小型広角カメラを用いた視線方向の推定

田中 恭友 *　　小池 英樹 *

概要. 近年，ウェアラブルなアイトラッカーの普及に伴い多くの分野において視線追跡の利用が進んでいる．しかし，スポーツ時の利用においては小型カメラを眼球周辺に配置することは危険が伴う．一方，全天球または全天周映像の撮影が可能な広角カメラの小型化が進んでいる．この小型広角カメラを胸部に固定することで，装着者の頭部と視野の映像を同時に撮影することができる．こうした映像から装着者が顔を向けている方向を推定することで，装着者の視線方向を推定することが可能である．本稿では，胸部に固定した小型広角カメラから視線方向を推定する手法の提案と評価結果について述べる．現在の実装では，固定の環境下でのみではあるが一定の精度で推定を行えることが確認された．今後は，多様な環境でも頑強に推定が行えるように改良を加えていき，推定結果を利用したアプリケーションを作成する予定である．

1　はじめに

近年，ウェアラブルなアイトラッカーの普及により，多くの分野で視線計測の活用が進んでいる．その中でも注目を集めているものとして，スポーツ分野への応用がある。例えば，スポーツにおいて周囲の環境をうまく認識するために，初級者と上級者では視線移動においてどのような違いがあるのかを調査することで，技術向上を支援するという試みがなされている．しかし，スポーツの種類によっては，アイトラッカーを利用することが困難であるものもある．アメリカンフットボールやスキーなど，頭部にゴーグルやヘルメットといったデバイスを装着するスポーツにおいては，小型カメラを装着することは不可能ではない．しかし，その他のスポーツ，例えばサッカーなどのスポーツにおいては，眼球周辺にカメラを配置することによる問題点が存在する．例えば，こうしたデバイスを眼球周辺に装着したままスポーツを行うことには危険性が伴う．また，小型カメラが常に視界に存在するために視野が一部制限されてしまうという問題点もある．こうした視野の制限は，プレイに支障を及ぼすだけではなく，本来の計測対象である自然な視線の移動を阻害してしまうことも考えられる．

一方で，複数のレンズを利用することで手軽に全天球映像の撮影が可能な広角カメラの普及が進んでいる．こうしたカメラは，従来のものに比べて小型化が進んでいるため，撮影者の身体に固定して撮影することが可能である．例として，半天球撮影が可能な小型広角カメラを撮影者の胸部に固定した際の画像を図 1 に示す．

図 1 を見るとわかるように，この映像にはカメラ装着者の頭部の一部（顎周辺）に加えて，装着者の視線の先にある物体も写り込んでいる．我々は，このような胸部カメラからの撮影映像を解析することで，カメラ装着者の視線方向の推定，並びにカメラ装着者の注視物体の推定が行えるのではないかと考えた．そこで本論文では，胸部に固定した小型広角カメラの映像から，装着者の顔方向を推定することにより，装着者のおおよその視点位置を類推することを目的とする．このとき，映像にはカメラ装着者の眼周辺は写っていないため，厳密に視線方向の推定を行うことは不可能である．しかし，人がある点を注視する際に視線の方向と頭部の方向には線形性の関係があることが Fang らによって示されている [2]．また，この関係性を利用することで頭部の方向から視線方向をより正確に推定する手法も提案されている [9]．したがって，頭部の方向の推定が行えれば一定の精度で視線方向を推定することは可能であるため，本論文では，カメラ装着者の顔方向をカメラ装着者の視線方向として扱うものとする．

図 1: 胸部に固定した半天球カメラからの映像例

2　関連研究

2.1　顔器官検出

目鼻や口など，人間の顔の中でも特徴的な領域の位置を検出する処理を顔器官検出という．

* 東京工業大学

映像中から顔器官を検出する手法には様々なものがあるが，その認識精度の高さから，近年では深層学習を利用した手法が注目されている [7]．これらの手法はリアルタイムでの検出が可能であり，顔の輪郭や目鼻などの器官を正確かつ頑強に検出することが可能である．また，横顔や遮蔽物がある場合などの顔の一部が隠れている場合であっても隠れた器官の位置を予測し検出が行える手法も提案されている [8]．

顔器官の位置を正確に検出することができれば，本稿で目的としている顔の向きの検出は容易に行える．しかし，本稿で認識対象とする映像では，カメラが頭部を俯瞰で撮影するため，顎部分が大きく映り，目鼻口などの器官は多少見える程度となる．すなわち，上記の手法で検出対象としている映像とは頭部の映像は大きく異なっており，顔器官検出の手掛かりとなる部位がない．したがって，上記の手法をそのまま適応するだけでは今回目的とする顔方向の推定は行えない．

2.2 ユーザ中心の動作認識

動作認識可能を特定の環境に寄らず行えるという利点があるため，小型カメラを身体に装着してユーザの動作情報を認識する研究が数多く行われている．Mistry らはカメラと小型プロジェクタを身体の前面に装着し，環境に縛られないハンドジェスチャによるインタラクションを提案した [5]．Helge らは装着者の全身が映りこむような位置に小型広角カメラを 2 台配置し，その映像から装着者の骨格情報を解析する手法を提案している [6]．Chan らは胸部に固定した広角カメラの映像からのジェスチャ認識手法を提案している [1] が，認識対象は手や足などであり，装着者の頭部は認識対象となっていない．

2.3 一人称視点映像の解析

小型のカメラを頭部に装着することで撮影した装着者の視点を再現した映像を一人称視点映像と呼ぶ．撮影者の注目物体が現れることや，動作の特徴が映像の動きに反映されるという特徴を持つため，一人称視点映像に対して様々な認識を行う研究が盛んに行われている [3][4]．

本研究で利用する全天周カメラの映像には，装着者の頭部だけでなく，注視領域も含まれている．従来の一人称視点映像の成果と本稿の貢献を組み合わせることで，カメラ装着者の注目物体とその物体に対するジェスチャの認識といった拡張が考えられる．

3 視線方向の推定

3.1 問題設定

本研究の目的は，映像中のユーザの注視点を推定するために，ユーザの顎部分の映像を手掛かりとしてユーザの視線方向を大まかに推定することである．本稿で認識の対象とする映像は半天球映像の撮影が可能なカメラである PIXPRO SP360 4K を使用して撮影した．このカメラを図 2 に示すように胸部に固定し撮影を行う．このカメラの撮影画角は水平方向 360°，垂直方向が　235°である．撮影した映像は図 1 に示したように，円周魚眼で撮影された映像になる．この映像中から装着者の頭部が映る部分を切り出し，認識対象の画像とする (図 3)．このとき推定に利用する領域は動的に変更することはなく，あらかじめ定めた固定の座標位置を利用する．

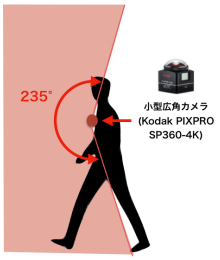


図 2: カメラの撮影範囲

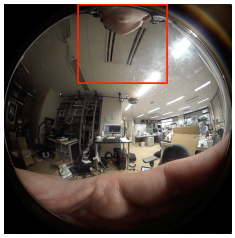


図 3: 認識対象領域

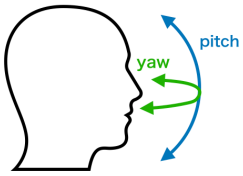


図 4: 顔方向の定義

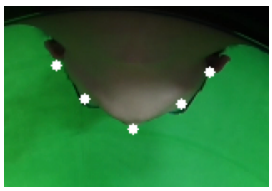


図 5: 顎の位置

こうして得られた装着者の頭部の画像を入力として，目的とする装着者の視線方向（顔方向）の推定を行う．視線方向は，装着者が真正面を向いている場合を $[0, 0]$ とした頭部の回転量 $[yaw, pitch]$ と定義する (図 4)．

3.2 認識手法

画像認識においては，CNN(Convolutional Neural Network) を利用した深層学習が一定の成果を挙げている．今回の問題においては，カメラの固定位置と認識対象（ユーザの顔）までにある程度の距離があること，また屋外環境での利用も想定していることから，頑強な検出のためには映像に現れる環境光などのノイズの影響を考えなければならない．そこで我々は，CNN を利用した機械学習を用いて顔方向の推定を行うことにした．

学習方法

目的の推定値は頭部の回転の 2 値 (ピッチ，ヨー) であるため，この問題は画像を入力とした回帰問題になる．更に，学習を効率よく進めるために，視線方向の手掛かりとして顎の輪郭 5 点の検出 (図 5) を

同時に行うことにした．

この問題を解くために利用するネットワークを図 6 に示す．このネットワークは，上述の推定対象の画像を 227[pixel]×227[pixel] に縮小したものを入力とし，ピッチ・ヨーの 2 値および顎の輪郭上の特徴点 5 点の座標値 10 個の計 12 個の値を出力するものである．

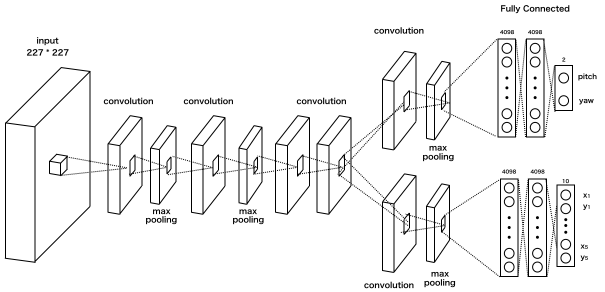


図 6: ネットワーク

学習に用いる損失関数 L には，目的とする顔方向の回転量の損失 L_{rot} と顎の特徴点の座標位置の損失 L_{jaw} の和を用いる．

$$L = L_{rot} + L_{jaw} \tag{1}$$

目的とする顔方向の回転量を $\boldsymbol{v} = [yaw, pitch]$ と表し，推定結果を $\boldsymbol{v}_e$，正解データを $\boldsymbol{v}_g$ とすると，L_{rot} は式 2 で表せる．

$$L_{rot} = \frac{1}{N}\sum_{i=0}^{N} ||\boldsymbol{v}_e - \boldsymbol{v}_g||^2 \tag{2}$$

一方，各顎の特徴点 p_i の画像中の座標値を $[x_i, y_i]$ とすると，今回の検出に使用する特徴点の数は 5 点なので，$\boldsymbol{j} = [x_1, y_1, ..., x_5, y_5]$ とすると L_{jaw} は

$$L_{jaw} = \frac{1}{5}\sum_{i=0}^{5} ||\boldsymbol{j}_e - \boldsymbol{j}_g||^2 \tag{3}$$

となる．以上の式を用いることで，L を最小化させるように学習を進めていく．

データセットの作成

本研究の目的は，多様な環境でもカメラ装着者の視線方向を推定することである．その前段階として，推定に最もノイズとして影響し得るだろう背景映像を単色の固定とした状態で目的とする推定が行えることを確認する．そのために，背景を単色の布で覆った環境（図 7）で撮影した画像（図 8）を学習させた．

また，正解データの取得方法について述べる．頭部の回転量はモーションキャプチャシステムの利用により正確な値を得ている．このとき，トラッキングに利用するマーカーが画像中に映り込まないように，カメラに映らない頭部の死角に配置している．

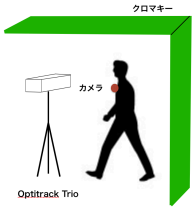


図 7: 撮影環境

図 8: 学習画像

顎の輪郭上の特徴点の座標は各フレームの画像に対して手作業で付与している．

4 認識結果

学習結果を利用して推定精度を確認するための実験を行なった．

この実験のための動画には学習用のデータセットに含まれていないものを利用している．60[FPS] で撮影した動画を検証用に 1500[Frame] 利用した．撮影に使用したカメラやカメラを固定する位置は学習に用いたデータセットと同じであり，正解データ (視線方向) の取得も学習用のデータセットと同じくモーションキャプチャの結果を利用している．

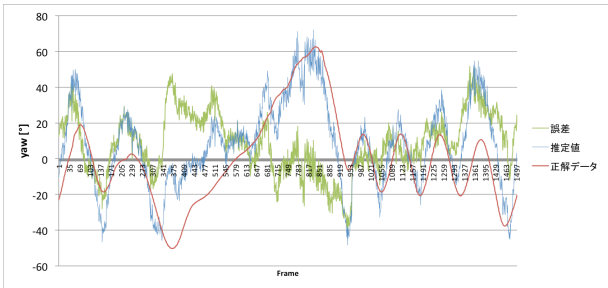
図 9 に推定結果のグラフを示す．yaw と pitch，どちらの推定においても，モーションキャプチャで取得した正解データの値とはほとんどのフレームにおいて誤差 20°以内に収まっている． yaw の推定においては誤差が 20〜40°の範囲になっているフレームもあるが，推定値と正解データの符号は一致している（右向きを左向きと誤推定していることはない）．すなわち，未だ認識制度に対しては改善の余地があるが，提案したネットワークの利用により目的とする視線方向のおおまかな推定を行える可能性が示されたといえる．

5 考察

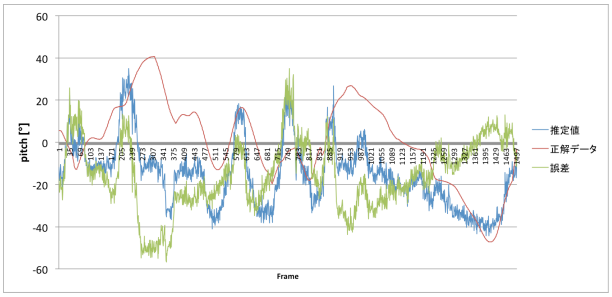
5.1 データセットの拡充

今回の検証に用いたデータセットは背景がコントロールされた状況においての検証であった．本来の目的である環境によらない認識を目指すためには背景や環境光が変化しても同じく認識を行える必要がある．そこで我々は，今回使用した映像にクロマキー合成を用いて様々な背景の場合のデータセットを作成することを考えている（図 10）．

また，今回のデータセットの作成にあたっては，顎の輪郭上の点の配置を手作業で行っていた．手作業でのラベリングはデータセットの拡大に対して多くのコストがかかるだけでなく，その質の低下の原因ともなる．そこで，人間の頭部を再現した 3D モデルを利用して学習データを作成する試みも進めている．



(a) yaw



(b) pitch

図 9: 検証結果

5.2 今後の展開

提案手法では，カメラ装着者の顔方向を視線方向として定義し，推定を行なった．推定した顔方向と全天周映像の座標値の対応付けを行うことで，カメラ装着者の注視点の映像も特定することができる．さらに，ユーザの後方にも全天周カメラを装着することでカメラ装着者を中心とした周囲 360 度の映像も撮影可能である．現在，この注視点の映像を用いた様々なアプリケーションを考案中である．例えば，スポーツプレイ中のユーザの注視点を特定し，上達のための示唆を与えるアプリケーションや，360 度映像中のユーザの注視点を記録するライフログシステムなどを考えている．

6 まとめ

本稿では小型広角カメラを用いたウェアラブルデバイスを提案し，そのデバイス上で装着者の視線方向をおおまかに推定するための手法を提案した．提案手法により装着者の頭部の映像が顎部分のみしか見えず，目鼻といった特徴的な器官が写っていない場合においても目的としている視線方向の推定を行える可能性が示された．現状の推定器では限られた環境下でしか推定が行えないため，よりデータセットの拡充および学習方法の改善により屋外環境でも頑強な推定を目指す．

(a) 元画像　(b) 背景合成後　(c) 背景合成後

図 10: クロマキー処理による学習データのかさ増し

参考文献

[1] L. Chan, C.-H. Hsieh, Y.-L. Chen, S. Yang, D.-Y. Huang, R.-H. Liang, and B.-Y. Chen. Cyclops: Wearable and Single-Piece Full-Body Gesture Input Devices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 3001–3009, New York, NY, USA, 2015. ACM.

[2] Y. Fang, R. Nakashima, K. Matsumiya, I. Kuriki, and S. Shioiri. Eye-Head Coordination for Visual Cognitive Processing. *PLOS ONE*, 10(3):1–17, 03 2015.

[3] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast Unsupervised Ego-action Learning for First-person Sports Videos. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pp. 3241–3248, Washington, DC, USA, 2011. IEEE Computer Society.

[4] M. Ma, H. Fan, and K. M. Kitani. Going Deeper into First-Person Activity Recognition. *CoRR*, abs/1605.03688, 2016.

[5] P. Mistry and P. Maes. SixthSense: A Wearable Gestural Interface. In *ACM SIGGRAPH ASIA 2009 Sketches*, SIGGRAPH ASIA '09, pp. 11:1–11:1, New York, NY, USA, 2009. ACM.

[6] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. EgoCap: Egocentric Marker-less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph.*, 35(6):162:1–162:11, Nov. 2016.

[7] Y. Sun, X. Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3476–3483, June 2013.

[8] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face Alignment Across Large Poses: A 3D Solution. *CoRR*, abs/1511.07212, 2015.

[9] 沖中大和, 満上育久, 八木康史. 人の眼球と頭部の協調運動を考慮した視線推定. 研究報告コンピュータビジョンとイメージメディア（CVIM）, pp. 1–8, may 2016.