Intuitive Gait Modeling using Mimetic-Words for Gait Description and Generation

Hirotaka Kato
Meijo University
1-501 Shiogama-guchi, Tenpaku-ku,
Nagoya, Aichi, Japan
shien249@ccmails.meijo-u.ac.jp

Takatsugu Hirayama
University of Human Environments
6-2 Kami-sanbonmatsu, Motojuku-cho,
Okazaki, Aichi, Japan
t-hirayama@uhe.ac.jp

Keisuke Doman
Chukyo University
101 Tokodachi, Kaizu-cho, Toyota, Aichi, Japan
kdoman@sist.chukyo-u.ac.jp

Ichiro Ide
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
ide@i.nagoya-u.ac.jp

Yasutomo Kawanishi
RIKEN
9-3 Kizugawa-dai, Kizugawa, Kyoto, Japan
yasutomo.kawanishi@riken.jp

Takahiro Komamizu
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
taka-coma@acm.org

Daisuke Deguchi
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
ddeguchi@nagoya-u.jp

Hiroshi Murase
Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan
murase@nagoya-u.jp

Abstract

Gait is one of the most familiar action for us, that is why we can distinguish slight difference of human gaits and perceive their impressions. However, the relationship has been never explored because of the absence of intuitive labels for the slight differences. In this paper, to solve this problem, we propose a intuitive gait model using Japanese mimetic-words. A mimetic-word has sound-symbolism, which means that there is an association between linguistic sounds and sensory experiences, and the phonemes of a mimetic-word is strongly related to the visual sensation. Thanks to the sound-symbolism, Japanese mimetic-words have a possibility of modeling gaits intuitively. Thus, we have previously proposed a method which describes gait with a mimetic-word. In this paper, in the opposite direction, we propose a method which generates gait from a mimetic-word, and confirm the effectiveness of the proposed intuitive gait model which consists of the phonetic-vector through evaluations of both the generation task and the description task.

1. Introduction

Gait is one of the most familiar action for us, that is why we can distinguish slight difference of human gaits and perceive their impressions such as "*strong-steps*" or "*light-steps*". Gait information is known to be useful for various tasks, such as anomaly detection or emotion recognition. In recent years, human pose and their movement information has become easier to analyze, thanks to high quality pose estimation methods such as OpenPose [2]. In the research field of gait analysis, researchers have been interested in "why" a person is walking so, such as because of the person's emotion, body condition, or intent. However, there is no study focusing on recognizing a person's walking action in more fine granularity than "running" or "walking". This task can be rephrased as recognition of "how" the person is walking. This is partly because there is no label capable of describing gait in a fine-grained manner.

On the other hand, more and more online events using remote meeting systems or streaming platforms are held in recent years. Accordingly, a method which can set intuitively a motion to an user's virtual avatar even for non-

2770-4319/22/$31.00 ©2022 IEEE
DOI 10.1109/MIPR54900.2022.00050

professionals is wanted. As for the fine-grained gait mentioned above, a method generating gait motion from intuitive query is needed. This also has not been explored because of the absence of such intuitive query, as the same reason as the description task.

In this paper, to solve both of these problems, we propose an intuitive gait model using Japanese mimetic-words. Mimetic-word is an expression used to verbally express the manner of a phenomenon intuitively, which is also called "ideophone" [3]. It is widely used in several languages in the World, e.g. Japanese, Bengali, Korean, and Tamil. Mimetic-words have an interesting property; sound-symbolism, which indicates that there is an association between linguistic sounds and sensory experiences. The phonemes of a mimetic-word are strongly related to the visual sensation when observing a gait, so the mimetic-words can describe the difference in the appearances of gaits in fine granularity using their rich vocabulary. In English, when we wish to properly express the aspect of gaits, we can use lexical verbs such as *stroll*, *stagger*, and so on. Meanwhile, in Japanese, when we wish to describe the slight difference of gaits, we can use mimetic-words adverbially. In addition, Japanese mimetic-words have a more simple and decomposable structure than those in other languages, such as "*noro-noro*" or "*yoro-yoro*", that is why we use Japanese mimetic-words for the intuitive gait modeling. Based on the decomposable structure and sound-symbolism, we propose the "phonetic-vector" and use it as the basis to represent the model.

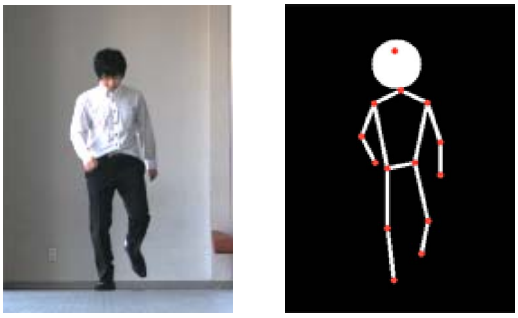
In the Japanese language, there are more than fifty gait-related mimetic-words according to a Japanese mimetic-word dictionary [9]. For example, *noro-noro* describes "slowly walk without having a vigorous intention to move forward," and *yoro-yoro* describes "walk with an unstable balance." Their difference of only one phoneme, i.e. /n/ or /y/, can represent the slight difference in gaits. Such associations are individual-invariant and linguistic-invariant similar to the famous Bouba/kiki-effect [10].

Based on such a background, we have previously proposed a method which describes gait with a mimetic-word [6]. In this paper, additionally, we introduce a method which generates gait from a mimetic-word, and confirm the effectiveness of the proposed intuitive gait model which consists of the phonetic-vector through evaluations of both methods.

2. Related work

There are a few studies on the quantization of a mimetic-word. Tomoto et al. proposed a method visualising impressions evoked by mimetic-words as a thesaurus map based on a handmade quantization table of phonemes [14]. Komatsu proposed a quantization table of impressions evoked from a mimetic-word based on subjective experiments and factor analysis [8]. Sakamoto et al. proposed a system which quantizes a mimetic-word with 35 scales of medical terms to help a medical doctor understanding the nuance of the mimetic-word complained by Japanese patients [11]. These quantization methods mainly aim to visualize or understand a mimetic-word itself. Our study is different from them in that mimetic-words are used to model another modality.

With regard to studies on exploring the relationship between a mimetic-word and other modalities, several studies focus on sound or textural images. Sundaram et al. proposed a semantic word-based similarity metric for clustering audio clips tagged with English onomatopoeias (mimetic-words of sound) [13]. Shimoda et al. proposed a method to recognize tactile images crawled from the Web with a mimetic-word query [12]. Meanwhile, in this paper, we focus on human gaits as visually dynamic states, and attempt to intuitively model human gaits using mimetic-words.



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

Figure 1. Example of body parts detection.

3. Dataset

For both tasks, we used the HOYO dataset[1], which has previously been constructed and made public by ourselves [6]. In this section, we introduce the dataset briefly.

This dataset includes gaits with various mimetic-word annotations. The gaits are viewed from front and back as shown in Figure 1(a). Gaits were recorded by RGB camera in 60 fps and their body-parts coordinates were automatically detected by Convolutional Pose Machine (CPM) [15], and then manually corrected. The number of body-parts is 14 which follows the CPM format as shown in Figure 1(b).

Then, annotators were asked to give the gaits three arbitrary mimetic-words in a free description form. Thirty annotators who are native Japanese University students in

[1] https://www.cs.is.i.nagoya-u.ac.jp/opensource/hoyo/

Table 1. Phoneme statistics of the entire freely described mimetic-words (upper: avg, lower: sd).

1st consonant	ϕ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	1.75	4.68	17.21	28.82	9.57	10.66	0.19	6.63	1.02	0.10	2.98	2.67	7.07	3.12	3.54
	2.07	4.33	15.80	13.23	8.46	9.89	0.63	7.42	2.71	0.46	4.08	3.59	10.39	2.63	2.85
1st vowel	/a/	/i/	/u/	/e/	/o/										
	11.65	2.18	36.61	12.86	36.70										
	6.57	2.45	12.83	7.67	14.33										
2nd consonant	ϕ	/k/	/s/	/t/	/n/	/h/	/m/	/y/	/r/	/w/	/g/	/z/	/d/	/b/	/p/
	15.49	16.17	12.52	20.92	0.59	0.03	0.59	0.16	28.57	0.58	0.02	0.08	0.42	3.71	0.16
	11.72	10.57	9.35	12.64	1.22	0.26	1.19	0.58	24.40	1.41	0.19	0.42	1.23	4.89	0.59
2nd vowel	/a/	/i/	/u/	/e/	/o/	/N/									
	39.02	10.20	13.31	2.87	21.54	13.05									
	13.62	7.46	7.90	3.08	13.15	11.36									

their twenties watched 146 videos showing gaits from the front and annotated each video with three mimetic-words they imagined. Fifteen annotators were assigned to each video and annotated. Here, the mimetic-words were restricted to the pattern of *ABCD-ABCD*, which is the most common pattern of Japanese mimetic-words. Note that *A* and *C* are consonants, and *B* and *D* are vowels.

Finally, 6,322 mimetic-words were collected except for 248 invalid words (e.g. typing error or not in the *ABCD-ABCD* pattern). Table 1 shows the phoneme statistics of the entire freely described mimetic-words.

In addition, we annotated the gaits as a phase information with the time stamp when a pedestrian grounded his/her foot.

4. Gait description and generation

4.1. Phonetic vector

In order to handle mimetic-words corresponding to gaits in a regression model, we have previously proposed expressing them in the form of "phonetic vector" in the description task [6, 7]. But in these papers, an evaluation of the phonetic vector itself was lacking. So in this paper, we also apply this to the generation task and evaluate it through both tasks.

As mentioned in Section 3, in our dataset, each gait sequence can be annotated with multiple mimetic-words. So we use the frequency vector of the appearance of each phoneme composing the mimetic-words corresponding to the gait as phonetic vector $\mathbf{v}$. The vector is composed of 41 dimensions because mimetic-words of the annotation are restricted to the pattern of *ABCD-ABCD*, where *A* and *C* consist of fifteen consonants, *B* consists of five vowels, and *D* consists of six vowels[2]. Let the frequency vector of phonemes *A, B, C*, and *D* be $\mathbf{v}_A$, $\mathbf{v}_B$, $\mathbf{v}_C$, and $\mathbf{v}_D$, respectively, the phonetic vector $\mathbf{v}$ is represented as $(\mathbf{v}_A, \mathbf{v}_B, \mathbf{v}_C, \mathbf{v}_D)$. Note that $\mathbf{v}_A$, $\mathbf{v}_B$, $\mathbf{v}_C$, and $\mathbf{v}_D$ are normalized so that the summation of each element becomes 1.

[2]In the Japanese language, a special phoneme /N/ sometimes appears except in the first phoneme (it is called syllabic nasal). Although, strictly speaking, it is not a vowel, in this paper, we handle it as a vowel for convenience.

4.2. Gait description

In this paper, we use the gait description method we have previously proposed [6, 7]. In this section, we explain it briefly (See [6] for the details).

In this description method, a sequence of arbitrary pairs of body-parts is used as an input feature and the phonetic vector is used as the supervisor of a regressor. After estimating a phonetic vector by the trained regressor, the estimated phonetic vector is converted into a mimetic-word with a nearest-neighbor-like scheme. This is because we aim to reflect human's intuitive impressions in the proposed gait model by explicitly intermediating the phonetic vector whose intuitiveness is guaranteed by sound-symbolism. Note that the input feature has 91 dimensions because the number of arbitrary combinations of body parts is ${}_{14}\mathrm{C}_2 = 91$.

Finally, a regression model learns the relationship of the input feature and phonetic vector $\mathbf{v}$. Let the space constructed by the phonetic vector be named "phonetic space", the procedure can be regarded as estimating the mapping of the kinetic feature space to the phonetic space.

4.3. Gait generation

In this paper, we propose a method for generating gait from a mimetic-word.

In contrast to the description task, the generation task is a conversion from a mimetic-word to a gait. Here, we set the task as follows: A single mimetic-word is input as a query, and a gait motion (time series of body-parts coordinates) is output. In this method, we introduce a two step architecture. The first step is the conversion from a single mimetic-word

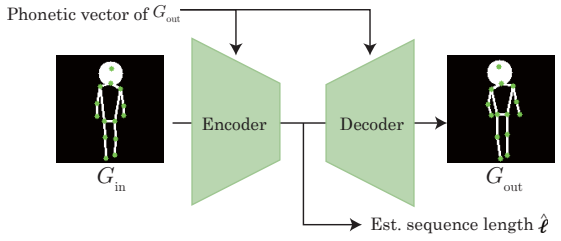


Figure 2. Overview of the module of the second step.

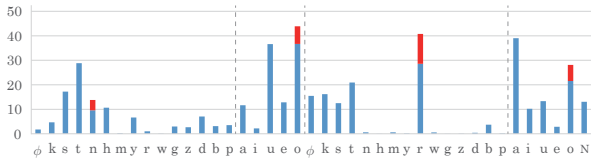


Figure 3. Example of complemented phonetic vector of *noro-noro*.

query to a phonetic vector, and the second is the generation from a phonetic vector to a gait motion.

The module of the second step is trained with the relationship between a gait motion and a phonetic vector. Here, we solve the conversion task as a style transfer task. Style transfer is originally an image conversion task proposed by Gatys et al. [4] They defined "style" as a texture of an image which means space-invariant feature, and "content" as a feature independent from the style. Inspired by this, we define the style of gait as the intuitive impression of gait, which can be assumed time-invariant during a few steps. For this module, an Encoder-Decoder model is used, and the phonetic vector is input as a style query using Adaptive Instance Normalization (AdaIN) [5]. In addition, we choose to train and estimate the fixed-length body parts coordination and the sequence length separately for stabilizing their training. An overview of this architecture is shown in Figure 2.

In this paper, the personality of gait is assumed to be independent from the intuitive impression and regarded as a content. So, in order to make the module learn consistency of the personality, we train the module under the following condition: Both the input gait of the encoder G_{in} and the gait of the supervisor (output of the decoder) G_{out} are of the same person.

As explained in Section 4.1, the phonetic vector is defined based on the frequency of phonemes of which an annotation mimetic-words consists. So, no one can calculate the phonetic vector from a single mimetic-word query as it is. To overcome this problem, the first step converts the single mimetic-word query to a phonetic vector using statistic complementation. Concretely, the complemented phonetic vector $\mathbf{v} = \bar{\mathbf{v}} + a\boldsymbol{\sigma} \circ \mathbf{q}(query)$. Note that $\bar{\mathbf{v}}$ and $\boldsymbol{\sigma}$ are the mean vector and the standard deviation vector of the entire annotation mimetic-words in the dataset, which are the same as the vectors in Table 1. Note that $\mathbf{q}(query)$ is a function converting an *ABCD-ABCD* pattern query into a 4-hot vector (corresponding bins of phonemes are activated), and $\circ$ represents Hadamard product. Here, a is a hyper parameter, and in this paper, we set it to 0.5, experimentally. An example of a complemented phonetic vector when a query *noro-noro* is given, is shown in Figure 3. The blue bars are the mean vector part, and the red bar is the emphasized part.

5. Experiments

5.1. Comparison implementations of phonetic vector

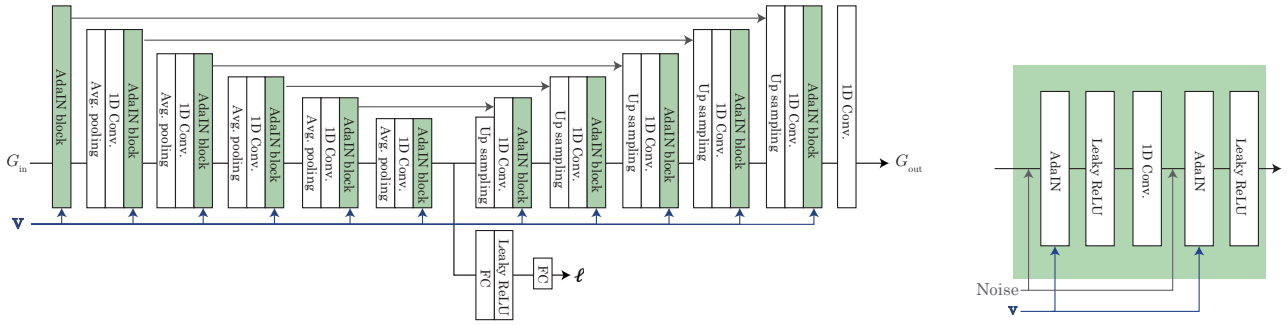
To evaluate the effectiveness of using the proposed phonetic vector, we compared three implementations of a phonetic vector: (1) phoneme-wise vectorization (proposed), (2) mora-wise vectorization, and (3) phoneme-wise vectorization with a conventional quantization table [8].

(1) is the proposed method explained in Section 4.1. It converts an *ABCD-ABCD* pattern mimetic-word into a 41 dimensional vector. (2) is another naïve idea of vectorization, which counts the frequencies of a mora (pair of a consonant and a vowel). It converts an *ABCD-ABCD* pattern mimetic-word into an 151 dimensional vector. (3) is a method that applies a conventional quantization table [8] to (1). It converts an *ABCD-ABCD* pattern mimetic-word into a 16 dimensional vector.

5.2. Implementation

In the description experiment, we subsampled the original gait sequences into that of 100 frames long and used them for training and estimating the regressor.

In the generation experiment, we subsampled the original gait sequences into that of one cycle length referring to the phase annotation explained in Section 3. Though there could be two types of subsamples which starts from the right foot or from the left foot, in this experiment, we used that from the right foot as training samples. For the training, the subsamples are stretched into a sequence of 128 frames long using smoothing spline, and the stretched subsample and its original frame length are input separately. For the generation, smoothing spline is applied to the generator outputs (the fixed length gait and the estimated frame



(a) Whole architecture.

(b) Detail of the AdaIN [5] block architecture.

Figure 4. Generation network architecture.

Table 2. Results of phonetic vector evaluations on the description task

Phonetic vector	Subjective score	MAE
(1) Phoneme-wise (Proposed)	4.751 ± 0.073	0.396
(2) Mora-wise	4.645 ± 0.064	0.750
(3) Phoneme-wise (With [8])	3.596 ± 0.072	0.367

Table 3. Results of phonetic vector evaluations on the generation task

Phonetic vector	Subjective score	MAE (coordinates)	MAE (length)
(1) Phoneme-wise (Proposed)	4.396 ± 0.023	0.539 (0.916 m)	11.27 frames
(2) Mora-wise	4.337 ± 0.037	1.040 (1.767 m)	22.88 frames
(3) Phoneme-wise (With [8])	3.428 ± 0.048	0.479 (0.814 m)	11.19 frames

length). Details of the implementation of the generation network architecture is shown in Figure 4.

Both architectures are implemented using Keras [1] and their parameters are decided experimentally.

5.3. Evaluation of phonetic vector implementation

For the description task, we evaluate the model with both subjective and objective criteria, in a leave-one-person-out cross validation manner based on the person ID in the HOYO dataset [6]. The former criterion is the average evaluation score of a subjective experiment; We presented a pair of a gait video and a generated mimetic-word to evaluators, and asked them how well the generated mimetic-word described the gait from seven levels of Likert scale. The evaluators were sixteen native Japanese University students, and five evaluators were assigned per question. The latter criterion is the estimation error of phonetic vectors with Mean Absolute Error (MAE). Note that the value of estimation error is regularized with the average norm of their ground-truth (4 for the implementation (1), 2 for (2), and 0.00793 for (3)) because the dimension of the phonetic vector differs per implementation. Experimental results of the description task is shown in Table 2.

For the generation task, we also evaluate the model with both subjective and objective criteria. The former criterion is the average evaluation score of a subjective experiment; We presented a pair of a query mimetic-word and a generated gait sequence which is visualized as in Figure 1(b) to evaluators, and asked them how well the mimetic-word described the generated gait from seven levels of Likert scale. The evaluators were 22 native Japanese University students. The number of the questions was 27, and the query mimetic-words were selected from words familiar to native Japanese speaker. For the latter criterion, gait generation from untrained phonetic vectors in the test set is evaluated with MAE of gait coordinates and sequence length. As the untrained data, we used flipped gaits of the subsamples starting from the left foot (because those starting from the right foot were all used for training, as explained in Section 5.2). Experimental results of the generation task is shown in Table 3. Note that the values in brackets of the coordinate column are conversion values under the following condition: the generated pedestrian is 1.7 m height.

Results in both Tables 2 and 3 show that the proposed method (1) achieved the better performance than (2) in terms of MAE, and (3) in terms of subjective score. From these results, we confirmed that the implementation of the proposed phonetic vector is a better model than the comparative implementations.

5.4. Evaluation of gait generation method

We evaluate the generated gaits through a more detail subjective experiment. In concrete, a case which

Table 4. Evaluation of gait generation.

	(1): Familiar	(2): Unfamiliar
Positive-pair	4.396	3.273
Negative-pair	4.047	3.202

shows pairs of generated gait and its query mimetic-word (positive-pair) is compared with a case which shows pairs of generated gait and an unrelated mimetic-word (negative-pair). We evaluate them under the following two presented word conditions: (1) the presented mimetic-words are familiar to a native Japanese speaker. (2) the presented mimetic-words are unfamiliar (generated from an arbitrary combination of phonemes). Note that these conditions only control the presented word, but not the query of the generator. For both conditions, 27 positive-pairs and 27 negative-pairs were prepared, and randomly shuffled, resulting in 54 questions for evaluation. The evaluators were 22 native Japanese University students.

The results are shown in Table 4. Although condition (1) got a significant deference ($p < 0.002$) between the results of positive and negative -pairs, condition (2) was not significant. These results imply that the proposed method can generate reasonable gait considering the query mimetic-word, but the ability of generating from unfamiliar mimetic-words is insufficient.

6. Conclusion

In this paper, we proposed a method which generates gait from a mimetic-word. Through evaluations of both the generation task and the description task [6], we confirmed the effectiveness of the proposed intuitive gait model based on the proposed phonetic-vector. The main limitation of the generation method is the lack of the ability of generation from unfamiliar mimetic-words. Furthermore, the generation method should be improved overall because it got less than five points out of seven also for the generation from familiar mimetic-words.

Future work includes using 3D motion, considering impression bias caused by pedestrian's appearance attributes, and designing an explicit one-to-many gait generation architecture which can explicitly generate various gaits from a single mimetic-word.

Acknowledgements

Parts of this work were supported by Grant-in-Aid for Scientific Research (22H03612).

References

[1] Keras documentation. `https://keras.io/`.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[3] C. M. Doke. *Bantu Linguistic Terminology*. Longmans, Green, London, UK, 1935.

[4] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[5] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. 16th IEEE Int. Conf. on Computer Vision*, pages 1501–1510, 2017.

[6] H. Kato, T. Hirayama, I. Ide, K. Doman, Y. Kawanishi, D. Deguchi, and H. Murase. More-natural mimetic words generation for fine-grained gait description. In *Proc. 26th Int. Conf. on Multimedia Modeling*, volume 2, pages 214–225, 2020.

[7] H. Kato, T. Hirayama, Y. Kawanishi, K. Doman, I. Ide, D. Deguchi, and H. Murase. Toward describing human gaits by onomatopoeias. In *Proc. 2017 IEEE Int. Conf. on Computer Vision Workshops*, pages 1573–1580, 2017.

[8] T. Komatsu. Quantifying Japanese onomatopoeias: Toward augmenting creative activities with onomatopoeias. In *Proc. 3rd ACM Int. Conf. on Augmented Humans*, number 15, pages 1–4, 2012.

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

[10] V. S. Ramachandran and E. M. Hubbard. Synaesthesia — A window into perception, thought and language. *J. Conscious. Stud.*, 8(12):3–34, 2001.

[11] M. Sakamoto, Y. Ueda, R. Doizaki, and Y. Shimizu. Communication support system between Japanese patients and foreign doctors using onomatopoeia to express pain symptoms. *J. Adv. Comput. Intell. Intell. Inform.*, 18(6):1020–1025, 2014.

[12] 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.

[13] 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.

[14] 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.

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