Vol.2021-NL-248 No.22
Vol.2021-CVIM-226 No.22
2021/5/21

人物姿勢と注視対象配置制約に基づく後ろ向き人物の注視領域推定

弓矢 隼大[1,a)]　出口 大輔[1]　川西 康友[2,1]　村瀬 洋[1]　細野 峻司[3]

概要：本研究では，画像中に後ろ向きで写る人物が何に注目しているか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の多くは，カメラで撮影した人物の顔領域を抽出し，そこから得られる視線や顔向きを手がかりにしている．しかしながら，後ろ向き人物からは顔向き等の情報が得られないため，このような手法は適用困難である．一方，我々人間は人物が後ろ向きであったとしても，見るべき対象の位置が限定されていれば，その姿勢からその人物がどれを見ていそうかを推測することができる．そこで，後ろ向き人物であっても取得可能な 3 次元骨格座標を手がかりとして注視領域推定を行なう手法を提案する．3 次元骨格座標から注視尤度を表すヒートマップを生成し，対象の配置を元に注視領域を推定する．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，人物の 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た評価を行い，提案手法の有効性を確認した．

1. はじめに

人物の行動や意図の理解において，その人物が何を注視しているかは大きな意味を持つ．このような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．平山らは，対象となる人物の顔画像から顔の向き及び視線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 [1]．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．一方，人物の顔画像を正面から観測できないような場合において，その人物の注視領域と人物の姿勢の関係が調査されている．Kawanishi らは，人物の姿勢と注視領域に何らかの関係性があることを報告している [2], [3]．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．加えて，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．一方，見ているシーンがわかっている場合，そこに存在する物体の配置や，各物体の大きさなどは，その人物がどれを見ているかを推定するのに重要な要素である．そこで本発表では，

- 物体の配置
- 物体の大きさなどによる物体領域の大きさの違い

を考慮した注視領域推定手法を提案する．具体的には，以下の 2 つの工夫により，後ろ向き人物であっても注視領域を精度良く推定可能な手法を実現する．1 つ目の工夫として，物体の配置を考慮した注視尤度を表すヒートマップを注視領域推定の中間表現として導入する．このヒートマップは，姿勢情報を入力とした逆畳み込みニューラルネットワークを用いて作成し，これによって物体の配置を加味した中間表現を得る．2 つ目の工夫として，ヒートマップから各物体領域に対応する平均尤度を求め，それに基づいて注視領域推定を行なう．これにより，物体の大きさの違いによって尤度の平均化の度合いが変わるため，精度の良い視領域推定が可能になる．

2. 関連研究

2.1 一般的な注視方向推定

Kellnhofer ら [4] は，様々な方向や状況で人物を撮影した一連の画像を用いた学習により，人物の注視方向を推定する手法を提案している．この手法では，屋内外の環境を全方位カメラで撮影した映像に対して，多数の人物の注視方

[1] 名古屋大学大学院 情報学研究科
[2] 理化学研究所ガーディアンロボットプロジェクト
[3] 日本電信電話株式会社 NTT メディアインテリジェンス研究所
[a)] yumiyah@murase.m.is.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.2021-NL-248 No.22
Vol.2021-CVIM-226 No.22
2021/5/21

向を 3 次元的にアノテーションした映像データセットを構築し，このデータセットを用いた学習によって高精度な推定を可能にしている．また，連続する複数フレームを入力とする LSTM によって推定精度の向上を図っている．しかし，対象人物を後ろから撮影した場合の注視領域推定は実現できていない．

2.2 後ろ向き人物の注視方向推定に関する研究

後ろ向き人物の注視方向推定手法として，Bermejo ら [5] は後ろ向き人物の頭部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，第三者視点カメラで撮影された単一フレーム画像から YOLO [6] によって抽出した後ろ向き人物の頭部領域を用いて注視方向を推定する．また，多様な人物の 3D モデルを作成し，仮想的に様々な環境下 (光源位置，角度，カメラ距離など) で後ろ向き画像を撮影し，それらを学習させることでカメラの配置や角度，照明条件，解像度などの影響による推定誤差の低減を図っている．推定誤差は横方向に 23 度，縦方向に 26 度程度であり，後ろ向き人物に対する注視方向推定としては比較的精度良く推定が可能である．しかし，人物と物体との距離によって注視方向と注視領域の対応関係は変わるため，注視方向のみでは注視対象が不明瞭である．そのため，注視領域を推定するためには注視対象と注視方向との関連付けが必要である．

2.3 人物の骨格情報を用いた注視領域推定に関する研究

Kawanishi ら [2], [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 [7] により取得した骨格情報を Deep Neural 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分類している．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．

3. 提案手法

3.1 提案手法の概要

提案手法は，Azure Kinect を用いて取得した後ろ向き人物の 3 次元関節座標を入力として人物の注視領域推定を行なう．Azure Kinect によって取得可能な 32 個のカメラ座標系の 3 次元関節座標[*2]を逆畳み込みニューラルネットワークに入力することで物体の配置を考慮した中間表現となる注視尤度ヒートマップを生成する．また，物体の大きさを考慮した注視領域推定を行なうため，注視対象物体の領域を制約として用いる．具体的には，注視尤度ヒートマップから算出される各物体の配置領域に制限した平均尤度を求めることで注視領域推定を行なう．

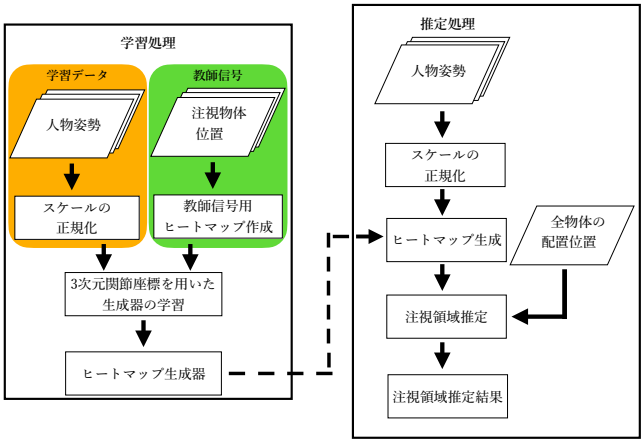

図 1 提案手法の処理手順

提案手法の処理手順を図 1 に示す．提案手法は学習段階と推定段階の 2 つに分けられる．学習段階では，3 次元関節座標と注視物体位置を入力として用い，3 次元関節座標は腰と首の関節間の距離が 1 になるように正規化処理を行なう．また，撮影時の注視物体位置から作成したヒートマップを教師データとする．以上の 3 次元関節座標とヒートマップの組を学習データとしてヒートマップ生成器の学習を行なう．推定段階では，ヒートマップ生成器の出力に対し，注視物体の配置制約を利用して注視領域の推定を行なう．

3.2 対象人物のサイズを考慮した 3 次元関節座標の正規化

人物の体格は個人により異なるため，同じ姿勢であっても得られる関節点の 3 次元関節座標は異なる．そこで，ヒートマップ生成器の学習において体格の影響を軽減するため，骨格のスケールを正規化する．具体的には，以下の処理により腰から首までの長さが全ての人物で同じ長さになるよう骨格全体のスケールを調整する．関節 i の 3 次元座標 $\mathbf{p}_i = [x_i, y_i, z_i]$ に対して，腰（$\mathbf{p}_1$）から首（$\mathbf{p}_3$）までの距離（式 (1)）が 1 になるように式 (2) を用いて正規化する．

$$d = \|\mathbf{p}_3 - \mathbf{p}_1\| = \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2 + (z_3 - z_1)^2} \quad (1)$$

$$\widehat{\mathbf{p}}_i = \frac{\mathbf{p}_i}{\|\mathbf{p}_3 - \mathbf{p}_1\|} = \left[\frac{x_i}{d}, \frac{y_i}{d}, \frac{z_i}{d}\right] \quad (2)$$

このようにして求めた正規化後の 3 次元関節座標の位置 $\widehat{\mathbf{p}}_i = [\widehat{x}_i, \widehat{y}_i, \widehat{z}_i]$ を用いることでヒートマップ生成器の学習を行なう．

[*2] https://docs.microsoft.com/ja-jp/azure/kinect-dk/body-joints. (2021/1/23 参照)

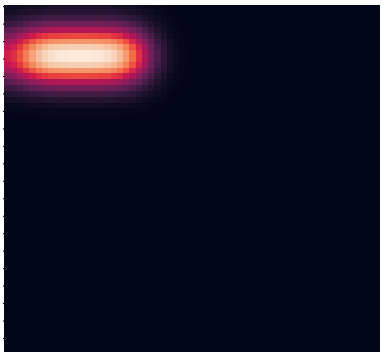

図 2 作成した教師信号用ヒートマップ

表 1 ネットワークの構成

input	Units	活性化関数
FullConnect	Units：2048	LeakyReLU
ConvTranspose1	Kernel：8 × 8 Stride：4 Channel：128	LeakyReLU
ConvTranspose2	Kernel：4 × 4 Stride：2 Channel：64	LeakyReLU
ConvTranspose3	Kernel：2 × 2 Stride：2 Channel：1	LeakyReLU
Output		Sigmoid

3.3 3 次元関節点座標を入力とした注視尤度を示すヒートマップ生成器

3 次元関節点座標を入力とした注視尤度を示すヒートマップ生成器について述べる．ヒートマップ生成器の入力は Azure Kinect によって取得した 32 個の 3 次元関節座標を並べた 96 次元ベクトルであり，物体領域ヒートマップを教師信号として注視尤度を表すヒートマップ生成器を学習する．

まず，教師信号である物体領域ヒートマップの作成について述べる．物体が存在する矩形領域の値を 1，それ以外の領域を 0 とした物体領域ヒートマップを作成する．その後，物体の矩形領域の輪郭部分は注視されにくいことを考慮し，ヒート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したものを教師データとする．作成した教師データの例を図 2 に示す．

提案手法で用いる逆畳み込みニューラルネットワークの構造を表 1 に示す．入力の関節点座標を並べた 96 次元ベクトルを全結合層に入力して 2,048 次元ベクトルに伸張し，4 × 4（128 チャネル）に変形して逆畳み込み層に入力する．全結合層および逆畳み込み層ではどちらも LeakyReLU を活性化関数に用いる．出力層では，60 × 60 の出力を Sigmoid 関数に入力し，出力値の範囲を $[0, 1]$ に制限する．生成器の学習には AdamW [8] を用い，出力ヒートマップと教師信号の物体領域ヒートマップの誤差が小さくなるようにネットワークのパラメータを学習する．なお，損失関数には平均二乗誤差（Mean Squared Error）を用いる．

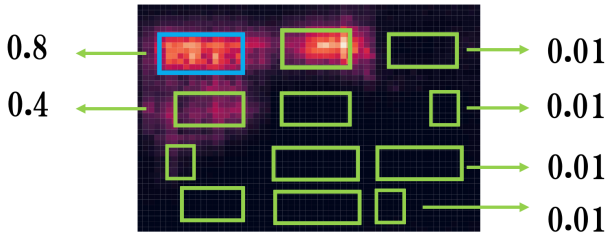

図 3 各物体領域の平均尤度

3.4 物体の配置制約に基づいた注視領域推定

姿勢情報を逆畳み込みニューラルネットワークに入力し，得られるヒートマップを用いて注視領域を推定する．まず，姿勢情報を入力として前節のネットワークにより注視尤度を表すヒートマップを生成する．次に，ヒートマップから各物体の配置領域の平均尤度を図 3 に示すように物体領域単位で算出する．最後に，各物体領域の平均尤度を比較して最も高い値をとる物体領域を推定結果とする．

4. データセット

本研究の目的は後ろ向き人物の 3 次元骨格座標から注視物体領域を推定することである．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は存在しない．そのため，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および内容について述べる．まず，4.1 節において撮影条件について述べ，次に 4.2 節において注視対象と人物の撮影手順について述べる．

4.1 撮影条件

本データセットは，コンビニエンスストアの棚に陳列された商品を人物が注視している様子を定点カメラで捉えるという状況を想定した．本データセットの画像撮影時には，指定した位置から被験者に棚上の商品を順番に自由な姿勢で注視させた．データセット撮影の様子を図 4 に示す．

高さ 120 cm× 横幅 180 cm の棚を高さ 30 cm× 横幅 60 cm の 12 箇所に分割した．注視対象の商品には，ペットボトル，缶，本，紙パックを用意した．各 3 種類ずつ用意し，各箇所に 1 種類ずつ商品を配置した．また，被験者が商品を注視する際の立ち位置を，棚からの距離 (0.5 m，1.0 m) と棚との位置関係 (左，中心，右) を組み合わせた計 6 箇所とした．図 6 に被験者の立ち位置及びカメラ位置を示す．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，解像度を 1,280 × 720 画素，フレームレートを 15 fps に設定した．

Vol.2021-NL-248 No.22
Vol.2021-CVIM-226 No.22
2021/5/21

図 4 撮影した注視の様子の例

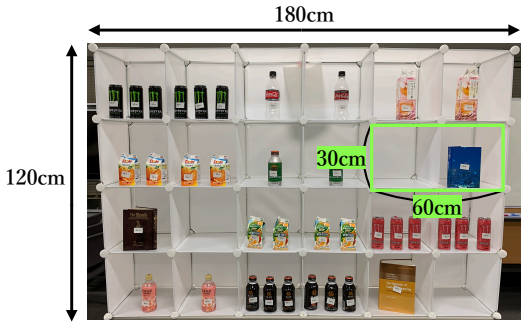

図 5 商品が配置された棚

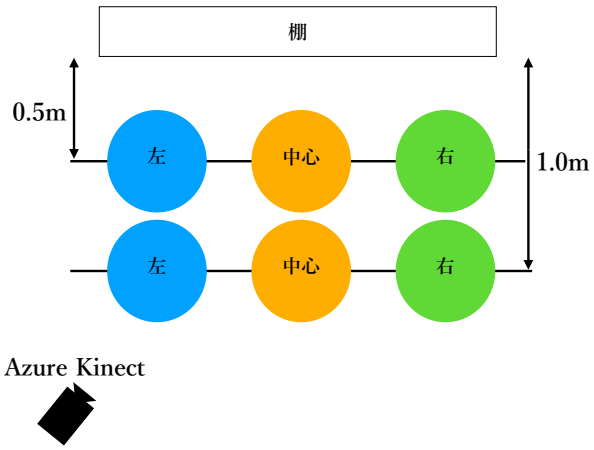

図 6 被験者の立ち位置の模式図

4.2 撮影手順

本節ではデータの撮影手順について述べる．前節で述べた 6 箇所の立ち位置から順に図 7 に示す順番に沿って注視を行なう．以降，図 7 に示すように 棚の 12 個の領域を「1」〜「12」と呼ぶ．被験者の立ち位置の順を図 8 に示す．上記の手順を 12 種類の物体に連続して行なう様子を撮影した．以上の撮影を 1 セットとし，各人物位置で 3 セットずつ撮影を行なった．7 名の被験者それぞれが上記タスクを行い，データセットを構築した．

機材の不備により 1 人分のデータの一部（右 −1.0 m）が破損していたため，6 人の完全なデータと 1 人の一部欠けた合計 7 人分のデータによりデータセットを作成した．

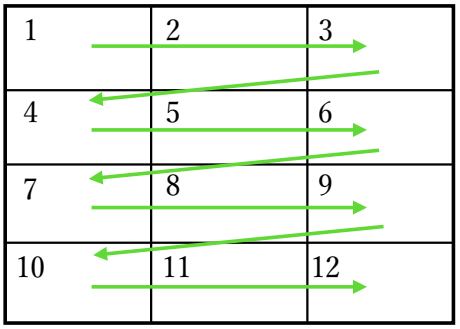

図 7 商品の注視順番図

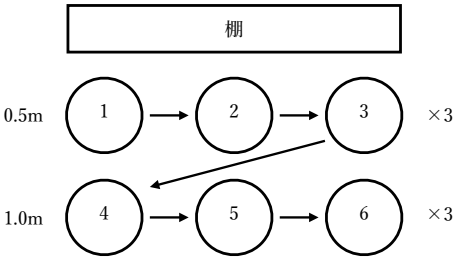

図 8 人物位置の順番図

5. 評価実験

姿勢情報から直接注視領域を分類する従来手法と提案手法の性能を比較した．

5.1 実験方法

本実験では，表 2 に示す 3 つの手法の比較を行なった．

従来手法は，人物の姿勢情報を入力とするニューラルネットワークを用いて注視領域の分類を行なう手法である．一方，提案手法 1 および 2 のいずれにも注視尤度ヒートマップを中間表現として用い，ヒートマップ生成器の構築には立ち位置毎のデータを用いた．使用した各位置における姿勢データのフレーム数を表 3 に示す．提案手法 2 は，ヒートマップ生成器から得られるヒートマップに対して各物体領域の尤度の平均値を算出し，平均値が最も高い物体領域を推定結果とする．一方，提案手法 1 はヒートマップ生成器から得られるヒートマップ上の最も高い値を有する領域を推定結果とする．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する交差検証

表 2 評価した手法

手法	ヒートマップの利用	分類手法
従来手法		姿勢情報から直接分類
提案手法 1	○	最も高い値を含む領域を採用
提案手法 2	○	注視対象の配置制約を利用

表 3 各位置毎の姿勢データのフレーム数

	左	中心	右
0.5m	14,602	13,437	14,153
1.0m	13,546	13,756	11,487

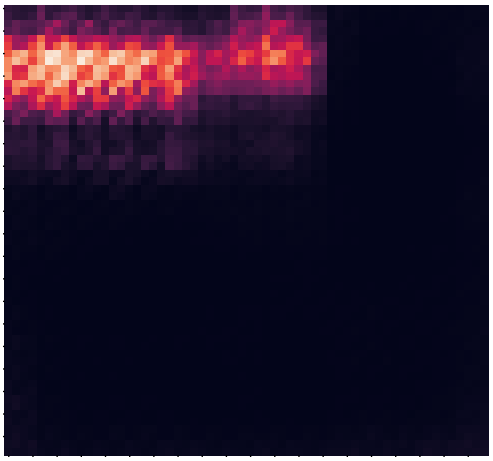

図 9 生成したヒートマップ例

表 4 距離 0.5m における平均正解率

手法	位置		
	左	中心	右
従来手法	21.9%	26.7%	17.7%
提案手法 1	33.7%	37.1%	25.8%
提案手法 2	**38.3%**	**41.3%**	**30.2%**

表 5 距離 1.0m における平均正解率

手法	位置		
	左	中心	右
従来手法	20.4%	19.7%	20.5%
提案手法 1	32.4%	30.2%	25.0%
提案手法 2	**36.9%**	**36.9%**	**29.2%**

を行ない，評価指標としては推定結果の正解率を採用した．

5.2 実験結果

図 9 に生成した注視尤度を示すヒートマップを示す．次に，表 4〜5 に各被験者の立ち位置ごとの注視推定結果の正解率を示す．正解率が最も高いものを赤字で示し，正解率は小数点第 2 位で四捨五入した．

従来手法との比較において，提案手法 1 および 2 の正解率は被験者の立ち位置が 0.5 m，1.0 m のいずれにおいても高くなることを確認した．また，提案手法 1 と提案手法 2 の比較においては，提案手法 2 の正解率が高いことを確認した．

また，表 6 および表 7 に各テストデータにおける提案手法 1 と提案手法 2 の正解率の差を示す．なお，Person7 の 1.0 m–右のデータは欠損のため表の一部に結果を記述していない．

これらの表より，どのテストデータにおいても平均して正解率は向上しており，全てのテストデータおよび距離で

表 6 距離 0.5 m での正解率 (%) の差分

	左	中心	右	平均値
Person1	6.4	5.4	6.4	6.1
Person2	5.8	5.5	5.7	5.7
Person3	3.5	3.8	2.1	3.1
Person4	4.8	2.0	3.1	3.3
Person5	5.6	2.1	4.3	4.0
Person6	1.5	1.7	5.1	2.8
Person7	4.8	9.4	3.9	6.0
平均	4.8	4.3	4.4	4.4

表 7 距離 1.0 m での正解率 (%) の差分

	左	中心	右	平均値
Person1	-1.3	7.5	7.4	4.5
Person2	5.2	6.0	2.9	4.7
Person3	3.0	4.3	3.9	3.7
Person4	5.1	3.4	3.0	3.8
Person5	7.8	9.9	4.8	7.5
Person6	8.4	9.8	3.1	7.1
Person7	3.1	6.2	—	4.7
平均	4.5	6.7	4.2	5.1

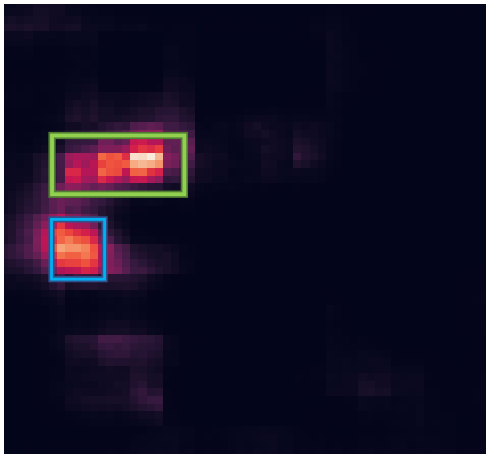

図 10 1.0m-左の Person1 のヒートマップ．緑色の枠が真の注視領域，青色の枠が提案手法 2 による推定された注視領域

の正解率向上は 4.8 ポイントとなっている．しかし，データ毎に正解率の向上度合いは異なっており，Person1 の位置 1.0 m–左においては低下が見られ，テストデータによって提案手法 2 で得られる改善幅が小さいことが確認できる．この原因を確認するため，改善幅が最も小さい Person1 の位置 1.0 m–左のヒートマップを図 10 に示す．この図は，提案手法 2 は誤推定したものの，提案手法 1 では正しい推定が得られた例である．この図から，提案手法 2 による誤推定する際，正解領域にピークを持つもののその周囲の値が低くなることがわかる．このことから，提案手法で用いた物体の配置制約によってピークを持つ物体領域の平均値が小さくなるためであると考えられる．

6. むすび

本発表では，棚に陳列された商品を顧客が注視している状況を想定し，後ろ向き人物の姿勢情報から注視領域を推定する手法を提案した．提案手法では，後ろ向き人物の 3 次元関節点座標から注視尤度ヒートマップを中間表現として生成し，そのヒートマップから注視領域を推定した．ヒートマップから注視領域を推定する際，物体の配置領域を制約として注視尤度の平均値を求め，その値が最大となる領域を注視領域とした．提案手法の有効性を確認するために，棚上の商品を注視している様子を Azure Kinect を用いて撮影し，骨格情報と注視物体を紐付けたデータセットを構築した．また，そのデータセットを用いた注視領域推定の実験を行なった．実験結果より，推定の中間表現としてヒートマップを導入した提案手法は従来手法である姿勢情報を入力としたニューラルネットワークを用いた分類する手法と比べ正解率が向上することを確認した．また，物体領域の配置制約を用いる提案手法 2 は，ヒートマップの最大値を注視領域として推定する提案手法 1 と比べ，平均 4.78 ポイント正解率が向上することを確認した．

今後の課題としては，安定したヒートマップ生成器の構築手法の検討，同じ姿勢で注視点のみが異なる人物への対応，多様な姿勢を含むデータセットへの拡張，人物位置の変化への対応，などが挙げられる．

謝辞 本研究の一部は科研費 (17H00745) による．

参考文献

[1] Takatsugu Hirayama, Yasuyuki Sumi, Tatsuya Kawahara, and Takashi Matsuyama. Info-concierge: Proactive multi-modal interaction through mind probing. In Proceedings of the 3rd Asia Pacific Signal and Information Processing Association Annual Summit and Conference (2011).

[2] 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 Pro - ceedings of the 24th International Conference on Pattern Recognition, pp. 1731–1736 (2018).

[3] 川西康友, 村瀬洋, 徐建鋒, 田坂和之, 柳原広昌. 屋内定点カ メラを用いたパンフレット閲覧項目推定システム. 精密工学会誌 Vol. 85, No. 5, pp. 463–468 (2019).

[4] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically un- constrained gaze estimation in the wild. In Proceedings of the 17th IEEE/CVF International Conference on Computer Vision, pp. 6912–6921 (2019).

[5] Carlos Bermejo, Dimitris Chatzopoulos, and Pan Hui. EyeShopper: Estimating shoppers' gaze using CCTV cameras. In Proceedings of the 28th ACM International Conference on Multimedia, pp. 2765–2774 (2020).

[6] Joseph Redmon and Ali Farhadi, YOLOv3: An incremental improvement, arXiv preprint arXiv:1804.02767, (2018).

[7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, No. 1, pp. 172–186 (2019).

[8] Ilya Loshchilov and Frank Hutter, Decoupled Weight Decay Regularization, arXiv:1711.05101, (2019).