入退室管理のための深層学習を用いた個人識別手法の構築

Personal identification by deep learning for entrance and exit management

精密工学専攻　33 号　戸田　哲郎
Tetsuro Toda

1. 序論

在室管理やセキュリティ，防犯，個別サービスなど，個人を識別することが必要となるシーンは様々なところにある．個人識別をするためによく用いられる方法として，RFID の利用や指紋認証，虹彩認証などがある．しかし，これらの方法は各個人が認証のための行動をとる必要があり，利用者の負担となっている．そのため，カメラから撮影された映像からの個人識別手法の研究が盛んに行われている．

近年では，CNN（Convolution Neural Network）を利用した顔識別 [1] が高い個人識別率を実現している．しかし，顔識別を行う場合には，画像中で顔が写っている必要があり，後ろ姿などでは識別ができない．そのため，顔識別により個人識別を行う場合，利用シーンが限られてしまうという問題がある．他の個人識別の研究として，人物のシルエット画像を用いて歩容を識別することで個人識別を行う手法 [2] がある．この手法では，対象の人物が鞄等の荷物を持っていたり，服装を変えたりすることで，シルエット画像による照合が失敗し，誤識別してしまう問題がある．そこで，本研究では骨格情報に着目する．骨格情報による個人識別はいくつか研究されており，その有用性が示されている [3, 4]．これらの手法では，いずれも RGB-D センサを用いており，3 次元の骨格情報から個人識別を行っている．本研究では，より一般的な単眼カメラからの骨格情報を用いた個人識別手法の構築を目指す．識別手法には，時系列データを扱うことができ，多くのデータから特徴を見つけ出して分類を可能とする深層学習を用いる．

また，本研究では個人識別を利用する対象として，入退出管理を考える．入退出管理では，個人の識別を行いつつ，部外者の認識も行うことが，セキュリティにおいて重要である．しかし，個人識別を多クラス識別問題として考えた場合，未知のクラスのデータが必ず既知のクラスに分類されてしまう問題がある．一般的に，未知クラスデータは入手困難であるため，未知クラスである部外者を部外者として学習することができない．そこで，本研究では，既知クラスのデータから疑似的に未知クラスのデータを生成し，部外者を部外者として認識する手法を提案する．

2. 個人識別手法

2.1 手法概要

本研究では，部外者の認識を行うために転移学習を用いる．転移学習とは，学習済みのモデルの一部を特徴抽出器として用いて，新たに別のネットワークを構成することである．つまり，本研究では，初めに既知のクラスのデータを用いて学習モデルを構成した後，入力層から中間層までを特徴抽出器として取り出す．次に，既知クラスのデータから疑似的に未知クラスのデータを生成した後，特徴抽出器に識別層を加えた新たなネットワークを用いて学習を行う．本研究では，一つ目に構成した学習モデルを第一学習モデル，二つ目に構成した学習モデルを第二学習モデルと呼ぶ．

2.2 骨格情報抽出

本研究では，Zhe らの人物姿勢推定手法 [5] を用いた OpenPose というライブラリを利用して骨格情報を抽出する．この手法は CNN を組み合わせて姿勢推定を行っている．画像を入力として，出力は特徴（肩，肘，膝等）の x, y 座標と認識の信頼度となっている．本手法では，COCO（Common Objects in Context）2016 のデータセットを教師データとした学習モデルを用いる．この学習モデルでは Fig. 1 に示す 18 点の特徴点を抽出できる．特徴点のデータ順を次式に示す．

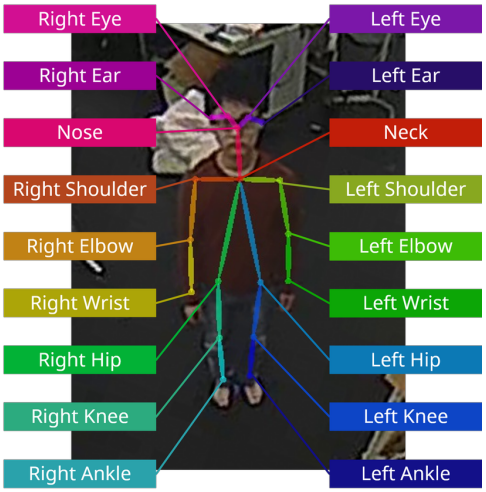


Fig.1 Skeleton information from OpenPose[5]

$$\boldsymbol{v} = \begin{bmatrix} x_1 & y_1 & c_1 & \cdots & x_{18} & y_{18} & c_{18} \end{bmatrix} \tag{1}$$

ここで x_n，y_n はそれぞれ n 番目の特徴点の x, y 座標である．また，c_n は特徴認識の尤度を表す．

2.3 データの前処理

深層学習への入力となるデータは 0 から 1 で正規化した方が一般によい学習結果が得られることがわかっている．また，x, y 座標は，人が画像中に写る位置によって変化してしまうため，識別の入力として適さない．そこで，前述の特徴 $\boldsymbol{v}$ に前処理を施す．初めに，人物の首の特徴点を中心として，それぞれの特徴点を極座標に変換する．次に，特徴 $\boldsymbol{v}$ を次式で表す特徴 $\boldsymbol{w}$ に変換する．

$$\boldsymbol{w} = \begin{bmatrix} r_1 & \theta_1 & c_1 & \cdots & r_{18} & \theta_{18} & c_{18} \end{bmatrix} \tag{2}$$

ここで，r_n，θ_n はそれぞれ首から特徴点へかけての動径と偏角を表す．ただし，動径と偏角の値は，それぞれ骨格抽出画像の対角線の長さと 2π で割って正規化する．

2.4 第一学習モデルの構築

第一学習モデルのネットワークの構成を Fig. 2 に示す．ネットワークは大きく 3 つに分かれている．以下，ネットワークの各構成とモデルの学習についてそれぞれ説明する．

2.4.1 CNN 層

一つ目は，CNN 層である．CNN 層の構成を Fig. 3 に示す．CNN 層は，二つの畳込み層と二つのプーリング層，一つの全結合層からなる．

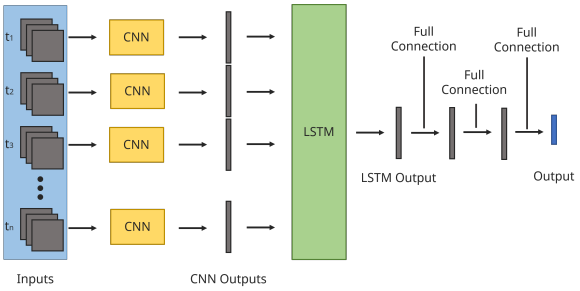


Fig.2 Constitution of first learned network

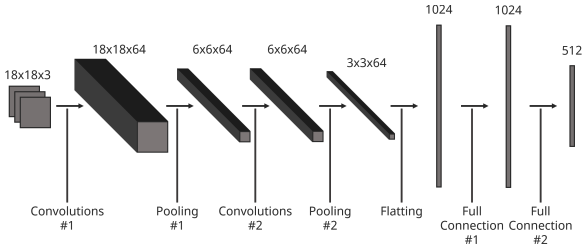


Fig.3 Constitution of CNN layers

入力は特徴 $\boldsymbol{w}$ を CNN の入力とするために変更した以下のものを用いる．まず，特徴 $\boldsymbol{w}$ の要素を次式で表すように種類別に分ける．

$$r = \begin{bmatrix} r_1 & \cdots & r_{18} \end{bmatrix} \tag{3}$$

$$\theta = \begin{bmatrix} \theta_1 & \cdots & \theta_{18} \end{bmatrix} \tag{4}$$

$$c = \begin{bmatrix} c_1 & \cdots & c_{18} \end{bmatrix} \tag{5}$$

次に，分けた特徴のそれぞれの直積を 3 チャンネルのデータとして用いる．これは $18 \times 18 \times 3$ の三階テンソル $\boldsymbol{F}$ であり，次式で表される．

$$\boldsymbol{F} = \begin{bmatrix} f_{1,1,n} & f_{1,2,n} & \cdots & f_{1,18,n} \\ f_{2,1,n} & f_{2,2,n} & \cdots & f_{2,18,n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{18,1,n} & f_{18,2,n} & \cdots & f_{18,18,n} \end{bmatrix} \tag{6}$$

ここで，$f_{i,j,n}$ は 3 チャンネル分のデータであり，次式で表される．

$$f_{i,j,1} = r_i r_j \tag{7}$$

$$f_{i,j,2} = \theta_i \theta_j \tag{8}$$

$$f_{i,j,3} = c_i c_j \tag{9}$$

この $\boldsymbol{F}$ を CNN の入力とする．

二つの畳込み層は，どちらも 5×5 のフィルタを一画素ずつストライドさせており，活性化関数は ReLU（Rectified Linear Unit）である．また，フィルタの数は，二つとも六十四個用意している．二つのプーリング層は，どちらも 3×3 のフィルタである．ただし，一つ目は三画素ずつ，二つ目は二画素ずつストライドさせている．全ての畳込み層とプーリング層でゼロパディングを行っている．全結合層では，$3 \times 3 \times 64$ 個のデータを 1 列のデータにフラット化した後に，1024 個のユニットに全結合させている．また，活性化関数には ReLU を用いる．過学習を防ぐために，プーリング層の後と全結合層の後にドロップアウトを行う．この CNN 層を複数並べ，時系列の骨格情報を入力として，次の層への入力とする．

2.4.2 LSTM 層

二つ目の層は，LSTM（Long short-term memory）[6] 層である．LSTM とは，時系列データを入力とできる RNN（Reccurent Neural Network）の一種である．本研究では，CNN で骨格情報から取り出された特徴の時系列データを LSTM への入力とすることで，人の歩き方の違いなどに着目した特徴抽出となる．

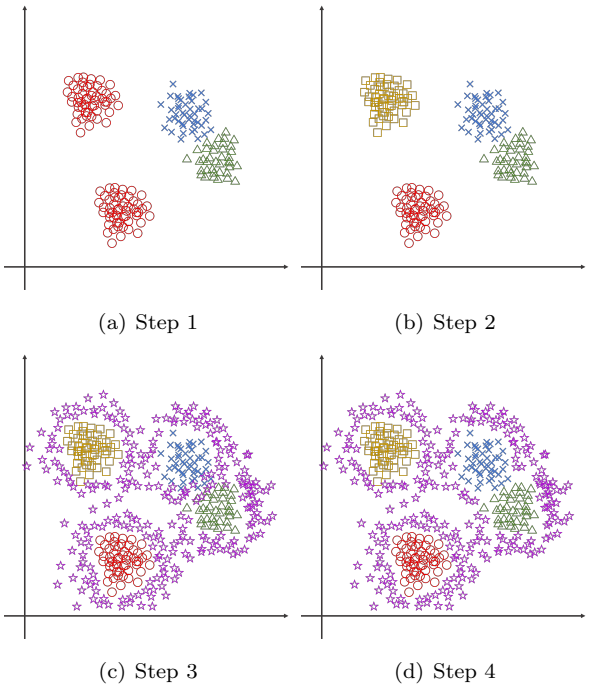


Fig.4 Generating unknown class data

2.4.3 識別層

三つ目の層は，識別層である．一つ目と二つ目の層で抽出した特徴を分類するために，全結合層を三つ繋げている．最後の全結合層はクラスの数だけユニットを用意する．活性化関数は最後の層以外では ReLU，最後の層では SoftMax 関数を用いている．

2.4.4 モデルの学習

損失関数には，ネットワークの出力と教師ラベルの交差エントロピーを用いる．また，最適化手法には，Adam（Adaptive Moment Estimation）[7] を用いている．学習時はミニバッチ学習を用い，学習が収束するまで行う．

2.5 特徴抽出器の取り出し

一般に，ネットワークの中間層は特徴の抽出を行っていると言われている．よって，学習モデルの中間層を取り出すことで，特徴抽出器として用いることができる．そこで，第一学習モデルが構成できたら，CNN 層と LSTM 層の部分を取り出し，これを特徴抽出器として用いる．以降，CNN 層や LSTM 層の重みやバイアスは変更しない．また，本研究では，この特徴抽出器で取得した特徴を深層特徴と呼ぶ．

2.6 未知クラスデータの生成

第一学習モデルでは，未知クラスのデータが入力されることを考慮していない．そのため，未知クラスのデータが入力されても必ず既知クラスのどれかに分類されてしまう．これは，既知クラスと未知クラスの間に識別境界が構成されていないことが原因であると考えられる．未知クラスの教師付きデータがあれば，第一学習モデルで部外者の学習が可能であるが，それらのデータの入手は困難である．そこで，既知クラスのデータを基に未知クラスのデータを生成し，その間に識別境界を構成することで部外者の認識を可能にする．未知クラスデータの生成は，以下の 4 ステップで行う．

2.6.1 ステップ 1

既知クラスのデータ群を前述の特徴抽出器に入力し，それぞれのクラスの深層特徴群を取得する（Fig. 4(a)）．Fig. 4(a) で

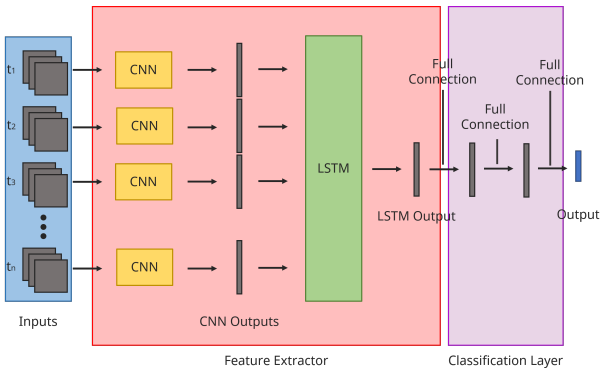
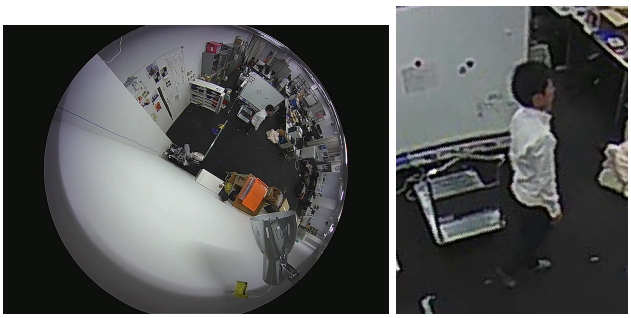


Fig.5 Constitution of second learned network



(a) Fisheye image (b) Corrected image

Fig.6 Correction of fisheye image

Fig.7 Targets of identification

は，2 次元の分布で示しているが，実際には 512 次元の特徴空間に分布している．

2.6.2 ステップ 2

取得したクラスごとの深層特徴群が等分散性を有している保証はない．そこで，次に，k-means 法によって深層特徴群のまとまった集団ごとに新たにラベル付けを行う（Fig. 4(b)）．

2.6.3 ステップ 3

次に，取得した深層特徴群の周囲に識別境界が構成されるように未知クラスデータを生成する．新ラベルごとに深層特徴群の要素の平均と標準偏差を計算する．次式が成り立つ要素 e を持つ深層特徴群を疑似的に生成する（Fig. 4(c)）．

$$e_{i,j} < \lambda_{i,j} - \sigma_{i,j} \tag{10}$$

$$e_{i,j} > \lambda_{i,j} + \sigma_{i,j} \tag{11}$$

ここで，i はクラスのラベル，j は要素の位置を表す．また，$\lambda_{i,j}$ と $\sigma_{i,j}$ はクラス i の深層特徴の j 番目の要素の平均と標準偏差である．

2.6.4 ステップ 4

生成した特徴が他のラベルの特徴と被っている場合，その特徴を削除する（Fig. 4(d)）．これにより，どのクラスにも属さないラベルだけを集めることができる．

2.7 第二学習モデルの構築

生成した未知クラスデータを用いて，部外者を認識できる学習モデルを構築する．ネットワークの構成を Fig. 5 に示す．入力から中間層までは第一学習モデルの一部から取り出した特徴抽出器で，中間層から出力層までが新たに学習を行う識別層である．ネットワークの出力は，既知クラスの数に部外者クラスを足した数だけある．各層の活性化関数やモデルの学習アルゴリズムは第一学習モデルと同様である．

3. 個人識別実験

3.1 実験条件

3.1.1 使用カメラ

本実験では，パナソニックの DG-SF438 という魚眼カメラを用いた．解像度は 1280×960 で使用した．カメラは床から高さ約 2.7[m] の位置に下向きに設置した．

魚眼カメラから取得した画像は，通常のカメラから取得した画像と比べて大きく歪んでいる（Fig. 6(a)）．そのため，魚眼画像をそのまま骨格推定の入力とすると，推定がうまくいかない．よって本研究では，取得した魚眼画像の人物の写った領域を次式で示す正像変換 [8] により歪みのない画像に変換した．

$$x = \frac{R(u\cos\alpha - v\sin\alpha\cos\beta + mR\sin\alpha\cos\beta}{\sqrt{u^2+v^2+m^2R^2}} \tag{12}$$

$$y = \frac{R(u\sin\alpha - v\cos\alpha\cos\beta + mR\cos\alpha\beta}{\sqrt{u^2+v^2+m^2R^2}} \tag{13}$$

ここで，x, y は魚眼画像での座標，u, v は変換後の座標である．また，R は魚眼画像の円の半径，α, β はそれぞれ方位角と仰角を表す．m は変換後の倍率を表し，次式で示す．

$$m = \frac{k}{\cos\beta} \tag{14}$$

これにより，遠くに写った人物が画像上で小さく見える分を補正するように画像が拡大される．以上の変換を行った画像（Fig. 6(b)）を骨格推定の入力とした．

3.1.2 使用データ

本実験では，Fig. 7 に示す六人を対象に識別を行った．六人全員が二十代の男性であり，極端な体格差はない．指定したルートを対象者が 7 往復歩いている動画を撮影し，それらの動画から骨格情報を推定し，データを用意した．データの 8 割を学習用，残りの 2 割を評価用のデータとした．また，六人の対象者のうち一名を部外者として扱い，学習時にはその部外者のデータを学習に使用しないこととした．

3.2 実験結果

初めに，第一学習モデルによる個人識別結果を混合行列で表したものを Fig. 8 に示す．混合行列では，対角成分が大きい値であるほど識別精度が高いことを表す．Fig. 8 では，対角成分がどれも大きく，適合率，再現率，F 値はそれぞれは 94.5%，93.9%，94.0%である．このことから，第一学習モデルによって骨格情報から個人を識別できていることがわかる．

次に，第二学習モデルによる個人識別結果を混合行列で表したものを Fig. 9 に示す．Fig. 9 について，0 から 4 までのラベルが既知のクラス，5 のラベルが未知のクラスが割り当てられている．全体での適合率，再現率，F 値はそれぞれは 87.8%，

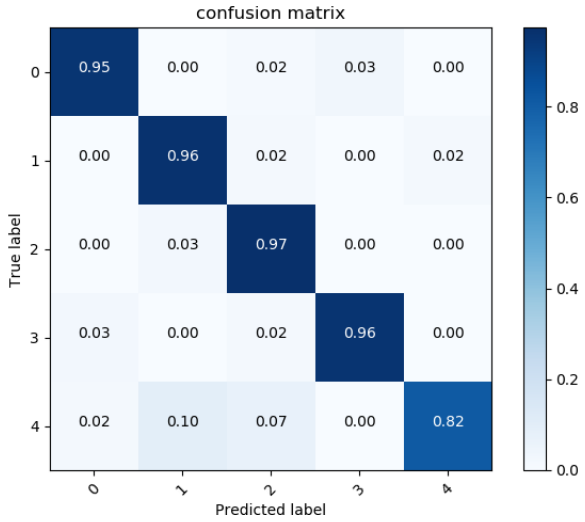


Fig.8 Confusion matrix of identification result by first learned model

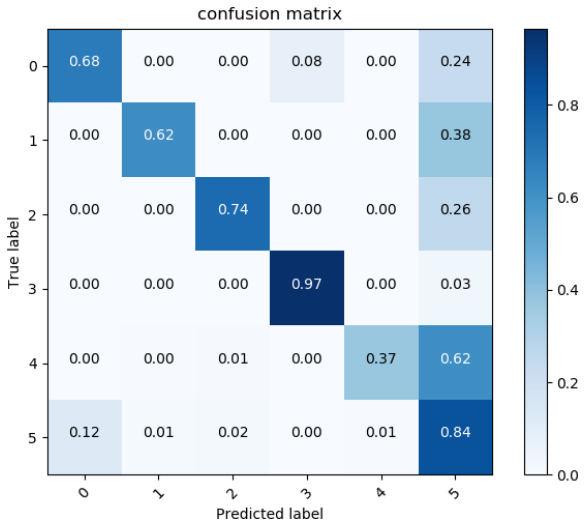


Fig.9 Confusion matrix of identification result by second learned model

69.9%，73.6%である．部外者のデータの 8 割以上が部外者として認識されており，提案手法により部外者を認識できていることがわかる．既知クラスのデータの結果について，第一学習モデルでの結果より精度は低下しているが，おおよそ識別できている．一部のクラスでは大幅に識別率が落ちてしまっている．これは，一部のクラスでは深層特徴の分布が複雑すぎて，疑似的に生成した未知クラスのデータに分類されるようになってしまったことが考えられる．しかし，ハイパーパラメータの再検討や，ラベル推定の尤度の結果から識別結果を改善できると考えている．

4. 結論

本研究では，骨格情報から深層学習によって個人を識別する手法を提案した．また，入退出管理において重要なタスクの一つである部外者の認識を行う手法を提案した．さらに，個人識別の実験を行い，提案手法の有効性を示した．今後は，本実験よりも既知のクラスを増やした場合の検証や，実際の入退出のシーンでの検証を行っていく．また，提案手法と顔識別や服装識別などを組み合わせて，認識率の高い入退出管理システムの構築を目指す．

参考文献

(1) Schroff, F., Kalenichenko, D., & Philbin, J., Facenet: A Unified Embedding for Face Recognition and Clustering, Proc. of the 2015 IEEE Conf. on CVPR, (2015) pp. 815-823.
(2) 槇原 靖, 佐川 立昌, 向川 康博, 越後 富夫, 八木康史, 周波数領域における方向変換モデルを用いた歩容認証, 情報処理学会論文誌コンピュータビジョンとイメージメディア, **48**-SIG1, (2007) pp.78-87.
(3) Sinha, A., Chakravarty, K., & Bhowmick, B., Person Identification Using Skeleton information from Kinect, Proc. Intl. Conf. on ACHI, (2013) pp. 101-108.
(4) Ball, A., Rye, D., Ramos, F., & Velonaki, M., Unsupervised Clustering of People from'Skeleton'Data, Proc. of the seventh annual ACM/IEEE international conf. on HRI, (2012) pp. 225-226.
(5) Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y., Realtime Multi-person 2d Pose Estimation Using Part Affinity Fields, Proc. of the 2017 IEEE Conf. on CVPR, (2017) pp. 1302-1310.
(6) Gers, F. A., Schmidhuber, J., & Cummins, F., Learning to forget: Continual prediction with LSTM, Conf. on ICANN, (1999) pp. 850-855.
(7) Kingma, D., & Ba, J., ADAM: A Method for Stochastic Optimization, Proc. of ICLR2015, (2015).
(8) 森 隆寛, 外村 元伸, 大住 勇治, 池永 剛, キュービック補間を用いた魚眼レンズ画像の高画質補正アルゴリズム, 情報科学技術フォーラム一般講演論文集, **5**-1, (2006) pp.7-8.