J-027

骨格情報の時空間パターンを用いたハンドジェスチャ認識

Hand Gesture Recognition Using Spatiotemporal Skeletal Features

河村 寛生 †　　中井 満 †

Hiroki Kawamura　Mitsuru Nakai

1. はじめに

ロボットと人が協働する空間において，円滑なコミュニーケーションを実現するために人の指示動作を認識することが重要である．人とロボットの意思疎通を目的とした従来研究として，指差しによってロボットを移動させる研究 [1] や，人間の動作をロボットに学習・獲得させる研究など [2] がある．これらの研究では，指差し動作や四角形・円形・三角形を描くような手先の軌道といった単純なジェスチャが用いられていた．

指示動作には日常的に使われる動作を用いることが望ましく，それによってより直感的かつ自然な操作が可能になると考えられる．本研究では,「親指を立てる動作」や「手を突き出す動作」などの指示や操作でよく使われる日常動作の認識を行う．RGB カメラの前で操作を行うことを想定し，立体的な動作を捉えるために骨格を推定する．両手の関節の 3 次元空間における骨格情報を時系列グラフとして表現し，これを入力として ST-GCN[3] によってジェスチャを認識する手法を提案する．

2. システムの構成

本研究で提案するシステムの流れを図 1 に示す．RGB 動画に対して骨格推定を行い，両手の関節の 3 次元座標を取得する．関節をノード，関節間のつながりをエッジとしたグラフ構造を ST-GCN の入力とし，ジェスチャを認識する．

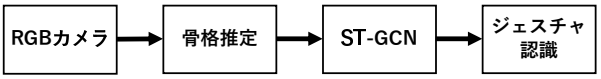

図 1: ジェスチャ認識システムの構成

2.1 両手の骨格グラフの作成

骨格推定は，Google 社が提供する機械学習フレームワークである MediaPipe[4] を用いて行う．これにより，動画像の各 RGB フレーム画像から，両手の 3 次元座標を推定できる．得られる手の骨格情報を図 2(a) に示す．手のランドマークは右手と左手それぞれの中心を原点とした相対座標になっている．また，全身のランドマークから右手首と左手首の位置関係が分かる．これをもとに手の骨格情報を平行移動し，右手と左手を同じ座標系に変換する．

2.2 ST-GCN

ST-GCN とは，グラフ畳み込みネットワーク（GCN）を時間方向へと拡張した動作認識手法である．ST-GCN で扱うグラフ構造の例を図 2(b) に示す．各フレームにおける関節間のつながりを考慮した空間グラフと，フレーム間の同一関節をつないで時間的な変化を考慮した時間グラフを組み合わせたネットワークである．空間グラフに対しては，GCN と同様に隣接行列や特徴行列を用いてグラフ畳み込み処理を行う．時間グラフに対しては，時系列の畳み込み処理を行う．

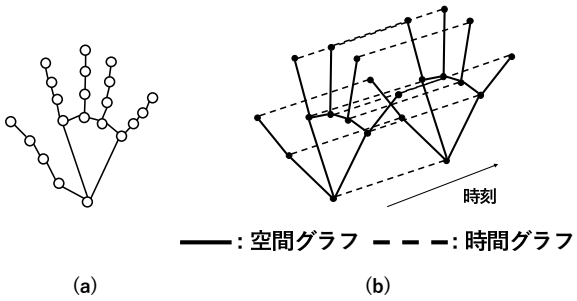

図 2: 手の骨格情報とグラフ構造

2.3 ST-GCN によるジェスチャ認識

手のグラフ構造の時空間パターンからハンドジェスチャを認識する流れを図 3 に示す．グラフ構造は合計 42 個の両手のランドマークで構成され，各ノードは 3 次元座標（x, y, z）を持つ．隣接関係は，図 2(a) に示すものと同様である．これを ST-GCN の入力情報とする．まず，1 層目では，$3 \times N$ 次元の重み行列を適用し，各ノードごとの特徴量を抽出する．ここで，N は次元数である．次に，空間グラフの畳み込みを行い，エッジで接続された隣接ノードの特徴量と自身の特徴量を統合することで，各ノードの特徴量を更新する．続いて，時間グラフの畳み込みを行い，異なるフレームにおける同一関節の特徴量を統合し，各ノードの特徴量を更新する．これを 1 層目の出力とし，2 層目の入力とする．2 層目以降は重み行列が $N \times N$ になることを除き，同様の処理を行う．これを L 層まで繰り返す．すべての ST-GCN 層での処理が終了した後，各ノードとフレームに対して平均値プーリングを適用し，特徴量を圧縮する．最終的に，圧縮された特徴量を $N \times M$ の全結合型ニューラルネットワークに入力し，各クラスのスコアを計算す

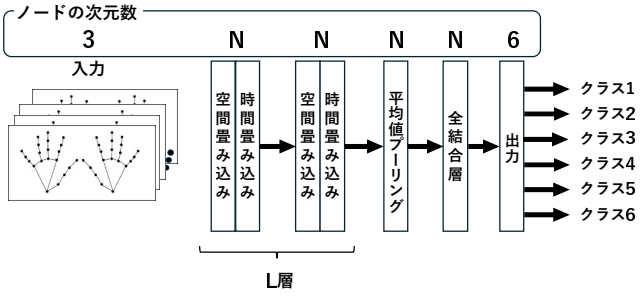

図 3: ST-GCN の構成

†富山県立大学, Toyama Prefectural University

図 4: 対象とするジェスチャ

る．ここで，M はジェスチャのクラス数である．出力されたスコアの中で最も高い値を持つクラスを当該ジェスチャであると決定する．

3. 実験

3.1 データ収集

カメラの正面に立った状態および座った状態でサンプルを取得した．RGB 動画の撮影には解像度 1280 × 800，フレームレート 30 fps のカメラを用いた．認識対象とするジェスチャを図 4 に示す．手を握って開く動作 (open)，手を突き出す動作 (stop)，手を上下させる動作 (up-down)，手を振る動作 (waving)，親指を立てる動作 (good)，指を差す動作 (pointing) の 6 種類である．撮影条件として，手に何も持たずに両手が完全に映ること，各フレームで 3 次元位置情報が取得できること，撮影開始時の初期状態を手を開いた状態にしておくこととした．「good」と「pointing」は片手の動作とし，その他は片手と両手の動作を含む．2 秒間（60 フレーム）のジェスチャを 1 サンプルとした．1 ジェスチャにつき 1 日 30 サンプルを収集した．被験者は大学生・大学院生の男性 4 人，女性 1 人の合計 5 人とし，4 日間に分けて行った．収集したデータの総数は 3600 サンプルとなった．

3.2 ST-GCN によるジェスチャ認識精度

6 種類のジェスチャに対し，ST-GCN のパラメータを変えて認識実験を行った．各ジェスチャ600 サンプルの合計 3600 サンプルを用いた．5 人分のデータのうち 1 人分を評価，残りを学習に用いて交差検証した．学習回数は最大 500 回，層の数は 3 層，バッチサイズは 16 に固定し，各層のノード数を 128，256，512 と変えて学習させた．結果として，ノード数が 512 かつ学習回数が 420 回のとき，認識率が最大となり，94.1 % となった．

3.3 ジェスチャ認識精度の詳細と考察

認識率が最大であった 420 エポックでの各ジェスチャに対する混同行列を図 5 に示す．「waving」以外のジェスチャに対して 90 %以上の認識率が得られた．特に「open」，「stop」，「up-down」はどれも 95 %以上の認識率となった．

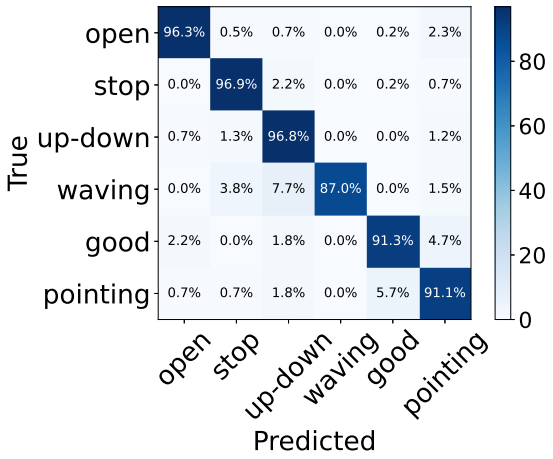

図 5: 各ジェスチャの混同行列

一方，「waving」の認識率は 87.0 %で他のジェスチャと比べて低く，また，「up-down」と誤ることが多かった．その原因として，手を振る動作における個人差が挙げられる．この動作には手首から先だけを使って小さく振る方法と，肘から先を使って腕を大きく振る方法がある．このような表現の仕方の違いにより，上下方向に対する特徴も捉えられてしまったことが考えられる．

また，「good」と「pointing」は相互に誤認識することが多かった．この原因として，どちらも 1 本の指を立てて他の指を握るという共通点が挙げられる．この 2 つのジェスチャの違いは立っている指が親指か人差し指かであるため，特徴を十分に捉えられなかったことが考えられる．

4. まとめ

ST-GCN を用いて，6 種類の日常動作の認識を行い，結果として 94.1 %の認識率を達成した．一方で，個人差や共通の構造による誤認識が確認された．今後は，データ拡張による認識精度の向上や，対応可能なジェスチャの種類を増やすことで，より柔軟で実用的なシステムの構築を目指す．

謝辞 本研究は JSPS 科研費 24K15050 の助成を受けて行った．

参考文献

[1] 篠塚晃希, 田村仁, “ハンドジェスチャによる 3 次元ポインティングを用いたロボット制御手法,” 第 20 回情報科学技術フォーラム, 2021.

[2] 滝澤和真, 大保武慶, “人とロボットとの協調的コミュニケーションに基づく描写的ジェスチャの獲得,” 第 37 回ファジィシステムシンポジウム 講演論文集, 2021.

[3] Sijie Yan, et al. “Spatial-Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition, ”Proc. of AAAI, 2018.

[4] Google for Developers, “MediaPipe” https://developers.google.com/mediapipe, 2024/2/8 閲覧.