1R-04

深度カメラを用いたどの角度からもジェスチャを認識する研究

謝 寧† 松澤 智史†
東京理科大学 理工学部 情報科学科†

1. はじめに

近年、モーションセンサ技術の発展が著しく、Google が Soli というレーダー波を用いたセンサ技術を開発した。より細かい手の動きを読み取る事ができ、ジェスチャ認識への注目がますます高まる。従来、様々なハンドジェスチャ認識を行う手法が提案されており、高い精度を記録するものもあるが、認識デバイスに対して特定の手の向きにおいて、認識率が大きく低下してしまう。

本研究では、16 のジョイントの 3 次元座標を記録したハンドデータを用いて、6 種類のジェスチャを学習し、これらのジェスチャの対し、どの角度からもジェスチャが認識される手法を提案する。

2. 基礎知識

2.1 再帰型ニューラルネットワーク (RNN)

主に時系列データに対して用いられる。ある層の出力が記憶され、次の層でも使われるようなニューラルネットワークである。

2.2 長・短期記憶 (LSTM)

長期的に記憶を保持でき、学習の安定性も兼ね備えた RNN である。

2.3 サポートベクターマシン (SVM)

教師あり学習によってパターンを認識する手法である。学習データの中で、他クラスに最も近い位置にいるものを基準とし、そのユークリッド距離が最大となる位置に境界線を引くことで、クラス間のマージンが最大になるよう分類を行う。

2.4 ハンドデータの定義

利用したハンドデータは ICVL Hand Posture Dataset[*1] を用いた。手のひら、親指の付け根、親指の中心、親指の先、人差し指の付け根、人差し指の中心、人差し指の先、中指の付け根、中指の中心、中指の先、薬指の付け根、薬指の中心、薬指の先、小指の付け根、小指の中止、小指の先を 16 のジョイントとしている。(図 1) それぞれの (x,y,z) 座標が与えられる。ただし、x,y は単位がピクセルに対して、z は mm である。

2.5 認識するジェスチャ

パーから人差し指のみに変える、人差し指のみから小指のみに変える、パーからグーに変える、グーのままひねる、グーのまま左右にスライドをする、パーからグーにひねりながら変えるの 6 つのジェスチャを認識する。(図 2)

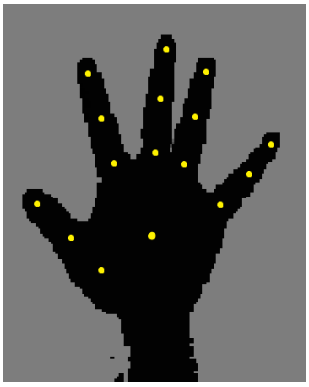

図 1: 16 ジョイントの位置
各点がジョイントの位置である。

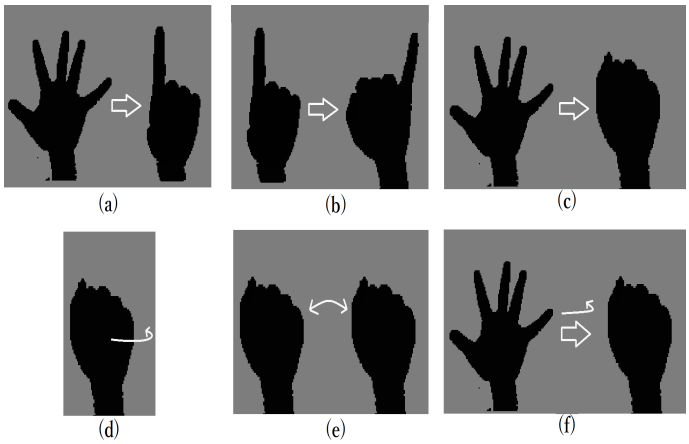

図 2: 認識する 6 つのジェスチャ

3. 関連研究

3.1 深度カメラを用いたジェスチャ認識システム

Ahn らの研究 [1] ではまず、閾値を設けて深度画像から手の部分を抽出した。次に、抽出した手の画像を、ヒストグラム平坦化で距離変換をし、掌をもとに骨格検出を行った。手を真上から見た場合、この細線化のアルゴリズムでは手が検出されない。そこで、深度の最小値を求め、この値に 55 を加えた数値を閾値として指先の検知を行った。55 は実験的経験によって求められた値である。最後に、SVM を用いてジェスチャの認識を行った。細線化のアルゴリズムでは、手の中心、掌の大きさ、腕と掌の軸、指の長さ、指の軸がデータとして必要となる。深度の最小値を用いたアルゴリズムでは、深度の最小値、指先の数、指先の面積と掌に対する割合がデータとして必要となる。この手法によって、一つのジェスチャを正面から見た場合と上からみた場合どちらからも、同じジェスチャと認識された。

この手法では、手の形のみを考慮しており、複数の

Gesture recognitioin in any direction with deep sensor
†Rei Sha, †Tomofumi Matsuzawa
†Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science

[*1] "ICVL Hand Posture Dataset"
https://labicvl.github.io/hand.html

形を組み合わせたりなどの動きに関しては考慮されていない。

3.2 RNN を用いたジェスチャ認識

手の骨格抽出をした二次元のジェスチャデータに対して、LSTM で時系列処理を行った。また、骨格情報と深度情報を独立して扱った。2 つの LSTM 層と、3 つの全結合層、1 つの出力層から成るネットワーク構造をしており、ミニバッチサイズがキーフレームの平均値である 32 で学習された。[2]

この手法では、データセットとして与えられた 14 つのジェスチャ以外を認識しようとすると精度が低下してしまう。また、カメラに対して手の向きが正面でなかった場合が考慮されておらず、動きとしては同じであるにもかかわらず認識できないジェスチャが生じてしまう。

4. 提案手法

本研究では、手の骨格データを用いて、手の形を変えたり、動かしたものを一つのジェスチャとして深度カメラを用いて様々なアングルから認識する手法を提案する。

4.1 手の形の分類

まず、ハンドデータを指の曲がっている位置の違いで分類し、ラベリングを行う。今回は、グーの状態を 0、人差し指のみ伸ばしている状態を 1、小指のみ伸ばしている状態を 2、パーの状態を 3 としてラベリングを行った。ラベル付けされたデータを教師データとテストデータに分割し、scikit-learn[*2] の one-versus-the-rest[*3] で多クラス SVM 分類を行う。

4.2 ジェスチャ認識

SVM で学習したモデルを時系列データとして扱い、LSTM で学習する。図 2 における (a) のような指の曲げ伸ばしのみのジェスチャの場合、変形する前と後を時系列データとして設定をする。(f) のようなひねる動作も加えられた場合、変形する前のデータを y 軸回転したものと、変形した後のデータを y 軸回転したものを時系列データとして間に加える。(図 3) 出力データは 6 つのカテゴリとして出力する。

5. 評価方法

Intel Realsense D435 の深度カメラを用いて手の骨格をタグ付けし、学習したモデルでテストを行い認識率の精度を求める。いくつかの手法と比較をし、考察する。

6. 結果

指を伸ばしきるや曲げきる、ひねりの動作は 180° 回しきるなど、メリハリのあるジェスチャに対しては認識が成功するが、中途半端な動作に関しては、認識が失敗となってしまった。(図 4) これは、手の形を分類する際に、指が半分まで曲がっている画像のラベリングを統一していなかったことが一因として考えられる。

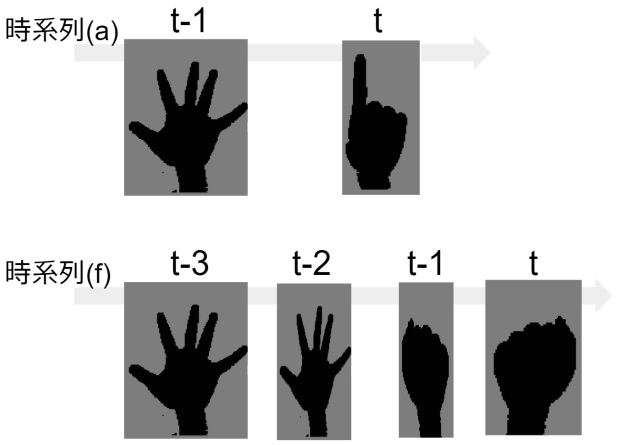

図 3: RNN による時系列データの例

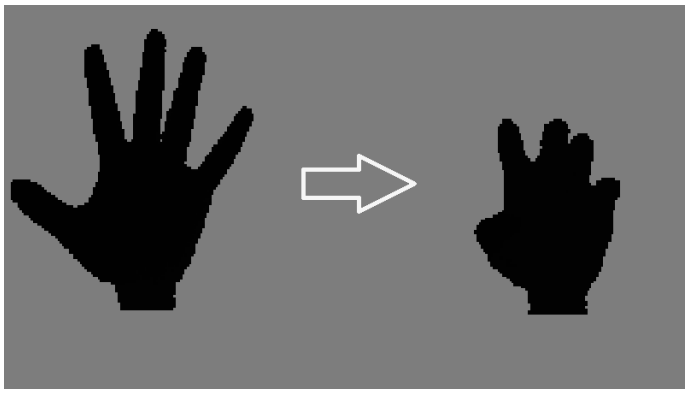

図 4: 認識失敗の例

7. まとめ

本研究では、深度カメラを用いて正面以外の角度のジェスチャも認識できるよう検討した。特に横向きのジェスチャを認識するのは、後ろの指が他の指によって隠されてしまうため、困難であった。しかし、深度カメラを用いて、手の骨格を抽出することで、手の回転動作などにも対応することが可能である。手の骨格に対するアノテーションは遮蔽が生じてしまい上手くデータがとれない場合があるため手動で行ったが、今後は、さらなるハードウェア発達により、効率よくデータがとれるようになるのではないかと期待できる。また、研究に関しては、より多くのパターンのジェスチャを認識し、汎用性を高める予定である。

参考文献

[1] Y.-K. Ahn, Y.-C. Park. The Hand Gesture Recognition System Using Depth Camera. In Proceeding of the 10th International Conference on Advances in Computer-Human Interactions(ACHI2017), pp.234-238, 2017.

[2] K. Lai, S.N. Yanushkevich. CNN+RNN Depth and Skeleton based Dynamic Hand Gesture Recognition. In Proceeding of the 24th International Conference on Pattern Recognition (ICPR2018). IEEE, pp3451-3456, 2018.

[*2] “scikit-learn” https://scikit-learn.org/stable/

[*3] “one-versus-the-rest” https://scikitlearn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html