一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
MVE2018-21(2018-09)

HOYO：オノマトペを付与した歩容データセット

加藤　大貴† 　平山　高嗣†† 　道満　恵介†††,† 川西　康友† 　井手　一郎†

出口　大輔††††,†村瀬　　洋†

† 名古屋大学　大学院情報学研究科〒 464–8601 愛知県名古屋市千種区不老町
†† 名古屋大学　未来社会創造機構〒 464–8601 愛知県名古屋市千種区不老町
††† 中京大学　工学部 〒 470–0393 愛知県豊田市貝津町床立 101
†††† 名古屋大学　情報戦略室 〒 464–8601 愛知県名古屋市千種区不老町
E-mail: †katoh@murase.is.i.nagoya-u.ac.jp, ††takatsugu.hirayama@nagoya-u.jp,
††††{kawanishi,ide,murase}@i.nagoya-u.ac.jp

あらまし　人間の歩行動作は多様なオノマトペで表現される．また，オノマトペには音象徴という性質があり，オノマトペから連想されるイメージはその音韻と強い関係があるとされる．このことから，音象徴に基づく「音韻空間」を歩容の特徴空間と対応付ければ，歩容の微妙な違いをオノマトペの音韻の違いで記述することができると考えられる．本報告では，その対応付けを行なうために構築した，多様なオノマトペを付与した歩容データセットを紹介する．
キーワード　オノマトペ，歩容，音韻，人体部位，音象徴性

HOYO: A Gait Dataset Annotated with Mimetic Words

Hirotaka KATO†, Takatsugu HIRAYAMA††, Keisuke DOMAN†††,†, Yasutomo KAWANISHI†,

Ichiro IDE†, Daisuke DEGUCHI††††,†, and Hiroshi MURASE†

† Graduate School of Informatics, Nagoya University Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
†† Institute of Innovation for Future Society, Nagoya University Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
††† School of Engineering, Chukyo University
101 Tokodachi, Kaizu-cho, Toyota-shi, Aichi, 470–0393 Japan
†††† Information Strategy Office, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
E-mail: †katoh@murase.is.i.nagoya-u.ac.jp, ††takatsugu.hirayama@nagoya-u.jp,
††††{kawanishi,ide,murase}@i.nagoya-u.ac.jp

Abstract　Gaits are expressed by various onomatopoeias according to their appearance. It is said that onomatopoeias have sound-symbolism and their phonemes are strongly related to their impressions. Thus, if the "phonetic-space" based on sound-symbolism can be linked with the feature-space of gaits, it is considered that subtle difference of gaits could be expressed as difference in phonemes. In this report, a gait dataset annotated with various onomatopoeias is introduced, for the purpose.
Key words　Onomatopoeia, gait, phoneme, body-parts, sound-symbolism

1. はじめに

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である．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技術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，人間が歩く際の身体運動の様子のことであり，その適切なアノテーション方法が提案されていないためである．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て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ると，人間が直感的に理解しやすい形で歩容の属性を記述し，計算機上で扱えるようにすることは重要である．

そこで，本報告では歩容の言語的な表現としてオノマトペに着目する．特に日本語においては，「のろのろ」，「つるつる」など，事象の様子を直感的に表現する手段として，多くのオノマトペが使用される [2]．

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 [3], [4]．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 [5]．ところが，質感画像 [6], [7] 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計算機が記述できれば，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 [8]．これをふまえ我々は，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，音韻の印象を表現する「音韻空間」へと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してきた [9], [10]．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．我々の提案手法では，オノマトペを音素単位に分解，定量化した「音韻ベクトル」に変換し，この音韻ベクトルと歩容の関係を回帰モデルにより学習する．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ことにより，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が可能となる．

しかし，この先行研究では，モデルの学習に用いたデータセットのアノテーションが 10 種のオノマトペの中から適当なものを選択する形式で行われており，ラベル（アノテーションされたオノマトペの種類）が 10 種しか存在していないという問題点がある．そのため，付与されたラベルが，歩容から真に想起されるオノマトペと厳密には一致していない可能性がある．また，回帰モデルを学習する際にも，目標変数（音韻空間上の点）が粗になってしまい，内挿外挿の学習に支障をきたす可能性がある．そこで，本報告ではより規模が大きなアノテーション実験を実施し，実験参加者にオノマトペを自由記述させることにより，より正確，より多様なオノマトペアノテーション付き歩容データセットの構築を行なう．

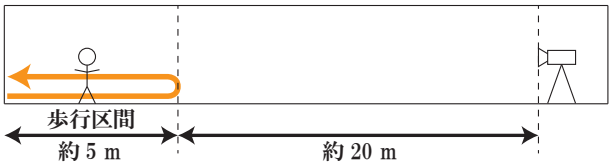


図 1　歩容映像の撮影状況

(a) 前面

(b) 背面

図 2　撮影した歩容映像の例

2. データセットの構築

2.1 撮　　影

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 社製のカメラ 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 1 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 1 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

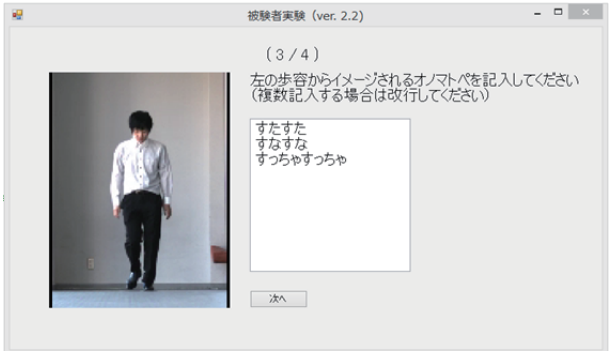
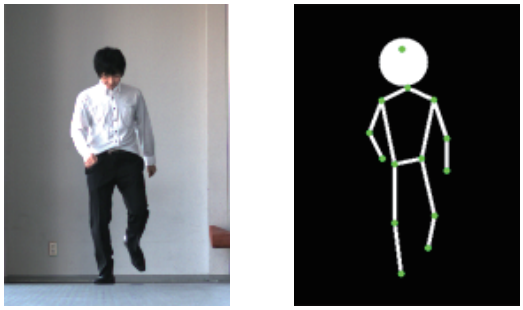


図 3　アノテーション実験で用いたインタフェース



(a) 元の映像フレーム　　(b) 歩容スケルトン

図 6　歩容スケルトンの例

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 回試行するようにした．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 [2] に掲載されているもののうち，構成する音韻の多様性を考慮して選択した．これは，より多様な歩容映像を収集するためである．この指示の種類を主観ラベルと呼ぶ．先行研究 [10] や本稿では主観ラベルを学習に用いることはないが，データセットに含める．歩行者は日本語を母語とする 20 代の男性 10 名であった．

映像はすべて 527 × 708 画素，60 fps で撮影し，最終的に 292 本の歩容映像を得た．撮影した映像の例を図 2 に示す．

2.2　アノテーション実験

アノテーション実験で用いたインタフェースを図 3 に示す．1. で述べた通り，先行研究のデータセットよりも多様で正確なオノマトペアノテーションを得るために，自由記述形式によるアノテーション実験を実施した．実験参加者には計算機上で歩容映像を見て，その歩容からイメージされるオノマトペを 3 つ記入するよう指示した．この際，アノテーション結果を取り扱いやすくするために，記入するオノマトペは ABAB 型（「すたすた」のように 2 音を 2 回繰り返す形式）のものに限定した．また，促音，拗音，長音が付与されたもの（「どっしどっし」等）と，第 2 モーラに撥音を用いたもの（「どんどん」等）は可とした．

実験には 2.1 節で得られた映像のうち，歩行者の前面を撮影した 146 本を用いた．参加者は，日本語を母語とする男女 30 名であり，映像 1 本あたり 15 人から回答を得た．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アノテーションをするものとした．

2.3　アノテーション結果

アノテーション実験の結果，6,570 件の回答を得て，うち 6,322 件が有効な回答であった．例えば，図 2(a) に示した歩容に対しては，おぞおぞ，がったがった，ぐらぐら（2），そろそろ，たらたら，ちょんちょん，とことこ，とたとた，とてとて，とろとろ（3），とんっとんっ，どんどん，のしのし（2），のそのそ，のろのろ，はくはく，ひょこひょこ，びくびく，ふらふら（8），ふわふわ，ぶらぶら，へいへい，へなへな，ぺすぺす，ほいほい，ゆっさゆっさ，ゆらゆら（3），よいよい，よたよた（2），よちよち，の計 45 件 31 種類の回答が得られた（括弧内の数字は重複した回答件数）．有効ではない回答としては，入力誤りと思われるものが 9 件，指示を無視して AA 型，ABCABC 型のオノマトペを回答したものが 239 件存在した．

このアノテーション結果を統合し，定量的に扱うために，回答を音素単位に分解・集計し，各音素の出現確率のベクトルとして表現する．例えば，前述の図 2(a) に示した歩容の例であれば，図 4 のようになる．図中の色の濃い部分は出現確率が高いことを示す．このベクトルは 46 次元の値からなり，最初の 15 次元が第 1 母音の種類，続く 5 次元が第 1 子音の種類，続く 15 次元が第 2 子音の種類，続く 6 次元が第 2 子音の種類，最後の 5 次元がそれぞれ，第 1 モーラ後の促音，第 2 モーラ後の促音，第 1 モーラ後の拗音，第 2 モーラ後の拗音，第 1 モーラ後の長音の出現確率を表している．第 1 子音，第 1 母音，第 2 子音，第 2 母音はそれぞれ音素の出現確率の和が 1 となっている．

このベクトルをすべての歩容映像に対して計算した結果の平均と標準偏差を図 5 に示す．平均値が大きい音素は，歩容を表現する上でよく使用される音素であるといえる．平均値が小さい割に標準偏差の大きい音素は，特定の歩容を表現するときに集中的に使われる音素であると言える．逆に，平均値が大きい割に標準偏差の小さい音素は，様々な歩容に満遍なく出現する汎用性の高い音素であると言える．

2.4　歩容スケルトンの作成

撮影実験参加者のプライバシーに配慮し，本データセットには元映像を含めずに歩容スケルトン，すなわち歩容映像から抽出した部位座標列を含める．抽出したスケルトンの例を図 6 に示す．これは，先行研究 [10] でも利用していた Convolutional Pose Machines （CPM）[11] を利用して検出した 14 箇所の部位座標列を基に，検出結果が誤っている部分を人手で修正したものである．

3.　構築したデータセットを用いた歩容の記述

本節では，上述のデータセットに我々が従来提案した手法 [10] を適用し，未知の歩容映像をオノマトペで記述する．

提案手法の処理手順を図 7 に示す．先行研究では，音韻ベクトルとして小松ら [12] の手法に基づきオノマトペを定量化した値を用いていたが，本報告では前述の音素出現確率ベクトルを

第1子音	ϕ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	0.0222	0	0.0222	0.2000	0.0889	0.3333	0	0.1778	0	0	0.0667	0	0.0222	0.0444	0.0222
第1母音	/a/	/i/	/u/	/e/	/o/										
	0.0667	0.0222	0.3556	0.0667	0.4889										
第2子音	ϕ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	0.1333	0.0889	0.1111	0.1333	0.0222	0	0	0	0.4667	0.0222	0	0.0222	0	0	0
第2母音	/a/	/i/	/u/	/e/	/o/	ん									
	0.4889	0.1333	0.0667	0.0222	0.2222	0.0667									
その他	っ1	っ2	ゃ1	ゃ2	ー										
	0.0444	0.0222	0.0667	0	0										

図 4　図 2(a) の歩容に対する音素出現確率ベクトル

第1子音	ϕ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	0.017488	0.046751	0.172095	0.288176	0.095725	0.106551	0.001891	0.066349	0.010189	0.000960	0.029794	0.026690	0.070741	0.031201	0.035368
	0.020717	0.043321	0.158045	0.132336	0.084562	0.098857	0.006323	0.074250	0.027074	0.004640	0.040823	0.035892	0.103883	0.026302	0.028459
第1母音	/a/	/i/	/u/	/e/	/o/										
	0.116503	0.021840	0.366055	0.128573	0.367016										
	0.065700	0.024541	0.128281	0.076689	0.143281										
第2子音	ϕ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	0.154895	0.161693	0.125186	0.209176	0.005906	0.000308	0.005940	0.001584	0.285725	0.005768	0.000155	0.000782	0.004213	0.037053	0.001603
	0.117239	0.105661	0.093543	0.126424	0.012158	0.002610	0.011916	0.005844	0.244002	0.014140	0.001872	0.004151	0.012290	0.048908	0.005917
第2母音	/a/	/i/	/u/	/e/	/o/	ん									
	0.390212	0.102027	0.133112	0.028726	0.215418	0.130488									
	0.136150	0.074600	0.079025	0.030824	0.131501	0.113627									
その他	っ1	っ2	ゃ1	ゃ2	ー										
	0.074560	0.038679	0.027047	0.008390	0.004823										
	0.065905	0.040127	0.037588	0.017667	0.013103										

図 5　データセット全体の音素出現確率ベクトル（各行上段が平均，下段が標準偏差）

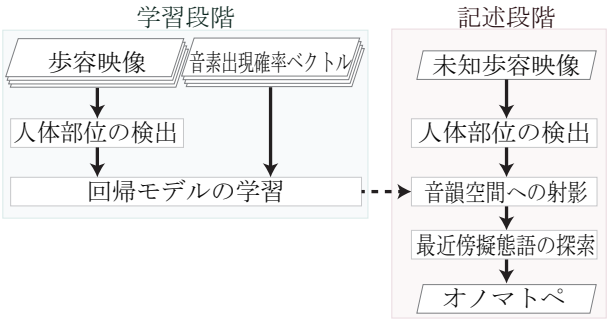
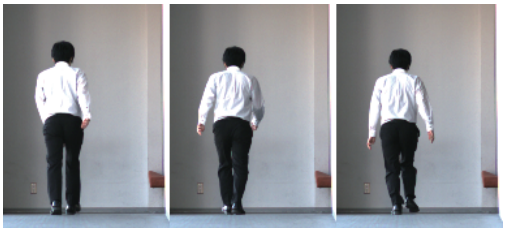


図 7　提案手法の処理手順

そのまま利用する．音韻ベクトルによって張られる空間を音韻空間と呼ぶ．この場合の音韻空間は各次元が $[0, 1]$ の値をとる $\mathbb{R}^{46}$ である．

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，スケルトンに含まれる 14 部位のうち全ての 2 部位の組み合わせにおける部位の相対距離系列を計算する．この相対距離系列を入力として回帰し，46 次元の音韻空間上に射影する．本手法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）を用いる．入力層で相対距離系列それぞれをチャンネルとみなしてチャンネル数 91，ユニット数 T の入力を受け付け，出力層は上述の 46 次元の音韻ベクトルを出力する．ここで，T は入力映像の長さである．本実験では $T = 100$ とし，データセットから部分系列を切り出して使用した．最後に，回帰モデルが出力した推定音韻ベクトルからオノマトペを生成する．本報告では，単純に第 1 子音，第 1 母音，第 2 子音，第 2 母音のそれぞれについて，最も値が大きい音素を選択することとする．促音等の付加要素については，値が 0.5 を超えた場合に付加するものとする．

(a) ののの



(b) すたすた

(c) とらとら

図 8　記述結果の例

以上の手法を用いて，未知歩容映像をオノマトペで記述する実験を行なった．学習・記述は 5 分割交差検定で実施した．記述結果例を図 8 に示す．また，推定された音素出現ベクトルと真値との平均 2 乗誤差は 0.167 であった．

4. むすび

本報告では，歩容をオノマトペで記述することを目指し，その学習のために，より正確で多様なオノマトペのアノテーションが付与された歩容映像データセットを構築した．

本報告では構築したデータセットによる記述結果例と平均 2 乗誤差のみを示したが，先行研究 [10] で実施したような主観評価実験を行ない，本データセットを用いた場合の記述精度の向上を確認することは今後の課題である．

なお，本データセットは以下の URL において後日公開予定である．

`http://www.murase.nuie.nagoya-u.ac.jp/~katoh/hoyo.html`

謝辞　本研究の一部は科研費および栢森情報科学振興財団の支援による．

文　献

[1] 川西康友，新村文郷，出口大輔，村瀬　洋，“サーベイ論文：画像からの歩行者属性認識，” 信学技報，PRMU2015-112，2015.
[2] 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.
[3] S. Hamano, The Sound-Symbolic System of Japanese, CSLI Publications, 1998.
[4] 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.
[5] 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，” 日本教育工学論，vol.29，pp.5–8，2005.
[6] 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 回人工知能学会全大 2L3-OS-09b-1，2017.
[7] W. Shimoda and K. Yanai, “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，” 人工知能学論，vol.30，no.1，pp.237–245，2015.
[9] H. Kato, T. Hirayama, Y. Kawanishi, K. Doman, I. Ide, D. Deguchi, and H. Murase, “Toward describing human gaits by onomatopoeias,” Proc, 2017 IEEE Int. Conf. on Computer Vision Workshop, pp.1573–1580, 2017.
[10] 加藤大貴，平山高嗣，道満恵介，川西康友，井手一郎，出口大輔，村瀬　洋，“音象徴性を利用したオノマトペによる歩容の記述，” 人工知能学論，vol.33，no.4，pp.B–HC2_1–9，2018.
[11] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition 2016, pp.4724–4732, 2016.
[12] 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，” 信学論（A），vol.J92-A，no.11，pp.752–763，2009.