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]　出口 大輔[†1]　川西 康友[†1]　村瀬 洋[†1]　細野 峻司[†2]

概要：本研究では後ろ向き人物の注視領域を推定する手法を提案する．人物の注視領域を推定する手法として，カメラで人物を撮影し，顔領域を取得して分析する手法が存在するが，顔領域の取得できない後ろ向き人物の場合には困難である．そこで，後ろ向き人物から取得可能な情報である 3 次元骨格座標を用いて注視領域を推定する．また、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 次元骨格座標を撮影することで姿勢情報と注視対象を紐付けたデータセットを作成し、提案手法の評価を行なった．

1. はじめに

人間の行動や意図を理解する上で注視は大きな意味を持っている．注視されている対象はどのような理由であれ，注意を惹きつけているからである．人物から取得可能な情報を用いて、何を注視しているかを推定することは、商品への興味度合いを調べることなどのマーケティングへの活用が期待できる重要なタスクである．

そこで，人物の注視領域推定をする手法について考える．対象とする人物の顔画像を取得可能な場合においては精度良く推定する手法が考案されている．[1] しかし，対象の人物が後ろ向きである場合，顔画像が取得出来ないため注視領域を推定することは難しい．そこで，後ろ向き人物から取得可能かつ，注視領域の違いと関係のある情報について考える．Kawanishi らの研究 [2] [3] によると，人物の姿勢と，注視領域には相関がある．確かに，物体を注視する際の人物の姿勢に注目すると屈んだり，頭を向けたりといった物体の位置に関連して姿勢が変化することがわかる．加えて，姿勢情報は後ろ向き人物からでも取得可能である．そこで，姿勢情報から注視領域を推定するモデルを提案する．この，物体の注視時にどのような姿勢をしているかを姿勢情報と定義する．しかし，姿勢情報を入力とした単純な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った場合，以下のような問題を考慮できない．

- 物体同士の配置関係
- 物体の大きさなどによる物体領域の大きさの違い

そこで，本研究では以下の 2 つの工夫を取り入れたモデルを構築する．1 つ目は，推定の中間表現として物体の配置領域と対応した注視尤度を示すヒートマップの導入である．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する．これによって物体の配置関係を加味した中間表現を作成する．2 つ目はヒートマップから各物体の領域における平均尤度を用いた注視領域の推定である．これによって物体の大きさの違いを加味した注視領域の推定を行う．以上を踏まえたモデルを用いて後ろ向き人物の注視領域推定を行う．

2. 関連研究

2.1 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 [4] 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 [5] 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ことが可能である．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，学習させることで，入力画像の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を抑えることを実現している．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高精度な推定が可能である．しかし，注視方向のみでは注視対象が不明瞭なため，注目度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

[1] 情報処理学会
IPSJ, Chiyoda, Tokyo 101–0062, Japan
[†1] 現在，名古屋大学大学院 情報学研究科
Presently with Nagoya Uniersity graduate of Informatics
[†2] 現在，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
Presently with NTT Media Intelligence Laboratories

必要である．

また，Kellnhofer ら [6] 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て学習させることで，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，屋内外の環境において，多くの人物の注視方向を 360 度の範囲で 3 次元的にアノテーションされた複数の短時間のビデオデータセットを作成し，学習に用いることで高精度な推定を可能にしている．また，複数の連続したフレーム画像群を入力とする LSTM を採用することで推定精度を向上させている．一方，応用例として対象の人物の注視領域の推定を行っているが，対象人物を横から撮影した場合の例であり，対象人物を後ろから撮影した場合の推定を行っていない．そのため，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．

2.2 人物の骨格情報を用いた注視領域推定に関する研究

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 [2][3] は画像上の人物から取得した骨格情報をもとに注視領域推定する手法を提案している．この手法では，注視領域の変化と対応して人物の姿勢が変化することに着目しており，OpenPose[7] を用いること画像から取得した骨格情報を入力とした Deep Neural Network を用い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．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．

3. 提案手法

3.1 提案手法の概要

後ろ向き人物の注視領域を推定するにあたり，後ろ向き人物からは顔領域を取得することができないという問題がある．そこで，Azure Kinect[8] によって取得した後ろ向き人物の 3 次元関節座標情報を用いることで注視領域推定を行なう．本研究では，物体の位置関係を加味した推定のために中間表現として 32 点の 3 次元関節座標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，注視尤度を示すヒートマップを生成する．次に，物体の大きさによる違いを加味するため，注視対象物体の配置状況を制約として用いる．具体的には注視尤度を示す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用いて注視領域の推定を行う．

ここで，提案手法の処理手順を図 1 に示す．提案手法は学習処理と推定処理の 2 つに分けられる．学習処理では，3 次元関節座標と注視物体位置を用いる．3 次元関節座標に対して正規化処理を行なう．撮影時に座標と対応付けした注視物体位置からヒートマップを作成する．以上の処理を施した 3 次元関節座標を学習データ，作成したヒートマップを教師信号としてヒートマップ生成器の学習を行なう．

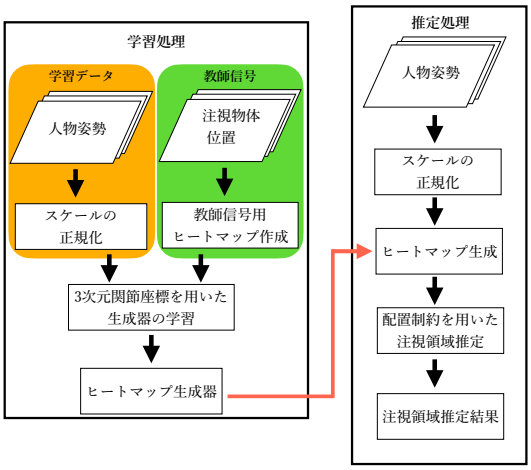

図 1 提案手法の処理手順

推定処理では，学習した生成器から出力されたヒートマップに対して注視物体の配置制約に基づいて，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

対象となる人物の身体の大きさには個人差があり，同じ姿勢であっても関節点の 3 次元関節座標が変わる．そこで，ヒートマップ生成器を姿勢変化のみに焦点を当てて学習させるため，体のサイズの違いを吸収するために骨格スケールを統一する．へそから腰にかけての距離は姿勢が変化しても保たれるため，スケールを統一する際の基準点としてへそと腰の座標点間の距離を採用する．各 i の 3 次元関節座標 $\mathbf{p}_i = [x_i, y_i, z_i]$ に対して，へそ ($\mathbf{p}_1$) から首 ($\mathbf{p}_3$) にかけての座標間距離を表す式 (1) が 1 になるように全身の関節点座標を式 (2) を用いて正規化をする．

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

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

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

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の概要，及びその学習について述べる．生成器は Azure Kinect によって取得した 32 点の 3 次元関節座標を並べた 96 次元ベクトルを入力，物体領域ヒートマップを教師信号として，注視尤度を示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．元の注視物体が配置されていた棚領域の画像サイズは 420 × 600 であり，10 分の 1 のスケールの物体

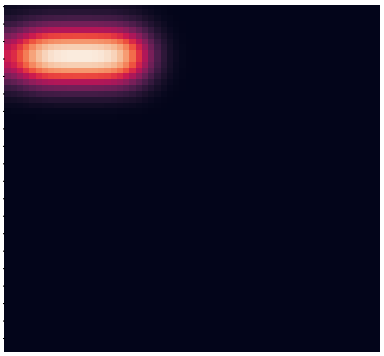

図 2 作成した教師信号用ヒートマップ

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

領域ヒートマップを作成する．具体的には，棚上の物体領域の部分を 1 とし，サイズが 42 × 60 のヒートマップを作成する．本研究で作成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．そこで作成したヒートマップを 60 × 60 に拡張し，拡張部分を 0 で埋める．その後，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 2 に示す．

作成した逆畳み込みニューラルネットワークの構成を表 1 に示す．入力である 96 次元ベクトルを FullConnect 層に入力し，2,048 次元ベクトルに伸張後，4 × 4(128 チャネル) に変形し Convolution Transpose 層に入力する．FullConnect 層及び Convolution Transpose 層どちらにも活性化関数として LeakyReLU を用いる．出力層では，60 × 60 の出力に活性化関数の Sigmoid 関数を適用して出力値を [0,1] の範囲に制限する．生成器における学習では AdamW[9] を用い，出力ヒートマップと入力ベクトルに対応付けされた物体領域ヒート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．なお，損失関数は平均二乗誤差 (Mean Squared Error) を用いる．

3.4 物体の配置制約に基づいた注視領域推定

姿勢情報を入力とした逆畳み込みニューラルネットワークにより生成したヒートマップから注視領域を推定する手法について述べる．姿勢情報から生成した注視尤度を示すヒートマップから各物体の配置領域に制限して平均尤度を算出する．3 各物体領域の平均尤度を比較して最も高かった物体領域を推定結果とする．

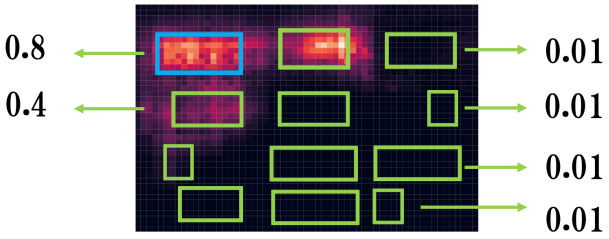

図 3 各物体領域の平均尤度

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．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順番に注視している様子を撮影した．撮影した様子を図 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 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 5 に示す．

また，被験者が商品を注視する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 箇所とする．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 8 名 (女性 2 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 15fps 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．

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．

図 4 撮影した注視の様子の例

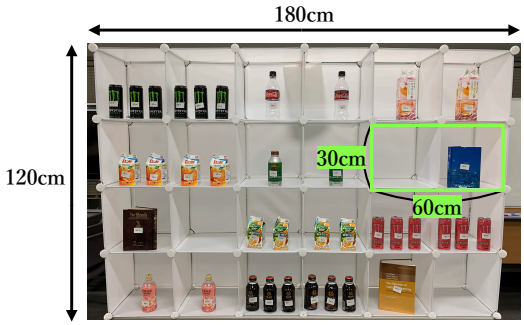

図 5 商品が配置された棚

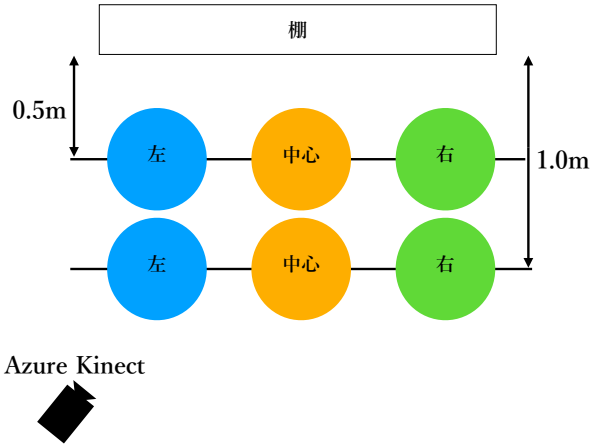

図 6 被験者の立ち位置の模式図

予め被験者には，自由な姿勢を取って注視を行なうよう教示した．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．各位置ごとで合計約 78000 フレームの 3 次元関節情報を取得した．8 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 人分のデータが取得できず，別の人物のデータの一部（右 −1.0m）が破損していたため，6 人

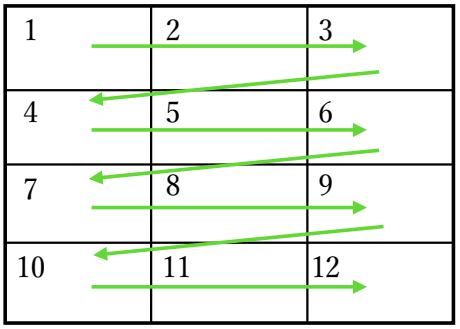

図 7 商品の注視順番図

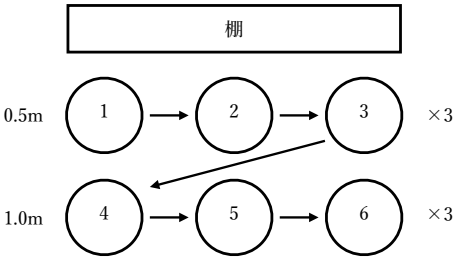

図 8 人物位置の順番図

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．

5. 評価実験

本研究において提案した，人物の姿勢情報から中間表現としてヒートマップを生成し，物体の配置領域の平均尤度から注視領域を推定する手法と従来手法である姿勢情報からの直接分類する手法を比較した．

5.1 実験方法

本実験では，生成したヒートマップから注視領域を推定する手法の比較を行なった．評価する手法を表 2 に示す．

従来手法は，人物の姿勢情報を入力としたニューラルネットワークを用いて注視領域を分類を行う．提案手法 1 及び 2 のいずれにおいても，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る．提案手法 2 では構築した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する．そして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 では，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した交差検証を行ない，評価指標としては推定結果の正

表 2 評価手法

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

解率を採用する．

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 3 〜表 4 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．正解率が高いものを赤文字で示す．また，正解率は小数点第 2 位で四捨五入した．

従来手法と比較し提案手法 1 及び 2 の正解率は 0.5 m，1.0 m ともに正解率が上回っていることが確認できる．また，提案手法 1 と提案手法 2 を比較すると提案手法 2 が上回っていることが確認できる．

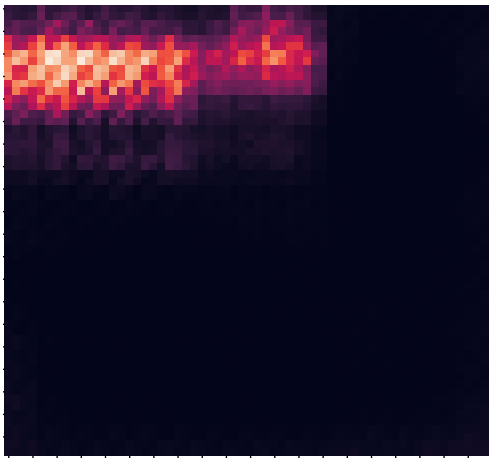

図 9 生成したヒートマップ例

表 3 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 4 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%

また，表 5 及び表 6 に各テストデータにおける提案手法 1 と提案手法 2 の正解率の差分を示す．なお，Person7 の 1.0m-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．

表を見ると，どのテストデータにおいても平均として正解率は向上しており，全てのテストデータ及び距離の正解率の平均向上値は 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

表 5 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 6 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

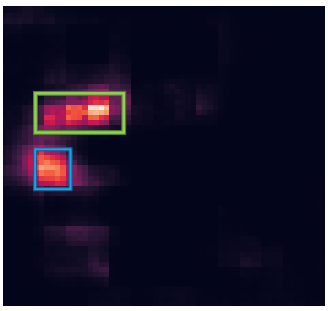

図 10 1.0m-左の Person1 のヒートマップ．緑色の枠が真の注視領域，青色の枠が提案手法 2 による推定された注視領域

の 1.0 m - 左においてはむしろ低下しており，テストデータによっては提案手法 2 による正解率の改善が小さいことが確認できる．その理由を調べるため，推定に用いるヒートマップの生成結果を確認する．まず，向上度合いが最も低い Person1 距離：1.0 m・位置：左のヒートマップを確認する．提案手法 2 が誤推定し，提案手法 1 が正しい推定を行ったヒートマップを図 10 に示す．

この図から提案手法 2 による誤推定が起こる場合，正解領域にピークを持つがその周囲の値が低くなっていることがわかる．このことから，誤推定は物体の配置制約を用いることによってピークを持つ領域の平均値が小さくなっているためであると考えられる．

以上より，提案手法 2 はヒートマップの特徴によっては誤推定を引き起こす場合もあるが，概ねの場合では配置制約を用いることで全般的に正解率を向上させており，注視領域推定において有効であると言える．

6. まとめ，今後の課題

本論文では，棚上の商品を顧客が注視している状況において，姿勢情報から後ろ向き人物の注視領域を推定する手法を提案した．注視する領域に合わせて姿勢が連動して動くことに着目し，姿勢情報を入力としたニューラルネットワークを用いて注視領域を分類する手法 [2] [3] が提案されているが，注視対象の配置関係や大きさを考慮していなかった．そこで本研究では，後ろ向き人物の 3 次元関節点座標から推定の中間表現として注視尤度を示すヒートマップを生成し，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に，物体の配置領域を制約として注視尤度の平均値を取り，比較して最大の領域を注視領域として判定した．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[8] 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以下である．

- 安定したヒートマップ生成
  提案手法では逆畳み込みネットワークを用いて，3 次元関節座標から注視尤度を示すヒートマップを作成した．しかし，格子状に値が分布している場合があった．これは逆畳み込みネットワーク ConvTP 層のアップスケール時に画素ごとに値が加算される回数が異なるためであると考えられる．安定したヒートマップを生成するようにモデルの改良を行なう必要がある．
- 同じ姿勢で異なる領域を注視している場合への対応
  姿勢情報から注視尤度を示すヒートマップを生成する際に，複数のピークを持つヒートマップが得られる場合がある．これは同じ姿勢でも異なる領域を注視しているためだと考えられる．しかし，単一フレームの姿勢情報からではこれを防ぐことは難しい．そのため，時系列でまとめた姿勢情報を生成器の学習に用いるなどの工夫を検討する必要がある．
- 多様な姿勢を含むデータセットへの拡張
  異なる領域を注視する際は姿勢が変わることを想定して，実験を行なったが，ある人物は頭部のみを動かして異なる領域を注視する場合があった．この人物のデータをテストデータとして推定した場合，正解率が著しく低下した．これは他の被験者は概ね姿勢を変化させながら注視を行なっていたため，頭部のみを動かしているようなデータを用いて学習することが出来ていなかったためであると考えられる．このように，人物によって姿勢変化には特徴があり，多様な姿勢に対応するためには多くのデータを用意する必要がある．
- 人物位置の変化への対応
  本研究では，人物の位置ごとにヒートマップ生成器を学習させ，位置ごとのテストデータを用いて注視領域の推定を行なった．実験のなかで，人物位置によって同じ人物でも大きく推定領域の正解率が異なる場合があった．また，実際のコンビニエンスストアでの注視領域推定を考えると，人物の位置は無数にあり，大量の生成器を位置ごとで用意するのは難しい．そのため，多様な位置に対応したデータセットを構築して生成器の入力に用いるなどの，人物の位置変化に柔軟に対応できる手法を検討する必要がある．

参考文献

[1] Takatsugu Hirayama, Yasuyuki Sumi, Tatsuya Kawahara, and Takashi Matsuyama. Info-concierge: Proactive multi-modal interaction through mind probing. In *The Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 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 *Proceedings of the 24th International Conference on Pattern Recognition*, pp. 1731–1736.

[3] 康友川西, 洋村瀬, 建鋒徐, 和之田坂, 広昌柳原. 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. Vol. 85, No. 5, pp. 463–468.

[4] 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.

[5] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement.

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

[7] 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.

[8] Microsoft. Azure kinect dk - ai モデルの開発.

[9] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization.