時空間骨格情報を用いた遠方歩行者のアイコンタクト検出

畑 隆聖[1,a)]　出口 大輔[1]　平山 高嗣[1]　川西 康友[2,1]　村瀬 洋[1]

概要

車両の運転において，歩行者からのアイコンタクトは自車への気付きを判断する重要な要素である．従来のアイコンタクト検出手法の多くは眼球計測に基づく直接的な視線推定に依存しており，道路環境のように車両と歩行者の距離が離れるような場合は視線推定が困難である．一方で我々人間は，視線が不明確な歩行者に対しても，顔向きや姿勢，振り向き等，歩行者の全身の情報や動きの情報から，こちらを見ているか判断している．本研究ではこの知見に倣い，顔向きと体の姿勢の関係性や，それらの時間変化を捉えることにより，眼球計測が困難な遠方歩行者に対してもアイコンタクトの有無を検出可能な手法を提案する．車載カメラ画像を用いた実験により，その有効性を示した．

1. はじめに

車両の運転において，歩行者が自車両の存在に気づいているかの判断は危険予測の観点から重要である．その気づきを判断するための重要な要素として，歩行者がこちらを見ているかどうか，すなわちアイコンタクトの有無がある．

これまでに，アイコンタクト検出と関連が深いタスクである視線推定に関して多くの研究がなされてきた．視線推定に関する従来研究の多くは，人物の眼球を近くで撮影した顔画像を入力とし，眼球計測や目の周りの外観に基づき直接的に視線を推定している [1,2] ．そのため，対象人物に対してカメラを非常に近い位置に配置する必要があり，道路環境のように車両と歩行者の距離が離れるような状況では，これらの手法を用いての視線推定は不可能である．よって，遠方で解像度が低くなる歩行者のアイコンタクト検出のためには，視線推定に依らない手法が必要である．

一方で我々が普段車両を運転する様子を振り返ると，視線が不明確な歩行者に対しても顔向きや姿勢の時間変化を加味してアイコンタクトの有無を判断していることに気づく．この知見に倣うことで，直接的な視線推定が困難な遠方歩行者に関しても，アイコンタクト検出が可能になるのではないかと考えた．例えば，図 1 の赤枠で示した時刻における歩行者のアイコンタクトの有無を考えてみよう．遠方で視線推定が困難な歩行者のアイコンタクトを判断する際，顔向きは重要な判断材料である．しかし，図 1 の (i) と (ii) のように顔向きが似ている場合でもアイコンタクトの有無は異なる可能性がある．ここで体の姿勢を考慮すると，図 1(i) の人物は顔と体が同方向であるのに対して，(ii) の人物は足から顔にかけて徐々にカメラ方向に姿勢が向く，すなわち「ねじれ」の状態であることがわかる．このことから，(i) の人物はこちらを見ておらず，(ii) の人物はこちらを見ていると推測できる．さらに，赤枠で示したフレーム単体だけでなく，それ以前のフレームも含めて見ると振り向きなどの動きが見て取れる．このような動きも考慮することで，赤枠の時刻でこちらを見ているかの判断がより明確になると期待できる．すなわち，顔向き，体の姿勢，それらの動きを複合的に考慮することによって，アイコンタクトの有無の高精度な判断が可能になると考える．

本研究では，顔向き，体の姿勢，それらの動き，を骨格の時系列的な変化（以下，骨格系列）として捉える．骨格系列は，人物の外観の違いに頑健であり，行動認識の研究等で広く用いられている．特に近年，関節点をノードとするグラフ構造として骨格系列を捉えるグラフ畳み込みが注目されている．代表的な手法の一つとして，ST-GCN [3] や MS-G3D [4] がある．これは，各関節座標の特徴だけでなく，関節間の空間的，時間的な接続関係も考慮するため，複雑な動作を認識できる．以上を踏まえ，本研究ではグラフ構造で表現された骨格系列を用いて顔と体の特徴的な時間変化を捉えるアイコンタクト検出手法を提案する．

2. 提案手法

2.1 概要

本稿では，複数フレームの骨格情報を利用することで歩行者のアイコンタクトを検出する手法を提案する．具体的には，各フレームの関節点をノードとする骨格系列グラフ畳み込みにより，歩行者の顔向き，体の姿勢，動きを考慮したアイコンタクト検出を実現する．

図 2 は，図 1 (ii) の歩行者画像系列を骨格系列で表したものである．青色は関節点を表し，赤色の実線が空間的な関節点間の接続関係，緑色の破線は同一関節点の時間的な

[1] 名古屋大学
[2] 理化学研究所 情報統合本部 GRP
[a)] hatar@vislab.is.i.nagoya-u.ac.jp

(i) アイコンタクト無　　(ii) アイコンタクト有

図 1　同様の顔向きの歩行者画像例

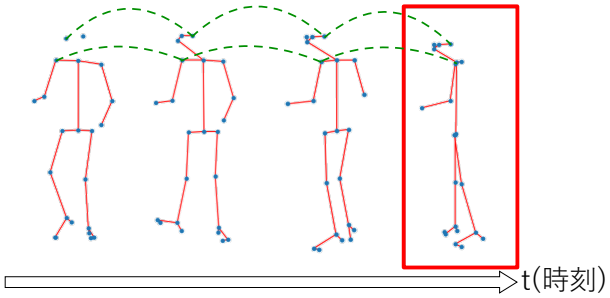

図 2　図 1 (ii) の骨格系列表現

対応を表す（図 2 では例として左耳および左肩の時間的対応関係を記載）．図 2 より，歩行者の顔向きや体の姿勢が関節点で表現されていることが見て取れる．このような歩行者の骨格系列の特徴を踏まえ，提案手法では複数フレームから抽出した関節点特徴量を用いることで，歩行者の顔向き，体の姿勢，動きを考慮してアイコンタクト検出を行う．

各関節点の特徴量はそれぞれ (x, y) 座標および推定信頼度 c の 3 次元特徴ベクトルで表す．しかし，単純にこれらの関節点の特徴量を用いるだけでは，関節点間の空間的な接続関係（図 2 の赤色実線）および，同一関節点の時間的な対応関係（図 2 の緑破線）を明示的に考慮することはできない．そこで提案手法では，行動認識の研究 [3,4] に倣い，これらの接続関係をグラフ構造を用いて表現する．フレーム毎の各関節点をノードとし，その空間的，時間的な接続関係を表す隣接行列を用いて骨格系列をグラフ表現する．この骨格系列に対してグラフ畳込みを行なうことによって，各関節点の特徴量の時空間的な関係を考慮した特徴量に変換する．これにより，単純に 3 次元の関節点を用いる場合に比べ，歩行者の顔向き，体の姿勢，動きを効果的に捉えるアイコンタクト検出を実現する．

2.2　骨格系列グラフ畳み込みの表現

T フレームから抽出した関節点の集合は

$$\mathbf{X} = \{\mathbf{x}_{t,n} \in \mathbb{R}^3 \mid t, n \in \mathbb{Z}, 1 \leq t \leq T, 1 \leq n \leq N\} \quad (1)$$

で与えられ，関節点の接続関係を 0，1 で表現した隣接行列 $\mathbf{A} \in \mathbb{R}^{(T\times N)\times(T\times N)}$ により骨格系列グラフを表現する．ここで，N は各フレームにおける人物骨格の関節点数である．関節点の集合 $\mathbf{X}$ および隣接行列 $\mathbf{A}$ の空間的な接続関係はフレーム単位での “骨格” の表現に対応し，$\mathbf{A}$ の時間的な接続関係を含めることで，“骨格系列” の表現となる．

T フレーム分の関節点をノードとして接続したグラフを $\mathbf{X}$，$\mathbf{A}$ を用いて表現し，Neural Network f_p により骨格系列に対してグラフ畳み込み（以下，骨格系列グラフ畳み込み）を行なう．骨格系列グラフ畳み込みは，重みパラメータ θ を用いて以下の式で表す．

$$\mathbf{x}_p = f_p(\mathbf{X}, \mathbf{A}, \theta) \quad (2)$$

これにより，関節間の空間的な接続関係やフレーム間での時間的な関係を考慮した特徴量 $\mathbf{x}_p$ に変換する．得られた $\mathbf{x}_p$ を用いて，T フレーム目についてアイコンタクト有/無の 2 クラス分類を行なう．

2.3　処理手順

図 3 に提案手法の処理手順を示す．

2.3.1　関節点特徴量の抽出

骨格推定器を用いて，各歩行者画像から N 個の関節点それぞれに対して関節点特徴量を取得する．ここでは，図 4 に示すような $N = 25$ 点の関節点を用いる．これによりある関節点 n の時刻 t での特徴量は，歩行者画像の左上を原点とするサブピクセル単位の x, y 座標，およびその推定の信頼度 c を要素とする 3 次元特徴ベクトル $(\widetilde{x}_{t,n}, \widetilde{y}_{t,n}, \widetilde{c}_{t,n})$ として得られる．$\widetilde{c}_{t,n}$ は $[0, 1]$ の範囲の実数値であり，$\widetilde{c}_{t,n} = 0$ の場合は関節点を検出できなかったものとみなして $\widetilde{x}_{t,n}$ と $\widetilde{y}_{t,n}$ の値はどちらも 0 とする（関節点の欠損）．そして，全ての関節点の特徴ベクトルを結合し，$\widetilde{\mathbf{p}}_t = (\widetilde{x}_{t,0}, \widetilde{y}_{t,0}, \widetilde{c}_{t,0}, ..., \widetilde{x}_{t,24}, \widetilde{y}_{t,24}, \widetilde{c}_{t,24})$ の 75 次元の特徴ベクトルとする．

時刻 t における L は，図 4 おける関節点 $5, 12$ を用いて，

$$L = \sqrt{(\widetilde{x}_{t,12} - \widetilde{x}_{t,5})^2 + (\widetilde{y}_{t,12} - \widetilde{y}_{t,5})^2} \quad (3)$$

により求め，関節点 n の特徴量を次式を用いて正規化する．次式により得られる正規化された特徴量の，T フレーム分の系列を $\mathbf{X}$ として用いる．

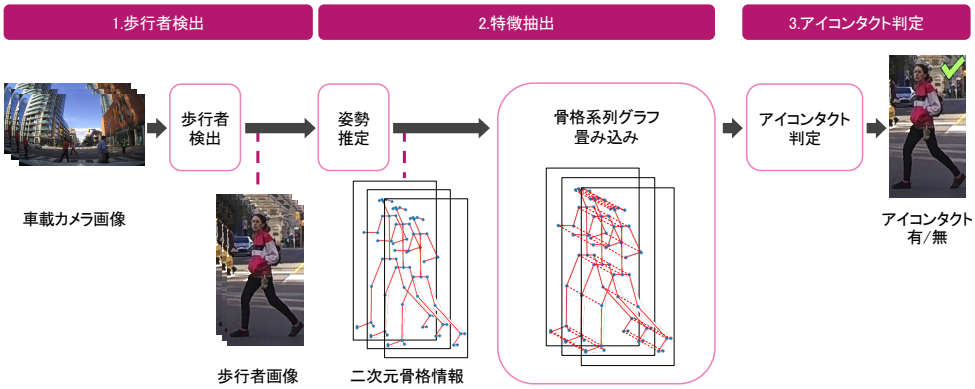

図 3　提案手法の処理手順

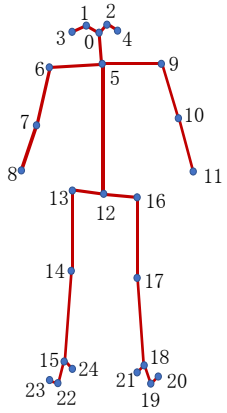

図 4　本研究で使用する関節点

$$\widehat{x}_{t,n} = \begin{cases} \frac{\widetilde{x}_{t,n} - \widetilde{x}_{t,12}}{L} & (\widetilde{c}_{t,n} > 0) \\ 0 & (\widetilde{c}_{t,n} = 0) \end{cases} \tag{4}$$

$$\widehat{y}_{t,n} = \begin{cases} \frac{\widetilde{y}_{t,n} - \widetilde{y}_{t,12}}{L} & (\widetilde{c}_{t,n} > 0) \\ 0 & (\widetilde{c}_{t,n} = 0) \end{cases} \tag{5}$$

$$\widehat{c}_{t,n} = \frac{\widetilde{c}_{t,n} - 0.5}{0.5} = 2\widetilde{c}_{t,n} - 1 \tag{6}$$

2.3.2　アイコンタクト判定

2.2 節で取得した $\mathbf{x}_p$ をアイコンタクト判定器に入力し，骨格系列の最後のフレームに対するアイコンタクトの有無を 2 クラス分類して出力する．アイコンタクト判定器は，単層の全結合型 Neural Network（以下，FCN）を用いる．

3.　実験

3.1　データセット

本実験では，車載カメラ画像として，Pedestrian Intention Estimation dataset（以後，PIE データセット）[7] を用いた．このデータセットは，157° の広角レンズを装備した車載カメラで撮影された，約 900,000 枚の画像系列からなる．PIE データセットには複数の歩行者が存在し，1,842 人の歩行者には，カメラを見ているか否か（アイコンタクトの有無），外接矩形，遮蔽具合などの情報が付与されている．

本実験では，歩行者検出処理を行う代わりとして，PIE データセットに付与されている矩形情報を歩行者検出結果として用いた．また，骨格推定器として OpenPose [6] の BODY_25 モデルを使用した．骨格系列グラフ畳み込みを行う Neural Network の f_p には，MS-G3D を用いた．また問題設定の簡単化のため，歩行者画像のうち以下の条件を満たす画像のみを評価の対象として用いた．

(1) 歩行者の遮蔽率が 25%以下
(2) 切り出した歩行者画像の縦サイズが 150 画素以上
(3) 切り出した歩行者画像の矩形が，元の車載カメラ画像の枠（$1,920 \times 1,080$）を超えない

さらに，アイコンタクト有/無のデータ数が同数となるように歩行者画像を抽出して用いた．データの内訳は，学習用 21,924 枚，評価用 2,126 枚，テスト用 12,546 枚である．

表 1　考慮する特徴量による正解率の比較 (5 回の平均正解率)

手法	モデル	使用する特徴量	1 frame	10 frames
比較 1 [8]	FCN	顔+体（胴体のみ）	85.5%	—
比較 2	FCN	顔+体（胴体のみ）	—	87.8%
提案 1	MS-G3D	顔のみ	85.6%	87.6%
提案 2	MS-G3D	顔+体	86.3%	87.6%
提案 3	MS-G3D	顔+体（胴体のみ）	86.5%	**88.6%**

3.2　顔，体，動きを用いること，骨格系列グラフ畳み込みを用いることの有効性の検証

本実験では，表 1 に示す 5 つの手法の比較を行なった．提案手法 1 は顔部分の 6 関節点（図 4 の 0，1，2，3，4，5）を用いた．一方，提案手法 2 は顔および体の 25 関節点を用いた．提案手法 3 および比較手法は，顔および体の胴体部（図 4 の 6，9，12，13，16）の 11 関節点のみを用いた．また 2 つの比較手法はグラフ表現を用いず，すべての関節点を 3 層の FCN に入力する手法である．各手法とも，損失関数はクロスエントロピー誤差，最適化関数は AdamW [5] を用いた．各手法に対し，単一フレームの骨格情報のみを用いる場合，10 フレーム分の骨格系列を用いる場合，それぞれアイコンタクト検出を行なった．ここで，比較手法 1 は，Belkada らの提案したアイコンタクト検出手法 [8] に対応しており，比較手法 2 はその時間方向への拡張である．本実験では，各試行をパラメータを変えて 5 回ずつ行ない，正解率の平均値を比較した．

実験結果を表 1 に示す．提案手法 3 で 10 フレーム入力した場合が最も正解率が高く，Belkada らの手法（単一フレーム FCN）に対して 3.1%向上することを確認した．

4.　考察

4.1　体の特徴について

3.2 節の実験結果について，表 1 に示したように，体の胴体部のみを加える提案手法 3 が，全身を用いた提案手法

(i) アイコンタクト有　　(ii) アイコンタクト有

図 5　体の特徴が胴体のみの場合に正しくアイコンタクト検出した例

図 6　グラフを用いた場合に正しくアイコンタクト検出した例

2 よりも高精度だった．この要因の一つとして，腕や脚の関節点位置は歩行に伴って変化が大きく，アイコンタクト検出する際のノイズになったのではないかと考えられる．提案手法 2 が誤判定したのに対し，提案手法 3 が正しくアイコンタクト有と判定できた例を図 5 に示す．これらの歩行者は，顔や胴体部の各関節点の位置は類似していることがわかる．一方で，腕や脚の各関節点の位置は歩行者間で大きく異なり，必ずしもアイコンタクトと相関しているとは限らない．故に，体のねじれの考慮には胴体のみで十分であり，腕や脚はノイズとなったのではないかと考える．

4.2　グラフについて

表 1 に示したように，グラフを用いる提案手法 3 の方がグラフを用いない比較手法よりも高精度であった．図 6 に，10 フレームの骨格系列を入力した場合において，FCN を用いた場合ではアイコンタクトの有無が誤判定され，MS-G3D では正しく判定された歩行者画像例（上半身のみ）を示す．この例は，振り向きにより顔の関節点の位置が変化している．グラフ構造では，時間方向に対して同一の関節点を結びつけて畳み込みを行なうため，フレーム間での各関節点の位置の変化量が明示的に考慮される．それにより提案手法では，より動きを捉えられているのではないかと考える．

5.　むすび

本研究では，骨格系列グラフ畳み込みによって，顔および体の特徴，動きを考慮したアイコンタクト検出手法を提案した．提案手法では，まず歩行者検出器により車載カメラ画像から歩行者画像を切り出し，骨格推定器によって関節点特徴量を取得する．そして，グラフ構造を表す隣接行列と関節点特徴量を合わせた骨格系列のグラフ畳み込みをすることで，時空間を考慮した特徴量を取得する．この特徴量を用いて，最終的なアイコンタクトの有無を判別する．

PIE データセットを用いた評価実験を行い，顔，体，動きを考慮することの有効性，ならびに骨格系列グラフ畳み込みを用いることの有効性を検証した．

実験の結果，アイコンタクト検出の正解率が，単一フレーム FCN の 85.5%に対して，提案手法は 88.6%と 3.1%向上したことを確認した．

今後の課題としては，アイコンタクトの有無の判定に特に重要な関節点の調査，歩行者周囲のコンテキストの考慮（他車両や他の歩行者の存在や位置関係，横断歩道などの道路環境），などが挙げられる．

謝辞 本研究の一部は JSPS 科研費 17H00745 による．

参考文献

[1] T. Baltrusaitis, A. Zadeh, Y.C. Lim, and L. Morency, “OpenFace 2.0: Facial Behavior Analysis Toolkit,” Proc. of the 13th IEEE International Conference on Automatic Face Gesture Recognition, pp.59–66, May. 2018.

[2] B.A. Smith, Q. Yin, S.K. Feiner, and S.K. Nayar, “Gaze Locking: Passive Eye Contact Detection for Human-Object Interaction,” Proc. the 26th Annual ACM Symposium on User Interface Software and Technology, pp.271–280, Oct. 2013.

[3] S. Yan, Y. Xiong, and D. Lin, “Spatial Temporal Graph Convolutional Networks for Skeleton based Action Recognition,” Proc. of the 32nd AAAI Conference on Artificial Intelligence, pp.7444-7452, Feb. 2018.

[4] Z. Liu, H. Zhang, Z. Chen, Z. Wang, W. Ouyang, “Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition,” Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.143-152, Jun. 2020.

[5] I. Loshchilov and F. Hutter, “Fixing Weight Decay Regularization in Adam,” Proc. of the Ninth International Conference on Learning Representations, pp.1–14, Feb. 2018.

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime Multi-person 2D Pose Estimation using Part Affinity Fields,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.43, no.1, pp.172–186, Jan. 2021.

[7] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, “PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction,” Proc. of the 2019 IEEE/CVF International Conference on Computer Vision, pp.6261–6270, Oct. 2019.

[8] Y. Belkada, L. Bertoni, R. Caristan, T. Mordan, and A. Alahi, “Do Pedestrians Pay Attention? Eye Contact Detection in the Wild,” arXiv preprint arXiv:2112.04212, Dec. 2021.