度の高い部分をその手話単語の動作位置として Spotting している．

上述の Spotting 手法 [2–4] は，いずれも連続手話動画内の既知単語の手話位置を高精度に特定できるものの，未知単語の位置は考慮されていない．そこで本研究では，未知単語の位置特定に向けて対応方法を検討する．

2.2 multi-stream neural networks を活用した高精度な手話単語認識手法

本研究で用いる手話単語認識モデルである，multi-stream neural networks を活用した手話単語認識手法 (MSNN) [8] について述べる．MSNN は，顔画像の局所領域により着目することで顔認識精度が向上するという報告 [17] を参考に，手話において重要である手や顔といった局所領域に着目した手法である．また，背景の影響が少ない骨格情報も含めて様々な情報を統合することで手話単語を認識する．具体的には，手話話者の上半身を映した全体画像や話者動作を表現するオプティカルフロー画像に加え，OpenPose [18] で抽出した左手画像，右手画像，顔画像と骨格情報も用いて，各入力情報から推定した単語認識結果を統合し，WLASL において高精度な単語認識精度を達成している．しかし，MSNN では，OpenPose を用いて局所領域や骨格情報を抽出しているが，WLASL のように比較的ゆっくり手話が表現される動画と異なり，手話ニュースのように高速で手話が表現される動画では，ブレ等の影響により局所

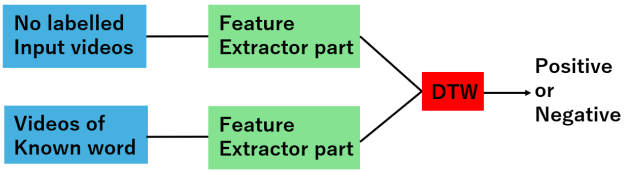

図 2: 軌跡を用いた既知単語・未知単語判定手法 [14] の概要図

領域や骨格情報の抽出精度が低くなる問題がある．そこで本研究では，MSNN のうち全体画像とオプティカルフロー画像の stream のみを用いる．つまり，本研究では全体画像とオプティカルフロー画像から I3D を用いて抽出した特徴量を元に，手話単語獲得に向けた特徴量を検討する．

2.3 軌跡を用いた既知単語・未知単語の判定手法

本節では，軌跡による手話単語の既知・未知判定する従来手法 [14] について述べる．判定手法の流れを図 2 に示す．この手法では，図 2 に示すようにまず手話単語動画より時空間特徴量を抽出し，各単語間の距離を Dynamic Time Warping (DTW) [19] で求めて判定する．具体的な処理を以下で述べる．

まず特徴抽出部について述べる．特徴抽出部は，手話動画のフレームをずらしながら I3D+Multilayer Perceptron(MLP) モデル [4] から時空間特徴量を抽出する．そして，この時空間特徴量を時系列順に並べて軌跡を作成する．これにより，手話動作の時系列情報を保持した手話単語の軌跡が作成可能となる．本稿では，以後この軌跡を軌跡特徴量と呼ぶ．

次に，判定部について述べる．上述の方法で求めた軌跡特徴量は手話単語動画ごとに長さが異なる時系列データである．このため [14] では，長さの異なる時系列データの類似度を測ることに適した手法である DTW を用いて，手話動画から作成した軌跡特徴量の類似度を測定し，以下で述べるように既知単語か未知単語かを判定する．

まず，調べたい単語の軌跡特徴量と学習データにある全ての既知単語の軌跡特徴量との距離 (軌跡間距離) を DTW によって算出する．次に，算出される軌跡間距離は各動画のフレーム数の影響を大きく受けるため，フレーム数で正規化する．具体的には，2 つの動画のフレーム数の平均の平方根で軌跡間距離を割る．これにより，フレーム数の違いにより算出される軌跡間距離に差が出るという問題を軽減する．最後に，算出した軌跡間距離の中での最小値を求め，閾値による既知・未知を判定する．この最小値が閾値より小さい場合は，既知単語の軌跡と類似した手話の軌跡特徴量であると考え，その手話単語動画は既知単語であると判定する．一方，閾値より大きい場合は，既知単語の軌跡特徴量とは類似しない手話の軌跡特徴量であると考え，その手話単語動画は未知単語であると判定する．

この従来手法では，手話単語の既知・未知を判定できる場合もあるが，同じ単語が表現されていても，手の形や表現速度等の話者間の細かな違いの影響を吸収できず，既知・未知判定が約 52%とランダムよりも少し良い程度の精度しか達成できなかったことが報告されている [14]．そこで本研究では，この問題に対し，クラスタリングを用いることで上述の細かな差異を吸収する時空間特徴量について検討する．

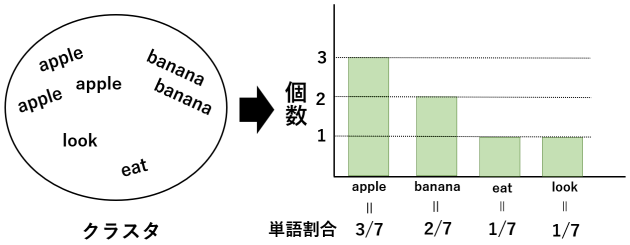

図 3: 単語割合ベクトルの作成方法

3. 提 案 手 法

本節では提案手法を構成する (1) 手話単語特徴量の作成方法，(2) 既知単語・未知単語の判定方法について順に説明する．

3.1 手話単語特徴量の作成方法

本節では，手話単語特徴量の作成方法について述べる．従来手法では十分に活用できなかった手話動作の特徴を活用するために，本研究では k-means 法を活用したクラスタリングをもとに手話単語特徴量を作成する．手話単語特徴量の作成方法は，4 つの手順からなり，以下でそれぞれ説明する．

まず，学習用の手話単語動画を一定フレームごとに分割し，各分割動画を既存の I3D モデル [8] によって特徴量ベクトルに変換する．特徴量ベクトルは，全体画像の入力によって得られる時空間特徴量とオプティカルフロー画像の入力によって得られる時空間特徴量を連結したものである．これにより，手話動作の視覚的特徴と動きの特徴を組み合わせた特徴量ができる．

次に，学習用の手話単語動画の特徴量ベクトルを全て用いて k-means 法でクラスタリングする．これにより，似た動作の特徴が集まったクラスタを作成できる．

そして，図 3 に示すように，各クラスタ内の単語の個数を数え，各クラスタ内に存在する単語割合を表現したベクトル (単語割合ベクトル) を作成する．このような単語割合ベクトルを作成することで，各クラスタ内に分布する手話単語の出現頻度が分かる．つまり，各クラスタがどの手話単語の動作が中心となって作られているかを表現できる．例えば，あるクラスタ内に図 3 のような単語の特徴量ベクトルが含まれているとする．このとき，“apple” の単語割合はクラスタ内の単語の総数が 7 個であるのに対して，“apple” は 3 個存在するため，3/7 となる．最終的には，各単語ごとの単語割合を求め，ベクトルとして表現する．もし，学習用の手話単語動画の語彙数が 100 の場合，この作成方法によって得られるベクトルは 100 次元となる．また，300 個のクラスタに分けた場合，300 個の単語割合ベクトルが作成される．

最後に，図 4 に示すように単語割合ベクトルを使ってテスト用の手話単語動画の手話単語特徴量を作成する．具体的には，まずテスト用の手話単語動画を分割した各分割動画を特徴量ベクトルに変換し，最も近いクラスタに割り当てる．次に，割り当てられた各クラスタに対応する単語割合ベクトルを順に足していく．そして，合計された単語割合ベクトルをフレーム数に

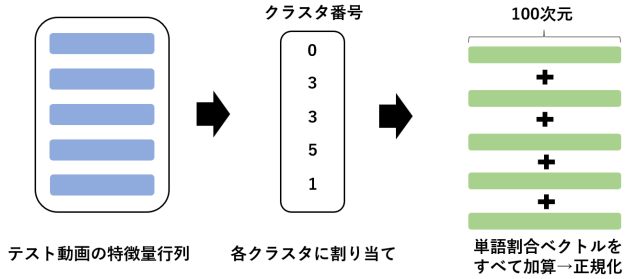

図 4: 手話単語特徴量の作成方法

よって正規化する．具体的には，テスト用の手話単語動画の分割動画数に対する平方根で割ることで正規化をする．これにより，動画のフレーム数の違いによる影響を弱めることができる．本研究では，この正規化したものを手話単語特徴量とする．このように手話単語特徴量を作成することで，手話動画がどの既知手話単語の要素を多く含んでいるかを表現できる．

3.2 既知単語・未知単語の判定方法

本研究では，上述の手話単語特徴量を基に，以下の 2 つ手順から既知単語・未知単語を判定する．まず，テスト動画の手話単語特徴量の要素の中で最大値を検索する．手話単語特徴量は既知単語による各クラスタの単語割合ベクトルを足し合わせたものであり，手話単語が同じであればその手話単語を多く含む単語割合ベクトルに割り当てられると考えられる．このため，手話単語特徴量内の各要素は学習データの各単語に対する確信度 ω を表現しているといえる．これより，確信度の最大値 $\omega_{\max}$ を用いることで既知単語・未知単語を判定することを考える．つまり，$\omega_{\max}$ が大きいほど，入力データが既知単語の可能性が高いといえる．一方で，$\omega_{\max}$ が小さいほど，既知単語と近しい特徴量が存在しないと考えられるため，入力データが未知単語の可能性が高いと判断できる．本研究では $\omega_{\max}$ に閾値 θ を設定し，$\omega_{\max} > \theta$ の場合は既知単語，$\omega_{\max} \leqq \theta$ の場合は未知単語と判定する．なお，この ω を用いることで，既知単語の認識にも応用でき，本研究ではこの手話単語特徴量が既知単語の認識にも有効であるのか検証する．

4. 実　　験

4.1 データセット

複数存在する手話単語認識のためのデータセットの中から，本研究では動画数，手話話者数，各クラス当たりの動画数の規模が大きい WLASL データセット [1] を使用する．WLASL の中には WLASL100，WLASL300，WLASL1000，WLASL2000 というサブセットがあり，各サブセット名に含まれている数字はそのサブセット内に含まれる手話単語の語彙数を表している．ここで，WLASL の各サブセットはクラス当たりの動画数が多い順にクラスをソートしたときの上位 $K(=\{100, 300, 1000, 2000\})$ 位のクラスで構成される．表 1 に各サブセットの詳細を示す．本研究では，このサブセットの内，WLASL100, WLASL300, WLASL2000 を主に用いる．また，データの分割に関しては，文献 [1] の分割方法に従う．つまり WLASL の各サブセット内のデータの割合は (学習 : 検証 : テスト) = (4 : 1 : 1) とする．

表 1: Details of datasets. Column “Mean” denotes the average number of videos per class.

Subset	#Class	#Video	Mean	#Signer
WLASL100	100	2,038	20.4	97
WLASL300	300	5,117	17.1	109
WLASL1000	1,000	13,168	13.2	116
WLASL2000	2,000	21,083	10.5	119

表 2: Baseline and Proposed methods accuracies (%).

	WLASL100			WLASL300		
Method	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10
Baseline	44.19	63.57	70.54	22.46	38.02	47.90
Proposed	62.40	81.01	87.98	50.90	65.27	71.26

本研究では，このサブセットのデータ分割を踏まえて以下のように既知単語と未知単語を設定する．WLASL100 の学習済みモデルを使用する場合，既知単語は WLASL100 のテストデータ，未知単語は WLASL2000 に含まれかつ WLASL100 に存在しない全データである．WLASL300 の学習済みモデルを使用する場合も同様に既知単語と未知単語を設定する．

4.2 既知単語の認識実験

提案手法の手順に従って作成した手話単語特徴量の性能を評価するために， WLASL100，WLASL300 を用いて既知単語の認識実験を行った．手法の精度評価には，Top-N 分類正解率を用いる．ここで，$N = \{1, 5, 10\}$ とする．本実験では，提案手法内のクラスタリング手法として k-means 法を用いる．本実験の結果としては，k を 100 から 1000 まで変更した予備実験で最も認識精度の高かった $k = 600$ の結果を示す．また，MSNN [8] で用いた I3D によって特徴量を 16 フレームごとに抽出し，この特徴量をもとに既知単語認識精度を評価した．

本実験では，比較対象として軌跡特徴量を用いた手法 [14] と既知単語の認識精度を比較した．軌跡特徴量を用いた手法の手話単語認識方法は以下の通りである．まず，調べたい手話単語動画の軌跡特徴量とすべて学習用の手話単語動画の軌跡特徴量との距離を DTW によって全て計算する．このうち，距離が最も小さい手話単語動画のラベルをその調べたい手話単語動画のラベルであると認識する．これをもとに，Top-N 分類正解率を求め，提案手法との認識精度を比較する．

WLASL100，WLASL300 での手話単語認識結果を表 2 に示す．どちらのサブセットでも Top-1 の認識精度に従来手法と比較して 20%程の開きがあった．これより，提案手法で作成した手話単語特徴量には手話単語の差異を表現し，従来手法である軌跡特徴量と比べて精度の高い既知単語認識ができていることが分かる．

4.3 既知単語・未知単語判定実験

本節では，既知単語・未知単語判定実験について述べる．手法の精度評価には，TP，FN，TN，FP を用いる．TP は既知単語を正しく判定できた個数，FN は既知単語を誤判定した個数，

表 3: Detection accuracy (%) of 100 known words in WLASL100 and 1900 unknown words in WLASL2000, excluding the 100 known words.

Method	TP	FN	TN	FP	Accuracy	Recall	Singularity
Baseline	145	113	10303	8754	54.09	56.20	54.06
Proposed	160	98	12395	6662	65.00	62.02	65.04

表 4: Detection accuracy (%) of 300 known words in WLASL300 and 1700 unknown words in WLASL2000, excluding the 300 known words.

Method	TP	FN	TN	FP	Accuracy	Recall	Singularity
Baseline	331	337	8256	7721	51.59	49.55	51.67
Proposed	387	281	10233	5744	63.80	57.93	64.05

TN は未知単語を正しく判定できた個数，FP は未知単語を誤判定した個数を表す．そして，これらの数値をまとめた指標として，正解率 (Accuracy)，再現率 (Recall)，特異度 (Singularity) の 3 つを用いる．正解率は既知単語と未知単語を正しく判定できた割合を，再現率は全既知単語のうち正しく既知単語と判定できた割合を，特異度は全未知単語のうち正しく未知単語と判定できた割合を表す．なお，閾値を変化させていき，最も精度の高い閾値の結果を表に載せている．

WLASL100 と WLASL300 を既知単語とした場合の実験結果をそれぞれ表 3 と 4 に示す．WLASL100 においては従来手法と比較して，どの指標においても提案手法が上回っていることが確認できる．これは，手話単語特徴量を用いることで，手話動作のクラスタの結果を集約することができ，既知単語・未知単語の判定に有効であったと考えられる．

4.4 未知単語のクラスタリング実験

本節では，提案する手話単語特徴量が，未知単語であっても同じ単語であれば類似する特徴量を持っているのかを調査した結果について述べる．具体的には，本研究では未知単語の動画から抽出された手話単語特徴量のみでクラスタリングした際に，同じ単語の手話単語特徴量がどれほど同じクラスタに属しているのかを調査する．これを調査するにあたり，本実験では，WLASL100 に存在せず，WLASL300 に存在する 200 単語を未知単語とした．そして，これらの単語の全動画から手話単語特徴量を作成し，語彙数と同じ $k = 200$ で k-means クラスタリングを実施した．また本実験では，対象未知単語の動画から抽出された手話単語特徴量がどのクラスタに属しているのかを調べ，そのクラスタ番号の出現回数を数えた．その上で，対象未知単語の全手話単語特徴量がどれほど出現回数上位 3 つのクラスタに属しているのか，その割合を調べた．

実験の結果を図 5 に示す．図 5 における横軸は，上述の割合であり，縦軸はその割合に属する単語数である．割合が大きいほど同じクラスタに対象未知単語の手話単語特徴量が集中していることを示す．結果から，未知単語の多くが上位 3 つのクラスタに手話単語特徴量が集まっていることが確認できる．また，図 6 と 7 にそれぞれ，未知単語の手話単語特徴量のクラス

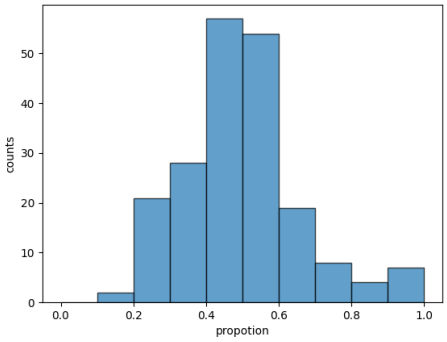

図 5: Clustering results of unknown words

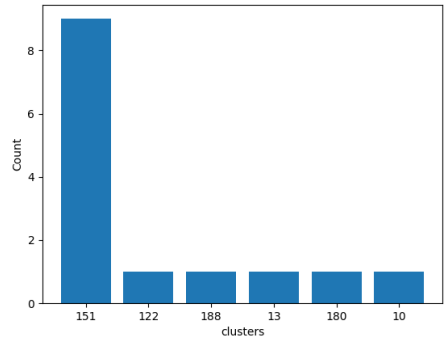

図 6: Successful Clustering Example (Word: Internet)

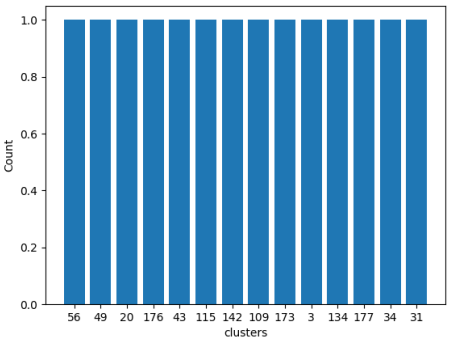

図 7: Failed Clustering Example (Word: humble)

内分散の小さいもの (成功例) と大きいものの (失敗例) ものを示す．ここで，横軸は手話単語特徴量が属するクラスタ番号，縦軸はそのクラスタ番号に属している対象未知単語の手話単語特徴量数を表している．図 6 より，“Internet” という単語は 1 つのクラスタに対してほとんどの手話単語特徴量が集中していることが確認できる．一方で，図 7 では “humble” という単語の手話単語特徴量が多数のクラスタに分散していることが分かる．ここで，humble の手話単語動画を確認したところ，動作のスピードが話者によって異なっていることが分かった．この動作スピードの違いが多数のクラスタに分散した原因であると考えられる．このため，今後はよりスピード等に考慮した手話単語特徴量をより考える必要があろう．

5. 考　　察

前節で行った実験を通じて，本研究で目指す未知単語の自動獲得に向けた課題を論じる．

本稿で用いた I3D の特徴量は，識別能力が比較的高い．しかし，既知単語と未知単語の判定に用いても，必ずしも優れた性能を発揮出来なかった．このことから，他の候補よりもスコアが高くなれば正解となる手話認識に比べて，距離が一定値以下か以上かで判定する既知単語と未知単語の判定は，より挑戦的なタスクである可能性がある．一方で，既知単語と未知単語の判定基準にニューラルネットワークなどの，より洗練された方法を使うことで，その性能が大きく改善される可能性もある．

本稿で用いなかった字幕情報の付いた手話動画を用いる影響として，以下の事が考えられる．これらの動画は手話が連続して行われるため，手話と手話の切れ目を同定することが求められる．類似する手話動作を集める際には，字幕情報を用いて候補を絞り込みできる事が性能改善の一助になる可能性もある．

6. おわりに

本研究では，自動での未知単語獲得を目指し，手話単語特徴量という未知単語に対応した特徴量を提案した．具体的には，手話単語の特徴を表現するために，動作特徴によるクラスタリング結果を単語割合ベクトルとして表現し，この単語割合ベクトルを割り当てられたクラスタごとに集計することで手話単語特徴量を作成した．実験より，提案手法である手話単語特徴量が従来手法よりも既知単語の認識精度と既知単語・未知単語の判定精度の両方を向上させることが確認できた．また，この手話単語特徴量を用いて未知単語のクラスタリングを行った結果，同じ手話単語の特徴量が少数のクラスタに集まっている様子を確認できた．一方で，同じ単語であっても多数のクラスタに分散している場合が見られた．これは手話動作のスピードの違いが多数のクラスタに分散した原因であると考えられる．

今後の課題として，動作スピードに配慮した特徴量の改良が挙げられる．また，このような手話単語特徴量と字幕情報を用いることによる未知単語の獲得手法の検討などが挙げられる．

謝辞　本研究は，JSPS 科研費#19K12023 ならびに 2023 年度 I-O DATA 財団研究助成の支援を受けて実施された．

文　　献

[1] D. Li, C. Rodriguez, X. Yu, and H. Li, “Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison,” Proc. of WACV, pp.1459–1469, 2020.

[2] P. Buehler, A. Zisserman, and M. Everingham, “Learning sign language by watching tv (using weakly aligned subtitles),” CVPR, pp.2961–2968, 2009.

[3] T. Pfister, J. Charles, and A. Zisserman, “Large-scale learning of sign language by watching tv (using co-occurrences),” British Machine Vision Conference, 2013. https://api.semanticscholar.org/CorpusID:9282404

[4] L. Momeni, G. Varol, S. Albanie, T. Afouras, and A. Zisserman, “Watch, read and lookup: learning to spot signs from multiple supervisors,” Proc. of ACCV, 2020.

[5] A.Z. Joao Carreira, “Quo vadis, action recognition? a new model and the kinetics dataset,” Proc. of CVPR, pp.6299–6308, 2017.

[6] Y. Min, A. Hao, X. Chai, and X. Chen, “Visual alignment constraint for continuous sign language recognition,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.11542–11551, Oct. 2021.

[7] A. Hao, Y. Min, and X. Chen, “Self-mutual distillation learning for continuous sign language recognition,” Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp.11303–11312, Oct. 2021.

[8] M. Maruyama, S. Ghose, K. Inoue, P.P. Roy, M. Iwamura, and M. Yoshioka, “Word-level sign language recognition with multi-stream neural networks focusing on local regions,” 2021.

[9] C.C. de Amorim, D. Macêdo, and C. Zanchettin, “Spatial-temporal graph convolutional networks for sign language recognition,” Proc. of ICANN, pp.646–657, 2019.

[10] N.C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, “Neural sign language translation,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[11] N.C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, “Sign language transformers: Joint end-to-end sign language recognition and translation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[12] A. Yin, T. Zhong, L. Tang, W. Jin, T. Jin, and Z. Zhao, “Gloss attention for gloss-free sign language translation,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.2551–2562, June 2023.

[13] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, “Scaling up sign spotting through sign language dictionaries,” International Journal of Computer Vision, vol.130, no.6, p.1416–1439, April 2022. http://dx.doi.org/10.1007/s11263-022-01589-6

[14] 山田, 大記，井上, 勝文，岩村, 雅一，P.P. Roy，吉岡, 理文，“時空間特徴量を用いた手話単語認識における未知単語判定手法の検討,” 情報処理学会研究報告コンピュータビジョンとイメージメディア (CVIM)，2022-CVIM-230，pp.156–163，2022.

[15] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision, vol.60, no.2, pp.91–110, 2004.

[16] S. Andrews, I. Tsochantaridis, and T. Hofmann, “Support vector machines for multiple-instance learning,” Proceedings of the 15th International Conference on Neural Information Processing Systems, p.577–584, NIPS’02, MIT Press, Cambridge, MA, USA, 2002.

[17] R. Adria, K. Petr, S. Simon, M. Wojciech, and T. Antonio, “Learning to zoom: a saliency-based sampling layer for neural networks,” Proc. of ECCV, pp.51–66, 2018.

[18] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y.A. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” Proc. of CVPR, pp.1302–1310, 2017.

[19] 櫻井保志，吉岡正俊，“ダイナミックタイムワーピングのための類似探索手法,” 情報処理学会論文誌，3 月 2004.