Vol.2022-CVIM-228 No.24
2022/1/28

骨格情報に基づく ATM 前異常行動検知に関する検討

川合諒[1]　吉田登[1]　劉健全[1]

概要：本論文では，振り込め詐欺対策などを目的とした，携帯電話の利用等，ATM 前での異常行動の検出について述べる．人物の特徴として，服装や性別等に左右されづらく，かつ個人情報を隠蔽可能な骨格情報を用い，あらかじめ登録されたサンプルとの類似度を回転不変な特徴量から算出することで異常行動を検出する．実験では，ATM に設置されたカメラを模した画角から撮影された人物の映像に対して認識を行い，一定の性能が得られたことを確認した．

キーワード：姿勢，骨格，人物行動認識，映像検索

Study on Abnormal Behavior Detection at ATMs based on Pose Information

RYO KAWAI[†1]　NOBORU YOSHIDA[†1]
JIANQUAN LIU[†1]

1. はじめに

近年，銀行における取引では ATM（現金自動預け払い機）の利用が主流になっている．昨今の人との接触を避けたいというニーズやネット専業銀行の台頭もあり，その傾向はさらに顕著になっているが，人を介さないことで，新たな問題も発生している．例えば，還付金詐欺に代表される，携帯電話で ATM の操作を指示して現金をだまし取る犯罪が多発していることや，立っての利用が前提となっているため車椅子利用者がスムーズに利用しづらいことなどである．

中でも詐欺被害の多発は大きな問題となっているため，ATM 周辺での携帯電話利用禁止の流れが強まっており，中には妨害電波を発することで携帯電話の利用自体をできなくする対策がなされている場合もある．しかし，例えばコンビニ ATM では，ATM を利用しない人も付近を通過する可能性があり，ATM 自体に携帯電話を利用して取引が行える機能が含まれている場合もあるため，携帯電話が全く使えなくなってしまうことはユーザーエクスペリエンスの低下につながる．そこで，その欠点を補う対策として，コンピュータにより自動で携帯電話による通話を発見し，迅速に対処するという方法が考えられる．

本研究では，自動での携帯電話による通話の検出に，人物の骨格情報を用いることを提案する．骨格情報は，人間の情報を首，肩，肘，膝等の関節点の位置情報に抽象化したものである．手で隠れやすい携帯電話ではなく，それを利用する人の姿勢に基づいて検出することで，未検出を減らすことを目指す．さらに，筆者らが独自に開発したオンデマンド行動検出の技術により，機械学習は行わずに当該姿勢を検出することを可能にした．また，学習を行わない本技術の利点として，データを追加して携帯電話の利用検知以外に検出対象を広げることが容易という点があることから，車いす利用者の検出も予備評価として合わせて実験した．

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 章で提案手法について説明する．4 章で実験用の映像を用いた評価結果と考察，5 章で車椅子利用者検出の予備評価結果と考察について述べ，6 章でまとめと今後の課題について述べる．

2. 関連研究

人物の骨格抽出は，Cao ら[1]による OpenPose を嚆矢として，現在も活発に議論されている分野である．対象者の体形や服装，周辺環境といった余分な情報は除去される一方，人の動きを識別可能な程度の情報は残る骨格情報は，行動認識に非常に適した情報であるといえる．実際，骨格情報に基づく行動認識の手法はいくつも提案されている．

Du ら[2]は，身体を四肢それぞれと背骨の計 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早くから提案している．一方，RNN ではなく Graph Convolutional Network（GCN）を用いたのが Yan ら[3]の手法であり，時系列で蓄積した姿勢情報に GCN を適用する Spatio Temporal GCN（ST-GCN）を提案し，高い精度を達成している．これらの改良手法も多く提案されており，例えば Shi ら[4]は，グラフ構造を層毎，行動毎に適応的に変え，関節点の位置だけでなく骨格の長さや方向の情報も用いる Two-Stream Adaptive Graph Convolutional Networks を提案，

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

ST-GCN を上回る精度を達成している．また，ニューラルネットワークによらないものとしては Weng ら[5]の手法が挙げられる．彼らは，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．

ただ，これまでに取り上げた手法は，ニューラルネットワークによるものはもちろんのこと，ST-NBNN でも各クラスの分類スコアを求める関数を得るため，認識対象とする行動の種類を決めたうえで前もって学習を行っておく必要がある．しかし，実環境での行動認識に応用するには，学習に十分なデータが用意できるとは限らないこと，また，顧客からのニーズは絶えず変化し得るものであるため認識対象としたい行動が増減する可能性があり，その都度学習をし直すことは合理的でないことなど，難しい点も多い．

3. 提案手法

前章で述べた従来研究の課題に対し，本研究では，吉田ら[6][7][8]によるオンデマンド行動検出を用いることを提案する．オンデマンド行動検出の概要を図 1 に示す．

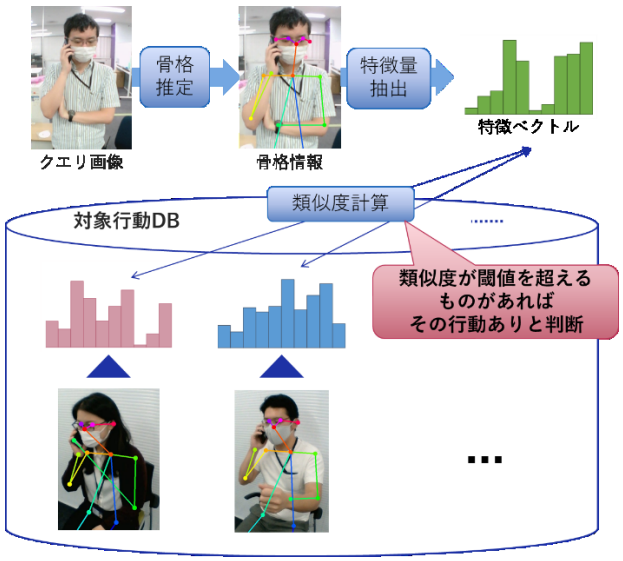


図 1 オンデマンド行動検出の概要

オンデマンド行動検出では，画像から得た骨格情報を変換して得られる，高さ情報にフォーカスした回転不変の特徴量（Normalized Vertical（NV）特徴量）を用いる．あらかじめ，識別対象としたい行動を映したサンプル映像を用意し，それから NV 特徴量を得ておく．クエリ画像が入力された際には，その画像から得られた骨格情報を NV 特徴に変換し，クエリとサンプル映像の同特徴の類似度を算出して，サンプル映像中に閾値以上の類似度のものがあれば当該行動をしているものと判断する．なお，学習によらないため，サンプル映像は最低 1 データあれば足りる．以下，ポイントごとに説明する．

3.1 骨格推定

まず，骨格推定について説明する．本研究では Pan ら[9]による NeoPose を用いる．骨格抽出の詳細の説明は[9]に譲り，ここでは，骨格を構成する情報について説明する．

NeoPose の骨格情報は，鼻，首と，左右の肩，肘，手首，股関節，膝，脚，目，耳の合計 18 点の 2 次元画像座標の値により構成される．ただし以下では，鼻と左右の目，耳の頭部 5 点について，これらのうち検出できた点の重心を代表させる形で 1 点に集約し，その他 13 点と合わせた 14 点を用いる．これは，頭部 5 点は，真正面を向いていない限りいずれかに隠蔽が発生する一方で，頭部 5 点の間の位置関係が大きく変わることはないためである．関節点のイメージ図を図 2 に示す．

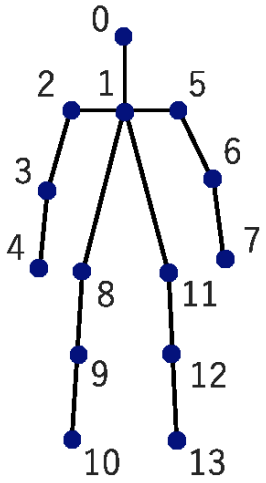


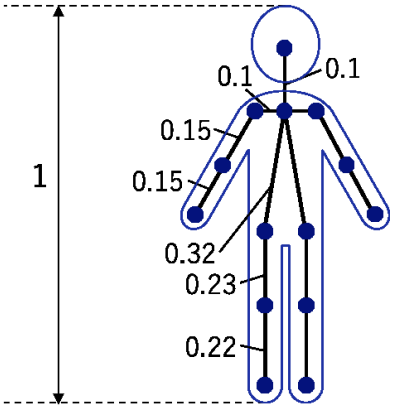
図 2 利用する骨格情報

3.2 回転不変特徴量への変換

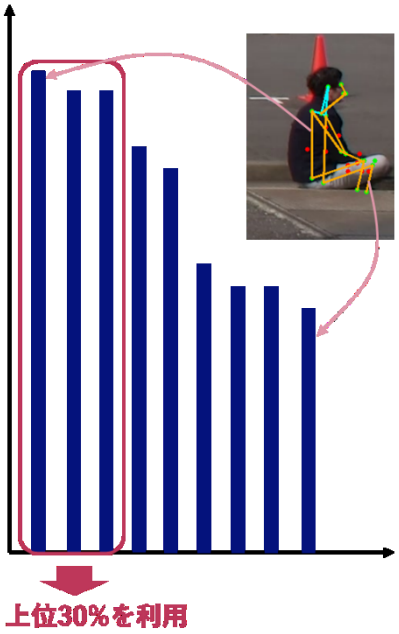
次に，NV 特徴量への変換の方法について説明する．本特徴量を求めるためには，先に特徴量の正規化の基準となる身長の推定を行ったのち，NV 特徴量に変換する．以下，順を追って記す．

3.2.1 身長の推定

ここでの身長とは，画像上のピクセル単位での身長であり，別の言い方をすれば，画像上に映っている人が，その場で直立したと仮定したときの足元から頭の先までのピクセル単位の長さである．この身長の求め方について，図 3 を参照しながら説明する．



(a) 平均人物モデル



(b) 各部位から算出する身長

図 3 モデルに基づく身長推定

まず，立っている人物の骨格情報を多数用意し，それぞれの人物の各関節を結ぶ骨格の長さと身長との比率を計算する．それを平均することで，身長を 1 としたときの平均的な人物の骨格の長さを算出する．(図 3(a))．推定の際には，検出された各関節点を結ぶ骨格の長さを図 3(a)で示した値で除することで，身長を推定する．もし検出できなかった関節点があっても，それ以外を用いて推定可能である．そして，図 3(b)に示すように，推定値の上位 30%の平均を，最終の推定値とする．上位の値を用いるのは，見た目の骨格は実際より短く見えることはあっても（カメラの光軸と骨格の向きが近い場合），実際より長く見えることはほぼないことから，長い方が信頼できるためである．

この値を基準に特徴量を正規化することで，対象者がとる姿勢にかかわらず，対象者が映る大きさのみによって正規化を行うことができる．

3.2.2 NV 特徴量の定義

ある骨格情報に対する NV 特徴量の要素$f_i(0 \le i \le 13)$は，

$$f_i = \frac{y_i - y_{\text{core}}}{h}$$

により求める．ここで，hは前節で推定した身長の値を，y_iは 14 個の関節点のy座標（添字iは図 2 上に記した数字に対応）を，y_{core}はベクトルの起点となる点のy座標を示す．本研究ではベクトルの起点は首の点とする．すなわち，$y_{\text{core}} = y_1$である．ただし，検出できなかった関節点については，$f_i = -1$とする．そして，これらを要素とする NV 特徴量$\boldsymbol{f}$を

$$\boldsymbol{f} = (f_0 \quad f_1 \quad \cdots \quad f_{13})^T$$

として定義する．

3.2.3 特徴量の類似度算出

2 つの NV 特徴量，$\boldsymbol{f}_a = (f_{a,0} \quad f_{a,1} \quad \cdots \quad f_{a,13})^T$と$\boldsymbol{f}_b = (f_{b,0} \quad f_{b,1} \quad \cdots \quad f_{b,13})^T$との間の類似度$\text{sim}(\boldsymbol{f}_a, \boldsymbol{f}_b)$は，$a, b$ともに検出できた関節点の添字の集合

$$I_{a,b} = \{i | 0 \le i \le 13, f_{a,i} \ne -1, f_{b,i} \ne -1\}$$

を用いて，

$$\text{sim}(\boldsymbol{f}_a, \boldsymbol{f}_b) = \left(1 + \frac{1}{|I_{a,b}|} \sum_{i \in I_{a,b}} |f_{a,i} - f_{b,i}|\right)^{-1}$$

により定義される．ここで$|I_{a,b}|$は集合$I_{a,b}$の要素数を示す．こうして得られた類似度が一定の閾値以上のとき，$\boldsymbol{f}_a$と$\boldsymbol{f}_b$はマッチしたとみなす．

4. 携帯電話利用検知の実験と考察

4.1 データセット

本研究では，図 4 のように携帯電話を利用している映像を評価用に撮影し，それらを用いて性能の評価を行った．サンプルとしてデータベースに登録した画像数と，テストに用いた画像数は表 1 のとおりである．

図 4 携帯電話利用の画像例

表 1 携帯電話利用の画像の枚数

	利用	利用せず
サンプル	136	
テスト	92	37

4.2 性能評価

性能評価の基準として，FPR（False Positive Rate）と FNR

（False Negative Rate）を用いる．定義式は TP（True Positive），TN（True Negative），FP（False Positive），FN（False Negative）のそれぞれのサンプル数を用いて以下のように定義される．0 に近づくほど性能が高いことを示す．

$$\mathrm{FPR} = \frac{FP}{FP + TN}$$

$$\mathrm{FNR} = \frac{FN}{FN + TP}$$

結果，FPR が 10.8%，FNR が 20.7%となった．なお，本評価結果における閾値は，定性的に最も安定していた 0.91 としている．概ね 8 割方以上の正解率で検出に成功していることがわかる．

検出成功例を図 5 に示す．

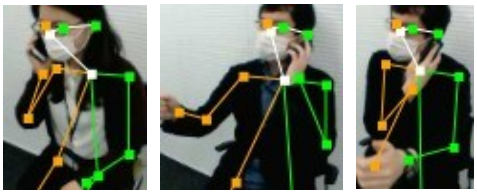
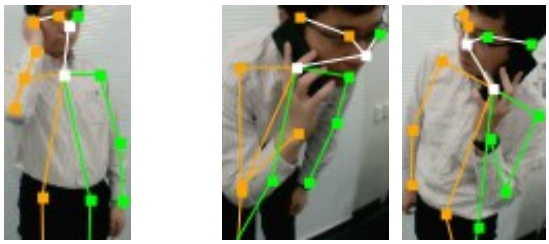


図 5 携帯電話利用検知の成功例

見た目の姿勢は大きく違うが，それでも正しく検出できていることがわかる．特に右側は，右手で持ちながら左耳に当てるという特殊な姿勢をしているが，正しく検出できている．これは，NV 特徴量は高さ情報に特にフォーカスした特徴量であることから，どちらの耳に当てていても，同じ手で持っている限り特徴量は類似したものになるためであり，このような特殊な姿勢がデータベースに含まれているわけではない．

次に，未検出・誤検出例を図 6 に示す．



(a) 誤検知　　(b) 未検知

図 6 携帯電話利用検知の失敗例

誤検知の例は耳を掻いていて，手首までの姿勢では携帯電話利用との見分けがつかない．これは提案手法の原理的に区別は不可能であり，携帯電話の検出などアプローチを変えない限り誤検出は避けられないと思われる．未検知の左側の画像は，腰を曲げて高齢者を想定した画像であるが，骨格抽出に失敗しているため検出に失敗してしまっている．現在の骨格抽出手法で想定していない姿勢および見え方になっている可能性がある．同じく右側の映像は，持ち方の問題か手首が少々下の方に検出されており，それにより未検出になっていると思われる．この改善策としては，現在は携帯電話使用のサンプルにマッチしなければ携帯電話を使用していないと判断しているのを，携帯電話不使用のサンプル（いわゆる Negative サンプル）も作ってより近いのはどちらかで判断するといったことが考えられる．

5. 車椅子利用者検出の予備評価

5.1 データセット

車椅子利用者検出の予備評価として，椅子に座っている人の検出の可能性について評価を行った．画像の例は図 7，画像数は表 2 のとおりである．

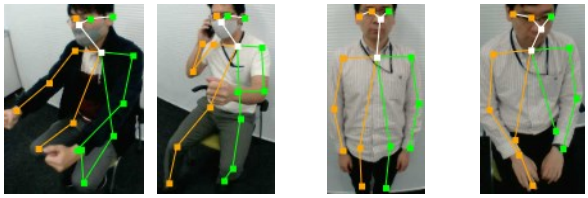
図 7 車椅子利用の画像例

表 2 車椅子利用者検出の画像の枚数

	利用	利用せず
サンプル	113	
テスト	42	97

5.2 性能評価

同じく FPR と FNR を算出したところ，閾値 0.95 において FPR が 53.6%，FNR が 21.4%となり，FP が非常に多い結果となった．検出成功例と失敗例を図 8 に示す．



(a) 成功　　(b) 誤検知　　(c) 未検知

図 8 車椅子利用者検出の成功例・失敗例

成功例の右側は携帯電話も利用している例だが，そういった違いには左右されずに検出に成功している．失敗例のうち，誤検知として多かったのが，図 8(b)のような例である．単に直立しているだけだが，俯角が比較的大きい撮影環境のために見た目上の下半身が小さく見え，NV 特徴量上での類似度が高くなってしまったものと考えられる．このような，俯角の影響によるロバスト性の低下への対処は今後の課題といえる．一方，未検出は図 8(c)のように足元に手があったり暗かったりすることが原因でそもそも骨格が検出できない場合が多かった．ATM にカメラを設置することを想定する場合，人とカメラとの距離は近くなるためこのような問題は実際に多く発生し得る．適切な設置位置の検討が重要であると考えられる．

6. まとめと今後の課題

本論文では，骨格情報に基づく ATM 前における異常行動検知について提案した．骨格情報を回転不変の特徴量に変換し，入力された骨格情報があらかじめ用意しておいた識別したい行動を映したサンプル画像から得られた特徴量と閾値以上の類似度があれば対象行動をとっているとみなす学習不要の手法である．

携帯電話の利用と車椅子の利用を対象に実験を行った結果，携帯電話の利用は，手と反対側の耳に当てるなど，イレギュラーな姿勢であっても正しく検出できることが確認できた一方，持ち方等により関節点位置が少しずれて検出されると未検出となるなどの問題も見られた．車椅子の利用は，足元が小さく見えることによる誤検知や，全く見えないことによる未検知の例が見られた．

今後の課題としては，腰を曲げた高齢者など，骨格抽出が失敗する例への対処や，Negative サンプルも検索対象に加え, Positive サンプルの検索結果と比較して判断するような新しい判別方法の導入などが挙げられる．

参考文献

[1] 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.
[2] 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.
[3] 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.
[4] Shi, Lei; Zhang, Yifan, Cheng, Jian, Lu, Hanqing. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019, pp. 12026-12035.
[5] 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.
[6] “NEC，数秒のサンプル映像を与えるだけで，ライブカメラの映像などから見つけたい行動を検出できる技術を開発”. https://jpn.nec.com/press/202104/20210405_01.html, (参照 2021-12-20).
[7] Yoshida, Noboru; Liu, Jianquan. On-demand action detection system using pose information. In: Proc. of the 29th ACM International Conference on Multimedia, 2021, pp. 2810-2812.
[8] Yoshida, Noboru; Liu, Jianquan. View-invariant feature using pose information and flexible matching algorithm for action retrieval. In; 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2021.
[9] Pan, Yadong; Kawai, Ryo; Yoshida, Noboru; Ikeda, Hiroo; Nishimura, Shoji. Training physical and geometrical mid-points for multi-person pose estimation and human detection under congestion and low resolution. SN Computer Science, 2020, 1.4: 1-8.