Vol.2016-CVIM-202 No.30
2016/5/12

時空間限定Dense Trajectoriesによるジェスチャ認識

山田 花穂[1]　吉田 武史[1]　鷲見 和彦[1]　波部 斉[2]　満上 育久[3]

概要：本研究では，集団検出を最終目的としたジェスチャ認識手法を提案する．近年，全身像からジェスチャを認識する手法として，Dense Trajectories 特徴が注目されているが，似ている動きの識別などに課題があった．そこで本研究では，身体のパーツ検出を拡張してジェスチャ認識領域を設定し，時間軸を一定の間隔で分割した各区間ごとに特徴点を追跡することにより，限定した時空間領域に対して特徴点軌跡を取得し，Dense Trajectories を用いたジェスチャ認識手法を高精度化した．評価実験より，時空間限定を行わない場合と比較して，指差し・うなずきなどのジェスチャ認識の精度が 20%以上向上することを確認した．

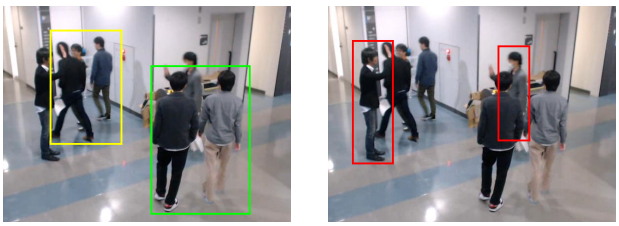
1. はじめに

近年，駅や街角などの公共空間において対象となる人物に応じた情報を提供する技術が普及している．情報提供を制御するための属性として，個人向けのものでは年齢・性別などが用いられ，JR に設置されている顧客に合わせて商品を勧める自動販売機等で実用化されている．集団向けのものでは個人の属性に加えて，家族・同僚のような集団の関係性を推定する必要があり，そのための集団検出及び関係性の推定技術は実用化に至っていない．そこで，関係性を推定するためには集団検出の高精度化が必要となる．

集団検出は従来から盛んに研究されている [1, 2]．Chamveha らは，歩行者ペアの相対的な位置・頭の向きなどを特徴量とした Decision Tree を用いて，集団を検出する手法を提案している [3]．しかし，同手法では人物間距離の影響が支配的であるため，集団検出において誤検出を招いてしまう場合が多く存在する．図 1 は同手法による集団検出の結果である．図 1(a) のように人物間の距離が近い 2 つの集団は正しく検出できる．一方，図 1(b) のように距離が遠い集団は正しく検出できない．ところが，互いに手をふり合うジェスチャに着目すると，集団と検出できる可能性が高い．従って，他人とのインタラクションを特徴付けるジェスチャ認識を新たな特徴量として加えることにより，人物間距離に依存しない集団検出が実現すると考えられる．

ジェスチャ認識は様々な切り口から研究が行われている [4–6]．その中でも人物領域周辺から抽出される特徴点軌跡を用いる手法は，その人物のジェスチャ認識に有効な情報を多く含む．そのため，特徴点軌跡を用いるジェスチャ認識手法はこれまでに数多く提案されている [6–8]．しかし，これらの手法は動画像の全領域及び全時間から特徴点を追跡しているものが多く，似ている動きの識別などに課題がある．

そこで本研究では，身体のパーツ検出を拡張してジェスチャ認識領域を設定し，時間軸を一定の間隔で分割した各区間ごとに特徴点を追跡することにより，限定した時空間領域に対してジェスチャを認識する手法を提案する．



(a) 正しく検出できる集団　(b) 正しく検出できない集団

図 1: 従来手法による集団検出の結果

2. 特徴点軌跡を用いるジェスチャ認識手法

人物領域周辺から抽出される特徴点軌跡には，人物行動に関する有効な情報が多く含まれている．そのため特徴点軌跡を用いるジェスチャ認識手法はこれまでに多く提案されている．例えば，KLT 法 [9] や Dense Trajectories [8] から特徴点軌跡を取得するジェスチャ認識手法がある．

2.1 KLT 法を用いるジェスチャ認識

高橋らは KLT 法 [9] によって取得される特徴点軌跡から固定次元の方向ヒストグラムを作成することで動作時間長

1 青山学院大学
Aoyama Gakuin University
2 近畿大学
Kinki University
3 大阪大学
Osaka University

© 2016 Information Processing Society of Japan

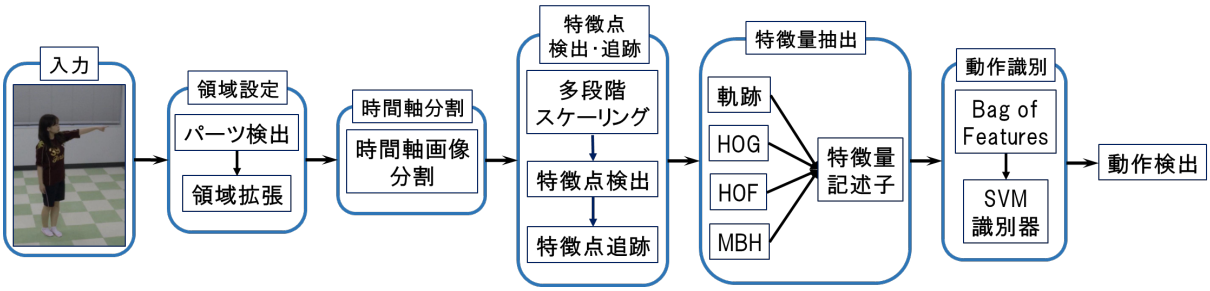


図 2: 提案手法の流れ

に依存しないジェスチャ識別を可能にした．これにより混雑映像でも「指をさす」,「抱擁する」などのジェスチャを検出可能にする手法を提案している [7]．しかし，同手法では対象人物の行動に回転を含む場合や，照明変化による輝度値の変化が激しい場合，特徴点の追跡に失敗することがある．

2.2 Dense Trajectories を用いるジェスチャ認識

Wang らは，Dense Trajectories を用いて密な特徴点の追跡を行い，得られた軌跡とその周辺領域の特徴量から BoF(Bag of Features) [10] と SVM(Support Vector Machine) [11] により，「手を叩く」,「握手する」などのジェスチャを認識する手法を提案している [8]．同手法では，他の特徴点軌跡を用いる手法よりも密な特徴点の追跡ができるため，微小な動きも捉えることが可能になる．

2.3 従来のジェスチャ認識手法の問題点

特徴点軌跡を用いる従来のジェスチャ認識手法は，動画像の全領域及び全時間から特徴点を追跡している．しかし，ジェスチャは「歩きながら指をさす」のように歩行などの動作と同時に発生するため，全領域から特徴点軌跡を取得すると，足などのパーツの軌跡がジェスチャを特徴的に表す身体のパーツの軌跡の妨げとなる．また「手をふる」と「指をさす」のように似ている動きを含むジェスチャが存在するため，全時間から軌跡を取得すると，互いの特徴量が近似したものになる．これらの問題は，他のジェスチャとの誤認識を高める原因となる．そこで本研究では，身体のパーツ検出を拡張してジェスチャ認識領域を設定し，時間軸を一定の間隔で分割した各区間ごとに特徴点を追跡することにより，限定した時空間領域に対して特徴点軌跡の取得を可能にする．

3. 提案手法

本研究では，身体のパーツ検出を拡張してジェスチャ認識領域を設定し，時間軸を一定の間隔で分割した各区間ごとに特徴点を追跡することにより，限定した時空間領域に対して特徴点軌跡を取得し，Dense Trajectories を用いたジェスチャ認識手法を提案する (図 2)．これにより，「指をさす」であれば腕領域，「うなずく」であれば頭領域のようにジェスチャごとに必要な領域を限定して特徴点の追跡を行うため，他の身体のパーツ領域の動きに左右されない特徴点軌跡の取得が可能となる．また似ている動きが含まれている場合でも，動きの違う箇所に限定した特徴量抽出ができるように時間を分割すれば，他のジェスチャとの誤認識が防げる．従って，Dense Trajectories によるジェスチャ認識手法の精度の向上が見込まれる．

3.1 認識するジェスチャの種類

認識するジェスチャは，2013 年に開催された青山学院大学相模原祭で撮影された 30 分程度の映像から，集団が行っていると判断できるジェスチャの回数が多いものを選択する．この結果を，表 1 に示す．同表より，回数が 2 桁以上となった「指をさす」,「うなずく」,「手をふる」の 3 種類とする．各ジェスチャの映像の長さは「指をさす」が「手を挙げ始めてから指をさすまでの間」,「うなずく」が「頭を上下に 2 回振る間」,「手をふる」が「手を左右に 2 回振る間」に設定する．

表 1: 集団が行っているジェスチャとその回数

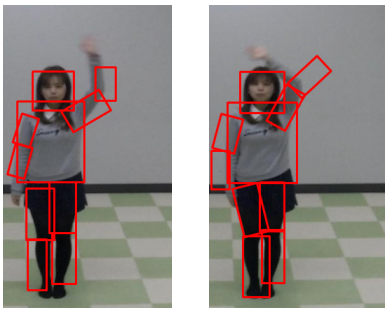
集団が行っているジェスチャ	個数
指をさす	40
うなずく	33
手をふる	11
身体に触れる	4
手を叩く	2
手を繋ぐ	2
肩を組む	1
その他	25

3.2 領域設定

領域設定処理では，各画像から頭・胴体・腕・足のパーツごとのジェスチャ認識領域を設定する．この時，腕は左右で距離があるため右腕と左腕のパーツ領域を別々に設定し，足は左右の距離が近いため両足で 1 つのパーツ領域を設定する．まず身体のパーツ範囲を検出するため，Rothrock らの and-or グラフ文法モデルと背景分割モデルを用いたパー

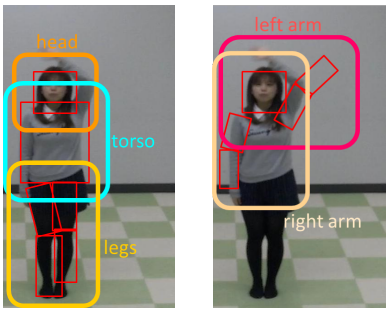
© 2016 Information Processing Society of Japan

ツ検出手法を使用する [12]．同手法によるパーツ検出の結果を図 3 に示す．これより，頭・胴体・足の検出率は高いが，腕の検出率は低いこと，また各パーツの検出範囲が実際のパーツ範囲より狭くなり，軌跡を捉えるには十分な領域とは言えないことが分かる．そこで本研究では，頭・胴体・足には図 4(a) のようにパーツ検出から得られた範囲に外接する矩形を 30%拡張した領域を設定する．両腕には図 4(b) のように頭と腕のパーツの検出範囲に外接する矩形を 30%拡張した領域を設定する．この拡張規模は，パーツ範囲を 10%から 60%の間で 5%ずつ拡張したジェスチャ認識領域を設定し，Dense Trajectories を用いてジェスチャを認識したときの全体の正解率を比較して選択した．このときの正解率の結果を図 5 に示す．同表より，正解率が最も高くなった 30%を拡張規模とする．



(a) 成功例　(b) 失敗例

図 3: パーツ検出の結果



(a) 頭・胴体・足　(b) 右腕・左腕

図 4: 領域設定

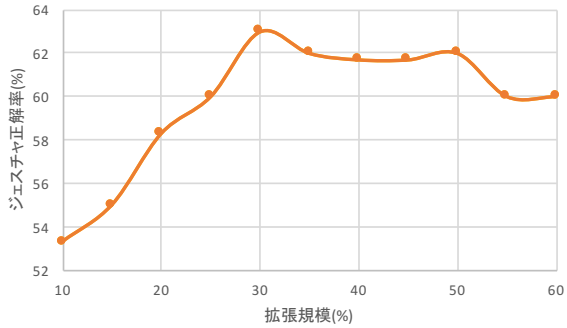


図 5: 拡張規模によるジェスチャ正解率

3.3 時間軸分割

時間軸分割処理では，図 6 のように時間軸を一定のフレーム間隔で分割する．必要となる分割数は対象ジェスチャによって変わるが，本研究では，似ている動きが多く含まれる「手をふる」と「指をさす」ジェスチャにおいて動きの違う箇所に限定した特徴量を抽出できる分割数を考える．2 つのジェスチャを比較したとき，動きに最も違いのある箇所は「指をさす」の動作中に行う「腕を挙げたまま静止している動き」である．20 人の被験者が「指をさす」動作にかかる時間を計測した．計測した結果を図 7 に示す．図中の水色が示すのは「腕を挙げるまでの時間」であり，黄色が示すのは「腕を挙げたまま静止している時間」である．また水色と黄色を合わると「指をさす」動作全体にかかる時間となる．ここから，静止時間全体に対して最終分割箇所の静止時間の占める割合が大きく，かつ最終分割箇所全体に対して最終分割箇所の静止時間の占める割合が大きくなる分割数が，「腕を挙げたまま静止している動き」に限定した特徴量を抽出できる．この 2 つの割合を分割数ごとに計算した結果を図 8 に示す．同表より，3 分割または 4 分割が適している．この 2 つの分割数によって時間軸分割し，Dense Trajectories を用いてジェスチャ認識したときの全体の正解率を比較した結果，3 分割のとき 70.0%，4 分割のとき 71.6%となった．そこで本研究では分割数を 4 分割とする．

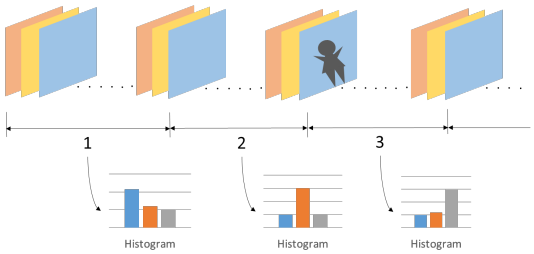


図 6: 時間軸分割処理

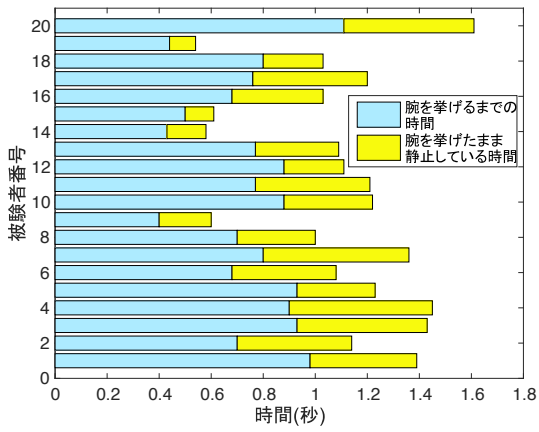


図 7: 「指をさす」の動作切り替え時間

© 2016 Information Processing Society of Japan

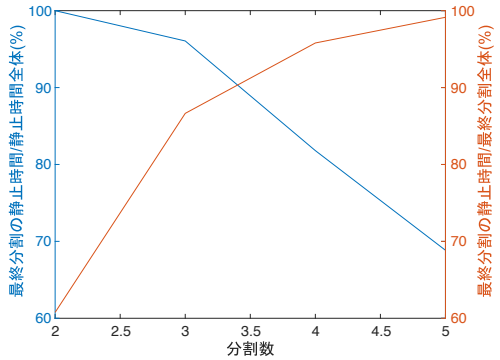
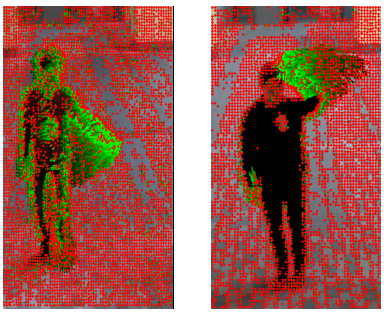


図 8: 各分割数の静止時間全体に対しての最終分割の静止時間と最終分割全体に対しての最終分割の静止時間の占める割合

3.4 特徴点検出・追跡

特徴点検出・追跡処理では，時空間で限定された領域に対して Dense Trajectories を用いて特徴点を追跡する．Dense Trajectories は，多段階スケーリング処理された画像の各スケールにおいて，Dense Sampling を用いて特徴点を検出し，Farneback Optical Flow [13] を用いたフロー抽出と，Median Filtering を用いたフローの対応付けから特徴点を追跡する．

Dense Trajectories を用いた特徴点追跡の結果を図 9 に示す．図中の赤色の点は検出した特徴点，緑色の線は特徴点を追跡した軌跡を表している．



(a) 指をさす　(b) 手をふる

図 9: Dense Trajectories を用いた特徴点追跡結果

3.5 特徴量抽出

特徴量抽出処理では，Dense Trajectories より得られた軌跡データから，軌跡，HOG(Histograms of Oriented Gradients) [14]，HOF(Histograms of Optical Flow) [15]，MBH(Motion Boundary Histograms) [16] を特徴量として抽出する．特徴量抽出方法を図 10 に示す．同図において，赤い線は追跡された動線を表している．この動線付近 32×32 ピクセルの周辺領域から 30 次元の軌跡特徴量 (動線の形状) を取得する．また，局所特徴量である HOG，HOF，MBH は図 10 のように，2(x 方向)×2(y 方向)×3(t 方向) の領域から取得した特徴量を連結することでそれぞれを抽出している．これより，96 次元 ($2 \times 2 \times 3 \times 8$) の HOG，108 次元 ($2 \times 2 \times 3 \times 9$) の HOF，$x$ 方向微分の MBHx，y 方向微分の MBHy を合わせた 192($2 \times 2 \times 3 \times 8 \times 2$) 次元の MBH を特徴量として記述する．そして，1 軌跡につきこれらの特徴量を全て組み合わせた 426 次元の特徴量を作成する．この特徴量はまず，身体の各パーツを時間軸分割した箇所から抽出する．そして，全分割箇所の特徴量を 1 つにまとめ，最終的には各パーツのジェスチャ認識領域の特徴量を作成する．

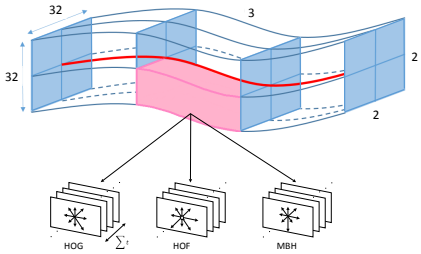


図 10: 特徴量抽出の方法

3.6 動作識別

動作識別処理では，得られた特徴量を基に BoF と SVM によりジェスチャを識別する．このとき「指をさす」と「手をふる」は腕領域，「うなずく」は頭領域がジェスチャ認識に必要となる．学習時は特徴量抽出処理によって抽出した頭領域と腕領域の特徴量を 1 つにまとめ，BoF によりベクトル化することで，各ジェスチャを識別する SVM 識別器を生成する．また評価時は，学習時と同様，頭領域と腕領域の特徴量を 1 つにまとめ，BoF によりベクトル化することで，SVM 識別器によりそのジェスチャを識別する．また，Wang らに倣って BoF のクラスタ数 k は 4,000 に設定する．

4. 実験

本手法の有効性を評価するため，従来手法 (Wang らの手法 [8]) と提案手法，従来手法に領域設定処理を加えた手法，時間軸分割処理を加えた手法の 4 つの精度を比較した．

4.1 実験条件

学習と評価に用いるデータは，図 13 のような公共空間に設置されているカメラを想定した角度・距離から 1 人につき 3 種類ジェスチャを撮影した映像を用いる．映像の撮影には GoPRO 社の HERO3+を使用し，解像度は Full HD(1920×1080) に設定した．各ジェスチャにつき学習用データに延べ 80 人，評価用データに延べ 20 人の動きの情報を使用した．学習用データは図 11 のように撮影した映像をジェスチャが見える最小の範囲で切り取った領域を用い，評価用データは図 12 のように実環境により近づけるため，映像の全域を用いることにした．

© 2016 Information Processing Society of Japan

図 11: 学習用データ

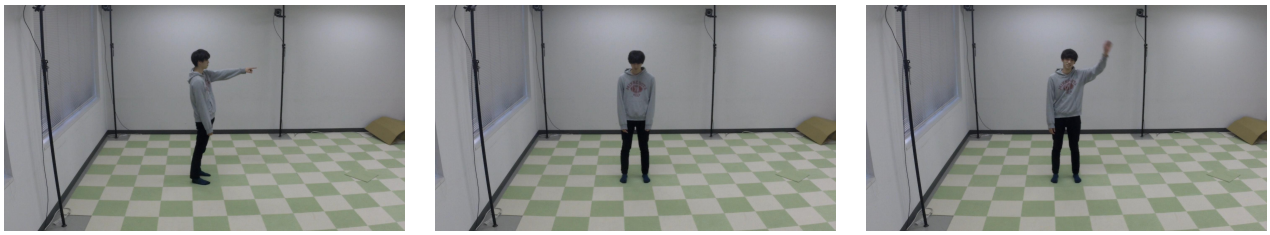


図 12: 評価用データ

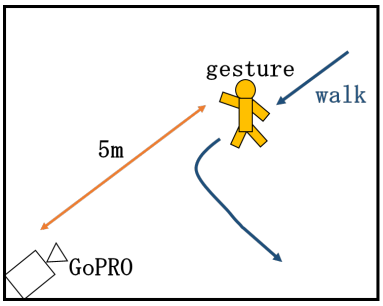


図 13: 撮影の様子

4.2 実験結果

実験結果を表 2 に示す．従来手法に領域設定処理を加えることにより「うなずく」の正解率が大きく向上したこと，時間軸分割処理を加えることにより全ジェスチャの正解率が向上したことから，全体で提案手法は従来手法より高い正解率を得た．また，「手をふる」の正解率は提案手法より従来手法に時間軸分割処理を加えた方が高くなった．

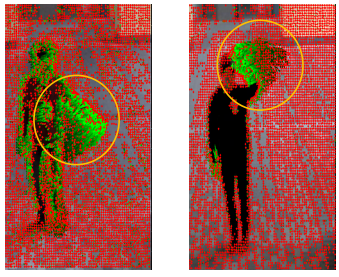
表 2: 各ジェスチャの正解率

行動	従来	領域設定のみ	時間軸分割のみ	提案
指をさす	45.0	50.0	70.0	80.0
うなずく	50.0	85.0	65.0	90.0
手をふる	65.0	55.0	80.0	70.0
全体	53.3	63.3	73.3	81.6

正解率 (%)

4.3 考察

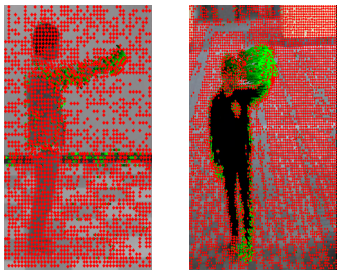
「うなずく」は他のジェスチャと比較すると動きが明らかに違うため，ジェスチャ正解率が高くなると考えられる．しかし，従来手法では頭以外の身体のパーツに動きがある場合，他のジェスチャと誤認識されることが多かった．提案手法では領域設定処理を加えることにより限定した領域から軌跡の取得が可能となり，正解率が向上した．また時間軸分割処理を加えることにより動きの違いを検出できたため，全ジェスチャの正解率が向上した．特に「指をさす」と「手をふる」は，図 14 の黄色の円で囲まれた箇所から分かるように，軌跡の似ている動きが含まれるため，従来手法では互いに誤認識することが多かった．提案手法では時間軸分割処理により，最終分割箇所に図 15 のような特徴点軌跡の違いが現れたため，正解率が向上した．さらに「手をふる」は頭と腕が重なっている場合，図 16(a) のように腕のパーツ検出だけでなく頭のパーツ検出も失敗する場合が存在する．これにより図 16(b) のように領域設定がうまくできず，他のジェスチャと誤認識されてしまった．従って提案手法よりも従来手法に時間軸分割処理を加えた方が高くなった．この解決法は，身体のパーツの重なりにも対応できるパーツ検出手法を考案することが挙げられる．



(a) 指をさす　(b) 手をふる

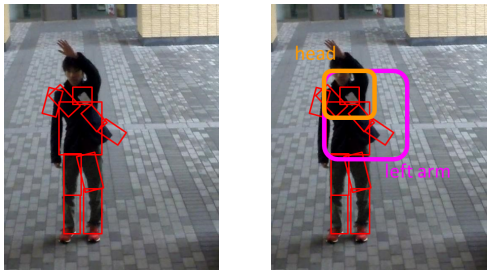
図 14: 軌跡の似ている動作例

© 2016 Information Processing Society of Japan



(a) 指をさす　(b) 手をふる

図 15: 最終分割箇所においての特徴点軌跡の違い



(a) パーツ検出失敗例　(b) 失敗例から領域設定

図 16: 領域設定の失敗例

5. おわりに

本研究では，身体のパーツ検出を拡張してジェスチャ認識領域を設定し，時間軸を一定の間隔で分割した各区間ごとに特徴点を追跡することにより，限定した時空間領域に対して特徴点軌跡を取得し，Dense Trajectories を用いたジェスチャ認識手法を提案した．評価実験より，時空間を限定をしない場合と比較して，「指をさす」，「うなずく」，「手をふる」のジェスチャ認識精度が 20%以上向上した．

今後の課題として，身体のパーツの重なりにも対応できるパーツ検出手法の考案が挙げられる．また，将来の研究として，より実環境に近づけたジェスチャ認識や，提案したジェスチャ認識を新しい特徴量として組み込んだ集団検出手法などが考えられる．しかし，集団検出をするためには，本研究データの他にどこに向かってジェスチャをしているかのようなジェスチャの向き情報などの付与が必要となる．

謝辞　本研究の一部は，科学技術振興機構 (JST) の戦略的創造研究推進事業 (CREST)「歩容意図行動モデルに基づいた人物行動解析と心を写す情報環境の構築」の支援によって行われた．

参考文献

[1] 岡本宏美，西尾修一，馬場口登，森井藤樹，萩田紀博：移動軌跡を用いた歩行者間の人間関係の推定 (テーマ関連セッション 8, コンピュータビジョンとパターン認識のための学習理論)，電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解， Vol. 108, No. 484, pp. 299–304 (2009).

[2] Ge, W., Collins, R. T. and Ruback, R. B.: Vision-Based Analysis of Small Groups in Pedestrian Crowds, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 34, No. 5, pp. 1003–1016 (2012).

[3] Chamveha, I., Sugano, Y., Sato, Y. and Sugimoto, A.: Social Group Discovery from Surveillance Videos: A Data-Driven Approach with Attention-Based Cues, *Proceedings of the British Machine Vision Conference* (2013).

[4] 西村拓一，向井理朗，野崎俊輔，　岡隆一：低解像度特徴を用いた複数人物によるジェスチャの単一動画像からのスポッティング認識，電子情報通信学会論文誌 D， Vol. 80, No. 6, pp. 1563–1570 (1997).

[5] 畠直志，岩井儀雄，谷内田正彦：動き情報と情報圧縮を用いたロバストなジェスチャ認識手法，電子情報通信学会論文誌 D， Vol. 81, No. 9, pp. 1983–1992 (1998).

[6] Matikainen, P., Hebert, M. and Sukthankar, R.: Trajectons: Action Recognition Through the Motion Analysis of Tracked Features, *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, IEEE, pp. 514–521 (2009).

[7] 高橋正樹，藤井真人，苗村昌秀，佐藤真一：特徴点軌跡に基づく監視映像からの人物行動検出 (テーマセッション, 映像処理と TRECVID)，電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解， Vol. 110, No. 414, pp. 111–116 (2011).

[8] Wang, H., Kläser, A., Schmid, C. and Liu, C.-L.: Dense Trajectories and Motion Boundary Descriptors for Action Recognition, *International journal of computer vision*, Vol. 103, No. 1, pp. 60–79 (2013).

[9] Shi, J. and Tomasi, C.: Good Features to Track, *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, IEEE, pp. 593–600 (1994).

[10] Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C.: Visual Categorization with Bags of Keypoints, *Workshop on statistical learning in computer vision, ECCV*, Prague, pp. 1–2 (2004).

[11] Boser, B. E., Guyon, I. M. and Vapnik, V. N.: A Training Algorithm for Optimal Margin Classifiers, *Proceedings of the fifth annual workshop on Computational learning theory*, ACM, pp. 144–152 (1992).

[12] Rothrock, B., Park, S. and Zhu, S.-C.: Integrating Grammar and Segmentation for Human Pose Estimation, *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, IEEE, pp. 3214–3221 (2013).

[13] Farnebäck, G.: Two-Frame Motion Estimation Based on Polynomial Expansion, *Image Analysis*, Springer, pp. 363–370 (2003).

[14] Dalal, N. and Triggs, B.: Histograms of Oriented Gradients for Human Detection, *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1, IEEE, pp. 886–893 (2005).

[15] Laptev, I., M arszałek, M., Schmid, C., Rozenfeld, B.：Learning realistic human actions from movies, *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, IEEE, pp. 1–8 (2008).

[16] Dalal, N., Triggs, B. and Schmid, C.: Human Detection using Oriented Histograms of Flow and Appearance, *Computer Vision–ECCV 2006*, Springer, pp. 428–441 (2006).

© 2016 Information Processing Society of Japan