一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2017-98(2017-10)

低解像度顔画像群からの集団の注目位置推定法の検討

児玉　祐樹† 　川西　康友† 　平山　高嗣† 　出口　大輔†† 　井手　一郎†
村瀬　　洋† 　永野　秀尚††† 　柏野　邦夫†††

† 名古屋大学 大学院情報学研究科　〒 464–8601 愛知県名古屋市千種区不老町
†† 名古屋大学 情報連携統括本部 情報戦略室　〒 464–8601 愛知県名古屋市千種区不老町
††† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所 〒 243-0198 神奈川県厚木市森の里若宮 3-1
E-mail: †kodamay2@murase.is.i.nagoya-u.ac.jp, ††ddeguchi@nagoya-u.jp,
†{kawanishi,ide,murase}@i.nagoya-u.ac.jp, †††{nagano.hidehisa,kashino.kunio}@lab.ntt.co.jp

あらまし　複数の人物が同時に注目している対象を知ることは重要である．しかし，複数人を同時に観測すると各人物画像の解像度は低くなるため，各人物の注目位置の高精度な取得は困難である．そこで我々は，複数人が同一の対象を同時に注視する状況を想定し，画像中の各人物の注目位置の低精度な推定結果を統合することで，集団の注目位置を高精度に推定する手法を考え，様々な統合方法について検討した．また，実験により，推定結果の様々な統計量を用いた統合や統合に用いる人数が集団の注目位置の推定精度に与える影響を分析した．
キーワード　注目対象位置推定，深層学習，低解像度画像，群衆解析

Study on Gaze Target Localization from Low-Resolution Faces of Group of People

Yuki KODAMA†, Yasutomo KAWANISHI†, Takatsugu HIRAYAMA†, Daisuke DEGUCHI††,
Ichiro IDE†, Hiroshi MURASE†, Hidehisa NAGANO†††, and Kunio KASHINO†††

† Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
†† Information Strategy Office, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
††† NTT Communication Science Laboratories, NTT Corporation
3-1, Morinosato Wakamiya, Atsugi-shi, Kanagawa, 243âĂŞ0198 Japan
E-mail: †kodamay2@murase.is.i.nagoya-u.ac.jp, ††ddeguchi@nagoya-u.jp,
†{kawanishi,ide,murase}@i.nagoya-u.ac.jp, †††{nagano.hidehisa,kashino.kunio}@lab.ntt.co.jp

Abstract　It is important to estimate what attracts many people. However, when we observe many people simultaneously, gaze estimation for each person independently is very challenging because of low-resolution. To avoid this problem, we introduce a gaze target localization method of a group of people, which integrates every gaze estimation results under the assumption that all of them are looking at the same object. We analyze the effects of several integration methods on the estimation accuracy, and also analyze the effects of changing the number of people to integrate on the estimation accuracy.
Key words　Gaze target localization, deep learning, low-resolution images, crowd understanding

1. ま え が き

近年，画像処理技術を用いて人物の視線方向や注視対象の推移などの情報（視線情報）を取得する研究が活発に行われている．視線情報は，小売店における消費者の購買行動の分析や，ディジタルサイネージの広告効果の測定など様々な用途への活用が期待されている．また，映画館における広告効果の測定，スポーツ観戦における観衆の注目行動の分析など，時空間を共

有する集団を対象とした応用も期待される．このような集団を対象とした応用（例えば，観衆のように複数の人物が同時に注目している対象の把握）を考えた場合，不特定多数の人物の視線情報を同時に観測する必要がある．その際に，コストや利用者の負担の少なさなどを考慮すると，人物が写った画像からその人物の視線情報を非接触で取得する手法が有用である [1], [2]．

1 人の人物の高解像度顔画像に対する視線推定手法は，これまでに多数提案されている [1], [3]．しかし，不特定多数の人物を撮影する場合には，各人物を高解像度で観測できるように複数のカメラを適切に設置することが困難であるため，この方法は現実的ではない．そこで，1 台の固定カメラを用いて複数の人物を同時に撮影する状況を想定する．このとき，1 台で広範囲を撮影する必要があるため，各人物の顔画像の解像度は低くなってしまう．一般に，低解像度の人物画像の視線方向を安定して推定することは難しいため，このような状況では 1 人の人物を対象とした従来手法では精度良く視線情報を取得することが困難である．

上述の問題に対して，複数の人物の視線情報を用いて集団に属する複数の人物が同時に注目している対象の位置を推定する手法も研究されている [4], [5]．しかし，Park らの手法 [4] では頭部カメラを用いているために，不特定多数の人物には適用できない，また，小嶋らの手法 [5] では推定する視線方向を水平方向に限定しているため，上下方向の推定が必要となる場面では適用できないなどの問題が存在する．一方，Sugano らはパブリックディスプレイに表示された映像に対するアテンションマップの作成において，1 台の固定カメラで撮影された複数の人物の視線情報を統合するアプローチをとっている [6]．このアプローチは注目対象の位置推定にも応用可能であると考えられる．

本研究では，顔や目画像の詳細な情報が得られないような遠方からの画像でも，多数の人物の情報を統合することで，これら集団の注目領域を精度よく推定することを目指す．そして，まず本報告では, 複数の人物が同一の対象を注視している状況を想定し，この集団を 1 台の固定カメラで撮影した画像中から 3 次元空間上で集団が注目している対象の位置（集団の注目位置）を推定する手法を検討する．この状況下では，図 1 のように個々の人物に関する視線推定結果が必ずしも正確でなくとも，集団全体の視線情報が高精度に推定可能であると考えられる．この仮説に基づいて，複数の人物を同時に撮影した低解像度顔画像群から得られる注目対象の位置の推定結果を統合することで，集団の注目位置を高精度に推定する手法を提案する．本報告では，統合を行う人数の増減による精度の変動の分析も行う．

注目位置の推定は, 空間的に離散的な位置を求めるもの [7], [8], 回帰分析などにより空間的に連続的な位置を求めるもの [4], [5] がある．ただし，空間的に離散的な位置を求める問題は，離散的な位置の取り方を細かくすると連続的な位置を求める問題に近似できる．本研究では，学習データの収集が比較的容易な前者の問題を主に扱う．そして，注目位置を既知とし，予め定めた N 箇所の注目位置のうち，注視している位置を識別する問題として定式化する．また，空間的に連続的な位置を求める問題の検討も行う．

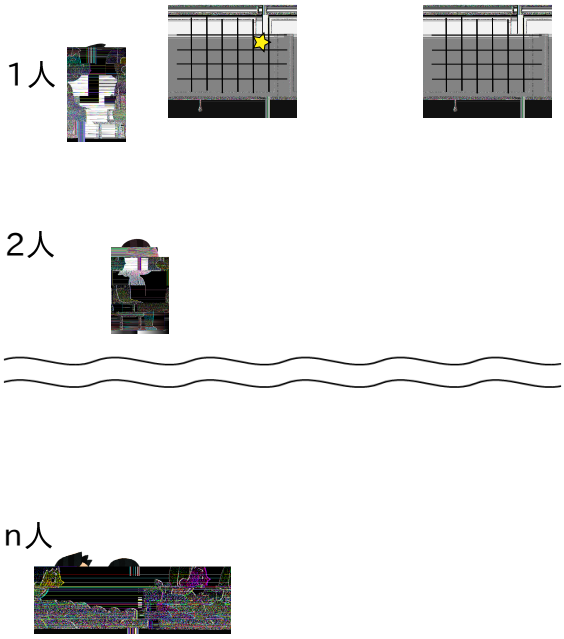


図 1　提案手法の概念図.


図 2　提案手法の処理手順.

2.　集団の注目位置推定

提案手法では，複数の人物が同一の対象を注視している状況を撮影した画像 I を入力とし，1 人の顔画像に対する注目位置識別器 $\boldsymbol{f}(d_i;\theta)$（弱注目位置識別器と呼ぶ）を用いて各人物の顔画像からそれぞれの注目位置を推定する．ただし $\boldsymbol{f}$ は，1 枚の顔画像 d_i を入力とし，パラメータ θ をもとに，各注目位置ごとの注視尤度 $\boldsymbol{p}_i = \boldsymbol{f}(d_i;\theta) = (p_{i1}, p_{i2}, ..., p_{iN})^T$ を出力する関数である．ここで，N は識別対象とする注目位置の数を表す．得られた各顔画像に対する推定結果を統合することで集団の注目位置 R を出力する．本手法は，学習段階と注目位置推定段階の 2 段階からなる．図 2 に提案手法の処理手順を示す．

2.1　学 習 段 階

事前に学習用画像 $\mathcal{I}_{\mathrm{train}} = \{I_1, I_2, ...\}$ を用意し，検出された各人物の顔画像 d_i に対する弱注目位置識別器 $\boldsymbol{f}(d_i;\theta)$ の最適なパラメータ $\hat{\theta}$ を学習する．まず，複数の人物が同一の対象を注視している状況を撮影した学習用画像 $\mathcal{I}_{\mathrm{train}}$ 中から各人物の顔を検出する．顔の検出は，特徴量として HOG 特徴 [9] を利

表 1　CNN のネットワーク構造.

Input	3 × 64 × 64
Conv. 1	Kernel：3 × 3, Channel：6, Maxpool：3 × 3
Conv. 2	Kernel：3 × 3, Channel：16, Maxpool：3 × 3
Conv. 3, 4	Kernel：3 × 3, Channel：24
Conv. 5	Kernel：3 × 3, Channel：16, Maxpool：3 × 3
F.C. 6, 7	Unit：256
F.C. 8	Unit：9

用し，画像ピラミッドに対するずらし照合により実現する(注1)．照合は，線形分類器を用いて顔か否かを判別する．

顔検出によって得られた各顔画像を $W \times H$ 画素に拡縮し，弱注目位置識別器構築のために用いる顔画像 $\{x_i\}_{i=1}^{Q}$ とする．ただし，Q は検出した顔画像の枚数を表す．i 番目の画像と，それに写っている人物が実際に注視していた注目位置 $\boldsymbol{r}_i = (0, 0, ..., 1, ..., 0)^T$ と対応付けて学習データ $T = \{(x_i, \boldsymbol{r}_i)\}$ を作成する．ただし，$\boldsymbol{r}_i$ は i 番目の人物が実際に注視していた注目対象位置に対応する要素が 1 でそれ以外の要素が 0 のベクトルである．この学習データ T を入力として Convolutional Neural Network（CNN）の学習を行う．学習の際，最終層で Softmax 関数を適用し，交差エントロピーによって定義される損失関数

$$L(\theta) = -\sum_i \boldsymbol{r}_i \cdot \log \boldsymbol{p}_i \tag{1}$$

を誤差関数とする．この損失関数 $L(\theta)$ を最小にする

$$\hat{\theta} = \arg\min_{\theta} L(\theta) \tag{2}$$

を求める．

表 1 に学習に使用したネットワーク構造の詳細を示す．このネットワーク構造は AlexNet [10] を参考にして構築した．ただし，AlexNet は，224 画素四方の入力画像を対象としており，本研究における入力画像のような低解像度画像には適していない．そこで，低解像度画像に対応するために，入力を 64 画素四方とし，各層のチャネル数および，カーネルサイズを縮小した．また，Pooling 層において過度な画像縮小を避けるため，ストライドを上下・水平方向ともに 1 とした．

2.2　注目位置推定段階

注目位置推定段階では，まず学習段階と同様に，複数の人物が同一の注目対象を注視している状況を撮影した評価画像 E から顔を検出して切り出す．そして，$W \times H$ 画素に拡縮した顔画像 x_i に対し，学習した弱注目位置識別器 $\boldsymbol{f}(d_i; \hat{\theta})$ を用いて各注目位置ごとの注視尤度 $\boldsymbol{p}_i = (p_{i1}, p_{i2}, ..., p_{iN})^T$ を算出する．その後，各人物に対して算出された尤度を統合し，最終的に推定した注目位置 R を出力とする．

2.2.1　各人物の注目位置推定

まず，複数人が同一の対象を注視している状況を撮影した評価画像 E 中から各人物の顔を検出する．検出には，学習段階と同様に HOG 特徴 [9] を利用した手法を用いる．

顔検出によって切り出した顔画像を学習段階と同様に $W \times H$ 画素に拡縮し，学習段階で構築した弱注目位置識別器 $\boldsymbol{f}(d_i; \hat{\theta})$ に入力する．その結果，各顔画像 $\{x_i\}_{i=1}^{M}$ に対して，注目位置ごとの注視尤度の算出結果

$$\boldsymbol{p}_i = (p_{i1}, p_{i2}, ..., p_{iN})^T = \boldsymbol{f}(x_i; \hat{\theta}) \tag{3}$$

が得られる．

2.2.2　集団の注目位置推定

得られた各顔画像 x_i（$i = 1, 2, ..., M$）から推定した各注目位置に対する注視尤度を並べたベクトル $\boldsymbol{p}_i = (p_{i1}, p_{i2}, ..., p_{iN})^T$（$i = 1, 2, ..., M$）を統合し，最終的な出力結果 $\hat{R}$ を得る．具体的には，統合手法として「注目位置別の注視尤度の中央値」と「注目位置別の注視尤度の平均値」および「注視尤度が最大値となる注目位置の多数決」を考える．

2.2.3　注目位置別の注視尤度の中央値

検出された顔 x_i（$i = 1, 2, ..., M$）それぞれに対して求めた注目位置別の注視尤度ベクトル $\boldsymbol{p}_i = (p_{i1}, p_{i2}, ..., p_{iN})^T$ を入力とし，次式により各注目位置の尤度を算出する．

$$\acute{p}_k = \underset{i=1,...,M}{\text{median}}\, p_{ik} \tag{4}$$

ここで，M は検出された顔の総数である．これにより，集団としての注目位置別の注視尤度 $\acute{\boldsymbol{p}} = (\acute{p}_1, \acute{p}_2, ..., \acute{p}_N)^T$ を得る．得られた集団としての注目位置別の注視尤度 $\acute{\boldsymbol{p}}$ から，注視尤度 $\acute{\boldsymbol{p}}$ の最大要素を求め，この要素に対応する位置を推定した注目位置 $R_{\text{med}} = \arg\max_k \acute{p}_k$ とし，最終的な出力結果とする．

2.2.4　注目位置別の注視尤度の平均値

「注目位置別の注視尤度の中央値」で求めた $\acute{\boldsymbol{p}}$ の代わりに，注目位置別の注視尤度 $\boldsymbol{p}_i = (p_{i1}, p_{i2}, ..., p_{iN})^T$ の平均

$$\acute{\boldsymbol{p}} = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{p}_i \tag{5}$$

を用いる．得られた集団としての注目位置別の注視尤度 $\acute{\boldsymbol{p}}$ から，注視尤度 $\acute{\boldsymbol{p}}$ の最大要素を求め，対応する注目位置 $R_{\text{ave}} = \arg\max_k \acute{p}_k$ を最終的な出力結果とする．

2.2.5　注視尤度が最大値となる注目位置の多数決

検出された顔それぞれに対して求めた注目位置別の注視尤度ベクトル $\boldsymbol{p}_i = (p_{i1}, p_{i2}, ..., p_{iN})$ を入力し，次式により顔画像ごとの推定位置を算出する．

$$V_i = \arg\max_k p_{ik} \tag{6}$$

次に，顔画像ごとの推定位置 V_i を式 (7) により投票する．

$$R_k = \sum_{i=1}^{M} X(V_i = k) \tag{7}$$

ただし，X は指示関数であり，$V_i = k$ の時に 1 を，そうでなければ 0 を返す．最後に，投票数が最大となる注目位置 $R_{\text{maj}} = \arg\max_k R_k$ を最終的な出力結果とする．

(注1)：Dlib C++ Library http://dlib.net/

3. 実　　験

3.1 データセットの作成

まず，12 名の被験者を撮影し，データセットを作成した．撮影の際には，特定の位置に固定した椅子に座った複数の被験者に，スクリーン上に提示した 1 つの視標を同時に注視させて撮影した．被験者が座る位置（被験者位置）は図 3 に示すようにスクリーンおよびカメラからの距離が異なる 2 種類の環境内で設定した．これによって，撮影される人物の解像度が異なる 2 種類の画像を得る．この解像度が異なる 2 種類の画像に対する結果を比較することで，解像度の変化に対する提案手法の頑健さを確認する．スクリーン中心の法線上の 360 cm（近距離）もしくは 660 cm（遠距離）離れた位置に 1 箇所の被験者位置を設定し，座席の中心の間隔が 60 cm となるようスクリーンから見て右側に 3 箇所，左側に 2 箇所の計 6 箇所を設定した．

図 4 に示すように，注目対象（視標）の表示位置は，壁から 40 cm の位置にあるスクリーン上の 9 箇所とし，十字の視標を 1 箇所ずつ順に表示した．図 4 に示すように，視標中心間の距離は縦 70 cm，横 93 cm であり，視標の大きさは縦 12 cm，横 8 cm とした．このとき近距離環境では，視標中心間の視角が縦約 11°，横約 15° であった．また，遠距離環境では，視標中心間の視角が縦約 6°，横約 8° であった．

スクリーンの真下に設置したカメラを用いて，被験者が視標を注視する様子を撮影した．カメラは地面から 94 cm の高さに設置した．使用したカメラは POINT GREY 社製の FL3-U3-13E4C-C であり，解像度は 1,280 × 1,024 画素，フレームレートは 60 fps，8 bit カラーの条件で撮影した．また，使用したレンズは 35 mm 換算で焦点距離 31 mm であった．各被験者が全ての被験者位置に 1 回は着席するようにした．図 5 に撮影した画像の例を示す．

3.2 実 験 方 法

評価実験において評価データに使用した顔画像は近距離環境・遠距離環境ともに 324 枚（6 名の被験者× 6 箇所の被験者位置× 9 箇所の視標表示位置）である．また，学習用データは被験者 6 名に対して検出された全ての顔と注目位置の組とした．具体的には，近距離環境では 18,256 組，遠距離環境では 16,908 組であった．さらに，全ての学習データに対し，顔画像の検出のずれを考慮して，上下左右 6 画素以内での検出位置のずらしおよび倍率 0.9 倍から 1.1 倍までの拡縮を無作為に施した．以上の方法で，CNN の学習のエポックごとに，学習データに含まれる顔画像と同じ枚数の新たな学習データを無作為に生成しながら弱注目位置識別器を構築した．

構築した弱注目位置識別器に評価データを入力することで，顔画像ごとに視標位置ごとの注視尤度を算出し，その推定結果の統合を行った．この時，結果の統合を行う人数を変化させ，結果の統合を行う人数の増加に伴う推定成功率の変化を調べた．テストデータには 6 人しか含まれていないが，7 人以上を統合する場合を考えるため，仮想的に同じ被験者位置に複数の被験者が存在する状況を設定し，異なる位置に座った同一被験者を異なる被験者として扱った．

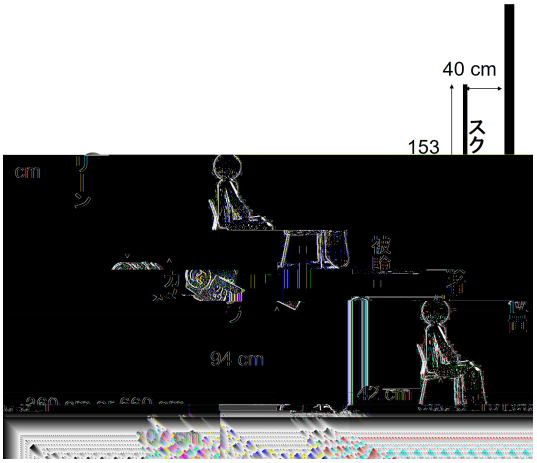


図 3　撮影環境の側面図.

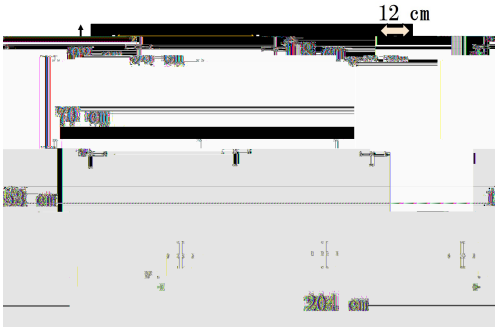


図 4　視標の表示位置.

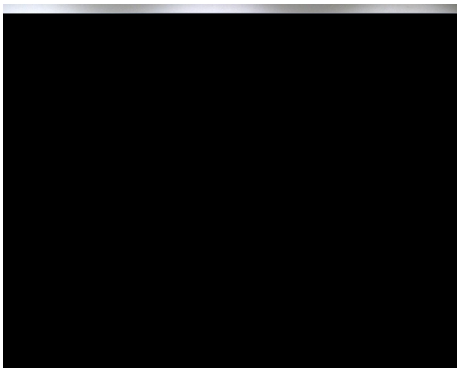


図 5　撮影画像の例.

3.3 結果および考察

図 6 および図 7 に統合する人数の増加にともなう推定成功率の変化を示す．その結果，いずれの統合手法でも統合を行う人数の増加に伴い，注目位置の推定精度が向上する傾向を確認した．遠距離環境において，統合を行わない場合の推定成功率は 28.40 %であったが，提案手法により 36 人の推定結果を統合することで 68.89 %まで推定成功率が向上した．図 6 および図 7 の結果から，撮影環境が近距離環境であるか遠距離環境であるかに関わらず，いずれの統合手法でも統合人数の増加に伴って注目位置の推定精度が向上する傾向が読み取れる．これより，提案手法は解像度の低下に頑健な手法であると考えられる．

4. 空間的に連続的な位置を求める問題の検討

3 章の実験では，空間的に離散的な位置を求める注目位置推定において，統合を行う人数の増加に伴い注目位置の推定精度が向上する傾向を確認した．そこで以下では，空間的に連続的

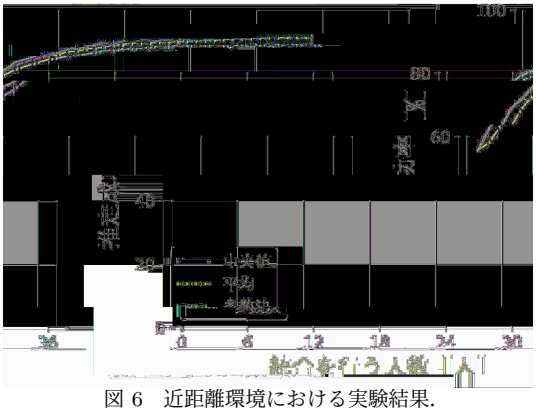


図 6　近距離環境における実験結果.

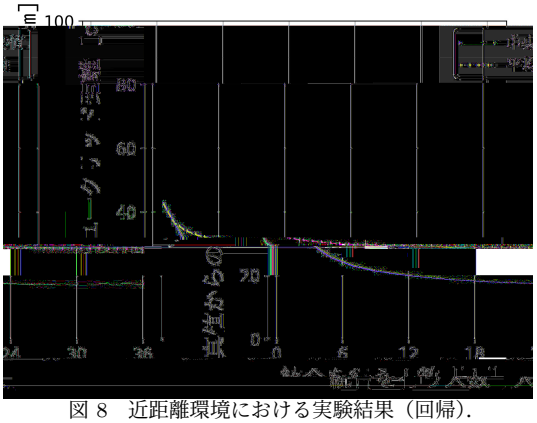


図 8　近距離環境における実験結果（回帰）.

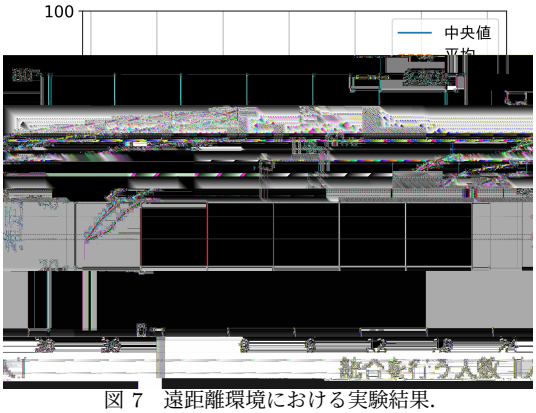


図 7　遠距離環境における実験結果.

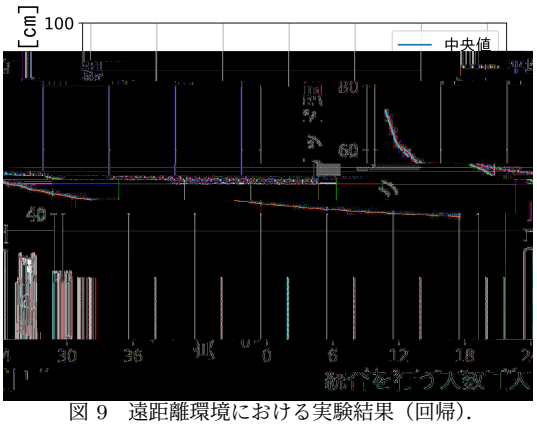


図 9　遠距離環境における実験結果（回帰）.

な位置を求める注目位置推定において，統合を行う人数の増加に伴う注目位置の推定精度向上の可能性に関する検討を行う．

4.1　推定器の作成

本検討では，検出された各人物の顔画像 d_i に対して，回帰によって空間的に連続的な位置であるスクリーン上の XY 座標 $\boldsymbol{r}_i = (x_i, y_i)^T$ を推定する推定器 $\boldsymbol{f}(d_i;\theta)$ を作成し，推定結果を統合する手法を考える．回帰手法としては，深層学習による回帰を用いる．具体的には，表 1 に示した提案手法のネットワーク構造の出力層である FC8 層の Unit 数を 2 チャネルに変更し，最終層の Softmax 関数を取り除く．また，学習の際の損失関数を，式 8 に示した交差エントロピーによって定義される損失関数から，平均二乗誤差によって定義される損失関数

$$L(\theta) = \frac{1}{n}\sum_{i}^{n}(\boldsymbol{r}_i - \hat{\boldsymbol{r}}_i)^2 \tag{8}$$

に変更する．

このようにして作成した推定器を用いることで，各人物の顔画像から，その人物が注視している位置 $\hat{\boldsymbol{r}}_i = (\hat{x}_i, \hat{y}_i)^T$ を推定できると考えられる．

4.2　統 合 手 法

作成した推定器による各顔画像に対する推定結果を統合することで，集団の注目位置 R を推定する．具体的な統合手法として，「座標の中央値」による統合と「座標の平均値」による統合を考える．

4.2.1　座標の中央値

検出された顔それぞれに対して求めたその顔画像の人物が注視している対象の XY 座標の推定結果 $\hat{\boldsymbol{r}}_i = (\hat{x}_i, \hat{y}_i)^T$ を入力とし，次式により集団の注目対象の XY 座標の推定結果 $R_{\mathrm{med\_regression}}$ を算出する．

$$\begin{aligned} R_{\mathrm{med\_regression}} &= \acute{\boldsymbol{r}} = (\acute{x}, \acute{y}) \\ \acute{x} &= \underset{i=1,\ldots,M}{\mathrm{median}}\ \hat{x}_i \\ \acute{y} &= \underset{i=1,\ldots,M}{\mathrm{median}}\ \hat{y}_i \end{aligned} \tag{9}$$

ここで，M は検出された顔の総数である．得られた集団の注目対象の XY 座標の推定結果 $R_{\mathrm{med\_regression}}$ を最終的な出力結果とする．

4.2.2　座標の平均

「座標の中央値」で求めた $\acute{r}$ の代わりに，座標の平均

$$R_{\mathrm{ave\_regression}} = \acute{\boldsymbol{r}} = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{r}_i \tag{10}$$

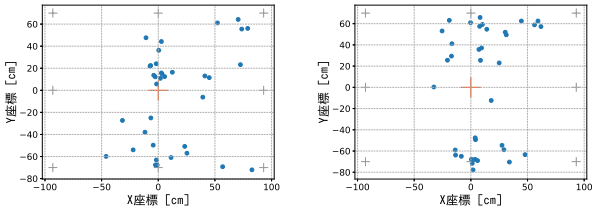
を用いる．得られた集団の注目対象の XY 座標の推定結果 $R_{\mathrm{ave\_regression}}$ を最終的な出力結果とする．

4.3　結果および考察

3.2 節と同様の実験方法で実験を行った．評価指標は真値と推定結果との Euclidean 距離 $D = ||\boldsymbol{r}_k - R_k||$ を用いる．

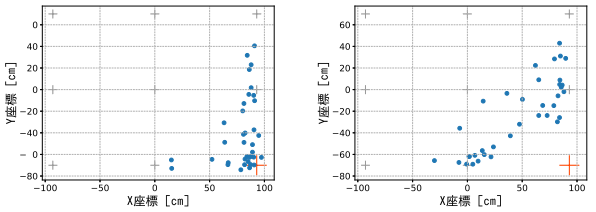
図 8 および図 9 に統合する人数の増加にともなう真値との Euclidean 距離の変化を示す．その結果，いずれの統合手法でも統合を行う人数の増加に伴い，真値との Euclidean 距離が短くなる傾向を確認した．遠距離環境において，統合を行わない場合の真値との Euclidean 距離は 72.29 cm であったが，提案手法により 36 人の推定結果を統合することで 39.33 cm まで短

図 10　右下の視標位置を注視する人物の顔画像に対する推定結果．
十字は視標位置の真値を表し，青い点が推定結果を表す．



(a) 全ての視標位置で学習した場合．(b) 中央の視標を学習から除いた場合．

図 11　中央の視標位置を注視する人物の顔画像に対する推定結果．
十字は視標位置の真値を表し，青い点が推定結果を表す．



(a) 全ての視標位置で学習した場合．(b) 右下の視標を学習から除いた場合．

図 12　右下の視標位置を注視する人物の顔画像に対する推定結果．
十字は視標位置の真値を表し，青い点が推定結果を表す．

くなった．

以上より，空間的に連続的な位置を求める注目位置推定においても，統合を行う人数の増加に伴って注目位置の推定精度が向上する傾向があると考えられる．しかし，作成した推定器による各個人の推定結果を確認すると，図 10 のように本来想定される真値を中心とした分布ではなく，真値およびその周辺の視標位置付近に分布していることが確認された．本実験で用いた統合手法は，推定結果が真値を中心とした分布となることを前提としている．そのため，この推定器による推定結果を統合しても，推定精度の向上は限定的であったと思われる．これは，推定器を作成する克晻鐮亻 槝蛤合