2017 IEEE International Conference on Computer Vision Workshops

Toward Describing Human Gaits by Onomatopoeias

Hirotaka Kato[1] Takatsugu Hirayama[2] Yasutomo Kawanishi[2] Keisuke Doman[3] Ichiro Ide[2]
Daisuke Deguchi[4] Hiroshi Murase[2]

[1]Graduate School of Information Science, Nagoya University, Aichi, Japan
[2]Graduate School of Informatics, Nagoya University, Aichi, Japan
[3]School of Engineering, Chukyo University, Aichi, Japan
[4]Information Strategy Office, Nagoya University, Aichi, Japan

[1]katoh@murase.is.i.nagoya-u.ac.jp
[2]{hirayama, kawanishi, ide, murase}@i.nagoya-u.ac.jp
[3]kdoman@sist.chukyo-u.ac.jp
[4]ddeguchi@nagoya-u.jp

Abstract

Native Japanese people can distinguish gaits based on their appearances and briefly express them using various onomatopoeias to express their impressions intuitively. It is said that Japanese onomatopoeias have sound-symbolism and their phoneme is strongly related to the impression of a motion. Thus, 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. This framework is expected to make human-computer interaction more intuitive. In this paper, we propose a method to convert the relative body-parts movements to onomatopoeias using a deep-learning based regression model. Through experiments, we confirmed the effectiveness of the proposed method, and discussed the potential of describing an arbitrary gait by not only existing onomatopoeias but also a novel one.

1. Introduction

Onomatopoeia is a formation of a word from a sound associated with what is named. Most English onomatopoeias are actually used for imitation of sounds such as *bow-wow* or *tic-tac*. The Japanese language is known to have a greater number of onomatopoeias not only to imitate sounds but also to represent feelings or intuitive impressions of various phenomena. Researchers have focused on Japanese onomatopoeias representing the texture of an object to understand the mechanism of cross-modal perception and apply it to information systems. Human motion, especially gait, is a visually dynamical state most commonly represented by onomatopoeias, but it still has not attracted attention from any researchers in the field of computer science. A native Japanese speaker can easily distinguish gaits based on their appearances and express them briefly using various onomatopoeias in order to express their impressions intuitively. In this paper, we focus on gaits and propose a computational method to convert the kinetic features to onomatopoeias.

Japanese onomatopoeias are referred to as sound-symbolic words, which involve the association between linguistic sounds and sensory experiences [4]. The phonemes of an onomatopoeia should be strongly related to the visual sensation when observing a gait so that the onomatopoeias can describe the difference in the appearance of gaits at a fine resolution [3]. In the Japanese language, it is said that there are more than fifty gait-related onomatopoeias. For example, according to a Japanese onomatopoeia dictionary [6], *noro-noro* means "slowly walk without having a vigorous intention to move forward," and *yoro-yoro* means "walk with an unstable balance." Their difference of only one sound, i.e. /n/ or /y/, can represent a slight difference in gaits. In addition, *suta-suta* means "walk with light steps without observing around," and *seka-seka* means "trot as being forced to hurry." As we can see from these examples, the phoneme /s/ seems to express an impression of fast, smooth, and stable motion. Such associations are individual-invariant and linguistic-invariant similar to the Bouba/kiki-effect [7].

Inspired by this cross-modal perception, we attempt to form a computational model that describes an arbitrary gait

2473-9944/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCVW.2017.185

by an onomatopoeia. If a phonetic space simulating the sound-symbolism can be constructed and associated with a kinetic feature space of gaits, difference of gait impressions can be computationally expressed as difference in phoneme. This framework enables us to assign not only the existing onomatopoeias but also a novel one generated from an arbitrary combination of phonemes to the gaits. The former is a classification task (see Section 5.2) and the latter is regarded as a kind of zero-shot translation task (see Section 5.3).

We can apply the proposed method to human-computer interaction. It makes the interaction more intuitive as Sakamoto et al. [8] revealed that their onomatopoeia quantification system is useful for communication between Japanese patients and foreign doctors. As an example of the application, an advanced driver assistance system can warn the driver of a specific pedestrian by intuitively expressing his/her gait by an onomatopoeia, which they should immediately pay attention to. Also, it can be used to understand eyewitness testimonies including an onomatopoeia of gait in an interactive surveillance system in order to identify pedestrians with such gaits in videos. Such applications which use onomatopoeias as an attribute of motion would be available for even those who do not speak the language owing to the sound-symbolism.

Our contributions are twofold: (1) We motivate the focus on onomatopoeias representing visually dynamic states, (2) We propose a computational model to associate the kinetic feature space of gait with the phonetic space simulating the sound-symbolism, and the model can also describe a gait by a novel onomatopoeia generated from an arbitrary combination of phonemes. The rest of the paper is composed as follows: Related work is introduced in Section 2. Section 3 describes the proposed method that converts human gaits to onomatopoeias. More concretely, it projects the kinetic features into the phonetic space using a deep learning based regression model. Section 4 introduces the dataset used for evaluation. Section 5 reports results from two experiments of classification and zero-shot translation, and in Section 6, we discuss the experimental results. Finally, the paper is concluded in Section 7.

2. Related work

There are some previous works on onomatopoeias associated with auditory, visual, and tactile modality in the field of computer science.

Sundaram et al. proposed a "meaning space" having the semantic word based similarity metric that can be used to cluster acoustic features extracted from audio clips tagged with English onomatopoeias [10]. They also constructed a latent perceptual space using audio clips categorized by high-level semantic labels and the mid-level perceptually motivated onomatopoeia labels [11]. Fukusato et al. proposed a method to estimate an onomatopoeia imitating a collision sound (e.g. "Bang") from physical characteristics of the objects [2]. Shimoda et al. demonstrated that Web images searched with different onomatopoeias have discriminable visual features [9]. Doizaki et al. proposed an onomatopoeia quantification system [1] which is based on sound-symbolism and quantifies onomatopoeias based on prior subjective evaluations using 26 opposing pairs of tactile adjectives such as "hard – soft".

These works target onomatopoeias imitating sounds or representing visually static states. Meanwhile, as mentioned in Section 1, in this paper, we attempt to computationally describe human gaits as visually dynamic states by onomatopoeias.

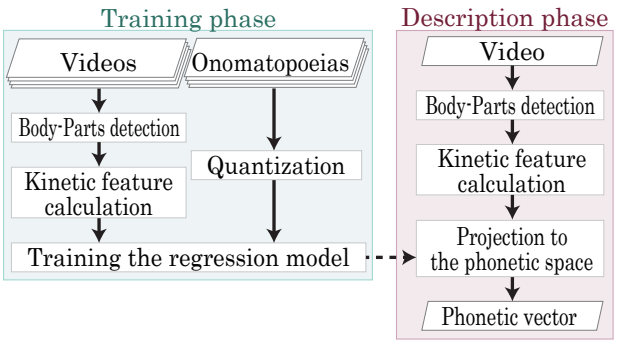


Figure 1. Procedure of the proposed method.

3. Gaits description by onomatopoeias

The procedure of the proposed method is shown in Figure 1. In our method, we map the kinetic features extracted from videos to the phonetic space by regression. It consists of the training phase and the description phase. In the following subsection, we explain in detail the kinetic features, the phonetic space, and the regression model.

3.1. Body parts detection and kinetic feature extraction

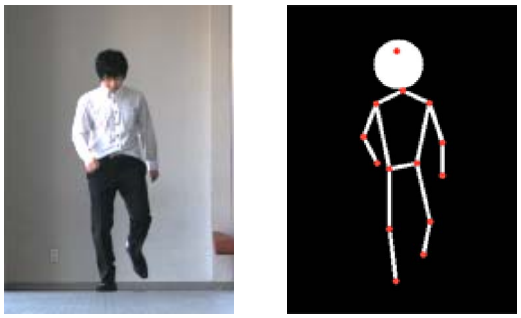
Li et al. proposed an algorithm for fine-grained classification of walking disorders arising from neuro-degenerative diseases such as Parkinson and Hemiplegia, by referring to relative body-parts movement [5]. In line with this work, we use kinetic features based on the relative movement of body parts.

As a preprocess, we need to detect the body parts of a pedestrian from videos. Here, we use Convolutional Pose Machines (CPM) [13]. CPM is an articulated pose estimation method based on a deep learning model. It can detect 14 parts of human body, and yields their pixel coordinates. Figures 2(a) and (b) show an example of the original video frame and the corresponding result of parts detection using CPM, respectively.

First, we apply CPM to each frame of an input video

Table 1. Phoneme quantization by eight attributes on impression of phoneme [12].

		Hardness	Intensity	Humidity	Fluidity	Roundness	Elasticity	Speed	Warmth
Vowel	/a/	0	1	−1	1	2	−2	−1	0
	/i/	2	2	0	0	−1	1	2	−1
	/u/	−1	−1	2	0	2	2	0	2
	/e/	1	−2	2	0	−2	0	0	2
	/o/	−1	2	0	1	2	0	−2	1
Consonant	/k/	2	2	1	0	0	0	2	−1
	/s/	2	0	1	2	0	0	2	−1
	/t/	2	1	2	2	0	1	−1	−2
	/n/	−1	0	2	−1	1	0	−2	2
	/h/	−2	−2	1	0	1	−1	−1	2
	/m/	−2	−2	1	0	2	0	−1	2
	/y/	−2	−1	0	1	2	1	0	0
	/r/	−1	−1	2	1	0	2	1	0
	/w/	−2	2	1	0	2	0	0	1
	/g/	3	3	0	−1	−1	0	1	−1
	/z/	3	1	0	1	−1	0	1	−1
	/d/	3	2	1	1	−1	1	−2	−2
	/b/	−1	−1	0	−1	0	−1	−2	2
	/p/	−3	−3	1	0	2	0	0	3
Contracted sounds		−1	−1	+1	±0	+1	+2	+2	+1
Doubled consonants		±0	±0	±0	±0	±0	±0	+1	±0



(a) Original video frame. (b) Parts detection result.

Figure 2. Example of body parts detection using CPM [13].

and obtain 14 sequences of pixel coordinates $\boldsymbol{P}(p,t) \in \mathbb{R}^2$. Here, $p \in \{0,\ldots,13\}$ indicates the index of each body part, and $t \in \{1,\ldots,T\}$ indicates the index of each video frame where the length of the input video is T. Next, we calculate the Euclidean distance $D_{p_1,p_2}(t)$ between arbitrary pairs of parts p_1 and p_2. Then, we calculate the human height $H(t)$, namely, the difference in y-coordinates between head and foot, and their average in the sequence $\bar{H}$. Finally, we divide all of $D_{p_1,p_2}(t)$ by $\bar{H}$, and obtain a sequence of the normalized body-parts distance $L_{p_1,p_2}(t)$. Note that the number of combinations of p_1 and p_2 under the condition $p_1 < p_2$ is ${}_{14}\mathrm{C}_2 = 91$.

3.2. Phonetic space

In order to construct the phonetic space, we need to quantize phonemes based on sound symbolism. The onomatopoeia quantification system by Doizaki et al. [1] seems to be an appropriate reference for our objective. Unfortunately, the parameters of the state-of-the-art system are not publicly available, and also, it is difficult to reconstruct the system accurately because we need to conduct a large scale subjective experiment. Thus, we decided to refer to another quantization proposed by Tomoto et al. [12], which is publicly available. They have argued each Japanese phoneme can be represented by an 8-dimensional vector consisting of eight attributes on the impression of phonemes: hardness, intensity, humidity, fluidity, roundness, elasticity, speed, and warmth.

Table 1 shows the correspondence between each phoneme and the values of an 8-dimensional vector [12]. This table covers all vowels, all consonants, contracted sounds, and double consonants in Japanese. The contracted sounds, e.g. /ky/, /sy/, are consonants accompanied by the consonant /y/. The double consonants, e.g. /kk/, /ss/, make the pronunciation of the preceding vowel shorter. When such consonant variation occurs, the vectors of contracted sounds and double consonants will have values modified by adding the weights to the values of the original consonant.

Based on the table, Tomoto et al. [12] proposed a 32-dimensional phonetic vector for "ABCD-ABCD"-type onomatopoeias, which composes the majority of ono-

Table 2. Selected onomatopoeias and their meanings [6].

Onomatopoeia	Meaning
suta-suta	Walk with light steps without observing around
noro-noro	Slowly walk without having a vigorous intention to move forward
yoro-yoro	Walk with an unstable balance
dossi-dossi	Walk with one's weight by stepping on the ground forcefully
seka-seka	Trot as being forced to hurry
teku-teku	Walk by firmly stepping on the ground for a long distance
tobo-tobo	Walk with dropping one's shoulder for a long distance
noshi-noshi	Walk with heavy steps forcefully
yota-yota	Walk with weak steps as with an elderly or a patient
bura-bura	Walk without having any intention

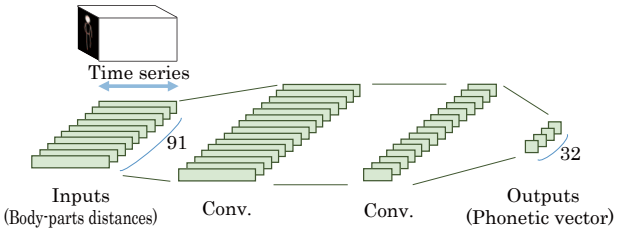


Figure 3. Overview of the regression model.

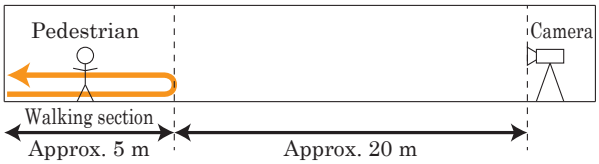


Figure 4. Video recording environment.

matopoeias representing gaits such as *suta-suta* and *yoro-yoro*. First, they decompose an "ABCD-ABCD"-type onomatopoeia into two sets of four phonemes. For example, *suta-suta* can be decomposed into "/s/, /u/, /t/, /a/" and "/s/, /u/, /t/, /a/." Next, they focus on only one set of the four phonemes and translate each phoneme into an 8-dimensional phonetic vector according to Table 1. Finally, they concatenate the four phonetic vectors, which yields a 32-dimensional phonetic vector. We use this phonetic vector to construct the 32-dimensional phonetic space.

3.3. Regression model

As a regression model, we use a 1-dimensional Convolutional Neural Network (CNN). Figure 3 shows the overview of the model. We aim to analyze the temporal change of the body-parts movement on multiple time-scales by using a temporal convolution process. The input data is 91 ($=_{14}\mathrm{C}_2$) sequences of body-parts distance $L_{p_1,p_2}(t)$, and the number of units in the input layer is T. Here, we handle each sequence as a channel. The supervisory data in the training phase and the output data in the description phase are 32-dimensional phonetic vectors. The network architecture of the regression model is explained in more detail in Section 5.2.

We project the kinetic features to the phonetic space by using this model. In other words, this process is regarded as a spatial transformation.

4. Dataset

We constructed a dataset that includes videos recording human gaits and their onomatopoeia labels. In this section, we introduce the video recording session and the onomatopoeia labeling experiment.

4.1. Video recording

The video recording was made over a single actor at a time. Figure 4 shows the environment of the video recording. The walking section was approximately five meters long.

Seven actors who were native Japanese University students in their twenties were asked to walk with a gait representing an onomatopoeia back and forth the walking section. Table 2 shows a list of instructed onomatopoeias and their meanings for reference. The ten onomatopoeias express typical impression of gaits, which were chosen from 56 onomatopoeias of gaits appeared in a Japanese onomatopoeia dictionary [6]. We recorded some ordinary gaits as well. Finally, we recorded 158 gaits (79 from front of the actors and the paired 79 from back).

The videos were taken at a rate of 60 fps, 527×708 pixels resolution, and 8-bit color. We used a USB 3.0 camera Flea3[1] produced by Point Gray Research, Inc. The camera was set approximately twenty meters away from the termination of the walking section to suppress the scale variation of body appearance due to walking along the optical axis of the camera.

[1]Sensor size: 2/3 inch, Focal length of lens: 35 mm.

Table 3. Result of subjective experiment.

Instruction to pedestrian	Number of labeled videos										Total
	suta	*noro*	*yoro*	*dossi*	*seka*	*teku*	*tobo*	*noshi*	*yota*	*bura*	
Ordinary	5	0	0	0	0	6	0	0	0	0	11
suta	11	0	0	0	4	2	0	0	0	0	17
noro	0	5	0	0	0	0	2	0	0	2	9
yoro	0	0	7	0	0	0	0	0	2	2	11
dossi	0	0	0	7	0	1	0	1	0	0	9
seka	1	0	0	0	3	0	0	0	0	0	4
teku	1	0	0	0	1	1	0	0	0	0	3
tobo	0	0	0	0	0	0	3	0	0	0	3
noshi	0	1	0	0	0	0	2	0	1	0	4
yota	0	0	2	0	0	0	2	0	0	0	4
bura	0	0	1	0	0	0	0	0	0	2	3
Total	18	6	10	7	8	10	9	1	3	6	78

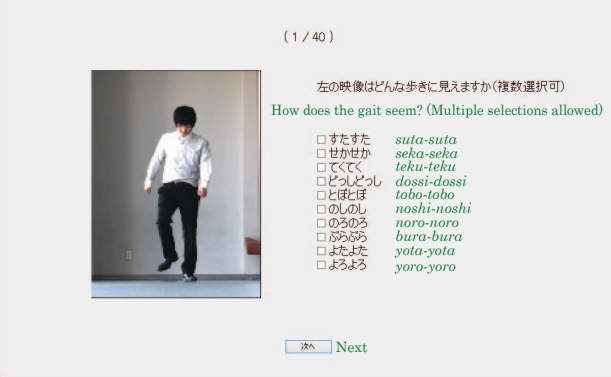


Figure 5. Annotation tool.

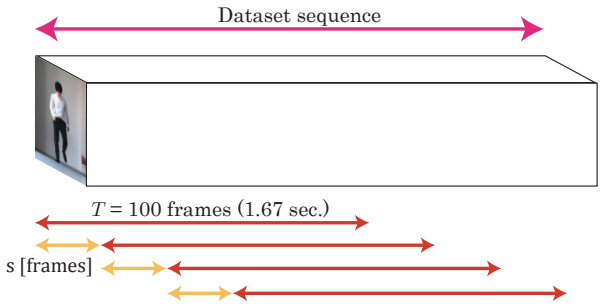


Figure 6. Extraction of samples from video sequence.

4.2. Onomatopoeia labeling

Although we asked the actors to walk with gaits representing specific onomatopoeias, they could not always move their bodies as they intended because they were not skilled actors. Thus, we should not use the instructed onomatopoeia as the ground-truth label of the corresponding gait. In order to annotate the gait recorded in each video, we conducted a subjective experiment and asked evaluators to annotate the videos with onomatopoeias.

Fourteen evaluators who were native Japanese University students in their twenties watched 79 videos showing the gaits from the front and annotated each video with zero or more labels which were selected from the ten onomatopoeias shown in Table 2. Figure 5 shows the annotation tool. Seven evaluators were assigned to each video. Each onomatopoeia was annotated when the majority of the evaluators selected it. The result is shown in Table 3, whose rows indicate the onomatopoeias instructed to the actors, and columns indicate the labeled results. Note that the total number of labels in Table 3 does not match the number of videos (= 79) because some videos were annotated with multiple or no onomatopoeias. Finally, the same label was assigned to the video showing the gaits from the back paired with the front.

If a video was annotated with various onomatopoeias, we assigned the onomatopoeia having the most votes to the video as the ground-truth label, and unfortunately, only a few videos were assigned to two classes *noshi-noshi* and *yota-yota* as shown in Table 3, so only the other eight classes were used in the following experiments.

5. Experiments

In this section, we report the result of two experiments. The first one is a multiclass-classification experiment for evaluating the performance of the regression model. The second one is a zero-shot translation experiment for describing gaits by an existing onomatopoeia or a novel one.

5.1. Sampling

Before conducting the experiments, we sampled training and test data from the dataset. Since the length of video sequences in the dataset are uneven, we sampled subsequences with a fixed length T [frames] defined in Section 3.1 by shifting the initial frame every s [frames]. As

Table 4. Network architecture of the proposed method using 1-dimensional CNN.

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: 32	

Table 5. Network architecture of the comparative method 1 using LSTM.

Input	Length: 100, Units: 91
Fully-connect	Units: 100
LSTM	Units: 100
Output	Units: 32

shown in Figure 6, we set the length T as 100 frames (1.67 sec.) so that it should contain one cycle of walking, i.e. two steps. The parameter s was changed according to each onomatopoeia class so that the number of samples should become roughly even among all the classes. In the following, we set s as five frames for the onomatopoeia class with the shortest sequence of videos.

5.2. Multiclass-classification on the phonetic space

To evaluate the regression model, we conducted multiclass classification on the phonetic space. A feature vector projected on the phonetic space was classified to one of the eight classes using a Nearest Neighbor method. The accuracy was calculated in a leave-one-actor-out cross-validation scheme that uses the samples extracted from the videos of an actor as test data and the others as training data until all the actors were used for testing.

We employed two comparative methods. One used a Long Short Term Memory (LSTM) model instead of the 1-dimensional CNN. Another applied a linear Support Vector Regression (SVR) to the following statistical features: temporal variance and kurtosis of the body-parts distance $L_{p_1,p_2}(t)$. The 1-dimensional CNN and the LSTM were implemented using Keras[2] whose network architectures are shown in Tables 4 and 5, respectively.

Table 6 shows the comparison in accuracy between the proposed and the comparative methods. Note that the baseline (chance level) accuracy is 0.125. We confirmed that the proposed method achieved the highest accuracy.

5.3. Zero-shot translation from gait to onomatopoeia

As a more challenging task, we conducted a zero-shot translation from a gait to an existing onomatopoeia or a novel one generated from an arbitrary combination of phonetic vectors using the proposed method. It can be carried out by a leave-one-onomatopoeia-out cross validation that uses the samples extracted from the videos assigned to an onomatopoeia class as test data and the others as training data until all the onomatopoeia classes are used for testing. For example, when we use *suta-suta* labeled gaits as test data, the regression model is trained using the gaits labeled with onomatopoeias other than *suta-suta*. Thus, *suta-suta* is regarded as an unknown onomatopoeia in this case. We verbalized each phoneme of test data projected on the phonetic space with a Nearest Neighbor method. More concretely, we divided the projected 32-dimensional phonetic vector into four 8-dimensional phonetic vectors, and selected a vowel or a consonant nearest to each of them based on Table 1. This process can generate a novel onomatopoeia by an arbitrary combination of phonetic vectors.

Table 6. Accuracies of multiclass-classification.

Method	Accuracy
Proposed (1-dimensional CNN)	**0.474**
Comparative 1 (LSTM)	0.334
Comparative 2 (Linear SVR)	0.338
Baseline (Chance level)	0.125

Tables 7 and 8 show examples of translation results by the proposed method and the representative onomatopoeia generated from the median vector of the projected vectors by each of the proposed and the comparative methods in each validation, respectively. In the tables, "*" indicates a novel onomatopoeia that does not appear in the Japanese onomatopoeia dictionary as gait-related onomatopoeias.

To evaluate the results of the zero-shot translation, we conducted a subjective evaluation. We showed the ground-truth and the representative onomatopoeias generated by the proposed and the comparative methods (see Table 8) to seven subjects who were native Japanese University students in their twenties, and asked them to select the more similar one to the ground-truth. The result is shown in Table 9. Note that when both methods generated the same onomatopoeia in the cases of *bura-bura* of the proposed method versus comparative method 1 and *teku-teku* of the proposed method versus comparative method 2, we regarded the selection rate as 0.5 because we could not judge the superiority of either methods. As shown in the table, the proposed method achieved better selection rate greatly over 0.5 for both comparison methods. We confirmed that the proposed method has the potential of describing an arbitrary gait by not only an existing onomatopoeia but also a novel one.

6. Discussion

The proposed method learns the relationship between the kinetic features and the phonetic features using a regression model to classify the gaits to the onomatopoeia classes.

[2]https://keras.io./

Table 7. Examples of zero-shot translation by the proposed method.

Ground-truth	Description	Distance
suta-suta	*seka-seka*	27.8
suta-suta	**teka-teka*	31.3
suta-suta	**sura-sura*	31.8
noro-noro	*yoro-yoro*	16.8
noro-noro	**yuro-yuro*	23.6
noro-noro	**roro-roro*	24.7
yoro-yoro	*noro-noro*	18.7
yoro-yoro	**nora-nora*	23.5
yoro-yoro	**toro-toro*	25.0
dossi-dossi	**tuko-tuko*	68.0
dossi-dossi	**toro-toro*	69.4
dossi-dossi	**sotto-sotto*	71.0
seka-seka	**sutta-sutta*	27.7
seka-seka	**sutto-sutto*	28.2
seka-seka	**tuto-tuto*	35.1
teku-teku	**sutta-sutta*	64.9
teku-teku	**totta-totta*	71.0
teku-teku	**tura-tura*	74.3
tobo-tobo	**roro-roro*	46.2
tobo-tobo	*noro-noro*	48.4
tobo-tobo	**toso-toso*	52.4
bura-bura	**rutyo-rutyo*	26.6
bura-bura	*noro-noro*	27.9
bura-bura	**moro-moro*	34.1

Table 8. Representative onomatopoeias generated from the median vectors of the projected vectors by the proposed and the comparative methods.

Ground-truth	Proposed	Comp. 1	Comp. 2
suta-suta	**teko-teko*	**toro-toro*	**toro-toro*
noro-noro	**toro-toro*	**tura-tura*	**tuso-tuso*
yoro-yoro	*noro-noro*	**tuso-tuso*	**tusa-tusa*
dossi-dossi	*toko-toko*	*yoro-yoro*	**suko-suko*
seka-seka	**sutta-sutta*	**toro-toro*	**toro-toro*
teku-teku	**toro-toro*	**sutta-sutta*	**toro-toro*
tobo-tobo	*yoro-yoro*	**tusa-tusa*	**turo-turo*
bura-bura	*noro-noro*	*noro-noro*	**toro-toro*

To obtain better accuracy of the multiclass-classification, it might be better off learning the direct relationship between the kinetic features and the onomatopoeia labels using a classification model. Here, we compare the multiclass-classification accuracy of the proposed method with a more end-to-end classification method. To realize the latter one, we replaced the output layer having 32 units of 1-dimensional CNN shown in Table 4 with eight units corresponding to eight onomatopoeia classes and trained the modified CNN from scratch. It can output one of eight onomatopoeias without using the phonetic space. As the result of multiclass-classification, we obtained 0.471 for the accuracy of the more end-to-end classification method, which was almost the same as the proposed method. The result implies the proposed method has more general capability of onomatopoeia expression with a higher degree of freedom.

Table 9. Selection rates of representative onomatopoeias generated by the proposed method.

Onomatopoeias	vs. Comp. 1	vs. Comp. 2
suta-suta	1.000	1.000
noro-noro	1.000	1.000
yoro-yoro	1.000	1.000
dossi-dossi	0.857	0.286
seka-seka	1.000	1.000
teku-teku	0.000	0.500
tobo-tobo	1.000	0.714
bura-bura	0.500	0.286
Average	0.795	0.723

Table 10. Comparison of classification results with a smaller dataset.

s [frames]	Proposed	Comp. 1	Comp. 2
5	0.474	0.334	0.338
10	0.429	0.299	0.341

Next, we analyzed whether the accuracy of the proposed method depends on the number of training data. We conduct the same experiment as in Section 5.2 but with smaller samples. In Section 5, we set the parameter s for sampling data as five frames for the onomatopoeia class with the shortest sequence of the videos. Here, we set the minimum of s at ten frames so that the number of samples becomes approximately 45% smaller. Table 10 shows the result. Comparative method 2 based on linear SVR kept the accuracy whereas in the other methods the accuracy was degraded. The result implies that the proposed method and comparative method 1 based on LSTM have the potential to improve the accuracy by enlarging the dataset size.

7. Conclusions

In this paper, we proposed a method for describing human gaits by onomatopoeias, which uses a kinetic feature, a phonetic space, and a 1-dimensional CNN regression model. We conducted two experiments, namely, multiclass-classification and zero-shot translation task, and confirmed the effectiveness of the proposed model.

The most critical limitation of this work is insufficient size of data. To solve the problem, we will attempt to acquire a larger data of onomatopoeia-labeled gaits using crowdsourcing and introduce transfer learning to train the CNN. Some temporal models (e.g. action recognition model) might be available for the latter approach.

Beyond describing human gaits by onomatopoeias, we will consider generating motions from onomatopoeias as the reverse framework to this work. It should be useful

for producers of computer graphics, operators of humanoid robots, and so on.

Acknowledgements

Parts of this work were supported by MEXT, Grant-in-Aid for Scientific Research and the Kayamori Foundation of Information Science Advancement.

References

[1] R. Doizaki, J. Watanabe, and M. Sakamoto. Automatic estimation of multidimensional ratings from a single sound-symbolic word and word-based visualization of tactile perceptual space. *IEEE Trans. on Haptics*, 10(2):173–182, 2017.

[2] T. Fukusato and S. Morishima. Automatic depiction of onomatopoeia in animation considering physical phenomena. In *Proc. 7th ACM Int. Conf. on Motion in Games*, pages 161–169. ACM, 2014.

[3] S. Hamano. *The Sound-Symbolic System of Japanese.* CSLI Publications, 1998.

[4] W. Köhler. *Gestalt Psychology: An Introduction to New Concepts in Modern Psychology*. WW Norton & Company, 1970.

[5] Q. Li, Y. Wang, A. Sharf, Y. Cao, C. Tu, B. Chen, and S. Yu. Classification of gait anomalies from Kinect. *Visual Computer*, pages 1–13, 2016.

[6] M. Ono. *Japanese Onomatopoeia Dictionary (in Japanese)*. Shogakukan Press, Tokyo, 2007.

[7] V. S. Ramachandran and E. M. Hubbard. Synaesthesia—A window into perception, thought and language. *Journal of Consciousness Studies*, 8(12):3–34, 2001.

[8] M. Sakamoto, Y. Ueda, R. Doizaki, and Y. Shimizu. Communication support system between Japanese patients and foreign doctors using onomatopoeia to express pain symptoms. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 18(6):1020–1025, 2014.

[9] W. Shimoda and K. Yanai. A visual analysis on recognizability and discriminability of onomatopoeia words with DCNN features. In *Proc. 2015 IEEE Int. Conf. on Multimedia and Expo*, pages 1–6, 2015.

[10] S. Sundaram and S. Narayanan. Analysis of audio clustering using word descriptions. In *Proc. 2007 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 769–772, 2007.

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

[12] Y. Tomoto, T. Nakamura, M. Kanoh, and T. Komatsu. Visualization of similarity relationships by onomatopoeia thesaurus map. In *Proc. 2010 IEEE World Congress on Computational Intelligence*, pages 3304–3309, 2010.

[13] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition 2016*, pages 4724–4732, 2016.