LSTMを用いた人流推定のための頭部トラッキング

原 佑輔[1]　内山 彰[1]　梅津 高朗[2]　東野 輝夫[1]

概要：本研究では，都市部における人流を歩道ごとに把握する目的で，深層学習による頭部検出に基づく手法を提案する．車載カメラ映像では歩行者や障害物による遮蔽が頻発するため，常に全ての歩行者を捉えられるとは限らない．そのため，連続する複数フレームでの頭部検出結果に対し，時空間的な位置関係及び画像特徴量による人物の同定を行い，映像中の歩行者の移動軌跡を推定する．歩道上に存在する歩行者の移動方向は車の進行方向に対して前方と後方に大別されるため，2 種類に分けて頭部を検出する．頭部検出では遮蔽が頻発する環境でも堅牢性の高い LSTM (Long Short-Term Memory) に基づく手法を適用する．提案手法の有効性を確認するため，実際に収集した車載カメラ映像に対し評価実験を行った．5 箇所の歩道に対し人数の平均絶対誤差率で評価を行ったところ，前方は 19.2%，後方は 10.6%となり，本手法の有効性を確認する事ができた．

1. はじめに

都市計画，安全支援，マーケティングなど，様々な目的において都市部における歩行者の分布および移動状況（人流）を把握することは重要である．例えば，把握した人流から人気のあるスポットを検出したり，混雑状況の監視・予測に基づく人流誘導を行うほか，災害時の帰宅困難者の人流に応じた救援計画の立案にも活用できると考えられる．

このような人流や人々の分布状況を把握するため，これまでに様々な手法が提案されている．例えばモバイル空間統計 [1] では携帯電話の通信統計情報を用いて区画毎の人口推定を行っている．また，混雑度マップ [2] では GPS 対応の携帯電話利用者から許諾を得て送信される位置情報の分布からの人口推定を行っている．しかし，いずれも 250m メッシュなど一定範囲ごとの人密度を推定するものであり，“ある道路の西側を駅方向に歩く人数” といったスポット的な人流を把握する試みは見当たらない．一方，防犯カメラ映像を用いて混雑状況を推定する手法 [3], [4] も存在する．これらの固定カメラを用いた手法で都市部全体の人流を把握するためには，膨大な数のカメラを設置する必要があり，設置場所やコストの制約上，現実的ではない．

そこで本研究では，近年普及が進んでいる車載カメラの映像を用いた歩道などのスポットレベルでの人流推定法を提案する．様々な道路を走行している複数の車両で撮影された画像から，歩道の歩行者を検出することで人流を推定し，各車両の位置情報と共にサーバーで集約・統合することで，都市部全体における人流把握の実現を目指す．このため，本研究では，(1) Long Short-Term Memory (LSTM) に基づく頭部検出法 [5] の適用による頭部検出精度の向上，および (2) それに基づく人流推定法の設計を行った．

歩行者の検出には安全運転支援を目的とした手法 [6], [7], [8] の適用が考えられるが，歩道上にいる歩行者は車と歩道の間に存在する植え込みや柵などによって身体の一部が遮蔽される事が多いため，これらの手法をそのまま適用することは困難である．一方，Stewart ら [5] は，Recurrent Neural Network の一種である LSTM を用いて人同士の重なりが生じる場合でも精度良く人検出を行う手法を提案している．多くの既存手法では，sliding window により総当たり的に物体検出を行ったり，Selective Search などにより得られた候補領域を CNN に入力して物体検出を行う．これが物体同士の重なる場合に検出漏れが増加する原因となるが，Stewart らの手法では，候補領域の検出処理を必要とせず，入力画像全体に対して人が存在する可能性が最も高い領域を一つずつ順番に検出することで，重なりが生じる場合でも高い精度を実現している．また，入力画像全体から歩行者の頭部検出を行うため，上半身や足，腕など，頭部以外の身体の一部だけでも画像に写っている場合では，頭部のみを利用した手法よりも高い性能を発揮する．このため，提案手法では Stewart らの手法をドライブレコーダー映像に対して適用し，歩行者頭部を検出する．その際，歩道上でほとんどの歩行者は前方と後方のいずれかにのみ移動するため，歩行者の頭部を前方と後方の 2 種

[1] 大阪大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
[2] 滋賀大学 経済学部
Faculty of Economics, Shiga University

図 1　システムアーキテクチャ

類に分類することで，歩行者の検出と同時に移動方向を推定する．このため，我々は Stewart らの手法を基に方向別の頭部検出器を構築した．

提案手法では人流推定を 2 ステップに分けて行う．まず，ドライブレコーダー映像から切り出した各フレーム画像に対して，方向別に頭部検出を行う．次に，この検出結果に依然として含まれる誤検出や検出漏れに対応するため，各フレームにおける検出結果同士の時空間的な関係性を考慮する．具体的には，一定期間内の複数フレームにおける検出結果に対して，同一歩行者であるか否かを検出された領域画像の色などを特徴量に用いて判定し，映像における検出結果の移動軌跡を推定する．これらの移動軌跡が想定される歩行者の移動軌跡に近い場合に，前方または後方に移動する歩行者が存在する物と見なす．

実際に大阪市内で収集したドライブレコーダー映像を用いて提案手法の性能を評価を行った．5 箇所の歩道に対し人数の平均絶対誤差率で評価を行ったところ，前方は 19.2%，後方は 10.6%となり，本手法の有効性を確認する事ができた．

2. 提案手法

2.1 概要

提案する人流推定法のシステムアーキテクチャを図 1 に示す．車載カメラの映像は，一部の協力ユーザから提供されるものとする．車載カメラとしては，ダッシュボードにマウントされたスマートフォンや一般のドライブレコーダーを想定している．ドライブレコーダーの中にはスマートフォンや車載器などと WiFi により接続できる製品が存在する．したがって，携帯通信網により外部ネットワークに接続されたスマートフォンや車載器をゲートウェイとすることで，ドライブレコーダーの映像をサーバーに送信できる．ただし，通信量をできるだけ抑えることが望ましいため，本研究ではドライブレコーダーで取得した映像に対して，スマートフォンや車載器で処理を行い，方向別の歩行者数を推定した後，その結果のみをサーバーに送信することを想定している．方向別歩行者数の推定結果は歩道の位置と撮影時刻とともにサーバーに送信される．サーバーでは，複数車両から送られてきた各地の歩道における人流を統合し，地図上にマッピングする．

歩道の位置は GPS などにより得られた車両位置と映像内の歩行者の検出位置（左側または右側）から容易に把握可能である．ただし映像内の歩行者が小さいと検出に十分な特徴量が得られないため，本研究では，走行車線に近い側の歩道（日本では左側）でかつ車両から一定距離内のみを推定対象とし，車両が歩道に最も近い車線を走行している時のみ，人流の推定を行うものとした．

図 2 に示すように提案手法では，人流推定を (1) フレームごとの方向別頭部検出，(2) フレーム間の時空間的関係性を考慮した人流推定の 2 ステップに分けて行う．ステップ (1) では，車載カメラにより撮影された映像の 1 フレーム（静止画）において，方向別の歩行者頭部位置の推定を行う．これは画像内の物体の場所とクラス（どこに何があるか）を決定することに等しく，画像処理の分野では Localization and Classification と呼ばれる問題である．これに対して，本研究では文献 [5] の手法を適用し，前方，後方の方向別頭部検出器を構築する．ステップ (2) では，ステップ (1) の検出結果における検出漏れや誤検出の影響を軽減しながら人流推定を行うため，時間的に連続する複数フレームにおける検出結果の位置関係や画像特徴量などの類似度に基づき，移動方向別の歩行者人数を推定する．

なお，本研究では，歩道上の歩行者は前方・後方のいずれかのみに移動するものとし，検出対象となる頭の方向を前方・後方の 2 種類に限定する．実際には交差点などにおいて道路側を向いて立ち止まっている人なども存在するが，これらの判別は困難なため，本研究では交差点付近を検出の対象外としている．

2.2 深層学習による方向別頭部検出

提案手法では，前方，後方の 2 種類の方向別頭部検出を行うため，Stewart らが提案した人検出法 [5] を適用する．図 3 に頭部検出の概要を示す．まず，縦 480 ピクセル，横 640 ピクセルの入力画像を GoogLeNet に与えることで，20×15 のセルそれぞれにおける 1024 次元のベクトルを特徴量として得る．各セルのベクトルは画像中の対応領域における特徴を要約しており，物体の位置情報も含んでいると考えられる．この各セルの特徴量を LSTM に入力することで，検出対象（頭部前方または後方）が存在す

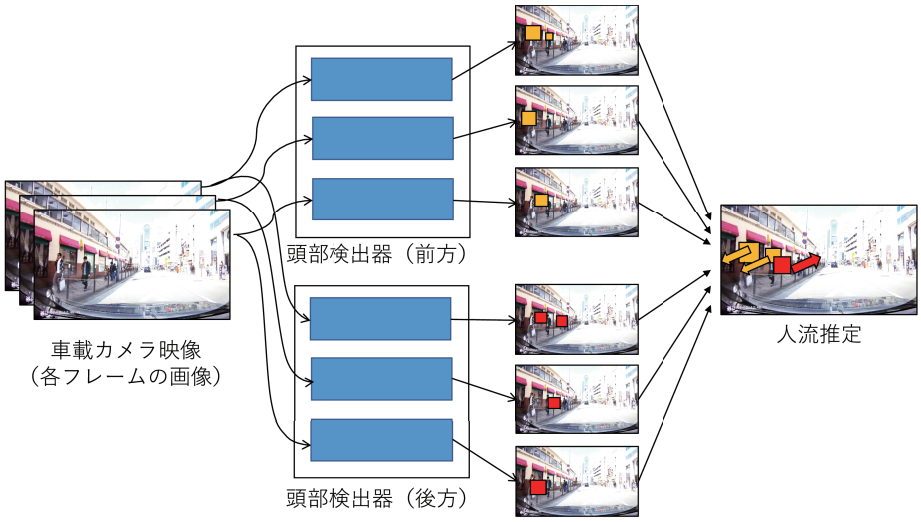

図 2　人流推定法の概要

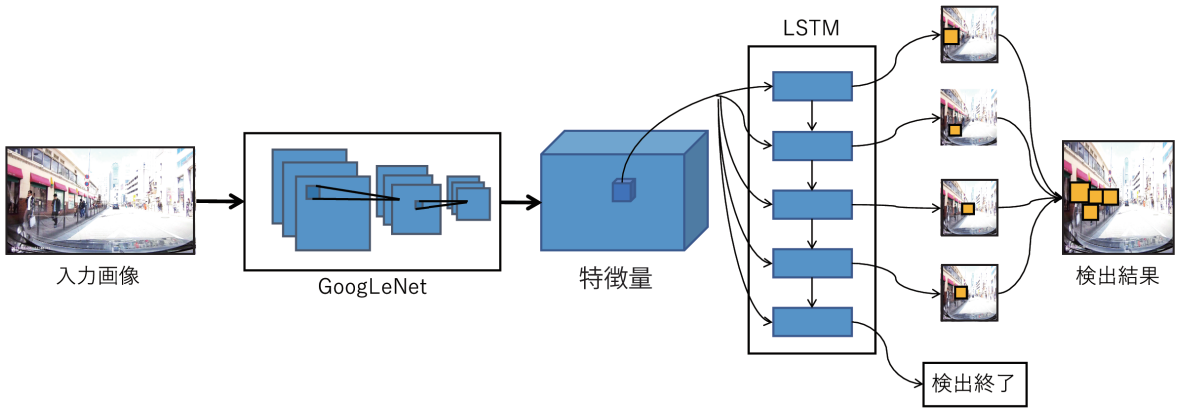

図 3　深層学習による方向別頭部検出の概要

る位置を bounding box として検出信頼度とともに出力する．LSTM では信頼度が高い bounding box から順に出力がなされる．この時，直前の出力結果を次のユニットに入力することにより，同一対象の重複検出が起こらないようにしている．これを信頼度が閾値 T 以上の bounding box が見つからなくなるまで繰り返す．最終的に，得られた複数の検出結果を統合することで，一つの入力画像に対する検出結果が得られる．学習は，文献 [5] で提案されている損失関数に従って行うものとした．

2.3　人流推定

複数フレームにおける方向別の頭部検出結果を統合し，歩行者 1 人 1 人の移動軌跡を推定することによって，人流推定を行う．各フレームにおける頭部検出結果は，その瞬間の画像特徴量のみを用いているため，誤検出や検出漏れが避けられない．そこで，頭部検出結果の時空間的な特徴や画像特徴を考慮することによって，誤検出や検出漏れに対する堅牢性の向上を図る．

フレーム t において検出された i 番目の bounding box を $b_i^t \in B^t$ とする．B^t はフレーム t で検出された bounding box の集合である．また，b_i^t により検出された人物の ID を $y(b_i^t)$ とする．提案手法では，b_i^t と b_j^u が同一人物（$y(b_i^t) = y(b_j^u)$）であるか否かを判定するため，類似度 $\mathrm{sim}(b_i^t, b_j^u)$ を以下のように定義する．

$$\mathrm{sim}(b_i^t, b_j^u) = w_1 l(b_i^t, b_j^u) + w_2 v(b_i^t, b_j^u) \quad (1)$$

ここで，$l(b_i^t, b_j^u), v(b_i^t, b_j^u)$ はそれぞれ b_i^t, b_j^u の位置関係，画像特徴量に基づき定義される類似度であり，w_1, w_2 は重みである．

位置関係に基づく類似度は，以下のように定義する．まず，時刻 t における頭部検出位置に対して，人および車両の移動速度を考慮して時間的に前後するフレームでの頭部位置を予測する．予測位置と実際に検出された頭部位置の距離に反比例した値を位置類似度とする．本来は GPS 情報から車速を得て推定移動座標を求める必要があるが，今回は簡単のため車速が一定であるものとし，映像から得られた固定の 1 フレームあたりでの移動座標を用いて予測位置を求めた．

画像特徴量に基づく類似度は，歩行者の服の色に着目し

図 4　実験で走行した道路 (大阪市茶屋町周辺)

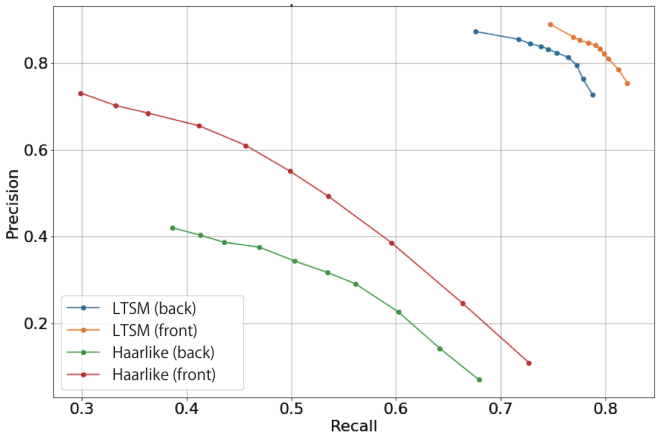

図 5　閾値の影響（Precision と Recall）

て以下のように定める．服の画像領域は，検出された頭部のウィンドウに対し，縦方向のウィンドウサイズを地面の方向に 3 倍した領域を対象とする．今回は画像をグレースケールに変換した後に色ヒストグラムを計算する．基準となる頭部の色ヒストグラムと，対象の頭部の色ヒストグラムの相関を計算することで，類似するヒストグラムであれば大きな値を出力するようになる．

同一人物の判定および移動軌跡の推定は以下の手順で行う．まず，類似度の定義に基づき，全ての隣接するフレーム間において，同一人物の判定を行う．具体的には，フレーム $t, t+1$ 間の bounding box ペアのうち，$\mathrm{sim}(b_i^t, b_j^{t+1})$ が最大のペア $(b_i^t, b_j^{t+1}) \in B^t \times B^{t+1}$ について，$\mathrm{sim}(b_i^t, b_j^{t+1})$ が閾値 S 以上であれば $y(b_i^t) = y(b_j^{t+1})$ と見なし，b_i^t, b_j^{t+1} をそれぞれ B^t, B^{t+1} から除外する．これを閾値を越えるペアが無くなるまで繰り返す．同様にして，フレーム t と $t+2, t+3, \ldots, t+W$ 間の bounding box ペアに対して，フレーム間隔を 1 ずつ増やしながら順に同一人物の判定を行う．これによって，ある程度の検出漏れを許容した同一人物判定を実現している．

以上のようにして各フレーム組の間で同一人物判定を行った後，

$$y(b_i^t) = y(b_j^u) \wedge y(b_j^u) = y(b_k^v)$$

であれば，bounding box b_i^t, b_j^u, b_k^v は全て同一人物の移動軌跡として推定する．最後に，同一人物と推定されたものが検出されたフレームが W 以下であったものは除外する．これにより誤検出を除外することが可能と考えられる．以上により歩行者一人一人の移動軌跡が推定されるため，歩道における移動方向別の歩行者数が得られる．

3. 性能評価

3.1 実験環境

提案手法の性能評価を行うため，大阪市茶屋町周辺の道路をドライブレコーダ (ユピテル社製 DRY-WiFiV5c) を設置した自動車で複数回通行し，映像を撮影した．撮影は，図 4 の地点 1 から地点 2 間の水色で示されている道路において，多くの人が行き交う休日の正午頃に行った．学習用データは前方の画像が 2560 枚， 後方の画像が 2695 枚である．

評価用データは頭部検出の評価には車載カメラ画像 100 枚，人流推定の評価には 5 箇所の歩道の動画を用いた．5 箇所の歩道の歩行者数は前方 93 人，後方 57 人である．評価指標には Precision, Recall, Accuracy を用いた．また，方向別頭部検出の比較対象として，Open CV により実装した Haarlike 特徴量を用いて学習用データから頭部を切り出して学習させて作成した検出器の場合との比較を行った．

学習を行う際のパラメータは文献 [5] に記載されている設定を用い，学習用ライブラリ及び学習ソフトは筆者らが公開しているものを用いた．式 (1) の頭部位置予測で用いるパラメータは $w_1 = w_2 = 1$ とした．学習に用いたワークステーションのスペックは CPU が Intel(R) Xeon(R) CPU E5-1680 v3 @ 3.20GHz，メモリ 128GB，GPU は GeForce GTX 1080 である．

3.2 方向別頭部検出の評価結果

図 5 に頭部検出における LSTM の閾値 T を変化させた時の Precision と Recall を示す．Haarlike 特徴量を用いた場合，Precision, Recall はそれぞれ 0.1〜0.7，0.7〜0.3 程度となっており，Precision が最も高い 0.7 程度の場合でも Recall が 0.3 程度まで低下しているため，十分な性能が出ているとは言い難い．一方で，提案手法の Precision, Recall はそれぞれ 0.7〜0.85，0.85〜0.68 程度に収まっている．これは前方，後方どちらの場合も共通であり，Haarlike 特徴量を用いた場合と比べて，提案手法が Precision, Recall ともに大きく上回っていることが分かる．このような結果となった理由は，人同士の重なりが頻発する場合においても，提案手法により誤検出や検出漏れを抑えることができているためと考えられる．

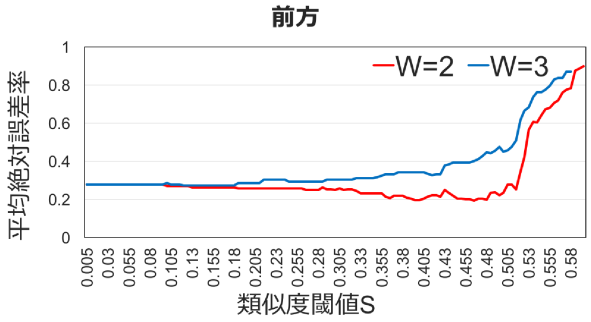

図 6　人流の誤差 (前方)

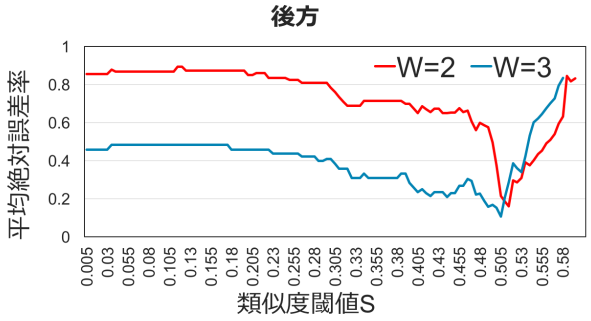

図 7　人流の誤差 (後方)

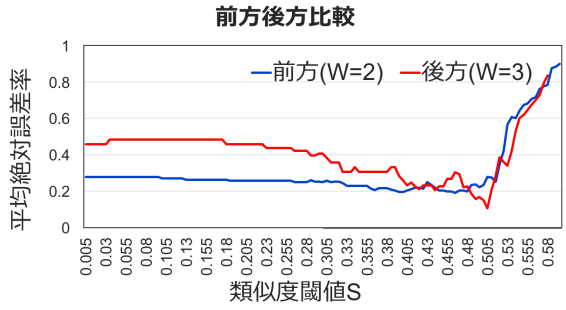

図 8　人流の誤差 (前方後方比較)

一方，Haarlike 特徴量と提案手法のどちらの場合でも，後方の性能は前方よりも低下している．この原因として，前方の場合は目や鼻，口など様々な顔の部位が画像に表れるため，検出に有益な特徴量が得られやすいが，後方の場合は髪の毛で頭部が覆われてしまい，画像から十分な特徴量を得られにくいことが考えられる．それでもなお，提案手法は方向別頭部検出において高い Precision, Recall を達成しており，その有効性が確認された．

3.3　移動軌跡推定

前節の評価で得られたフレームごとの方向別頭部検出結果に対して，移動軌跡推定を行い，人流の推定を行った．交差点から交差点までの歩道 5 箇所についての人数の絶対誤差率の平均で評価を行った．類似度の閾値 S と最小の検出フレーム W を変化させて，方向別に結果をプロットした．S は大きくなるほど同一人物の判定が厳しくなるパラメータであり，パラメータ W は大きくなるほど 1 人とカウントするのに多くのフレームで検出する必要がある．

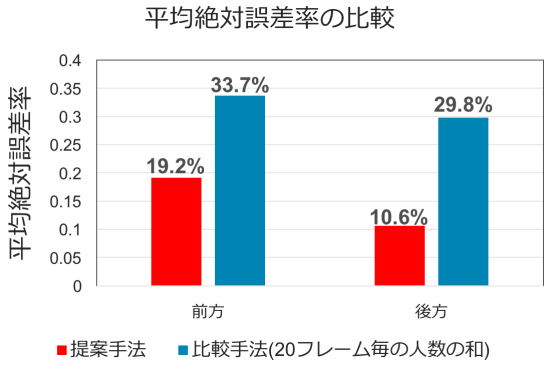

図 9　ナイーブ手法との比較

図 6 に前方の結果を示す．W が増加するほど平均誤差率が大きくなっていることが分かる．これは，前節の頭部検出の結果から前方の頭部検出は後方と比べ誤検出が少ないため，W を大きくすることで真値である歩行者も除外していると考えられる．

図 7 に後方の結果を示す．こちらは W が増加するほど平均誤差率が小さくなっていることが分かる．これは，後方の頭部検出は前方と比べ誤検出が多いので W を大きくすることで誤検出が除外できていると考えられる．

図 8 に前方と後方の比較結果を示す．類似度閾値 S が小さいときは前方の方が後方と比べ，平均誤差率が小さくなっている．これは，S が小さい場合は誤検出であっても同一人物であると紐付けをする場合が多くなり，結果として検出器の性能差が人流の推定性能に表れていると考えられる．

図 9 にナイーブ手法との比較を示す．ナイーブ手法として，ダブルカウントが起こらないフレーム間隔での頭部検出結果の人数の和を取るといったものを用いた．フレーム間隔は 20 フレームとした．この結果から提案手法の方が平均誤差率が小さくなっており，移動軌跡を推定することで誤検出，検出漏れが補正されていることがわかり，本手法の有効性が確認された．

4.　おわりに

本研究では，街中を走行する車両の車載カメラ映像を利用した歩道レベルでの人流推定法を提案した．車載カメラ映像では歩行者や障害物による遮蔽が頻発するため，常に全ての歩行者を捉えられるとは限らない．そのため，連続する複数フレームでの頭部検出結果に対し，時空間的な位置関係及び画像特徴量による人物の同定を行い，映像中の歩行者の移動軌跡を推定する．歩道上に存在する歩行者の移動方向は車の進行方向に対して前方と後方に大別されるため，2 種類に分けて頭部を検出する．頭部検出では遮蔽が

頻発する環境でも堅牢性の高い LSTM (Long Short-Term Memory) に基づく手法を適用する．提案手法の有効性を確認するため，実際に収集した車載カメラ映像に対し評価実験を行った．また，その結果を時系列的に処理を行い，検出位置の特徴と服の色の特徴を用いて同一人物判定を行い，人物の移動軌跡推定を行い歩道の人流推定を行った．5 箇所の歩道に対し人数の平均絶対誤差率で評価を多なったところ，前方は 19.2%，後方は 10.6%となり，本手法の有効性を確認する事ができた．

今後，様々な場所におけるデータに対する評価や，車速が変化した場合でも対応できるように改良を進めていく予定である．

謝辞

本研究は JSPS 科研費 JP26220001，JP26700006，JP16K00123 の助成を受けたものです．

参考文献

[1] 寺田雅之，永田智大，小林基成：モバイル空間統計における人口推計技術 (社会・産業の発展を支える 「モバイル空間統計」：モバイルネットワークの統計情報に基づく人口推計技術とその活用)，NTT DoCoMo テクニカル・ジャーナル， Vol. 20, No. 3, pp. 11–16 (2012).

[2] 株式会社ゼンリンデータコム：混雑度マップ，http://lab.its-mo.com/densitymap/.

[3] Silveira Jacques Junior, J., Musse, S. and Jung, C.: Crowd Analysis Using Computer Vision Techniques, *IEEE Signal Processing Magazine*, Vol. 27, No. 5, pp. 66 – 77 (2010).

[4] Wu, Z., Thangali, A., Sclaroff, S. and Betke, M.: Coupling detection and data association for multiple object tracking, *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1948–1955 (2012).

[5] Stewart, R., Andriluka, M. and Ng, A. Y.: End-To-End People Detection in Crowded Scenes, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

[6] Wöhler, C., Anlauf, J. K., Pörtner, T. and Franke, U.: A Time Delay Neural Network Algorithm for Real-Time Pedestrian Recognition, *Proceedings of International Conference on intelligent vehicle*, pp. 247–251 (1998).

[7] Papageorgiou, C., Evgeniou, T. and Poggio, T.: A Trainable Pedestrian Detection System, *Proceedings of Intelligent Vehicles*, pp. 241–246 (1998).

[8] Lee, K.-H., Hwang, J. N., Okapal, G. and Pitton, J.: Driving recorder based on-road pedestrian tracking using visual SLAM and Constrained Multiple-Kernel, *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 2629–2635 (2014).