Vol.2022-CG-188 No.28
Vol.2022-DCC-32 No.28
Vol.2022-CVIM-231 No.28
2022/11/19

人物姿勢に着目した後ろ向き人物の注視領域推定手法の検討

弓矢 隼大[1,a)]　出口 大輔[1]　川西 康友[2,1]　村瀬 洋[1]

概要：本研究では，画像中に後ろ向きに写る人物が注視している商品領域を推定する手法を提案する．これまでに研究されてきた人物の注視領域推定手法は，人物の顔領域から得られる視線や顔向きを利用している．対象の人物が後ろ向きの場合，既存の技術を適用できないが，我々人間は姿勢から対象の人物が画像内のどの領域を見ているかを推測できる．ただし，人の姿勢のとり方には個人差が存在し，姿勢からの注視領域推定には曖昧性が存在する．そこで，後ろ向き人物から取得可能な 3 次元骨格座標を利用する姿勢の取り方の個人差を考慮した注視領域推定手法を提案する．提案手法の性能を評価するため，人物が棚上の物体を注視している様子を撮影し，3 次元骨格座標と注視対象を紐付けたデータセットを構築した．このデータセットを用いて提案手法の有効性を確認した．

1. はじめに

人物が何に注視を向けているかを推定する注視領域推定は，マーケティングにおける商品への興味度合いの調査といった様々な活用が期待される重要な技術である．このような背景から，画像中の人物の注視領域を推定する手法がいくつか提案されている．Fridman ら [1] は，対象となる人物の顔画像から顔の向きを抽出することで注視領域を精度良く推定する手法を提案している．しかし，対象となる人物が後ろ向きの場合は顔画像が取得出来ず，注視領域を推定することができないという問題がある． 図 1 に後ろ向きの人物の画像の例を示す．これを見ると，何かを注視している人物の姿勢は，その対象の位置によって頭の向きを変化させたり，低い位置の対象の場合は屈んだ姿勢を取るといったように，注視対象によって姿勢が変化することがわかる．また，姿勢は Azure Kinect[*1] 等を用いることで後ろ向き人物からでも取得可能である．しかし，注視時の姿勢のとり方は人によって異なることが多く，同じ領域を注視していても異なる姿勢を取る場合が多く存在する．異なる姿勢だと同じ領域を注視していても異なる特徴をもつため，単純に姿勢から注視領域を推定することは難しい．そこで本研究では，姿勢の個人差を考慮しつつ後ろ向き人物の注視領域を推定可能な手法を提案する．具体的には，注視領域ラベルを用いた距離学習によって，同じ領域を注視している姿勢情報が特徴空間上で近づくような埋め込み表現に変換し，それを利用して個人差を考慮した注視領域推定を行う．

本研究の注視領域推定における貢献点は以下である．

- 顔情報が取得できない後ろ向き人物に対する姿勢情報を用いた注視領域推定の実現
- 距離学習を利用した姿勢の個人差を考慮した特徴空間上で同じクラスの特徴同士が近づくような埋め込み方法の提案

2. 関連研究

2.1 一般的な注視方向推定

Nonaka ら [2] は，人の持つ視線，頭，体の協調性に着目し，頭の位置や姿勢の時系列的な情報から注視方向を推定する手法を提案している．この手法では，複数の状況下で撮影された 3 次元視線をアノテーションした監視カメラ映像データセットを構築し，それを用いて頭と体の向きの尤度と視線方向の条件付き分布をニューラルネットワークでモデル化する．それによって，オクルージョンがあるような状況でも 3 次元視線方向を推定可能にしている．しかし，時系列情報に依存しており，単一フレームの情報から推定することはできていない．

2.2 人物の骨格情報を用いた注視領域推定

Kawanishi ら [3] は，画像上の人物から取得した骨格情報を用いて注視領域を推定する手法を提案している．この手法では注視領域に応じて人物姿勢が変化することに着目し，OpenPose [4] により取得した骨格情報を Deep Neural

[1] 名古屋大学 情報学研究科
[2] 理化学研究所 GRP
[a)] yumiyh@vislab.is.i.nagoya-u.ac.jp

[*1] Microsoft. Azure kinect dk AI モデルの開発 (2021/1/23) https://azure.microsoft.com/ja-jp/services/kinect-dk.

Vol.2022-CG-188 No.28
Vol.2022-DCC-32 No.28
Vol.2022-CVIM-231 No.28
2022/11/19

(a) 左上を注視 (人物 A)　**(b)** 左下を注視 (人物 A)　**(c)** 左下を注視 (人物 B)

図 1: 棚上のある物体を見ている様子

Network の入力とすることで注視対象であるパンフレットの 4 つの領域のうちどれを見ているかを推定している．このことから，姿勢情報からでもある程度注視領が推定可能であることがわかる．しかし，単純なクラス分類問題として定式化しているため，物体の配置を陽に扱えていない．

2.3 距離学習

距離学習は，特徴空間上において意味的に同じデータ同士を近くの位置に，意味的に異なるデータ同士は遠くの位置へ写像するような埋め込み表現を学習する手法である．距離学習における代表的なアプローチとしてはアンカーデータ，それと同じクラスのポジティブデータ，異なるクラスのネガティブデータという 3 組のデータを利用して埋め込み表現を学習するものである．ネットワークモデルはアンカーデータとポジティブデータ間の距離がアンカーデータとネガティブデータ間の距離より小さくなるように学習を行う．また，距離学習における損失関数や 3 組のデータのサンプリング方法は学習の効率や推定精度の良さに大きく影響を及ぼすため，様々な手法が提案されている．[5], [6]

本研究では，この枠組みを利用することで姿勢の個人差を考慮するような特徴空間上での埋め込み表現を獲得する．

3. 提案手法

本研究では，姿勢情報から後ろ向き人物の注視領域を姿勢の個人差を考慮して推定する手法を提案する．

図 1(a) を見ると，人物は左上を見ていると直感的に推測できる．また，図 1(a) と異なる領域を注視している図 1(b) の姿勢に着目すると，頭部向き，腰や足の曲げ方などにおいて図 1(a) 異なる特徴を持つことが確認できる．このことから，同じ姿勢であれば同じ領域を見ており，逆に異なる姿勢ならば異なる領域を見ているといえる．以上より，人はこのような姿勢の違いに着目して注視領域を推測していると考えられる．本研究ではその姿勢のとり方の違いに着目し、姿勢情報を利用した後ろ向き人物の注視領域推定を行う．具体的には，3 次元姿勢情報から棚上の注視領域を表す尤度マップを生成し，尤度マップから注視領域を推定する．

しかし，異なる人物が左下の同一領域を見ている図 1(a) と図 1(c) とを比較すると，同じ領域を注視しているにも拘わらず，異なる姿勢を取っていることが確認できる．このことから，同じ領域を注視している場合でも人によって姿勢の取り方には個人差が存在することがわかる．そのため単純に，姿勢情報を用いるだけでは姿勢の個人差を考慮することができない．そこで，本研究では，注視領域をクラスとして扱い，それに従って姿勢が特徴空間上で分離されるような特徴に変換する．具体的には，3 次元姿勢情報を距離学習を用いて学習した Encoder に入力し，注視領域クラス毎に分離されるような特徴空間に写像する．これにより，姿勢の個人差の影響を抑えた埋め込み表現を獲得する．この埋め込み表現を用いることで個人差を考慮した注視領域推定を行う．単純に姿勢情報を用いた場合と比べ，姿勢の個人差による推定の曖昧性を抑えることが可能になる．

提案手法の概要を図 2 に示す．姿勢情報から尤度マップを推定するためのモデルは姿勢を埋め込み表現へ変換する Encoder 部と埋め込み表現から尤度マップを推定する逆畳み込みネットワーク部で構成される．

3.1 姿勢情報から埋め込み表現に変換する Encoder

姿勢情報を姿勢の個人差を考慮した埋め込み表現へ変換する Encoder の概要及びその学習について述べる．Encoder は，Azure Kinect から取得した 21 個のカメラ座標系の 3 次元関節座標（以下，姿勢情報）を入力，姿勢情報と対応付いた注視領域を示すラベルを教師信号として，距離学習の枠組みを用いて特徴空間上の埋め込み表現を学習する．

まず，教師信号について述べる．棚上に配置された商品の種類毎に領域を分割し，全 12 領域を注視領域とする．そ

Vol.2022-CG-188 No.28
Vol.2022-DCC-32 No.28
Vol.2022-CVIM-231 No.28
2022/11/19

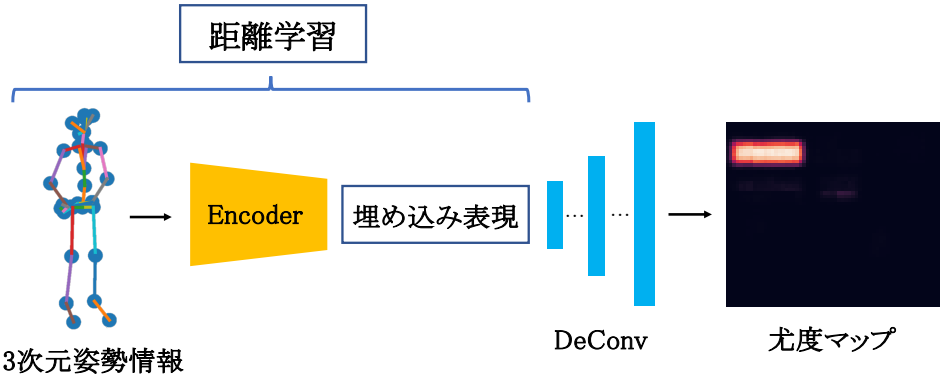

図 2: モデルの構成

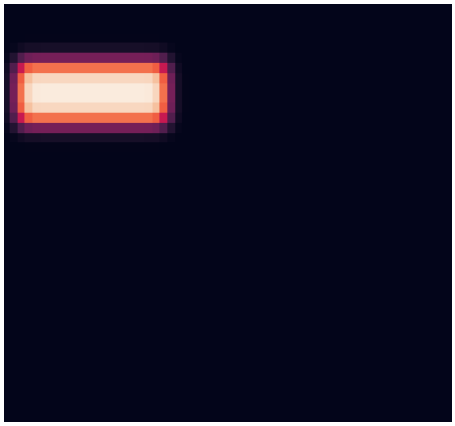

図 3: 教師信号用物体領域マップの例

して，学習データに対応した注視領域を教師信号として，学習に利用する．

Encoder に対する距離学習では，同じラベルを持つ姿勢情報同士を特徴空間上で近づけ，異なるラベルを持つ姿勢情報同士を特徴空間上で遠ざけるように学習を行い，姿勢情報を姿勢の個人差に影響されない埋め込み表現に変換する．なお，入力は 21 個の 3 次元関節座標を 63 次元ベクトルとして，Encoder を通じて 4 次元ベクトルに変換する．

距離学習の際，学習データから Triplet を取り出すときのサンプリングには Easy Positive Triplet Mining [7] を用いる．学習の最適化には AdamW[8] を用い，損失には複数ラベルを考慮することができる NTXentLoss[9] を用いる．

3.2 尤度マップ生成器

特徴空間上で埋め込まれた特徴を入力とした注視尤度を示す尤度マップ生成器の概要及びその学習について述べる．

生成器は前節で述べた Encoder によって写像された特徴を入力とし，棚上の物体領域マップを教師信号として注視尤度を表す尤度マップを生成するよう学習する．まず教師データについて述べる．棚上の物体領域の部分を 1 とし，サイズが 40×60 の物体領域マップを作成する．提案手法で利用する逆畳み込みネットワークの都合上，ネットワークの出力サイズを縦横が等しい正方形にする必要性があるため，物体領域マップを 64×64 に拡張し，拡張部分を 0 で埋める．その後，物体領域マップに対してガウシアンフィルタ ($\sigma = 3$) を適用して輪郭部分をぼかす処理をする．作成した教師データ用ヒートマップの例を図 3 に示す．

生成器における学習では AdamW [8] を 用い，入力に対応付けされた物体領域マップと出力尤度マップの誤差が小さくなるように，損失には平均二乗誤差 (MSE) を用いてネットワークのパラメータを学習する．

3.3 推定処理の手順

学習済みのモデルを用いた推定処理の手順を示す．

(1) 姿勢情報の変換
学習済みの Encoder に対し，21 点の 3 次元関節座標を 63 次元のベクトルとして入力し，4 次元の埋め込み表現を得る．

(2) 尤度マップの生成
Encoder から得た 4 次元の埋め込み表現を学習済み逆畳み込みニューラルネットワークに入力し，64×64 の尤度マップを得る．

(3) 注視領域の推定
逆畳み込みニューラルネットワークから得た尤度マップの棚上の各物体領域にの平均尤度を算出し， その平均尤度を比較して最も高かった物体領域を注視領域として，出力する．

4. データセット

本研究では，後ろ向き人物の 3 次元骨格座標から，注視物体領域の推定を目的としている．しかしながら，このようなタスクを対象とした公開データセットが存在しない

Vol.2022-CG-188 No.28
Vol.2022-DCC-32 No.28
Vol.2022-CVIM-231 No.28
2022/11/19

図 4: 撮影した棚の配置

ことから，独自にデータセットの構築を行った．本節ではデータセット作成における撮影条件及び内容について述べる．本研究では，コンビニエンスストアにおいて棚上に配置されている商品のいずれかを注視している人物を，定点カメラによって撮影している状況を想定する．被験者が自由な姿勢をとり，指定された位置から棚上の商品を注視している様子を撮影した．高さ 120 cm × 横幅 180 cm の棚を高さ 30 cm × 横幅 60 cm の 12 領域に分割し，各領域に 1 種類ずつ商品を配置する．実際の棚の様子を 図 4 に示す．また，被験者が商品を注視する位置は棚からの距離 0.5 m かつ棚との位置関係が正面の場合とする．実験参加者は 20 代の 7 名（女性 1 名，男性 6 名）であった．Azure Kinect は，解像度は 1280 × 720 画素，フレームレートは 15 fps となるよう設定して撮影を行なった．注視対象としたのはペットボトル，缶，本，及び紙パックである．各 3 種類ずつ用意し，上記で述べた 12 領域に 1 種類ずつ配置した．フレーム数は 15,228 である．

5. 実験

本研究で提案した，3 次元関節座標を姿勢の取り方の個人差を考慮した特徴空間に写像し，写像された特徴量から注視尤度を表す尤度マップを生成する手法と，姿勢から直接注視尤度を示す尤度マップを生成するベースライン手法を比較した．

5.1 実験方法

本実験では，姿勢から注視尤度を示す尤度マップを生成する提案手法の性能評価を行なった．

実験では，全 7 人分のデータから 6 人分を学習データ，1 人分をテストデータとした交差検証を行なった．評価指標には，以下の 2 つを用いる．1 つ目は，推定した尤度マップにおいて各物体領域の尤度の平均値を算出し，そして平均値が最も高い物体領域が真値と合っているかを示す正解率である．2 つ目は，推定した尤度マップにおけて最も尤度の平均値が高い領域と真値となる領域との距離の平均値である．

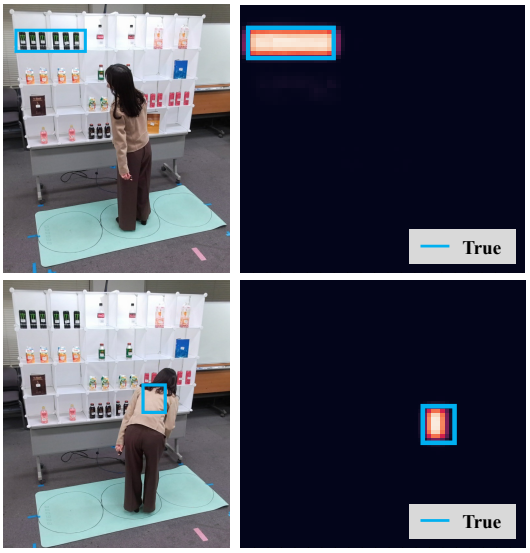

図 5: 撮影した画像と推定した尤度マップ

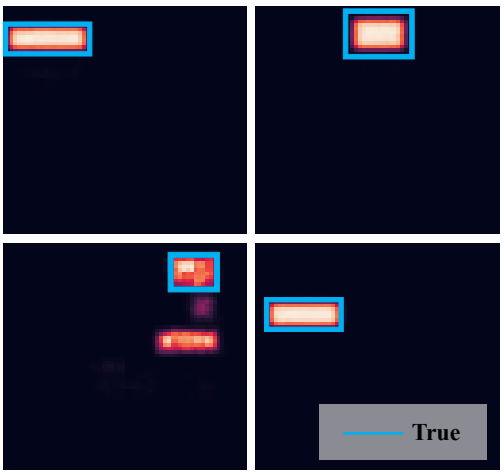

図 6: 提案手法によって推定した尤度マップ

5.2 実験結果と考察

図 5 に撮影した画像とその姿勢から提案手法によって生成した尤度マップを示す．表 1 に提案手法のすべてのテストデータにおける評価結果の平均値を示す．表 1 より，姿勢情報を特徴空間を経由して尤度マップを推定する提案手法の第 1 位正解率は 34.26% であり，第 3 位正解率までを加味すると 66.13% の正解率が得られることを確認した．

図 6 に示す提案手法が生成した尤度マップを見ると，生成された尤度マップは正解付近にピークを持つことがわかる．

このことから，提案手法は姿勢の個人差を考慮した特徴量から尤度マップを推定することによって，個人差による姿勢の曖昧性を抑えた推定を実現していると考えられる．提案手法は姿勢情報を特徴空間上に写像することによって，姿勢の個人差による注視領域の曖昧性を抑制した推定を実現しており，注視領域推定において一定の効果が得られることを確認した．

Vol.2022-CG-188 No.28
Vol.2022-DCC-32 No.28
Vol.2022-CVIM-231 No.28
2022/11/19

表 1: 正解率と推定誤差での評価結果

	第 1 位 正解率	第 2 位 正解率	第 3 位 正解率	推定誤差
提案手法	34.26％	55.02％	66.13％	0.33 m

6. まとめ

本研究では，画像中に後ろ向きで写る人物が注目している商品領域を推定する手法を提案した．3 次元関節座標を姿勢の取り方の個人差を考慮した特徴空間上の埋め込み表現に変換し，その特徴量から注視尤度を表す尤度マップを生成することで，姿勢のとり方の個人差による注視領域の曖昧性を抑えた推定を行う．提案手法の有効性を確認するために，3 次元関節座標と注視対象を対応付けたデータセットを構築し，それを用いた注視領域推定実験を行なった．実験の結果，姿勢の個人差による注視領域の曖昧性を抑制した注視領域推定が可能であることを確認した．

謝辞 本研究の一部は科研費 (17H00745) による．

参考文献

[1] Fridman、L., Langhans, P, Lee, J. and Reimer, B., Driver Gaze Region Estimation without Use of Eye Movement. IEEE Intelligent Systems, vol. 31, no. 3, pp. 49-56, (2016)
[2] Nonaka, S., Nobuhara, S., and Nishino, K., Dynamic 3D Gaze from Afar:Deep Gaze Estimation from Temporal Eye-Head-Body Coordination In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition , pp. 2192-2201 (2022).
[3] Yasutomo Kawanishi, Hiroshi Murase, Jianfeng Xu, Kazuyuki Tasaka, and Hiromasa Yanagihara. Which content in a booklet is he/she reading? Reading content estimation using an indoor surveillance camera. In Proceedings of the 24th International Conference on Pattern Recognition, pp. 1731-1736.
[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. Vol. 43, No. 1, pp.172-186.
[5] Chopra, Sumit and Hadsell, Raia and LeCun, Yann. Learning a similarity metric discriminatively, with application to face verification 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition
[6] Wang, Jian and Zhou, Feng and Wen, Shilei and Liu, Xiao and Lin, Yuanqing. Deep metric learning with angular loss Proceedings of the IEEE international conference on computer vision 2017
[7] Xuan, Hong, Abby Stylianou, and Robert Pless. Improved embeddings with easy positive triplet mining. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020.
[8] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization.
[9] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In International conference on machine learning. PMLR, 2020. p. 1597-1607.