Vol.2023-CVIM-233 No.58
2023/3/3

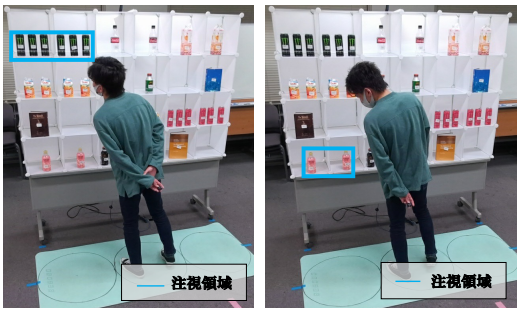
個人差吸収型距離学習を用いた後ろ向き人物姿勢からの注視領域推定

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

概要：本報告では，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，顔領域を用いる既存の手法を用いることは難しい．そこで，人物の姿勢に着目する．個人差はあるものの，どこを見ているかによって姿勢は変化し，後ろ向きの人物からでも姿勢情報は取得可能である．そこで，我々は後ろ向き人物の姿勢から注視領域を推定する手法の開発を行なってきた [2]．しかし，単純に姿勢を用いるだけでは，姿勢のとり方の個人差による特徴の変化に対応することが難しい．そこで，注視領域が同じ場合は特徴同士が近づき，異なる場合は実空間上での領域間の距離と同じだけ特徴同士が離れるような埋め込み表現に変換し，その特徴を用いて推定することで姿勢の個人差による影響を抑え，精度の向上を図る．提案手法の性能を評価するため，3 次元骨格座標と注視対象を紐付けたデータセットを使用し，提案手法の有効性を確認した．

1. はじめに

人物が何を注視し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取得した顔の特徴点から注視領域を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の特徴点を抽出できず，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(a) と (b) 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．つまり，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．我々はこのような姿勢の変化に着目した手法を提案している [2]．しかし，人物 A の注視の様子（ 図 1(b) ）と人物 B の注視の様子（ 図 1(c) ）を比べると，注視時の姿勢のとり方は人によって個人差があり，同じ領域を注視していても異なる姿勢を取っていることがわかる．このように，人によって同じ領域を注視していても異なる姿勢をとることがあるため，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，個人差による姿勢のとり方の違いを吸収するために，物体領域間の実空間上での距離関係を加味した特徴に変換する姿勢埋め込みモジュールを導入する．これ



(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

[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

により，入力された 3 次元姿勢は，人によって姿勢のとり方が異なっていたとしても，その注視している物体領域に対応した特徴に変換され，結果として姿勢の個人差を吸収した埋め込み表現を獲得することができる．この姿勢埋め込みモジュールは，深層距離学習を用いて，物体領域間の実空間上の距離関係を学習する．姿勢埋め込みモジュールによって個人差を吸収した埋め込み表現に姿勢情報を変換する．そして，その埋め込み表現を入力とする尤度マップ推定モジュールで注視領域を表す尤度マップを推定し，物体領域を参照して最も尤度の高い領域を注視領域として出力する．

提案する注視領域推定の貢献は以下の通りである．

- 顔情報が取得できない後ろ向き人物に対する姿勢を用いた注視領域推定の実現
- 深層距離学習を用いて物体領域間の実空間上の距離関係を反映するような個人差吸収埋め込み表現の提案

2. 関連研究

2.1 一般的な注視方向推定

Kellnhofer ら [3] 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視線方向推定を可能にしている．そして，連続する複数フレームを入力とするニューラルネットワークの一種である Long Short-Term Memory（LSTM）を用いることによって推定精度の向上を図っている．応用例として，注視領域推定へ適用する試みも行なわれているが，顔を正面から捉えられるようにカメラを配置した状況下であり，人物の目の特徴に大きく依存しているため，目の情報を取得できない後ろ向き人物への適用は難しい．あくまで，注視方向推定の手法であるため，後ろ向き人物に対して焦点を当てた注視領域推定手法が必要である．

Nonaka ら [4] 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，人物画像から，頭と体の向きの条件付き分布と視線方向の条件付き分布をそれぞれ LSTM でモデル化する．人物画像を入力して頭と体の向きを推定し，その頭と体の向きから視線方向を推定している．また，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ており，それを学習及び検証に使用している．人の視線，頭，体の向きの協調性を利用することで，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，時系列情報を推定に用いるため，単一フレームの情報から推定することはできない．

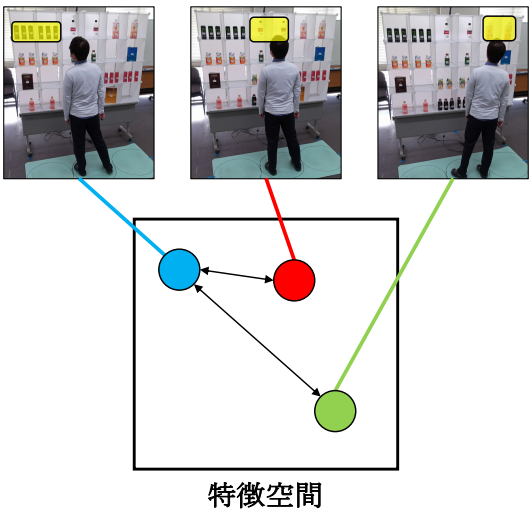


図 2: 実空間上の物体領域間の距離関係を反映した特徴の例

2.2 人物の骨格情報を用いた注視領域推定

人物の骨格情報を用いて注視領域を推定する手法として，Kawanishi ら [5] 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 [6] により取得した骨格情報を単純な 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推定している．人物とカメラの距離によって分類精度は変化するが，60 %から 80 %の分類精度を実現している．このことから，姿勢情報からでもある程度注視領域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，物体の配置関係を加味した推定はできていない．

3. 提案手法

人物が後ろ向きである場合は顔領域が取得出来ないため，目や顔向き等の注視に非常に密接な関係を持つ情報が得られず，注視領域を推定することは難しい．これに対して，本報告では姿勢情報を用いることで後ろ向き人物の注視領域推定を実現することを目的とする．

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，姿勢情報には同じ領域を注視している場合でも姿勢のとり方に個人差があり，特徴が大きく異なることがある．そのため，単純に姿勢を推定に用いるだけではその個人差による特徴の変化に対応することは難しい．我々 [2] は深層距離学習を用いて姿勢を識別的な特徴に変換して尤度マップを推定する手法を提案している．しかし，単純に深層距離学習を用いて異なる領域を注視している姿勢を変換した場合，異なる領域を見ている姿勢同士を離す処理の際に実空間上の距離関係を加味することができない．その

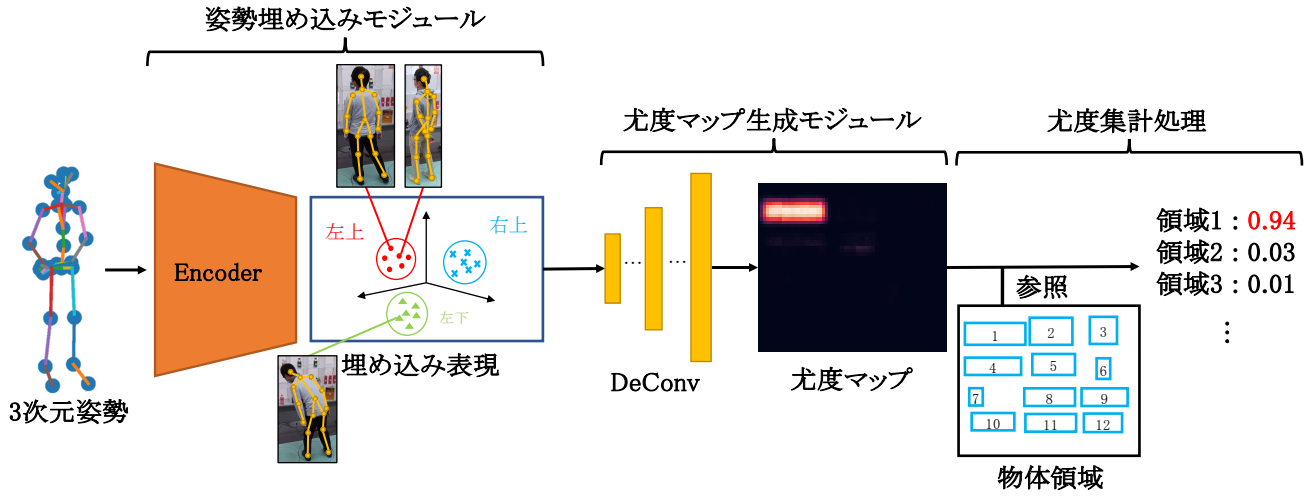


図 3: モデルの構成

ため，異なる領域が実空間上で近い場合と遠い場合を区別せずに特徴同士を離す処理を行なってしまう．そこで，この個人差による姿勢のとり方の違いを吸収するために，物体領域間の実空間上の距離関係を反映した特徴に変換する姿勢埋め込みモジュールを導入する．姿勢埋め込みモジュールは，深層距離学習を用いて学習されるエンコーダであり，3 次元姿勢を入力として，図 2 に示すような物体領域の実空間上の距離関係を加味した埋め込み表現に変換する．これにより，3 次元姿勢は人によって姿勢のとり方が異なっていたとしても，その注視している物体領域に対応した特徴に変換され，姿勢の個人差を吸収した埋め込み表現を獲得することができる．

その特徴を入力とする尤度マップ推定モジュールで注視領域を表す尤度マップを推定し，物体領域を参照して最も尤度の高い領域を注視領域として出力する．このようにして，後ろ向き人物の姿勢を用いた注視領域推定を実現する．

提案手法の概要を図 3 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埋め込みモジュールと埋め込み表現から尤度マップを推定する尤度マップ生成モジュール，予め得た物体位置を参照して，各物体ごとの注視尤度を計算する尤度集計処理から構成される．

提案モデル全体は，姿勢埋め込みモジュールの損失 L_e と尤度マップ生成モジュールの損失 L_d の和 L を最小化する End-to-end で学習する．その損失関数 L は式 (1) で表される．

$$L = L_e + L_d \tag{1}$$

3.1 姿勢埋め込みモジュール

提案手法では，深層距離学習の枠組みに基づき，姿勢と注視物体領域のラベルを用いて，姿勢の個人差を吸収できる特徴空間に埋め込むエンコーダを学習する．

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．学習の際，実空間上での物体領域の距離関係に近づくような損失を与えることで，特徴空間上での特徴間の距離を実空間における距離を反映したものにする．これにより，同じ領域を見ている姿勢は特徴空間上で近づくため，姿勢の個人差を抑えた特徴に変換することができる．なお，シーン画像中には，人が注視するようなオブジェクトが複数存在するため，その注視対象になりえる物体領域を整数値で表される ID として与え深層距離学習に用いる．エンコーダへの入力は 21 個の 3 次元関節座標を連結した 63 次元であり，エンコーダを通じて 4 次元ベクトルに変換する．学習のための損失関数として 2 つの特徴間の距離と実空間上の距離との間の絶対値誤差をモジュールの損失 L_e として用いる．損失 L_e は式 (2) で定義される．

$$L_e = \left| \|\mathbf{f} - \hat{\mathbf{f}}\| - \|\mathbf{r} - \hat{\mathbf{r}}\| \right| \tag{2}$$

ここで，$\mathbf{f}$ と $\hat{\mathbf{f}}$ は 2 つの姿勢をエンコーダに入力し得られた特徴を表し，$\mathbf{r}$ と $\hat{\mathbf{r}}$ は 2 つの姿勢が注視している領域の中心点を表す．この損失を小さくすることで，特徴空間における特徴間での距離関係が実空間における距離関係を反映するように学習する．

距離学習の際，学習データからサンプルを取り出すときのサンプリング手法には Triplet Margin Miner を用いる．

3.2 尤度マップ生成モジュール

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．

尤度マップ生成器の入力は，前節で述べた姿勢埋め込みモジュールによって変換された特徴 $\mathbf{f}_i$ であり，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

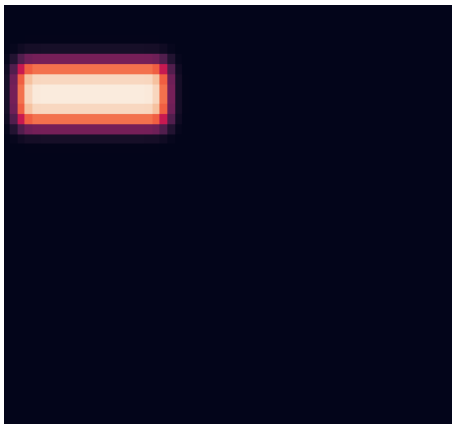


図 4: 教師信号用物体領域マップの例

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．元の注視物体が配置されていた棚領域の画像サイズは 400×600 であり，本研究ではその 10 分の 1 のスケールの物体領域マップを作成する．棚上の物体領域の値を 1 それ以外の位置の値が 0 である，40×60 画素の物体領域マップを作成する．尤度マップ生成モジュールで用い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．そのため，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．作成した教師データ用尤度マップの例を図 4 に示す．ネットワークには，画像生成の分野で用いられる逆畳み込みニューラルネットワークを用い，出力尤度マップと入力姿勢に対応付けされた物体領域尤度マップと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．

生成器における学習では，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，平均二乗誤差 (MSE) を損失に用いてネットワークのパラメータを学習する．姿勢から生成された尤度マップに対する MSE 損失（L_d）を式 (3) に示す．

$$\begin{aligned} L_d &= \mathrm{MSE}(\mathbf{m}, \widehat{\mathbf{m}}) \\ &= \frac{1}{K^2} \sum_{j=1}^{K^2} (y_j - \widehat{y}_j)^2 \qquad (3) \\ y_j &\in \mathbf{m}, \ \ \widehat{y}_j \in \widehat{\mathbf{m}} \end{aligned}$$

$\mathbf{m}_i$ は生成器より得られた i 番目の尤度マップ，$\widehat{\mathbf{m}}_i$ は i 番目の真値の物体領域マップを表す．K は尤度マップのサイズを表し，本提案では $K = 64$ である．

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

学習済みのエンコーダに対し，21 点の 3 次元関節座標を連結した 63 次元のベクトルを入力し，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 画素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に対して尤度集計処理を行ない，棚上の各物体領域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

本報告では，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ことから，独自にデータセットの構築した．このデータセットの詳細については [2] にゆずる．

5. 実験

3 次元姿勢を姿勢埋め込みモジュールに入力し，姿勢の個人差を吸収した特徴量に変換した後に注視尤度を表す尤度マップを生成する提案手法と，3 次元姿勢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を分類問題として推定する比較手法 1 と我々の従来手法 [2]（比較手法 2）を評価した．

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を評価した．

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により評価した．また，物体領域数は 12 とした．評価指標には，以下の 2 つを用いた．1 つ目は，正解率で，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．2 つ目は，推定誤差で，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で示される．

5.2 実験結果と考察

表 1 に全てのテストデータにおける実験結果の平均値を示す．第 1 位正解率は 33.97 %であり，第 3 位正解率までを加味すると 65.30 %の正解率が得られており，比較手法 1 と比べ，第 1 位正解率は 17.33 ポイント，推定誤差は 0.15 m 向上していることがわかる．実空間上の距離関係を加味していない比較手法 2 と比べ，第 1 位正解率は 7.63 ポイント，推定誤差は 0.07 m 向上していることがわかる．図 5 と 図 6 に示す提案手法が生成した尤度マップを見る

表 1: 正解率と推定誤差での評価結果

手法	第 1 位 正解率	第 2 位 正解率	第 3 位 正解率	推定誤差
比較手法 1	16.64 %	34.59 %	44.51 %	0.47 m
比較手法 2	26.34 %	44.30 %	58.82 %	0.39 m
提案手法	**33.97** %	**54.65** %	**65.30** %	**0.32** m

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，提案手法は実空間上の距離関係を加味した埋め込み表現から尤度マップを推定することによって，姿勢の個人差による影響を抑え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．

6. まとめ

本提案では，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影響を抑えた注視領域推定を実現した．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実施した．実験の結果，姿勢の個人差による影響を抑えた注視領域推定が可能であることを確認した．

謝辞 本提案の一部は科研費 (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] 弓矢隼大, 出口大輔, 川西康友, 村瀬洋. 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. 研究報告コンピュータビジョンとイメージメディア（CVIM）, no. 28, pp. 1-5 2022.
[3] Kellnhofer, P. Recasens, A., Stent, S., Matusik, W. and Torralba, A. Gaze360: Physically unconstrained gaze estimation in the wild. In Proceedings of the IEEE International Conference on Computer Vision, pp. 6912-6921, 2019.
[4] 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 Conference on Computer Vision and Pattern Recognition , pp. 2192-2201 2022.
[5] Kawanishi, Y., Murase, H., Xu, J., Tasaka, K., and Yanagihara, H. Which content in a booklet is he/she reading? Reading content estimation using an indoor surveillance camera. In Proceedings of the International Conference on Pattern Recognition, pp. 1731-1736, 2018.
[6] Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. IEEE Transactions on Pattern Analysis Machine Intelligence, vol.43, no.01, pp.172–186, 2021.
[7] Chopra, Sumit and Hadsell, Raia and LeCun, Yann. Learning a similarity metric discriminatively, with application to face verification In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 539–546, 2005.
[8] Wang, Jian and Zhou, Feng and Wen, Shilei and Liu, Xiao and Lin, Yuanqing. Deep metric learning with angular loss In Proceedings of the IEEE international conference on computer vision, pp. 2593-2601, 2017.

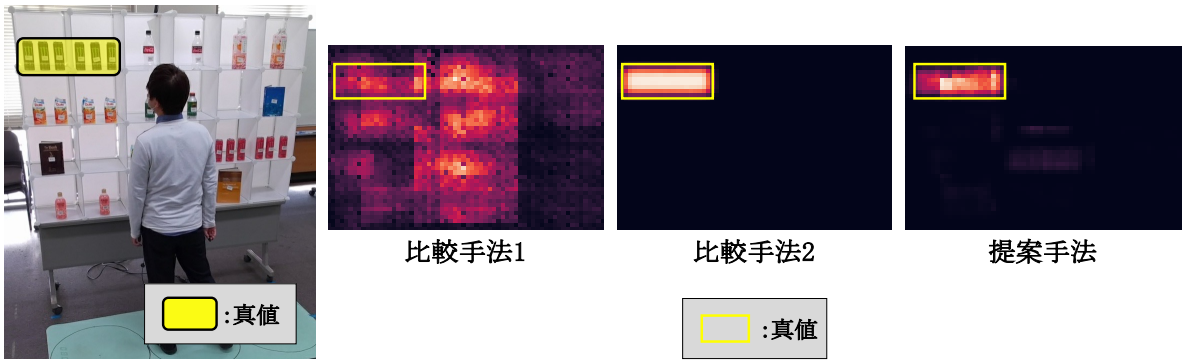


図 5: 左上を注視している人物の推定結果の例

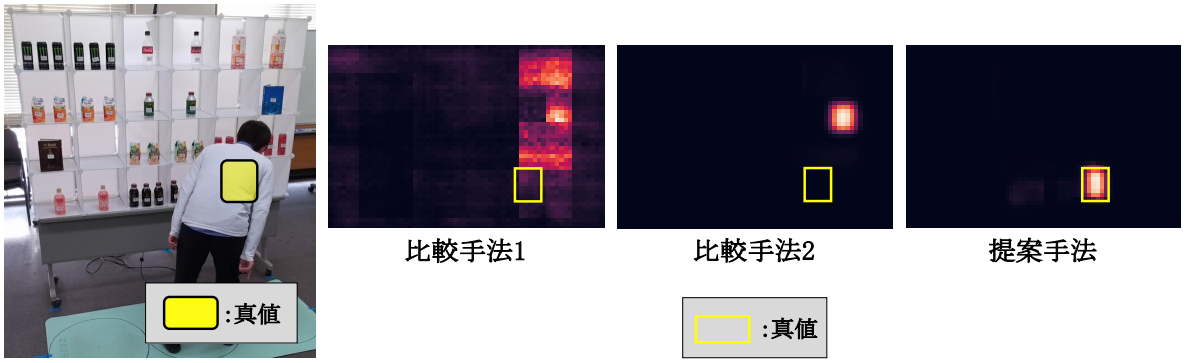


図 6: 右下を注視している人物の推定結果の例