Annotation Generation From IMU-Based Human Whole-Body Motions in Daily Life Behavior

Wataru Takano, *Member, IEEE*

***Abstract*—This article describes a stochastic framework for integrating human whole-body motions with natural language. Human whole-body motions in daily life are measured by inertial measurement units (IMU) and subsequently encoded into motion primitives. Sentences are manually attached to the human motion primitives for their descriptions. Two aspects of semantics and syntactics are represented by stochastic modules. One stochastic module trains the linking of motion primitives to words, and the other module represents word order in the sentence structure. These two modules are helpful toward converting human whole-body motions into descriptions, where multiple words are generated from the human motions by the first module, and the second module searches for syntactically consistent sentences consisting of the generated words. The proposed framework is tested on a large dataset of human whole-body motions and their descriptive sentences. The linking of human motions to natural language enables robots to understand observations of human behavior as sentences.**

***Index Terms*—Human motion, language, probabilistic modeling.**

I. Introduction

HUMANS are animals that create structure. To understand the things we encounter in the world, we categorize, classify, and arrange them. We use languages of words and symbols to do so, and it would be no exaggeration to say that these words and symbols support humanity's highly advanced society and culture.

The function of language is to put the world in order, to maintain that order, and, from that, to create a new order. Our society is complex, but we can view it in structural terms by understanding it as a combination of symbols. When the world is organized as low-information symbols, it can be conveyed through conversational language, or recorded for future generations as characters. New combinations of words can furthermore lead to the creation of new ideas and social order. The results of our history, where we understand social order, maintain it, and recreate a new order, become the advanced knowledge systems of humanity.

The path toward incorporating knowledge systems into robots lies in transforming the world into language. Robots that will coexist with us in our daily lives have a particular need for representing human behavior as language. Extensive research has investigated how to describe the bodily movement of humanoid robots through numerical modeling. A typical approach is imitation learning, transforming human bodily movements into robotic movements that are learned and recorded. Taking the continuous information of bodily motion as a parameter set for a mathematical model, expressed as a discrete representation, provides movement primitives, which can be regarded as the first step toward symbolization. Such movement primitives are used to plan, produce, and control full-body robot motion in accordance with the mathematical model. They can be used not only in techniques for creating human-like movements, but also in motion recognition by comparing motion primitives with observed motions.

Language has semantic (meaning) and syntactic (rules) aspects. Motion primitives successfully express relations between the signified and the signifier by attaching mathematical models to actually moving bodies, but they do not take into account the syntactic processing of language. By combining meaning with syntactic processing, we can raise language parsing from the word level to the textual level. Thus, by parameterizing intertextual word orders for grammatical rules and unifying them with motion primitives, we can advance the study of transforming movement into language.

Our daily lives are filled with various movements and linguistic representations. The robots that will permeate our society must overcome this diversity by developing intelligence that understands various movements as language. One key to overcoming this diversity lies in enormous datasets related to human behavior and linguistic expressions. This article uses movement data from everyday life, captured via wearable motion sensor suits, along with a collection of texts describing these data, to create such an enormous dataset. Using a mathematical model that statistically learns movements and texts, a system is constructed that translates a wide variety of movements into texts, and the system's effectiveness is demonstrated through experiments for quantitative evaluations.

II. Related Work

Mathematical modeling of human or robotic motions, typified by imitative learning and programming by demonstration, is a potential approach toward constructing artificial intelligence based on motion data. There are two popular approaches to motion modeling: dynamical systems [1]–[4] and stochastic

Manuscript received June 9, 2019; revised October 11, 2019; accepted December 12, 2019. Date of publication January 9, 2020; date of current version January 14, 2020. This work was supported by a Grant-in-Aid for Challenging Research (Exploratory) from the Japan Society for the Promotion of Science under Grant 17K20000. This article was recommended by Associate Editor X. Hu. *(Corresponding author: Wataru Takano.)*

The author is with the Center for Mathematical Modeling and Data Science, Toyonaka 560-8531, Japan (e-mail: takano@sigmath.es.osaka-u.ac.jp).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2019.2960630

systems [5]–[9]. Dynamical systems represent motions as differential equations in the state space, while stochastic systems represent motion transitions and distributions. Both encode motions into model parameters. This encoding implies that continuous motion data are discretely represented as points in the parameter space, and that points can be defined as motion symbols, which are helpful as motion recognizers and motion synthesizers. Motion recognizers classify observations of human motion into the motion symbol most similar to the observation, and motion synthesizers generate robot motions similar to training motions according to dynamics embedded in the mathematical model. However, it is not easy for humans to use motion recognizers or synthesizers, because mathematical models do not provide an intuitive interface; humans cannot understand motion symbols from model parameters alone. It is, thus, important to connect motion symbols with natural language to improve ease of use [10], [11].

Several studies on connecting motions with language have been conducted in the fields of robotics and computer graphics. Rose *et al.* proposed the concepts of "verbs" and "adverbs" in the motion space in which groups of similar motions are defined as verbs, and differences between motions within the same group are defined as adverbs [12]. Adverbs are parameterized to control interpolation between motions, and consequently synthesize new motions with consistency between verbs and adverbs. Arikan *et al.* presented a method for synthesizing character motions from word queries [13]. A database of motions with verb labels attached to frames in the motions is constructed. This method uses dynamic programming to search for continuous motion sequences while satisfying constraint conditions on input verb labels attached to these motions. We were inspired by computation in machine translation, and applied the statistical translation model to motion sequences and their relevant verb labels [14], [15]. This model allows conversion of motions to verb labels and vice versa. Nakamura *et al.* presented a method for categorizing physical information from multiple sources, including robot-mounted visual, audio, and tactile sensors [16]. Categories of multimodal data are connected to words in the statistical model [17]. These approaches succeeded in linking real-world physical data to words, but have not reached the level of sentences, in which the words are put together in accordance with grammatical rules.

Sugita and Tani presented a novel approach to integrating motions and language, where two neural networks for motions and sentences share parameters [18]. This allows synthesizing motions from sentences. Ogata *et al.* extended this framework to generate sentences describing motions [19]. Multiple sentences are created from a motion, and each of these sentences is subsequently converted to a motion in the same manner as in Sugita and Tani. An appropriate sentence is selected by comparing the generated motions with the original motion. This article has been expanded to accommodate a variety of motions and sentences [20], [21]. We have also developed a framework for integrating human or robotic whole-body motions with natural language by the statistical method [22]–[24]. This framework consists of two statistical modules: one module trains associations between motions and words, and the other trains arrangements of words in sentences. This allows for bidirectional mapping between motions and their relevant sentences. This mapping was validated by an experiment on a small dataset of human motions and attached sentences. Additionally, this framework was extended to handle manipulated objects to improve the generation of correct and detailed sentences from human behaviors involving whole-body motion and object data [25], [26].

In recent years, deep learning techniques [27] have outperformed state-of-the-art machine learning frameworks in computer vision [28] and natural language processing [29], and have gradually come to be applied to problems in robotics [30], [31]. Plappert *et al.* applied deep learning techniques to link human whole-body motions to natural language [32]. In their method, one recurrent neural network computes context from motion, and another receives the context and previous word to predict the following word. Iteration of this process synthesizes a word sequence providing a description of the motion. Ahn *et al.* used the deep learning techniques of generative adversarial networks to establish a motion synthesizer from language [33]. This framework consists of a generator and a discriminator based on recurrent neural networks. The generator encodes a sentence into a text feature, from which it generates a human motion. The discriminator also encodes a sentence into a text feature and differentiates actual motions from generated motions from consideration of text features. The generator and discriminator mutually develop in a way such that the generator creates a realistic human motion and the discriminator differentiates between realistic and generated motions. Yamada *et al.* proposed a deep learning technique for bidirectional translation between actions of motions and visual perception on the one hand and sentences on the other hand [34]. A pair of recurrent neural networks encodes actions into action features and decodes actions from those features. Another pair of recurrent neural networks encodes sentences into text features, and decodes sentences from those features. These recurrent networks separately train action and sentence data, and are additionally modulated to minimize error between these two features. This modulation matches actions to language. These deep learning techniques do not convert each motion data into multiple sentences, but into a single relevant sentence. This article proposes a probabilistic framework to generate multiple sentences with probabilities of their describing the motion. Additionally, each motion is represented by a compact module of a hidden Markov model (HMM). Since the motion modules are mutually independent, a module for the new motion can be easily put into this framework without retuning the parameters of other modules. In the deep learning frameworks, all the motion patterns are embedded in a large-sized neural network. When a new training motion pattern is given, all the parameters of the neural network need to be modulated.

III. Synthesis of Sentences From Motions

A. Motion Representation

Measurement techniques for human whole-body motions, as typified by optical motion-capture systems or inertial measurement unit (IMU) sensors, have been developed. Although optical motion capture systems can accurately measure the location

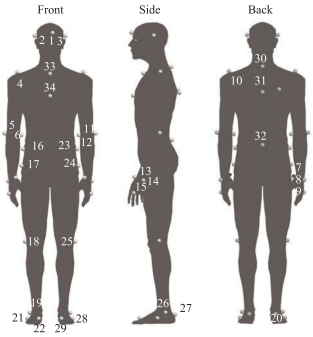

Fig. 1. In total, 34 virtual markers are attached to a human performer. IMU sensor data are converted to positions in a trunk coordinate system.

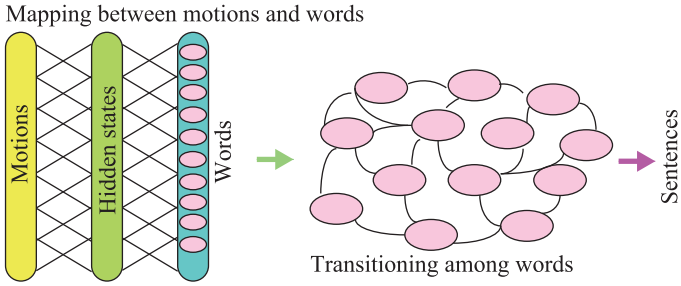

Fig. 2. Overview of mapping from motion to language. Motions are connected to words via latent states in a stochastic model. Transitions among words are also represented in a stochastic model. Motion inputs are converted to multiple words, based on knowledge of relations between motions and words. Sentences are subsequently created from those words according to knowledge of word transitions.

of multiple markers attached to a performer, the measurement space is limited to a controlled laboratory with installed capture cameras, making it difficult to record human motions under a wide variety of situations. In contrast, IMU sensors attached to a performer measure linear acceleration, angular velocity, and magnetic fields. These sensor outputs are converted to the posture of a human character through forward and inverse kinematic computations, thereby estimating whole-body joint angles and positions. IMU sensors do not limit measurement range, and are superior to optical motion-capture systems in terms of measurement area. We, therefore, adopt IMU sensors to measure human whole-body motions to create a large dataset of human motions.

In total, 17 IMU sensors are attached to a performer. The resulting sensor data are converted to locations of 34 virtual markers attached to the performer, as shown in Fig. 1. Human whole-body posture is represented by a feature vector whose elements are positions of the virtual markers in a trunk coordinate system. Human whole-body motion is consequently represented as a sequence of feature vectors. The sequence is encoded into a set of parameters for a HMM, which is referred to as a motion symbol. Observations of human whole-body motion are classified into the motion symbols most likely to generate those observations. In this way, the motion symbols function as motion recognizers.

B. Mapping Between Motions and Words

We describe a statistical framework for conversion from human whole-body motions to descriptive sentences. As shown in Fig. 2, our framework uses two modules, one for mapping between human motions and words and one for word arrangement.

Mappings between human motions and words are represented by a probabilistic graphical model, where nodes in the first layer are motion symbols, nodes in the third layer are words, and hidden states in the second layer connect motion symbols with words. Fig. 3 shows the mapping architecture. The connection is parameterized by the probability $P(s|\lambda)$ of hidden state s being generated by motion symbol λ and the probability $P(\omega|s)$ of word ω being generated by hidden state s. The optimal parameters can be computed by an EM algorithm that maximizes the probability $P(\boldsymbol{\omega}|\lambda)$ of a set of words included in sentence $\boldsymbol{\omega}$

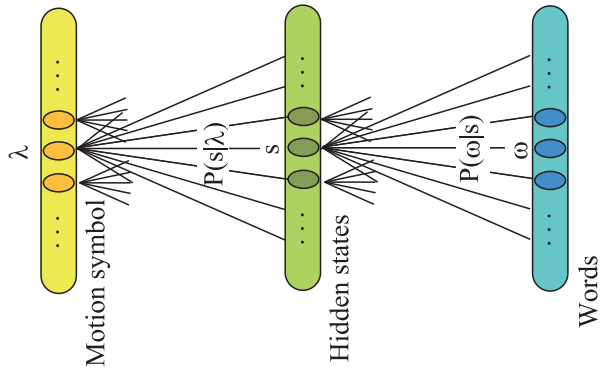

Fig. 3. Mappings between motions and words are represented by a probabilistic graphical model. Motions are connected to words via a hidden state by the probability of the hidden state being generated by the motion and the probability of the word being generated by the hidden state.

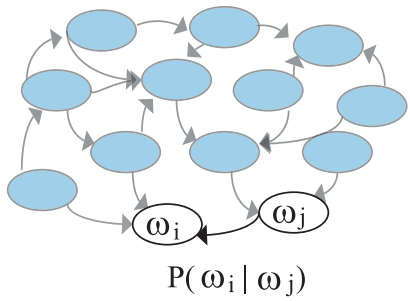

Fig. 4. Sentence structures are represented by transitions among words, parameterized according to the word N-gram.

being generated by motion symbol λ

$$\mathcal{P} = \sum_k \log P(\boldsymbol{\omega}^{(k)}|\lambda^{(k)}) \tag{1}$$

where the kth human motion datum is classified as motion symbol $\lambda^{(k)}$, and sentence $\boldsymbol{\omega}^{(k)}$ is manually attached to this datum. Since sentence $\boldsymbol{\omega}^{(*)}$ is a sequence of words with length m as

$$\boldsymbol{\omega}^{(*)} = \left(\omega_1^{(*)}, \omega_2^{(*)}, \ldots, \omega_m^{(*)}\right) \tag{2}$$

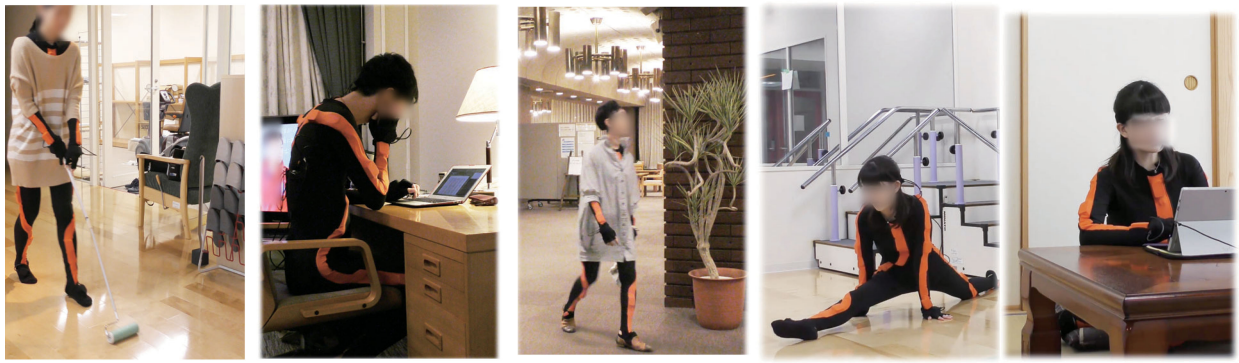

Fig. 5. IMU sensors are attached to performers, whose whole-body motions in daily life are recorded.

the probability $P(\boldsymbol{\omega}^{(*)}|\lambda^{(*)})$ is calculated as

$$\log P(\boldsymbol{\omega}^{(*)}|\lambda^{(*)}) = \sum_{i=1}^{m} \log P(\omega_i^{(*)}|\lambda^{(*)}). \tag{3}$$

The EM algorithm alternates between an E-step and an M-step. The E-step estimates the conditional probability distribution $P(s|\lambda, \omega)$ of hidden state s given motion symbol λ and word ω as

$$P(s|\lambda, \omega) = \frac{P(\omega|s)P(s|\lambda)}{\sum_s P(\omega|s)P(s|\lambda)}. \tag{4}$$

The M-step optimizes probability parameters $P(s|\lambda)$ and $P(\omega|s)$ using the estimated $P(s|\lambda, \omega)$ derived in the E-step, as

$$P(s|\lambda) = \frac{\sum_\omega n(\lambda, \omega)P(s|\lambda, \omega)}{\sum_{\omega,s} n(\lambda, \omega)P(s|\lambda, \omega)} \tag{5}$$

$$P(\omega|s) = \frac{\sum_\lambda n(\lambda, \omega)P(s|\lambda, \omega)}{\sum_{\lambda,\omega} n(\lambda, \omega)P(s|\lambda, \omega)} \tag{6}$$

where $n(\lambda, \omega)$ is the number of correspondences between motion symbol λ and word ω as observed in the training dataset. Derivations of the E-step and M-step are described in detail in [24].

C. Learning of Word Sequences

Sentences are word sequences, and sentence structures are assumed to be extracted by transitions among words. Sentences are also represented by a probabilistic graphical model in the same manner as are mappings between motions and words. Fig. 4 shows this graphical model, where nodes and edges denote words and transitions among words, respectively. In the word N-gram model, the graphical model is parameterized by the interword relation probability $P(\omega|\boldsymbol{\omega}_{1:N-1})$ of transitioning from a sequence of $N-1$ words, $\boldsymbol{\omega}_{1:N-1}$ to word ω, and an initial node probability $P(\omega)$ of starting at ω. The optimal probability $P(\omega|\boldsymbol{\omega}_{1:N-1})$ can be computed such that the probability $P(\boldsymbol{\omega})$ of training sentence $\boldsymbol{\omega}$ being generated is maximized. The objective function $\mathcal{Q}$ is

$$\mathcal{Q} = \sum_k \log P(\boldsymbol{\omega}^{(k)}) \tag{7}$$

$$\begin{aligned}\log P(\boldsymbol{\omega}^{(*)}) = {} & \log P(\omega_1^{(*)}) + \log P(\omega_2^{(*)}|\omega_1^{(*)}) \\ & + \cdots + \log P(\omega_{N-1}^{(*)}|\boldsymbol{\omega}_{1:N-2}^{(*)}) \\ & + \sum_{i=N}^{m} \log P(\omega_i^{(*)}|\boldsymbol{\omega}_{i-N+1:i-1}^{*})\end{aligned} \tag{8}$$

where $\boldsymbol{\omega}_{j:k}^{(*)}$ is a sequence of words from the jth to the kth position in sentence $\boldsymbol{\omega}^{(*)}$. The optimal probability $P(\omega|\boldsymbol{\omega}_{1:N-1})$ is derived as

$$P(\omega|\boldsymbol{\omega}_{1:N-1}) = \frac{n(\boldsymbol{\omega}_{1:N-1}, \omega)}{n(\boldsymbol{\omega}_{1:N-1})} \tag{9}$$

where $n(\boldsymbol{\omega}_{1:N-1})$ is the number of observations of word sequence $\boldsymbol{\omega}_{1:N-1}$ among the training sentences, and $n(\boldsymbol{\omega}_{1:N-1}, \omega)$ is the number of observations of word ω following that word sequence.

D. Conversion From Motion to Sentences

Human whole-body motion is converted to a descriptive sentence by integrating two probabilistic graphical models, one for mappings between human motions and words and one for word arrangement. The conversion is formulated as a search for the most probable sequence of words for the human whole-body motion. Specifically, the probability $P(\boldsymbol{\omega}|\boldsymbol{x})$ of word sequence $\boldsymbol{\omega}$ being generated from human whole-body motion $\boldsymbol{x}$ is rewritten as

$$P(\boldsymbol{\omega}|\boldsymbol{x}) = \sum_\lambda P(\boldsymbol{\omega}|\lambda)P(\lambda|\boldsymbol{x}) \tag{10}$$

$$= \sum_\lambda P(\boldsymbol{\omega}|\lambda)\frac{P(\boldsymbol{x}|\lambda)P(\lambda)}{P(\boldsymbol{x})}. \tag{11}$$

Note that the prior probability of $P(\boldsymbol{x})$ has no effect on searching for the sentence. The assumption that prior probabilities of $P(\lambda)$ are equivalent for all motion symbols leads to searching for the

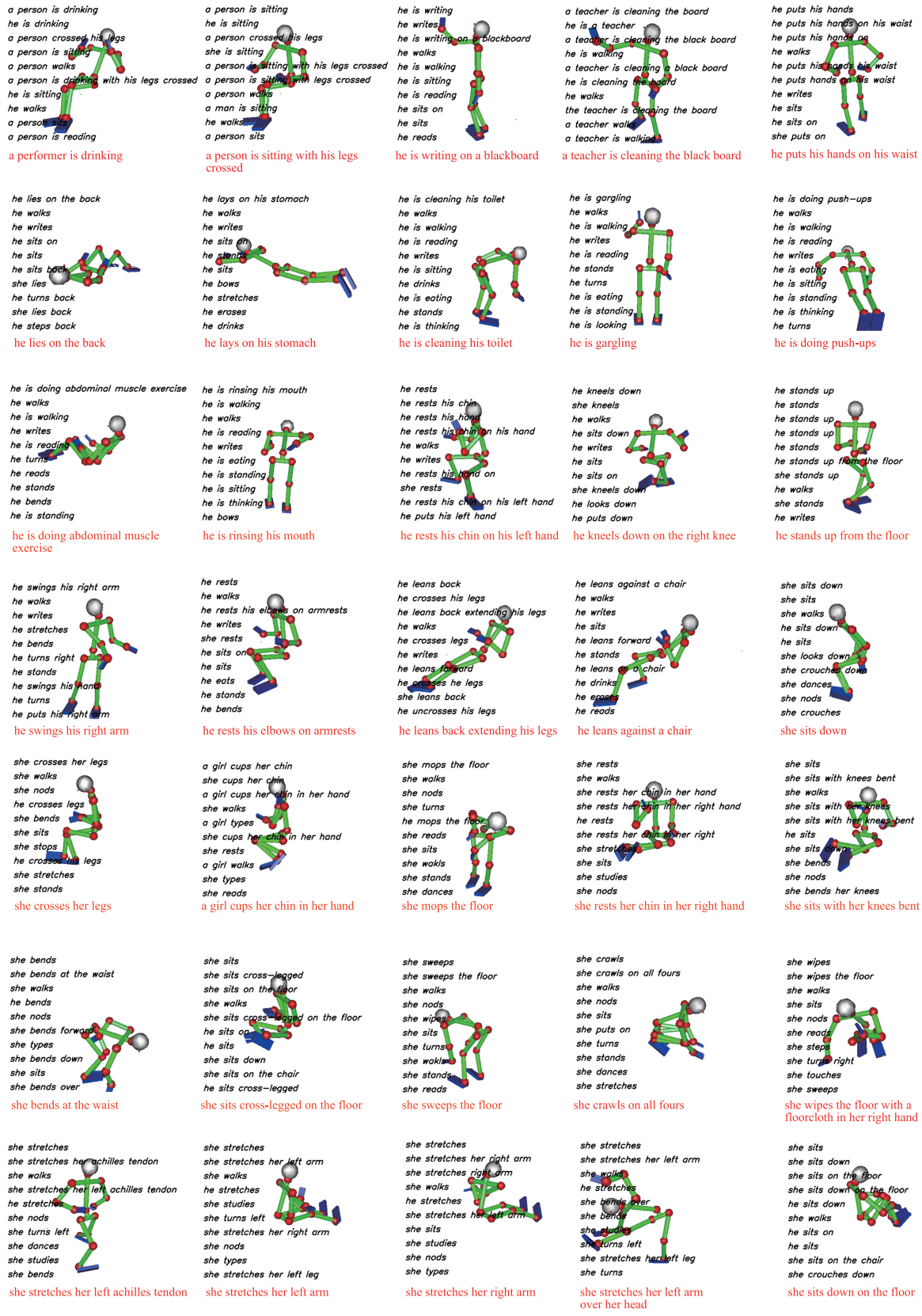

Fig. 6. Ten most likely sentences are synthesized from human whole-body motions. The synthesized sentences are displayed in black, and correct sentences are displayed in red.

TABLE I
BLEU SCORES FOR SENTENCE GENERATION

	Rank									
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
$\mathcal{H}_{40}$	0.342	0.333	0.329	0.333	0.348	0.358	0.357	0.350	0.350	0.346
$\mathcal{H}_{400}$	0.596	0.540	0.442	0.403	0.377	0.368	0.360	0.350	0.347	0.356
$\mathcal{H}_{4000}$	0.604	0.501	0.409	0.380	0.365	0.362	0.353	0.344	0.340	0.333

$\mathcal{H}_{40}$, $\mathcal{H}_{400}$, and $\mathcal{H}_{4000}$ denote mappings between motions and sentences, with the number of hidden states being set to 40, 400, and 4000, respectively.

sentence with the highest probability, as

$$\arg\max_{\boldsymbol{\omega}} P(\boldsymbol{\omega}|\boldsymbol{x}) = \arg\max_{\boldsymbol{\omega}} \sum_{\lambda} P(\boldsymbol{\omega}|\lambda) P(\boldsymbol{x}|\lambda) \tag{12}$$

$$= \arg\max_{\boldsymbol{\omega}} \sum_{\lambda} P(\boldsymbol{\omega}|\omega_1, \omega_2, \ldots, \omega_m) \times P(\omega_1, \omega_2, \ldots, \omega_m|\lambda) \times P(\boldsymbol{x}|\lambda). \tag{13}$$

The search decomposes into three probabilities: the probability of sentence $\boldsymbol{\omega}$ being created from words $\{\omega_1, \omega_2, \ldots, \omega_m\}$ by the graphical model of the word arrangement in (8); the probability of words $\{\omega_1, \omega_2, \ldots, \omega_m\}$ being generated from motion symbol λ by the graphical model mapping between human motions and words in (3); and the probability of whole-body motion $\boldsymbol{x}$ being generated by motion symbol λ. Equation (13) can be efficiently solved by Dijkstra's algorithm.

IV. EXPERIMENTS

We tested the proposed approach for synthesizing sentences to describe human behavior on a dataset of human whole-body motions with descriptions attached to those motions. Human whole-body motions were measured using commercial IMU sensors (Xsens Technologies), shown in Fig. 5. In total, 17 IMU sensors with a sampling rate of 120 Hz were attached to a performer output estimating their positions from measurements of linear acceleration and angular velocity. Position data were retargeted to a human character with 34 degrees of freedom, and the positions of virtual markers fixed on the human character were computed by inverse kinematics and forward kinematics computations. The virtual markers were arranged over the character's whole body according to Helen Hayes marker placement [35]. IMU measurements were converted to a posture feature vector whose elements are locations of virtual markers in the trunk coordinate system, and human whole-body motions were expressed as sequences of posture features. We recorded human whole-body motions of five male and eight female performers. The recorded motion data consisted of 84 868 680 frames, equivalent to 707 239 s.

Human whole-body motions were manually annotated. A total of 62 207 sentences with 3349 different words were collected for the motion descriptions. Additionally, motion segments described by the sentences were collected by manually detecting their boundaries along the human whole-body motion data. Each motion segment was encoded into HMM parameters for a motion symbol. This implies that the number of motion symbols was the same as the number of sentences. This manual annotation and segmentation created a large dataset of pairs of motion symbols and descriptive sentences.

We set the number of hidden states in the probabilistic graphical model for mapping between motions and words to 4000. This model trained this mapping from the dataset of motions and descriptive sentences, as described above. The probabilistic graphical model for sentence structure also trained word 4-grams in the same training sentences. We subsequently tested sentence generation from the human whole-body motions by integrating these two models. We computed the probabilities of human whole-body motions being generated by motion symbols, and searched for the most probable sentences in consideration of the probabilities of words being generated from the motion symbols and probabilities of transitions among the words. Fig. 6 qualitatively shows several examples of sentence generation. We computed multiple sentences, sorted them in descending order of probability, and displayed the ten sentences with the highest probability. As these examples show, correct or nearly correct sentences were generated to describe the human motions with consistency in mappings between motions and words on the one hand and grammatical structure on the other. Short sentences are more likely to be highly ranked; longer sentences comprise more words, creating more probabilities for generating a word from a motion and for transitioning among words, so the resultant probability is lower. Complicated sentences, thus, tended to have middle ranks.

We varied the number of hidden states in the probabilistic graphical model for mapping between motions and words, and quantitatively evaluated the generation of sentences from human whole-body motions. The number of hidden states was set to 40, 400, and 4000. As the evaluation index, we chose the bilingual evaluation understudy (BLEU) score, a common measure for machine translation performance [36]. BLEU scores are formulated based on word N-gram matching between a reference sentence and a generated sentence as

$$\text{BLEU} = \text{bp} \exp \sum_{n=1}^{N} \frac{\log p_n}{N} \tag{14}$$

$$\text{bp} = \min \left\{ 1, \exp \left(1 - \frac{l_r}{l_g} \right) \right\} \tag{15}$$

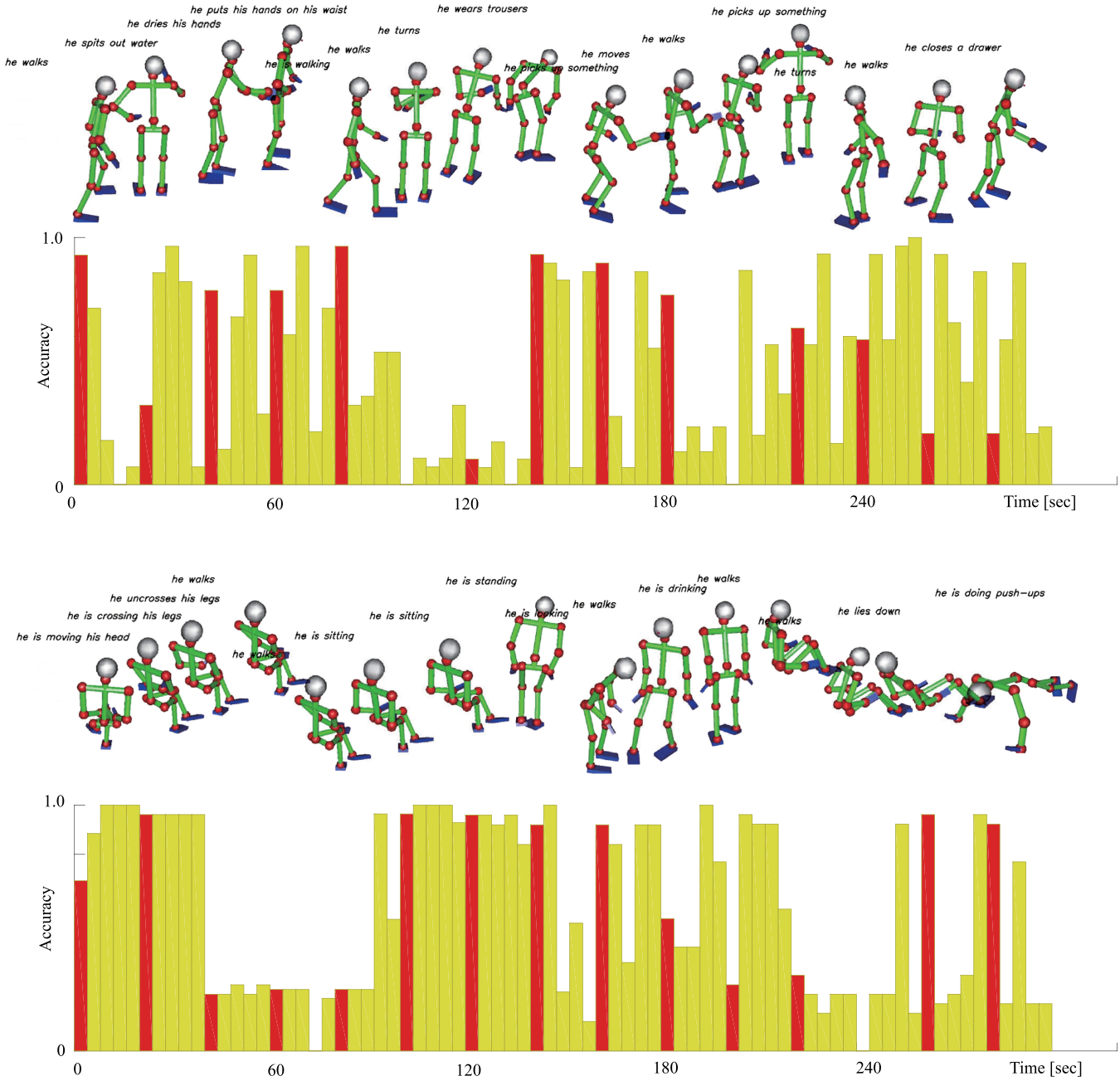

Fig. 7. Accuracy of sentence generation for human whole-body motion. Generated sentences are manually judged as correct or incorrect through a crowdsourcing framework. Accuracy rates for human motion in a snapshot are displayed in red.

TABLE II
RESULTS FOR MANUAL EVALUATIONS OF SENTENCE GENERATION

Number of label "true"	254,116
Number of label "false"	185,846
Accuracy rate	0.578

where p_n is the ratio of the number of word n-grams in a generated sentence matching a word n-gram in its reference sentence, bp is a brevity penalty, and l_r and l_g are the lengths of the reference and generated sentences, respectively. BLEU scores range from 0 to 1, with higher scores indicating closer similarity between the reference and generated sentences. In this article, N is set to 4. Table I shows average BLEU scores for cases where the number of hidden states is set to 40, 400, and 4000. The model with 400 hidden states outperformed the model with 40 hidden states in terms of sentence generation in a range from first to sixth ranks. The model with 4000 hidden states achieved an average BLEU score of 0.604, outperforming the model with 400 hidden states (BLEU score 0.596) in generating first-rank sentences. The first-rank sentence is the most important as it is the most likely to be generated from a motion. However, the model with 400 hidden states achieved better performance in

generating second- to tenth-rank sentences than did the model with 4000 hidden states.

In addition, we used crowdsourcing to perform evaluations of sentence generation with 4000 hidden states in the mapping between motions and sentences. A 3.0 s sliding window for motion segments was converted to relevant sentences regarding human whole-body motions. Note that motion segments for training were partially clipped among human whole-body motions, and that most of the sliding windows were not included in the training dataset. We created videos containing character motions and their respective generated sentences, and uploaded them to YouTube. We asked Yahoo users in a crowdsourcing framework to watch these YouTube videos, and to provide a "true" label when sentences were judged as correct or a "false" label otherwise. We effectively collected many answers, with 254 116 labeled true and 185 846 labeled false. Table II shows statistics for the collected answers. Our proposed method, thus, achieved an accuracy rate of 0.578 for sentence generation, according to the subjective assessment. Fig. 7 shows a temporal profile of accuracy rates and human whole-body motions. In the top panel, a walking motion was observed at 4 s, and this motion was converted to the sentence "he walks." In total, 26 true labels and 2 false labels were attached to this motion–sentence pair, a high level of "correct" judgments for the generated sentence. A gargling motion followed, and was converted to the sentence "he spits out water." In total, 9 true labels and 19 false labels were attached to this sentence generation. The "spitting out water" motion was observed around the same time as the "gargling" motion. A sliding window for the motion partially included a "spitting out water" motion, which was translated as "he spits out water." In the bottom panel in Fig. 7, the motions "crossing legs," "sitting on a chair," "standing," "looking down," "lying on the floor," and "doing push-ups" were correctly translated into the sentences "he is crossing his legs," "he is sitting," "he is standing," "he is looking," "he lies down," and "he is doing push-ups" at 20, 100, 120, 140, 160, 260, and 280 s. These generated sentences scored high accuracy rates above 0.8 according to the subjective assessment performed via crowdsourcing.

V. Conclusion

The results of this article are as follows.

1) We used IMU sensors to record many human whole-body daily life motions. By annotating these human motions, we created a large dataset of human whole-body motions matched with descriptive sentences. In total, we created 84 868 680 frames of human whole-body motions and attached 62 207 sentences to these data.
2) We created a framework for matching human whole body motions with descriptive sentences. Mappings between human motions and relevant words, and transitioning among words in sentences, are represented by probabilistic graphical models. Linking between these two models allows human whole-body motions to be translated into descriptive sentences.
3) The proposed framework was optimized from a large training dataset of human whole-body motions and relevant sentences. We conducted an experiment on generating sentences from human whole-body motions. Our method computes probabilities of observed human motion being generated by motion symbols, probabilities of words being associated with those motion symbols, and probabilities of word transitions. We could, therefore, compose reasonable sentences with the highest resultant probability. Sentence generation was evaluated through two assessments: BLEU scores measuring matching rate between generated and reference sentences, and subjective judgments in a crowdsourcing framework. The established dataset proved to be useful, verifying the validity of the proposed method for integrating human motions with sentences to describe human whole-body motions.

References

[1] M. Okada, K. Tatani, and Y. Nakamura, "Polynomial design of the nonlinear dynamics for the brain-like information processing of whole body motion," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2002, pp. 1410–1415.
[2] A. J. Ijspeert, J. Nakanishi, and S. Shaal, "Learning control policies for movement imitation and movement recognition," *Neural Inf. Process. Syst.*, vol. 15, pp. 1547–1554, 2003.
[3] J. Tani and M. Ito, "Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 33, no. 4 pp. 481–488, Jul. 2003.
[4] H. Kadone and Y. Nakamura, "Symbolic memory for humanoid robots using hierarchical bifurcations of attractors in nonmonotonic neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2005, pp. 2900–2905.
[5] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura, "Embodied symbol emergence based on mimesis theory," *Int. J. Robot. Res.*, vol. 23, no. 4, pp. 363–377, 2004.
[6] A. Billard, S. Calinon, and F. Guenter, "Discriminative and adaptive imitation in uni-manual and bi-manual tasks," *Robot. Auton. Syst.*, vol. 54, pp. 370–384, 2006.
[7] T. Asfour, F. Gyarfas, P. Azad, and R. Dillmann, "Imitation learning of dual-arm manipulation task in humanoid robots," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2006, pp. 40–47.
[8] D. Kulic, H. Imagawa, and Y. Nakamura, "Online acquisition and visualization of motion primitives for humanoid robots," in *Proc. 18th IEEE Int. Symp. Robot Human Interactive Commun.*, 2009, pp. 1210–1215.
[9] K. Sugiura, N. Iwahashi, H. Kashioka, and S. Nakamura, "Active learning of confidence measure function in robot language acquisition framework," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2010, pp. 1774–1779.
[10] Y. Cheng *et al.*, "Modeling natural language controlled robotic operations," in *Proc. IEEE Int. Conf. CYBER Technol. Autom., Control, Intell. Syst.*, 2017, pp. 1072–1077.
[11] Y. Cheng, Y. Shi, Z. Sun, and L. Dong, "Analytic approach for robot control using natural language in dynamic environment," in *Proc. IEEE Int. Conf. Inf. Autom.*, 2018, pp. 1503–1508.
[12] C. Rose, B. Bodenheimer, and M. F. Cohen, "Verbs and adverbs: Multidimensional motion interpolation," *IEEE Comput. Graph. Appl.*, vol. 18, no. 5 pp. 32–40, Sep./Oct. 1998.
[13] O. Arikan, D. A. Forsyth, and J. F. O'Brien, "Motion synthesis from annotations," *ACM Trans. Graph.*, vol. 22, no. 3 pp. 402–408, 2003.
[14] W. Takano, K. Yamane, and Y. Nakamura, "Capture database through symbolization, recognition and generation of motion patterns," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2007, pp. 3092–3097.
[15] W. Takano, D. Kulic, and Y. Nakamura, "Interactive topology formation of linguistic space and motion space," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2007, pp. 1416–1422.
[16] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal object categorization by a robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2007, pp. 2415–2420.
[17] T. Nakamura, T. Nagai, and N. Iwahashi, "Bag of multimodal LDA models for concept formation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 6233–6238.
[18] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adaptive Behav.*, vol. 18, no. 1 pp. 33–52, 2005.

[19] T. Ogata, M. Murase, J. Tani, K. Komatani, and H. G. Okuno, "Two-way translation of compound sentences and arm motions by recurrent neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2007, pp. 1858–1863.
[20] H. Arie, T. Endo, S. Jeong, M. Lee, S. Sugano, and J. Tani, "Interactive learning between language and action: A neuro-robotics experiment," in *Proc. 20th Int. Conf. Artif. Neural Netw.*, 2010, pp. 256–265.
[21] T. Ogata and H. G. Okuno, "Integration of behaviors and languages with a hierarchical structure self-organized in a neuro-dynamical model," in *Proc. IEEE Symp. Series Comput. Intell.*, 2013, pp. 94–100.
[22] W. Takano and Y. Nakamura, "Integrating whole body motion primitives and natural language for humanoid robots," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2008, pp. 708–713.
[23] W. Takano and Y. Nakamura, "Statistically integrated semiotics that enables mutual inference between linguistic and behavioral symbols for humanoid robots," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 646–652.
[24] W. Takano and Y. Nakamura, "Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions," *Int. J. Robot. Res.*, vol. 34, no. 10 pp. 1314–1328, 2015.
[25] W. Takano, Y. Yamada, and Y. Nakamura, "Generation of action description from classification of motion and object," *Robot. Auton. Syst.*, vol. 91, pp. 247–257, 2017.
[26] W. Takano, Y. Yamada, and Y. Nakamura, "Linking human motions and objects to language for synthesizing action sentences," *Auton. Robots*, vol. 43, no. 4 pp. 913–925, 2019.
[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553 pp. 436–444, 2015.
[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst*, 2012, vol. 25, pp. 1106–1114.
[29] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27, pp. 3104–3112.
[30] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *J. Mach. Learn. Res.*, vol. 17, pp. 1–40, 2016.
[31] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 3389–3398.
[32] M. Plappert, C. Mandery, and T. Asfour, "Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks," *Robot. Auton. Syst.*, vol. 109, pp. 13–26, 2018.
[33] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh, "Text2action: Generative adversarial synthesis from language to action," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 5915–5920.
[34] T. Yamada, H. Matsunaga, and T. Ogata, "Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4 pp. 3441–3448, Oct. 2018.
[35] M. P. Kadaba, H. K. Ramakrishnan, and M. E. Wootten, "Measurement of lower extremity kinematics during level walking," *J. Orthopaedic Res.*, vol. 8, no. 3 pp. 383–392, 1990.
[36] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.