萌芽論文

Exploratory Research Paper

音象徴性を利用したオノマトペによる歩容の記述

Describing Gaits by Onomatopoeias with Sound Symbolism

加藤 大貴 Hirotaka Kato
名古屋大学大学院情報学研究科 Graduate School of Informatics, Nagoya University
katoh@murase.is.i.nagoya-u.ac.jp

平山 高嗣 Takatsugu Hirayama
名古屋大学未来社会創造機構 Institutes of Innovation for Future Society, Nagoya University
takatsugu.hirayama@nagoya-u.jp

道満 恵介 Keisuke Doman
中京大学工学部 School of Engineering, Chukyo University
kdoman@sist.chukyo-u.ac.jp

井手 一郎 Ichiro Ide
名古屋大学大学院情報学研究科 Graduate School of Informatics, Nagoya University
ide@i.nagoya-u.ac.jp

川西 康友 Yasutomo Kawanishi
（同　上）
kawanishi@i.nagoya-u.ac.jp

出口 大輔 Daisuke Deguchi
名古屋大学情報戦略室 Information Strategy Office, Nagoya University
ddeguchi@nagoya-u.jp

村瀬 洋 Hiroshi Murase
名古屋大学大学院情報学研究科 Graduate School of Informatics, Nagoya University
murase@i.nagoya-u.ac.jp

keywords: onomatopoeia, gait, phoneme, body-parts, sound-symbolism

Summary

There are few studies on the motion attribute of gaits because of the absence of appropriate semantic labels describing motion. We focus on onomatopoeias to describe the motion of gaits. The Japanese language is known to have a greater number of onomatopoeias in its vocabulary. Especially, the human gait is one of the most commonly represented phenomenon by onomatopoeias expressing its visually dynamic state. It is said that Japanese onomatopoeias have sound-symbolism and their phonemes are strongly related to the impression of various phenomena. Because of this, Japanese people can distinguish gaits based on their appearances and express their impressions on them briefly and intuitively using various onomatopoeias. In addition, a previous study revealed relative body-parts movement is associated with onomatopoeias when we describe gaits. Inspired by these studies, we considered that if a phonetic space based on sound-symbolism can be associated with the kinetic feature space of gaits, subtle difference of gaits could be expressed as difference in phoneme. In this paper, we propose a framework to convert the relative body-parts movements to any onomatopoeia using a regression model. This framework is expected to make human-computer interaction more intuitive. Through experiments, we confirmed the effectiveness of the proposed framework, and discussed the potential of describing an arbitrary gait by not only existing onomatopoeias but also a novel one.

1. ま え が き

近年，機械学習による歩行者の属性認識に関する研究が盛んに行なわれている．歩行者画像からの年齢，性別，服装や持ち物など，歩行者の見えに関する様々な属性認識タスクに目覚ましい進歩がみられる [*1] 一方で，歩容の属性に関する研究は少ない．歩容とは，人間が歩行する様子，すなわち歩行者の動きを表現する語である．これは従来，歩行者の動きを記述するための適切な方法が提案されていないことが原因の１つに挙げられる．しかし，車載カメラ映像から特徴的な動きをしている歩行者を検出し，それを運転者に伝達したり，監視カメラ映像から特定の動きをしている人物を検索するなどの応用を考えたとき，人間が直感的に理解しやすい形で歩容の属性を記述し，計算機上で取り扱えるようにすることは重要である．

そこで，本論文では歩容を記述しうる表現方法としてオノマトペに着目する．オノマトペとは，擬音語及び擬態語と呼ばれている言語表現の総称である [小野 07]．日本語においては,「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手段として，多く

*1 http://rap.idealtest.org/

のオノマトペが使用される．一方，英語においては，オノマトペは擬音語のみであり，擬態語を含まないと定義されることが多い．また，擬音語には *bow-wow*，*tic-toc* 等，日本語の擬音語と類似した形式のものも多く存在するが，擬態語は *trot*（急いで歩く），*stroll*（ぶらぶら歩く）など様態動詞（様子が語感として動詞に含まれたもの）がほとんどであり，日本語の擬態語に比べると間接的な表現である [呂 04, 小倉 16]．

オノマトペは音象徴性という性質を持ち，その音響的印象が事象の様態と対応するため，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 [Hamano 98, 田守 99]．そのため，オノマトペは論理的な表現が容易ではない印象を端的に他者に対して伝えるために有効な手段であると考えられている．例えば藤野らは，人間が運動感覚を学習する場合などにオノマトペの利用が効果的であると指摘している [藤野 05]．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，神原らが開発した，「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といったような描画システム [神原 08] や，Fukusato らが提案するアニメーション映像への擬音語の重畳表示システム [Fukusato 14] のように，ヒューマンコンピュータインタラクションへの応用性の高さも示されている．

ところが，音響信号 [比屋根 98, 石原 03, Sundaram 07, Sundaram 08] や質感画像 [權 17, Shimoda 15] に関しては工学的な研究例も存在する一方で，映像とオノマトペとの関係はほとんど検討されてこなかった．歩容をオノマトペという直感的な表現を用いて記述できれば，例えば車載カメラ映像から「ふらふら」，「よろよろ」している歩行者を検出し，さらに自動車運転者の注意誘導のための直感的な音声提示に利用したり，監視カメラ映像中から「どっしどっし」歩いている歩行者を検索するなどの応用が期待できる．

鍵谷らは CG 映像作成ソフトウェアを用いて作成した粘性をもつ液体の映像を被験者に提示し，映像と，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 [鍵谷 15]．これをふまえ本論文では，歩容と，オノマトペを構成する音韻との間にも同様の関連性が存在すると仮定し，それを利用して歩容をオノマトペで記述することを考える．そのために，音韻の印象を表現する「音韻空間」に歩容の特徴を射影し，その音韻空間上で歩容を取り扱う枠組みを提案する．

オノマトペはその音象徴性ゆえに，人は辞書にないような新しいオノマトペを即興で作って直感的に様子を表現することもできる．そのような場合には，事前に特定のオノマトペをラベルとして教師あり学習を行なう単純な手法では対応できない．本手法ではオノマトペそのものではなく音韻との関係性を獲得するため，最終的に（1）既定のオノマトペの中から最も近いオノマトペを判別するだけでなく，（2）その歩容をよく表現する新たなオノマトペを生成することまでも，同じ枠組みで実現できる．

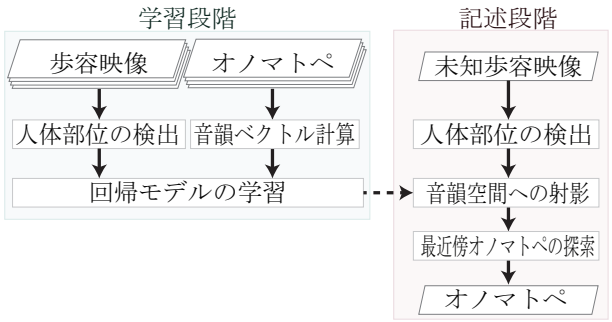

図 1 歩容映像の音韻空間への射影手法の処理手順

本論文では，まず 2 章で音韻空間を用いた歩容の記述手法について述べる．次に 3 章で実験に用いるデータセットの作成方法について述べる．更に 4 章で，提案する枠組みの妥当性を検証するための実験及び結果について述べる．特に，上記の（1）について 4·3 節，（2）について 4·4 節で詳しく述べる．最後に 5 章で本論文をまとめ，今後の課題について検討する．

2. 音韻空間を用いた歩容の記述手法

提案手法の処理手順を図 1 に示す．学習には，オノマトペがアノテーションされた歩容映像を用いる．まず，事前処理として，2·1 節で述べるようにして歩容映像から人体の部位検出を行なう．一方，2·2 節で述べるようにしてオノマトペを定量化し，音韻ベクトル（音韻空間上の点）に変換する．そして，2·3 節で述べるようにしてそれらの対応関係を学習し，回帰モデルを構築する．記述の際は，オノマトペが未知な映像から同様に人体部位の検出を行ない，学習した回帰モデルを用いて音韻空間へと射影する．最後に，2·4 節で述べるようにして，推定された音韻ベクトルをオノマトペへと変換する．以下，各処理の手順について詳述する．

2·1 人体部位の検出

Li らは異常歩容の検出のために人体部位の動きを利用する手法を提案しており，特徴的な歩容を捉える特徴量として，人体部位の動きが有用であることを示唆している [Li 18]．一方，杉山ら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い，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，体の部位の相対的な運動に着目することが重要であると示唆している [杉山 11]．これらをふまえ，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こととした．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．

まず，入力された歩容映像の全フレームに対して検出処理を行ない，各部位の位置座標系列 $C(p,t)$ を得る．ここで，$p \in \{0,\ldots,P-1\}$ は部位数を P とした時の各部位の識別子である．また，$t \in \{1,\ldots,T\}$ はフレーム番号である．ここで T は映像長である．

得られた位置座標系列 $C(p,t)$ から，任意の 2 部位 $p_1,p_2 \in \{0,\ldots,P-1\}$ のすべての組み合わせにおける部位間相対距離系列 $D_{p_1,p_2}(t)$ 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 $\bar{H}$ で除することにより，正規化された部位間相対距離系列 $L_{p_1,p_2}(t)$ を得る．

$$L_{p_1,p_2}(t) = \frac{D_{p_1,p_2}(t)}{\bar{H}} \tag{1}$$

$$\bar{H} = \frac{1}{T}\sum_{t=1}^{T} H(t) \tag{2}$$

ここで，p_1 と p_2 の組み合わせは ${}_P\mathrm{C}_2$ 通りである．

2·2 音韻空間の構築

本手法では，ABAB 型（「すたすた」,「のろのろ」等，同じ 2 音が 2 回繰り返される型）を記述の対象とし，オノマトペを構成する 4 つの音素（「すたすた」であれば/s/，/u/，/t/，/a/の 4 つ）それぞれを N 次元のベクトルで定量化し，得られた 4 つのベクトルを結合した $4N$ 次元のベクトルによって張られる空間を音韻空間と定義する．なお，本論文では音素の列を音韻と定義し，この N 次元のベクトルを音素ベクトル，結合した $4N$ 次元のベクトルを音韻ベクトルと呼ぶ．学習時には，映像に対応付いたオノマトペを音素ごとに定量化・結合し，音韻空間上の 1 点として表現する．

ここで，本論文で取り扱う ABAB 型オノマトペは，濁音，半濁音、拗音が付加されたもの，および「どっしどっし」のように第 2 モーラ目の前に促音が付加されたものを含み，撥音，長音は含まないこととする．ただし，伊藤ら [伊藤 15] が秋山らの音素定量化手法 [秋山 11] を用いて「ガッシャーンガッシャーン」のような複雑な形式のオノマトペを ABAB 型として扱っているように，より広義の ABAB 型を扱える音素定量化手法を利用することで，本手法が対応する ABAB 型オノマトペの範囲は容易に拡張可能である．なお，本論文の評価実験では音素の定量化には既存の手法を用いる．これに関しては 4·2 節で詳述する．

2·3 回帰モデルの構築

2·1 節で得た正規化された部位間相対距離系列 $L_{p_1,p_2}(t)$ を説明変数，2·2 節で求めた対応するオノマトペから生成された音韻ベクトルを目的変数として回帰し，相対距離系列を音韻空間上に射影する回帰モデル f を構築する．未知の歩容映像から計算された相対距離系列をこのモデルに入力することで，音韻ベクトル $\hat{v}$ を推定できる．

$$\hat{v} = f(L) \tag{3}$$

ここで，入力 L は 1 つの歩容映像から得られた相対距離系列をまとめたもので，$T \times {}_P\mathrm{C}_2$ の行列である．本論文の評価実験では回帰モデルとしていくつかの深層学習モデルを用いる．詳細な入力形式はモデルによって異なるため，詳細は 4·2 節で述べる．

2·4 音韻ベクトルからオノマトペへの変換

本手法では，用途に応じて 2 種類の変換手法を提案する．1 つは，変換先の候補となるオノマトペの種類があらかじめ定まっている場合の手法であり，これを全体最近傍法と呼ぶ．推定された音韻ベクトルを $\hat{v}$，変換先の候補となる ABAB 型オノマトペの集合を O として，

$$\arg\min_{o \in O} ||\hat{v} - Q(o)|| \tag{4}$$

を満たすオノマトペ $o \in O$ を出力とする．ここで，Q はオノマトペを音韻ベクトルに変換（定量化）する関数であり，$||\cdot||$ は任意の距離尺度である．すなわち，候補のオノマトペのうち音韻空間上で最も近いものを選択する手法である．これは，辞書に載っているオノマトペのみを出力候補にしたい場合や，例えば「すたすた」と「のろのろ」のうち近い方を判定したい場合などに利用できる．

もう 1 つは，変換先のオノマトペとして，任意の音素の組み合わせを許す場合の手法であり，これを音素最近傍法と呼ぶ．まず，推定された $4N$ 次元の音韻ベクトルを N 次元の音素ベクトル 4 つに分解する．分解された音素ベクトルを $\hat{v}_k (k = 1,2,3,4)$，変換先の候補となる音素の集合を O_k として，

$$\arg\min_{o_k \in O_k} ||\hat{v}_k - q(o_k)|| \tag{5}$$

を満たす音素 $o_k \in O_k$ を求める．そして，推定された 4 つの音素 o_1,o_2,o_3,o_4 を順に結合して得られたオノマトペを出力とする．ここで，q は音素を音素ベクトルに変換（定量化）する関数である．これにより，未知の歩容映像に対し，その動きをよく表現する新たなオノマトペを生成できる．

3. データセットの作成

本章では，評価実験で用いるオノマトペがアノテーションされた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 節で歩容映像の撮影方法について述べる．次に，3·2 節で第三者の評価に基づいて歩容映像にオノマトペをアノテーションする手順について述べる．

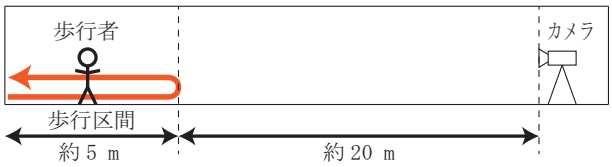

図 2 歩容映像の撮影状況

3·1 撮 影 方 法

歩容映像として，歩行者の前面及び背面を撮影した．側面からの撮影を行なわなかったのは，十分な長さかつ解像度の映像を撮影するためには歩行者に合わせてカメラを移動させる，複数台のカメラを設置する等の大規模な撮影環境の構築が必要となるためである．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 138 mm であった．歩容映像の撮影状況を図 2 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．

図 2 に示すように，撮影実験協力者は 1 回の試行で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．各試行において，通常の歩行,「すたすた」,「のろのろ」,「よろよろ」,「どっしどっし」,「せかせか」,「てくてく」,「とぼとぼ」,「のしのし」,「よたよた」,「ぶらぶら」の 11 種類のうち，実験者が指定した 1 種類を表現するよう指示した．協力者及び試行によって異なるオノマトペを指示し，各協力者が 6〜16 回の試行を行なうようにした．これらのオノマトペは，歩行に関するオノマトペとしてオノマトペ辞典 [小野 07] に掲載されているもののうち，2·2 節で述べた定義による ABAB 型であるものの中から，構成する音韻の多様性を考慮しながら選択した．歩行者は日本語を母語とする 20 代の男性 7 名であった．

映像はすべて 527 × 708 画素，60 fps で撮影した．撮影実験は複数回に分けて実施した．当初は，通常の歩行,「すたすた」,「のろのろ」,「よろよろ」,「どっしどっし」を対象として撮影を行ない，その後，オノマトペの種類を増やすために,「せかせか」,「てくてく」,「とぼとぼ」,「のしのし」,「よたよた」,「ぶらぶら」を追加した 11 種類を対象とした．その結果，通常の歩行,「すたすた」,「のろのろ」,「よろよろ」,「どっしどっし」を表現した歩容の映像を各 22 本,「せかせか」,「てくてく」,「とぼとぼ」,「のしのし」,「よたよた」,「ぶらぶら」を表現した歩容の映像を各 8 本ずつ，合計 158 本の映像を得た．

表 1 第三者評価によるアノテーション結果

教示	アノテーションされた映像数										合計
	すた	のろ	よろ	どっし	せか	てく	とぼ	のし	よた	ぶら	
通常	4					5					9
すた	7				2	1					10
のろ		5					2			1	8
よろ			7							2	9
どっし				7		1		1			9
せか					3						3
てく	1					1					2
とぼ							3				3
のし		1					2		1		4
よた			2				2				4
ぶら			1							2	3
合計	12	6	10	7	5	8	9	1	1	5	64

3·2 第三者評価に基づくアノテーション

データセットの撮影時に，歩行者には特定のオノマトペを表現するように指示したが，撮影実験協力者がイメージ通りに体を動かせるとは限らないため，得られた歩容は客観的に見てそのオノマトペを表現できているとは限らない．そこで，第三者による評価に基づいて，改めて歩容映像に対するオノマトペのアノテーションを行なった．

実験には 3·1 節で得られた映像のうち，歩行者の前面を撮影した 79 本を用いた．日本語を母語とする 20 代の男女 14 名に対して映像を提示し，その映像に対応すると思うオノマトペを，前述の 10 種の中から，複数回答を許して選択させた．また，アノテーション実験の参加者には撮影実験協力者が含まれており，自身の歩容に対して評価を行なう場合も存在するが，アノテーション実験を撮影実験から 1ヶ月以上期間をあけて実施することで影響を低減した．実験参加者の負担軽減のため，映像 79 本を 2 つのセットに分け，映像 1 本あたり 7 名から回答を得た．そして，その過半数である 4 名以上が対応付いていると回答したオノマトペを映像にアノテーションした．過半数票の獲得をアノテーションの条件とすることにより，アノテーションの客観性を確保している．ここで，複数のオノマトペが過半数票を獲得した場合は最多得票のオノマトペをアノテーションし，最多得票オノマトペが複数存在する映像及びいずれのオノマトペも過半数票を獲得しなかった映像はデータセットから除外することとした．各オノマトペに対応付いた映像の数を表 1 に示す．表 1 の各行が，歩行者に対して教示したオノマトペ，各列が第三者によってアノテーションされたオノマトペである．ここで，歩行者の背面を撮影した映像については，対になる前面を撮影した映像と同じオノマトペをアノテーションするものとした．

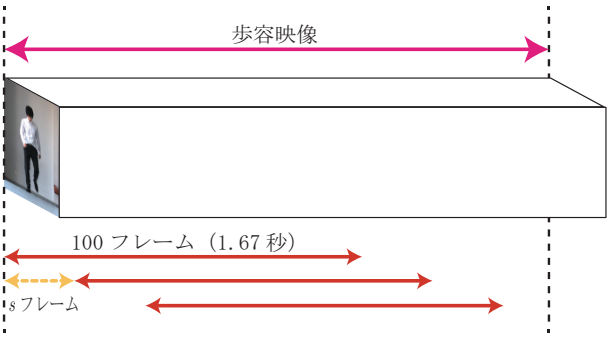

図 3　標本の切り出し方法

図 4　CPM [Wei 16] で検出される人体の 14 部位

4. 評　価　実　験

本章では，2 章で提案した歩容の映像とオノマトペを対応付ける枠組みの妥当性を検証するための評価実験について述べる．まず 4·1 節で，実験標本の作成方法について述べる．次に，4·2 節で，実験に際する手法の実装方法について述べる．次に 4·3 節で，音韻空間上の多クラス分類実験，すなわち既定のオノマトペの中で最も近いものを判別するタスクの評価について述べる．更に 4·4 節で，歩容に対する任意のオノマトペの生成実験，すなわちその歩容を最もよく表現する新奇的なオノマトペを生成するタスクの評価について述べる．最後に 4·5 節で，実験結果について考察する．

4·1　実験に用いる標本の作成

3 章で作成したデータセットは映像長が一定ではないため，これを実験で扱いやすくするために，元の映像から固定長の部分映像を標本として切り出した．具体的には，図 3 に示すように，開始フレームを s フレームずつずらしながら 100 フレーム分（約 1.67 秒）の映像を順次切り出した．すなわち回帰モデルに入力する映像長 $T = 100$ とした．この際，各オノマトペの標本数を大まかに揃えるため，標本数が最も少なくなる（延べ映像長が最も短い）オノマトペの標本を $s = 5$ として切り出し，それ以外のオノマトペは適宜 s の値を増加させることで標本数を調整した．映像長を 100 フレームとしたのは，歩容の 1 周期（2 歩）が十分収まる長さであるためである．

表 2　実験に用いた CNN の構造

Input	Units: 100	Channels: 91
Convolution 1	Kernel size: 10	Channels: 128
	Max-pooling size: 10	
Convolution 2	Kernel size: 10	Channels: 128
	Max-pooling size: 10	
Output	Units: $4N_p$	

表 3　実験に用いた LSTM の構造

Input	Length: 25, Units: 91
Fully-connect	Units: 100
LSTM	Units: 100
Output	Units: $4N_p$

4·2　実　　装

本論文では，人体の部位検出処理に Convolutional Pose Machines （CPM）[Wei 16] を利用した．CPM は，深層学習モデルを用いた姿勢推定手法であり，入力画像に対して，図 4 に点で示した人体の部位 14 か所の位置座標を検出する．この場合，$P = 14$ であり，回帰モデルの入力の次元数は ${}_{14}\mathrm{C}_2 = 91$ となる．

また，本実験では音韻空間の構築に用いる音素の定量化手法として，既存の手法を用いた．音素の定量化手法は複数存在するが，[Doizaki 17] のように，具体的なパラメータが公開されていないものもあるため，パラメータが公開されている手法のうち，小松らが提案している 8 次元属性ベクトル [小松 09]，および秋山らが提案している 4 次元属性ベクトル [秋山 11, Komatsu 12] を用いた．小松らが提案するベクトルは，日本語音素の構成要素であるすべての母音，子音に対して「硬さ」,「強さ」,「湿度」,「滑らかさ」,「丸さ」,「弾性」,「速さ」,「温かさ」の 8 項目の属性で構成されている．各属性がとる値は，音韻がその属性に与える影響の大きさによって 2，1，0，−1，−2 のいずれかの値をとる．なお，これらの値はすべて経験的に定義されている．秋山らの手法は，1 つの音素に対して「キレ・俊敏さ」,「柔らかさ・丸み」,「躍動感」,「大きさ・安定感」の 4 つの属性値を割り当てたもので，これらは被験者実験と因子分析によって決定された実数値である．全体最近傍法および音素最近傍法で用いる距離尺度として，Euclidean 距離の 2 乗を用いた．

また，本論文では回帰モデルとして，深層学習モデルの一種である 1 次元の CNN（Convolutional Neural Network）および LSTM（Long Short Term Memory）を用いた．CNN は入力層で正規化された部位間相対距離系列 $L_{p_1,p_2}(t)$ それぞれをチャネルとみなしてチャネル数 91，ユニット数 100 の入力を受け付け，出力層は上述の $4N$ 次元の音韻ベクトルを出力する．$4N$ は小松らの手法の場合 32，秋山らの手法の場合 16 である．実験で用いた CNN および LSTM の構造をそれぞれ表 2，表 3 に示す．ここで，LSTM に関しては入力の系列長を 25 とし，

表 4 音韻空間上での多クラス分類結果

定量化手法	CNN	LSTM
[小松 09]	0.474	0.334
[秋山 11]	0.388	0.294

表 5 [小松 09]+CNN により生成されたオノマトペの例

真値	生成オノマトペ	距離
すたすた	せかせか	27.8
すたすた	てかてか	31.3
すたすた	すらすら	31.8
のろのろ	よろよろ	16.8
のろのろ	ゆろゆろ	23.6
のろのろ	ろろろろ	24.7
よろよろ	のろのろ	18.7
よろよろ	のらのら	23.5
よろよろ	とろとろ	25.0
どっしどっし	つこつこ	68.0
どっしどっし	とことこ	68.4
どっしどっし	そっとそっと	71.0
せかせか	すったすった	27.7
せかせか	つっとつっと	29.7
せかせか	つたつた	39.7
てくてく	すったすった	64.9
てくてく	とろとろ	72.1
てくてく	つらつら	74.3
とぼとぼ	ろろろろ	46.2
とぼとぼ	のろのろ	48.4
とぼとぼ	とそとそ	52.4
ぶらぶら	るちょるちょ	26.6
ぶらぶら	のろのろ	27.9
ぶらぶら	もろもろ	34.1

標本のフレームを間引いて入力している．これらの深層学習モデルの実装には Keras[*2] を用い，パラメータは実験的に設定した．

4·3 音韻空間上の多クラス分類

本節では，音韻空間上の歩容の多クラス分類実験について述べる．本実験では，クラスとして，データセットに含まれる 10 種類のオノマトペのうち，主観評価実験によりラベル付けされた映像数が少なかった「のしのし」及び「よたよた」を除いた 8 種類のオノマトペを用いた．データセットを学習用と評価用に分け，学習用データで回帰モデルを学習し，それを用いて評価用データを音韻空間に射影した．そして，音韻空間上で全体最近傍法により多クラス分類を行なった．評価は，歩行者別の Leave-one-person-out 交差検証で行なった．すなわち，歩行者 7 名中 6 名の歩容を学習用データとし，1 名の歩容を評価用データとする試行を 7 回行なった．評価指標にはミクロ平均正解率（Accuracy）を用いた．すなわち，全 fold の結果を合算した後に正解率を算出した．

実験の結果を表 4 に示す．なお，8 クラス分類であるため Chance rate は 0.125 である．いずれの回帰モデル，いずれの定量化手法による空間でも，Chance rate を大きく上回り，提案手法の有効性を確認した．また，CNN を用いた場合の方が LSTM を用いた場合よりも良い結果となった．これに関しては 4·5 節で考察する．

*2 https://keras.io./

表 6 写像から真値までの平均距離（誤差）

定量化手法	[小松 09]		[秋山 11]	
回帰モデル	CNN	LSTM	CNN	LSTM
すたすた	42.8	42.0	6.3	4.7
のろのろ	34.3	35.9	16.1	20.4
よろよろ	32.9	39.4	10.4	13.2
どっしどっし	87.0	90.4	37.3	39.7
せかせか	36.6	51.9	6.2	5.9
てくてく	87.5	68.6	14.8	13.1
とぼとぼ	58.4	54.6	10.3	9.3
ぶらぶら	44.8	48.3	9.9	8.0
平均	**53.0**	53.9	**13.9**	14.3

4·4 歩容に対する任意のオノマトペの生成

本節では，音素最近傍法を用いた歩容に対する任意のオノマトペの生成実験について述べる．本実験でも 4·3 節と同様に「のしのし」及び「よたよた」を除いた 8 種類のオノマトペのデータセットを用いた．評価は Leave-one-onomatopoeia-out 交差検証で行なった．すなわち，オノマトペ 8 種中 7 種を学習用データとし，1 種を評価用データとしてオノマトペを生成する試行を 8 回行なった．生成結果の例と，その射影から真値までの音韻空間上の距離を表 5 に示す．また，4·3 節の各手法で生成実験を行なった場合の，射影から真値まで音韻空間上での平均距離（誤差）の比較を表 6 に示す．

どちらの定量化手法においても，CNN により生成されたオノマトペの方が，LSTM で生成されたオノマトペより真値に近いことがわかる．しかしベースラインが存在しないため，この結果だけでは絶対的に手法の有効性を確認することができない．また，小松らの定量化手法と秋山らの定量化手法では値が取りうる範囲が異なるため，音韻空間上の距離を直接比較することができない．

そこで，生成されたオノマトペを元となった映像とともに評価者に提示し，生成オノマトペがどの程度映像中の歩容を表現できているかを問う主観評価実験を実施した．評価者は日本語を母語とする 20 代男性 9 名であった．回答方法は 7 段階の Likert scale を用いた．主観評価実験で用いたインタフェースを図 5 に示す．各映像に対して，真値，[小松 09]+CNN，[小松 09]+LSTM，[秋山 11]+CNN，[秋山 11]+LSTM の各条件で生成したオ

図 5 主観評価に用いたインタフェース

表 7 生成オノマトペの主観評価結果

	[小松 09]	[秋山 11]
真値	6.002 ± 0.048	
CNN	4.418 ± 0.075	4.075 ± 0.075
LSTM	3.025 ± 0.075	4.011 ± 0.076
ランダム	4.014 ± 0.081	

ノマトペ，ランダムに生成したオノマトペの 6 種類について評価した．なお，ランダムに生成したオノマトペには促音，撥音，拗音は含まれず，各音韻は等確率で出現するものとした．

各条件での評価値の平均と標準誤差を表 7 に示す．[小松 09] の定量化手法を用い，回帰モデルとして CNN を用いた場合が真値に次いで良い評価を得ていることがわかる．ここで，提案手法による生成の中で最も良い評価を得た [小松 09]+CNN の評価と，ベースラインであるランダムな生成結果の評価との有意差を検定した．Carifio らの議論 [Carifio 08] を参考とし，評価者ごとに回答の値を平均したものを標本として検定を実施した．両分布に正規性が認められなかったため，Wilcoxon の符号順位検定を適用した．各群の標本数は 9，[小松 09]+CNN の中央値は 4.45，ランダムの中央値は 4.16 であった．検定の結果，$p < 0.01$ で有意差が認められた．

4·5 考　　察

4·3 節における音韻空間を用いた分類は，単純な教師ありの多クラス分類問題を，学習に用いないオノマトペも認識できるように拡張したものと考えることができる．そこで，音韻空間を用いずに単純な End-to-end 型の多クラス分類を行なった結果と，4·3 節において最も良い精度を得た [小松 09]+CNN の結果を比較した．前者は，表 2 に示した CNN の出力層のユニット数を 8 とし，8 クラス分類問題として学習を行なった．そのほかの条件は 4·3 節に準じている．結果として，End-to-end 型の 8 クラス分類では正解率が 0.471 となった．4·3 節における [小松 09]+CNN の正解率は 0.474 となっており，ほぼ同等の結果が得られたと言える．これは，音韻空間の構成に適切な定量化手法を用いれば，音韻空間上での分類は単純な多クラス分類の妥当な拡張となることを示唆している．

表 8 学習標本数を減らした場合の比較

定量化手法	[小松 09]		[秋山 11]	
回帰モデル	CNN	LSTM	CNN	LSTM
$s = 5$	0.474	0.334	0.388	0.294
$s = 10$	0.429	0.299	0.333	0.287
$s = 20$	0.371	0.262	0.306	0.183
$s = 30$	0.321	0.206	0.309	0.164

次に，学習データの量に関して分析を行なう．データセットを新規に追加せずに標本数を増やすことは難しいため，逆に減らして実験することにより分類性能の変化を調べた．具体的には，4·3 節では標本切り出し時の最小ずらし幅を $s = 5$ としたが，これを $s = 10, 20, 30$ に変更し 4·3 節と同じ実験を行なった．これにより標本の切り出し密度が低くなり，学習標本数はそれぞれ約 45%，69%，78%減少した．比較した結果を表 8 に示す．いずれの回帰モデル，いずれの定量化手法の場合でも，標本を減らすことによりほぼ単調に分類性能が低下していることがわかる．逆に言えば，これ以上に標本数を増やすことができれば，本手法はさらなる性能向上が期待できるということを示唆している．また，いずれの定量化手法を用いた場合も，CNN で $s = 30$ とした場合が LSTM で $s = 5$ とした場合とほぼ同等の性能を示しており，これは CNN が LSTM よりも少ない標本数でも学習可能であることを示唆している．

5. むすび

本論文では，人体部位間の相対的な動きと音韻との関係性を利用し，任意の歩容をオノマトペにより記述する手法を提案した．提案手法では，人体部位の相対的位置関係に基づく特徴を抽出し，回帰を利用して音韻空間へと射影した．さらに，音韻空間上で最近傍法を用いることにより，多クラス分類問題としてオノマトペを選択することが可能なことを示し，さらには学習に用いない任意のオノマトペによって歩容を記述できる可能性も示した．

しかし，4·3 節の実験において，多クラス分類問題の正解率が 50%と必ずしも高くないことから，音韻空間の構築に用いる音素の定量化手法や回帰モデルなど，内部処理の更なる改良が求められる．本論文で用いた CNN は単純な構造であり，まだ改良の余地が多分に存在すると考えられる．同時に，データ不足に対応するため，転移学習の利用も検討する必要がある．さらに，4·4 節の生成実験に関しては，記述結果の主観評価の規模が十分でない問題がある．さらに大規模な評価実験を実施し，提案手法の有効性を検証することは今後の課題である．

本論文では撮影環境上の都合から，正面及び背面から撮影した歩容映像を用いたが，オノマトペの種類によっ

ては，側面から撮影した映像を用いた方が特徴を捉えやすい可能性がある．あるいは，3 次元姿勢推定を利用し，3 次元的な部位の位置関係を利用することも有用である可能性がある．また，人は新奇的なオノマトペを，一般的なオノマトペよりも受け入れ難いと感じる可能性がある．そのため，言語モデル等を用いて，より自然なオノマトペの生成を目指すことも重要である．今後は，オノマトペの音韻が対応する動きについて詳細に分析し，その知見に基づいて提案手法を一般化することで，あらゆる動きをオノマトペで記述したいと考えている．さらに，本論文では歩容からオノマトペを生成する手法について検討したが，映像特徴空間と音韻空間の対応を利用して，Generative Adversarial Networks（GAN）等を用いて，逆にオノマトペから歩容を生成することも検討したい．

謝　　辞

本研究の一部は栢森情報科学振興財団，科研費及び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の助成を受けて実施された．

◇ 参　考　文　献 ◇

[秋山 11] 秋山 広美, 小松 孝徳, 清河 幸子 : オノマトペから感じる印象の客観的数値化方法の提案, 情報処理学会研究報告 2011-HCI-142-23 (2011)

[Carifio 08] Carifio, J. and Perla, R.: Resolving the 50-year debate around using and misusing Likert scales, *Medical Education*, Vol. 42, No. 12, pp. 1150–1152 (2008)

[Doizaki 17] Doizaki, R., Watanabe, J., and Sakamoto, M.: Automatic estimation of multidimensional ratings from a single sound-symbolic word and word-based visualization of tactile perceptual space, *IEEE Trans. on Haptics*, Vol. 10, No. 2, pp. 173–182 (2017)

[藤野 05] 藤野 良孝, 井上 康生, 吉川 政夫, 仁科 エミ, 山田 恒夫 : 運動学習のためのスポーツオノマトペデータベース, 日本教育工学会論文誌, Vol. 29, pp. 5–8 (2005)

[Fukusato 14] Fukusato, T. and Morishima, S.: Automatic depiction of onomatopoeia in animation considering physical phenomena, in *Proc. Seventh Int. Conf. on Motion in Games*, pp. 161–169 (2014)

[Hamano 98] Hamano, S.: *The Sound-Symbolic System of Japanese*, CSLI Publications (1998)

[比屋根 98] 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)

[石原 03] 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)

[伊藤 15] 伊藤 惇貴, 加納 政芳, 中村 剛士, 小松 孝徳 : オノマトペの音象徴属性値の調整のための一手法, 人工知能学会論文誌, Vol. 30, No. 1, pp. 364–371 (2015)

[鍵谷 15] 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)

[神原 08] 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)

[小松 09] 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)

[Komatsu 12] Komatsu, T.: Quantifying Japanese onomatopoeias: Toward augmenting creative activities with onomatopoeias, in *Proc. Third Augmented Human Int. Conf.*, p. 15 (2012)

[權 17] 權 眞煥, 川嶋 卓也, 下田 和, 坂本 真樹 : DCNN を用いた画像の質感認知―音象徴性からのアプローチ―, 第 31 回人工知能学会全国大会 2L3-OS-09b-1 (2017)

[Li 18] Li, Q., Wang, Y., Sharf, A., Cao, Y., Tu, C., Chen, B., and Yu, S.: Classification of gait anomalies from Kinect, *Visual Computer*, Vol. 34, No. 2, pp. 229–241 (2018)

[呂 04] 呂 佳蓉 : 英語のオノマトペの象徴メカニズム, 言語科学論集, No. 10, pp. 99–116 (2004)

[小倉 16] 小倉 慶郎 : 日英オノマトペの考察 : 日英擬音語・擬態語の全体像を概観する, 大阪大学日本語日本文化教育センター授業研究, No. 14, pp. 23–33 (2016)

[小野 07] 小野 正弘 : 擬音語・擬態語日本語 4500 オノマトペ辞典, 小学館 (2007)

[Shimoda 15] Shimoda, W. and Yanai, K.: A visual analysis on recognizability and discriminability of onomatopoeia words with DCNN features, in *Proc. 2015 IEEE Int. Conf. on Multimedia and Expo*, pp. 1–6 (2015)

[杉山 11] 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)

[Sundaram 07] Sundaram, S. and Narayanan, S.: Analysis of audio clustering using word descriptions, in *Proc. 2007 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 2, pp. 769–772 (2007)

[Sundaram 08] Sundaram, S. and Narayanan, S.: Classification of sound clips by two schemes: Using onomatopoeia and semantic labels, in *Proc. 2008 IEEE Int. Conf. on Multimedia and Expo*, pp. 1341–1344 (2008)

[田守 99] 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 (1999)

[Wei 16] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y.: Convolutional pose machines, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition 2016*, pp. 4724–4732 (2016)

〔担当委員：折田 奈甫〕

2017 年 12 月 20 日　受理

著　者　紹　介

加藤　大貴

2016 年 名古屋大学工学部電気電子･情報工学科卒業．2018 年 同大学院情報科学研究科博士前期課程修了．現在，同研究科博士後期課程在学中．

平山　高嗣

2005 年大阪大学大学院基礎工学研究科博士後期課程修了．博士（工学）．同年より京都大学大学院情報学研究科特任助教．2011 年より名古屋大学大学院情報科学研究科特任助教．2012 年より同助教．2014 年より同特任准教授．2017 年より同大学未来社会創造機構特任准教授．顔画像認識，注視行動分析，視覚的注意の計算モデル，マルチモーダルインタラクションに関する研究に従事．電子情報通信学会，情報処理学会，ヒューマンインタフェース学会，IEEE，ACM 各会員．

道満　恵介

2012 年 名古屋大学大学院情報科学研究科博士後期課程修了．博士（情報科学）．2011 年 日本学術振興会特別研究員 DC2，2012 年 日本学術振興会特別研究員 PD，同年 中京大学情報理工学部講師，2013 年より同大学工学部講師．この間，2013〜16 年 名古屋大学大学院情報科学研究科招聘講師兼任，2017 年より同大学院情報学研究科招聘講師兼任．電子情報通信学会，IEEE，日本栄養改善学会 各会員．

井手　一郎（正会員）

2000 年 東京大学大学院工学系研究科電気工学専攻博士課程修了．博士（工学）．同年 国立情報学研究所助手．2004 年より名古屋大学大学院情報科学研究科助教授・准教授．2017 年より同大学院情報学研究科准教授．電子情報通信学会，情報処理学会各シニア会員，映像情報メディア学会，IEEE，ACM 各会員．

川西　康友

2012 年 京都大学大学院情報学研究科博士後期課程修了．博士（情報学）．同年 京都大学学術情報メディアセンター特定研究員．2014 年 名古屋大学未来社会創造機構特任助教．2015 年 同大学院情報科学研究科助教．2017 年 同大学院情報学研究科助教．現在に至る．防犯カメラ・車載カメラ映像を対象とした，人物検出・追跡・検索を含む人物画像処理に関する研究に従事．2011 年度 PRMU 研究奨励賞受賞．IEEE ITS Society Nagoya Chapter Young Researcher Award 受賞．IEEE, 電子情報通信学会各会員．

出口　大輔

2006 年 名古屋大学大学院情報科学研究科博士後期課程修了．博士（情報科学）．2004〜06 年まで日本学術振興会特別研究員．2006 年 名古屋大学大学院情報科学研究科研究員，同年 同大学院工学研究科研究員，2008〜12 年まで同大学院情報科学研究科助教，2012 年より同大学情報連携統括本部情報戦略室准教授．現在に至る．CARS2004 Poster Award，CADM2004 大会賞，2006 年 日本医用画像工学会奨励賞，2006 年 日本コンピュータ外科学会講演論文賞．電子情報通信学会，IEEE 各会員

村瀬　　洋

1980 年 名古屋大学大学院工学研究科修士課程修了．同年 日本電信電話公社（現 NTT）入社．1992 年から 1 年間 米国コロンビア大学客員研究員．2003 年から名古屋大学大学院情報科学研究科教授．2017 年から同大学院情報学研究科教授．工学博士．1994 年 IEEE-CVPR 最優秀論文賞，1996 年 IEEE-ICRA 最優秀ビデオ賞，2001 年 高柳記念奨励賞，2002 年 電子情報通信学会業績賞，2003 年 文部科学大臣賞，2004 年 IEEE Trans. MM 論文賞，2010 年 前島密賞，2012 年 紫綬褒章，ほか受賞．IEEE，電子情報通信学会，情報処理学会各フェロー．