Vol.2018-CVIM-210 No.26
2018/1/18

VR環境における多様なつかみ動作の認識

天野 祐嗣[*1]　　小室 孝[*1]

Recognition of Various Hand Grasping in Virtual Reality

Yuji Amano[*1] and Takashi Komuro[*1]

Abstract – 本論文では、手の姿勢や動きが制約されることのない、**VR** 環境における多様なつかみ動作を認識する手法を提案する。手の姿勢を経由せずにつかみ動作を直接認識することで、仮想物体の位置や形状に適した自由なつかみ方で操作できるようになる。**HMD** とデプスカメラを組み合わせた装置を作成し、手のつかみ動作に特化したデータセットの作成を行った。つかみ動作の認識には **CNN** を用いることで、ホールドアウト検証において 97.9 % の認識率を達成した。提案手法を **VR** 環境に統合することでより自然なつかみ方で仮想物体とインタラクションできる可能性を示した。

Keywords : virtual reality, hand action recognition, convolutional neural network

1 はじめに

近年、Virtual Reality (VR; 仮想現実感) のための Head Mounted Display (HMD) やコントローラを組み合わせた製品が開発、販売されている。仮想空間内に表示される仮想物体に対してつかむなどの操作をするために、手による操作を認識する手法が必要とされる。これらの製品では、ユーザの手による操作をコントローラのボタンなどの操作に対応付けて検出している。そのため、ユーザは行いたい操作をコントローラの操作に変換する必要があり、自然な操作の妨げとなることがある。この問題を解決するために、手による操作を直接認識するための手法が求められている。

ビジョンベースの手法で手の正確な姿勢を推定することは困難であることが知られている。手の自由度が高く、多様な姿勢を取ることができることがこの問題を引き起こしている。また、手の形状から自己遮蔽が起きやすいことも原因の一つである。つかみに限定しても多様な姿勢が存在することが知られている**??**。

グローブやマーカーを装着することでこの問題の単純化を図った研究が存在する。Wang らは手の部位によって色が異なるグローブを装着することで、手の姿勢推定を単純化する手法を提案した [1]。Buchmann らは手に AR マーカーをつけることで、数か所の関節の位置を取得する手法を提案した [2]。これらの研究では、専用の装置を用いることで手の姿勢推定を単純化することに成功した。しかしながら、これらの手法は使用の度に装置の装着が必要なため、煩わしさを感じることがある。

CNN (Convolutional Neural Network) を用いた深層学習は、画像を扱うタスクにおいて state-of-the-art の性能を示すことが知られている [3]。Tompson らは CNN を用いることでリアルタイムで動作する、手の姿勢推定の手法を提案した [4]。手の画像から数か所の関節の位置を推定し、手の姿勢を最適化問題として復元して得ることができる。Sinha らは CNN を手の画像の次元削減に用いることで低次元の特徴量を計算し、対応する手の姿勢をデータベースから選択する手法を提案した [5]。

手の姿勢から操作を認識するためには、もう一段階の推定処理が必要である。手の姿勢が十分正確に推定できていない場合、操作を正しく認識することができない。そのため、手の姿勢を経由することなくジェスチャを認識する手法が研究されている。Ren らはデプスカメラを使用して、10 種類の姿勢と動きのパターンをジェスチャとして認識する手法を提案した [6]。Song らはモバイル端末に組み込まれたカメラを使用して、7 種類のジェスチャを認識する手法を提案した [7]。Molchanov らは RGB-D カメラとレーダーを組み合わせた照明変動に強い、11 種類のジェスチャ認識の手法を提案した [8]。

特定の姿勢と動きのパターンをジェスチャとして認識するため、手の姿勢や動きが予め定められたパターンに限定される。このことが自然な操作の妨げとなる恐れがあり、制約がないことが望ましい。人間の行動認識の研究では、姿勢や動きが統一されていない行動を認識する手法が存在する。Wang らはオプティカルフローに寄って計算される人間の動きの軌道を用いて、人間の行動を認識する手法を提案した [9]。Wang らは人間の動きの軌道を CNN に入力することで、従来の特徴量を用いた手法より高い性能を示すことを明らかにした [10]。これらの研究で姿勢や動きが統一されていない人間の行動を認識できることが示されたが、

[*1]埼玉大学 / Saitama University

手に特化した手法は知られていない。

そこで、手の姿勢や動きが制約されることのない、VR 環境における多様なつかみ動作を認識する手法を提案することを目的とする。本研究では、手による操作としてつかみを対象とする。つかみ動作を直接認識することで、仮想物体の位置や形状に適した自由なつかみ方で操作できるようになる。多様なつかみ動作を認識するために深層学習を用いる。深層学習を用いることで、手の階層的な構造を考慮した特徴量を抽出して認識が行われることが期待できる。

2　深層学習によるつかみ動作の認識

深層学習を用いたつかみ動作の認識手法を提案する。入力としてデプスカメラから得られるデプス画像を使用する。前処理としてデプス画像の 2 値化を行い、手が存在する奥行き範囲に含まれる部分とその他の背景部分に分離するして使用する。デプス画像にはユーザの手が写っている場合と写っていない場合が考えられる。手が写っている場合、つかみ、非つかみの 2 通りに分けることができる。デプス画像に対して CNN を適用し、手を含まない、つかみ、非つかみの 3 クラスに分類することでつかみ動作の認識を行う。

CNN の構造を表 1 に示す。入力は 80×60 ピクセルの 1 チャンネルグレースケール画像である。6 層の畳み込み層 (Conv) に続いて Global Average Pooling 層 (GAP)、全結合層 (FC) で 3 クラスへの分類を行う。畳み込み層の後段では Batch Normalization を行う。アクティベーション関数には ReLU を用いる。

表 1　ネットワーク構造

Type/Stride	Filter Shape	Input Size
Conv/s1	$3 \times 3 \times 4$	$80 \times 60 \times 1$
Conv/s2	$3 \times 3 \times 8$	$80 \times 60 \times 4$
Conv/s2	$3 \times 3 \times 16$	$40 \times 30 \times 8$
Conv/s2	$3 \times 3 \times 32$	$20 \times 15 \times 16$
Conv/s2	$3 \times 3 \times 64$	$10 \times 8 \times 32$
Conv/s2	$3 \times 3 \times 128$	$5 \times 4 \times 64$
GAP	3×2	$3 \times 2 \times 128$
FC	128×3	$1 \times 1 \times 128$

3　評価実験

3.1　データセット

手のつかみ画像に特化したデータセットを作成した。画像の撮影に使用する装置を図 1 に示す。VR 環境を提供する HMD の上方にデプスカメラを固定した装置を使用した。HMD は Oculus Rift を用いた。デプスカメラは SoftKinetic DepthSense DS325 を用いた。

図 1　撮影に使用した装置

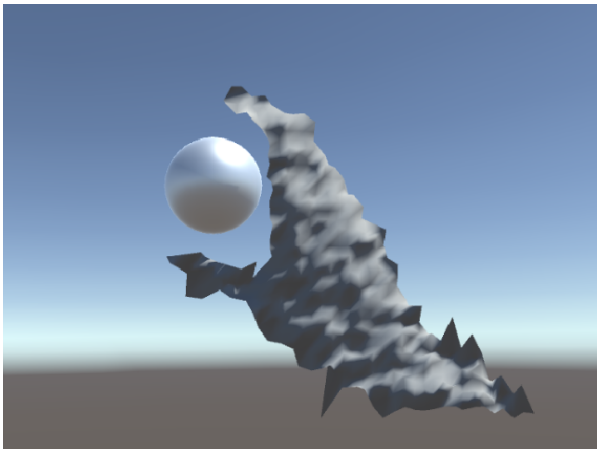

図 2　撮影中の仮想空間内の様子

60 fps で 320×240 ピクセルのデプス画像を取得でき、$10 - 100$ cm の近距離の範囲で使用できる。

手の画像の撮影は 9 人の参加者 (男性 9 人) に対して行った。撮影中の仮想空間内の様子を図 2 に示す。参加者は着座した状態で撮影に参加した。仮想物体は視野の中心に固定されて表示される。3 cm 相当の球の仮想物体を使用した。また、参加者の手も現実空間との位置と整合性が取れるように VR 環境内に表示した。HMD とデプスカメラの位置をキャリブレーションすることで実現した。参加者が VR 内に表示されている仮想物体をつかむとき、はなすときの手をそれぞれ撮影した。このとき、参加者に対して仮想物体の具体的なつかみ方に指示は行わないことで、多様なつかみ方が含まれるようにした。VR 環境のレンダリングには Unity を使用した。

撮影したデプス画像を手を含まない、つかみ、非つかみの 3 クラスに分類し、80×60 ピクセルにリサイズ、2 値化する前処理を行ったものをデータセットとした。装置の装着位置と腕の長さの関係から、2 値化の閾値を 80 cm に設定した。

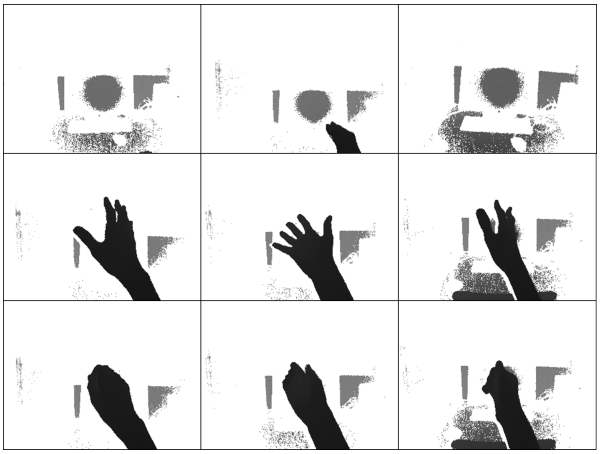

図 3　使用した画像例 (上段: 手を含まない画像、中段: つかみ画像、下段: 非つかみ画像)

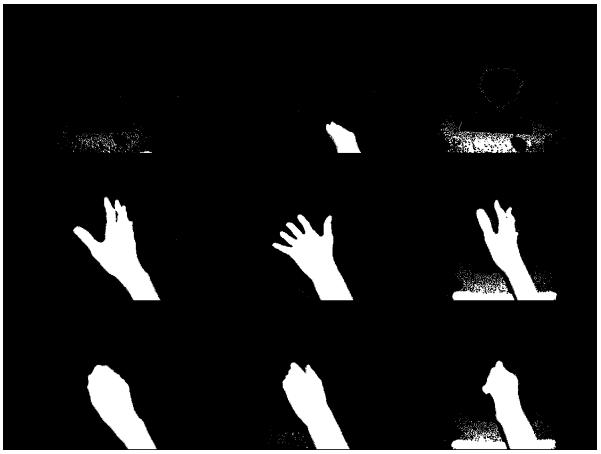

図 4　2 値化画像例 (上段: 手を含まない画像、中段: つかみ画像、下段: 非つかみ画像) 閾値: 80 cm

3.2　訓練

訓練には手を含まない画像として約 4 千枚、つかみ画像として約 1 万 6 千枚、非つかみ画像として約 1 万 6 千枚を用いた。訓練に用いた画像の例を図 3、2 値化した画像を図 4 に示す。上段は手を含まない画像、中段はつかみ画像、下段は非つかみ画像である。2 値化の前処理によって背景の多くが除去することができている。CNN は Chainer (v2.1.0) を用いて実装した。最適化手法として Adam を使用し、50 エポックで終了した。

3.3　認識率

データセットをランダムに 2 分割したホールドアウト検証とユーザに対する一つ抜き交差検証で認識率の評価を行った。

入力画像に対して 3 クラス中の正しいクラスを出力できた場合を正解とすると、手のつかみ動作の認識率はホールドアウト検証で 97.8 %、一つ抜き交差検証で 71.9 % であった。ホールドアウト検証による分類結果の混同行列を表 2、一つ抜き検証による混同行列を表 3 に示す。手を含まない画像に対する認識率が低い。手がカメラの撮影範囲に入るときや、出るときは手の一部のみが画像内に写るため、識別に失敗すると考えられる。加えて、つかみ画像、非つかみ画像に対して手を含まない画像の数が少ないことも原因の一つと考えられる。

表 2　ホールドアウト検証による混同行列

正解 \ 出力	手を含まない	つかみ	非つかみ
手を含まない	96.0 %	3.95 %	0.00 %
つかみ	0.06 %	97.9 %	2.04 %
非つかみ	0.00 %	1.91 %	98.1 %

表 3　一つ抜き検証による混同行列

正解 \ 出力	手を含まない	つかみ	非つかみ
手を含まない	66.5 %	24.9 %	8.58 %
つかみ	11.3 %	71.5 %	17.2 %
非つかみ	4.14 %	22.3 %	73.5 %

一つ抜き交差検証によるユーザごとの認識率を表 4 に示す。ユーザごとの認識率にばらつきがあり、一部のユーザ (ユーザ 3) で特に低い認識率を示した。これは、本研究で作成したデータセットで訓練を行った CNN では、ユーザの手の形状やつかみ方の癖による個人差を吸収できていないためであると考えられる。データセットに含まれる手の画像のユーザ数を増やすことで、個人差に対する汎化性能を向上させることができると考えられる。十分なユーザ数によるデータセットであれば、ホールドアウト検証による認識率に近づくと思われる。

表 4　一つ抜き交差検証によるユーザごとの認識率

1	2	3	4	5
72.9 %	77.7 %	56.7 %	72.5 %	81.1 %
6	7	8	9	
70.2 %	73.6 %	76.3 %	63.9 %	

訓練中の認識率の変化を図 5 に示す。実線がホールドアウト検証による認識率、点線が一つ抜き交差検証による認識率である。一つ抜き交差検証による認識率は、ユーザごとの結果の平均をとった。ホールドアウト検証による認識率は訓練開始時から上昇し、30 エポック程度で上限に達している。一方、一つ抜き交差検証による認識率は訓練開始時からほとんど上昇していない。これは、データセットが不十分であるため個人差を吸収できていないことが原因であると考えられる。

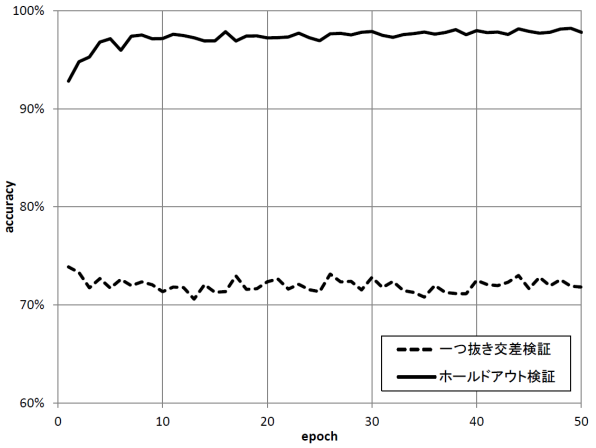

図 5　訓練中の認識率の変化

3.4　実行時間

手のつかみ動作の認識にかかる時間の測定を行った。実行時間は 1,000 枚の処理にかかる時間の平均から算出した。計測には Intel Core i7 7700 CPU、NVIDIA GeForce GTX 1070 を搭載したコンピュータを使用した。実行時間は 76.3 ms、337 fps であった。リアルタイム動作の目安である 60 fps に届いているため、VR 環境におけるインタラクションの手法として利用することができると思われる。

4　おわりに

本研究では、手の姿勢や動作が制約されることのない、VR 環境における多様なつかみ動作を認識する手法を提案した。つかみ動作を直接認識することで、ユーザの自由なつかみ方の操作を認識することができる。また、VR 環境においてはつかみ動作の対象となる仮想物体の位置や形状が既知であるため、位置や形状に合わせたより自然なつかみ方を学習することができる可能性がある。今後の課題として、つかみ動作の認識を VR 環境に統合して VR 内に表示された仮想物体とのインタラクションを実現することや、評価実験を行って使いやすさの評価を行うことが挙げられる。

参考文献

[1] Robert Y Wang and Jovan Popović. Real-time hand-tracking with a color glove. In *ACM Transactions on Graphics*, Vol. 28, p. 63, 2009.

[2] Volkert Buchmann, Stephen Violich, Mark Billinghurst, and Andy Cockburn. Fingartips: gesture based direct manipulation in augmented reality. In *Proceedings of the ACM International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia*, pp. 212–221, 2004.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[4] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, Vol. 33, No. 5, p. 169, 2014.

[5] Ayan Sinha, Chiho Choi, and Karthik Ramani. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4150–4158, 2016.

[6] Zhou Ren, Junsong Yuan, and Zhengyou Zhang. Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *Proceedings of the ACM International Conference on Multimedia*, pp. 1093–1096, 2011.

[7] Jie Song, Gábor Sörös, Fabrizio Pece, Sean Ryan Fanello, Shahram Izadi, Cem Keskin, and Otmar Hilliges. In-air gestures around unmodified mobile devices. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*, pp. 319–329, 2014.

[8] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Kari Pulli. Multi-sensor system for driver's hand-gesture recognition. In *Proceedings of the IEEE Conference and Workshops on Automatic Face and Gesture Recognition*, Vol. 1, pp. 1–8, 2015.

[9] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551–3558, 2013.

[10] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4305–4314, 2015.