Vol.2020-CVIM-221 No.36
2020/3/17

姿勢に基づく人物行動認識に関する基礎検討

川合諒[1]　吉田登[1]　潘雅冬[1]　西村祥治[1]

概要：本論文では，防犯カメラ映像を用いた人物行動の認識手法について，基礎検討の結果を報告する．人物の特徴としては，対象人物の服装や映像の撮影条件等に左右されにくいものとして姿勢の形状に着目し，それを一定時間蓄積して用いる．学習フェーズでは，距離学習により，類似した行動が近くに，異なる行動が遠くに位置するような特徴空間への変換を学習し，照合フェーズでは，学習結果を用いて姿勢情報を変換したうえで，得られた特徴を SVM により分類する．実験では，駅のプラットフォームやコンコースを模して撮影した映像を用いて認識を行い，一定の性能が得られたことを確認した．

キーワード：姿勢，人物行動認識，Metric Learning

Basic Study on Human Behavior Recognition based on Pose Features

RYO KAWAI[†1]　NOBORU YOSHIDA[†1]
YADONG PAN[†1]　SHOJI NISHIMURA[†1]

1. はじめに

近年，駅や空港などの多くの人が集まる場所において，事件や事故を未然に防止するニーズが高まっている．事件や事故を未然に防ぐ有効な手段として，それらの予兆となる人の動き（以下行動と称する）を発見し，実際にそれらが起こる前に適切に対処することが挙げられる．例えば，駅で泥酔するなどによりふらつきながら歩いている人がいれば，ホームからの転落を予測して保護することで，人身事故を防ぐことができるし，店内で店員などの動きをうかがうように周りを見回している人がいれば，事前に声をかけて万引き等の犯罪を防止することができる．

また，特定の行動の発見は，困っている人への手助け等，いわゆる「おもてなし」の面でも有効である．例えば，観光地の道路上で周りをキョロキョロと見回している人がいれば，道に迷っている観光客である可能性が高いため，迅速に発見して案内することで，その土地に対する満足度も向上し得る．

しかしながら，そのような各種の行動の発生を人手で監視することを考えると，人の目で監視できる範囲は限られているため，当該行動を確実に発見するには，監視すべき範囲に比例した多数の人員の確保が必要となる．とはいえ，監視だけのために多数の人員を確保することは人件費などの面から現実的ではなく，駅であれば駅員など，その場所に通常配置されているスタッフが，彼らの周りで発生した事案にのみ対処しているのが現状である．

そこで，防犯カメラで撮影された映像をもとに，コンピュータにより自動で各種の行動を検出することが考えられる．著者らもこれまでに，ふらつき歩行にフォーカスを当て，Two-stream CNN を用いた認識を行っている[1]．しかしこの手法では，特定の環境では人の往来がある程度あっても一定程度の認識性能が得られるものの，いくつかの問題があった．すなわち，対象者の体形や服装の多様性，背景や照明条件の違い等，周囲の環境やその他の条件によって大きく精度が低下し得るという点や，対象行動が変わったり増えたりしたときには，学習データの準備やパラメータを調整しながらの学習等を最初からやり直さなければならないという点などである．

そこで本研究では，姿勢情報を基にした行動認識について提案する．姿勢情報は，人間の情報を首，肩，腰，膝等の関節点の位置情報に抽象化したものである．姿勢情報は，行動の認識に必要な情報は残しつつ，服装や周囲の背景といった認識に余分な情報は取り除いた効率的な情報であるといえる．そういった姿勢情報を一定時間蓄積して 1 つの姿勢ベクトルとしたうえで，Metric Learning により低次元な特徴ベクトルを得て，それを SVM により分類することで行動を識別する．それにより，多様な環境下でも安定的な行動の識別を行うことを可能にする．

最終的には，あらゆる人間の行動を区別して表現可能な特徴ベクトル空間を構築し，認識の対象とする行動が変わったり増えたりしても，対象行動の低次元な特徴ベクトルを少数学習させるだけで識別できるようにすることを目指す．本論文ではその基礎検討として，与えられた複数の行動に対する認識性能を検証するとともに，学習データから一部行動を除外することで，未知の行動に対する現状の認識性能についても検証を行った．

本論文の構成は以下のとおりである．2 章で映像からの行動認識についての関連研究を紹介し，3 章で本研究の目

1 日本電気株式会社 バイオメトリクス研究所
Biometrics Research Laboratories, NEC Corporation

指す形について説明する．4 章で姿勢情報を用いた行動の識別手法の説明を行い，5 章において実際の映像を用いた実験について説明と結果の考察を行う．そして 6 章でまとめと今後の課題について述べる．

2. 関連研究

本章では，カメラ映像からの行動認識の手法についての関連研究を概観する．

Simonyan と Zisserman[2]は，2 種類の CNN（Convolutional Neural Network）を組み合わせる Two-stream CNN を提案した．得られた画像をそのまま入力し，見た目の特徴を学習する Spatial Stream と，オプティカルフローを計算して入力し，時系列に沿った動きの特徴を学習する Temporal Stream の 2 種類からなり，それぞれの CNN を用いて識別を行い，両結果を統合することにより精度高く対象人物の行動を認識することを可能にした．この手法に関しては，現在に至るまで様々な改善手法や拡張手法が提案されている[3][4][5].

しかし，前章でも述べたように，対象者の体形や服装，背景や照明条件の違いにより，認識性能が大きく低下し得るという問題がある．実際 He ら[6]は，映像中の認識対象の人物を矩形でマスクしても，残った背景のみに基づいて対象人物の行動を一定程度認識することが可能であることを示し，行動認識に背景が及ぼす影響が非常に大きいことを主張している．

そのような問題に左右されない手法として，姿勢情報を基にした行動認識の技術が複数提案されている．例えば Du ら[7]は，身体を両腕，両脚と背骨の計 5 つのパーツに分け，階層構造を持つ多数の Bidirectional Recurrent Neural Network（BRNN）で学習させることで行動認識を行う手法を提案したほか，Yan ら[8]は時系列で蓄積した姿勢情報に Graph Convolutional Network（GCN）を適用する Spatio Temporal GCN を提案している．また，ニューラルネットワークによらないものとして，Weng ら[9]は，特徴点マッチングベースで物体等の認識を行う NBNN（Naive Bayes Nearest Neighbor）を時系列に拡張した Spatio-Temporal NBNN を姿勢情報に適用して行動認識を行う手法を提案している．近年では，Pan らの手法[10]や Openpose としてライブラリ化されている Cao らの手法[11]など，可視画像から姿勢情報を高精度に抽出する手法が多く提案されていることもあり，活発に議論されている分野である．

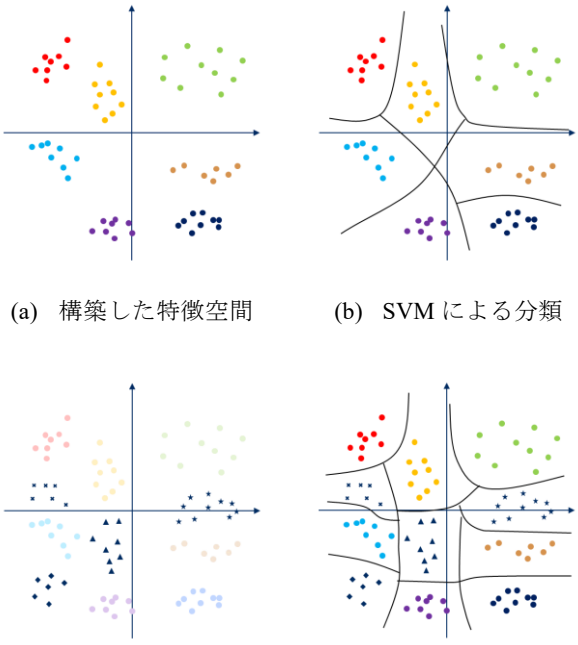
ただ，これまでに取り上げた手法はいずれも，認識対象とする行動の種類をあらかじめ決めておく必要があり，対象行動を変えたり増やしたりする際には，最初から学習をやり直す必要がある．これは，監視業務を考えるうえでは大きな問題になり得る．例えば，万引きの予兆となる行動を認識させることを考える．そのような行動は多岐にわたるため，それらをあらかじめ網羅的に対象行動に含めておくことは非常に困難で，監視業務を進めていく中で新たに万引きの予兆となり得る行動を発見する可能性がある．そのような行動を発見するたびに毎回大量のデータを使って学習をさせることは，非常に多くの時間の浪費になってしまう．

3. 本研究の目指す形

前章で述べた従来研究の課題に対し，本研究では，Metric Learning と SVM の組み合わせにより，最初に学習を行った後は，対象行動を増やす際でも学習に必要なデータを最小限にすることを目指す．

理想的なイメージを図 1 を参照しながら説明する．

1. 複数の種類の行動をより低次元な特徴ベクトル空間に射影する（図 1(a)）．
2. これにより得られる特徴ベクトルは，類似した行動は近くに，異なった行動は離れて分布するベクトル空間を形成し，SVM により容易に分類が可能になる（図 1(b)）．
3. さらに，当該ベクトル空間上では，未知な行動が入力されても，既存の行動とは適当な距離を保ちながら，似た行動同士は近くに分布する（図 1(c), 既存の行動は薄く示している）．
4. そのため，再度一から大量のデータを使って学習させることなく，SVM によって容易に各行動を分類できる（図 1(d)）．



(a) 構築した特徴空間　(b) SVM による分類

(c) 未知の行動の入力例　(d) 未知行動を含む分類

図 1 本研究の目指す形のイメージ

以上を本研究の最終目標とする．

4. 提案手法

4.1 概略

提案手法の概略について，図 2 を参照しながら説明する．人物の動作の情報を得るため，一定時間の姿勢情報の系列を用いる．認識にあたっては，まず 168 次元の姿勢ベクトルを Deep Metric Learning により 40 次元の特徴ベクトルに変換し，その特徴ベクトルに対して Support Vector Machine（SVM）による分類（Support vector classification）を行って行動の種類を認識する．以下，姿勢情報のベクトル化の手順と Deep Metric Learning について，詳しく説明する．

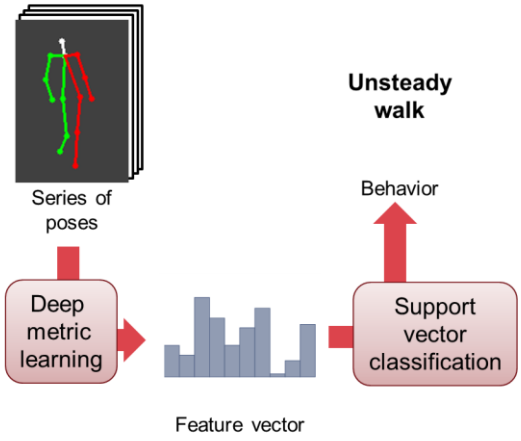


図 2 姿勢情報に基づく行動認識の流れ

4.2 姿勢情報のベクトル化の手順

姿勢情報は，Openpose の出力に基づき，鼻，首，および両肩，両肘，両手，両股関節，両膝，両足，両目，両耳の計 18 のキーポイントにより表現される．キーポイントのイメージ図を図 3 に示す．

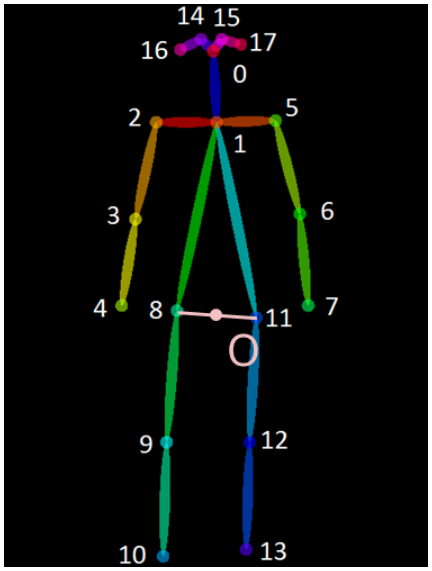


図 3 キーポイント

これらのキーポイントは，他の人物や物体によって多少隠蔽されても推定される位置に検出されるが，大きく隠蔽があると検出できず欠損となる．このような欠損データがある場合は，基本的には今回の実験のデータには含まない一方，鼻と両目，両耳は例外として，少なくとも 1 点が検出されていればよいものとした．この理由は，他の人物や物体がなくても，これらのキーポイントでは欠損が容易に起こり得るためである．たとえば，当該人物が後ろ向きに映っている場合は鼻や両目が映らないため検出されないし，横向きに映っている場合は目や耳が片側しか映らないため反対側のキーポイントが検出されない．

得られた姿勢情報を，以下の手順でベクトル化する．

まず，前述の通り検出点数にばらつきのある鼻と両目，両耳のキーポイントについて，検出できているキーポイントの重心点を算出し，その点を頭部のキーポイントP_0とする．

また，首，右肩，右肘，右手，左肩，左肘，左手，右股関節，右膝，右足，左股関節，左膝，左足のキーポイントを順に$P_1, P_2, \dots, P_{13}$とし（添字は図 3 内の数字と一致する），線分P_8P_{11}の中点を原点O（図 3 中にも表記），Oから$P_k(0 \leq k \leq 13)$へのベクトルを$\boldsymbol{p}_k = (x_k \quad y_k)^T$，線分$P_iP_j$の長さ$\|\boldsymbol{p}_j - \boldsymbol{p}_i\|$を$l_{i,j}$とおく．

そして，このままでは人の位置や大きさが統一されていないため，姿勢情報を正規化する．頭から左右の足先までのキーポイントを順にたどった長さの平均，

$$l = l_{0,1} + \frac{l_{1,8} + l_{8,9} + l_{9,10} + l_{1,11} + l_{11,12} + l_{12,13}}{2}$$

を基準として，$\boldsymbol{p}_k$を正規化したベクトル

$$\boldsymbol{p'}_k = \begin{pmatrix} x'_k \\ y'_k \end{pmatrix} = \frac{\boldsymbol{p}_k}{l} = \begin{pmatrix} x_k/l \\ y_k/l \end{pmatrix}$$

を得る．そして，フレーム単位の姿勢ベクトル$\boldsymbol{v}$を

$$\boldsymbol{v} = \begin{pmatrix} \boldsymbol{p'}_0 \\ \vdots \\ \boldsymbol{p'}_{13} \end{pmatrix} = \begin{pmatrix} x'_0 \\ y'_0 \\ \vdots \\ x'_{13} \\ y'_{13} \end{pmatrix}$$

として定義する．ここからわかるように，$\boldsymbol{v}$は 28 次元のベクトルとなる．このベクトル$\boldsymbol{v}$を一定フレーム蓄積し，それらを積み上げて 1 つの姿勢ベクトル

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{v}_{-p(n-1)} \\ \vdots \\ \boldsymbol{v}_{-p(1)} \\ \boldsymbol{v}_{-p(0)} \end{pmatrix}$$

を得て，この姿勢ベクトル$\boldsymbol{V}$により行動を識別していく．ここで，$\boldsymbol{v}_{-t}$は$\boldsymbol{v}$のt秒前のフレームにおける姿勢ベクトルを表し，$\boldsymbol{v}_0 = \boldsymbol{v}$である．また，$n$は蓄積するフレーム数，$p(k)(0 \leq k \leq n-1)$は$k$番目の特徴が何秒前の映像かを定める$k$についての関数である．本論文では，以下$p(k) = 0.2k, n = 6$として

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{v}_{-1.0} \\ \boldsymbol{v}_{-0.8} \\ \boldsymbol{v}_{-0.6} \\ \boldsymbol{v}_{-0.4} \\ \boldsymbol{v}_{-0.2} \\ \boldsymbol{v}_{0} \end{pmatrix}$$

を用いる．言い換えれば，1 秒前から現在まで，0.2 秒刻みで 6 フレーム分の姿勢情報を 1 つの特徴ベクトル$\boldsymbol{V}$として用いる．$\boldsymbol{v}_{-t}$が 28 次元のベクトルであることから，$\boldsymbol{V}$が 168 次元のベクトルとなることがわかる．

4.3 Deep Metric Learning による特徴ベクトルの導出

次に，Deep Metric Learning により，姿勢ベクトル$\boldsymbol{V}$ を特徴ベクトルに変換する．本研究では，Triplet loss[12]を用いた Metric Learning を行う．Triplet loss は，図 4 に示すように，基準となる Anchor サンプルと，それと同じクラスに属する Positive サンプル，それらと異なるクラスに属する Negative サンプルの 3 組を入力とする．

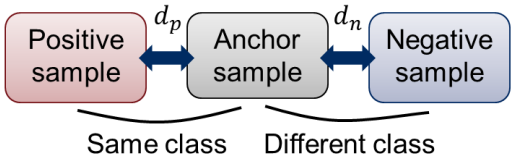


図 4 Triplet loss に用いる 3 つのサンプル

そして，Anchor・Positive 間の距離d_p，Anchor・Negative 間の距離d_n，およびそれらの離し具合を調整するパラメータであるマージンmを用いて，Loss

$$L = \max(d_p - d_n + m, 0)$$

が小さくなるように学習を進める．本研究では，距離としてはユークリッド距離を用い，図 5 に示すようなネットワーク構造により学習を進め，40 次元の特徴ベクトルを得る．

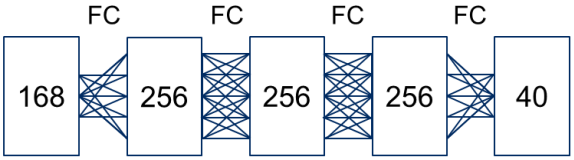


図 5 ネットワーク構造

5. 実験と考察

5.1 データセット

本研究では，2 通りの方法により姿勢情報のデータベースを作成し，データセットとした．1 つ目は，カメラ映像に OpenPose を適用し，姿勢検出を行う方法である．事業場内の屋内駐車場で撮影した映像と，駅のプラットフォームやコンコースを模した場所で撮影した映像に対し，この方法により姿勢検出を行った．2 つ目は，モーションキャプチャシステムにより 3 次元の姿勢情報を取得し，カメラ設置条件を仮定してそれを 2 次元情報に変換する方法である．これらの方法で 12 通りの行動の姿勢情報を取得し，学習および評価を行った．データセットの例を図 6 に，具体的な行動の種類とデータ数を表 1 に示す．なお以下では，各行動を表 1 内の行動名の前に記したアルファベットのラベルで呼ぶことがある．

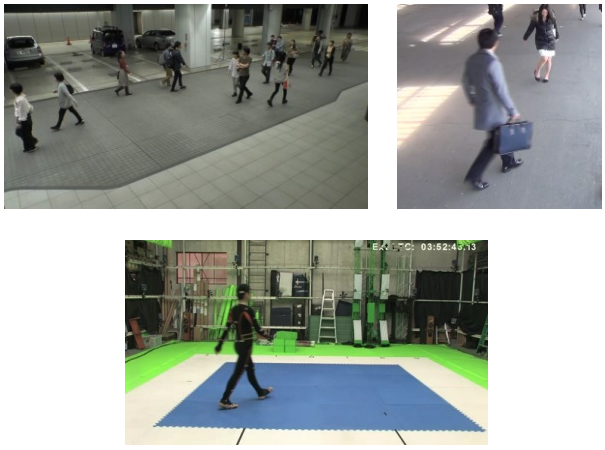


図 6 データセットの例

表 1 各データの枚数

行動	Metric	SVM
A: 通常歩行	25154	3602
B: ふらつき歩行	61976	3756
C: 歩きスマホ	44028	6035
D: うつむき歩行	17996	5827
E: 小走り	6999	3157
F: 倒れ	25696	3416
G: キョロキョロ	30668	3592
H: 座り込み	35072	4128
I: 歩行の中断	7712	3616
J: 歩行中断後スマホ操作	10400	5088
K: 静止	14752	4944
L: 後ろ向きのふらつき歩行	4800	3584

Metric の列に示しているデータで Metric Learning を行った後，SVM の列に示しているデータで SVM の 4 分割交差検証を行って評価する．ラベル A～E はカメラ映像とモーションキャプチャシステムのデータの両方があるが，同 F～L はモーションキャプチャシステムのデータのみからなる．なお，すべての行動は被験者に当該行動をするように指示をして撮影および取得したものである．

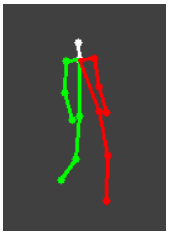
5.2 性能評価

全 12 種類で認識を行った際の混合行列を表 2 に示す．表内のアルファベットは行動のラベルである．赤色の濃さは認識率の高さに対応している．認識率の単位は%である．

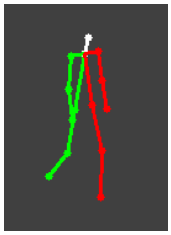
表 2 全行動による性能評価の結果

正解＼推定	A	B	C	D	E	F	G	H	I	J	K	L
A	50.6	9.1	3.0	31.0	1.7	0.4	0.1	0.0	1.7	0.1	0.0	2.3
B	12.1	49.7	2.3	8.3	4.0	5.2	1.8	0.5	1.0	0.3	0.2	14.8
C	2.2	1.5	81.5	8.7	4.7	0.1	0.2	0.0	0.0	0.1	0.2	0.7
D	23.6	5.5	9.6	56.3	2.6	0.1	0.0	0.0	1.0	0.5	0.0	0.8
E	1.6	1.1	3.5	3.2	89.7	0.0	0.0	0.0	0.0	0.0	0.0	0.9
F	0.9	2.2	0.0	1.6	0.0	91.0	0.2	0.5	0.5	0.4	0.1	2.8
G	0.1	4.3	0.0	0.0	0.0	0.0	71.3	0.3	2.3	0.1	20.0	1.7
H	0.0	0.2	0.0	0.0	0.0	0.6	0.1	98.8	0.0	0.0	0.0	0.3
I	1.7	0.2	0.0	2.4	0.0	0.0	3.0	0.0	70.2	18.0	4.4	0.1
J	0.2	0.4	0.5	1.2	0.0	0.0	0.3	0.0	17.3	74.5	5.6	0.1
K	0.2	0.9	0.0	0.0	0.0	0.0	33.3	0.0	2.1	9.6	53.9	0.0
L	3.0	10.3	1.5	1.6	1.2	5.8	2.2	0.0	4.4	0.2	0.9	68.8

全体として，7〜8 割程度以上の認識率を達成した．特に，座り込みや倒れなど，他の行動と動きの違いがはっきりしている行動の認識率は 9 割を超えている．その一方，通常歩行とうつむき歩行，キョロキョロと静止を誤りやすい傾向があることがわかる．これら 4 行動の姿勢を可視化した例を図 7 に示す．

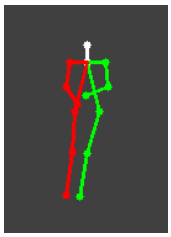


(a) 通常歩行



(b) うつむき歩行

(c) キョロキョロ



(d) 静止

図 7 類似した行動の姿勢の例

見比べるとわかるように，通常歩行とうつむき歩行，キョロキョロと静止はかなり類似した姿勢となっており，この姿勢情報からの識別はかなり困難であるといえる．これらの行動に関していえば，顔向きが重要な要素である一方，頭部のキーポイントを 1 点に集約したことで，その顔向きの情報がほとんど消えてしまっていることが大きな原因として考えられる．鼻のキーポイントがなければ後ろ向き，左耳のキーポイントがなければ右向きなど，顔向きによりどのキーポイントが検出されるかが変わるため，検出されなかったキーポイントの情報も姿勢情報に含めることで，性能が改善する可能性がある．

また，ふらつき歩行に関しては通常歩行，うつむき歩行，倒れ，後ろ向きのふらつき歩行等，多くの行動と誤りやすい傾向があることがわかる．これに関しては，現在の識別方法では一連の行動から 1 秒間を抜き出して識別を行うことになるが，ふらつき歩行のような不安定な動きから 1 秒を抜き出すと，通常歩行やうつむき歩行，倒れ，後ろ向きのふらつき歩行など，様々な別の行動に類似したものになっている可能性がある．識別のために蓄積する時間を延ばすことでより高い精度で識別できる可能性があるが，それによって副作用も発生し得るため，適切な時間幅について慎重に検証する必要がある．

次に，未知の行動に対する認識精度を評価する．各行動のうち，うつむき歩行を除いて Metric Learning を行ったうえで，全 12 行動に対して SVM による認識を行ったときの混合行列を表 3 に示す．

表 3 うつむき歩行を除外した場合の性能評価の結果

正解＼推定	A	B	C	D	E	F	G	H	I	J	K	L
A	73.8	14.3	5.5	4.3	1.6	0.0	0.4	0.0	0.0	0.0	0.1	0.0
B	15.2	75.1	1.0	2.3	1.7	0.2	0.5	0.2	0.0	0.2	0.3	3.3
C	3.4	0.9	85.0	5.6	2.8	0.2	0.2	0.1	0.0	0.1	0.1	1.6
D	16.6	7.4	15.0	32.8	4.1	3.6	0.3	0.0	11.4	3.4	0.4	5.0
E	3.0	0.7	3.5	1.5	90.1	0.0	0.0	0.0	0.0	0.0	0.0	1.2
F	0.0	1.3	0.5	2.2	0.0	92.8	0.1	1.0	0.3	0.6	0.0	1.2
G	0.1	4.0	0.0	0.2	0.0	0.1	63.1	0.5	2.7	0.1	27.6	1.6
H	0.0	2.5	0.0	0.0	0.0	1.1	0.0	95.9	0.0	0.0	0.0	0.5
I	0.0	0.0	0.0	5.4	0.0	0.0	2.6	0.0	68.4	19.4	3.8	0.2
J	0.0	0.9	0.7	2.7	0.0	0.2	1.2	0.0	14.6	75.8	3.6	0.2
K	0.0	1.2	0.0	0.3	0.2	0.0	34.5	0.0	2.0	5.6	55.1	1.2
L	0.0	3.2	0.9	9.0	0.7	1.8	2.2	0.4	1.9	0.4	1.2	78.5

Metric Learning から除外されているうつむき歩行の認識率は 32.8%となり，全 12 行動で Metric Learning を行ったとき（表 2）の 56.3%と比べて下がってはいるものの，誤って他の行動と認識されたもののうち最も割合が高いものでも通常歩行の 16.6%であり，その 2 倍の割合で正解していることがわかる．また，他の行動を誤ってうつむき歩行と認識する割合も全て 10%に満たない．うつむき歩行以外の 11 行動から構築された特徴空間上でも，うつむき歩行がある程度固まって分布できており，Metric Learning の効果が表れているといえる．今後，行動の種類をさらに増やしていくことで，より多くの未知の行動にも対応した特徴空間が構築可能になると見込まれる．

6. まとめと今後の課題

本論文では，姿勢情報をもとにした行動認識について提案した．姿勢情報を一定時間蓄積して 1 つの姿勢ベクトルとしたうえで，Triplet Loss を用いた Metric Learning により低次元な特徴ベクトルに変換し，それを SVM により分類することで行動を識別する手順をとった．実験の結果，12 種類の行動に対して，概ね 7〜8 割程度以上の性能が得られ，特に動きの違いがはっきりしている行動では 9 割を超える認識率となった．一方，顔向きの違い以外が類似している行動や，1 秒間に切り取った際に他の行動と類似している行動に対して性能が低下する傾向が見られた．また，1 つの行動を除外して Metric Learning を行ったうえですべ

ての行動で SVM を行っても，除外した行動が正しいクラスに割り当てられる割合が最も高く，Metric Learning による効果が表れていることが確認できた．

今後の課題としては，顔向きの違いを区別できる姿勢情報の表現方法の検討，識別対象とする適切な時間幅の検証のほか，頭部以外の情報欠損に対する対処，見え方の違いに対する頑強性の確保等が挙げられる．また，既存の手法との認識性能の比較も行っていく必要がある．

そのほか現在，Metric Learning において類似度を考慮する方法も検討中である．例えばうつむき歩行と歩きスマホは異なる行動であるが見た目の類似度は高い．そのような行動と，歩行と寝込みのように見た目が全く異なる行動とを同じ基準で距離を離すように学習させるのではなく，類似度によって離す度合いを適応的に変えることで，より頑健な特徴空間が得られる可能性がある．

このような手法を導入したり，これまで用いてきた 12 種類の行動よりもさらに多様な姿勢をとっているデータを学習データに追加するなどして対応可能な範囲を広げたりすることでより普遍的な特徴空間を構築し，最終目標としている，あらゆる人間の行動を区別して表現可能な特徴ベクトルの生成を目指したいと考えている．

参考文献

[1] 川合諒; 細井利憲; 小西勇介. 人工生成データの学習による人の重なりに頑強なふらつき歩行認識. 第 24 回画像センシングシンポジウム (SSII2018), IS1-17, 2018.
[2] Simonyan, Karen; Zisserman, Andrew. Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, 2014, pp. 568-576.
[3] Kapidis, Georgios, et al. Multitask Learning to Improve Egocentric Action Recognition. In: Proc. of the IEEE Int. Conf. on Computer Vision Workshops, 2019.
[4] Crasto, Nieves, et al. MARS: Motion-augmented RGB stream for action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2019, pp. 7882-7891.
[5] Xu, Huijuan; Das, Abir; Saenko, Kate. Two-stream region convolutional 3d network for temporal activity detection. IEEE trans. on Pattern Analysis and Machine Intelligence, 2019, 41.10: 2319-2332.
[6] He, Yun, et al. Human action recognition without human. In: Proc. European Conference on Computer Vision 2016 Workshops, 2016, pp. 11-17
[7] Du, Yong; Wang, Wei; Wang, Liang. Hierarchical recurrent neural network for skeleton based action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2015, pp. 1110-1118.
[8] Yan, Sijie; Xiong, Yuanjun; Lin, Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conf. on artificial intelligence, 2018.
[9] Weng, Junwu; Weng, Chaoqun; Yuan, Junsong. Spatio-temporal naive-bayes nearest-neighbor (ST-NBNN) for skeleton-based action recognition. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2017, pp. 4171-4180.
[10] Pan, Yadong; Nishimura, Shoji. Multi-Person Pose Estimation with Mid-Points for Human Detection under Real-World Surveillance. In The 5th Asian Conf. on Pattern Recognition (ACPR2019), 2019.
[11] Cao, Zhe, et al. Realtime multi-person 2d pose estimation using part affinity fields. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2017, pp. 7291-7299.
[12] Wang, Jiang, et al. Learning fine grained image similarity with deep ranking. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2014, pp. 1386-1393