Proceedings of the 2020 IEEE/SICE
International Symposium on System Integration
Honolulu, Hawaii, USA, January 12-15, 2020

Sentence Generation from IMU-based Human Whole-Body Motions in Daily Life Behaviors

Wataru Takano

***Abstract*— This paper presents a probabilistic approach toward integrating human whole-body motions with natural language. Human whole-body motions in daily life are recorded by inertial measurement units (IMU) and subsequently encoded into motion symbols. Sentences are manually attached to the human motion primitives for their annotation. Two aspects of semantics and syntactics are represented by probabilistic graphical models. One probabilistic model trains the linking of motion symbols to words, and the other model represents sentence structure as word sequences. These two models are useful toward translating human whole-body motions into descriptions, where multiple words are associated from the human motions by the first model, and the second model searches for syntactically consistent sentences consisting of the associated words. The proposed approach was tested on a large dataset of human whole-body motions and sentences to annotate these motions. The linking of human motions to natural language enables robots to understand observations of human behavior as sentences.**

I. Introduction

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 structured society and culture. Humans are animals that create structure.

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 understand it as a combination of symbols, and view it in structural terms. Since the world is organized as symbolic representations with low information, it can be efficiently 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 with symbols and language being useful tools become the advanced knowledge systems of humanity.

The path toward incorporating knowledge systems into robots lies in transforming the world into language. Robots have a particular need for representing human behavior as language such that they can coexist with us in our daily lives. Extensive research has investigated how to describe the whole body motions of humanoid robots through mathematical modeling. A popular approach is imitation learning or programming by demonstration. It transforms human whole body motions into robot motions that are consequently recorded as a mathematical model. Taking the continuous information of human motion as a parameter set for a mathematical model, expressed as a discrete representation, provides motion symbols, which can be regarded as the first step toward symbolization. Such motion symbols are used as a motion synthesizer to plan, produce, and control robot whole body motion in accordance with the mathematical model. They can be used not only in techniques for synthesizing human-like motions, but also in techniques for recognizing human motions by comparing observed motions with motion symbols.

Language has semantic and syntactic functions that denote meaning and rules respectively. Motion symbols successfully represent relations between the signified and the signifier by attaching mathematical models to actually human or robot motions, but they have not taken into account the syntactic function of language. By unifying meaning with syntactic processing, we can raise language parsing from the word level to the textual level. Thus, by representing intertextual word orders for grammatical rules as a mathematical model, and combining them with motion symbols, we can advance the study of transforming motion into language.

Our daily lives are full of various human motions and linguistic expressions. The robots must overcome this diversity by developing intelligence that understands various motion as language such that they can permeate our society. One key approach to overcoming this diversity lies in enormous datasets related to human behavior and linguistic expressions. This study uses human whole body motion data from everyday life, captured via wearable motion sensor suits, along with a collection of text sentences annotating these data, to create such an enormous dataset. Using probabilistic graphical models that combine human whole body motions and text sentences, we constructs a novel system that translates a wide variety of human whole body motions into sentences, and its validity is demonstrated through experiments for quantitative evaluations.

II. Related Work

A framework of imitative learning or programming by demonstration is a popular approach toward constructing artificial intelligence based on motion data. There are two potential methods for motion modeling: stochastic systems [1][2][3][4][5] and dynamical systems [6][7][8]. Stochastic systems represent motions as motion transitions and distributions while dynamical systems represent motions

*This research was supported by a Grant-in-Aid for Challeging Research (Exploratory) (17K20000) from the Japan Society for the Promotion of Science.

Wataru Takano is with the Center for Mathematical Modeling and Data Science, Mathematics and Computer Science, Osaka University, 1-3 Machikaneyamacho, Toyonaka, Osaka, Japan takano@sigmath.es.osaka-u.ac.jp

978-1-7281-6667-4/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on June 21,2020 at 01:30:55 UTC from IEEE Xplore. Restrictions apply.

as differential equations in the state space. Both encode motions into model parameters. This encoding implies that continuous motion data are discretely compressed into points in the parameter space, and that points can be defined as motion symbols. The motion symbols are used as motion synthesizers and motion recognizers. Motion synthesizers generate robot motions similar to training motions according to dynamics embedded in the mathematical model. Motion recognizers classify observations of human motion into the motion symbol most similar to the observation. However, it is not easy for humans to use motion synthesizers or recognizers, because mathematical models do not provide an intuitive interface. To put it briefly, humans cannot understand motion symbols from model parameters alone. It is thus necessary to form a bridge between motion symbols and natural language to improve ease of use.

We have presentend a probabilistic framework to integrate human or robot whole-body motions with natural language [9][10][11]. This framework is built on two probabilistic graphical models: one model trains associations between motions and words, and the other trains arrangements of words in sentences. This realizes bidirectional mapping between motions and their relevant sentences. This mapping was validated by an experiment on a dataset of human motions and attached sentences that were collected in a laboratory with well controlled experimental conditions. Additionally, this framework was extended to take into account manipulated objects to allow for the generation of correct and detailed sentences from human behaviors involving whole-body motion and object data [12]. An approach to integrating motions and language was presentend by Sugita and Tani in a different direction [13]. Two recurrent neural networks for motions and sentences share parameters. The shared parameters allow synthesizing motions from sentences. This framework was extended to generate sentences describing motions by Ogata et al [14]. Several sentences are composed 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 chosen by comparing the generated motions with the original motion. This research has been expanded to accommodate a wide variety of motions and sentences [15].

In recent years, deep learning techniques [16] have outperformed state-of-the-art machine learning frameworks in computer vision [17] and natural language processing [18], and have gradually come to be applied to problems in robotics [19][20]. Plappert et al. presented the application of deep learning techniques to link human whole-body motions to natural language [21]. In their framework, one recurrent neural network estimates context from motion, and another receives the context and previous word to predict the following word. Iteration of this process results in the synthesis of a word sequence providing a description of the motion. The deep learning techniques of generative adversarial networks was also used to design a motion synthesizer from language [22]. This framework is built on a generator and a discriminator using 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 by taking into consideration the text feature. The generator and discriminator develop in a mutual manner such that the generator creates a realistic human motion and the discriminator differentiates between realistic and generated motions. Yamada et al. presented a deep learning technique for bidirectional translation between actions including motions and visual perception on the one hand and sentences on the other hand [23]. One autoencoder in recurrent neural networks encodes actions into action features and decodes actions from those features. Another autoencoder encodes sentences into text features, and decodes sentences from those features. These autoencoders separately train action and sentence data, and are additionally tuned to minimize error between these two features. This tune matches actions to language.

III. Integration of Motions and Language

A. *Motion Representation*

Optical motion-capture systems and inertial measurement unit (IMU) sensors have been developed for measurement of human whole-body motions. Optical motion capture systems can accurately measure the position of several markers attached to a performer, but the measurement space is constrained to controlled laboratory with installed capture cameras. It is difficult to record human motions under a wide variety of situations. In contrast, IMU sensors are attached to a performer and measure linear acceleration, angular velocity, and magnetic fields even when the performer moves out of a laboratory. Robotics computational algorithms, such as forward kinematics and inverse kinematics, convert these sensor outputs to the posture of a human character, thereby estimating whole-body joint angles and positions. IMU sensors do not limit measurement range, and have an advantage over optical motion-capture systems in terms of measurement area. We therefore used IMU sensors to measure human whole-body motions to build a large dataset of human motions.

Seventeen IMU sensors are attached to a performer. The resulting sensor data are transformed to positions of 34 virtual markers attached to the performer [11]. Human whole-body posture is expressed 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 Hidden Markov Model (HMM), which is referred to as a motion symbol. Observations of human whole-body motion are classified into the motion symbols that is the most likely to generate those observations. In this way, the motion symbols can be used as motion recognizers.

B. *Mapping between Motions and Words*

We describe a probabilistic approach toward converting from human whole-body motions to descriptive sentences.

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on June 21,2020 at 01:30:55 UTC from IEEE Xplore. Restrictions apply.

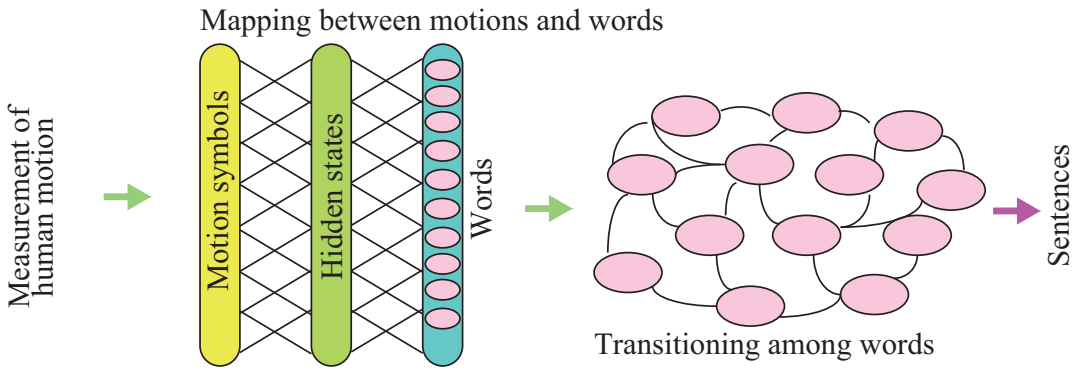

Fig. 1. Overview of mapping from motion to language. A motion symbol is connected to words via latent states in a probabilistic graphical model. Transitions among words are also represented in a probabilistic graphical model. An observation of human whole body motion is classified into a motion symbol. The motion symbol is converted to multiple words, based on knowledge of relations between motion symbols and words. Sentences are subsequently created from those words according to knowledge of word transitions.

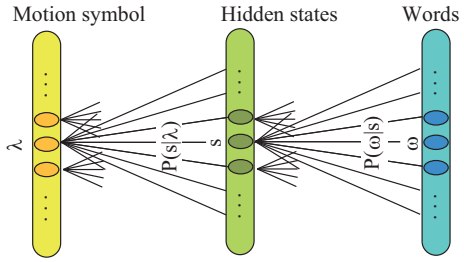

Fig. 2. Mappings between motions and words are represented by a probabilistic graphical model. Motions are connected to words via a hidden state by the probability of the hidden state being generated by the motion and the probability of the word being generated by the hidden state.

As shown in Fig. 1, our approach uses two models, 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 the first layer is composed of motion symbols, the third layer is composed of words, and the second layer is composed of hidden states that connect motion symbols with words. Figure 2 shows the mapping architecture. The connection is described by parameters indicating 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 found by an EM algorithm that maximizes the probability $P(\boldsymbol{\omega}|\lambda)$ of generating a set of words included in sentence $\boldsymbol{\omega}$ from motion symbol λ;

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

where the k-th 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

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

the probability $P(\boldsymbol{\omega}^{(*)}|\lambda^{(*)})$ is assumed to be 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 $P(s|\lambda,\omega)$ 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 estimate $P(s|\lambda,\omega)$ 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 [11].

C. Word Sequences for Natural Language

Sentences are word sequences, and sentence structures are assumed to be extracted by transitions among words. Sentences are thereby represented by a probabilistic graphical model. Figure 3 shows this graphical model whose nodes and edges respectively denote words and transitions among words. In the word N-gram model, the graphical model is described by the inter-word 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

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on June 21,2020 at 01:30:55 UTC from IEEE Xplore. Restrictions apply.

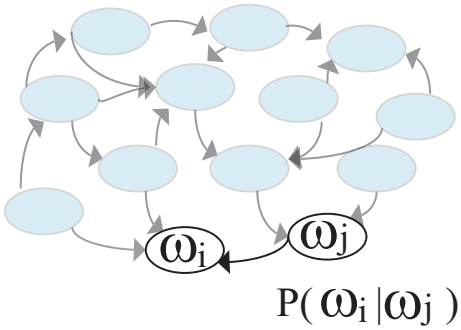

Fig. 3. Sentence structures are assumed to be represented by transitions among words. In case of word bigrams, a node and an edge are a word and a transition between two word respectively.

$P\omega$ of starting at ω. The probability $P(\omega|\boldsymbol{\omega}_{1:N-1})$ can be optimized such that the probability $P(\boldsymbol{\omega})$ of generating training sentence $\boldsymbol{\omega}$ is maximized. The objective function $\mathcal{Q}$ is written as

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

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

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

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

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. Translation from Motion to Sentences

Human whole-body motion is translated to a sentence to describe the motion by combining two probabilistic graphical models, one for mappings between human motions and words and one for word arrangement. The translation 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 generating word sequence $\boldsymbol{\omega}$ from human whole-body motion $\boldsymbol{x}$ is rewritten as

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

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

Note that the prior probability of $P(\boldsymbol{x})$ has no effect on searching for the sentence. We assume that prior probabilities of $P(\lambda)$ are equivalent for all motion symbols. The following problem of searching for the sentence with the highest probability is obtained.

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

$$\begin{aligned} &= \arg\max_{\boldsymbol{\omega}} \sum_\lambda P(\boldsymbol{\omega}|\omega_1, \omega_2, \cdots, \omega_m) \\ &\quad \times P(\omega_1, \omega_2, \cdots, \omega_m|\lambda)P(\boldsymbol{x}|\lambda). \end{aligned} \quad (13)$$

It is decomposed into three probabilities: the probability of sentence $\boldsymbol{\omega}$ being created from words $\{\omega_1, \omega_2, \cdots, \omega_m\}$ by the graphical model of the word arrangement in Eqn. 8; the probability of words $\{\omega_1, \omega_2, \cdots, \omega_m\}$ being generated from motion symbol λ by the graphical model mapping between human motions and words in Eqn. 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.

Fig. 4. IMU sensors are attached to performers, whose whole-body motions in daily life are measured.

IV. Experiments

We tested the proposed approach for generating sentences to describe human behavior on a dataset of human whole-body motions with annotations attached to those motions. Human whole-body motions were measured using commercial IMU sensors (Xsens Technologies), shown in Fig. 4. Seventeen IMU sensors with a sampling rate of 120 Hz were attached to a performer. Their positions were estimated from measurements of linear acceleration and angular velocity. Position data were retargeted to a posture of a human character with 34 degrees of freedom. The positions of virtual markers fixed on the human character were computed by inverse kinematics and forward kinematics algorithms. The virtual markers were placed over the character's whole body according to Helen Hayes marker placement [24]. IMU measurements were converted to a resultant posture feature vector whose elements are positions of virtual markers in the trunk coordinate system. Human whole-body motions were subsequently expressed as sequences of posture features. We recorded human whole-body motions of 5 male and 8 female performers. The recorded motion data included 84,868,680 frames, equivalent to 707,239 seconds.

Human whole-body motions were manually annotated. A total of 62,207 sentences with 3,349 different words were collected for the motion annotations. Additionally, motion segments described by the sentences were collected by

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on June 21,2020 at 01:30:55 UTC from IEEE Xplore. Restrictions apply.

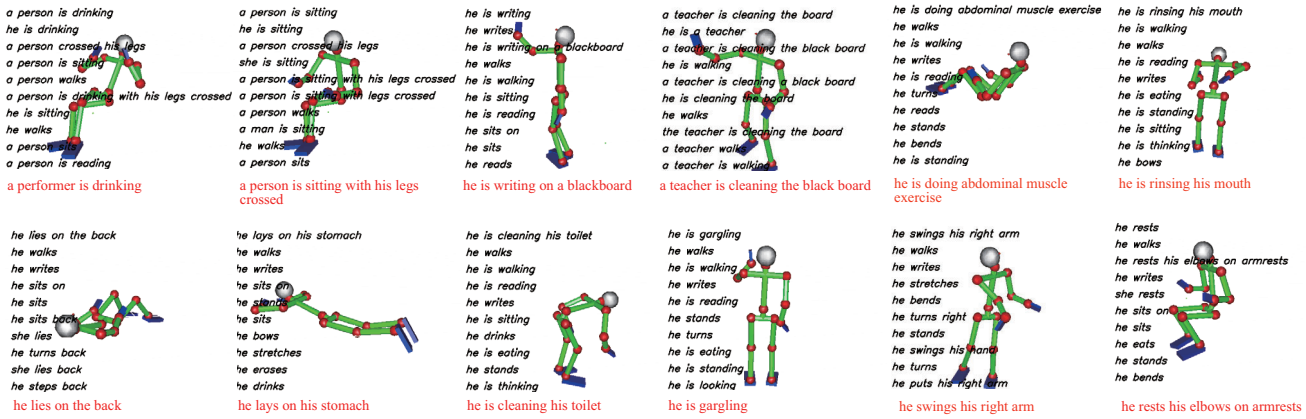

Fig. 5. Ten most likely sentences are generated from human whole-body motions. The generated sentences are displayed in black, and correct sentences are displayed in red.

manually detecting their boundaries along the human whole-body motion data. Each motion segment was encoded into an individual HMM for a motion symbol. This implies that the number of motion symbols was equal to the number of sentences. This manual annotation and segmentation created a large dataset of pairs of motion symbols and sentences to annotate the motions.

We set the number of hidden states in the probabilistic graphical model for mapping between motions and words to 4,000. This graphical model trained the mapping from the dataset of motions and annotation 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 probability of generating words from the motion symbols and probability of transitioning among the words. Figure 5 qualitatively shows several examples of sentence generation. We searched for multiple sentences, sorted them in descending order of probability, and displayed the ten sentences with the highest probability. These examples show that correct or nearly correct sentences were generated to describe the human motions with semantic and syntactic consistencies. 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 performed a quantitative evaluation of the generation of sentences from human whole-body motions. The number of hidden states was set to 40, 400, and 4000. We chose the bilingual evaluation understudy (BLEU) score as the evaluation index. It is commonly used for machine translation performance [25]. BLEU scores are formulated based on word N-gram matching between a reference sentence and a generated sentence as

$$BLEU = bp \exp \sum_{n=1}^{N} \frac{\log p_n}{N} \quad (14)$$

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

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. l_r and l_g are the lengths of the reference and generated sentences, respectively. bp is a brevity penalty. BLEU scores range from 0 to 1, and higher scores indicates closer similarity between the reference and generated sentences. In this evaluation, N was set to 4. Table I shows average BLEU scores. The model with 400 hidden states achieved higher BLEU scores than the model with 40 hidden states in a range from 1st to 6th ranks. The model with 4000 hidden states achieved an average BLEU score of 0.604, outperforming the model with 400 hidden states (BLEU score 0.594) in generating 1st-rank sentences. However, the model with 400 hidden states achieved better performance in generating 2nd- to 10th-rank sentences than did the model with 4000 hidden states.

V. Conclusions

The conclusions we obtained from this research are as follows.

1) We recorded many human whole-body daily life motions by using IMU sensors. We additionally annotated these human motions. we thereby created a large dataset of human whole-body motions and sentences annotating the motions. In total, we recorded 84,868,680 frames of human whole-body motions and attached 62,207 sentences to these data.
2) We presented a framework for combining human whole body motions with sentences. Mappings between human motions and relevant words, and transitioning among words in sentences, are represented

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on June 21,2020 at 01:30:55 UTC from IEEE Xplore. Restrictions apply.

TABLE I

BLEU SCORES FOR SENTENCE GENERATION. $\mathcal{H}_{40}$, $\mathcal{H}_{400}$, AND $\mathcal{H}_{4000}$ DENOTE MAPPINGS BETWEEN MOTIONS AND SENTENCES, WITH THE NUMBER OF HIDDEN STATES BEING SET TO 40, 400, AND 4000, RESPECTIVELY.

	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

by probabilistic graphical models. A bridge between these two models makes it possible to translate human whole-body motions into sentences for annotations.

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 three probabilities; probability of observed human motion being generated by motion symbols, probability of words being associated with those motion symbols, and probability of word transitions. We could consequently compose reasonable sentences with the highest resultant probability. We performed a quantitative evaluation of sentence generation based on BLEU scores measuring matching rate between generated and reference sentences. The established dataset proved to be useful, verifying the validity of the proposed method for integrating human motions with sentences to annotate human whole-body motions.

REFERENCES

[1] T. Inamura, I. Toshima, H. Tanie, and Y. Nakamura, "Embodied symbol emergence based on mimesis theory," *International Journal of Robotics Research*, vol. 23, no. 4, pp. 363–377, 2004.

[2] A. Billard, S. Calinon, and F. Guenter, "Discriminative and adaptive imitation in uni-manual and bi-manual tasks," *Robotics and Autonomous Systems*, vol. 54, pp. 370–384, 2006.

[3] T. Asfour, F. Gyarfas, P. Azad, and R. Dillmann, "Imitation learning of dual-arm manipulation task in humanoid robots," in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 40–47.

[4] D. Kulic, H. Imagawa, and Y. Nakamura, "Online acquisition and visualization of motion primitives for humanoid robots," in *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication*, 2009, pp. 1210–1215.

[5] K. Sugiura, N. Iwahashi, H. Kashioka, and S. Nakamura, "Active learning of confidence measure function in robot language acquisition framework," in *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1774–1779.

[6] M. Okada, K. Tatani, and Y. Nakamura, "Polynomial design of the nonlinear dynamics for the brain-like information processing of whole body motion," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2002, pp. 1410–1415.

[7] A. J. Ijspeert, J. Nakanishi, and S. Shaal, "Learning control policies for movement imitation and movement recognition," *Neural Information Processing System*, vol. 15, pp. 1547–1554, 2003.

[8] J. Tani and M. Ito, "Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment," *IEEE Transactions on Systems, Man and Cybernetics Part A: Systems and Humans*, vol. 33, no. 4, pp. 481–488, 2003.

[9] W. Takano and Y. Nakamura, "Integrating whole body motion primitives and natural language for humanoid robots," in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2008, pp. 708–713.

[10] ——, "Statistically integrated semiotics that enables mutual inference between linguistic and behavioral symbols for humanoid robots," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2009, pp. 646–652.

[11] ——, "Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions," *International Journal of Robotics Research*, vol. 34, no. 10, pp. 1314–1328, 2015.

[12] W. Takano, Y. Yamada, and Y. Nakamura, "Linking human motions and objects to language for synthesizing action sentences," *Autonomous Robots*, vol. 43, no. 4, pp. 913–925, 2019.

[13] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adaptive Behavior*, vol. 18, no. 1, pp. 33–52, 2005.

[14] 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 *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 1858–1863.

[15] T. Ogata and H. G. Okuno, "Integration of behaviors and languages with a hierarchical structure self-organized in a neuro-dynamical model," in *Proceedings of the IEEE Symposium Series on Computational Intelligence*, 2013, pp. 94–100.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems 25*, 2012, pp. 1106–1114.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of Advances in Neural Information Processing Systems 27*, 2014, pp. 3104–3112.

[19] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, pp. 1–40, 2016.

[20] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2017, pp. 3389–3398.

[21] M. Plappert, C. Mandery, and T. Asfour, "Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks," *Robotics and Autonomous Systems*, vol. 109, pp. 13–26, 2018.

[22] H. Ahn, T. Ha, Y. Choi, H. Yoo, and S. Oh, "Text2action: Generative adversarial synthesis from language to action," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2018, pp. 5915–5920.

[23] T. Yamada, H. Matsunaga, and T. Ogata, "Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3441–3448, 2018.

[24] M. P. Kadaba, H. K. Ramakrishnan, and M. E. Wootten, "Measurement of lower extremity kinematics during level walking," *Journal of Orthopaedic Research*, vol. 8, no. 3, pp. 383–392, 1990.

[25] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

Authorized licensed use limited to: TIANJIN UNIVERSITY. Downloaded on June 21,2020 at 01:30:55 UTC from IEEE Xplore. Restrictions apply.