Vol.2019-CVIM-217 No.13
2019/5/30

深層学習を用いた三人称視点映像からのアイコンタクト識別

大嶋 佑紀[1,a)]　中澤 篤志[1]　西田 豊明[1]

概要：アイコンタクトは介護従事者の介護スキルの評価や ASD(自閉症スペクトラム) の早期発見などに利用できる指標であるため，自動検出しようとする試みがいくつか存在する．従来は一人称視点映像から検出する手法が知られていたが，本研究では三人称視点映像からアイコンタクト識別する手法を提案する．本手法の識別器は，シーン中の 2 者間の視線方向・目領域画像・相互の位置関係を入力とする畳み込みニューラルネットワーク（Convolutional Neural Network）である．複数人が対話を行う動画データセットに対して実験を行った結果，相互の視線と位置関係のみを用いるよりも，両者の目領域画像を加えることで精度が向上することが確認できた．ここから，画像から検出された視線方向以外の画像特徴が，アイコンタクトの識別に有用であることが示唆される．

Eye contact discrimination from third person video with deep learning

Yuki Ohshima[1,a)]　Atsushi Nakazawa[1]　Toyoaki Nishida[1]

1. はじめに

アイコンタクトは，二人の人物が互いに目と目を見合わせる行為のことであり，非言語コミュニケーションの一つである．会話の際に相手の目を見ることは礼儀とされているように、アイコンタクトは社会生活において重要な役割を持っている．

介護の現場においても，アイコンタクトは重要な役割を持っている．認知症の介護の際，暴言や暴力が介護を妨げる要因となることが報告されており [1]，これらの状況を減らす方法の一つとして，ユマニチュード [2] という介護ケア手法がある．ユマニチュードでは，「見つめる」「話しかける」「触れる」「立たせる」の四技能が重視されているが，その中の「見つめる」は，介護従事者が被介護者とどれだけアイコンタクトできるかが良い介護をすることの指標となることを意味しており，アイコンタクト検出は介護スキルを測定するための指標として用いられる．

また，ASD（自閉症スペクトラム）の早期発見にもアイコンタクト検出は役立てられる．ASD はアイコンタクトをとることが難しい場合が多いということから，乳幼児の視線情報を用いた ASD の早期発見ができるということが報告されている [3]．

このような背景から，アイコンタクト検出を一人称視点映像から行おうとする試みがある [4], [5]．この研究では，頭部装着型カメラで撮影した一人称視点映像をもとにアイコンタクトの自動検出を行う．

しかしながら，すべての介護現場において頭部装着型カメラを用いることは，機器の入手可能性や介護作業の妨げになるという点から，現実的とは言えない．

それに対して，通常撮影される視点の映像（三人称視点映像）から人物間のアイコンタクト検出を行うことできれば，これらの問題が解決できると考えられる．また，複数人が映っている映像からアイコンタクトをしている二者を検出するといった使い方も可能であるため，汎用性が高くなることが見込まれる．

よって本研究では，三人称視点映像からのアイコンタクト検出手法を提案する．顔検出は HyperFace[6] などの精度の高い既存手法があるので，これらで得られた顔検出システムから得られた顔画像を入力としてアイコンタクト識別をする手法を考える．

[1] 京都大学 Kyoto University
[a)] ohshima@ii.ist.i.kyoto-u.ac.jp

2. 関連研究

2.1 アイコンタクト識別に関する先行研究

一人称視点映像からのアイコンタクト検出を行った研究として，Ye らの手法 [5] が挙げられる．Ye らは，大人と乳幼児との間のアイコンタクト検出を行った．その手法は，まず，大人に Pivothead という頭部装着型カメラデバイスを装着させて乳幼児の一人称視点映像を撮影し，その映像の各フレームに対して顔検出を行い，顔のクリッピングを得る．次に，顔のクリッピングを混合ガウシアンモデルを用いて 3 クラスにクラスタリングする．クラスタリングされた顔画像から顔ランドマーク点推定により目領域のクリッピングを得る．顔方向特徴に対するクラスタリングの処理を挟んでいるのは，似通った目領域画像で顔方向が違う場合に対処するためである．顔検出には OMRON OKAO Vision が，顔ランドマーク点推定には IntraFace[7] がそれぞれ用いられた．その後，目領域画像の特徴抽出を行い，得られた特徴量と真値から Random Forest を学習させることにより識別器を構成している．特徴抽出の部分は，LBP(Local Binary Pattern), CNN, HOG(Histograms of Oriented Gradients) の 3 通りの手法について検証している．

また，Ye らはこの単一フレームに対するアイコンタクト検出モデルの他に，複数フレームを用いた時系列学習を行うモデルも提案している．このモデルでは，各フレームから求めた尤度を複数個まとめて linear-chain CRF に入力するようになっている．

Ye らは別の研究 [8] で，アイコンタクトが大人と乳幼児との双方向のコミュニケーションであることに着目した手法を提案している．この手法では，大人に，シーンカメラとアイカメラを備えた眼鏡型カメラデバイスを装着させることで，前方の乳幼児と装着者の目を同時に撮影し，双方の視線を考慮したアイコンタクト検出を可能にした．

Mitsuzumi らは，一人称視点映像を用いた顔検出におけるフレームアウトの問題に着目し，顔パーツが隠れている場合にもロバストな手法を提案した [9]．この手法では，顔検出を行わずにアイコンタクト検出を行う．具体的には，まず，sliding window によって両目領域の候補を挙げ，それらを両目領域画像かどうかを学習させた SVM の識別器にかける．両目領域画像であると判定された画像に対してランドマーク点推定を行い，左右の目領域画像を得る．左右の目領域画像を別々の CNN に通した後，それらをまとめて 3 層の全結合層に入力している．これらのニューラルネットを学習させることにより識別器を得る．また，ランドマーク点から Random Forest 回帰によって顔方向特徴を得て，それを目画像から得られた特徴と合わせて全結合層に入力するモデルも提案している．

Zhang らは，一人称視点映像の単一フレームからのアイコンタクト検出に視線方向推定を用いる手法を提案した [10]．この手法のモデルは，まず対象フレームに対して OpenFace[11] による顔ランドマーク点推定を行い，それを Zhang らの他の研究の手法 [12] による CNN を用いた視線方向特徴抽出器にかけて視線方向特徴を得る．その後，視線方向特徴を OPTICS アルゴリズム [13] を用いて対象物ごとにクラスタリングし，クラスターの中からカメラ位置に最も近いクラスターを正解のクラスターとして選択する．このクラスターの視線方向特徴を重み付きの SVM を学習させた分類器にかけることでアイコンタクト識別を行う．

本研究の位置付けは，Zhang らの手法 [10] のように単一フレームからの視線方向推定結果を用いてアイコンタクト識別を行う手法をベースライン手法とし，Mitsuzui らの手法 [9] をもとに目画像を入れる識別器へと拡張した手法と，Ye らの研究 [5] で行われた時系列化を Ye らとは異なる方法で実現する手法を提案するというものになる．

2.2 視線方向推定に関する先行研究

Recasens らは，三人称視点の画像から視線方向推定を行った [14]．手法のモデルは，視線方向を推定したい人物が映っているシーン画像とその人物の顔画像のクリッピングおよび顔の二次元位置座標を入力として，注視の尤度をヒートマップとして出力する．シーン画像と顔画像はそれぞれ別々の畳み込み層に入力されて特徴抽出を行い，シーン画像からはシーン全体での顕著性を表したヒートマップを出力し，顔画像と顔の位置からは注視点のヒートマップを出力する．これらを要素ごとの積を求めることで統合して注視点のヒートマップを作成する．最後にこのヒートマップから shifted Grid によって視線方向ベクトルの推定を行う．ヒートマップの出力に，シーン全体のヒートマップと注視点推定という 2 つの異なる経路を用いているのは，人間の注視行動が，まず人物の視線方向を推定し，その後視線の先にある目立った対象物を探すというプロセスになっているという仮説に基づく．

Chong らは，Recasens らの手法を拡張することで識別精度の向上を実現した [15]．この手法のモデルは，顔画像の畳み込みから得られた特徴から注視角を推定する補助タスクを追加している．この部分のロスには，L_1-loss に実際の視線と注視ベクトルの推定値との余弦距離を加えたものを利用している．また，推定された注視角は，シーン画像のヒートマップとともに全結合層に入力され，注視尤度推定に利用されている．これにより，注視角に基づいてヒートマップの値を操作できるようになり，人物がどこも見ていない場合はヒートマップの値を 0 に近づけることで，対象物がシーン画像の外にある場合にも対処可能となった．

本研究は，フレーム全体とその一部のクリッピングと位置情報を入力する点がこの研究と似ているが，アイコンタ

クト識別を問題にしているという点が異なる．視線方向推定では，人が物を見る状態を対象にしているのに対し，アイコンタクト識別は人と人がお互いを見合う状態を対象にしているという違いがあるので，本研究ではこの違いを生かして，双方の視線に関する情報を入力する手法を提案する．また，本研究では時系列化を行うという点もこの研究と異なる点である．

3. 手法

3.1 提案手法の概要

まず，本研究で識別しようとするアイコンタクトの定義を述べておく．ある人物 A が他のある人物 B の目周辺を見ている状態を”A から B へのゲイズがある”と定義する．すると，人物 A と人物 B がアイコンタクトをしている状態は，”A から B へのゲイズと B から A へのゲイズがともにある”と言い換えることができる．本研究では，双方向のゲイズ識別器を組み合わせる形でアイコンタクト識別器を構成する．ゲイズ識別器の処理の流れは図 1 の通りである．

アイコンタクト識別器の処理の流れは図 1 のようになる．入力は，アイコンタクト識別する二者の時間的に連続した顔画像であり，出力は，アイコンタクトありを意味する 1 とアイコンタクトなしを意味する 0 の二値である．顔画像にランドマーク点推定と視線方向推定を行う．ランドマーク点推定で得られた目の周辺のランドマーク点から，目領域を検出し，目画像と目の中心位置を得る．視線方向推定の結果得られたベクトルと目画像および目の中心位置座標をもとに，畳み込みニューラルネットを用いた分類器によってアイコンタクト識別を行う．分類器の出力はゲイズあり・ゲイズなしの 2 クラスの尤度となり，あらかじめ設定した閾値以上となっているクラスに分類する．分類器の学習は，人がラベル付けした教師データつきの訓練データセットに対して誤差逆伝播によりニューラルネットを学習させることで行う．顔ランドマーク点推定と視線方向推定にはどちらも OpenFace[11] を用いる．

畳み込みニューラルネットを用いた分類器の部分は，以下の 4 通りを用意した．まず，視線方向と人物の目の位置を入力とする Baseline algorithm を作成した．次に，Mitsuzumi らの目画像を利用した一人称視点映像からのアイコンタクト検出の研究 [9] のように，目画像を利用する手法を定義し，これを提案手法 1 とした．さらに，Ye らの手法で時系列化による高精度化が実現したことを踏まえて，目画像を時系列データとして扱うように提案手法 1 を拡張した手法（提案手法 2）を定義した．また，アイコンタクトが二者間のコミュニケーションであることに着目して，目画像・位置・視線方向を相互に入力として与える手法（提案手法 3）を定義した．

3.2 OpenFace による顔ランドマーク点推定と視線方向推定

OpenFace の顔ランドマーク点推定は CLNF[16] を利用する．視線方向推定では，ランドマーク点推定で得られた目と瞳孔の位置をもとに，左右の目の視線方向ベクトルを別々に計算している．平面の顔画像から，瞳孔の中心を通るベクトルを算出し、眼球球との交点を計算することで三次元カメラ座標上の瞳孔位置を割り出し，眼球中心から瞳孔位置へのベクトルを計算することで視線方向ベクトルを求めている．

まず，左目のランドマーク点を $\boldsymbol{e_0}, \boldsymbol{e_1}, \cdots, \boldsymbol{e_5}$ として，目を含む矩形の 4 頂点 $\boldsymbol{r_{00}}, \boldsymbol{r_{01}}, \boldsymbol{r_{11}}, \boldsymbol{r_{10}}$ を以下の式で設定する．

$$\begin{aligned} \boldsymbol{r_{00}} &= (e_{0_x} - margin_x, \min_{i=1,2}(e_{i_y} - margin_y)) \\ \boldsymbol{r_{01}} &= (e_{0_x} - margin_x, \max_{i=4,5}(e_{i_y} + margin_y)) \\ \boldsymbol{r_{11}} &= (e_{3_x} + margin_x, \max_{i=4,5}(e_{i_y} + margin_y)) \\ \boldsymbol{r_{10}} &= (e_{3_x} + margin_x, \min_{i=1,2}(e_{i_y} - margin_y)) \end{aligned} \tag{1}$$

ただし，今回は $\boldsymbol{margin} = (5, 5)$ とした．この矩形画像をグレースケール化した後，32 × 32 に resize し，さらに GCN により以下の式の通り正規化したものを目画像特徴として用いる．

$$x \leftarrow \frac{x - \overline{x}}{\sigma} \tag{2}$$

3.3 畳み込みニューラルネットによるアイコンタクト識別

3.3.1 視線方向と位置を用いたニューラルネットによる分類 (Baseline algorithm)

この手法では，OpenFace による視線方向特徴と目の位置座標のみを入力とする全結合ニューラルネットを学習させることによって分類器を得る．分類器の入力は，OpenFace を用いて求めた人物 A の視線方向推定結果の 6 ユニット (gaze estimation)，人物 A の目の中心の座標位置の 2 ユニット (eye locationA) および人物 B の目の中心の座標位置の 2 ユニット (eye locationB) である．視線方向は，人物 A の両目の視線方向ベクトル (x, y, z) を合わせたものを入力し，目の位置は，それぞれの人物のよく見えているほうの目の中心の二次元座標位置 (x, y) を入力する．よく見えているほうの目は，カメラから見て右側にいる人物に対しては左目，左側にいる人物に対しては右目を選択するようにした．これらの入力は，一つにマージされ，10 ユニットの全結合層 2 層を経て 2 ユニットの出力層へつながっている．中間層の後には LeakyReLU を活性化関数として用いており，出力層の後には Softmax 関数を用いている．分類器の出力は各クラスの尤度であり，この値と真値の 0, 1 とのクロスエントロピーをロスとして誤差逆伝播によりニューラルネットを学習する．

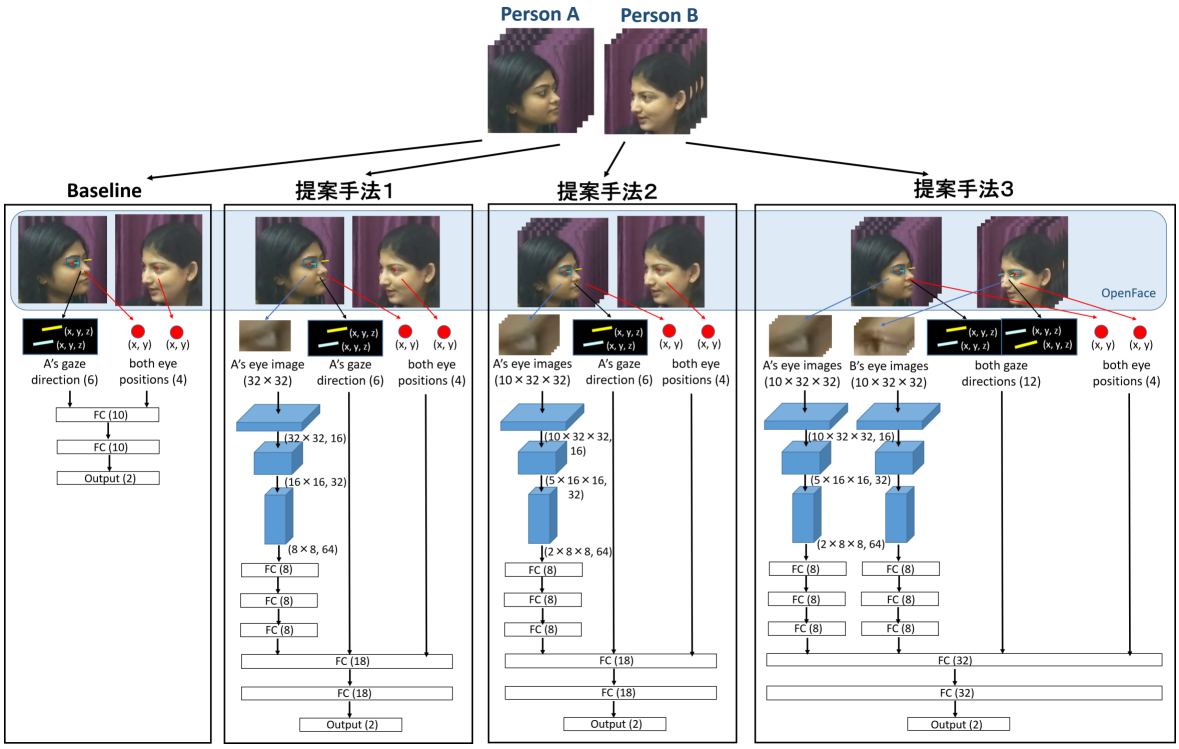

図 1　ゲイズ識別器の概要図

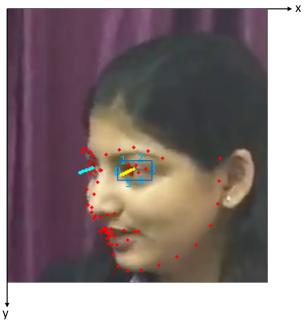

図 2　顔ランドマーク点推定 (赤点) と視線方向推定 (右目：水色，左目：黄色) の適用例．青枠は切り出す矩形領域

3.3.2　視線方向・位置・目画像を用いたニューラルネットによる分類 (提案手法 1)

この手法では，Baseline algorithm の分類器の入力に，目画像を加える．加える目画像は，人物 A の顔画像から 3.2 節の通りにクリッピングされたものであり，Baseline algorithm と同じ方法で片方の目のみを選択して入力とする．分類器の入力は，この目画像特徴 32 × 32 と，視線方向 6 ユニット，位置 2 ユニットの A, B 二つ分である．目画像は畳み込み層 2 層を重ねたものを直列に 3 つ，プーリング層を 1 層ずつ介して連結されたネットワークに入力される．フィルターの数は，初めの 2 層では 16 個，中間の 2 層では 32 個，最後の 2 層では 64 個である．畳み込み層のフィルターのサイズはいずれも 3 × 3 であり，いずれもゼロパディングを行う．プーリング層はプーリングサイズ 2 × 2，ストライド (2, 2) で MaxPooling を行い，画像のサイズを半分にしている．各畳み込み層の後には LeakyReLU 関数による活性化が行われる．この目画像の畳み込みの部分のネットワーク構造は，Mitsuzumi らの手法のネットワーク構造を参考にした．畳み込み処理によって得られた特徴は，1024 ユニット, 124 ユニット, 8 ユニットの 3 層の全結合層を通って他の入力とマージされる．1024 ユニットの全結合層にはドロップアウトを入れている．その後，10 ユニットの全結合層 2 層を経て 2 ユニットの出力層へつながっている．中間層の後では LeakyReLU で活性化を行い，出力層の後では Softmax 関数を用いて尤度を出力している．分類器は，尤度が 0.5 より大きければ 1 を，そうでなければ 0 を出力する．

3.4　時系列推定への拡張

3.4.1　視線方向・位置・時間的に連続した目画像を用いた畳み込みニューラルネットによる分類 (提案手法 2)

この手法では，提案手法 1 の二次元畳み込みを三次元畳み込みに変更し，入力目画像も，過去 9 フレーム分と合わせて 10 枚をまとめて入力する．分類器の入力は，人物 A の 10 × 32 × 32 の目画像時系列特徴と，視線方向 6 ユニット，位置 2 ユニットの A, B 二つ分である．ネットワーク構造については，提案手法 1 とほとんど同じであるが，三次元畳み込みを行うために，畳み込み層のフィルターサイズが 3 × 3 × 3 に，プーリング層のプーリングサイズが 2 × 2 × 2 に，ストライドが (2, 2, 2) にそれぞれ変更され

る．全結合層 3 層を通った後に他の入力とマージされる部分も提案手法 1 と同様であるが，この手法では 1024 ユニットと 124 ユニットの全結合層にドロップアウトを入れている．マージされた後の処理については提案手法 1 と同様である．

3.4.2 相互の視線方向・位置・時間的に連続した目画像を用いた畳み込みニューラルネットによる分類 (提案手法 3)

提案手法 1, 2 の手法では，人物 A から人物 B へのゲイズ分類器に，位置については人物 A, B 両方の情報を入力していたが，目画像および視線方向については人物 A のもののみを用いていた．提案手法 3 では，人物 B の目画像と視線方向の情報も入力に追加する．人物 A, B の目画像特徴は別々の CNN に入力され，ユニット数 1024, 124, 8 の全結合層を経由して他の入力とマージされる．その後，ユニット数 32, 32, 2 の 3 層の全結合層へつながっている．その他の細かい処理やパラメータに関しては提案手法 2 と同様である．

4. 実験

4.1 データセット

YouTube にアップロードされている動画の中から 5 本を選び，アイコンタクト識別用のデータセットとする．使用した動画 a〜e のタイトルと URL は以下の通りである．

a. KNOW YOUR SENIOR(XIMB-XUB) - IN CONVERSATION WITH MANALI MADHUCHHANDA
`https://www.youtube.com/watch?v=KOVp-K5-dR0`

b. GIRL TALK | Boys, 'Does he like me?' & First Dates! with AmeliaLiana
`https://www.youtube.com/watch?v=5wWHS8toEj0`

c. HCA Auction - Just Chatting about the Auction
`https://www.youtube.com/watch?v=lPSzK\_wGACk`

d. Watch heart-melting moment Alzheimer's patient recognises her daughter and says 'I love you'
`https://www.youtube.com/watch?v=xKBcE4KhFPc`

e. 福耳伝説 34　(株) 河野設備　河野昭久氏
`https://www.youtube.com/watch?v=bZ2Y4FrVeOI`

各動画から時間的に連続した，人物の顔が映っているフレームをほぼ同数ずつ取り出してデータセットとした．各動画でカメラから見て右側の人物を right，左側の人物を left で表し，それぞれの人物が相手の目周辺を見ている（ゲイズありの）フレームに対しては 1，そうでない（ゲイズなしの）フレームに対しては 0 の真値をつけた．ただし，d の動画の左側の人物は，眼鏡を着用しているため，今回のデータセットからは省くことにした．また，e の動画に関しては，4 人の人物が映っているため，右の 2 人にのみ注目した．

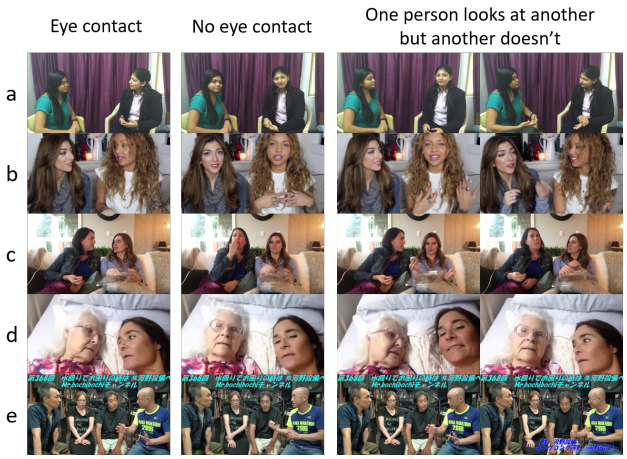

図 3　データセットの例

	Baseline	提案手法 1	提案手法 2	提案手法 3
a-left	0.40253	0.74708	0.71709	**0.76708**
a-right	0.48950	0.62746	**0.79140**	0.66178
b-left	0.33287	**0.87643**	0.72996	0.76536
b-right	0.73203	**0.95939**	0.87227	0.93579
c-right	0.21333	0.94407	**0.96630**	0.94227
c-left	0.21001	0.79499	**0.80930**	0.78846
d-right	0.20432	0.50775	0.70270	**0.73441**
e-left	0.76307	0.58280	**0.94335**	0.79540
e-right	0.14928	0.23026	**0.73842**	0.69277
average	0.39165	0.69351	**0.80754**	0.78564

表 1　正答率

4.2 評価方法

提案手法のゲイズ識別器の性能評価を，4.1 節のデータセットを用いて行う．評価指標には，以下の式で定義される正答率 (Accuracy)，適合率 (Precision)，再現率 (Recall)，F 値 (F-measure) を用いる．

検証方法には，交差検証を採用した．例えば，a-left と a-right のテストには，b, c, d, e の動画を訓練データとして用いる．このようなテストを b, c, d, e に対しても行い，正答率，適合率，再現率の平均値を求める．また，平均適合率と平均再現率から平均の F 値を算出する．各検証で，訓練データをさらにランダムに 8 割と 2 割に分割し，8 割を訓練に，2 割をバリデーションデータとして用いる．

4.3 実験結果

各手法ごとの正答率の比較は表 1 に，適合率の比較は表 2 に，再現率の比較は表 3 に，F 値の比較は表 4 に示した通りとなった．

正答率と F 値の比較では，提案手法 2，提案手法 3，提案手法 1，Baseline algorithm の順に精度が高いという結果になった．再現率の比較では，提案手法 1，提案手法 3，提案手法 2，Baseline algorithm の順に精度が高いという結果になった．適合率の比較では，提案手法 2，提案手法

	Baseline	提案手法 1	提案手法 2	提案手法 3
a-left	0.40253	**0.93424**	0.71705	0.73115
a-right	0.50802	0.59113	**0.89966**	0.69343
b-left	0.31875	**0.93427**	0.88889	0.92419
b-right	0.56842	**0.91855**	0.69146	0.87407
c-right	0.21213	0.81659	0.90407	**0.96336**
c-left	0.20575	0.40000	**0.93333**	0.32692
d-right	0.72527	0.86822	0.88280	**0.95930**
e-left	0.33045	0.30676	**0.81683**	0.47326
e-right	0.12448	**0.13416**	0.09266	0.11448
average	0.29794	0.53699	**0.77317**	0.71011

表 2　適合率

	Baseline	提案手法 1	提案手法 2	提案手法 3
a-left	**1.0000**	0.39983	0.49089	0.66639
a-right	0.84434	**0.93964**	0.67789	0.63659
b-left	**0.97585**	0.65532	0.16685	0.28101
b-right	0.06879	0.93376	**0.95924**	0.89299
c-right	**1.00000**	0.94924	0.94078	0.73635
c-left	**1.00000**	0.00350	0.07343	0.02972
d-right	0.02935	0.46287	**0.73010**	0.70209
e-left	0.27617	**1.00000**	0.89350	0.95848
e-right	**1.00000**	0.98347	0.13223	0.22865
average	0.59328	**0.65936**	0.60724	0.61553

表 3　再現率

	Baseline	提案手法 1	提案手法 2	提案手法 3
a-left	0.57401	0.56000	0.58280	**0.69727**
a-right	0.63436	0.72571	**0.77319**	0.66379
b-left	0.48054	**0.77032**	0.28096	0.43098
b-right	0.12273	**0.92609**	0.80363	0.88343
c-right	0.35001	0.87794	**0.92206**	0.84740
c-left	**0.34128**	0.00694	0.13615	0.05449
d-right	0.05642	0.60383	0.79922	**0.81078**
e-left	0.30088	0.46949	**0.85345**	0.63365
e-right	0.22140	**0.23611**	0.10897	0.15257
average	0.39667	0.59192	**0.68023**	0.65945

表 4　F-measure

3，提案手法 1，Baseline algorithm の順に精度が高いという結果になった．

4.4　考察

まず，Baseline algorithm の検証結果から，平均正答率が 0.39 と低いため，視線方向と位置の情報のみからはゲイズ識別できないということがわかる．再現率が 1 と高い数値を出しているデータもあるが，これは識別結果のほとんどがゲイズありに偏っているためであり，学習がうまく行えていない．実際，そのようなデータでは，適合率が 0.40，0.21，0.20，0.12 と低く，平均適合率も 0.29 と低い．この要因としては，視線方向の特徴空間でのクラス分けを行うことが難しいということが挙げられる．図 4 は，視線方向

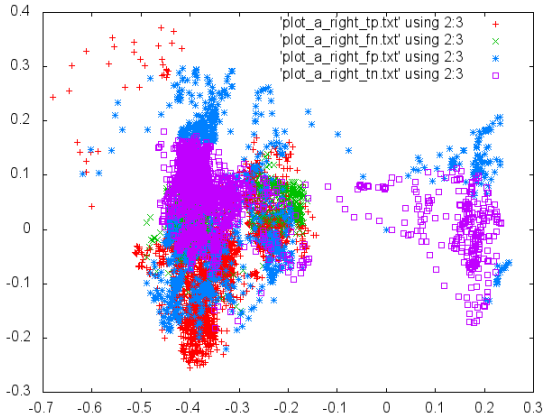

図 4　視線方向と識別結果の関係．視線方向 (x, y) に対する分類結果の分布を，赤：TP，青：FP，緑：FN，紫：TN で表す．Positive と Negative がきれいに分離していないため識別が困難であることがわかる

のみを用いた分類器（Baseline algorithm の入力から位置を除いたもの）の分類結果を表したものであり，識別する人物の右目の視線方向ベクトルの x 成分を横軸に，y 成分を縦軸にしている．学習には a-right を用いており，同じく a-right を評価に用いたときの TP，FP，FN，TN に当たる点をそれぞれ赤，青，緑，紫でプロットしている．この推定結果は，正答率 5 割程度であった．この数値は，訓練と評価のデータが全く同じ学習の正答率としてはかなり低いものである．この図では，x の値が 0 から 0.3 の点の集合はカメラの正面を向いている点の集合であり，x が負の値である点の集合は，フレーム中で視線が左を向いている点の集合であるが，x が負である点の大部分で真値がゲイズありの赤・緑とゲイズなしの青・紫とが左側で重なってしまっている．z 成分については，ほとんどの点が 0.9 付近であったため，これも分離には寄与しない．このことから，視線方向のみの情報からゲイズありとゲイズなしを分離することは不可能であることがわかる．位置の情報は，それ自体はゲイズあり・なしには関係せず，視線方向に制約を与える形で用いられるため，図 4 の試行のように同一動画から作られたデータで位置がほとんど固定の場合に視線方向から識別できなければ，位置を入れたとしても結果は変わらないということが考えられるため，Baseline algorithm ではゲイズ識別はできないという結論が得られる．

次に，目画像を入れた場合（提案手法 1）の検証結果では，Baseline algorithm と比べて平均正答率が 0.3 上がり，チャンスレートである 0.5 を超えた．平均適合率・平均再現率ともに 0.5 を超えているため，学習が Baseline algorithm と比べてうまく回っていることがわかる．この要因としては，目画像を入れることによって，目を閉じている場合をゲイズなしに分類できるようになったことが挙げられる．図 5 は右側の人物のゲイズ識別結果が，Baseline algorithm では FP に，提案手法 1 では TN になったフレームである．

図 5　Baseline algorithm で識別に失敗したフレーム

これらのフレームでは目を閉じているため，ゲイズなしが正解であるが，視線方向推定のベクトルは瞼を突き抜けて出てしまっており，その方向が相手の人物を向いているために Baseline algorithm ではゲイズありに分類されたと考えられる．目画像の情報を入れることで，目が閉じている場合を学習することができ，このようなフレームをゲイズなしに分類することで精度が改善したと考えられる．

目画像の時系列化（提案手法 2）の検証結果からは，時系列化の効果が確認できる．目画像の三次元畳み込みによってゲイズの時系列的な特徴が学習できるようになったために精度が向上したと考えられる．

提案手法 3 と提案手法 2 の比較からは，双方向の視線方向と目画像を入力にしたことが精度向上に寄与しないことが確認された．これは，ほとんどのデータで，片方の人物は相手を見ているがもう一方は相手を見ていないという状況が多いことが原因であると考えられる．そのため，片側のゲイズの推定で，相手の視線に関する情報が邪魔になっている可能性がある．また，過学習の傾向もあるため，同調のような複雑なインタラクションを学習するためにはデータが少なすぎたことも原因であると考えられる．

4.5　まとめ，今後の課題

本研究では，従来のアイコンタクト識別が一人称視点映像を用いていたことに対して，三人称視点映像からアイコンタクト識別を行うという新しいタスクを定義し，それを解くための方法として，視線方向・目の位置・目画像を入力とする CNN による識別方法（提案手法 1）を提案した．また，この手法を改良し，目画像を時系列データとして用いる手法（提案手法 2）と，識別したい人物だけでなく相手の視線に関する情報も入力に加える手法（提案手法 3）を提案した．そして，これらの手法と，視線方向と目の位置のみを用いるベースライン手法を実装してゲイズ識別器を作成し，それらの性能を F 値で評価したところ，提案手法 2，提案手法 3，提案手法 1，ベースライン手法の順に識別性能が高いことを確認し，提案手法の有用性を確認した．

本研究の提案手法の問題点としては，OpenFace による顔ランドマーク推定ができないフレームに対してはゲイズ識別ができないという点が挙げられる．システムの実際の運用を考えると，顔が全く見えていないフレームに対しても，厳密なゲイズ識別でないにしろ，顔が相手の方を向いているというアイコンタクトに関する手がかりが得られたほうが有用な場面が多く，コミュニケーションを捉える上では重要であると考えられる．そのため，相手の顔を見ているという新しいクラスを設けたほうがよいと思われる．また，本研究で用いたデータセットは，二人の人物がカメラに対して向かい合って座っている動画が大半で，実際の介護現場に近いとは言えず，三人以上が映っている映像に対して本研究の手法をそのまま適用することは難しい．さらに，OpenFace の顔ランドマーク点推定の訓練データは，介護場面で必要な高齢者のデータが少ないために高齢者の顔画像に対して使いづらいという問題点もある．そのため，この手法をすぐに実環境に適用することは難しいといえる．このような実際の介護環境を見据えた改良を行っていくことが今後必要になると考えられる．

謝辞　本研究は科研費 17H01779, 26249029, 15H02738, および, JST CREST, JPMJCR17A5 の支援を受けている．

参考文献

[1] for Health, N. I. and Britain), C. E. G.: *Dementia: Supporting People with Dementia and Their Carers in Health and Social Care: Quick Reference Guide* (2006).
[2] イヴジネスト，ロゼットマレスコッティ，ジェロームペリシエ，本田美和子，辻谷真一郎：Humanitude(ユマニチュード)「老いと介護の画期的な書」，株式会社トライアリスト東京 (2014).
[3] Jones, W. and Klin, A.: Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism, *Nature*, Vol. 504, No. 7480, p. 427 (2013).
[4] 沖野祐介，中澤篤志，本田美和子，石川翔吾，竹林洋一，西田豊明：頭部装着型カメラを用いた介護スキル評価 (医用画像)，電子情報通信学会技術研究報告= IEICE technical report: 信学技報，Vol. 116, No. 39, pp. 95–100 (2016).
[5] Ye, Z., Li, Y., Liu, Y., Bridges, C., Rozga, A. and Rehg, J. M.: Detecting bids for eye contact using a wearable camera, *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1, IEEE, pp. 1–8 (2015).
[6] Ranjan, R., Patel, V. M. and Chellappa, R.: Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 1, pp. 121–135 (2019).
[7] Xiong, X. and De la Torre, F.: Supervised descent method and its applications to face alignment, *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, pp. 532–539 (2013).
[8] Ye, Z., Li, Y., Fathi, A., Han, Y., Rozga, A., Abowd, G. D. and Rehg, J. M.: Detecting eye contact using wearable eye-tracking glasses, *Proceedings of the 2012 ACM conference on ubiquitous computing*, ACM, pp. 699–704 (2012).
[9] Mitsuzumi, Y., Nakazawa, A. and Nishida, T.: DEEP eye contact detector: Robust eye contact bid detection using convolutional neural network, *British Machine Vision Conference 2017 (BMVC 2017)* (2017).

[10] Zhang, X., Sugano, Y. and Bulling, A.: Everyday eye contact detection using unsupervised gaze target discovery, *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, ACM, pp. 193–203 (2017).
[11] Baltrušaitis, T., Robinson, P. and Morency, L.-P.: Openface: an open source facial behavior analysis toolkit, *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, IEEE, pp. 1–10 (2016).
[12] Zhang, X., Sugano, Y., Fritz, M. and Bulling, A.: It' s written all over your face: Full-face appearance-based gaze estimation, *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, pp. 2299–2308 (2017).
[13] Ankerst, M., Breunig, M. M., Kriegel, H.-P. and Sander, J.: OPTICS: ordering points to identify the clustering structure, *ACM Sigmod record*, Vol. 28, No. 2, ACM, pp. 49–60 (1999).
[14] Recasens, A., Khosla, A., Vondrick, C. and Torralba, A.: Where are they looking?, *Advances in Neural Information Processing Systems*, pp. 199–207 (2015).
[15] Chong, E., Ruiz, N., Wang, Y., Zhang, Y., Rozga, A. and Rehg, J. M.: Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency, *European Conference on Computer Vision*, Springer, pp. 397–412 (2018).
[16] Baltrusaitis, T., Robinson, P. and Morency, L.-P.: Constrained local neural fields for robust facial landmark detection in the wild, *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 354–361 (2013).