Vol.2018-EC-50 No.17
2018/12/21

歩行足音を用いた人手を介さない個人識別システムの初期的検討

堀 佑貴[1,a)]　安藤 崇央[2,b)]　福田 晃[2,c)]

概要：近年，歩行足音から個人識別を行うための研究が進められている．従来の研究では連続する足音から 1 歩分の足音を手動で切り出す必要があるなど，足音解析に人手を介す必要があった．これに対し本稿で提案するシステムでは，1 歩分の足音の波形を自動的に切り出すためにルールベースの手法を採用した．この機械的に切り出した波形に対して，メル周波数スペクトログラムを算出し 4 種の識別方法を用いてその識別精度の比較を行なった．その結果，水増しデータも学習に用いた CNN を利用する事で人手による足音の切り出しが必要な従来手法と同等以上の精度で識別可能であったことを報告する．

キーワード：足音，個人識別，CNN(Convolutional Neural Network), Augmentation, NMF(Non-negative Matrix Factorization)

1. はじめに

近年，人々の歩行方法から人物の推定を行う歩行認識と呼ばれる研究が盛んである．歩行認識によって人物の推定が可能になれば，セキュリティシステムの一環への導入が期待でき，性別や年齢などの情報まで認識できれば，新たな入客情報収集システムとしての活用も期待できる．

歩行認識に関するこれまでの研究は，そのほとんどが視覚情報やセンサ情報を用いたものであり，音響情報のみを用いた人物識別に関する研究は少ない．しかし，歩行に基づく音響情報のみから人物識別を行う事ができれば、カメラを用いた手法に比べ，屋内監視におけるプライバシーに関する問題に考慮する事ができる．また，より安価なハードウェアで暗闇にも対応できるという利点があるため，音響情報のみの歩行認識について研究を行うことは重要であると考える．これまでに行われたいくつかの研究では，この研究領域に対する有用性について確かめられている．

まず，[1] ではメル周波数ケプストラム解析，歩行間隔，およびスペクトル包絡線の類似度を特徴量として使用し，k 平均法を用いて 5 人の被験者に対し分類を行っている．[2] では [1] の特徴量の他に，ラウドネス，シャープネス，ゆらぎの強さ，粗さのような心理音響的特徴を加え識別精度が向上するか検証している．また，[3] では動的時間伸縮法を用いて 10 人の被験者に対し認識精度の検討を行っている．

また，他にも 1 歩分の足音で切り出した足音信号波形に対し，パワースペクトルのスペクトル包絡によって分類を行ったもの [4]，STFT(Short-Time Fourier Transform) を特徴量としたもの，およびウェーブレット変換と SVM(Sapport Vector Machine) を用いたもの [5] など様々な報告がなされている．これらの研究では，最適な特徴量の抽出についてが議論の中心となっており，足音信号波形を検出し切り出すまでの流れまで含めて言及されたものは少ない．多くは足音開始時間を既知としていたり，人手を介し足音信号波形について切り出しを行ったデータを用いて検討を行っている．

一方，1 歩分の足音信号波形を検出する方法について議論された研究としては [6] があるが，これは人物推定を目的としたものではなく，様々な環境音のコーパスにおける 1 歩分の足音の検出までに留まっている．厳密に 1 歩分を切り出された足音信号を入力とする場合には高精度な識別が可能な手法であっても，足音信号波形の検出や切り出し方の厳密さが失われると識別精度が低下することが十分に考えられる．機械的・自動的に足音信号を検出・切り出しを行うシステムでは，人手を介す場合に比べ検出・切り出し精度が下がることが容易に想像される．したがって，歩行足音を用いた人手を介さない個人識別システムを実現す

[1] 九州大学大学院 システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
[2] 九州大学大学院 システム情報科学研究院
Faculty of Information Science and Electrical Engineering, Kyushu University
[a)] hori@f.ait.kyushu-u.ac.jp
[b)] ando.takahiro@f.ait.kyushu-u.ac.jp
[c)] fukuda@f.ait.kyushu-u.ac.jp

表 1　録音条件

場所	九州大学 伊都キャンパス 総合学習プラザ
床	木造
被験者	20 代男性 3 名，40 代男性 1 名
履物	異なるスニーカー
マイクロフォン	audio-technica AT-VD6
測定時間	学習用データ 60 秒，検証用データ 20 秒
サンプリング周波数	44100Hz

るには，足音信号波形の検出および切り出しによる影響まで考慮した識別手法を採用する必要がある．

そこで本稿では，厳密な 1 歩分の足音の切り出しを必要としない個人識別可能なシステム実現のための初期的検討として，4 種の識別手法について検討を行う．

本稿の構成は以下の通りである．2 節では足音の採取方法と足音の検出および切り出し手法について示し，3 節では得られた 1 歩分の足音信号波形に対しどのような特徴量を得るか検討する．4 節では本稿で比較を行う 4 つの識別手法について示し，5 節では実験による識別結果の比較を行い，提案するそれぞれの識別手法に対する評価を示す．最後に 6 節でまとめとする．

2. 足音検出手法

本節では，足音の採取方法と足音の検出および切り出し手法について示す．

2.1 足音の採取

足音を採取した際の条件について表 1 に示す．また，録音風景について図 1 に示す．採取場所の床から 5cm の場所にマイクロホンを固定し，被験者にその場で歩行してもらい足音を採取した．被験者は 22〜24 歳の男性 3 名と 40 代男性 1 名の計 4 人である．audio-technica AT-VD6 を用いて，オーディオインターフェースを経由し PC で録音した．また，それぞれ異なる履物（スニーカー）を履いて録音を行った．録音データに関して，一人あたり 60 秒の歩行音声を学習データとし，それとは別に 20 秒分の歩行音声を録音し，最終的な評価に用いている．採取した足音信号波形の例として，40 代男性の学習データの最初の 5 歩分を図 2 に示す．

2.2 1 歩分の足音検出手法

録音した足音信号波形から，足音の部分とそうでない部分を分離する方法として，実効値を用いたしきい値処理による手法を用いる．録音した音声波形を $x(t)$，その時間長を T とした場合の実効値 x_{rms} は以下の式で表される．

$$x_{rms} = \sqrt{\frac{1}{T}\int_0^T x(t)^2 dt} \qquad (1)$$

振幅値が，式 1 により算出された実効値より下回り 100ms

図 1　録音風景

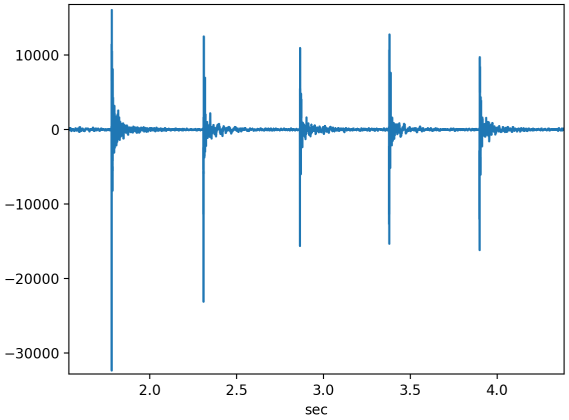

図 2　40 代男性による 5 歩分の足音波形

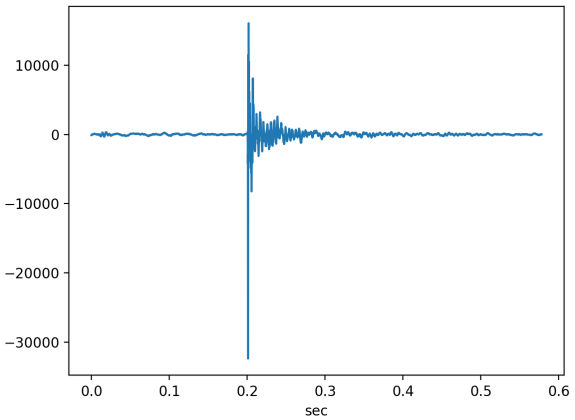

図 3　切り出した 1 歩分の足音波形

経過した時点で 1 歩分の足音として音声の切り出しを行なう．この手法を用いることで，機械的に 1 歩分の足音を検出でき，1 歩分の足音の切り出しを自動化できる．例として図 2 に示した波形から，提案した手法によって 1 歩分を取り出した際の波形を図 3 に示す．

3. 特徴量の選択

本節では 2.2 節で得られた 1 歩分の足音信号波形に対しどのような特徴量を得るか検討する．

3.1 特徴量の選択

[5] では周波数スペクトルの時間変化に着目した解析によって，1 歩分の足音に対して被験者や履物および場所による差異が見られた事が報告されており，[7] では人間が歩行音によって個人を識別できる事が示されていた．

そこで本研究ではこれらを参考にし，周波数スペクトルの時間変化に着目し，人間の聴覚特性を考慮した特徴量としてメル周波数スペクトログラムを採用する．スペクトログラムとは，信号波形について短時間フーリエ変換を行い，時間ごとに並べたものである．縦軸は周波数，横軸が時間を表し，濃淡が振幅値を表す．

今回は 23ms の時間窓長で 1/8 ずつオーバーラップを行いスペクトログラムを算出した．このスペクトログラムに対してメルフィルタバンクを適応し対数を取ることで，人間の音高知覚に調整した特徴量であるメル周波数スペクトログラムを算出する．図 3 に対するメル周波数スペクトログラムを算出した結果を図 4 に示す．

3.2 非負値行列因子分解 (NMF) を用いた特徴量抽出

3.1 節で特徴量としてメル周波数スペクトログラムを採用すると述べた．しかし，メル周波数スペクトログラムを単純に入力とすると，特徴量の次元が大きく 4 節で検討する手法ではうまく識別ができない可能性が考えられる．そこで，重要な周波数スペクトルだけの時間変化を表す特徴量を得ることを考え，そのための手法として本稿では非負値行列因子分解 (Non-negative matrix factorization：NMF) を用いた特徴量抽出を行う．

NMF は，有用な特徴量を抽出することを目的とする教師なし学習手法である．実世界には画像やパワースペクトルなど，非負の値によって表されるデータが多く，主成分分析などの多変量解析では，所与のデータを複数の加法的な成分に分解することを目的とするのと同様，NMF もそれぞれのデータが非負値であるという前提を置いた上で分解を行う．NMF は顔画像からパーツを抽出することを目的とした手法であったが，近年ではスペクトログラムを画像と見立てることで，音源分離 [8] や自動採譜 [9] など様々な領域で適用されている．NMF では与えられたデータ行列 $X = [x_1, \cdots, x_N] \in \mathbb{R}^{M \times N}$ に対して，指定した特徴数 K のもとで基底ベクトルを並べた非負の行列 $H = [h_1, \cdots, h_K] \in \mathbb{R}^{M \times K}$ と非負の結合係数行列 $U = [u_1, \cdots, u_N] \in \mathbb{R}^{K \times N}$ により

$$X \simeq HU \tag{2}$$

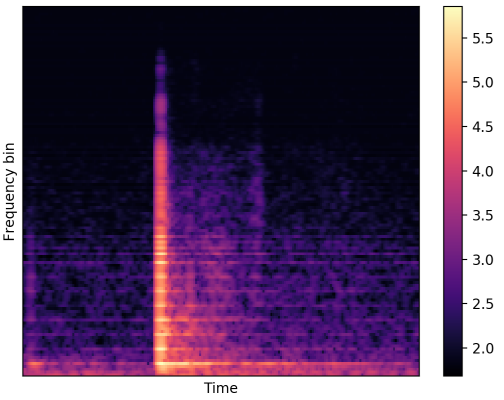

図 4　1 歩分のメル周波数スペクトログラム

で近似する行列を求める手法である．指定する特徴数は一般に与えられたデータ行列 X のランクより低いものである．そのため，NMF ではより低いランクの行列でデータ行列 X を近似することになり，その場合に求まる基底行列と係数行列が重要な意味をもつとして考える．スペクトログラムに NMF を適用する場合，H は頻出する周波数を表す基底スペクトルを，U は各基底スペクトルに対する時間ごとのアクティビティを表す．

3.1 節で述べた通り，1 歩分の足音において，周波数スペクトルの時間変化に着目することで差異が見られたことから，本研究では NMF において各基底スペクトルに対する時間ごとのアクティビティを表す U について特徴量として抽出する．また，予備実験の結果，特徴数を 5 に設定した場合に最良の結果が得られたことから，本稿では特徴数を 5 個に設定した NMF を採用している．

4. 識別手法

本節では，検討する識別手法について述べる．本稿では SVM を用いた識別手法と，CNN を用いた識別手法の二つの識別手法について検討を行う．そのうち，SVM を用いた識別手法では，メル周波数スペクトログラムをそのまま入力とした場合と，NMF を用いて特徴量の次元削減を行ったものを特徴量として使用する場合の 2 通りの手法を検討した．CNN を用いた識別手法では，2.2 節で取得した 1 歩分の足音信号波形についてデータの水増しを行っていないものと行ったものに対して，メル周波数スペクトログラムを入力とした 2 通りの手法で検討を行う．

4.1 SVM を用いた識別手法

SVM はクラス分類手法においてよく知られたものであり，画像認識やテキスト分類，音声認識などの多様な分野で優れた性能を示すことが報告されている．音響データである歩行足音を入力とする識別システムにおいても，高い

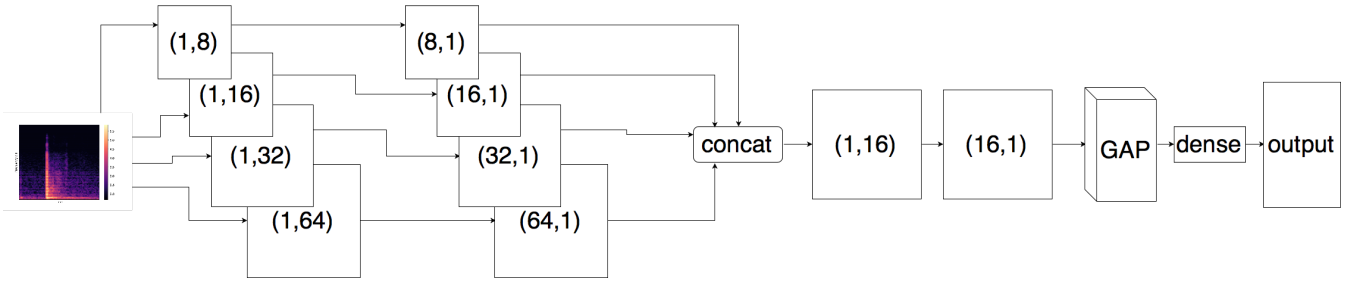

図 5 本稿で提案する CNN アーキテクチャ

精度の識別が期待できることから，本稿においても SVM を用いた識別手法を実装し，その識別精度についての検討を行う．

SVM は，学習データ $\boldsymbol{x}_i$ とそれに対応する正負のラベルをもつ y_i をペアとした n 個の訓練データに対して，重み $\boldsymbol{w}$ を用いて次式の超平面を求める 2 値線形分類器である．

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^T \cdot \boldsymbol{x} + b \tag{3}$$

ここで b はバイアス項である．この分類超平面と学習データが最も接近する場合のマージンについて最大化する分類超平面を求める．マージンの最大化は，式 (5) の条件で式 (4) を最大化する双対問題を解くことで達成される．

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} y_i y_j \alpha_i \alpha_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{4}$$

$$\sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \leq a_i \leq C \tag{5}$$

ここで C は正の定数であり，式 (4) の $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ は Kernel 関数と呼ばれ次式で表される．

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{\Phi}(\boldsymbol{x_i}) \cdot \boldsymbol{\Phi}(\boldsymbol{x_j}) \tag{6}$$

SVM は 2 値線形分類器であるが，非線形変換 $\boldsymbol{\Phi}$ を用いて高次元空間に写像しその高次元空間での線形分離を考えることで実質的な非線形分離を行う．この Kernel 関数を用いることで，莫大な計算を避け元の空間で直接解くことができる．Kernel 関数には線形カーネルや多項式カーネル，シグモイドカーネルなど多様な種類が存在する．本稿においてメル周波数スペクトログラムを入力とした識別手法では線形カーネルを，NMF を用いた識別手法では Radial Basis Function(RBF) カーネルを用いた．これら識別手法に対するカーネル関数の決定は，グリッド探索を用いて他のカーネル法と比較し，最良の結果を示したことから採用した．最終的に学習データではない未知のデータ $\boldsymbol{x}$ に対する分類は，以下の式 (7) によって決定される．

$$f(\boldsymbol{x}) = sgn(\sum_{i=1}^{l} (\alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x})) + b) \tag{7}$$

以上，本稿で識別精度を検討する SVM を用いた識別手法についてまとめる．本稿では以下の 2 種について検討を行う．

- Kernel 関数に線形カーネルを用いた，メル周波数スペクトログラムを入力とする SVM による識別手法
- Kernel 関数に RBF カーネルを用いた，メル周波数スペクトログラムから NMF により抽出した特徴量を入力とする SVM による識別手法

4.2 CNN を用いた識別手法

CNN(Convolutional Neural Network) は画像認識などの領域で，近年よく用いられる手法である．音声認識の分野においても，環境音の認識 [10] など幅広い対象に適用されており，これまでの従来手法に比べて高い精度を示すことが多い．一方で，音響情報のみを用いた足音からの個人識別に対して CNN を用いた研究は今まで行われておらず，本稿ではこの有用性についても含め検討を行う．

一般に，CNN の入力には時系列データをそのまま入力する場合と，スペクトログラムに変換したものを入力とする場合の 2 通りが主に考えられる．今回は，他手法との検討のしやすさも考えメル周波数スペクトログラムを入力として識別精度を求めた．図 5 に本稿で使用する CNN アーキテクチャを示す．今回用いるアーキテクチャでは，入力であるメル周波数スペクトログラムに対し，それぞれ長さの違うフィルタにより畳み込みを行なった後，それらを結合しさらに畳み込み，分類を行う．周波数と時間の両方に対し畳み込む長さを変えたこのようなフィルタを適用することで，より頑健な識別が行えるよう考慮した．

また，本研究では被験者が 4 人と少なく学習データの不足が考えられたため，モデルにおけるパラメータ数の削減を狙い Pooling 層には Global Average Pooling(GAP) を使用し，その後 Softmax 関数を繋ぎ分類を行なっている．モデルの学習に関して，評価関数には Multi-class logarithmic loss を，最適化関数には Adam を用いている．

さらに，本稿では 3.3 節で述べた通り，識別精度を更に向上させるために水増ししたデータを用いて学習を行った CNN についても検討する．データの水増しでは 1 歩分の足音を切り出した音声に対し，録音時の雑音の影響を考慮しガウス雑音の付与を行ったもの，微妙な接地時間のずれ

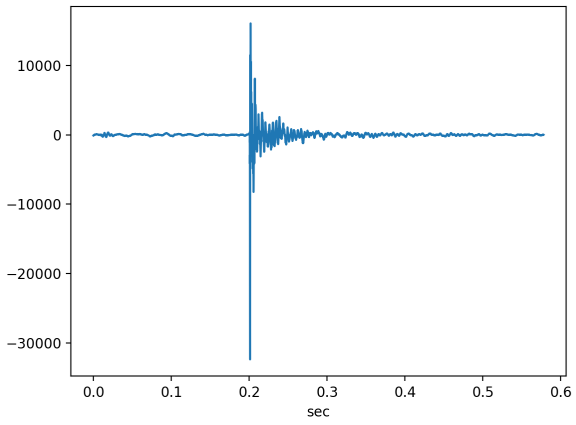

図 6　ガウス雑音を追加した足音波形

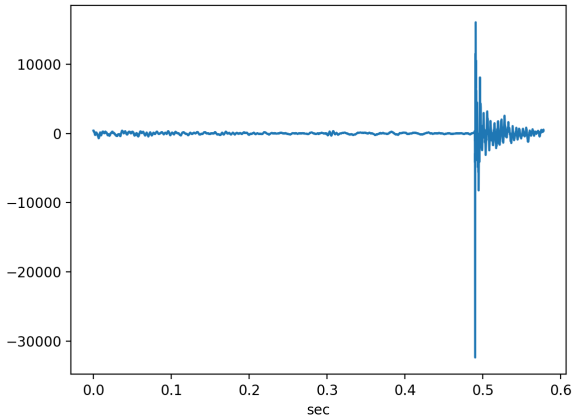

図 7　時間シフトを行った足音波形

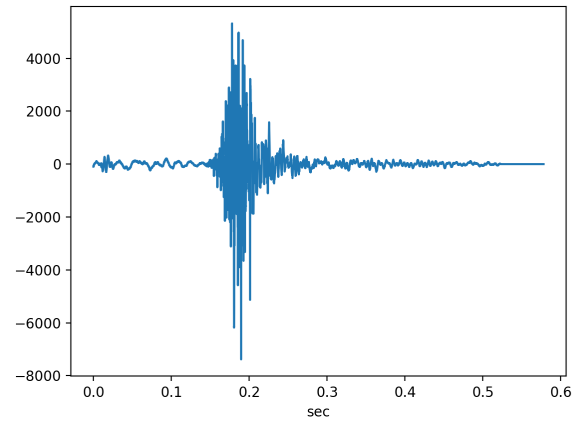

図 8　タイムストレッチを行った足音波形

を考慮し 1 歩分の足音を含む領域に対しタイムストレッチを行ったもの，1 歩分の足音切り出しにおける切り出し位置のバラつきを考慮し時間軸をシフトさせたもの，の 3 通りの手法に加え，ガウス雑音を付与しさらに時間軸をシフトさせたものの 4 通りのデータの水増しを行った．ガウス雑音を付与したものでは，付与したガウス雑音の強度の違いにバリエーションを持たせた 30 通りの音声と，タイムストレッチを行ったものでは 0.8 から 1.2 倍の間で 10 通りの音声を作成した．また，時間軸のシフトをしたものでは，それぞれ 1 から 100 でデータ数を割りその商で折り返した 100 通りの音声と，ガウス雑音と時間シフトの両方をおこなったものでは，一定のガウス雑音を付与し 2 から 11 の除算の商によって折り返した 10 通りの音声を作成した．それぞれの 3 の音声について水増しを行った結果の音声波形を図 6, 7, 8 に示す．これにより 1 歩分の足音音声から合計 200 個の模擬足音音声の生成を可能にした．また，この水増しを行ったデータに対して mixup[11] を用いて更にデータの水増しを行った．

表 2　それぞれの手法による識別精度

	A	B	C	D
SVM のみ	0.97	0.83	0.3	1.0
NMF+SVM	0.82	0.52	0.70	0.95
CNN	0.97	0.93	0.33	1.0
Augmentation+CNN	1.0	1.0	0.94	1.0

以上，本稿で識別精度を検討する CNN を用いた識別手法についてまとめる．本稿では以下の 2 種について検討を行う．

- 各被験者の 60 秒分の足音音声に対するメル周波数スペクトログラムのみを学習した CNN
- 各被験者の 60 秒分の足音音声に対するメル周波数スペクトログラムに加え，1 枚につき 200 枚のデータ水増しおよび mixup によるデータ水増しを行い学習を行った CNN

5. 識別結果

前節で示した 4 つの識別手法について，被験者 4 人における新たに録音した 20 秒分のデータに対する識別精度を比較する．使用したデータは学習データと同じように，2.2 節で述べた方法によって 1 歩分の音声データとして切り出しを行い，それをメル周波数スペクトログラムに変換し入力とした．被験者 4 人それぞれの，20 秒の間に観測された足音全てに対し識別を行い，以下の式 (8) によって識別精度を算出した．ここで N_i はそれぞれの被験者に対する観測された足音数であり，N_i^{true} は正確に識別できた足音数である．

$$(accuracy) = \frac{N_i^{true}}{N_i} \tag{8}$$

これによって算出したそれぞれの識別精度について表 2 に示す．ここで 20 代男性を A, B, C，40 代男性を D と表す．結果を見ると，水増しありの CNN を用いた手法がどの被験者に対しても最も良い精度を示しており，被験者 3 人に対して完全に識別ができていることがわかる．一方，完璧には識別ができなかった被験者についても，他手法では著

しく識別精度が落ちており，4 つの識別手法の中では格段に良い精度を記録した．これは足音音声の切り出しを手動で行っていた従来手法 [3],[4],[12] と比較しても同等以上の精度であり，この手法に対する有用性が確認できる．また，メル周波数スペクトログラムをそのまま入力とするよりも，NMF で特徴量の抽出を行ったり学習データの水増しを行うことで全体的な精度の向上が見られることがわかった．一方，CNN による特徴量抽出だけでは足音の切り出し方による影響をあまり低減できず，結果的に水増しありの CNN と比べると識別精度が低くなることがわかった．

6. まとめ

本稿では人手を介さない歩行足音による個人識別システムの初期的な検討として，しきい値処理により自動的に切り出しを行った 1 歩分の足音に対して 4 つの識別手法について比較検討を行なった．被験者 4 人の 20 秒分の足音について識別を行なった結果，水増し学習を行った CNN が最良の結果を示し，従来手法と比較しても同等以上の認識精度が得られた．本稿で提案した手法は，足音の切り出し位置の同定を必要としない．したがって，この手法を用いることで人手を介さずに自動で足音から個人識別を行うシステムの実現が可能であると考えられる．

今後の課題として，被験者を増やした際の識別精度の検証，靴や環境など違いによる識別精度への影響の検証，また，ある特定の人物のみ認識を行うシステムについての検討などが考えられる．

謝辞 本研究は，科研費 JP15H05708 の助成を受けたものである．

参考文献

[1] Yasuhiro, S., Takasuka, T. and Yasukawa, H.: Personal identification using footstep detection (2004).
[2] Itai, A. and Yasukawa, H.: Footstep Recognition with Psyco-acoustics Parameter, *APCCAS 2006 - 2006 IEEE Asia Pacific Conference on Circuits and Systems*, pp. 992–995 (2006).
[3] Itai, A. and Yasukawa, H.: Footstep classification using simple speech recognition technique, *2008 IEEE International Symposium on Circuits and Systems*, pp. 3234–3237 (2008).
[4] 田中元志，井上　浩：足音スペクトルの比較による木造家屋内の歩行認識に関する一検討，電気学会論文誌. C， Vol. 122, No. 3, pp. 525–526（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/130006845751/⟩ (2002).
[5] 安田浩大，田中元志，井上　浩：周波数スペクトルの時間変化に着目した足音の解析，計測自動制御学会東北支部第 237 回研究集会, 2007,（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/10029998326/⟩ (2007).
[6] She, B.: Framework of footstep detection in indoor environment, *ICA2004*, (online), available from ⟨https://ci.nii.ac.jp/naid/10019715622/⟩ (2004).
[7] Mäkelä, K., Hakulinen, J. and Turunen, M.: ICAD03-144 THE USE OF WALKING SOUNDS IN SUPPORTING AWARENESS, *Proceedings of International Conference on Auditory Display*, pp. 144–147 (2003).
[8] Ozerov, A. and Févotte, C.: Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation, *IEEE Trans. Audio, Speech & Language Processing*, Vol. 18, No. 3, pp. 550–563 (online), available from ⟨https://doi.org/10.1109/TASL.2009.2031510⟩ (2010).
[9] 亀岡弘和：非負値行列因子分解とその音響信号処理への応用，日本統計学会誌，Vol. 44, No. 2, pp. 383–407（オンライン），DOI: 10.11329/jjssj.44.383 (2015).
[10] Piczak, K. J.: Environmental sound classification with convolutional neural networks, *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6 (2015).
[11] Eaton-Rosen, Z., Bragman, F., Ourselin, S. and Cardoso, M. J.: Improving Data Augmentation for Medical Image Segmentation, *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, pp. 1–3 (2018).
[12] 板井陽俊，安川　博：ケプストラムと DTW を用いた歩行足音識別，情報処理学会研究報告数理モデル化と問題解決（MPS），Vol. 2007, No. 86, pp. 61–64（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/110006403652/⟩ (2007).