Vol.2016-CVIM-202 No.25
2016/5/12

人体部位の相対的位置関係を利用したオノマトペ歩容映像の識別に関する検討

加藤 大貴[1,a)]　平山 高嗣[2,1]　川西 康友[1]　道満 恵介[4]　出口 大輔[3,1]　井手 一郎[1,2]　村瀬 洋[1,2]

概要：オノマトペは，物事の様子を端的に表現する言葉として，口語表現で頻繁に使用される．直感的な印象を計算機に伝える手段としても有効であると考えられ，インタフェースの入出力手段として利用する研究が盛んになりつつある．従来研究では，主に擬音語や，質感を表す静的なオノマトペが対象とされ，動きを表すオノマトペに関する研究は少ない．そこで，本報告では，人間の歩行動作が多様なオノマトペで表現されることに着目し，オノマトペ表現に対応した歩容映像を識別する手法を検討する．具体的には，人体部位の相対的位置関係に基づく特徴を利用し，オノマトペ表現ごとに 2 クラス識別器を構築する．

1. はじめに

口語表現では「のろのろ」,「つるつる」,「しゃかしゃか」など，事象の様子を直感的に表現する言葉として，オノマトペが使用される．オノマトペは擬音語および擬態語と呼ばれている言語表現の総称である [1]．他の言語と比べ，日本語はオノマトペの種類が圧倒的に多く，その使用範囲も多岐にわたり，多用されることが知られている．

オノマトペは，その音響的印象が事象の様態と対応しているため，人間はオノマトペに対して共通のイメージを想起するとされている [2]．そのため，オノマトペは論理的な表現が容易ではない直感的な印象を端的に他者に伝えるための有効な手段であると考えられている．

また，オノマトペは直感的な印象を計算機に伝える手段としても有効であると考えられており，近年オノマトペを直感的なインタフェースの入出力手段として利用する研究が盛んになりつつある [3], [4]．神原ら [3] は，「オノマトペン」という描画システムを開発している．これは例えば，「ぎざぎざ」と発話しながら線を描くことで，「ぎざぎざ」な線を描くことができるインタフェースであり，利用者に直感的な操作環境の提供を実現している．また，小松ら [4] は，利用者がオノマトペを入力すると，そのオノマトペに合致したロボットの動作記述作業を自動で行なうシステムを開発し，一般人には敷居が高い作業を直感的に行なえるようにしている．このように，オノマトペをインタフェースの入出力手段として用いることの有用性が示されつつある．

オノマトペを用いてより幅広い応用を実現するためには，画像，映像，音声のような，一般に広く普及している視聴覚メディアとオノマトペが対応付けられることが望ましい．このうち音声については，オノマトペの中でも擬音語が特に関連深く，音響分野において，音響信号と擬音語を対応付ける研究が行なわれている [5], [6]．また，画像は質感を表すオノマトペ（「つるつる」,「さらさら」など）と特に関連深く，画像特徴を用いて画像と質感を表すオノマトペを対応付ける研究が行なわれている [7]．一方，映像については動きを表すオノマトペが関連深いと考えられるが，これに関する研究はあまり行なわれていない．しかし，例えば多数の人間が映っている監視カメラ映像から，「ふらふら」歩行している人を検出したいという状況を想定した場合，映像を入力としてオノマトペ表現に対応した人の動きを識別する技術が求められる．このように，動きを表すオノマトペを用いた応用では，映像とオノマトペを対応付けて識別する技術が必要であることが想定される．

これに関連して，映像とオノマトペの関係性を被験者実験により分析した研究は従来から行なわれている．鍵谷ら [8] は，液体の粘性に注目し，CG 映像作成ソフトウェアを用いて，映像作成時の動粘度パラメータと，作成された映像から想起されるオノマトペを構成する音韻の種類に関連性があることを明らかにしている．また，映像から切り

[1] 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University
[2] 名古屋大学実世界データ循環学リーダー人材養成プログラム
Graduate Program for Real-World Data Circulation Leaders, Nagoya University
[3] 名古屋大学情報戦略室
Information Strategy Office, Nagoya University
[4] 中京大学工学部
School of Engineering, Chukyo University
[a)] katoh@murase.m.is.nagoya-u.ac.jp

出された静止画像から想起されるオノマトペについても同様に分析し，映像と静止画像では関連性がある音韻の種類が異なるという知見を示している．杉山ら [9] は，犬型ロボットの歩行シミュレータを用いて，被験者にオノマトペを表現したロボットの歩行パターンを設計させる実験を行なっている．その結果として，擬態語オノマトペ（「ぴょんぴょん」，「てくてく」など）の動きを設計する場合は，個人によらず似たような歩行動作となることを明らかにしている．また，動きに対応したオノマトペの種類を人間が判別するためには，肩と足，右足と左足など，体の相対的な運動を見ることが重要であると示唆している．しかしこれらの研究は，あくまで心理的観点から映像とオノマトペの関係を分析した研究であり，オノマトペに対応した映像の機械的な識別を可能にするものではない．

また，動きを表すオノマトペは歩行動作と関連が深い [1]．そのため，歩行動作を分析することで，動きを表すオノマトペに対応する一般的な動き特徴を抽出できる可能性がある．そこで本研究では，オノマトペ表現に対応した映像として，人間が歩いている様子である「歩容」の映像に注目する．そして，オノマトペの種類を識別するためには，体の相対的な運動を見ることが重要という杉山ら [9] の知見に基づいて，動画像解析による特徴抽出とその機械学習により，オノマトペ表現に対応した歩容映像を識別する．なお，歩容の識別に関する従来研究の多くは，年齢性別などの属性認識を目的としたものであり [10], [11]，オノマトペの識別を試みた研究例は存在しない．

「オノマトペ表現に対応した歩容」の定義は 2 通り考えられる．1 つは，歩行者本人が「すたすた」歩いているつもりの歩容，とする主観的定義である．もう 1 つは，歩容を見た第 3 者が「すたすた」歩いていると感じた歩容，とする客観的定義である．本研究では，擬態語オノマトペは異なる人に対しても同じ印象を想起させるという杉山ら [9] の知見に基づき，歩行者全員がオノマトペに対して共通の印象に基づいて歩容を表現することができると仮定し，歩行者による主観的な歩容を「オノマトペ表現に対応した歩容」と定義する．

2. オノマトペ歩容映像の識別手法

一般に，歩容映像は複数のオノマトペ表現と対応付く可能性がある．例えば，「のろのろ」していて，かつ「ふらふら」しているような場合である．このような事例を考慮すると，未知の歩容映像が与えられたとき，オノマトペ表現と歩容映像を 1 対 1 で対応付けるよりも，その歩容映像が対応する全てのオノマトペ表現を識別できる方が良い．そのためには，ある特定のオノマトペ表現に対応した歩容映像を識別する 2 クラス識別器を複数個用意する方法が考えられる．そこで本研究では，未知の歩容映像がある特定のオノマトペ表現と対応しているか否かを識別する 2 クラス識別器を構築することを目的とする．

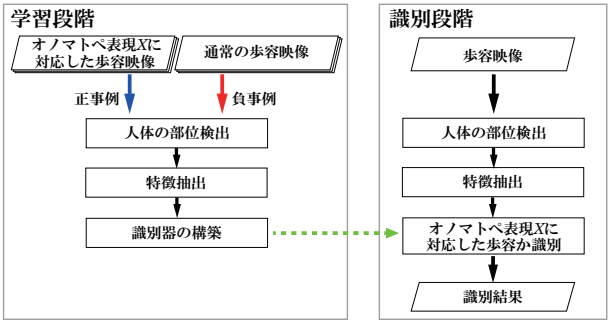
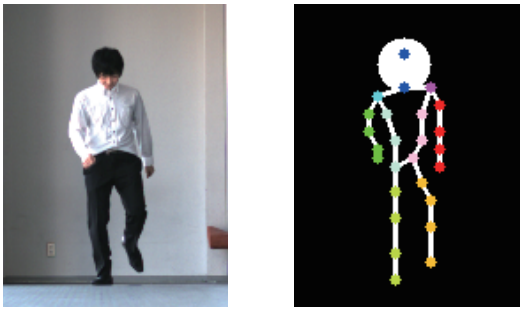


図 1　提案手法におけるオノマトペ表現 X の識別手順



(a) 元の映像フレーム　(b) 部位検出結果

図 2　FMP による部位検出結果の例

本手法では，識別対象とするオノマトペ表現が N 種類の時，2 クラス識別器を N 個構築することによって，歩容映像を識別する．識別手順は，学習段階と識別段階の 2 段階からなる．提案手法における，あるオノマトペ表現 X に関する 2 クラス識別手順を図 1 に示す．ここで，オノマトペ表現 X は N 種類のオノマトペ表現のうちの 1 つを表す．学習段階では，各オノマトペ表現に対応する歩容映像から特徴を抽出し，2 クラス識別器を構築する．識別段階では，入力された歩容映像から特徴を抽出し，学習段階で構築した N 個の識別器を用いて，対応するオノマトペ表現を識別する．

個々の 2 クラス識別器は，入力された歩容映像がある特定のオノマトペ表現に対応しているか否かを識別する．これらの 2 クラス識別器によって，入力映像と N 種類のオノマトペ表現の対応関係について重複を許して求める．本手法では，人体の周期運動に基づく特徴と，人体部位の相対的位置関係に基づく特徴を用いる．以降，各処理について詳述する．

2.1 学習段階

2.1.1 人体の部位検出

本手法では，人体部位の相対的位置関係に基づく特徴を用いる．そのために，事前処理として映像から人体の部位を検出する．ここでは，部位検出処理に Flexible Mixture of Parts （FMP）[12] を利用する．FMP は，人検出と姿

勢推定を同時に行なう手法であり，入力画像に対して，人体の部位 26 か所の位置座標を得られる．入力映像中のあるフレームに対して FMP を実行した結果を図 2 に示す．図 2(a) が元の映像フレームであり，図 2(b) が検出結果をグラフ表現で可視化したものである．図 2(b) 中のノードは，それぞれ検出された部位の位置を示す．

事前処理では，入力された歩容映像の全フレームに対して FMP を適用し，各部位の位置座標系列 $P(p,t)$ を得る．ここで，p は各部位の識別子であり，$p \in \{0,\ldots,25\}$ である．t はフレーム番号であり，映像長を T とすると，$t \in \{1,\ldots,T\}$ である．

得られた各部位の位置座標系列 $P(p,t)$ にはノイズが多く含まれるため，時間方向に平滑化することでその影響を軽減する．平滑化された各部位の位置座標系列 $P'(p,t)$ は，平滑化窓幅を S とすると次式で表される．

$$P'(p,t) = \begin{cases} \dfrac{1}{S}\displaystyle\sum_{\tau=t-S+1}^{t} P(p,\tau) & (t > S) \\ \dfrac{1}{t}\displaystyle\sum_{\tau=1}^{t} P(p,\tau) & (\text{otherwise}) \end{cases} \tag{1}$$

得られた位置座標系列 $P'(p,t)$ から，26 部位のうち全ての 2 部位 p_1，p_2 の組み合わせにおける部位の相対距離系列 $D_{p_1,p_2}(t)$ を計算する．相対距離の計算には Euclidean 距離を用い，単位は画素（pixel）とする．

また，各フレームにおける頭の y 座標と足の y 座標の差 $H(t)$ を計算し，映像全体での $H(t)$ の平均 $\bar{H}$ を求める．そして，すべての $D_{p_1,p_2}(t)$ を $\bar{H}$ で割って正規化することにより，正規化された部位の相対距離系列 $L_{p_1,p_2}(t)$ を得る．

$$L_{p_1,p_2}(t) = \frac{D_{p_1,p_2}(t)}{\bar{H}} \tag{2}$$

$$\bar{H} = \frac{1}{T}\sum_{t=1}^{T} H(t) \tag{3}$$

2.1.2 人体の周期運動に基づく特徴の抽出

オノマトペ表現の違いによる歩行周期の違いを捉えるために，2.1.1 で導入した部位の相対距離系列 $L_{p_1,p_2}(t)$ を用いて，人体の周期運動に基づく特徴を抽出する．まず，$L_{p_1,p_2}(t)$ を対称拡張法により周期拡張し，周期 $2T$ の相対距離系列 $\tilde{L}_{p_1,p_2}(t)$ を得る．対称拡張法は，短い周期的系列を拡張するための一般的な手法であり，$\tilde{L}_{p_1,p_2}(t)$ と，$\tilde{L}_{p_1,p_2}(t)$ を時間方向に反転させた $\tilde{L}_{p_1,p_2}(T-t+1)$ を交互に繰り返して結合することによって系列を拡張する．これに対し，時間方向に窓幅 W の Fast Fourier Transform（FFT）を適用することにより，$Q_{p_1,p_2}(\omega)$ が得られる．周期拡張を行なうことによって，FFT の窓幅を大きくすることができるため，FFT の分解能を高めることができる．さらに，$Q_{p_1,p_2}(\omega)$ のパワースペクトル $R_{p_1,p_2}(\omega)$ を次式により求める．

$$R_{p_1,p_2}(\omega) = \frac{Q_{p_1,p_2}(\omega)Q^*_{p_1,p_2}(\omega)}{W} \tag{4}$$

ここで $Q^*_{p_1,p_2}(\omega)$ は $Q_{p_1,p_2}(\omega)$ の複素共役である．そして，全ての p_1,p_2 の組み合わせ 325 通りに対して $R_{p_1,p_2}(\omega)$ を足し合わせた $C(\omega)$ を次式で求める．

$$C(\omega) = \sum_{p_1,p_2} R_{p_1,p_2}(\omega) \tag{5}$$

本手法では，このようにして得られた $C(\omega)$ のうち，低周波成分を表す部分を特徴として抽出する．これは，歩行動作の 1 周期（2 歩）に要する時間が 1 秒前後であるため，$C(\omega)$ の 1 Hz 付近の低周波成分に注目することによって，オノマトペ表現の違いによる歩行周期の違いを捉えられると想定したためである．本研究では，$C(\omega)$ のうち，$\omega \in \{1,\ldots,512\}$ の 512 成分を連結したものを特徴量とする．この次元数は予備実験の結果に基づいて決定した．なお，$C(0)$ は直流成分を表すため使用しない．

FFT の周波数分解能は W と入力映像のフレームレート F の比で決まる．周波数分解能を一定にするためには，W の値を F の値に応じて決定する必要がある．一般に，入力映像のフレームレートが F [fps] の場合，W/F が 1 Hz に対応する．そこで，$1 \le \omega \le 512$ の 512 成分が 1 Hz 付近の成分を表すようにするために，フレームレート F に対して W/F の値が 512 の約半分である $4,096/15$ となるように W の値を定める．例えば，フレームレートが 30 fps であれば窓幅を 8,192 フレーム，60 fps であれば窓幅を 16,384 フレームとする．

2.1.3 人体部位の相対的位置関係に基づく特徴の抽出

2.1.1 で導入した部位の相対距離系列 $L_{p_1,p_2}(t)$ を用いて，部位の相対的位置関係に基づく特徴を抽出する．これにより，大きく動く部位など，局所的な特徴を捉えられると想定する．まず，全ての $L_{p_1,p_2}(t)$ について，時間方向の分散 V_{p_1,p_2} と尖度 K_{p_1,p_2} を次式により計算する．

$$V_{p_1,p_2} = \frac{1}{T}\sum_{t=1}^{T}(L_{p_1,p_2}(t) - \mu_{p_1,p_2})^2 \tag{6}$$

$$K_{p_1,p_2} = \frac{1}{TV^2_{p_1,p_2}}\sum_{t=1}^{T}(L_{p_1,p_2}(t) - \mu_{p_1,p_2})^4 \tag{7}$$

ここで μ_{p_1,p_2} は $L_{p_1,p_2}(t)$ の時間方向の平均であり，次式で計算される．

$$\mu_{p_1,p_2} = \frac{1}{T}\sum_{t=1}^{T} L_{p_1,p_2}(t) \tag{8}$$

そして，得られた全ての分散 V_{p_1,p_2} と尖度 K_{p_1,p_2} を連結したものを特徴量とする．$p_1 < p_2$ を満たす p_1 と p_2 の組み合わせは ${}_{26}\mathrm{C}_2 = 325$ 通りである．よって，部位の相対的位置関係に基づく特徴量の次元数は，分散について 325 次元，尖度について 325 次元の合計 650 次元となる．

表 1　識別対象として選択したオノマトペの辞書上の意味（[1] より引用）

オノマトペ	意味
すたすた	足どり軽く見向きもしないで
のろのろ	にぶい動きでなかなか進まず
よろよろ	今にも倒れそうな足取りで
どっしどっし	体重をかけ力強く踏みつけて

2.1.4　識別器の構築

2.1.2 と 2.1.3 で述べた特徴量を連結して得られた 1162 次元の特徴量を用いて識別器を構築する．

識別器はオノマトペ表現の種類ごとに構築する．各 2 クラス識別器は，入力された歩容映像が特定のオノマトペ表現に対応しているか否かを判定する．本研究では，各識別器は特定のオノマトペ表現に対応した歩容映像を正事例として，通常の歩容映像を負事例として学習する．

負事例としては，他のオノマトペ表現に対応した歩容映像を用いることも考えられるが，学習のための歩容映像をあらゆるオノマトペ表現について網羅的に集めるのは現実的でない．そこで本手法では，歩行者自身がオノマトペに対応しないと考える通常の歩容を撮影した歩容映像を負事例として用いる．

なお，識別器としては Support Vector Machine（SVM）を用いる．

2.2　識別段階

図 1 の右側に示すような手順で，オノマトペ表現に対応した歩容映像を識別する．学習段階と同様に，入力された未知の歩容映像から FMP によって人体の部位を検出し，特徴抽出を行なう．そして，得られた特徴量を N 個全ての識別器に入力し，それぞれで 2 クラス識別を行なう．このようにして，入力歩容映像と N 種類のオノマトペ表現との対応関係を重複を許して求める．

3.　データセット作成

本節では，評価実験で使用するオノマトペ表現に対応する歩容映像データセットの作成方法について述べる．まず，3.1 で識別対象とするオノマトペ表現の選択基準について述べる．次に，3.2 でオノマトペ表現に対応する歩容映像の撮影方法について述べる．

3.1　識別対象とするオノマトペ表現の選択

本研究では，識別対象の「動きを表すオノマトペ」として，「すたすた」，「のろのろ」，「よろよろ」，「どっしどっし」の 4 種類を選択した．これらは，歩行に関するオノマトペとしてオノマトペ辞典 [1] に記載されており，それぞれ表 1 に示すような意味を持つ．これらは，各オノマトペから想起される動きに違いが表れると想定して選択した．また，各オノマトペの音韻が類似しないようにした．

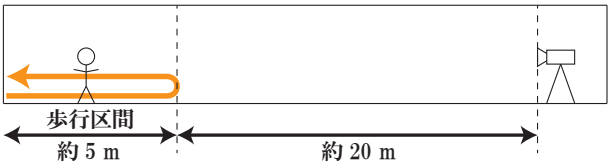


図 3　歩容映像の撮影状況の模式図

3.2　オノマトペ表現に対応した歩容映像の撮影方法

歩容映像として，歩行者を正面および背面から撮影した．奥行き方向の移動による歩行者の大きさの変化を最小限に抑えるために，歩行者から十分に離れた位置にカメラを設置した．撮影には Point Gray Research 社のカメラ Flea3 を用いた．カメラレンズの焦点距離は 35 mm，センサの大きさは 2/3 inch であり，35 mm 判換算焦点距離は約 89 mm であった．歩容映像の撮影状況の模式図を図 3 に示す．歩行区間は約 5 m，歩行区間とカメラとの距離は約 20 m とした．

被験者に対して，「通常の」歩行，「すたすた」した歩行，「のろのろ」した歩行，「よろよろ」した歩行，「どっしどっし」した歩行の 5 種類を表現するように指示した．その際に，表 1 に示した各オノマトペに関する辞書上の意味も参考として提示した．被験者は日本語を母語とする 20 代の男性 5 名であった．

図 3 に示すように，被験者はまずカメラに近づく向きに歩き，歩行区間の端に達したところで一旦静止し，180 度向きを変えてカメラから離れる向きに歩いた．このような試行を 5 種類の歩行動作に対して上記の順番で行ない，それを 2 回繰り返すことで，被験者 1 人あたり，正面と背面合わせて 20 本の歩容映像を撮影し，5 人から合計 100 本の歩容映像を得た．以下，このようにして得られた歩容映像のうち，「通常の」歩行を撮影した映像を，「通常の歩容映像」，「○○」した歩行を撮影した映像を「オノマトペ表現○○に対応した歩容映像」と呼ぶ．映像はすべて 527 × 708 画素，60 fps で撮影した．人やオノマトペ表現の種類によって歩く速さが異なるため，映像によって映像の長さは異なる．最も短い映像は 100 フレーム，最も長い映像は 500 フレームであった．

4.　評価実験

3 節で作成したデータセットを用いて，提案手法の有効性を検証するための評価実験を行なった．本節ではまず，4.1 で実験標本の作成方法について述べる．次に，4.2 で実験における評価方法について述べる．そして，4.3 で実験結果について報告し，それに基づいて 4.4 で考察する．

4.1　学習・識別に用いる標本の作成方法

ここでは，識別器の学習や識別に用いる入力映像を標本とする．撮影した歩容映像は，各オノマトペ表現に対し，1

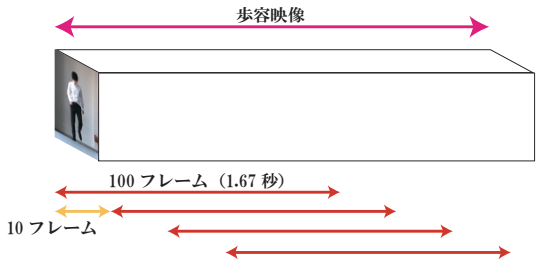


図 4　標本の切り出し方法

人あたり 4 本，5 人合わせて 20 本である．各映像をそのまま 1 つの標本とすると，学習標本数が不足するため，各映像中の区間をずらしながら，複数の映像を切り出して標本とした．標本の切り出し方法を図 4 に示す．具体的には，1 つの標本を 100 フレームの映像とした（$T = 100$）．そして，開始フレームを 10 フレームずつずらしながら 100 フレーム分の映像を順次切り出した．この処理により，「通常」の標本を 227 本，「すたすた」の標本を 95 本，「のろのろ」の標本を 447 本，「よろよろ」の標本を 473 本，「どっしどっし」の標本を 270 本得た．

4.2　評価方法

2 節で述べたように，本手法ではオノマトペ表現の種類ごとに識別器を構築する．そのため，評価も識別器ごとに行なう．

オノマトペ表現 X の識別器は，学習時には通常の歩容映像を負の学習標本として使用し，X 以外のオノマトペ表現に対応する歩容映像は使用しなかったが，識別時には，オノマトペ表現 X に対応する歩容映像を正の試験標本として，X 以外のオノマトペ表現に対応する歩容映像および通常の歩容映像を負の試験標本として使用した．学習時に X 以外のオノマトペ表現に対応した歩容映像を使用しなかったのは，2.1.4 でも述べたように，実用上は学習段階であらゆるオノマトペ表現の歩容映像を網羅的に集めることが現実的ではないためである．また，負の試験標本は正の試験標本よりも数が多くなるため，負の試験標本の数と，正の試験標本の数がほぼ同数になるように均等に間引いた．このとき負の試験標本は，オノマトペ表現の種類ごとに標本数がほぼ同数になるようにした．

各識別器の構築および評価は，Leave-one-person-out 交差検定で行なった．データセットは歩行者 5 人の映像からなるため，5 分割となる．SVM のカーネルには線形カーネルを採用し，特徴量を [0,1] に正規化した．評価尺度は各識別器の再現率（Recall），適合率（Precision），F 値（F-measure）と分類率（Classification rate）とした．再現率は，正の試験標本のうち，クラスを正しく判定できたものの割合を示す指標である．適合率は，識別器が正と判定した標本のうち，実際に正の試験標本だったものの割合を示す指標である．F 値は，再現率と適合率の調和平均である．これらは，主に正の試験標本を正しく識別できたかを見る指標であり，検出などの応用を考慮する上で重要である．一方，分類率は全ての試験標本のうち，クラスを正しく判定できたものの割合を表す指標である．これは正の試験標本だけでなく，負の試験標本も含めて正しく識別できたかを見る指標であり，純粋に識別器の性能を測る上で重要である．

比較評価のために，人体の周期運動に基づく特徴のみを用いた手法（比較手法 1），人体部位の相対的位置関係に基づく特徴のみを用いた手法（比較手法 2），HOF（Histogram of Optical Flow）を特徴として用いた手法（比較手法 3）の 3 つを比較手法とした．用いた HOF の次元数は 65536 次元であり，これを PCA により 256 次元に圧縮したものを特徴量とした．なお，提案手法，比較手法 1 および比較手法 2 では FMP の処理速度の都合により，映像の解像度を 128×172 画素に縮小した．FMP が映像中から人を検出できないフレームは存在しなかった．また，部位座標の平滑化窓幅 S は 8 フレームとした．データセットのフレームレートは 60 fps である（$F = 60$）ため，FFT の窓幅 W は 16,384 フレームに設定した．

4.3　実験結果

評価実験により得られた識別器ごとの再現率，適合率，F 値および分類率を表 2〜表 6 に示す．表 6 に示すように，全ての識別器の平均では提案手法がどの評価指標においても，すべての比較手法を上回った．これにより，人体部位の相対的位置関係を基にした特徴と，人体の周期運動に基づく特徴を組み合わせて使用することの有効性が確認できた．また，人体部位の相対的位置関係を基にした特徴がオノマトペ表現に対応した歩容映像の識別に有用であることが示され，杉山ら [9] の知見を支持する結果となった．

ただし，識別器ごとに注目すると，表 4 の「よろよろ」識別器は全ての指標で提案手法が比較手法 3 を下回った．また，表 5 の「どっしどっし」識別器は適合率で提案手法が比較手法 3 を下回った．しかし，F 値や分類率は改善していることから，「どっしどっし」識別器でも提案手法は有効である．また，表 2 の「すたすた」識別器は適合率で比較手法 1 を下回り，表 3 の「のろのろ」識別器は適合率と分類率で比較手法 2 を下回った．しかし，比較手法 1 と比較手法 2 は提案手法のサブセットであり，表 6 に示す全識別器の平均においては提案手法がこれらを上回っていることから，2 種類の特徴を組み合わせることにより安定した識別性能を得られるようになったと考えられる．

4.4　考察

4.4.1　「よろよろ」識別器に関して

評価実験の結果，「よろよろ」識別器を除く識別器においては，ほぼ全ての指標で，提案手法が比較手法を上回った．

表 2 「すたすた」識別器の再現率，適合率，F 値，分類率

	再現率	適合率	F 値	分類率
提案手法	**0.811**	0.740	**0.774**	**0.779**
比較手法 1	0.716	**0.782**	0.747	0.775
比較手法 2	0.747	0.683	0.714	0.721
比較手法 3	0.678	0.557	0.611	0.605

表 3 「のろのろ」識別器の再現率，適合率，F 値，分類率

	再現率	適合率	F 値	分類率
提案手法	**0.647**	0.610	**0.628**	0.648
比較手法 1	**0.647**	0.431	0.517	0.446
比較手法 2	0.631	**0.623**	0.627	**0.655**
比較手法 3	0.624	0.573	0.597	0.611

表 4 「よろよろ」識別器の再現率，適合率，F 値，分類率

	再現率	適合率	F 値	分類率
提案手法	0.799	0.673	0.730	0.726
比較手法 1	0.835	0.490	0.618	0.520
比較手法 2	0.732	0.659	0.693	0.699
比較手法 3	**0.970**	**0.751**	**0.847**	**0.836**

表 5 「どっしどっし」識別器の再現率，適合率，F 値，分類率

	再現率	適合率	F 値	分類率
提案手法	**0.907**	0.778	**0.838**	**0.838**
比較手法 1	0.637	0.612	0.624	0.647
比較手法 2	0.796	0.749	0.772	0.783
比較手法 3	0.574	**0.838**	0.681	0.749

表 6 平均再現率，平均適合率，平均 F 値，平均分類率

	平均再現率	平均適合率	平均 F 値	平均分類率
提案手法	**0.791**	**0.700**	**0.743**	**0.748**
比較手法 1	0.709	0.579	0.637	0.597
比較手法 2	0.727	0.678	0.702	0.715
比較手法 3	0.712	0.680	0.695	0.700

「よろよろ」識別器で比較手法 3 が提案手法を上回ったのは，「よろよろ」した歩容は体が左右に大きく揺れる，すなわち絶対座標の変化が特徴的であり，相対座標を用いている提案手法よりも，絶対座標を用いている比較手法 3 の方がこの特徴をよく捉えられたためだと考えられる．しかし，比較手法 3 は人体の位置変化に対して頑健ではない．一方，提案手法は相対座標を特徴としているため，人体の位置変化に頑健であると考えられる．よって，より実環境に近い実験条件では，提案手法が有効となると考えられる．

4.4.2 歩容の個人差：主観的定義の正当性

本研究では，歩行者全員がオノマトペに対する共通の印象に基づいて歩容を表現するという仮定のもと，歩行者によって主観的に表現された歩容を「オノマトペ表現に対応した歩容」と定義し，データセットの作成と評価実験を行なった．

実験の結果，平均的に高い識別性能が得られたことから，この仮定はある程度正しかったと考えられる．しかし，識別結果を詳細に調査すると，交差検定によって特定の被験者を試験する場合において極端に悪い識別結果が得られることがあった．表 7 に「のろのろ」識別器の，表 8 に「よろよろ」識別器の交差検定結果の内訳を示す． ここで，表中の誤検出率とは，負の試験標本を正と誤って判定した割合を表す．このうち，表 7 の「のろのろ」識別器の被験者 A に注目すると，誤検出率（「のろのろ」に対応する歩容映像以外を正と判定した割合）が著しく高いことが分かる．この場合，識別器は被験者 B，C，D，E の「のろのろ」に対応した歩容映像を正事例として，被験者 B，C，D，E のオノマトペ表現に対応しない歩容映像を負事例として学習している．よって，被験者 A の歩容映像が全体的に正と判定されたということは，被験者 A の歩容が，客観的に見ると基本的に「のろのろ」している可能性が考えられる．同じことが，表 7 の「よろよろ」識別器の被験者 A にもいえる．

表 7 「のろのろ」識別器の交差検定結果の内訳

	再現率	誤検出率			
	のろのろ	対応なし	すたすた	よろよろ	どっしどっし
A	0.99 (113/114)	0.71 (27/38)	0.33 (8/24)	1.00 (35/35)	1.00 (36/36)
B	0.36 (35/97)	0.15 (4/27)	0.28 (5/18)	0.52 (13/25)	0.06 (2/33)
C	0.33 (24/73)	0.15 (3/20)	0.00 (0/20)	0.47 (9/19)	0.00 (0/22)
D	0.84 (61/73)	0.15 (5/33)	0.00 (0/17)	0.80 (16/20)	0.14 (3/21)
E	0.62 (56/90)	0.00 (0/44)	0.00 (0/16)	0.58 (19/33)	0.00 (0/27)

表 8 「よろよろ」識別器の交差検定結果の内訳

	再現率	誤検出率			
	よろよろ	対応なし	すたすた	のろのろ	どっしどっし
A	1.00 (138/138)	0.53 (20/38)	0.54 (13/24)	0.92 (35/38)	1.00 (36/36)
B	0.74 (73/98)	0.26 (7/27)	0.22 (4/18)	0.45 (15/33)	0.88 (29/33)
C	0.77 (72/94)	0.13 (4/30)	0.00 (0/20)	0.12 (3/25)	0.05 (2/44)
D	0.41 (32/78)	0.00 (0/33)	0.00 (0/17)	0.16 (4/25)	0.00 (0/21)
E	0.97 (63/65)	0.00 (0/22)	0.00 (0/16)	0.28 (5/18)	0.26 (7/27)

逆に，表 7 の「のろのろ」識別器の被験者 B や被験者 C に注目すると，再現率が著しく低いことが分かる．これらは，被験者 B や被験者 C が考える「のろのろ」が，他の 3 人が考える「のろのろ」と異なる可能性を示唆している．

このように，歩容の表現に関して，基本的に個人差は小さいが，平均的な表現から逸脱した表現をする人が一部存在する，という仮説が立てられる．そのような人は統計的

には外れ値であり，外れ値を学習標本として用いると，識別器の性能は低下してしまうと考えられる．その対策として，平均的な表現から逸脱した人の歩容映像は学習標本から除去することが考えられる．そのためには，学習標本に基づいて，オノマトペ表現ごとに平均的な歩容をモデル化する必要がある．また，本報告中の評価実験（主観的定義）の実験結果と，被験者実験により客観的に歩容映像とオノマトペ表現を対応付けた場合（客観的定義）の実験結果とを比較することにより，歩容の個人差が識別性能に与える影響を定量的に分析する必要がある．

5. むすび

本報告では，動きを表すオノマトペ表現に対応した映像として歩容映像に注目し，オノマトペ表現に対応した歩容映像を識別する手法を検討した．提案手法では，映像から人体の部位座標を検出した結果を基に，人体の周期運動に基づく特徴と人体部位の相対的位置関係に基づく特徴を抽出し，それらの特徴量を学習した 2 クラス識別器をオノマトペ表現の数だけ構築した．2 クラス識別器はそれぞれ，入力された歩容映像が特定のオノマトペ表現に対応しているか否かを判定した．

実験では，「すたすた」，「のろのろ」，「よろよろ」，「どっしどっし」の 4 種類のオノマトペ表現を識別対象とした．そして，HOF を用いた手法と比較することで，提案手法の有効性を評価した．評価指標として，再現率，適合率，F 値および分類率を用いた．全ての識別器の平均では，どの指標においても提案手法が比較手法を上回り，提案手法の有効性を確認した．

今後の課題として，客観的定義に基づくデータセットの作成，音韻と映像特徴量の関係性の分析などが挙げられる．1 節で述べた通り，本研究では歩行者全員がオノマトペに対して共通の印象に基づいて歩容を表現できると仮定し，歩行者によって主観的に表現された歩容を「オノマトペ表現に対応した歩容」と定義した．すなわち，オノマトペに対する主観的な印象が，客観的な印象と等しいと仮定した．しかし，4.4.2 で述べたように，同じオノマトペ表現に対して，平均的な歩容とは異なる歩容を表現する人もいる可能性が実験により示唆された．そのため，被験者実験により第 3 者が歩容映像を見ることで感じるオノマトペ表現を付与することで，その歩容が対応する客観的なオノマトペ表現を調査する必要がある．そして，それが主観的なオノマトペ表現と一致しない場合には，その客観的なオノマトペ表現を真値として，評価実験を行なう必要がある．

また，提案手法では，オノマトペ表現に対応した歩容映像から映像特徴を抽出することで識別を可能にした．これは，動きを表すオノマトペと，映像特徴量との間に関係性があることを示している．オノマトペには音響的印象が事象の様態と対応する性質があることを考慮すると，オノマトペを構成する音韻が，映像特徴量と何かしらの関係性を持つ可能性がある．音韻と映像特徴量との関係性が明らかになれば，学習に用いなかったオノマトペに対応した歩容映像や，新たに登場したオノマトペに対応した映像の識別が可能になると考えられる．

謝辞　データセットの撮影にご協力頂いた諸氏に感謝する．また，本研究の一部は科研費による．

参考文献

[1] 小野正弘：擬音語・擬態語日本語 4500 オノマトペ辞典，小学館 (1999).
[2] 田守育啓，ローレンススコウラップ：オノマトペ—形態と意味—，くろしお出版 (2007).
[3] 神原啓介，塚田浩二：オノマトペン，インタラクティブシステムとソフトウェアに関するワークショップ（WISS）2008 予稿集，pp. 79–84 (2008).
[4] 小松孝徳，秋山広美：ユーザの直感的表現を支援するオノマトペ表現システム，電子情報通信学会論文誌（A），Vol. J92-A, No. 11, pp. 752–763 (2009).
[5] 石原一志，坪田　康，奥乃　博：日本語の音節構造に着目した環境音の擬音語への変換，電子情報通信学会技術研究報告，SP2003-38 (2003).
[6] 比屋根一雄，澤部直太，飯尾　淳：単発音のスペクトル構造とその擬音語表現に関する検討，電子情報通信学会技術研究報告，SP97-125 (1998).
[7] Shimoda, W. and Yanai, K.: A visual analysis on recognizability and discriminability of onomatopoeia words with DCNN features, *Proc. 2015 IEEE Int. Conf. on Multimedia and Expo*, pp. 1–6 (2015).
[8] 鍵谷龍樹，白川由貴，土斐崎龍一，渡邊淳司，丸谷和史，河邉隆寛，坂本真樹：動画と静止画から受ける粘性印象に関する音象徴性の検討，人工知能学会論文誌，Vol. 30, No. 1, pp. 237–245 (2015).
[9] 杉山雄紀，近藤敏之：ロボットの歩行動作設計によるオノマトペ・情報表現の共通理解，第 25 回人工知能学会全国大会，1C1-OS4a-4 (2011).
[10] 奥村麻由，槇原　靖，八木康史：大規模歩容データベースを用いたガウス過程回帰による年齢推定の評価，情報処理学会研究報告，Vol. 2011-CVIM-195, No. 33, pp. 1–8 (2011).
[11] 万波秀年，槇原　靖，八木康史：歩容における性別・年齢の分類と特徴解析，電子情報通信学会論文誌（D），Vol. 92, No. 8, pp. 1373–1382 (2009).
[12] Yang, Y. and Ramanan, D.: Articulated human detection with flexible mixtures-of-parts, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 12, pp. 2878–2890 (2013).