3 次元姿勢推定による警察官の手信号認識の学習データ量の削減

M2021SC013 手塚悠
指導教員：河野浩之

1 はじめに

警視庁によると, 事故や停電などで信号機に問題があった際は警察官による交通整理が行われ, 信号機よりも優先するという決まりがある. そのため, 警察官による手信号を認識するシステムは自動運転車に備えておくべき機能であるといえる. そこで本研究ではカメラ映像を用いて警察官の手信号の認識を行う. 体の各部位の座標を推定し, 手信号の分類には取得した座標の時系列での動きをニューラルネットワークで学習させ実施する.

2 警察官の手信号認識とその他のジェスチャー認識の先行研究

警察官の手信号以外にも手話等のハンドサイン認識やヨガや筋トレといった運動時の姿勢支援等, 数多くの研究が実施されている. ジェスチャー認識には一般的に 2 種類の手法があり, 1 つは CNN ベースで人体の状態から直接検出を行うものと, もう 1 つは目や鼻, 腰, 足などの骨格点の座標を抽出してその座標データを元に分類を行うものである.

警察官による手信号認識の研究では小野ら [1] が人体の骨格座標を 18 箇所の点で表すことができる OpenPose を用いて骨格座標を二次元座標で抽出し、LSTM による学習で認識を行う手法で 98.0% の精度を達成した. その際, 2,146.2 秒の学習データと 1,424.8 秒のテストデータの計 3,571 秒のデータを使用した.

次に, Taeseung ら [2] は骨格座標を使用せずに CNN ベースで腕の方向を検出し RNN に入力して分類を行うという手法をとった. この研究では 4 種類の RNN を比較して GRU が優れているものの層を増やすことで Bi-LSTM の方がより高いテスト精度を得られており, 実験の結果 90.8% の精度を達成した. また, データセットとして FPS が 32.5 の 104,654 フレームと約 3,220.0 秒のデータで構成されている.

また, 手信号だけでなく手話の認識では松田ら [3] がアルゼンチンの手話データセット LSA64 を用いて実験を行い, CNN ベースで手の動きや顔の表情を処理し, その時系列データを Bi-LSTM で分類を実施. その際, 手話動作のフレームごとにピクセル内のオプティカルフローを計算しフロー画像を用いることで腕の部分にのみ着目するようにして実験を行い, 94.0% を達成した. LSA64 は 3,200 個のデータで構成され, 全部で 64 単語の手話動作を行っており, 1 動画当たり約 1 秒のため合計 3,200.0 秒ほどのデータセットである.

3 手信号認識のデータ量削減の提案手法

3.1 先行研究の課題

従来の 2 次元座標を用いたジェスチャー認識では, 同じ動きであっても対象をどこから見ているかによって骨格座標が変わってくる。日本の警察官の手信号は車の進行方向が体の正面に対して平行である場合も垂直である場合も動作は同じであり, 警察官の体の向きによってのみ意味が変化するため, 区別して学習を行う必要がある. さらに交差点の形は常に 90 度で交差する場合のみではないため, これに対処するとなるとあらゆる角度からの座標データが必要になってしまう. データ量が増えることの問題として, Curtis ら [4] は ImageNet 等の 10 個の大規模データセットにおいて平均して 3.3% 以上の誤差があるとしており, 原因として大規模データセットのデータ収集やラベル付けは外部で労働者を雇うため, ミスが発生するリスクがあることが考えられる.

そこで本研究では 3 次元姿勢推定機能を用いてカメラ映像から奥行きの座標を推定し, 肩の座標から体の向き (角度) を計算することで体の向きによらず分類ができる. また, 手招き動作の分類に使用する骨格座標の取得数を減らすという 2 種類の手法によってデータの削減を図った. これにより, データ数が減ればその分だけミスも減らせるため有効であると考えられる. ジェスチャーの認識には 3 次元座標データのほかに時系列も考慮する.

3.2 警察官の手信号

警察官の手信号は, 故障や停電などで信号機に不具合が発生した場合などに行われるもので, 警察官が腕を横にしている際, 体の向きに対して進行方向が垂直の場合は赤信号で平行の場合は青信号を表す. また, 両腕を上にあげている際は体の向きに対して進行方向が垂直の場合は赤信号なのは変わらずに, 平行の場合は青信号ではなく黄信号を表す. また, 腕を横にしている際, 手招き動作が行われる.

本研究では 3 次元座標データから腕の位置が「横」であるのか, 「上」であるのか, 「その他」であるのかを 3 クラスで判定する. そして腕が「横」である時, 手のひらの時系列データを使い手招き動作をしているのかどうかを分類する. 位置と時系列データの分類はニューラルネットワークを用いて学習を行っていく. なお, 手招き動作の分類には角度の算出による回転処理は行わない.

この研究の目的として体の向きによらない検出が主なため右折については取り扱わない. また, 手信号をしている人物が警察官であることを前提として実験を行っていくため, 警察官かどうかの判定も同様に実施しない.

3.3　骨格座標の抽出方法

人物の骨格点の抽出には MediaPipe や OpenPose が使われることが多くこれらを用いた研究が一般的である．基本的にはどちらも高精度な検出が可能であり，3 次元座標推定も実行できる．MediaPipe は 1 人の検出しかできないが 3 次元座標推定の場合では OpenPose も現状は 1 人の検出にしか対応していない．パフォーマンス面では MediaPipe の方がより高 FPS でモーションブラーに対して堅牢であり[*1]，実際の道路での走行を想定すると緊急事態に備える必要もあるため本研究では MediaPipe を使用する．

MediaPipe を用いて骨格点の 3 次元座標データの取得を実施すると図 1 のように出力される．z の所の数値は z 座標を表している．

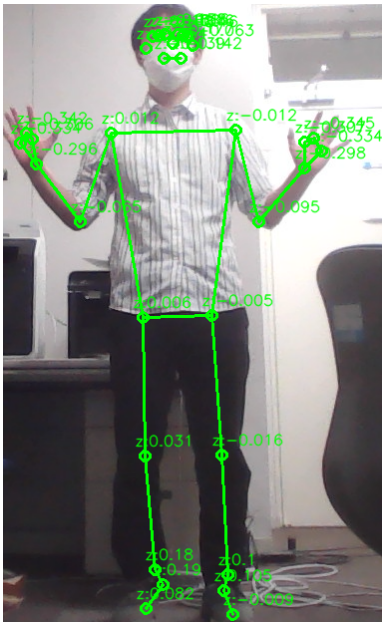

図 1　骨格座標データの抽出結果

3.4　3 次元座標の回転処理

MediaPipe から取得した 3 次元座標データ K_i が i=33 箇所の各点で得られる．これらの点は左肩の座標 (i=11) を基準として体の向き θ を算出する．カメラの高さを肩の高さと等しくすることで，両肩の高さは常に一定なため y 座標を無視した x, z 平面での座標変換で計算が可能になる．数式 (1) のように肩の 2 点間の x, z 座標から

$$tan\ \theta = \frac{K(x)_{12} - K(x)_{11}}{K(z)_{12} - K(z)_{11}} \tag{1}$$

を計算しこれをもとに角度 θ を求める．しかし，θ は x 軸からの角度を計算するため θ の範囲は ($-45 \leq\ \theta\ \leq 45$) で出力される．このままだと体の向きが正面を向いた時と後ろを向いた時で同じ角度であるため，右肩と左肩の差を取った時に x, z 座標の値が正か負であるかを判定して適切な数を足すことで角度を ($0 \leq\ \theta\ \leq 360$) 度で表す．そうして求まった θ を用いて右肩以外のすべての点に対して

$$K' \begin{pmatrix} K(x)_i' \\ K(y)_i' \\ K(z)_i' \end{pmatrix} = \begin{pmatrix} K(x)_i cos(-\theta) - K(z)_i sin(-\theta) \\ K(y)_i \\ K(x)_i sin(-\theta) + K(z)_i cos(-\theta) \end{pmatrix} \tag{2}$$

数式 (2) を計算することで常に正面を向いた状態での骨格点抽出が可能となる．手信号の認識の際には角度と動作を組み合わせることで手信号の状態を認識する．

次に時系列データとして利用できるように N フレーム間の座標データも取得する．手信号をする際，下半身の座標はほとんど動かなく，また，頭の座標は指示には直接関係がない．腕の部分についても掌の座標や指の座標はそこまで差がないと考えられるため，実際に時系列データを取得するのは左腕の手のひらの座標のみとする．これによりさらなるデータの削減を目指す．

4　手信号認識のデータ量削減の実験

4.1　実験の手順と 3 次元座標データの取得

実験では撮影した約 25 分の動画データの内，約 16 分の動画から骨格座標と時系列データを抽出し，クラス分けを行ったものをデータセットとして用意する．次に，取得したデータに対してモデルの学習を行う．学習したモデルを用いて残りの動画から分類結果を算出する．

MediaPipe による骨格点推定をして 3 次元座標データを抽出すると表 2 のように得られる．これは右肩の座標を基準として各点の座標との差を取り，正規化を行っている．座標データで取得されるデータ数は 1 行当たり，クラスと 33 箇所の各部位 x, y, z 座標で合計 100 個の要素で構成されている．次に，時系列データで取得されるデータ数は 1 行当たり，クラスと左手のひらの N フレーム間の x, y, z 座標で 3N+1 個の要素で構成されている．

表 1　右肩の座標を基準とした時の骨格座標の抽出結果

左肩の座標		
x	y	z
-0.2397	-0.0113	0.1756
-0.2394	-0.0110	0.1788
-0.2487	-0.0098	0.1895
-0.2620	0.0102	0.1720

また，分類をする際 FPS が約 20 で安定しているため時系列データとして取得するフレーム N は，手を横にして手招きをする動作の手招き 1 回分の時間である約 0.8 秒から N=16 として判断する．次に得られた N の値をもとに左手のひらの時系列データを取得する．このデータを用いて手を横にしている際の手招き動作の検出を行う．

[*1] HearAI(https://www.hearai.pl/post/14-openpose/) を参照

4.2　座標データと時系列データのモデル学習

取得した座標データに対して手の位置が横に伸ばしている時を“side”で, 上にあげている時を“up”で, それ以外を“other”として 3 クラスを分類する. パラメータ数が 2,663 で 7 層の多層パーセプトロンで手の位置の分類学習を実施する. データセットとして 981 個の座標データを使って学習を行うと検証データに対する損失が 0.031, 精度は 0.996 という結果が得られた.

次に座標データによる手の位置が“side”と判定される際, 手招き動作をすると“beckoning”かそれ以外の状態を“other”として 2 クラスを分類する. パラメータ数が 1,280 で 5 層の LSTM で手招き動作の分類学習を実施する. 座標データの時と同様に 1,050 個の時系列データを用意し学習を行うと, 検証データに対する損失が 0.39718, 精度が 0.82028 という結果が得られた. 次の節以降は実際に動画に対して精度を算出していくがその際, 精度向上のため Python の Counter 関数を用いてフレーム中に判定されたものの中で最多要素を分類結果として出力させる. これによって作成したモデルよりも高精度な分類が可能となる.

4.3　座標データでの手の位置の分類実験

前節の学習モデルを用いて座標データでの手の位置の分類実験を行う. 訓練データとは別に用意した 3 分 40 秒 (約 4,500 フレーム分) の動画を実験データとして分類の精度を算出した結果の一部分を図 2 に載せた. このグラフは縦軸の数値が 1 の時“side”, 2 の時“up”, 0 の時“other”を表しており, 横軸は時間をフレームで表している. 途中 300 フレーム位の時に 0 となっているのは手を下に下げている時なので正しく判定されている. 約 1,000 フレーム以降の手を上にしている箇所も正しく判定されているが, 800 フレームあたりの 2 箇所が手を上にしていると誤判定されてしまった. 全体を通して誤判定されてしまった箇所はあるもののタイミングは正しく出力されており, また, 全 4,371 フレーム中 4,350 フレームが正しく分類されているため精度は 0.9952 という結果となった.

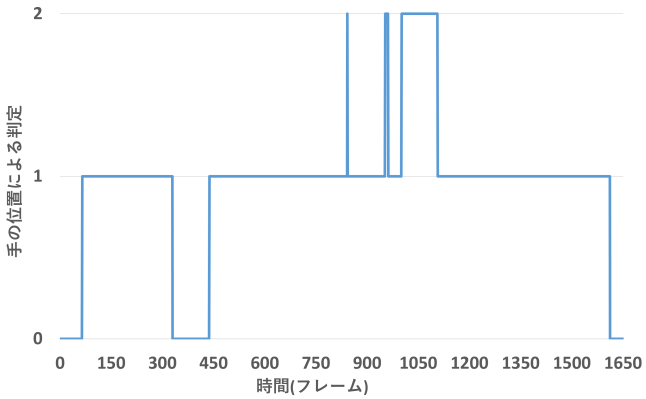

図 2　座標データでの手の位置の分類実験

4.4　時系列データによる手招き動作の分類実験

次に時系列データを用いて認識実験を行う. 実験データとして 8 分の動画約 8,000 フレームを用いて分類の精度を算出した結果, 図 3 のように出力された. これは体の向きが 45 度の際の計測結果で, 実験全体を通して分類と切り替わりのタイミングが一番正しく判定できたものである. 図 3 の縦軸については手招き動作の認識で上の図は 0 の時“other”で, 1 の時“beckoning”を表し, 下の図は角度を表しており, 横軸は時間をフレームで表している. これによると正確に手招きの判定ができていることがわかる. また, 角度の図は 0 と 360 度を行き来するため (0〜360) 表示を (−180〜180) 表示に変更している. 右手と左手の手招き動作の切り替わりによって約 20 度くらいのばらつきが発生し, これは左手の手招きをする時に若干左に向くため少し角度が小さくなってしまう. さらに, それぞれの手招き動作一回でも小さい山ができている. 実際に実験を行っている時の様子を図 4 に載せた.

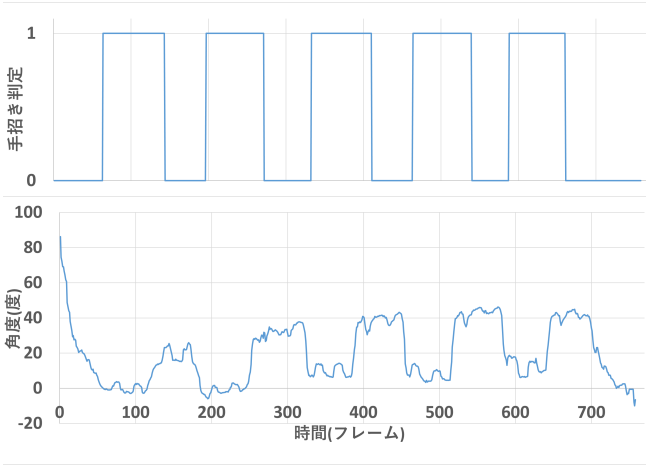

図 3　体の向きが 45 度の際の分類結果

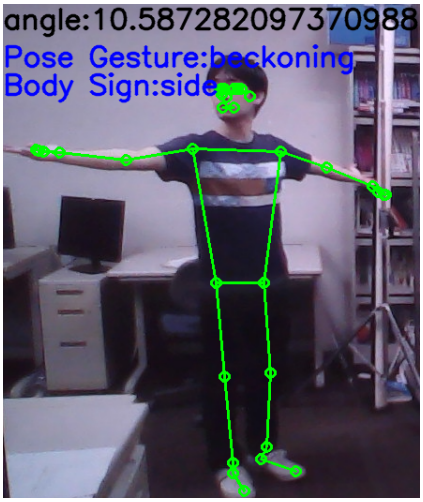

図 4　体の向きが 45 度の際の認識の様子

表 2　角度ごとの分類結果

角度 (度)	0		45		90		135		180		225		270		315	
分類	手招	その他	手招	その他	手招	その他	手招	その他	手招	その他	手招	その他	手招	その他	手招	その他
手招	652	23	382	0	591	42	286	64	478	1	554	0	502	31	796	0
その他	5	440	0	376	139	758	30	359	9	548	10	576	71	538	0	519
フレーム	1120		758		1530		739		1036		1140		1142		1315	
精度	0.975		1.000		0.882		0.873		0.990		0.991		0.911		1.000	
													全体		8780	0.952

次に, うまく認識をすることができなかった時の実験結果を図 5 に示す．これは体の向きが 90 度の時のもので, 上と下の図は 45 度の時と同様で真ん中の図は理想の波形を示している．縦軸と横軸の意味も同じで, 波形を見ると 45 度の時と比較して正しく分類できていないものが多くなってしまった．体の向きが 90 度であるため実際には 70〜110 度位で推移していることが望ましいが, 角度の図を見ると 40〜70 度と 40 度ほどずれが発生してしまった．

表 2 に 45 度と 90 度の時を含めた 45 度ごとのすべての認識結果と精度をまとめた．2〜4 行目は分類の結果の混同行列を, 5 行目はフレームの合計数を, 6 行目は角度ごとの精度を, 7 行目は全体の角度の精度を表している．これを見ると全体を通して 95% を達成している．表の中で精度が低いものの特徴として体が横を向き手のひらが手前に来ると奥側の肩の座標を正しく推定できず誤分類を起こしてしまうということが挙げられる．

しかし, 体が 270 度になると左の手のひらが体に隠れてしまう懸念があったがそれでも 91% と高い精度が得られた．これにより, 最大 40 度ほどのずれが発生するもののあらゆる方向での高精度な分類が可能である．

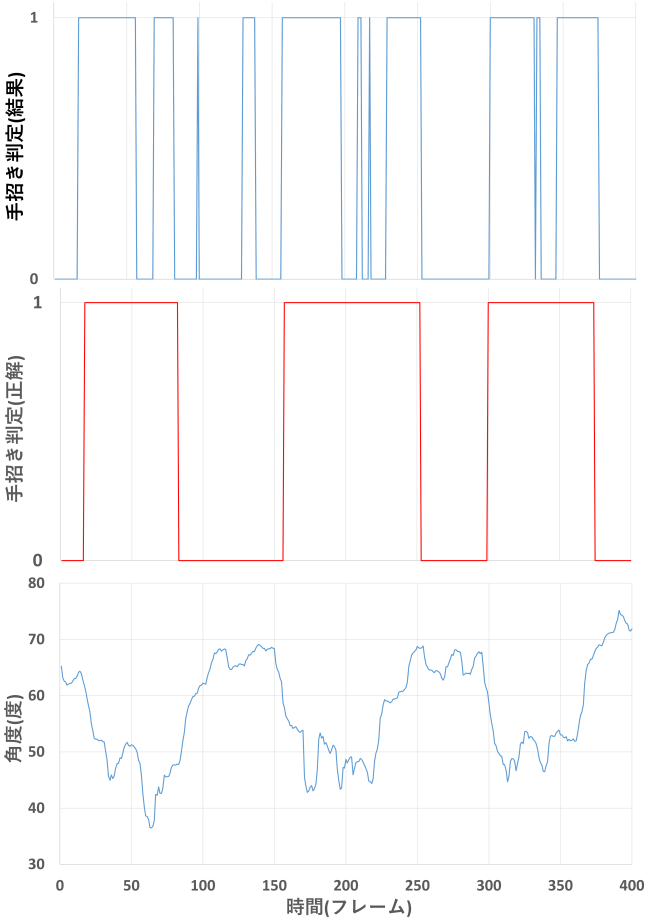

図 5　体の向きが 90 度の際の分類結果

5　むすび

本研究では警察官の手信号の認識について, 3 次元座標を用いることによって体の向きを求めることと, 座標データで手の位置を判定し時系列データで手招き動作の認識を行い 95 %の精度を達成した．また, [1] の研究で用いた 1 時間の動画データに対して本研究では 25 分と半分以下であることと, 手招き動作の検出に用いる骨格点を 18 箇所ではなく左手のひらのみで実施しつつも大きく精度を損なわないことで先行研究よりも少ないデータで警察官の手信号の認識ができた．

これにより手信号が行われている道路の形状がわかれば, 両肩からの体の向きと現在の手の位置と左手のひらの状態から指示を認識することができる．

参考文献

[1] 小野晋太郎, 木田侑, “自動運転のための警察官の手信号の認識システム,” 生産研究, 72 巻, 2 号, pp. 101-106, 2020.

[2] Taeseung Baek, Yong-Gu Lee, “Traffic control hand signal recognition using convolution and recurrent neural networks,” Journal of Computational Design and Engineering, Volume 9, Issue 2, pp.296–309, April 2022.

[3] 松田啓佑 飯塚博幸, “手話動作分類における RCNN モデルの性能評価と内部状態解析,” The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018.

[4] Curtis G. Northcutt, “Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks,” 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks