Vol.2018-CVIM-211 No.3
2018/3/1

一人称視点映像における人物位置予測

八木 拓真[1,a)]　マンガラム カーティケヤ[2]　米谷 竜[1]　佐藤 洋一[1]

概要：一人称視点映像中に観測された人物の将来位置を予測する新たな問題に取り組む．具体的には，ウェアラブルカメラを用いて撮影した映像中に映る人物の短いクリップが与えられた時に，当該人物が近い将来撮影者の視野のどの地点に映るかを予測する問題を考える．一人称視点映像では撮影者自身の動きが映像中の自己運動として現れ，また固定視点映像に比べ映る人物の姿形を大きく捉えられるという特徴がある．そこで，本研究では対象人物の位置履歴，姿勢及び映像中の自己運動の 3 つの手掛かりの時系列を入力とする，畳み込みニューラルネットワーク (CNN) を用いた予測手法を提案する．人物位置予測を主目的とした一人称視点映像データセットは存在しなかったため，独自に計 4.5 時間のデータセットを構築し，提案手法の評価を行った．その結果，今回作成したデータセット及び公開データセットにおいて提案手法が既存の固定視点映像を対象としたモデルに比べ少ない誤差で予測を行えることを示し，自己運動及び姿勢情報が予測精度の向上に寄与することを実証した．

1. はじめに

ウェアラブルカメラから撮影された映像を解析する一人称ビジョンの分野において，視覚障碍者支援が 1 つの有望な応用先として注目されている．一人称ビジョンは，カメラ装着者自身が視界に捉えている環境に近い映像が撮影できるため，装着者の代わりに環境を知覚し，次に何をすべきかを補助・決定するシステムとして利用できる．

本研究では，装着者の周辺に多数の歩行者がいるような混雑環境におけるナビゲーションに注目する．ウェアラブルカメラで撮影した映像フレーム中に観測された人物の動きから，当該人物が続いてどの方向に歩いていくかを予測することができれば，その予測を装着者に通知することで衝突を回避することが可能となる．混雑環境においてこうした装着者ナビゲーションを実現するための第一歩として，本研究では一人称視点映像における人物位置予測という新しいタスクを提案する (図 1)．本タスクを解決するにあたり，一人称視点映像に特有の以下の特徴に着目する．

(1) **装着者の動きが映像中の自己運動として観察される：** 歩いている人物のそばを装着者が前進しながら通り過ぎる状況を考える．この際，映像と人物は装着者が前進しながら徐々に視野の下方向に移動する．もしこのとき装着者がと人物とが相対する形であったならば，相手は衝突を避けるためにわずかに移動方向を変えるだろう．こうした相互作用は撮影者自身の挙動に応じて発生するため，撮影者の動きを反映した自己運動は位置予測に寄与すると考えられる．

(2) **人物の姿勢及びその変化が近い将来動く方向を反映する：** 人物姿勢は，映像中の人物の将来の移動方向を決定づける有望な手がかりの 1 つである．当該人物が歩く方向を変えようとする際，その意図は首や足の向きなどに反映される．一人称視点映像には固定視点映像と比べて人物が大きく映るという特性があるため，姿勢情報を自然に組み込むことが可能である．

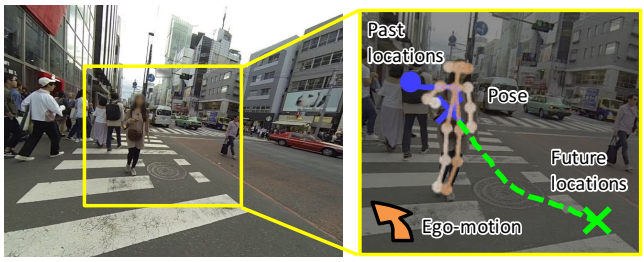

図 1: 一人称視点映像における未来位置予測．

以上の知見を元に，我々は対象人物の移動履歴，姿勢及び映像中の自己運動に基づく，一人称視点映像のための新しい人物位置予測手法を提案する (図 1)．具体的には，上記 3 つの手掛かりの時系列を入力として受け取り，対象人物が近い将来視野中に現れる位置の系列を出力する深層ニューラルネットワークとしてモデル化する．具体的には，各種手掛かりの時間変化を捉えるために畳み込みニューラ

[1] 東京大学
The University of Tokyo
[2] インド工科大学
Indian Institute of Technology
[a)] tyagi@iis.u-tokyo.ac.jp

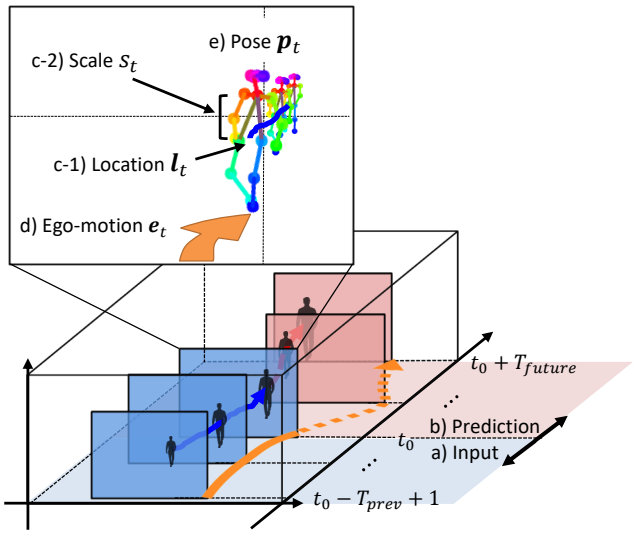

図 2: 問題設定．a) 対象人物の T_{prev} フレーム分の観測が与えられた時，b) 続く T_{future} フレームにおける将来位置を予測する．提案手法は c-1) 位置，c-2) スケール，d) 装着者の自己運動及び e) 姿勢情報を予測の手掛かりとする．

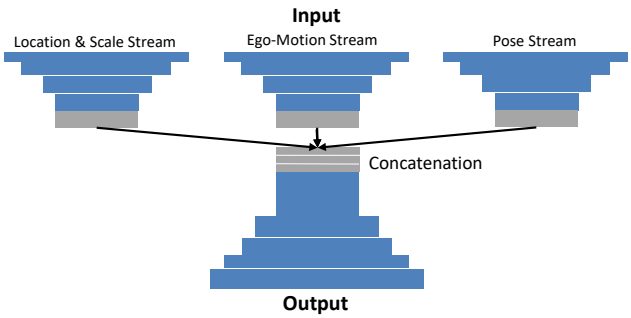

図 3: 提案モデルの概要．青いブロックは畳み込み層または逆畳み込み層を，灰色のブロックは中間特徴量を表す．

ルネットワーク (CNN) を使用し，時間方向に関する畳み込み・逆畳み込み構造を導入する．

提案手法の有効性を示すため，我々は新しいデータセットとして First-Person Locomotion (FPL) データセットを構築した．本データセットは多様な地点で撮影されたおよそ 5,000 人分の歩行を含む．我々は実験を通じて我々の手法が一人称視点映像中の人物の将来位置を十分に予測できることを実証した．また，公開されている一人称視点映像データセット [3] においても，提案手法が十分な性能を発揮することを確認した．

2. 提案手法

2.1 概要

本節では，一人称視点映像における人物位置予測問題を定式化する．図 2 に問題設定を示す．まず，予測対象人物が単眼 RGB カメラを用いて撮影した一人称視点映像中のとあるフレームに映っている状況を考える．我々の目標は，続くフレーム群において対象人物が撮影者 (カメラ) の視野中のどの地点に映るか (図 2 青線) を現在のフレームを含む過去の観測 (図 2 赤線) から予測することである．

$\boldsymbol{l}_t \in \mathbb{R}^2_+$ をフレーム t における人物の画像平面上の位置とする．現在のフレーム t_0 に続く T_{future} フレームにおける対象人物の将来位置系列 $L_{\text{out}} = (\boldsymbol{l}_{t_0+1} - \boldsymbol{l}_{t_0}, \boldsymbol{l}_{t_0+2} - \boldsymbol{l}_{t_0}, ..., \boldsymbol{l}_{t_0+T_{\text{future}}} - \boldsymbol{l}_{t_0})$ を予測することが直接的な目標である．

この際，技術的な焦点となるのは L_{out} の予測に有効な特徴量として何が使えるかという点である．図 2 に示すように，c-1) 位置及び c-2) 人物のスケールは将来位置を予測するための直接的な手がかりである．スケールは，カメラから対象人物までのおおよその奥行き情報を内包する．第 1 章で述べた通り，本研究で特に注目するのは d) 装着者の自己運動及び e) 対象人物の姿勢情報である．上記 3 つの手掛かりを適切に組み合わせたモデルとして，我々は 3 つの手掛かりに関する個別のストリームを持つ畳み込みニューラルネットワーク (CNN) を提案する (図 3)．

入力側の各ストリームは入力部を除き同一の構造を持つ．各畳み込み層は時間方向への畳み込み (1 次元) を複数回行う．各ストリームを通って抽出された中間特徴量を出力層の手前で結合し，時間方向の逆畳み込み層からなる単一の出力ストリームを経て L_{out} を生成する．ネットワークの最適化は誤差逆伝播法を用いて end-to-end で行われる．続いて，各ストリームの入力特徴量をより詳細に定義する．

2.2 位置スケール特徴

L_{out} を予測するための直接的な手がかりとなるのが画像平面上の位置履歴である．位置履歴は直進運動などの単純なケースの予測に大きく寄与するが，実世界上での距離関係を必ずしも正しく反映しない．例えば，同じ速度で歩いている人であっても，装着者から遠い人物の画像平面上での動きは小さく，近い人物の動きは大きく観測されてしまう．こうした見かけ上の動きと実世界での動きを適切に紐づけるためには，装着者から対象人物までの物理的距離を表す情報を何らかの形で得る必要がある．本研究では，人物のスケールが物理的距離を近似するとみなして，位置とスケールを組み合わせた特徴量を位置スケール特徴量として使用する．

具体的には，$L_{\text{in}} = (\boldsymbol{l}_{t_0-T_{\text{prev}}+1}, \ldots, \boldsymbol{l}_{t_0}) \in \mathbb{R}^{2\times T_{prev}}$ を予測直前までの位置情報の系列とする．続いて，各フレームでの位置 $\boldsymbol{l}_t \in \mathbb{R}^2_+$ にスケール情報 $s_t \in \mathbb{R}_+$ を加え，$\boldsymbol{x}_t = (\boldsymbol{l}_t^\top, s_t)^\top$ を得る．最終的にストリームに入力する位置スケール特徴量を $X_{\text{in}} = (\boldsymbol{x}_{t_0-T_{\text{prev}}+1}, \ldots, \boldsymbol{x}_{t_0}) \in \mathbb{R}^{3\times T_{prev}}$，対応する出力を $X_{\text{out}} = (\boldsymbol{x}_{t_0+1} - \boldsymbol{x}_{t_0}, \ldots, \boldsymbol{x}_{t_0+T_{\text{future}}} - \boldsymbol{x}_{t_0}) \in \mathbb{R}^{3\times T_{future}}$ とする．

2.3 自己運動特徴

位置スケール特徴 X_{in} は対象人物の動きを明示的に反映

しているものの，一人称視点映像では映像中の自己運動により画像平面上の位置 $\boldsymbol{l}_t$ が大きく変化する場合があるため，X_{in} から X_{out} を予測するのは必ずしも容易ではない．

一方，映像中の自己運動そのものは装着者と対象人物との相互作用が起こる状況下で対象人物の動きと密接に関わっている．例えば，装着者と対象人物が相対して，互いに何もしなければ衝突するという状況があったとき，装着者は対象人物の動きを見ながら (意識的か無意識的かに関わらず) 自らの歩行速度と方向を変更する．対象人物もまた装着者の動きを見ながら回避行動をとる．この際，映像中の自己運動それ自身が位置予測の重要な手がかりとなる．

そこで，我々はカメラ装着者が過去どのように動いてきたかを自己運動特徴量として反映することを提案する．まず，フレーム t のカメラ投影中心を原点とするカメラ座標系を基準として，フレーム $t-1$ の座標系から t の座標系への回転行列 R_t 及び並進ベクトル $\boldsymbol{v}_t$ が区間 $[t_0 - T_{\mathrm{prev}} + 1, t_0]$ について与えられたとする．これらはフレーム間の局所的な運動を表しているが，区間全体での大局的な運動を捉え切れていない．そこで $[t_0 - T_{\mathrm{prev}} + 1, t_0]$ 中の各フレーム t について，フレーム $t_0 - T_{\mathrm{prev}}$ のカメラ座標系を基準とした $t - (t_0 - T_{\mathrm{prev}})$ フレーム分のカメラの並進・回転運動の累積値を取る：

$$R'_t = \begin{cases} R_{t_0 - T_{prev} + 1} & (t = t_0 - T_{\mathrm{prev}} + 1) \\ R_{t-1}\, R'_t & (t > t_0 - T_{\mathrm{prev}} + 1), \end{cases} \tag{1}$$

$$\boldsymbol{v}'_t = \begin{cases} \boldsymbol{v}_t & (t = t_0 - T_{\mathrm{prev}} + 1) \\ R_t^{-1}\boldsymbol{v}_t + \boldsymbol{v}'_{t-1} & (t > t_0 - T_{\mathrm{prev}} + 1), \end{cases} \tag{2}$$

各フレームの特徴量は，R'_t をロールピッチヨー角に変換した $\boldsymbol{r}'_t$ 及び $\boldsymbol{v}'_t$ を結合した 6 次元ベクトルとして構成し，それを時間方向に並べた行列 E_{in} を最終的な自己運動特徴量とする：

$$\boldsymbol{e}_t = ((\boldsymbol{r}'_t)^\top, (\boldsymbol{v}'_t)^\top)^\top \in \mathbb{R}^6, \tag{3}$$

$$E_{\mathrm{in}} = (\boldsymbol{e}_{t_0 - T_{\mathrm{prev}} + 1}, \ldots, \boldsymbol{e}_{t_0}) \in \mathbb{R}^{6 \times T_{prev}}. \tag{4}$$

2.4 姿勢特徴

一人称視点映像は，固定視点映像と比べ眼前の人物を大写しにすることができる．結果，対象人物の姿勢をより精密に認識することが可能となり，それを将来の移動方向を見積もるための強力な手がかりとして利用できる．

本研究では姿勢特徴量として身体部位の位置系列を使うことを提案する．具体的には，眼，肩，腰，脚などを含む複数部位の位置系列を使用する．これを V 個の身体部位の 2 次元位置を結合した $2V$ 次元ベクトル $\boldsymbol{p} \in \mathbb{R}^{2V}_{+}$ として定義し，入力特徴量はそれを時間方向に並べた行列 $P_{\mathrm{in}} = (\boldsymbol{p}_{t_0 - T_{\mathrm{prev}} + 1}, \ldots, \boldsymbol{p}_{t_0}) \in \mathbb{R}^{2V \times T_{prev}}$ とする．

図 4: First-Person Locomotion データセット．

3. 実験

提案手法の有効性を確認するため，我々は新しい一人称視点映像データセットを構築し，その上で学習・評価を行った．また，一人称視点映像の公開データセットである Social Interaction データセット [3] においても評価を行い，装着者が歩行中に様々な行動を行う状況において提案手法がどのように振る舞うかを検証した．

3.1 データセット

提案手法の有効性を評価するために，我々は First-Person Locomotion (FPL) データセットを構築した．本データセットは胸部に装着したウェアラブルカメラ (GoPro HERO3 Black) より撮影されたおよそ 4.5 時間分の一人称視点歩行映像からなる．図 4 に出現フレームの例を挙げる．映像は主に混雑した街中で撮影され，データセット全体を通じて約 5000 人が検出された．

各訓練サンプルは $(X_{in}, E_{in}, P_{in}, X_{out})$ の組として与えられる．X_{in} は位置及びスケール，E_{in} は自己運動，P_{in} は姿勢，X_{out} は $\boldsymbol{x}_{t_0}$ に対する相対座標の系列である．入力である X_{in}, E_{in}, P_{in} の区間は $[t_0 - T_{\mathrm{prev}} + 1, t_0]$，出力である X_{out} の区間は $[t_0 + 1, t_0 + T_{\mathrm{future}}]$ とする．本研究においては，映像のフレームレートを 10 fps とし，$T_{prev} = T_{future} = 10$，すなわち 1 秒間の観測から 1 秒後までの位置を予測する問題とした．

撮影映像からの各サンプルの生成は次のように行った．まず，歪み除去を行った各フレームについて OpenPose[2] を用いて人物検出を行った．続いて，フレーム間の人物を対応づけるため，ホモグラフィ変換によって位置合わせを行ったフレームの組に対し Kernel Correlation Filter[4] を適用し人物を追跡した．予め設定した時空間領域内に追跡結果と実際の検出が見つかった場合検出同士を結び付け，そうでない場合は追跡を終了した．追跡の結果，多数の短

Layer type	Channel	Kernel size	Output size
Input Streams (Location-scale, ego-motion, and pose)			
Input	-	-	$D \times 10$
1D-Conv+BN+ReLU	32	3	32×8
1D-Conv+BN+ReLU	64	3	64×6
1D-Conv+BN+ReLU	128	3	128×4
1D-Conv+BN+ReLU	128	3	128×2
Output Stream			
Concat	-	-	384×2
1D-Conv+BN+ReLU	256	1	256×2
1D-Conv+BN+ReLU	256	1	256×2
1D-Deconv+BN+ReLU	256	3	256×4
1D-Deconv+BN+ReLU	128	3	128×6
1D-Deconv+BN+ReLU	64	3	64×8
1D-Deconv+BN+ReLU	32	3	32×10
1D-Conv+Linear	3	1	3×10

表 1: 提案ネットワークの構造. 入力次元数 D はストリーム毎に異なり，位置スケールストリームでは 3，自己運動ストリームでは 6，姿勢ストリームでは 36 である.

い軌跡片 (tracklets) を得た．次に，軌跡片の集合について 1) ある軌跡片の終了時の検出と別の軌跡片の開始時の検出の視覚特徴が類似しており，2) 軌跡片同士が時空間的に十分に近い場合に統合する処理を行った．視覚類似度の算出には Faster R-CNN[14] の中間層の特徴ベクトル同士のコサイン類似度を用いた．

最終的に得られた各軌跡片について，OpenPose[2] を用いて抽出した 18 点，各 2 次元の身体部位検出をそのまま使用し，各フレームの検出について $\boldsymbol{p}_t \in \mathbb{R}^{36}$ $(V = 18)$ を得た．位置及びスケールは共に身体部位より計算され，位置 $\boldsymbol{l}_t$ は左右の腰の中間地点として，スケール s_t は首の位置と左右の腰の間の距離とした．$\boldsymbol{e}_t$ の算出には教師なしカメラ姿勢推定器 [17] を使用した．以上の処理を経て，様々な長さを持つ約 5,000 の軌跡片を得た．これを固定長のサンプルに変換するため，スライディングウィンドウを用いて固定長の系列を軌跡片から 2 フレーム間隔で取り出し，最終的に約 50,000 サンプルを生成した．全検出 (約 830,000 検出) のうち，十分な長さを追跡し軌跡片の生成に実際に使われたのは約 200,000 検出 (24.1%) であった．

3.2 提案手法の実装

提案手法の詳細なネットワーク構造を表 1 に示した．各ストリームは $D \times 10$ 次元の行列を入力として受け取り，4 つの畳み込み層 (各畳み込み層には Batch Normalization (BN) [6] と ReLU 関数 [12] が続く) を通る．続いて，得られた 128×2 次元の特徴量はチャンネル方向に結合され，出力ストリームへの入力とした．出力ストリームは 2 つの BN と ReLU 関数が続く 1×1 畳み込み層，4 つの BN と ReLU 関数が続く逆畳み込み層及び恒等活性化関数を持つ単一の 1×1 畳み込み層からなり，最終的に 3×10 次元の行列を予測として出力する構造となった．

ネットワークの学習にあたって，X_{in}, X_{out} をそれぞれ平均 0，分散 1 を持つように正規化した．P_{in} についても位置及びスケール不変性を持たせるため，各要素の平均と大きさを左右の腰の中心及びスケールによってそれぞれ正規化した．また，訓練時の前処理としてランダムにサンプルを水平反転した．損失関数には平均二乗誤差 (MSE) 関数を使用した．最適化には Adam[7] を使用し，ミニバッチの大きさを 64 として 17,000 回更新を行った．学習率は 0.001 を初期値とし，5,000 回毎に値を半分とした．

3.3 比較手法

一人称視点映像における人物位置予測に関する先行研究はこれまで存在しなかったため，本研究では次の 3 手法を比較手法として選定した：

- **ConstVel**: 観測時の平均速度・方向に従って等速直線運動する．具体的には，X_{in} の平均速度及び方向を計算し，$t_0 + T_{\text{future}}$ フレーム目にどの位置に移動するかを計算した．
- **NNeighbor**: テスト系列が与えられた時，訓練系列からその $L2$ 距離が小さい k 種類の系列を抽出し，その出力系列の平均をテスト系列に対する予測とした．近傍数 k は 16 とした．
- **Social LSTM** [**1**]: 固定視点映像に対する最新の手法の 1 つとして採用した．ただし，一人称視点映像に対して性能を確保するためにモデル構造を若干改変し，入力と出力には提案手法と同様スケール情報を追加した．また，ガウス分布に基づく出力はしばしば失敗したため，X_{out} を直接予測するものとした．近傍サイズ N_o は 256 とした．

3.4 評価方法

テスト誤差の評価には 5 分割交差検証を用いた．データセットの分割は同一映像が別々のスプリットに入らないよう配慮した．単一のスプリットあたりの学習所要時間は 1 枚の NVIDIA TITAN X を用いて約 1.5 時間であった．さらに，より詳細な評価のため，収集したサンプルを対象人物と装着者との関係性の種類に従って 3 種類に分割した：

- **toward**：装着者と対象人物が相対する
- **away**: 装着者と対象人物が同一の方向に動く
- **across**：装着者の前を対象人物が横切る

評価指標には，人物位置予測の先行研究 [1] に倣い最終予測誤差 (FDE, final displacement error) を採用した．FDE は最も遠い予測である $\boldsymbol{l}_{t_0+T_{future}}$ とその正解値との L2 距離として定義される．

Method	Relation type			
	Toward	Away	Across	Average
ConstVel	178.96	98.54	121.60	107.15
NNeighbor	165.78	89.81	123.83	98.38
Social LSTM[1]	173.02	111.24	148.83	118.10
Ours	**109.03**	**75.56**	**93.10**	**77.26**

表 2: 提案手法及び各比較手法における平均最終予測誤差 (平均 FDE) の比較．単位は画素．各フレームの大きさは全て 1280 × 960 画素であった．各列は **Toward, Away, Across** の各条件に属するサンプルに関する平均 FDE 及びテストデータ全体に対する平均 FDE である．

Features	Relation type			
	Toward	Away	Across	Average
L_{in}	147.23	80.90	104.85	88.16
X_{in}	126.64	79.09	102.98	81.86
$X_{\text{in}} + E_{\text{in}}$	122.16	76.67	99.39	79.09
$X_{\text{in}} + P_{\text{in}}$	113.33	78.55	100.33	80.57
Ours ($X_{\text{in}} + E_{\text{in}} + P_{\text{in}}$)	**109.03**	**75.56**	**93.10**	**77.26**

表 3: 提案手法から一部の特徴量を取り去った場合の平均最終予測誤差 (平均 FDE) の比較．各列の数字の意味は表 2 に同じ．

3.5 実験結果

定量的評価

表 2 に FPL データセットにおける平均最終予測誤差 (FDE) の比較を示す．各手法，いずれの種類のサンプルについても，1 秒後の人物の位置をフレーム幅の 15%以下に収まる程度の誤差であった．その中で，提案手法 (**Ours**) が他の手法に対して明確な改善があることを確認した．一人称視点映像においては人物の相対的な移動方向及び速度が大きく変化することから，**ConstVel** 及び **NNeighbor** では全体的に低い性能に留まった．また，**Social LSTM** についても，我々のデータセットでは良好な性能を発揮することができなかった．これは，映像中の自己運動によって将来の位置が強く条件づけられる本タスクにおいて，適切に時系列データの依存関係をとらえることが出来なかったためであると推察される．歩行パターンの種類毎に比較した場合，対象人物が装着者に向かって近づいてくる **Toward** 条件の方が他の条件に比べ歩行パターンの潜在的な種類が多く，平均予測誤差が大きかった．

誤差解析

提案手法の性質をより詳細に把握するため，個別のサンプル毎の誤差分布の解析を行った．提案手法では，約 73%のサンプルが 100 画素 (水平視野角換算で約 10°) 以下に留まり，300 画素 (水平視野角換算で約 30°) 以上の顕著な誤差を示したサンプルはわずか 1.4%であった．

定性的評価

図 5 に提案手法及び比較手法のいくつかの予測例を示す．(a), (b), (c) はそれぞれ **Toward**, **Across**, **Away** 条件に従うサンプルである．大きな自己運動 (装着者の右折) を伴う (b) において，他の手法がいずれも誤った予測を行った中，提案手法は自己運動を補正し，概ね正しく予測を行った．また，例 (e) では装着者の方向転換に加え対象人物もゆるやかに歩行方向を変えている難しいサンプルの予測例を示したが，自己運動及び人物姿勢を考慮した提案手法では予測に成功した．

Method	Relation type			
	Toward	Away	Across	Average
ConstVel	173.75	176.76	133.32	170.71
NNeighbor	167.11	159.26	148.91	162.02
Social LSTM [1]	240.03	196.48	223.37	213.59
Ours	**131.94**	**125.48**	**112.88**	**125.42**

表 4: Social Interaction データセットにおける平均最終予測誤差 (平均 FDE) の比較．フレームの大きさは 1280 × 960 画素または 1280 × 720 画素であった．各列の数字の意味は表 2 に同じ．

各特徴量の影響

手法間の比較に加え，スケール，自己運動，姿勢の各特徴量が性能にどのように影響を及ぼすのかを検討した．具体的には，画像平面上の位置 L_{in}，位置スケール特徴量 X_{in} のみを用いる場合，位置スケール+自己運動 $X_{in} + E_{in}$，位置スケール+姿勢 $X_{in} + P_{in}$ の 4 つを比較した．結果，前者 2 つは単一のス入力ストリーム，後者 2 つは 2 つの入力ストリームを持つ構造となった．表 2 に結果を示す．比較実験の結果，各特徴量は性能向上に独立して寄与していることを確認した．特に，スケール情報の追加 ($L_{in} \rightarrow X_{in}$) 及び姿勢情報の追加 ($X_{in} \rightarrow X_{in} + P_{in}$) が **Toward** 条件の性能向上に，自己運動情報の追加 ($X_{in} \rightarrow X_{in} + E_{in}$) が **Away** 条件の性能向上に強く寄与していることが分かる．

失敗例

図 6 に失敗例を示す．入力区間が終了した直後に発生した急な方向転換の影響を受け，提案手法及び比較手法の双方が予測に失敗した．

3.6 Social Interaction データセットにおける評価

最後に，我々は公開されている一人称視点映像データセットの 1 つである Social Interaction データセット [3] に対して性能評価を行った．本データセットはテーマパーク上で撮影された複数の一人称視点映像からなり，歩行のみならず友人・店員・他の入場客との会話，行列待ち，食事などの一般的な社交シーンが含まれている．また，FPL データセットが胸部撮影であるのに対し本データセットは頭部撮影であることから，より一般的かつ難しいデータセットである．歩行とは関係のないシーンが多数含まれていたため，今回我々は手動で会話中を含む歩行シーンを抽出し，約 10,000 サンプルからなる部分データセットを構築した．

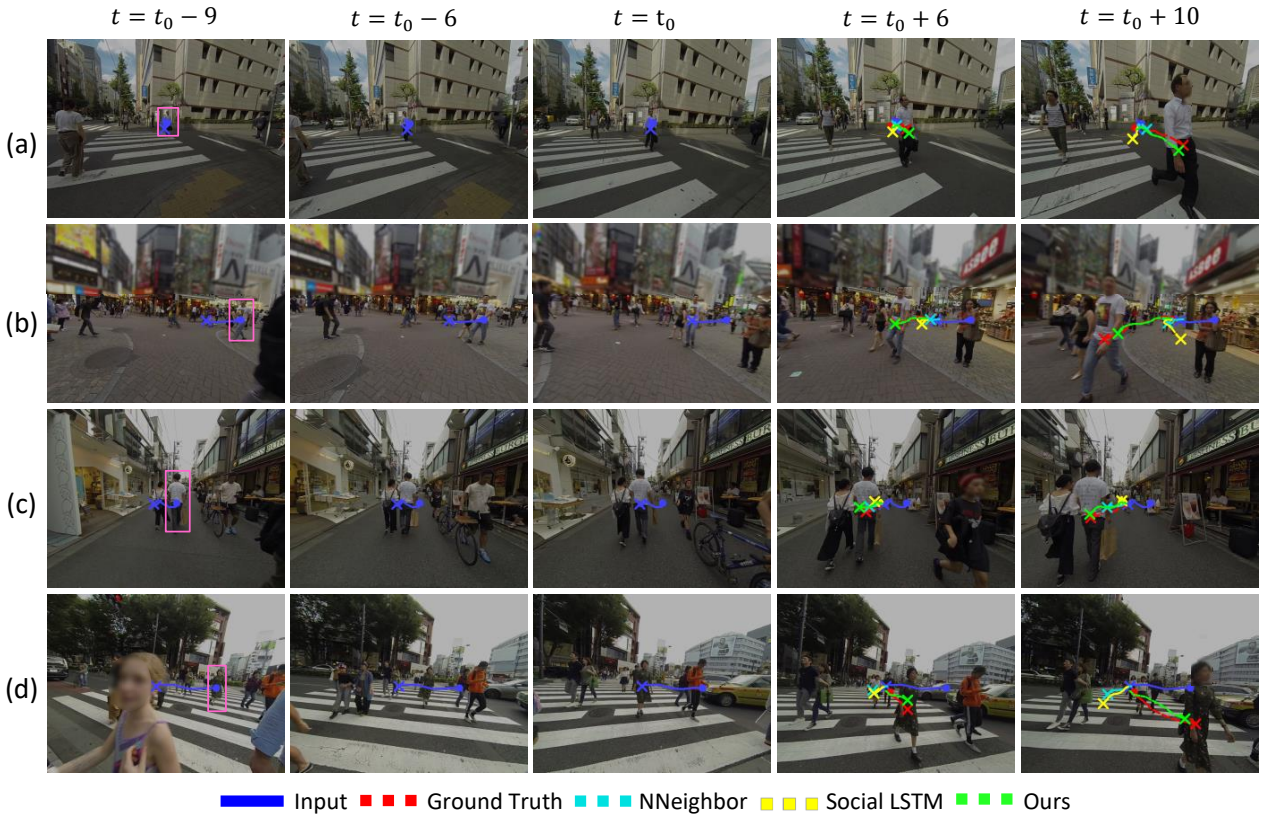

図 5: 手法間の予測結果の比較．左列の四角で囲った領域中の人物が予測対象．

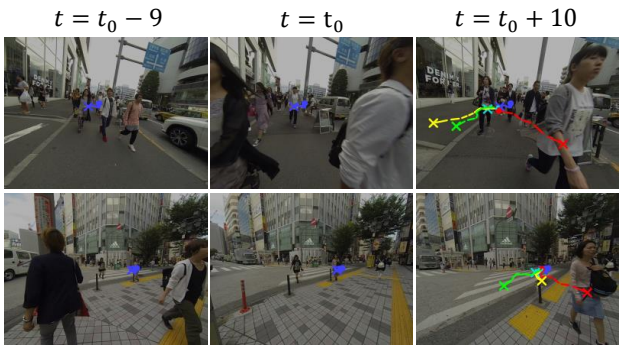

図 6: 失敗例．各図の矢印は図 5 に同じ．

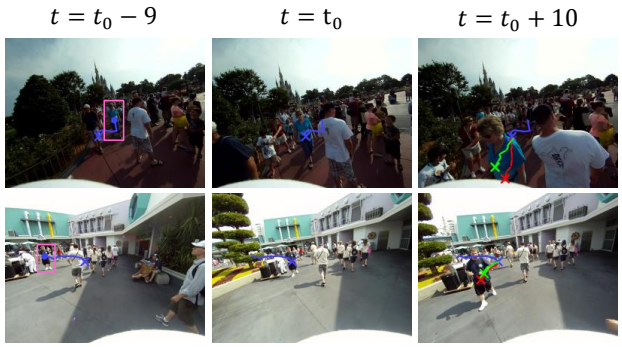

図 7: Social Interaction データセット [3] における予測例．青線が入力系列，赤線が正解，緑線が提案手法による予測を表す．

FPL データセットと同様提案手法及び比較手法双方に関し 5 分割交差検証を用いた評価を行った．

本データセットではカメラを頭部に装着したことに由来する大きな自己運動が多数観測された．そのため，FPL データセットで使用したカメラ姿勢推定器 [17] を使用した場合有効な推定結果を得ることができなかった．そこで，2.3 節に示した特徴量に代わり，オプティカルフローを自己運動特徴量として使用した．具体的には，[5] を用いて計算したオプティカルフローを縦横 3×4 グリッドに分割し，各グリッドの平均フローを結合して計 24 次元のベクトルとした．この自己運動特徴量を使用した提案手法を用いて FPL データセットについて学習したところ 79.15 FDE (1.89px 悪化) を得た．訓練は上記の学習済みモデルから再学習する形で行い，200 回重みの更新を繰り返した．最適化には学習率 0.002 の Adam を使用した．

学習結果を表 4 に示す．性能は FPL データセット (表 2) と比較すると全体的に劣っているが，本データセットにおいても提案手法が既存手法と比べ良好な性能を示した．図 7 に提案手法の実際の予測例を示した．

4. 関連研究

一人称ビジョンにおいて位置予測タスクを取り扱った研究として，装着者自身の将来位置を予測する試み [13] があるが，一人称視点映像中に映る人物の位置予測に焦点を当

てた研究はこれまで存在しなかった．Su ら [16] は，複数のバスケットボール選手のプレイの様子を一人称視点映像として記録し，位置を含む選手の将来行動を推定する手法を提案しているが，彼らの手法は複数の一人称視点映像が存在し，シーンの 3 次元復元が可能な程度に十分な量の映像があるという前提を置いている．対照的に，我々の提案手法は単一単眼の RGB カメラの入力のみで動作し，場所場面を問わず使用できるという点で異なっている．

一般の人物位置予測の研究はこれまで盛んにおこなわれてきているが [1], [8], [9], [10], [11], [15]，いずれの手法も固定視点かつ背景が変化しないことを前提としており，一人称視点映像特有の自己運動に対応できるような設計ではない．一方，我々の手法は映像中の自己運動を予測性能向上のために有効に活用している．

5. 結論

本研究では，一人称視点映像における人物位置予測という新しいタスクを提案した．また，自作データセット及び公開データセットにおける実験を通じて装着者の自己運動及び対象人物の姿勢が一人称視点映像における位置予測に際立って寄与することを示した．今後の展開として，装着者自身の位置予測 [13] を取り入れることで，より正確な予測を実現することが考えられる．

謝辞

本研究の一部は JST CREST JPMJCR14E1 の支援を受けた．

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] Cao, Z., Simon, T., Wei, S.-E. and Sheikh, Y.: Real-time Multi-Person 2D Pose Estimation using Part Affinity Fields, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291 – 7299 (2017).

[3] Fathi, A., Hodgins, J. K. and Rehg, J. M.: Social interactions: A first-person perspective, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1226–1233 (2012).

[4] Henriques, J. F., Caseiro, R., Martins, P. and Batista, J.: High-speed tracking with kernelized correlation filters, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 3, pp. 583–596 (2015).

[5] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A. and Brox, T.: FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2462 – 2470 (2017).

[6] Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, *International Conference on Machine Learning*, pp. 448–456 (2015).

[7] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *CoRR*, Vol. abs/1412.6980 (online), available from ⟨http://arxiv.org/abs/1412.6980⟩ (2014).

[8] Kitani, K. M., Ziebart, B. D., Bagnell, J. A. and Hebert, M.: Activity forecasting, *Proceedings of the European Conference on Computer Vision*, pp. 201–214 (2012).

[9] Kooij, J. F. P., Schneider, N., Flohr, F. and Gavrila, D. M.: Context-based pedestrian path prediction, *Proceedings of the European Conference on Computer Vision*, pp. 618–633 (2014).

[10] Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H. S. and Chandraker, M.: DESIRE: Distant Future Prediction in Dynamic Scenes With Interacting Agents, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 336–345 (2017).

[11] Ma, W.-C., Huang, D.-A., Lee, N. and Kitani, K. M.: Forecasting Interactive Dynamics of Pedestrians With Fictitious Play, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 774 – 782 (2017).

[12] Nair, V. and Hinton, G. E.: Rectified Linear Units Improve Restricted Boltzmann Machines, *Proceedings of the International Conference on Machine Learning*, pp. 807–814 (2010).

[13] Park, H. S., Hwang, J.-J., Niu, Y. and Shi, J.: Egocentric Future Localization, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4697–4705 (2016).

[14] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, *Advances in Neural Information Processing Systems*, pp. 1–9 (2015).

[15] Robicquet, A., Sadeghian, A., Alahi, A. and Savarese, S.: Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes, *Proceedings of the European Conference on Computer Vision*, pp. 549–565 (2016).

[16] Su, S., Pyo Hong, J., Shi, J. and Soo Park, H.: Predicting Behaviors of Basketball Players From First Person Videos, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1501–1510 (2017).

[17] Zhou, T., Brown, M., Snavely, N. and Lowe, D. G.: Unsupervised Learning of Depth and Ego-Motion from Video, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1851 – 1860 (2017).