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

信学技報
TECHNICAL REPORT OF IEICE.

局所領域に着目したMulti-stream Neural Networksによる手話単語認識

丸山瑞己 [a)†]　Shuvozit Ghose[b)††]　井上勝文 [c)†]　Partha Pratim Roy[d)††]　岩村雅一 [c)†]
吉岡理文 [c)†]

† 大阪府立大学 大学院工学研究科，〒 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)]maruyama@sig.cs.osakafu-u.ac.jp, [b)]shuvozit.ghose@gmail.com,
[c)]{inoue, masa, yoshioka}@cs.osakafu-u.ac.jp, [d)]proy.fcs@iitr.ac.in

あらまし　近年，手話認識に関する研究が広く行われおり，様々なアプローチが提案されている．中でも，行動認識のタスクのために提案された I3D ネットワークを用いた手法は，大規模な手話認識データセットにおいて最も高い認識率を達成している．I3D を用いた従来手法では，話者の全身もしくは上半身の外観情報のみを観測しているが，手話認識では手の形状や顔の表情のような局所的な情報や，体と手の位置関係が重要な意味を持つ．そこで，本研究では，手話認識において重要な要素である局所的な情報に加えて，体に対する手の位置を捉えるために骨格情報を追加で用いる．すなわち，既存の I3D ネットワークに，局所領域画像パッチを入力するストリームと，骨格情報を入力とするストリームを加えた Multi-stream 構造のモデルによって，手話認識精度の向上を図る．大規模な手話データセットである WLASL を用いた実験の結果，提案手法は従来手法に比べ，一位認識率において約 15%の向上を達成した．

キーワード　手話認識，ニューラルネットワーク，深層学習，I3D，ST-GCN，WLASL，MS-ASL

1. はじめに

手話は，聴覚障害者にとって他者とのコミュニケーションを図るための重要な手段の一つである．しかし，手話は語彙が多く，表現が複雑であることから，習得に多くの時間と労力が必要で，誰しもが容易に習得できるものではない．そのため近年，聴覚障害者と健常者との間のコミュニケーションの壁を取り払うことを目的として，手話単語を自動で認識する研究が広く行われている [1]〜[3]．本研究では，よりコミュニケーションを円滑にするために手話単語認識精度の向上を目指す．手話単語認識では，素早く細かい手や体の動き，手の形状，顔の表情，体と手の位置関係など，多くの複雑な要素を捉える必要がある．さらに，手話の中には，同じような体の動きでも手の形状が違うことで異なる単語を表すものがあることから，手話単語認識は難しいタスクであると言える．

従来手法では，話者の全身もしくは上半身の外観情報を用いて手話単語を認識しているが，手話にとって重要な要素である手や顔の情報を捉えきれないという問題があると考えられる．現在，手話単語認識において最も高い精度を達成している手法は，行動認識タスクのために提案された I3D ネットワーク [4] を用いた手法 [2], [3] である．これは，I3D が画像の全体を観測することで理解できる行動認識タスクを得意としており，手話動画の全体的な時空間特徴を捉えることができた結果である．しかし，手話認識にとって重要な手の形状や顔の表情などの局所的な特徴や，体に対する手の位置という情報を十分に捉えられておらず，その精度は依然として低い．

そこで，本研究では，手と顔の局所領域や体と手の位置関係に着目した手話単語認識手法を提案する．顔認識の研究において，顔全体に加えて目などの局所領域に着目することで，認識精度が向上すると報告されている [5]．これに基づき，話者の全身もしくは上半身の外観情報だけでなく，手話認識において重要な手や顔の局所領域に着目する．それに加えて，体と手の位置関係を捉えるために骨格情報も導入する．この骨格情報の導入は，外観情報のみを用いた手法において危惧される背景や話者の見た目などの情報の影響を軽減することも期待でき，さらなる精度の向上が望める．

局所領域の情報と骨格情報を，画像全体の外観情報と効果的に組み合わせるために，Multi-stream Neural Networks を採用する．Multi-stream Neural Networks とは，画像の外観情報やオプティカルフローのような動きの情報など，異なる情報をそれぞれ別々のネットワークで学習し，各ネットワークからの出力を組み合わせることで認識結果が得られるものである [6]〜[8]．複数の情報が互いに補い合うことで，認識精度の向上が見られる．本研究では，従来の I3D ネットワークに，手と顔の局所領域画像を入力とするストリームと，骨格情報を入力とするストリームを加えた Multi-stream 構造のモデルにより，手話単語認識精度の向上を図る．

2. 関連研究

本節では，まず，従来の手話単語認識手法について述べる．

次に，本研究の提案手法に用いる 2 つのネットワーク，I3D と ST-GCN について述べる．

2.1 手話単語認識手法

初期の手話認識手法では，ハンドクラフト特徴量を用いている．手話の空間的な表現を捉えるために，Scale-Invariant Feature Transform (SIFT) 特徴量を用いた手法 [9]，Histogram of Oriented Gradients (HOG) 特徴量を用いた手法 [10]，動きの軌跡や速度を用いた手法 [11] などが存在する．そして，それらの特徴量の手話動画内での時間的関係をモデル化するために，Hidden Markov Models (HMM) を用いた手法 [12] や，Dynamic Time Warping (DTW) を用いた手法 [13] がある．

近年は，Deep Neural Network を用いた手話認識手法が広く見られ，入力に用いるデータの違いから，大きく 2 種類のアプローチに分けられる．1 つ目は手話動画の外観情報を用いる手法 [2], [3], [14], [15] で，2 つ目は手話を表現する話者の骨格情報を用いる手法 [2], [16], [17] である．

入力に手話動画の外観情報を用いるアプローチは更に 2 種類に分けられる．2D Convolutional Neural Network (2DCNN) を用いる手法 [2], [3] と 3D Convolutional Neural Network (3DCNN) [2], [3], [14], [15] を用いる手法である．2DCNN を用いる手法では，まず，2DCNN によって入力動画フレームの空間的な特徴を抽出する．そして，Recurrent Neural Network (RNN) を用いて，入力動画の時間的な特徴を捉える．Li らの手法 [2] では，2DCNN に ImageNet [18] で事前学習した VGG16 [19] を採用し，抽出された特徴量を Gated Recurrent Unit (GRU) に入力している．一方，3DCNN は入力動画に対して，空間情報 (2D) と時間情報 (1D) をまとめて，3D の畳み込みを行うことで，時空間特徴を同時に捉えられる．Li ら [2]，Joze ら [3] は 3DCNN のネットワークに ImageNet 及び Kinetics [4] で事前学習した I3D を採用し，それぞれのデータセットにおいて最も高い認識精度を達成している．

手話話者の骨格情報を用いるアプローチも，外観情報を用いる手法と同様に，2 種類に分けられる．1 つ目は RNN を用いて時間特徴を抽出する手法 [2], [16] で，2 つ目は Graph Convolutional Network (GCN) [20] を用いて時空間特徴を抽出する手法 [2], [17] である．GCN とはグラフ構造を入力とする畳み込みニューラルネットワークである．RNN を用いる手法では，まず，OpenPose [21] 等を用いて人物の骨格推定を行う．OpenPose とは画像中の人物の骨格点の座標及び尤度を取得できる手法である．そして，推定された 2 次元の骨格点座標を連結してベクトルにし，RNN の入力として用いる．GCN を用いる手法においても，まず，人物の骨格推定を行う．推定された骨格データをグラフ構造と捉えると，関節点がノード，各関節間の関係がエッジに対応する．この骨格グラフを GCN の入力として用いて，手話認識を行う．

本研究では，手話単語認識の精度向上を目的とした提案手法のネットワークに，画像ベースの行動認識や手話単語認識において高精度を誇る I3D [4] と，骨格ベースの行動認識において高精度を誇る ST-GCN を用いる．

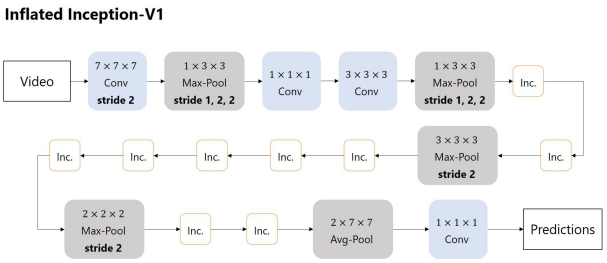

図 1 I3D network architecture [4].

2.2 I3D

I3D [4] は Carreira らによって提案された 3DCNN ベースのネットワークである．I3D の概観を図 1 に示す．従来の 3DCNN [22] は，パラメータ数の多さから，深い構造にすることが困難であった．そこで，I3D は，Inception Network [23] の 2D フィルタ及びプーリング層を 3 次元に拡張することで，大幅にパラメータを削減し，より深い構造を可能にした．さらに，2D フィルタの 3 次元拡張というアイデアは，ImageNet のような大規模な画像データセットでの事前学習を可能にし，高精度な 2D 画像分類モデルの重みを I3D の重みの初期値として用いることができる．そして，Kinetics のような大規模な動画データセットでファインチューニングすることで，動画から時空間特徴を学習することができる．

行動認識タスクにおいて，全体画像とオプティカルフロー画像をそれぞれ別々の I3D に入力し，その出力を組み合わせることで認識結果を取得するという Two-stream I3D が提案されている．この Two-stream I3D の認識精度は，全体画像のみを入力する I3D の認識精度よりも高いと報告されている [4]．これは，オプティカルフロー画像の入力により，行動認識において重要な人物の動きにより焦点を当てられた結果である．

手話単語認識の分野においては，Li らの研究 [2]，Joze らの研究 [3] で，複数の手法の認識精度を比較しており，I3D を用いた手法が最も高い精度を達成している．しかし，その認識精度は，実用化に至るほど高い精度ではなく，さらなる精度向上の余地がある．これらの手法では，話者の全身もしくは上半身の外観情報のみを用いて手話単語を認識しており，手話において重要な手や顔の局所特徴を十分に捉えられないことが問題であると考えられる．

2.3 ST-GCN

動画像の背景情報の影響を受けないロバストなモデルの構築を目的として，近年，骨格情報を用いた行動認識の研究が行われている．これらの手法では，骨格データをグラフ構造とみなし，グラフ構造を入力とする畳み込みニューラルネットワークである GCN を用いている．その中で，Spatial Temporal Graph Convolutional Network (ST-GCN) [24] は骨格データを入力に用いて行動認識を行うために，Yan らによって提案されたネットワークであり，高い識別精度を誇る．ST-GCN の概観を図 2 に示す．この手法では，骨格データを 2 つのグラフ構造と捉える．1 つは，同一フレーム内の関節点を結ぶことで，関節間の関係を考慮する空間グラフ，もう 1 つはフレーム間の同一

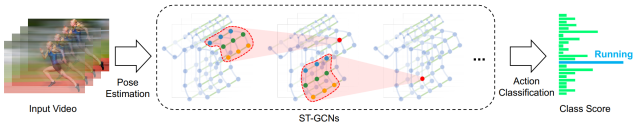

図 2　An overview of ST-GCN network architecture [24].

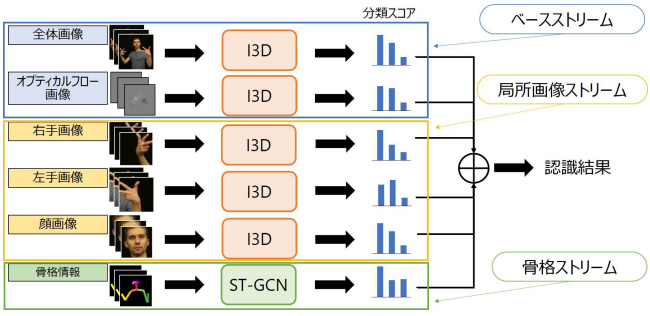

図 3　Overview of the proposed method.

関節点を結ぶことで時間方向の変動を考慮する時間グラフである．これにより，骨格点の空間的特徴及び時間的特徴の両方を同時に学習することができ，人物の行動認識が可能である．

手話単語認識においても，ST-GCN を用いた研究が行われており，人物の体や手の骨格データをグラフ構造とみなすことで，骨格点の時空間特徴の抽出を可能としている [17]．しかし，骨格情報が保持する情報量は画像の外観情報から得られる情報量より乏しいため，骨格情報のみ用いた手法の認識精度は外観情報を用いた手法よりも精度が低い [2], [3]．そのため，本研究では，外観情報に加えて骨格情報も用いることで，背景情報などの影響を軽減し，精度の向上を図る．

3. 提 案 手 法

提案手法の概要を図 3 に示す．本節では提案手法を構成する (1) ベーシストリーム，(2) 局所画像ストリーム，(3) 骨格ストリームの 3 つを順に説明する．各ストリームは個別に学習し，テスト時に各ストリームから得られる分類スコアの平均を取り，最も分類スコアの高いクラスを最終的な認識結果とする．

3.1 ベーシストリーム

手話単語認識の従来手法 [2], [3] では，手話動画から切り出した各フレーム画像全体のみを入力するシングルストリームの I3D が用いられる．Carreira らの研究 [4] から，行動認識のタスクにおいては，オプティカルフロー画像も入力に用いて，Two-stream の構造を取ることで，認識精度が向上することが分かっている．行動認識と同様に，手話認識においても人物の動きは非常に重要な情報の一つであるため，Two-stream 構造が手話認識性能を向上させると容易に想像できる．したがって，本研究では，全体画像とオプティカルフロー画像を入力に用いる Two-stream I3D を，提案手法のベースとして用い，このストリームをベーシストリームと呼ぶ．提案手法ではオプティカルフローの計算に TV-L[1] アルゴリズム [25] を用いる．

3.2 局所画像ストリーム

手話の中には，手の動きは同じで，形状だけが異なる手話が

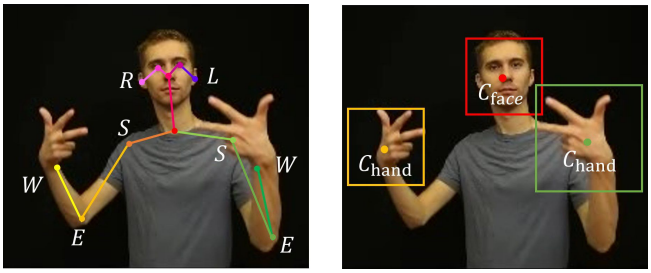

図 4　An example of face and both hands bounding boxes extraction.

存在するため，手話認識において手の形状は最も重要な要素の一つである．さらに，手話を読み取る際，顔の情報も重要になる．聴覚障害者は，話者の表情や口の動きも捉えることで手話を読み取る．このように，手話において，手や顔の局所領域の情報は重要な意味を持つ．そのため，手話の動画フレームから手と顔の領域を切り出した局所領域画像パッチを I3D に入力することで，モデルが手の形状や顔の表情といった局所特徴量を捉えられるようにする．このように，全体画像からでは捉えきれない局所的な特徴を学習することを目的としたストリームを局所画像ストリームと呼ぶ．

図 4 に手及び顔の局所領域画像の切り出し例を示す．手は肘から手首へ向かった先の延長線上にあるという認識のもと，OpenPose で推定した骨格点座標を用いて，以下の数式の通り手領域のバウンディングボックスを取得する．

$$\overrightarrow{EC_{\text{hand}}} = 1.33 \times \overrightarrow{EW} \tag{1}$$

$$w_{\text{hand}} = h_{\text{hand}} = 1.2 \times \max(|\overrightarrow{EW}|, 0.9 \times |\overrightarrow{SE}|) \tag{2}$$

顔領域に関しても，OpenPose の骨格座標を用いて，以下の数式の通りバウンディングボックスを取得する．

$$\overrightarrow{RC_{\text{face}}} = 0.5 \times \overrightarrow{RL} \tag{3}$$

$$w_{\text{face}} = h_{\text{face}} = 1.5 \times |\overrightarrow{RL}| \tag{4}$$

ここで，C はバウンディングボックスの中心点，S, E, W はそれぞれ肩，肘，手首点を表し，R, L はそれぞれ右耳，左耳の位置を表す．また，w, h はバウンディングボックスの幅及び高さを表す．局所画像パッチを切り出した後，224×224 ピクセルにリサイズした右手，左手，顔の画像パッチを入力に用いて，3 つの I3D を個別に学習する．

3.3 骨格ストリーム

手話を識別する際，上で挙げた要素以外にも重要な要素がある．それは，体に対して手がどの位置にあるのかという情報と，手が指さす方向という情報である．これらの情報に焦点を当てて学習するために，我々は手話話者の骨格情報を用いる．さらに，骨格情報を用いることで，背景の情報や話者の見た目といった情報による影響を軽減できると考える．

骨格情報の取得には，OpenPose を用いる．推定された骨格点の内，図 5 に記した，体 5 点，左右の手それぞれ 11 点，計 27 点の骨格点を抜粋する．ただし，推定に失敗した骨格点につ

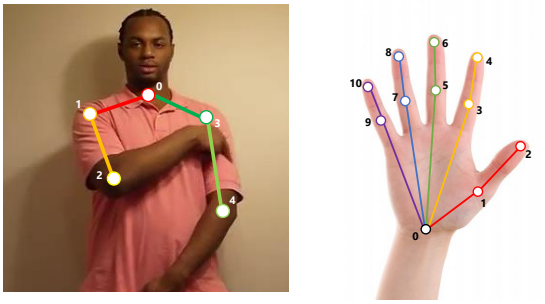

図 5　27 keypoints inputted to ST-GCN in Skeleton Stream. 5 keypoints refer to body and 11 keypoints refer to each hand [17].

いては，その座標を $(0, 0)$ とする．そして，これらの骨格点の座標をすべて連結させ，54 次元のベクトルとし，ST-GCN に入力する．本研究では，この骨格情報を入力に用いるストリームを骨格ストリームと呼ぶ．

4. 実　　験

4.1 データセット

本実験で用いるデータセットについて説明する．手話単語認識のためのデータセットは数多く存在する．その中でも動画数，手話話者数，各クラス当たりの動画数において規模の大きい WLASL [2] と MS-ASL [3] データセットを使用する．これらのデータセットの各動画は，1 人のネイティブの手話話者もしくは通訳者が 1 つのアメリカ手話単語を表現したものである．どちらのデータセットもそれぞれ 4 つのサブセットを持つ．WLASL の中には WLASL100，WLASL300，WLASL1000，WLASL2000 というサブセット，MS-ASL の中には MS-ASL100，MS-ASL200，MS-ASL500，MS-ASL1000 というサブセットがあり，各サブセット名の後の数字はそのサブセットに含まれるクラス数を表す．これらのサブセットは各データセットにおいて，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 位のクラスで構成される．ここで，WLASL に対しては $K = \{100, 300, 1000, 2000\}$，MS-ASL に対しては $K = \{100, 200, 500, 1000\}$ である．本実験では，WLASL のすべてのサブセットと MS-ASL100 を使用して，提案手法の認識性能を評価する．表 1 に各サブセットの詳細を示す．MS-ASL に関して，動画データのダウンロードに使用するリンクがいくつか切れているため，元のデータセットの約 25%少ないデータしか取得できない．それに伴い，1 クラス当たりの平均動画数と話者数も，元のデータセットより少ない．そのため，MS-ASL を用いた実験結果は既存の文献に記載の数値と直接比較が出来ないので，注意が必要である．元のデータセットからの減少の具合は，表 1 の MS-ASL100 の行中の括弧内に示す．また，データの分割に関しては，どちらのデータセットに対してもそれぞれの著者が公開している分割に従い，その割合は学習：バリデーション：テスト＝ 4 : 1 : 1 である．

表 1　Details of datasets. We use #Class, #Video and #Signer to denote the numbers of classes, videos and signers, respectively. Column “Mean” denotes the average number of videos per class. The numbers in parentheses denote the relative size of the dataset we obtained compared with its original.

Subset	#Class	#Video	Mean	#Signer
WLASL100 [2]	100	2,038	20.4	97
WLASL300 [2]	300	5,117	17.1	109
WLASL1000 [2]	1,000	13,168	13.2	116
WLASL2000 [2]	2,000	21,083	10.5	119
MS-ASL100 [3]	100	4,315 (-25%)	43.2 (-25%)	163 (-26 signers)

4.2 定 量 評 価

本実験での学習及びテストの実装方法は [2], [3] に従う．すなわち，前処理として，WLASL 及び MS-ASL 内のすべての動画フレームに対して，前者は YOLOv3 [26] によって，後者は SSD [27] によって人物のバウンディングボックスを検出する．そして，バウンディングボックスの対角サイズが 256 ピクセルになるように各フレームをリサイズした後，バウンディングボックスを中心とした 256×256 ピクセルの正方形を切り出す．空間的なデータ拡張のために，学習フェーズでは，各フレームからサイズが 224×224 ピクセルの正方形画像パッチをランダムに切り出し，さらに，0.5 の確率でランダムに水平方向の反転処理を行う．アメリカ手話において，鏡写しでも手話の意味は変わらない．そのため，この水平反転処理は，どちらの表現にも対応できるように学習することが目的である．そして，時間的なデータ拡張のために，連続 64 フレームをランダムに動画から選択する．ただし，64 フレームより少ない動画に対しては，動画の最初か最後のどちらか 1 フレームをランダムに選択して複製することで，足りない分を補う．I3D を学習する際は，初期学習率が 10^{-3}，weight decay の値が 10^{-7} の Adam を用いる．ST-GCN を学習する際は，初期学習率が 0.01，weight decay の値が 10^{-4} の Adam を用いる．すべてのモデルはいずれのデータセットでも 200 エポック学習する．テスト時は，動画の全フレームをモデルに入力する．

本実験では，提案手法の精度を評価するために，2 つのベースラインと比較する．1 つ目のベースラインは，全体画像のみを入力に用いる I3D である．これは，Li ら [2]，Joze ら [3] の研究で用いられ，前者は WLASL で，後者は MS-ASL で state-of-the-art を達成している．2 つ目のベースラインは，入力に全体画像とオプティカルフロー画像を用いる Two-stream I3D である．つまり，図 3 内のベースストリームの部分のみを採用したモデルである．これら 2 つのベースラインをそれぞれ，Baseline1，Baseline2 と呼ぶ．また，局所画像ストリームの有無，骨格ストリームの有無のすべての組み合わせで実験を行い，これらのストリームが手話認識精度の向上に有効であるかどうかを確かめる．提案手法の精度評価には，Top-N classification accuracy を用いる．ここで，$N = \{1, 5, 10\}$ とする．

WLASL の 4 つのサブセットでの実験結果を表 2 に示す．

表 2 Recognition accuracy (%) on four subsets of WLASL dataset. Columns “Flow,” “Local” and “Skeleton” denote optical flow images, local image patches and skeletal information, respectively.

				WLASL100			WLASL300			WLASL1000			WLASL2000		
Model	Flow	Local	Skeleton	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10	Top 1	Top 5	Top 10
Baseline1 [2]				65.89	84.11	89.92	56.14	79.94	86.98	47.33	76.44	84.33	32.48	57.31	66.31
Baseline2	✓			77.55	91.25	94.92	66.96	87.61	92.03	56.35	83.03	88.77	38.67	68.43	76.39
Ours1		✓		76.60	89.13	92.80	66.34	88.46	92.43	56.91	84.55	89.83	41.01	74.46	81.85
Ours2			✓	71.07	90.13	93.42	65.10	85.49	90.64	53.75	80.01	86.56	37.65	67.61	75.97
Ours3		✓	✓	77.48	92.38	95.72	69.99	89.91	93.50	60.85	86.98	91.41	45.12	79.17	85.65
Ours4	✓	✓		80.38	93.38	95.97	73.07	**90.85**	94.44	62.76	88.02	92.32	45.30	79.63	85.75
Ours5	✓		✓	78.05	91.63	95.42	69.77	88.61	92.33	58.68	84.05	90.25	41.94	73.16	81.68
Ours6	✓	✓	✓	**81.38**	**94.13**	**96.05**	**73.43**	90.19	**94.83**	**63.61**	**88.98**	**92.94**	**47.26**	**81.71**	**87.47**

表 3 Recognition accuracy (%) on MS-ASL100.

				MS-ASL100		
Model	Flow	Local	Skeleton	Top 1	Top 5	Top 10
Baseline1				73.69	91.38	93.87
Baseline2	✓			82.46	94.66	96.61
Ours1		✓		75.12	91.33	93.87
Ours2			✓	75.61	92.38	95.07
Ours3		✓	✓	76.69	92.65	95.35
Ours4	✓	✓		83.84	94.69	96.18
Ours5	✓		✓	**84.22**	94.77	96.48
Ours6	✓	✓	✓	83.86	**94.86**	**96.66**

表 2 の Baseline1 と Baseline2 の結果から，オプティカルフロー画像を付け加えるだけで，従来手法 [2] の精度を上回っていることが分かる．これは，RGB 画像の外観情報だけでは捉えきれない動きの情報を学習することが手話単語認識に有効である事を示す．このような結果は行動認識の分野における Carreira らの研究 [4] でも示されている．さらに，表 2 から，WLASL のすべてのサブセットにおいて，オプティカルフロー画像，局所領域画像パッチ，骨格情報を入力に用いた Ours6 が最も高い認識精度を誇ることが分かる．このことから，ベースラインのモデルが捉えられない手話特徴を，提案手法では学習できることが確かめられた．次に，局所画像ストリームの有無（例えば，Baseline1 と Ours1 や，Baseline2 と Ours4）を比較すると，いずれのサブセットにおいても，局所画像ストリームがあるモデルの方が認識精度が高いという結果が得られた（Baseline1 と Ours1 の比較において，例えば WLASL100 の Top-1 認識率は 65.89%から 76.60%に 10.71%向上した）．この結果から，単語レベルの手話認識において，我々が提案する局所画像ストリームは効果的であることが分かる．さらに，骨格ストリームの有無（例えば，Baseline1 と Ours2 や，Baseline2 と Ours5）を比較する．局所画像ストリームを付け加えたときの精度の上がり幅と比較すると，骨格ストリームによる認識率の向上は少し劣るものの，骨格ストリームの導入により認識精度は向上した（Baseline1 と Ours2 の比較において，例えば WLASL100 の Top-1 認識率は 65.89%から 71.07%に 5.18%向上した）．これより，骨格ストリームの付与も手話認識において有効な手法であることが示された．

WLASL 以外のデータセットにおける提案手法の有効性を確かめるために，MS-ASL を用いて実験を行った．MS-ASL100 での実験結果を表 3 に示す．Top-1 accuracy では Ours5，Top-5，10 accuracy では Ours6 の精度が最も高いという結果が得られた．WLASL での実験と同様に，局所画像ストリーム及び骨格ストリームの有無で比較すると，どちらのストリームについても付与したモデルの方が精度が高いことが分かる．よって，本研究で提案する Multi-stream モデルが手話認識に有効であることが示された．しかし，WLASL での実験結果と比較すると，MS-ASL では，局所画像ストリームと骨格ストリームを付与することでの認識精度の向上度合いが小さいことが分かる．これは，MS-ASL に含まれるデータのバリエーションの豊富さによる影響であると考えられる．図 6 に WLASL と MS-ASL のデータの一例を示す．WLASL に含まれる動画は，図 6 (a) のように，話者は正面を向き，カメラは話者の腰から上を写しているものに限られている．一方，MS-ASL には，図 6 (b)(c) のような，話者がカメラに対して横向きである動画や，話者がアップで写る動画が含まれる．我々は手領域の検出に，OpenPose で取得した肩，肘，手首の 3 点を使用するため，話者がアップで写る場合には，肘点の推定ができず，手領域抽出に失敗する．また，横からの視点の動画から切り取られた局所画像パッチや骨格推定結果は，正面視点のそれらのデータの様子とはかけ離れたデータとなる．このようなデータが局所画像ストリームや骨格ストリームの学習に影響を及ぼし，認識率向上の妨げになると考えられる．

そこで，今後の課題として，手領域画像の抽出方法の変更が挙げられる．話者がアップに写る場合でも，手が写っていれば手領域を正確に取得できるような手法へ改善することで，より多くのバリエーションのデータに対応することができ，さらなる手話認識精度の向上が望めると考える．

(a) WLASL

(b) MS-ASL (View from the side)

(c) MS-ASL (Only the part from the chest up)

図 6　Samples of WLASL [2] and MS-ASL. [3]

5. おわりに

本研究では，手話単語認識精度の向上を目指し，局所特徴に着目した Multi-stream 構造を持つ認識手法を提案した．具体的には，手話認識において重要な，手の形，顔の表情，体に対する手の位置という局所的な情報を捉えるために，局所画像ストリームと骨格ストリームを導入した．実験より，提案手法である Multi-stream モデルが手話単語認識精度を向上させることが確認できた．さらに，局所画像ストリームと骨格ストリームの有効性を検討するために，それぞれの有無による認識精度の比較を行った．その結果，どちらのストリームにおいても，付け加えたモデルの方が認識精度が高く，これらのストリームが手話認識に有効であることが確かめられた．一方で，横視点の動画や人物がアップの動画などバリエーションの豊富なデータを含む MS-ASL における実験では，大きな認識精度の向上は見られなかった．そこで，今後の課題として，様々なデータセットに対応できるようなモデルの設計による，さらなる手話認識精度の向上が挙げられる．

謝辞　本研究は，JSPS 科研費#19K12023 の補助による．

文　献

[1] D. Li, X. Yu, C. Xu, L. Petersson, and H. Li, “Transferring cross-domain knowledge for video sign language recognition,” Proc. of CVPR, pp.6205–6214, 2020.

[2] 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.

[3] H. Vaezi Joze and O. Koller, “MS-ASL: A large-scale data set and benchmark for understanding american sign language,” Proc. of BMVC, p.100, 2019.

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

[5] 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.

[6] Z. Tu, W. Xie, Q. Qin, R. Poppe, R.C. Veltkamp, B. Li, and J. Yuan, “Multi-stream CNN: Learning representations based on human-related regions for action recognition,” Pattern Recognition, vol.79, pp.32–43, 2018.

[7] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, “Action recognition with dynamic image networks,” IEEE Trans. PAMI, vol.40, no.12, pp.2799–2813, 2018.

[8] J. Zang, L. Wang, Z. Liu, Q. Zhang, G. Hua, and N. Zheng, “Attention-based temporal weighted convolutional neural network for action recognition,” Proc. of AIAI, pp.97–108, 2018.

[9] F. Yasir, P.W.C. Prasad, A. Alsadoon, and A. Elchouemi, “SIFT based approach on bangla sign language recognition,” Proc. of IWCIA, pp.35–39, 2015.

[10] S. Liwicki and M. Everingham, “Automatic recognition of fingerspelled words in british sign language,” Proc. of CVPR Workshops, pp.50–57, 2009.

[11] C. Monnier, S. German, and A. Ost, “A multi-scale boosted detector for efficient and robust gesture recognition,” Proc. of ECCV, pp.491–502, 2015.

[12] T. Starner and A. Pentland, “Real-time american sign language recognition from video using hidden markov models,” Proc. of ISCV, pp.265–270, 1995.

[13] J.F. Lichtenauer, E.A. Hendriks, and M.J.T. Reinders, “Sign language recognition by combining statistical DTW and independent classification,” IEEE Trans. PAMI, vol.30, no.11, pp.2040–2046, 2008.

[14] J. Huang, W. Zhou, H. Li, and W. Li, “Sign language recognition using 3D convolutional neural networks,” Proc. of ICME, pp.1–6, 2015.

[15] Y. Ye, Y. Tian, M. Huenerfauth, and J. Liu, “Recognizing american sign language gestures from within continuous videos,” Proc. of CVPR Workshops, pp.2064–2073, 2018.

[16] S.-K. Ko, J.G. Son, and H. Jung, “Sign language recognition with recurrent neural network using human keypoint detection,” Proc. of RACS, pp.326–328, 2018.

[17] 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.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” IJCV, vol.115, no.3, pp.211–252, 2015.

[19] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” Proc. of ICLR, 2015.

[20] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. van denBerg, I. Titov, and M. Welling, “Modeling relational data with graph convolutional networks,” Proc. of ESWC, pp.593–607, 2018.

[21] 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.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” Proc. of ICCV, pp.4489–4497, 2015.

[23] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” Proc. of CVPR, pp.1–9, 2015.

[24] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” Proc. of AAAI, pp.7444–7452, 2018.

[25] C. Zach, T. Pock, and H. Bischof, “A duality based approach for realtime TV-L^1 optical flow,” Pattern Recognition, pp.214–223, 2007.

[26] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” arxiv:1804.02767, 2018.

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A.C. Berg, “SSD: Single shot multibox detector,” Proc. of ECCV, pp.21–37, 2016.