Vol.2017-CVIM-207 No.11
2017/5/10

CNNを用いた隠蔽に頑健なアイコンタクト検出

三鼓 悠[1,a)]　中澤 篤志[1]　西田 豊明[1]

概要：本論文は顔画像に隠蔽などがあった場合でも頑健にアイコンタクトを検出する手法を提案する．アイコンタクトは人同士のコミュニケーション解析において重要であり，一人称視点映像からアイコンタクトを検出する試みがあるが，隠蔽による顔検出の失敗に弱いという問題があった．本研究では，目領域をHoG と SVM を用いて検出し，その領域を畳み込みニューラルネットワーク（CNN）で学習することでアイコンタクト識別を行う．実験では顔に何も着用していない映像とマスクを着用した映像に対して手法を適用した．その結果，前者については本手法は従来法と同程度であり，後者では従来法よりも頑健に検出できた．

キーワード：アイコンタクト，一人称視点映像，畳み込みニューラルネットワーク

Robust eye contact detection for occlusion using CNN

Yu Mitsuzumi[1,a)]　Atsushi Nakazawa[1]　Toyoaki Nishida[1]

Abstract: We propose a method to detect eye contacts robustly even when the face image contains occlusions. Eye contact is important in human-to-human communication analysis, and there are some attempts to detect eye contacts from first person view video, but there was a problem that it is vulnerable to face detection failure due to occlusion. In this study, we detect eye region using HoG and SVM and identify eye contacts by convolutional neural network (CNN). In the experiment, we apply our method to the image wearing nothing on the face and the mask worn on the face. As a result, for the former, this method was comparable to the conventional method, and in the latter it was more robust than the conventional method.

Keywords: Eye contact, First person view video, Convolutional Neural Network

1. はじめに

アイコンタクトとは視線を合わせることであり，非言語的なコミュニケーションの 1 つである．アイコンタクトを検出することは健常者のコミュニケーションの解析に有益である．

自閉症スペクトラム（ASD）の人はアイコンタクトが困難な割合が多い．また乳幼児の視線情報には ASD の早期発見に有用であるという報告がある [1] ことから，ASD の早期発見のためにアイコンタクトの成否を利用することもある．

また，認知症ケアスキルの 1 つとしてもアイコンタクトは重要視されており [2]，介護者の一人称視点映像からアイコンタクトを評価する研究 [3] が行なわれている．

従来法では図 1（a）のように，まず顔全体の検出をし，顔方向を推定，次に顔パーツ点の検出を行いアイコンタクトを検出する．しかし画像中に顔の一部分しか写っていない場合には顔検出が行えないためアイコンタクト検出に失敗する．

本研究ではアイコンタクト検出の手法を図 1 のようにする．すなわち顔全体の検出器によらず顔の一部分の検出器を用いることで頑健にアイコンタクトを検出する手法を提案する。具体的には画像中の目領域のみからアイコンタクトを検出する。ここでは目領域の画像について畳み込みニューラルネットワーク（CNN）を用いて認識することで従来よりも大まかな目領域検出でもアイコンタクトの検出

[1] 京都大学，〒 606-8501 京都府京都市左京区吉田本町
[a)] mitsuzumi@ii.ist.i.kyoto-u.ac.jp

を可能にした．

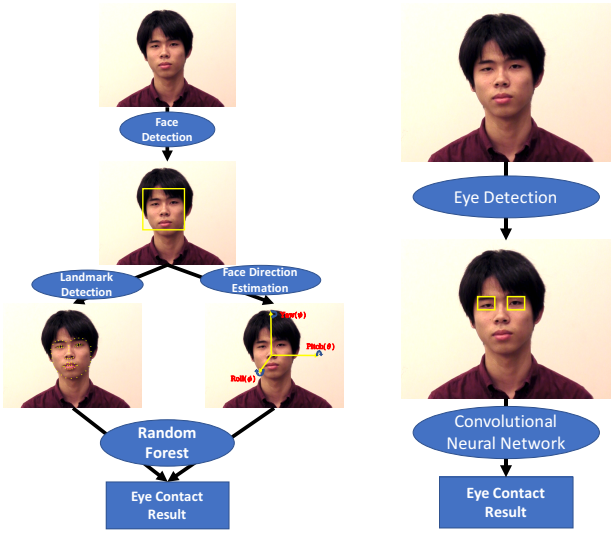

図 1　従来法 [3][4] と提案法の違い．従来法はまず顔全体を検出し，目領域と顔方向からアイコンタクトを認識する．提案法では，目領域のみから CNN によりアイコンタクトを認識するので隠蔽に頑健である．

2. 関連研究

アイコンタクトを検出する研究として Ye らの研究 [4] がある．Ye らは幼児の顔画像を顔方向で 3 つにクラスタリングし，それぞれのクラスタごとの RandomForest 分類器に目画像を入力することで識別を行う．介護シーンに対してアイコンタクトの検出を行った研究として沖野らの研究 [3] があるが，これは顔方向特徴と目画像特徴を合わせて入力特徴とし RandomForest 分類器により識別を行う手法である．顔方向を用いていない手法としては Smith らの手法 [5] が挙げられる．ここでは顔全体から両目の縁を検出することで目画像を得，特徴抽出し，主成分分析と多変量解析により次元圧縮したのちに SVM で識別を行う．アイコンタクト検出手法ではないが顔画像からの視線推定を行う手法として Zhang らの手法 [6] の手法がある．これは片目画像と正規化した顔方向を CNN に入力することで視線方向を推定する手法である．

これらの手法は顔方向を推定する，或いは顔全体から目のパーツ検出を行うという方法を採っているため，顔の目以外の部分が隠蔽された場合について適用が困難になる．特に介護シーンについては介護従事者と被介護者の顔の距離が近いため，顔検出がうまくいかないフレームが一人称視点映像に多く含まれる．本研究では両目が対象画像に含まれていれば適用可能であるため，このような問題を解決するためには有効であると考える．

3. 提案法

3.1 両目検出

画像中から King らの手法 [7] により，両目画像領域を学習済みの SVM を用いて検出する．この SVM の学習には機械学習ライブラリ Dlib[8] を用いた．

次に検出された両目画像領域中にある目や眉といった目及びその周辺の顔パーツの輪郭を推定する．輪郭推定には Kazemi らの手法 [9] を用いる．輪郭推定器も同様に Dlib を用いて学習した．

3.2 アイコンタクト識別

3.2.1 特徴抽出

得られた目の輪郭より目画像を切り出す．切り出す領域は図 2 の緑枠のように，片目の外接矩形より上下左右に 10px ほど余白をとった領域である．これは目の輪郭推定にはノイズが多かったことを考慮したためである．この領域画像をグレースケール化，60×36 にリサイズしたものを左右それぞれ $\boldsymbol{i}^{\boldsymbol{Left\text{-}crop}}$，$\boldsymbol{i}^{\boldsymbol{Right\text{-}crop}}$ とする．

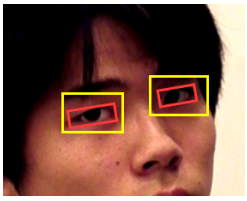

図 2　目画像特徴

3.2.2 識別器

識別器 $P(y_t|\boldsymbol{X_t})$ の値によりアイコンタクトの検出を行う．すなわち y_t の値は $\{0,1\}$ の 2 値であり $y_t = 0$ はアイコンタクトしていない，$y_t = 1$ はアイコンタクトしているとする．$P(y_t = 1|\boldsymbol{X_t})$ の値が閾値 τ を超えた場合にアイコンタクトをしていると識別する．

時刻 t において，$\boldsymbol{i}^{\boldsymbol{Right\text{-}crop}}$，$\boldsymbol{i}^{\boldsymbol{Left\text{-}crop}}$ を GCN（Global Contrast Normalization）により正規化し，特徴 $\boldsymbol{i}^{\boldsymbol{Left\text{-}crop\text{-}GCN}}{}_t$，$\boldsymbol{i}^{\boldsymbol{Right\text{-}crop\text{-}GCN}}{}_t$ とすると入力特徴は式 1 となる．

$$\boldsymbol{X_t} = (\boldsymbol{i}^{\boldsymbol{Left\text{-}crop\text{-}GCN}}{}_t, \boldsymbol{i}^{\boldsymbol{Right\text{-}crop\text{-}GCN}}{}_t) \qquad (1)$$

識別器は畳み込みニューラルネットワークを学習することにより得る．ネットワーク構造は図 3 の左に示す．左右の目画像は別々の畳み込み層で処理される．並列な畳み込み層は同じ構造であり，畳み込み層を 2 つ，プーリング層を 1 つ連結したものを，2 つ直列に連結し 512 ユニットの全結合層へと連結する．すなわち全結合層のユニット数は 1024 個である．プーリング層では Max pooling を用い，出力される画像は元の半分となる．活性化関数には出力層以外では Leaky ReLU 関数を，出力層では Softmax 関数を

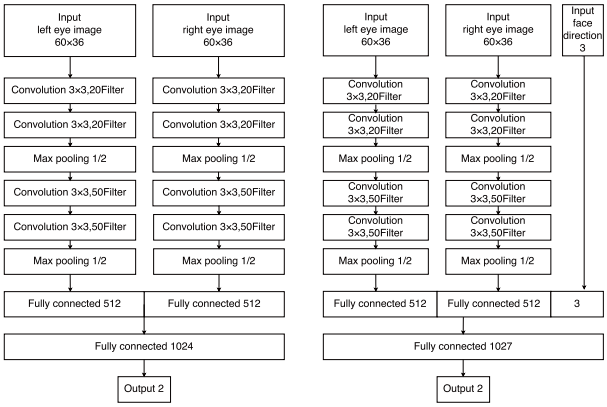

図 3　左:提案法のネットワーク，右:提案法に顔方向入力を加えたネットワーク

図 4　マスク非着用時、マスク着用時の画像

利用する．全結合層の間には確率 50%のドロップアウト層を用意した．

4. 評価実験

4.1 データセット

撮影方法　被験者は椅子に座り，そこから 1.5m 程度離れた位置にカメラを設置して撮影した．1 フレームのサイズが 1920 × 1080[pixel] で，人物の顔の大きさはおおよそ 350 × 350[pixel] ほどになる．1 秒あたり 30 フレームの間隔で撮影し，1 つの動画につき 1 分間撮影した．

学習データ　被験者 1 人につき，何も顔に着用せずに，カメラに対して視線を送り続ける映像とカメラから視線を外し続ける映像の 2 種類を合計 5 人の人物から撮影した．

評価データ　被験者 1 人につき，図 4 のように何も顔に着用しない，マスクのみ着用した 2 パターンについてカメラに対して視線を送り続ける映像と視線を外し続ける映像の 2 種，計 4 動画を合計 2 人の人物から撮影した．

4.2 手法比較実験

4.2.1 比較対象

実験では以下の 4 手法を比較した．図 5 に比較対象の入力特徴，識別器をまとめた．

従来法　顔検出を行い，Baltruaitis らの手法 [10] により顔方向と顔ランドマーク点を推定，顔方向特徴 $\boldsymbol{r1_t}$ と図 2 の赤枠の目画像特徴を白色化したもの $\boldsymbol{i^{Left\text{-}red\text{-}whiten}}_t, \boldsymbol{i^{Right\text{-}red\text{-}whiten}}_t$ を入力として RandomForest によりアイコンタクト識別を行う．ただし赤枠には片目を含む画像のうち最も小さい長方形であるという制約がある．識別器の入力特徴は式のようになる．

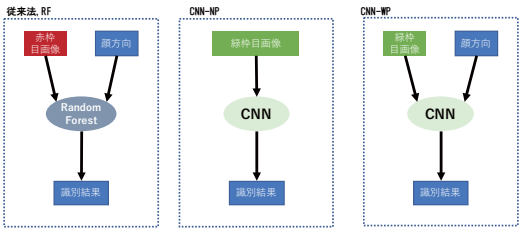

図 5　比較対象

$$\boldsymbol{X_t^{Conventional}} = \begin{pmatrix} \boldsymbol{i^{Left\text{-}red\text{-}whiten}}_t \\ \boldsymbol{i^{Right\text{-}red\text{-}whiten}}_t \\ \boldsymbol{r_t} \end{pmatrix}^{\mathrm{T}} \tag{2}$$

RF　3.1 節により両目領域を検出，目と眉の輪郭を推定する．得られた目と眉の輪郭より図 2 の赤枠の目画像特徴を白色化したもの $\boldsymbol{i^{Left\text{-}red\text{-}whiten'}}_t, \boldsymbol{i^{Right\text{-}red\text{-}whiten'}}_t$ と顔方向特徴 $\boldsymbol{r2_t}$ を抽出し，RandomForest によりアイコンタクト識別を行う．なお顔方向特徴 $\boldsymbol{r'_t}$ の推定は輪郭との関係を RandomForest 回帰により学習し関数を得ることにより行う．識別器の入力特徴は式のようになる．

$$\boldsymbol{X_t^{RF}} = \begin{pmatrix} \boldsymbol{i^{Left\text{-}red\text{-}whiten'}}_t \\ \boldsymbol{i^{Right\text{-}red\text{-}whiten'}}_t \\ \boldsymbol{r'_t} \end{pmatrix}^{\mathrm{T}} \tag{3}$$

CNN-NP　本研究の手法．識別器の入力特徴は式のようになる．

$$\boldsymbol{X_t^{CNN\text{-}NP}} = \boldsymbol{X_t} \tag{4}$$

CNN-WP　本研究の入力特徴に顔方向特徴 $\boldsymbol{r'_t}$ を加えた手法．用いるネットワークは図 3 の右側となる．なお顔方向特徴の抽出方法は手法 RF と同じである．識別器の入力特徴は式のようになる．

$$\boldsymbol{X_t^{CNN\text{-}WP}} = \begin{pmatrix} \boldsymbol{i^{Left\text{-}crop\text{-}GCN}}_t \\ \boldsymbol{i^{Right\text{-}crop\text{-}GCN}}_t \\ \boldsymbol{r'_t} \end{pmatrix}^{\mathrm{T}} \tag{5}$$

4.2.2 評価方法

評価は各手法によって得られたあるフレームのアイコンタクトの尤度 $P(y|\boldsymbol{X_t^j})$ と閾値 τ の間に

$$P(y=1|\boldsymbol{X_t^j}) > \tau \tag{6}$$
$$j \in \{Conventional, RF, CNN\text{-}NP, CNN\text{-}WP\}$$

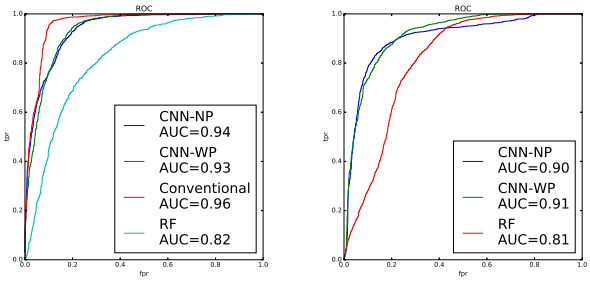

図 6 左:マスク非着用時の人物 A の結果，右:マスク着用時の人物 A の結果

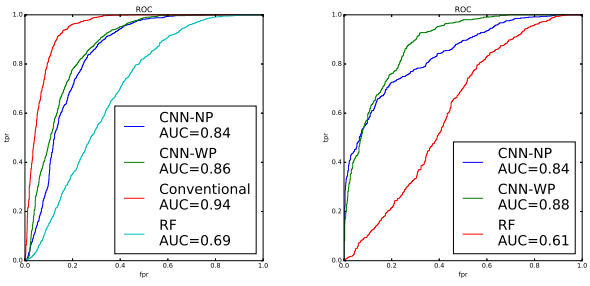

図 7 左:マスク非着用時の人物 B の結果，右:マスク着用時の人物 B の結果

が成立している場合にそのフレームでアイコンタクトが成立している，成立していないとした場合にそのフレームではアイコンタクトをしていないとし，閾値 τ を変化させることで ROC 曲線を求め評価する．ROC 曲線の評価については曲線下の面積 AUC（Area Under the Curve）を算出することによって行う．また閾値 $\tau = 0.5$ とした場合の再現率，適合率，F 値も算出し評価方法とする．

4.2.3 実験結果，考察

人物 A についての実験結果は図 6，表 1，人物 B についての実験結果は図 7，表 2 である．

まず RF，CNN-NP，CNN-WP の手法について評価すると CNN-NP，CNN-WP はほぼ同程度の精度で最もよく，それらに比較すると RF は精度が良くなかった．

RF は従来手法とは目検出に Dlib を用いるところが異なる手法だが，精度は従来法より低くなった．主な原因として目領域検出及びパーツ推定の精度があげられる．RF では両目領域検出し，その限られた矩形内で輪郭推定を行うため精度が下がる．一方で，従来手法では Baltruaitis らの手法 [10] を用いて目の検出を行っており，より正確な輪郭の推定が可能である．用いる目画像特徴が，目を含む長方形の中で最も小さいものであることにより，輪郭の推定に誤差が生じることで目画像の一部が抽出できない場合があり，実験結果のようになった原因と考える．

CNN-NP,CNN-WP と RF について比較すると手法として異なる点は用いた目画像特徴と識別器が異なる．目画像については CNN-NP と CNN-WP は目の外接矩形にある程度余白を設けて抽出しているが，RF は目を含むもののうちもっとも面積の小さい長方形から抽出している．そのため前者ではパーツ検出の精度に強いが，後者では誤差により目画像特徴の抽出に失敗するという点がある．また前者では CNN を識別器の学習に用いたため，畳み込み層で入力目画像の位置ずれに対してもより頑健であったことが考えられる．

CNN-NP と CNN-WP とを比較すると，異なる点はネットワークの全結合層での顔方向特徴の有無であるが，あまり差がないことから顔方向を推定することが必要ではないことが考えられる．従来手法では顔方向を必要としていたがこれは目の画像特徴の正規化により画像からは顔方向の情報がなくなっていることから生じているが，この 2 手法利用した目の画像特徴の事前処理はリサイズのみであるため両目画像の組には顔方向の情報が残ったと推測する．

従来法と提案法の比較だが，マスクなしの場合では提案法は従来手法ほどの精度にはならなかった．原因としてはランドマーク点推定に誤差が生じることにより切り出す矩形内での目の位置がずれ，そのずれが大きくなるとアイコンタクトに必要な特徴抽出が畳み込み層でうまく行うことができなくなると考える．マスクありの場合について，従来手法では目の検出ができないことからアイコンタクト検出ができないが，そのような映像についても提案法では目

	マスク非着用時			
	適合率	再現率	F 値	目検出率
従来法	**0.9028**	**0.9581**	**0.9296**	3627 / 3627
RF	0.7903	0.6367	0.7052	3625 / 3627
CNN-NP	0.8279	0.9139	0.8688	3625 / 3627
CNN-WP	0.8669	0.8233	0.8445	3625 / 3627
	マスク着用時			
	適合率	再現率	F 値	目検出率
従来法	-	-	-	11 / 3618
RF	0.7489	0.7304	0.7395	3596 / 3618
CNN-NP	0.8077	**0.8881**	**0.8460**	3596 / 3618
CNN-WP	**0.8388**	0.8384	0.8386	3596 / 3618

表 1 実験結果（人物 A）

	マスク非着用時			
	適合率	再現率	F 値	目検出率
従来法	**0.8821**	**0.8567**	**0.8692**	3630 / 3630
RF	0.6539	0.3460	0.4525	3088 / 3630
CNN-NP	0.8081	0.5928	0.6839	3088 / 3630
CNN-WP	0.8341	0.4665	0.5984	3088 / 3630
	マスク着用時			
	適合率	再現率	F 値	目検出率
従来法	-	-	-	2 / 3623
RF	0.6141	0.3096	0.4117	1639 / 3623
CNN-NP	0.8854	**0.5816**	**0.7020**	1639 / 3623
CNN-WP	**0.9089**	0.4697	0.6193	1639 / 3623

表 2 実験結果（人物 B）

の検出が可能であり，アイコンタクトの検出もマスクなしと同程度にできた．

4.3 人の識別能力との比較実験

提案法では左右の目画像のみを与えてアイコンタクト識別を行ったが，同様の画像のペアを人が識別した場合について識別器と比較した．

4.3.1 実験設定

以下のような設定で実験を行った

- アンケートをとることで人の識別結果を得た．
- アンケートでは提案法の識別器に与える左右の目画像のペアについてアイコンタクトをしているかどうかを答える質問を 50 問用意した．
- アンケートは 13 人から回答を得た．
- アンケートで出題した目画像に提案法を適用することで識別器の結果を得た．

4.3.2 実験結果，考察

実験結果は図 8 のようになった．正答率では人の識別能力よりも実装した識別器の能力の方が高いという結果になった．人は通常顔全体から顔方向を認識し，黒目の位置と合わせてアイコンタクトを認識しているため，目画像のみに限られた場合にはアイコンタクトの識別が通常よりも難しいものになったと考えられる．しかし目画像のペアにも顔方向の情報は残っているため，識別器はアイコンタクト識別ができた．また人はアイコンタクトを時系列で捉えているが，このアンケートでは時系列要素が排除されているため通常のアイコンタクトの識別よりも困難になった点も識別器より劣る結果となった要因ではないかと考える．

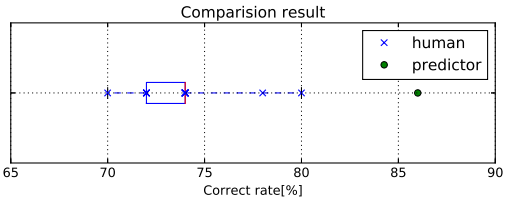

図 8 識別器，人の識別の正答率

5. 結論，今後の課題

本研究では従来法では目以外の顔パーツの隠蔽によってアイコンタクト検出ができない映像について，目領域を検出し領域画像を利用することでアイコンタクト検出を行うシステムを提案した．両目検出器を用いて両目を含む領域を検出し，パーツ推定をして左右の目領域を抽出，CNN を用いて識別する方法を実装した．

実験では従来法と提案法のアイコンタクト検出精度の比較，及び従来法では困難な映像について提案法の適応が可能かの検証を行った．精度比較においては従来手法が可能な映像について提案法では同程度か少し低いという結果となったが，従来手法の適用可能外の映像についても提案法の適用が可能であり，本研究の有効性を確認した．

本研究では実験環境での検証を行ったが，実環境での提案法の検証は今後の課題となる．

参考文献

[1] Jones, W. and Klin, A.: NIH Public Access, Vol. 504, No. 7480, pp. 427–431 (online), DOI: 10.1038/nature12715.Attention (2014).
[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] 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).
[4] 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 2015*, (online), DOI: 10.1109/FG.2015.7163095 (2015).
[5] Smith, B. A., Yin, Q., Feiner, S. K. and Nayar, S. K.: Gaze locking: passive eye contact detection for human-object interaction, *Proceedings of the 26th annual ACM symposium on User interface software and technology*, ACM, pp. 271–280 (2013).
[6] Zhang, X., Sugano, Y., Fritz, M. and Bulling, A.: Appearance-based gaze estimation in the wild, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 07-12-June, pp. 4511–4520 (online), DOI: 10.1109/CVPR.2015.7299081 (2015).
[7] King, D. E.: Max-Margin Object Detection, (online), available from ⟨http://arxiv.org/abs/1502.00046⟩ (2015).
[8] King, D. E.: Dlib-ml: A Machine Learning Toolkit, *Journal of Machine Learning Research*, Vol. 10, pp. 1755–1758 (2009).
[9] Kazemi, V. and Sullivan, J.: One millisecond face alignment with an ensemble of regression trees, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874 (online), DOI: 10.1109/CVPR.2014.241 (2014).
[10] Baltrušaitis, 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*, pp. 354–361 (online), DOI: 10.1109/ICCVW.2013.54 (2013).