Vol.2021-NL-248 No.31
Vol.2021-CVIM-226 No.31
2021/5/21

胸部方向を考慮した車載映像における歩行者の経路予測

福田 太一[1,a)]　延原 章平[1,2]　西野 恒[1]

概要：歩行者の経路予測は自動運転において進路決定や衝突回避を正確に行う上で不可欠である．本研究では，車載映像における歩行者の経路予測問題に取り組む．歩行者の姿勢は経路予測において重要な手がかりの 1 つであるが，低解像度の車載映像から正確に推定することは難しい．本論文では，低解像度の映像からも頑健に推定することが可能な胸部方向を経路予測の手がかりとして用いることを提案する．経路予測を行うニューラルネットワークは歩行者の位置・深度・エゴモーションに加え，胸部方向を入力として受け取り，将来の位置を出力する．胸部方向を用いないモデルとの比較実験により，胸部方向が幅広いシーンで予測精度向上に寄与していることを示した．

1. 歩行者の経路予測

自動運転技術は物流における無人配送や自動タクシー・バスなどの応用範囲があり，運転タスクからの解放や人員不足解消に役立つ技術である．現在，自動ブレーキや車線維持のアシストに加え，高速道路など特定の状況における自動運転は実用化されつつある．しかしながら，市街地などの複雑な環境に対応できる自動運転は実現できていない．

市街地における自動運転の難点は，市街地では高速道路と異なり歩行者との関係を考慮する必要がある点である．例えば，市街地では歩行者が横断歩道を渡ろうとしているなら停止する，小さい子供の飛び出しを警戒してスピードを落とすといったように状況に応じた複雑な対応が要求される．このような自動車と歩行者のインタラクションは市街地での自動運転技術の実現に必要不可欠である．一方で，歩道や横断歩道，時には車道を縦横無尽に移動する歩行者の動きを捉えることは難しい．

歩行者の動きを予測することを目的として，コンピュータビジョンの分野では歩行者の経路や意図を予測する多くの研究がなされてきた [1]．Mangalam らは一人称視点で撮影された映像から予測を行い [2]，Dendorfer らは監視カメラのような定点映像から予測を行う研究を行った [3]．これらの研究によって，歩行者の経路を大まかに予測することが可能になったが，未だ完全な自動運転を実現できるほどの精度で予測を行うことはできていない．

本研究は自動運転への応用を意識し，図 1 のように車載映像における歩行者の経路予測を行う．車載映像には歩行者の位置はもちろん，奥行きや路面情報，自動車自身の動きなど歩行者の経路予測における多くの手がかりが存在する．理想的には画像から得られるすべての情報を用いたいが，計算資源の観点から現実的ではない．通常は手がかりとなる情報の中でも重要度が高いと考えられるものに絞ってモデルへの入力とする．入力する情報の選び方により予測精度の向上を目指すことができる．

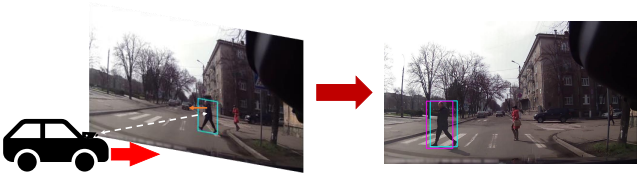

図 1　車載映像における歩行者の経路予測．車載映像から得られる複数の情報から，将来の歩行者の位置を予測する．本研究では歩行者の位置・深度・エゴモーション・胸部方向を手がかりとして経路予測を行う．シアンのバウンディングボックスが歩行者の位置の真値，マゼンタのバウンディングボックスが予測結果を表している．

歩行者の姿勢は経路予測における重要な手がかりの 1 つであるので，モデルへの入力として有用であると考えられる．しかし，車載映像が低解像度であるため歩行者の 2 次元姿勢の推定精度は低く，推定された姿勢を直接モデルの入力として与えても予測精度向上は困難である．胸部方向は 2 次元姿勢に比べ低解像度でも推定しやすい単純な情報でありながら，歩行者の姿勢から得られる情報を経路予測に必要なレベルで十分に表現している．なぜなら，歩行者の経路予測において腕の振り方などの各関節点の局所的な情報よりも，歩行者の進行方向などの姿勢全体からわかる大域的な情報が重要だからである．歩行者の姿勢情報とし

[1] 京都大学大学院 情報学研究科
[2] JST さきがけ
[a)] tfukuda@vision.ist.i.kyoto-u.ac.jp

て胸部方向を手がかりとすることで解像度の低い車載映像における歩行者の経路予測の頑健なモデルを設計することが可能であると考えられる．

本研究では手がかりとなる情報として歩行者のバウンディングボックス・深度・エゴモーションに加え，歩行者の胸部方向を用いる．深度は歩行者と自動車の距離を，エゴモーションは歩行者と自動車の相対的な位置関係の変化を表す情報である．胸部方向は地面平面に射影された歩行者の胸部の向きとして定義される．予測に用いるモデルは，上記の 4 種類の情報を入力として受け取り，歩行者の将来のバウンディングボックスを予測するニューラルネットワークとして設計される．

胸部方向の有用性を検証するために，JAAD データセット [4] で学習したモデルを用いて以下の評価実験を行った．入力から深度・エゴモーション・胸部方向のいずれかを除いた場合における予測精度を JAAD データセットを用いて評価した．胸部方向を入力として与えることで精度が最も向上し，歩行者の経路予測において胸部方向が深度・エゴモーションに比べ有用性が高いことが示された．予測フレームを変化させた場合における評価も JAAD データセットを用いて行い，特に長期の予測で精度向上に寄与することがわかった．Caltech Pedestrian データセット [5] に対しても胸部方向を用いないモデルと用いるモデルの予測精度比較を行い，モデルの汎化性能を検証した．学習時と異なるデータセットに対しても精度向上が見られ，幅広いシーンに対して胸部方向の有用性が確認された．我々は胸部方向を用いた歩行者の経路予測モデルを通して，完全な自動運転に必要な高精度での経路予測の実現に貢献した．

2. 関連研究

本章では，まず一人称視点における経路予測に関する研究について代表的なものを紹介する．次に，本研究で入力として用いる胸部方向・深度・エゴモーションを推定する研究について述べる．

2.1 一人称視点における経路（行動）予測

一人称視点における経路予測問題は，自動運転を想定して車載映像を対象にした研究が多く，学習ベースの様々な手法が存在する．Mangalam ら [2] は，車載映像に映る歩行者の将来の位置を関節点レベルで予測する手法を提案した．車載映像が低解像度であるため，機械学習モデルを用いて推定された姿勢にはノイズが発生し，精度の低下や一部関節点の消失が見られた．彼らは，得られた姿勢のうち不確かであるものからノイズを除去するネットワークを提案し，機械学習モデルを用いることにより発生する不確かさを低減した．姿勢予測ネットワークは Global stream と Local stream の 2 ストリームモデルとして設計され，歩行者の姿勢の大域的な情報と局所的な情報を別々に学習することによって姿勢という複雑な対象の予測を可能にした．

未来における歩行者の位置や経路を予測するのではなく，意図を予測する研究も存在する．Liu ら [6] は歩行者が近い将来に道路を渡るか渡らないかという未来の行動により意図を定義し，それを車載映像から予測するモデルを提案した．歩行者と他のオブジェクト（自動車，他の歩行者，信号機など）との空間的な関係をグラフ畳み込み演算によりグラフ構造としてモデル化した．グラフを時系列的に結合することにより，時空間的な歩行者とオブジェクトとの関係から歩行者の意図を推論した．一方で，車載映像でなく歩行者の視点から他の歩行者の経路予測を行う研究も存在する．八木ら [7] は歩行者視点の一人称視点データセットを作成し，そのデータセットを用いて歩行者視点の経路予測モデルを提案した．

本研究では，自動運転への応用を目指して，車載カメラから取得した映像を用いて歩行者の経路予測を行う．本手法で提案するネットワークは歩行者の大域的な動きを予測する問題を解くため，Mangalam ら [2] の Global stream を基にした構造を用いる．

2.2 方向推定

画像から人間の向いている方向を推定する問題は歩行者の経路予測だけでなくロボットや VR など様々な応用先を持つ．胸部方向推定では人間の胴体の向きを推定する．既存の胸部方向のアノテーションを持つデータセット [8] には規模の小ささやアノテーション精度の低さという問題があった．Wu ら [9] は，様々なシーンで撮影された人物画像を含む COCO データセット [10] に対して，正確な人間の胸部方向のアノテーションを行った MEBOW データセットを作成した．彼らは MEBOW データセットを学習に用いた汎化性能の高い胸部方向推定モデルを提案した．3 次元姿勢推定において胸部方向推定結果を用いた損失を与えることにより高精度な推定を行えるようになることも示した．

2.3 深度・エゴモーション推定

深度推定はカメラに対する画像上の各オブジェクトとの距離を推定する問題であり，これにより 2 次元画像から 3 次元情報を得ることが可能になる．エゴモーション推定は動画中でのカメラ自身の移動を推定する問題である．Casser ら [11], [12] は車載映像から深度推定を行う教師なしモデルを提案した．推定された深度およびエゴモーションを用いて連続するフレームの画像を変換し，変換された画像と元画像との差分を損失とすることにより教師なしでの学習を可能にした．推論時は単一の画像に対しても深度を推定することができる．

Vol.2021-NL-248 No.31
Vol.2021-CVIM-226 No.31
2021/5/21

3. 胸部方向を考慮した経路予測

車載映像によりある歩行者が観測されているとするとき，現在のフレームまでの観測により得られた情報をもとに将来の歩行者の位置を予測することが本研究の目的である．歩行者の位置はバウンディングボックス $\boldsymbol{b} = [x_{tl}, y_{tl}, x_{br}, y_{br}]$ で表現される．(x_{tl}, y_{tl})，(x_{br}, y_{br}) はそれぞれ画像座標上におけるバウンディングボックスの左上 (top left)，右下 (bottom right) の座標を表している．本研究では，歩行者の経路予測における胸部方向の有用性を評価するために，歩行者の現在までの位置は既知であるとして真値を用いる．歩行者のバウンディングボックスを画像から検出する研究も存在する [5] が，検出結果を用いると歩行者検出の誤差を考慮する必要があり問題が複雑化してしまう．

その他の情報として，既存の手法により観測画像から推定された歩行者の胸部方向・深度・エゴモーションを用いる．以上の 4 種類の過去情報を入力とし，将来のバウンディングボックス座標を出力するようなニューラルネットワークを学習することにより歩行者の経路を予測する．

3.1 ネットワーク構造

図 2 にネットワーク構造を示す．ネットワークの入力は過去 t_p フレームのバウンディングボックス・胸部方向・深度・エゴモーションである．以上の入力からネットワークは未来 t_f フレームのバウンディングボックスを予測する．バウンディングボックス座標は画像の解像度に応じて 0 から数千の範囲を持つ値である．ネットワークの入出力の値が大きくなることは予測精度の低下の原因になりうるため，入出力はバウンディングボックスの前フレームとの差分を用いる．また，ネットワークの入出力が歩行者のバウンディングボックスの大きさの違いにより大きく変化することを避けるため，過去 t_p フレーム間のバウンディングボックスの面積の平均

$$S = \frac{1}{t_p} \sum_{t=t_0-t_p+1}^{t_0} (x_{br} - x_{tl})(y_{br} - y_{tl}) \tag{1}$$

を求め，入出力のバウンディングボックスを S を用いて正規化する．ここで，t_0 は現在の時刻である．

各フレームごとに得られる情報から予測に必要な特徴量を抽出するためにフレームエンコーダを導入する．フレームエンコーダへの入力は各フレーム $t = t_0 - t_p + 1, ..., t_0$ に対し，正規化されたバウンディングボックスの差分 $\Delta\boldsymbol{b}_t$，胸部方向の期待値 $\boldsymbol{o}_t$，バウンディングボックスの中心における深度 d_t，過去 3 フレーム間のエゴモーション $\boldsymbol{e}_t = [\boldsymbol{e}_{t-2\to t-1}, \boldsymbol{e}_{t-1\to t}]$ を連結したベクトルである．ここで，

$$\Delta\boldsymbol{b}_t = \frac{\boldsymbol{b}_t - \boldsymbol{b}_{t-1}}{\sqrt{S}} \tag{2}$$

$$d_t = \boldsymbol{D}_t \left(\frac{x_{br} + x_{tl}}{2}, \frac{y_{br} + y_{tl}}{2} \right) \tag{3}$$

である．$\boldsymbol{D}_t$ は画像全体の深度を，$\boldsymbol{D}_t(x, y)$ は画像座標系における点 (x, y) での深度を表す．また，$\boldsymbol{e}_{t-1\to t} = [\Delta_x, \Delta_y, \Delta_z, \theta_x, \theta_y, \theta_z]$ であり，Δ は各成分方向の平行移動を，θ は回転を表す．フレームエンコーダはフレーム間で重みを共有している．フレームエンコーダの出力は圧縮されたベクトルであり，これが QRNN エンコーダ-デコーダへの入力となる．QRNN デコーダの出力をフレームデコーダに入力することにより，予測されたバウンディングボックスの差分 $\{\hat{\Delta\boldsymbol{b}}_t\}_{t=t_0+1}^{t_0+t_f}$ を得る．バウンディングボックスの予測結果 $\{\hat{\boldsymbol{b}}_t\}_{t=t_0+1}^{t_0+t_f}$ は

$$\hat{\boldsymbol{b}}_t = \boldsymbol{b}_{t_0} + \sqrt{S} \sum_{t'=t_0+1}^{t} \hat{\Delta\boldsymbol{b}}_{t'} \tag{4}$$

で計算できる．

ネットワークの損失としては予測されたバウンディングボックスと真値のバウンディングボックスの L1 距離

$$L = \frac{1}{t_f} \sum_{t=t_0+1}^{t_0+t_f} |\boldsymbol{b}_t - \hat{\boldsymbol{b}}_t| \tag{5}$$

を用いる．学習時には，誤差逆伝搬により，フレームエンコーダ，QRNN エンコーダ-デコーダ，フレームデコーダの重みを更新する．

3.2 胸部方向推定

人間の姿勢は，進行方向を予測するうえで非常に大きな手がかりとなる．精度の高い 2 次元姿勢推定をおこなうにはある程度十分な解像度の人間の画像が要求される．しかし，車載映像に映る歩行者はカメラとの距離が遠いと低解像度になるため推定精度が低下してしまい，推定された姿勢を直接ネットワークの入力として用いても精度の向上はあまり見込めない．Liu ら [6] は，推定した姿勢を加えることによる意図予測の精度の向上は見られなかったとしている．そのため，手がかりとして姿勢よりも単純な情報が必要である．

胸部方向は 2 次元姿勢に比べて非常に単純な情報でありながら，姿勢の持つ進行方向情報を損なわない有用な手がかりであると考えられる．本研究では高い汎化性能を持つ胸部方向推定モデルである MEBOW [9] を用いることで進行方向についての手がかりを得る．MEBOW によって推定される胸部方向の定義を図 3 に示す．人間の両肩の関節点を結ぶベクトルと胴体を表すベクトルに対し，外積を取ることにより 3 次元胸部方向が得られる．MEBOW は 3 次元胸部方向を地面に射影した 2 次元胸部方向を推定するモデルである．画像平面に対し奥側が 0° で手前側が 180° になっており，時計回りを正方向の回転とする．本モデルへの入力はバウンディングボックスにより切り取られた歩行

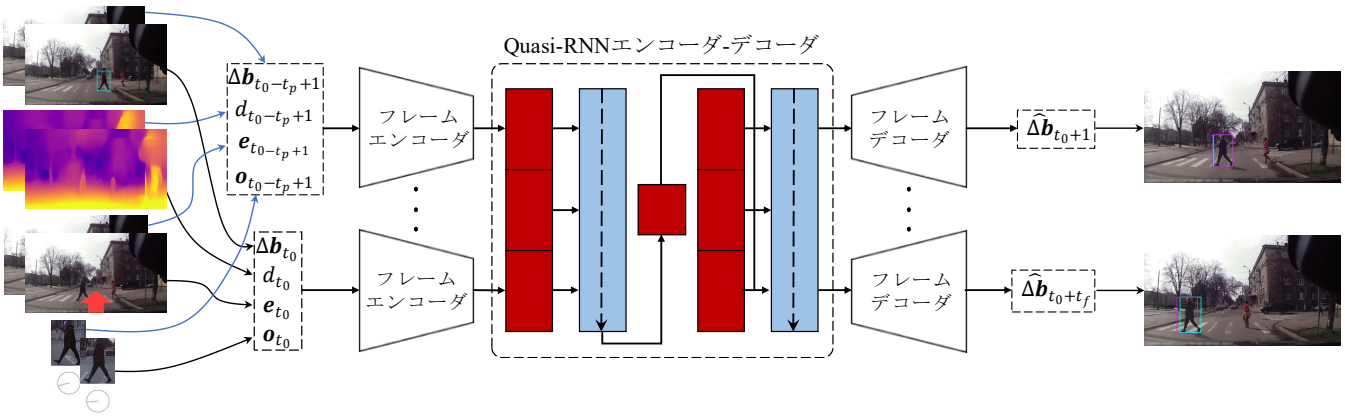

図 2 経路予測ネットワークの構造．入力は，上から，歩行者のバウンディングボックスの前フレームとの差分を正規化したもの $\Delta \boldsymbol{b}_t$，バウンディングボックス中心の深度 d_t，エゴモーション $\boldsymbol{e}_t$，胸部方向の期待値 $\boldsymbol{o}_t$ である．出力は予測フレーム t_f の正規化されたバウンディングボックスの差分 $\hat{\boldsymbol{b}}_t$ である．フレームエンコーダ，デコーダはそれぞれ重みを共有している．QRNN エンコーダ-デコーダにより時系列的処理を行う．

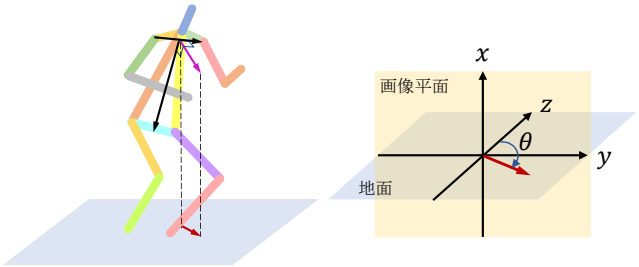

図 3 3 次元胸部方向は両肩を表すベクトルと胴体を表すベクトルの外積（マゼンタのベクトル）により得られる（左図）．MEBOW[9] では 3 次元胸部方向ベクトルの地面への射影である 2 次元胸部方向を推定する．画像奥側である z 軸正方向を 0° とし，時計回りを正方向の回転とする（右図）．

者の画像である．出力は $\boldsymbol{p} = [p_0, p_1, ..., p_{71}] (\Sigma_{i=0}^{71} p_i = 1.0)$ であり，入力画像中の人物の胸部方向を表現するヒートマップを表している．より具体的には，ヒートマップの各要素 p_i は胸部方向が $[(i \times 5 - 2.5)°, (i \times 5 + 2.5)°]$ の範囲である確率を表す．

得られたヒートマップ $\boldsymbol{p}$ に対し方向の期待値を表すベクトル $\boldsymbol{o} = [o_x, o_y]$ を計算する．$i = 0, 1, ..., 71$ に対し $\theta_i = i \times 5°$ とすると，

$$o_x = \sum_{i=0}^{71} p_i \cos \theta_i$$
$$o_y = \sum_{i=0}^{71} p_i \sin \theta_i \quad (6)$$

により $\boldsymbol{o}$ が得られる．$\boldsymbol{o}$ は予測ネットワークの入力となる．方向ベクトルの期待値をとることによってヒートマップでの方向推定の確からしさに対応する重みづけを行うことができる．例えば，ヒートマップすべての要素が同じ値を持つ推定結果が得られたとき，それはモデルがその人物の向いている方向を推定できなかったことを意味する．このとき，$\boldsymbol{o} = \boldsymbol{0}$ となる．逆にある要素のみの値が 1 であるようなヒートマップが得られたとき，$|\boldsymbol{o}| = 1$ となりこれが $|\boldsymbol{o}|$ の最大値となる．

3.3 深度・エゴモーション推定

人間は目に写った映像から奥行きを考慮した上で歩行者の動きを予測する．例えば近くにいる歩行者の動きは大きく，遠くにいる歩行者の動きは小さいというように，奥行きを知ることにより得られる知識は大きい．奥行きの情報はバウンディングボックスの大きさからもある程度推測することが可能だが，バウンディングボックスの大きさは歩行者の身長により異なるので，その大きさだけで奥行きを測ることは信頼性に欠ける．本研究では歩行者の自動車との距離をより正確に測るために，深度を予測ネットワークの入力とする．また，自動車自身の動きによって歩行者のバウンディングボックスの位置は変化する．エゴモーションによる歩行者の位置の相対的な変化を考慮する必要があるため，エゴモーションも予測ネットワークの入力とする．

深度・エゴモーション推定には，教師なし学習モデルである Struct2Depth[11], [12] を用いる．エゴモーションと深度が正しく推定されることで隣接するシーン間で画像の再構成を行うことができる．本モデルは隣接フレーム画像の再構成誤差を最小化するような損失を導入し，モデルの学習を行うことで教師なしでの深度・エゴモーション推定を可能にしている．画像列をモデルに入力することで，画像全体の深度 $\boldsymbol{D}_t$ および前フレームに対するエゴモーション $\boldsymbol{e}_{t-1 \to t}$ を得る．Struct2Depth により得られた出力を，図 2 のようにバウンディングボックス・胸部方向とともにフレームエンコーダへの入力として用いる．

3.4 Quasi-Recurrent Neural Network

本研究では，時系列データ処理のアーキテクチャとして Quasi-Recurrent Neural Network(QRNN)[13] を用いる．QRNN は Convolutional Neural Network のプーリング層で時系列情報を伝搬することにより Recurremt Neural Network のように過去情報を考慮したネットワークの出力を可能にしている．時系列処理を行うニューラルネットワークには，過去の情報を必要とするため系列長が長いと計算時間がかかるという問題がある．QRNN は並列計算可能な畳み込み演算と単純なプーリング演算による時系列情報伝搬により構成されているため，長い入力に対しても計算時間を抑えて処理することができる．

本研究では QRNN を用いたエンコーダ-デコーダを用いることで系列入力から系列出力を得る．図 2 において，フレームエンコーダの出力は QRNN エンコーダに入力され，QRNN エンコーダは過去の情報から得られた隠れ変数を QRNN デコーダに受け渡す．QRNN デコーダは QRNN エンコーダの隠れ変数を受け取ることで，過去の情報から将来の歩行者の位置を予測する．

4. 評価実験

本章では，歩行者の経路予測における胸部方向の有用性を検証するための評価実験の内容およびその結果について述べる．予測ネットワークに与える入力の 1 つを提案手法から取り去ることによる予測精度の変化を評価した．予測フレームの変化にともなう予測精度の変化や異なる評価用データセットに対するモデルの汎化性能も評価した．

4.1 データセット

本実験では Joint Attention in Autonomous Driving(JAAD) データセット [4] を用いて予測ネットワークの学習を行い，JAAD データセットおよび Caltech Pedestrian データセット [5] を用いて評価を行った．

JAAD データセットは自動運転の文脈における運転者と，歩行者や他の運転者との関係の研究のために作られたデータセットである．自動車のフロントカメラで撮影された 30 fps，5 ∼ 10 秒の動画 346 本で構成されている．動画の総フレーム数は 82032 である．北アメリカおよびヨーロッパの複数の場所で撮影されており，様々な天候条件のもとで日常的な市街地の走行シーンを表現している．動画中に登場する歩行者の数は 2793 人であり，そのすべての歩行者に対しフレームごとにバウンディングボックスが与えられている．動画の解像度は 1280×780 のものが 10 本，1920×1080 のものが残りの 336 本である．

Caltech Pedestrian データセットは道路シーンにおける歩行者検出のためのデータセットである．自動車のフロントカメラから撮影された 30 fps，解像度 640×480 の動画により構成されている．総フレーム数は約 250000 で，歩行者数は約 2300 人である．すべての歩行者に対しフレームごとにバウンディングボックスのアノテーションが行われている．歩行者のラベルは 1 人の歩行者を指す ‘Person’ と複数の歩行者のグループを指す ‘People’ に分けられており，‘People’ はグループ全体を 1 つのバウンディングボックスによりアノテーションしている．そのため，本研究では歩行者のバウンディングボックスとして ‘Person’ ラベルが付与されたもののみを用いた．

同じ歩行者が訓練用データと評価用データの両方に入ることを避けるために，それぞれ異なる動画の歩行者データを用いた．

4.2 評価指標

本研究では評価指標として Intersection over Union (IoU) を用いる．IoU は 2 つのバウンディングボックスの和領域に対する積領域の割合を指し，予測されたバウンディングボックスと真値のバウンディングボックスの重なり具合を表す指標である．IoU は予測されたバウンディングボックスと真値のバウンディングボックスが完全に一致するとき最大値 1 をとり，2 つが全く重ならない場合に最小値 0 をとる．IoU はすべての予測フレームに対し計算することができる．各予測フレームに対し IoU を計算し，データセット全体における平均 (IoU-average) と予測フレームのうち最後のフレームのみの IoU の平均 (IoU-last) を求めた．

4.3 学習の詳細

入力フレーム数は $t_p = 30$ とした．出力フレーム数は $t_f = 6, 30, 60$ とした．深度およびエゴモーションの推定には Struct2Depth [11], [12] を用いた．モデルは KITTI データセット [14] で学習済みのものを用い，3 枚の画像の組を入力とすることで深度・エゴモーションを得た．胸部方向推定には MEBOW [9] を用いた．胸部方向アノテーションが与えられた MEBOW データセットにより学習済みのモデルを使用した．方向推定ネットワークへの入力はバウンディングボックスにより切り取った歩行者の画像である．経路予測ネットワークのフレームエンコーダには 2 層の線形層を使用した．1 層目の線形層は入力を 8 次元出力に圧縮し，2 層目は 8 次元入力を 4 次元出力に圧縮する．QRNN エンコーダ-デコーダ [13] の隠れ変数は 8 次元であり，畳み込み層のカーネルサイズは 2，層数は 2 とした．フレームデコーダは 1 層の線形層であり，QRNN デコーダの最後の隠れ変数（8 次元）を 4 次元に圧縮して，バウンディングボックスの差分を出力する役割を持つ．エポック数は 200 とした．学習率は 0.01 から始まり 50 エポックごとに 0.1 倍するようなスケジューリングを行った．バッチサイズは 128 とした．最適化アルゴリズムとしては Adam[15] を用いた．Adam はニューラルネットワークの最適化アルゴリズムとして広く使われている手法である．モデルの

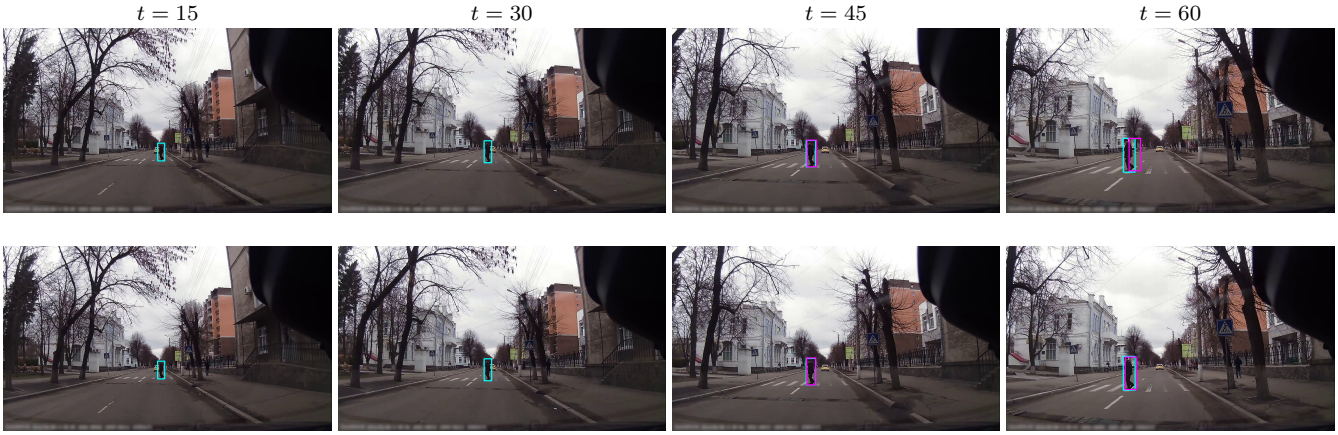

図 4 胸部方向なしモデル（上段）および胸部方向ありモデル（下段）による予測結果．左側（$t = 15, 30$）に観測フレーム，右側（$t = 45, 60$）に予測フレームを示す．シアンのバウンディングボックスが真値，マゼンタのバウンディングボックスが予測結果を表す．胸部方向なしモデルは観測から時間が経過するほどバウンディングボックスのずれが大きくなっている．胸部方向ありモデルでは観測から時間が経過してもバウンディングボックスのずれは小さい．

訓練には JAAD データセット [4] の 300 クリップを用い，評価には JAAD データセットの 46 クリップと，Caltech Pedestrian データセット [5] を用いた．

4.4 入力変化に対する予測精度の評価

本節では予測フレーム $t_f = 30$ としたときの，人間の方向を入力することによる予測精度の変化を評価する．深度とエゴモーションについても入力として与えた場合と与えなかった場合についての予測精度の変化を評価する．表 1 に入力変化に対する IoU-average，IoU-last の変化を示す．IoU-last のほうが IoU-average よりも低くなっている．これは入力から時間的に離れるほど誤差が蓄積され，予測精度が下がることを表している．IoU-average は提案手法，深度なし，エゴモーションなし，胸部方向なしという順で IoU が大きいという結果になった．IoU-last は深度なし，提案手法の値が最も高く，次がエゴモーションなし，最も精度が低かったのが胸部方向なしという結果になった．本節で比較した入力の中で，歩行者の胸部方向は歩行者の経路を予測する上で最も有用な手がかりであると考えられる．図 4 にそれぞれ胸部方向なしのモデルと提案したモデルによる同一シーンの予測結果を示す．$t = 1$ を予測開始フレームとしている．$t = 60$ の時点で，胸部方向なしモデルの予測バウンディングボックスには真値に対しずれが見られるが，胸部方向ありのモデルではほとんどずれが見られないことがわかる．歩行者の動きとエゴモーションが同時に観測される自動運転における一般的なシーンで，胸部方向を与えることにより予測精度が向上することが示されている．

表 1 入力を変化させた場合の IoU の変化．IoU-average，IoU-last ともに提案手法で最大値を取っており，胸部方向なしモデルで IoU が最小値を取っていることから，胸部方向が比較した 3 つの入力の中で最も重要な手がかりであると言える．

	IoU-average	IoU-last
胸部方向なし	0.706	0.489
エゴモーションなし	0.712	0.504
深度なし	0.713	**0.505**
提案手法	**0.714**	**0.505**

4.5 予測フレーム変化に対する予測精度の評価

表 2 に予測フレーム $t_f = 6, 30, 60$ としたときの胸部方向なしモデルと提案手法における IoU を示す．表の各要素は IoU-average/IoU-last を表している．すべての時刻で提案手法のほうが IoU が高いことがわかる．時刻ごとの IoU の変化に関しては，予測フレームが大きくなればなるほど胸部方向なしモデルと提案手法で IoU の差が大きくなっていることがわかる．特に，$t_f = 60$ における IoU-last は胸部方向なしモデルが 0.155 であるのに対し提案手法は 0.204 と大きな差が見られる．胸部方向は特に長期的な予測に対してより有用であると言える．胸部方向は歩行者の将来の進行方向に関する意図を含んでいるため，より長期の予測で精度の向上に寄与していると考えられる．予測フレームの増加に伴う IoU の減少は，フレームごとの誤差が (4) 式により蓄積されたことを示唆している．

4.6 Caltech Pedestrian データセットでの評価

JAAD データセットで訓練したモデルに対し Caltech Pedestrian データセットを用いて評価を行った．$t_f = 30$ とした．結果を表 3 に示す．IoU-average，IoU-last ともに方向入力を与えたモデルのほうが高くなるという結果となり，異なるデータセットに対しても胸部方向入力は

表 2 予測フレームを $t_f = 6, 30, 60$ としたときの IoU の変化．表中の各数値は IoU-average/IoU-last を表している．すべての予測フレームにおいて IoU-average，IoU-last ともに提案手法のほうが大きな値を取っている．予測フレームが小さいほうが IoU の値が大きく，予測フレームが大きいほうが胸部方向ありとなしのモデル間の IoU の差は大きい．

t_f	6	30	60
胸部方向なし	0.900/0.843	0.706/0.489	0.510/0.155
提案手法	**0.904/0.847**	**0.714/0.505**	**0.532/0.204**

意味を持つことが示された．しかし，表 1 と比較すると，IoU-average，IoU-last の両方で数値が 0.1 以上減少していることもわかる．これは訓練を JAAD データセットで行ったためであると考えられる．

より具体的には，データ画像の解像度や縦横比が異なることで既存の手法による深度・エゴモーション，胸部方向推定に影響を及ぼし，それを入力とする予測ネットワークの出力が変化したことが原因であると考えられる．また，JAAD データセットと Caltech Pedestrian データセットは画像の解像度が異なるため，予測すべきバウンディングボックスのスケールは異なる．しかし，3.1 節で述べた通りバウンディングボックスは面積がおよそ 1 になるように正規化し入力されている．そのため，2 つのデータセットの間の解像度の差によるバウンディングボックスのスケールの変化は大きな問題にはならないと考えられる．

4.7 失敗例

図 5, 図 6 に失敗例となる予測結果の画像を示す．$t_f = 30$ としたモデルにおいて観測開始時刻を $t = 1$ としたときの観測画像および予測結果画像である．上段が観測画像，下段が予測画像である．シアンのバウンディングボックスが真値を，マゼンタのバウンディングボックスが予測値を表している．

図 5 について，2 つのバウンディングボックスはほとんど重なっておらず，予測が正しく行えていないことが読み取れる．このような予測結果が得られた原因は，観測フレームでは歩行者は横断歩道の前で停止していたが，予測フレームで歩き出したことであると考えられる．本研究で提案したモデルは歩行者のバウンディングボックス・胸部方向・深度・エゴモーションを与え，エンコーダ-デコーダアーキテクチャで時系列処理をおこなう単純な構造を持つ．そのため，横断歩道で歩行者が待っていれば自動車は止まり，自動車が止まれば歩行者は歩き出すといったような自動車，歩行者間のインタラクションを学習しきれていないことが考えられる．

図 6 についても真値と予測値でバウンディングボックスの位置が大きくずれている．この画像については，歩行者は観測フレームと予測フレームで歩行者の移動速度は大きく変化していない．しかし，観測フレームでのエゴモーションと予測フレームのエゴモーションが異なっている．これが誤った予測の原因であると考えられる．

表 3 Caltech Pedestrian データセットを用いた IoU の評価結果．IoU-average，IoU-last ともに提案手法が大きな値を取っている．表 1 と比較すると胸部方向ありとなしの 2 モデルとも IoU が減少している．

	IoU-average	IoU-last
胸部方向なし	0.592	0.352
提案手法	**0.597**	**0.357**

以上のことから，このモデルは歩行者が突然移動したり自動車が突然停車するといったような急な変化にともなうバウンディングボックスの位置の変化の予測が正しくできていない場合が多いと考えられる．逆に，図 4 のように正しく予測できている場合は歩行者や自動車が一定の速度で移動している状況下であることが多い．

5. 結論

本研究では車載映像における胸部方向を考慮した歩行者の経路予測モデルを提案し，経路予測における胸部方向の有用性を検証するために 3 つの定量的評価を行った．胸部方向なし，深度なし，エゴモーションなしのモデルと提案手法の比較では，胸部方向を与えた場合に最も IoU が向上した．このことから胸部方向は歩行者の経路予測における重要な手がかりであると言える．予測フレームを変化させた場合では $t_f = 60$ のときに胸部方向の有無での IoU の差が最も大きく，胸部方向は長期的な予測に特に有用であることがわかった．訓練時と異なる Caltech Pedestrian データセットに対する予測でも胸部方向なしのモデルに比べ IoU が向上し，胸部方向を用いたモデルは多様なシーンで歩行者の経路予測の精度を向上させることを示した．我々は様々な側面から歩行者の経路予測における胸部方向の有用性を示すとともに，自動運転技術に不可欠な歩行者の経路予測の精度向上に貢献した．

今後の課題は，提案手法での失敗例を改善することである．歩行者や自動車が観測されるフレームと予測されるフレームで異なる動きをしているときに正しく歩行者の位置が予測できていなかった．この問題を解決するには，例えば自動車が停止すれば歩行者が歩き出すといった自動車と歩行者のインタラクションを明示的にモデル化するような手法が必要になる．加えて，信号機などさらに多くのオブジェクトとのインタラクションを考えることにより更に予測精度が向上すると考えられる．このようなモデル化として Liu ら [6] のように各オブジェクト間の関係をグラフとして定式化する方法が有用であると考えられる．

謝辞 本研究の一部は JSPS 科研費 17K20143, 20H05951, JST さきがけ JPMJPR1858 の助成を受けたものです．

図 5 提案したモデルでの予測の失敗例．観測フレームでは歩行者が歩道で停止しているためネットワークはそのまま停止すると判断して予測を行ったが，実際は予測フレームで歩行者が歩き出したため真値と予測値でバウンディングボックスが大きくずれていると考えられる．

図 6 提案したモデルでの予測の失敗例．歩行者の動きは観測フレームと予測フレームで大きく変化していないが，エゴモーションがその前後で変化したためにバウンディングボックスの予測値と真値が大きくずれたと考えられる．

参考文献

[1] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L. and Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 961–971 (2016).

[2] Mangalam, K., Adeli, E., Lee, K.-H., Gaidon, A. and Niebles, J. C.: Disentangling human dynamics for pedestrian locomotion forecasting with noisy supervision, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2784–2793 (2020).

[3] Dendorfer, P., Osep, A. and Leal-Taixé, L.: Goal-GAN: Multimodal Trajectory Prediction Based on Goal Position Estimation, *Proceedings of the Asian Conference on Computer Vision* (2020).

[4] Kotseruba, I., Rasouli, A. and Tsotsos, J. K.: Joint attention in autonomous driving (JAAD), *arXiv preprint arXiv:1609.04741* (2016).

[5] Dollar, P., Wojek, C., Schiele, B. and Perona, P.: Pedestrian detection: An evaluation of the state of the art, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 34, No. 4, pp. 743–761 (2011).

[6] Liu, B., Adeli, E., Cao, Z., Lee, K.-H., Shenoi, A., Gaidon, A. and Niebles, J. C.: Spatiotemporal relationship reasoning for pedestrian intent prediction, *IEEE Robotics and Automation Letters*, Vol. 5, No. 2, pp. 3485–3492 (2020).

[7] Yagi, T., Mangalam, K., Yonetani, R. and Sato, Y.: Future person localization in first-person videos, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7593–7602 (2018).

[8] Andriluka, M., Roth, S. and Schiele, B.: Monocular 3d pose estimation and tracking by detection, *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 623–630 (2010).

[9] Wu, C., Chen, Y., Luo, J., Su, C.-C., Dawane, A., Hanzra, B., Deng, Z., Liu, B., Wang, J. Z. and Kuo, C.-h.: MEBOW: Monocular Estimation of Body Orientation In the Wild, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3451–3461 (2020).

[10] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C. L.: Microsoft coco: Common objects in context, *European conference on computer vision*, Springer, pp. 740–755 (2014).

[11] Casser, V., Pirk, S., Mahjourian, R. and Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pp. 8001–8008 (2019).

[12] Casser, V., Pirk, S., Mahjourian, R. and Angelova, A.: Unsupervised monocular depth and ego-motion learning with structure and semantics, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2019).

[13] Bradbury, J., Merity, S., Xiong, C. and Socher, R.: Quasi-recurrent neural networks, *arXiv preprint arXiv:1611.01576* (2016).

[14] Geiger, A., Lenz, P., Stiller, C. and Urtasun, R.: Vision meets robotics: The kitti dataset, *The International Journal of Robotics Research*, Vol. 32, No. 11, pp. 1231–1237 (2013).

[15] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014).