Action Description From 2D Human Postures in Care Facilities

Wataru Takano and Haeyeon LEE

***Abstract*—This article describes a novel approach to classification of whole-body motions from estimated human postures in 2D camera images and subsequent generation of their relevant descriptions. The motions are encoded into stochastic motion models referred to as motion primitives. Words are connected to the motion primitives, and word n-grams are represented stochastically. More specifically, the motion observation is classified into the motion primitive, from which its relevant words are generated. These words are arranged in a grammatically correct order to make the descriptions for the observation. This approach was tested on actions performed by older adults in the care facility and its validity was demonstrated.**

***Index Terms*—Gesture, posture and facial expressions, probability and statistical methods, recognition.**

I. Introduction

THERE are high expectations for a human-centered society in which social problems will be solved through new social systems that connect real and virtual worlds. The virtual world of the Internet has formed a useful collective intelligence by organically linking contents via language. Human intelligence in the real world has a history of constantly developing knowledge systems by describing, recording, communicating, and passing on information about the world via language. In other words, language is the tool by which we fuse the virtual and real worlds.

Cameras have become pervasive in daily life as a form of social infrastructure, and observations of human behavior are increasing in society. An enormous amount of data related to human behavior is recorded at various times and places. Expansion of the data utilization and support fields will rely on how behavioral data is constructed and managed in a way that is easy to reuse. In a rapidly aging society, there are limits on the extent to which the young can accompany and care for the elderly. Against this backdrop, how can the accumulation of behavioral data and cutting-edge technologies contribute to tackling this pressing social issues?

Manuscript received September 8, 2019; accepted January 3, 2020. Date of publication January 9, 2020; date of current version January 22, 2020. This letter was recommended for publication by Associate Editor H. Zha and Editor C. Cadena Lerma upon evaluation of the reviewers' comments. This work was supported by the Grant-in-Aid for Challenging Exploratory Research under Grant 17K20000 from the Ministry of Education, Culture, Sports, Science and Technology. *(Corresponding author: Wataru Takano.)*

W. Takano is with the Center for Mathematical Modeling and Data Science, Mathematics and Computer Science, Osaka University, Osaka 560-8531, Japan (e-mail: takano@sigmath.es.osaka-u.ac.jp).

H. LEE is with Partner Robot Div, Toyota Motor Corporation, Toyota 471-8571, Japan (e-mail: haeyeon_lee_aa@mail.toyota.co.jp).

Digital Object Identifier 10.1109/LRA.2020.2965394

In this study, we estimated 2D human postures accompanying behavior as recorded by cameras in care facilities, and developed a method for classifying behavior. Further, we introduce a technique for generating descriptions to link behavior classifications with language in order to convert behavioral data into an intuitively easy-to-understand textual form. By using machines to generate descriptions of behavior in care facilities, it becomes possible to automatically create care journals. This would allow the daily situation at facilities to be ascertained by reviewing care journals, without the need for replaying videos. Further, detecting descriptions associated with unusual behavior or emergency situations would lead to support systems capable of providing prompt notifications when needed.

II. Related Work

There has been considerable research on mathematical modeling of human or robot motions. The models can be divisible into two groups; dynamical system approach [1]–[4] and stochastic approach [5]–[10]. A dynamical system approach encodes the motions data into differential equations, and a probabilistic approach encodes the motion data into the state transitions and data distributions in the states. Both of them represent the continuous motion data as a set of model parameters, which can been seen as a discrete point in a multidimensional parameter space. It implies that the motion data is compressed into the mathematical model in a symbolic representation which is referred to as a motion symbol. The motion symbol is helpful not only to synthesize a human-like motion but also to recognize an observation of human actions. More specifically, a recognizer finds a model that is the closest to the observation, and the observation is consequently classified into this model category. The mathematical model is not intuitive representation. When we receive the response of the model into which the observation is classified, we cannot understand the motion category such as "walking," "running," or "squatting. It is important to connect the motion model to human-readable representation typified by language.

In the robotics and computer graphics, researchers aim at connecting motions to language. A novel concept of "verb" and "adverb" was proposed in the motion space [11]. "verb" is a group of similar motions, and "adverb" is difference between two motions in the same group. "adverb" is parameterized, and this parameter controls the interpolation between these motions. It consequently synthesizes a new motion from queries of "verb" and "adverb". Arikan *et al.* also proposed a method to create character motions from inputs of words [12]. A database of

motions and verb labels attached to their frames is established. This method finds a continuous sequence of motions to which the input of verb labels is attached. A dynamic programming can efficiently search fpr a sequence of motions while satisfying two constraints of the continuity and attachment of the verb labels. We were inspired by the research on machine translation based on the probabilistic model, and applied the translation model to pair of two sequences; human motions and verb labels attached to the motions [13], [14]. This approach makes it possible to convert an observation of human action to its relevant word sequence.

Sugita and Tani integrated motions and language by using two neural networks for the motions and sentences with parameters being shared [15]. It allows for motion synthesis from sentences. Ogata *et al.* extended this framework to synthesize sentences from motions [16]. Multiple candidates of sentence are created from a motion, and each candidate is subsequently converted to a motion in the same manner as Sugita and Tani. An appropriate candidate is selected by comparing the generated motion with the original motion. This research has been continued to handle a variety of motions and sentences [17], [18]. We have also presented a framework to integrate human/robot full body motions with natural language in the probabilistic method [19]–[21]. It is composed of two probabilistic graphical models; one model trains the relation between motions and words in charge of semantic function, and the other model trains the arrangement of words along sentences in charge of syntactic function. They allow for the bidirectional mapping between the motions and their relevant sentences. Furthermore, this framework was extended to handle manipulated objects to improve the generation of correct and detailed sentences from human behaviors of full body motion data and object data [22].

In recent years, deep learning techniques [23] have made tremendous progress in computer vision [24] and natural language processing [25], and gradually have been used to solve problems in robotics [26], [27]. Plappert *et al.* applied deep learning techniques to connect human full body motions to natural language [28]. One recurrent neural network computes context from motion, and another recurrent neural network receives the context and previous word, and subsequently predicts the following word. The iteration of this process results in a sequence of the words for the description of the motion. T. Yamada *et al.* presented a bidirectional translation between visual motion perception and sentence based on a deep learning technique [29]. A pair of recurrent neural networks encodes the perception into features, and decodes the perception from its relevant feature in the manner of auto-encoder. Another pair of recurrent neural networks encodes the sentence into text feature, and decodes the sentence from the feature. The recurrent networks train pairs of perception and sentence, and are subsequently tuned such that the error between these two features is minimized. This tuning process binds the perception to language. Ahn *et al.* also adopted the deep learning techniques of a generative adversarial network to construct a motion synthesizer from language [30]. This framework is composed of a generator and a discriminator based on recurrent neural networks. The generator extracts a text feature from a sentence and generates a human motion from the text feature. The discriminator also extracts a text feature from a sentence and differentiate a real motion from the generated motion with consideration of the text feature. These generator and discriminator develop in the mutual way that the generator creates a realistic human motion and that the discriminator differentiates between a realistic motion and a generated motion. These previous approaches have dealt with three dimensional data in order to aim not only at sentence generation from motions but also at motion synthesis from sentences. More specifically, a technique to recover three dimensional human postures from camera images containing human actions is absolutely essential for the methods mentioned above. Although the three dimensional data of human postures would decrease the ambiguity for action classification, three dimensional posture estimation is challenging especially from monocular camera images. When targeting only at generating sentences from human motions, it may not need to estimate three dimensional postures, and a technique of generating sentences from two dimensional postures in camera images would be easy-to-implement.

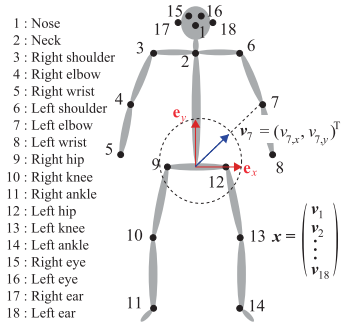

Fig. 1. Keypoints and features for whole-body motions.

III. Calculation of Descriptions From Whole-Body Motions

A. Motion Representation

We use a camera to capture human whole-body motion, and in the captured images we detect the positions of the 18 feature points shown in Fig. 1 [31]. A 2D coordinate plane is created, taking the midpoint between "left hip" and "right hip" as the origin, Then, a line drawn between these two locations is taken as the x-axis, and a line passing through the origin and "neck" is taken as the y-axis. Note that the y-axis is adjusted to be orthogonal to the x-axis. Let $\boldsymbol{e}_x$ be the unit vector in the x-direction, and let $\boldsymbol{e}_y$ be the unit vector in the y-direction. Further, let $\boldsymbol{v}_k$ be a unit vector from the origin in the direction of the kth feature point. Vector $\boldsymbol{x}$ is a feature vector for full-body posture from unit direction vectors $\boldsymbol{v}_k$, where $k = 1, 2, \ldots, 18$.

$$\boldsymbol{x} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{18}]^T \tag{1}$$

This feature vector is invariant to body shape and translational position of the subject.

Full-body motion is represented as a time-series signal for feature vector $\boldsymbol{x}$. This time-series signal $\boldsymbol{X}$

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T] \tag{2}$$

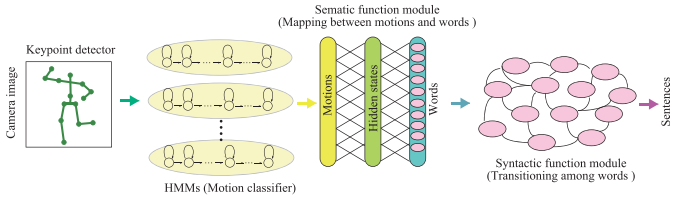

Fig. 2. Keypoints on a performer in camera images are detected for a feature vector time-series, which is classified into the relevant motion pattern. Several words are associated from the motion pattern, and these words are arranged in a grammatical valid order to sentence describing the motion.

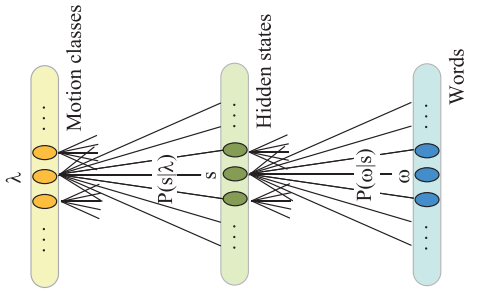

Fig. 3. A semantic function module is established from three layers; motions, hidden states, and words. The connections among these layers are represented by a probability of a hidden state being generated from a motion, and a probability of a word being generated from a hidden state.

is learned using a Hidden Markov model (HMM λ) for each motion pattern. Here, "learning" refers to optimizing parameters to maximize the probability $P(\boldsymbol{X}|\lambda)$ that time-series feature vector $\boldsymbol{X}$ for each motion pattern is generated from the HMM λ. The HMM can be used as a motion classifier that finds an HMM with maximal probability of generating a feature vector time-series $\boldsymbol{X}$ for the observed motion in a set S_λ of HMMs.

$$\lambda_{\mathcal{R}} = \arg\max_{\lambda \in S_\lambda} P(\boldsymbol{X}|\lambda) \tag{3}$$

B. Linking Motions to Language

A framework in statistical mathematics for associating classified motions with language comprises two modules for semantic and syntactic functions [21]. Figure 2 shows a pipeline from camera images containing a performer to a sentence describing a behavior.

A semantic function module expresses an association between motions and words in descriptions as a generation probability for each word under the conditions of the motion classification results. The semantic function module is a probabilistic graphical model where a node denotes a motion class λ, a hidden state s or a word ω, and an edge denotes a probability $P(s|\lambda)$ of generating a hidden state s from a motion class λ, or a probability $P(\omega|s)$ of generating a word ω from a hidden state s. The probabilities, $P(s|\lambda)$ and $P(\omega|s)$, are optimized by EM algorithm such that words in a sentence are the most likely to be generated from a motion class over training pairs of a motion class and a sentence. An objective function is given as

$$\mathcal{P} = \sum_k \log P(\boldsymbol{\omega}^k|\lambda^k). \tag{4}$$

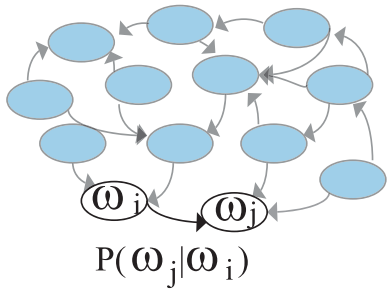

Fig. 4. A sentence structure is assumed to be represented by a transition among words. It is established on word N-gram.

Since a sentence $\boldsymbol{\omega}$ includes some words $\omega^{(*)}$

$$\boldsymbol{\omega} = \left(\omega_1^{(*)}, \omega_2^{(*)}, \ldots, \omega_m^{(*)}\right), \tag{5}$$

this objective function is modified by dividing a sentence into a set of words.

$$\log P(\boldsymbol{\omega}^{(*)}|\lambda^{(*)}) = \sum_{i=1^m} \log P(\omega_i^{(*)}|\lambda^{(*)}) \tag{6}$$

The EM algorithm alternates an estimation step (E-step) and a maximization step (M-step). The E-step estimates the probability of a hidden state given a motion class and a word. The estimate is given as

$$P(s|\lambda,\omega) = \frac{P(\omega|s)P(s|\lambda)}{\sum_s P(\omega|s)P(s|\lambda)}. \tag{7}$$

The M-step optimizes the probabilities,$P(s|\lambda)$ and $P(\omega|s)$, under the estimate $P(s|\lambda,\omega)$ Optimal probabilities can be computed as

$$P(s|\lambda) = \frac{\sum_\omega n(\lambda,\omega)P(s|\lambda,\omega)}{\sum_{s,\omega} n(\lambda,\omega)P(s|\lambda,\omega)} \tag{8}$$

$$P(\omega|s) = \frac{\sum_\lambda n(\lambda,\omega)P(s|\lambda,\omega)}{\sum_{\lambda,\omega} n(\lambda,\omega)P(s|\lambda,\omega)} \tag{9}$$

where $n(\lambda,\omega)$ is the number of correspondences between motion class λ and word ω being observed in the training dataset.

The syntactic function module expresses rules for word arrangements in a description as a probability of transitioning among words in a sentence. This module is also a probabilistic model established on a word N-gram. Figure 4 illustrates the model when N is set to two. This model has a probability $P(\omega|\boldsymbol{\omega}_{1:N-1})$ of a word ω following a sequence of $N-1$ words $\boldsymbol{\omega}_{1:N-1}$. The probability $P(\omega|\boldsymbol{\omega}_{1:N-1})$ is optimized such that a training sentence is the most likely to be generated by this model over the training dataset. The objective function is given as

$$\mathcal{Q} = \sum_k \log P(\boldsymbol{\omega}^{(k)}). \tag{10}$$

An optimal probability $P(\omega|\boldsymbol{\omega}_{1:N-1})$ can be computed as

$$P(\omega|\boldsymbol{\omega}_{1:N-1}) = \frac{n(\boldsymbol{\omega}_{1:N-1},\omega)}{n(\boldsymbol{\omega}_{1:N-1})}, \tag{11}$$

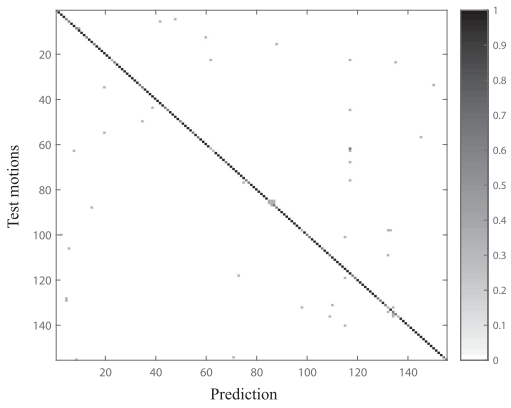

Fig. 5. Confusion matrices show the accuracies of motion classification for 465 motion data.

where $n(\boldsymbol{\omega}_{1:N-1})$ is the number of word sequence $\boldsymbol{\omega}_{1:N-1}$ being observed in the training dataset, and $n(\boldsymbol{\omega}_{1:N-1}, \omega)$ is the number of the word sequence $\boldsymbol{\omega}_{1:N-1}$ being followed by word ω.

From the motion recognition results, an optimal sentence for describing motions can be created based on an index combining the probability of each word being generated according to the semantic function module and the probability of a sentence being formed by the rearrangement of words according to the syntactic function module. More specifically, an observation of a human motion is classified into a motion class λ, and it subsequently searches for a sentence $\boldsymbol{\omega}$ with the largest probability $P(\boldsymbol{\omega}|\lambda)$ of the sentence $\boldsymbol{\omega}$ being generated from the motion class λ. This search problem is rewritten as searching for a sentence with the largest probability as formulated by

$$P(\boldsymbol{\omega}|\lambda) = P(\omega_1, \omega_2, \ldots, \omega_m|\lambda)P(\boldsymbol{\omega}|\omega_1, \omega_2, \ldots, \omega_m). \quad (12)$$

The first term $P(\omega_1, \omega_2, \ldots, \omega_m|\lambda)$ is computed according to the semantic function module, and the second term $P(\boldsymbol{\omega}|\omega_1, \omega_2, \ldots, \omega_m)$ is computed according to the syntactic function module. This search problem can be efficiently solved by Dijkstra's algorithm.

IV. Experiments

A. Preliminary Experiment

We measured the movements of one subject in a laboratory. He was asked to perform designated motions. There were 155 motion patterns with three trials in each motion pattern. We created datasets by dividing the data from 465 trials into three sets, ensuring that each dataset includes motion data for each motion pattern. We used two datasets for training data, and the remaining one dataset for test data. We conducted cross-validation test through repetition in which each dataset was used as test data once.

Figure 5 shows the results of classifying motions performed by one subject. Each dataset to test included 155 as described above. 130, 147, and 140 were correctly classified in case of testing the first, second and third dataset respectively. The classification rates results in 84%, 95% and 90%. To evaluate the validity of HMMs for motion classification on our dataset, we tested an extended graph neural network (Spatial Temporal Graph ConvNet :ST-GCN) as the motion classifier in the same manner as described above [32]. We set the batch size of the ST-GCN and the number of training epochs to 64 and 50 respectively. Top-1 accuracy was 5.7% and top-5 accuracy was 28%. The size of training motion data may be too small for ST-GCN to achieve the high classification rates. This experiment demonstrates that the HMMs construct a reasonable motion classifier on a small dataset of human motions.

Furthermore, we conducted experiments to generate sentences describing motions from the result of motion classification. Multiple sentences were manually attached to each motion pattern. The number of attached sentences was 756, and the number of different words in these sentences was 256. We resultantly created training dataset with pairs of the motion pattern and the sentence. After training the relation between the motion pattern and sentence, and the sentence structure as word-4 grams from this dataset, we tested the sentence generation. Figure 6 shows 10 sentences with a high probability for each motion. When the subject was crossing his legs on a chair, the sentences "a manager crosses his legs," "a coach crosses his legs," and "a student crosses his legs" were generated as the first, second and third candidate. The sentence "a student descends the stairs" was correctly generated when the subject was going down the stairs. The sentence "a student ties the shoe lases" was generated from the motion of tying the shoe lases. Qualitatively correct sentences were generated from full body motions as shown in Fig. 6. We also quantitatively evaluated the generation of sentences from the human motions. We adopted the evaluation index of "BLEU" (Bilingual Evaluation Understudy) that is popularly used to measure the performance of machine translation [33]. A BLEU score is computed based on the word N-gram matching between the reference sentence and generated sentence as follows,

$$BLEU = bp \exp \sum_{n=1}^{N} \frac{\log p_n}{N} \quad (13)$$

$$bp = min\left\{1, \exp\left(1 - \frac{l_r}{l_g}\right)\right\} \quad (14)$$

where p_n is the ratio of the number of word n-gram in the generated sentence that matches a word n-gram in the reference sentence to the number of word n-gram in the generated sentence. bp denotes the brevity penalty. l_r and l_g are the lengths of the reference sentence and the generated sentence, respectively. The BLEU score ranges from 0 to 1 and high BLEU score indicates large similarity between the reference sentence and the generated sentence. In this experiment, N was set to 4. Table I shows average BLEU scores over the generated sentences from all the human motions to be tested. Note that several reference sentences were attached to each motion pattern. We compared a sentence generated from an individual motion pattern with all reference sentences attached to it, and computed BLEU scores on the references, of which the largest BLEU score was selected to compute the average BLEU score. The generation of sentences ranked at the first and the second achieved an average BLEU score of 0.84 and 0.68 respectively.

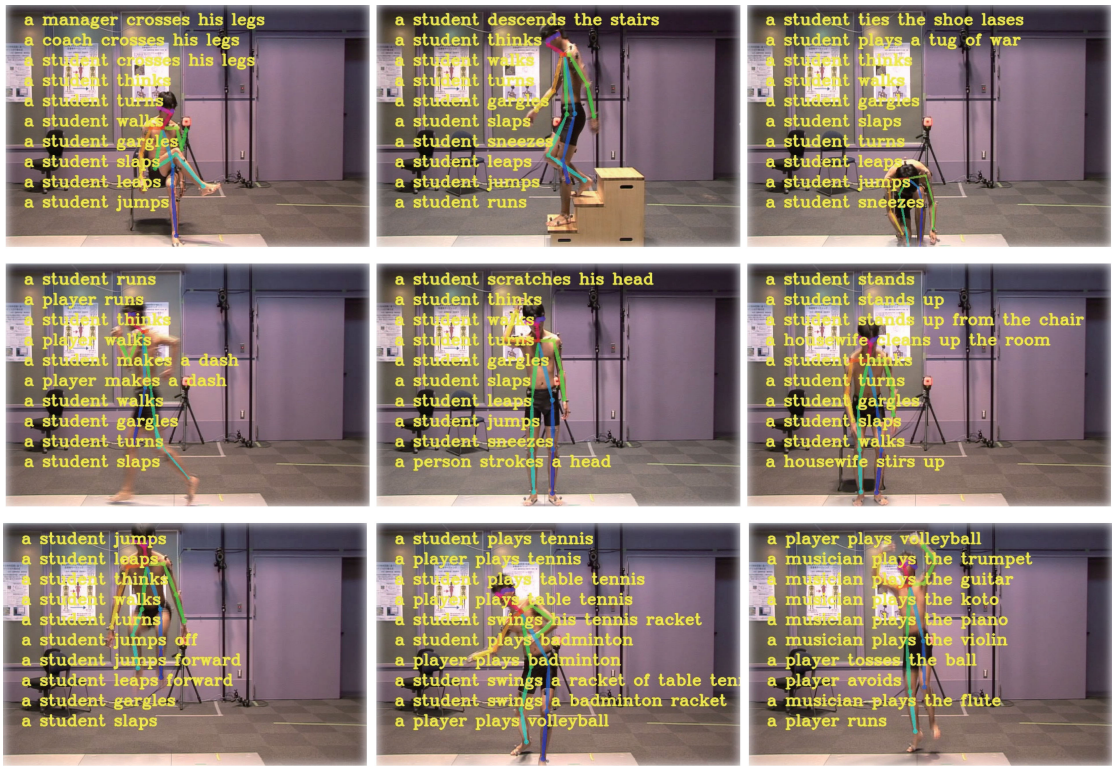

Fig. 6. Motions performed by one subject are recorded in a laboratory. These motions are classified into motion classes, and subsequently converted to descriptions.

TABLE I
BLEU SCORES OF SENTENCE GENERATION IN THE LABORATORY CONDITION. GENERATED SENTENCES RANKED IN THE TOP 10 WERE EVALUATED BY BLEU SCORES

Rank									
1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
0.84	0.68	0.49	0.42	0.40	0.39	0.38	0.38	0.37	0.38

TABLE II
21 DIFFERENT KINDS OF MOTION PATTERNS

Motion index	Motion label
1	sit at a table
2	turn in a wheelchair
3	talk on a wheelchair
4	get up from a bed
5	slide from a bed
6	move from a bed to a wheelchair
7	move from a wheelchair to a bed
8	lie on a bed
9	move in a wheelchair
10	move from a chair to a wheelchair
11	stand up from a wheelchair
12	move from a sofa to a wheelchair
13	eat at a table
14	drink
15	be fed
16	open a door
17	wash hands
18	walk along the corridor
19	walk along the corridor holding on to the railing
20	fall over
21	lose a balance

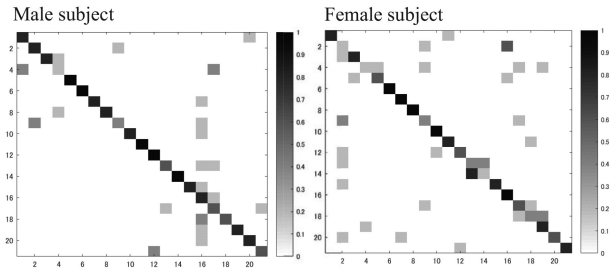

Fig. 7. Confusion matrices shows the accuracies of motion classification for the individual participants.

B. Experiment in a Care Facility

We used a camera to capture the movements of one older man and one older woman in a care facility. Both of them were in good health and were asked to perform several behaviors frequently observed in care facilities. There were 21 motion patterns, listed in Table II, and five trials were captured for each motion pattern for each participant. We prepared datasets by dividing the data from the 210 trials into 10 sets, ensuring that each dataset included motion data for each motion pattern. Nine datasets were used as training data, and the remaining one dataset as test data. We performed cross-validation through repetition in which each dataset was used as test data once.

Figure 7 shows the results of classifying motions for each participant. There were 105 test data for each participant. Among motion data for the male participant, 80 were correctly classified, for a classification rate of 76%. Among motion data

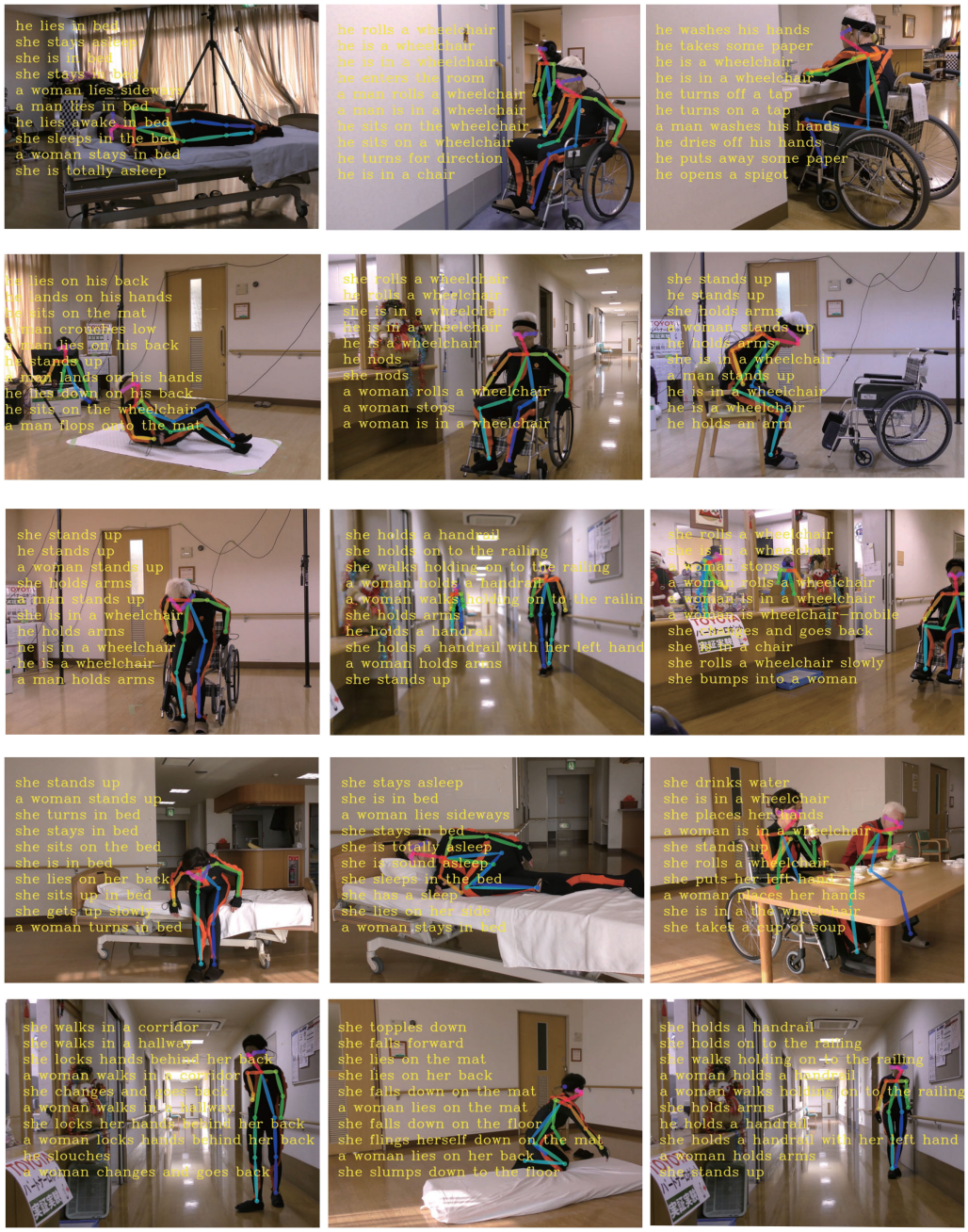

Fig. 8. Behaviors are classified into motion classes, and subsequently converted to descriptions.

TABLE III
BLEU SCORES OF SENTENCE GENERATION IN A CARE FACILITY. GENERATED SENTENCES RANKED IN THE TOP 10 WERE EVALUATED BY BLEU SCORES

Rank									
1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
0.93	0.88	0.93	0.92	0.78	0.73	0.81	0.75	0.71	0.72

for the female participant, 70 were correctly classified, for a classification rate of 67%. As an example of misclassification, a "move in a wheelchair" motion by the male participant was classified as a "turn in a wheelchair" motion. Both motions were related to movement in a wheelchair, so the misclassification was likely related to the similar postures. Similarly, a "turn in a wheelchair" motion by the female participant was classified as an "open a door" motion. The "open a door" motion is performed while sitting in a wheelchair, so this too may be a result of similar posture while sitting in a wheelchair.

We conducted experiments to generate linguistic expressions describing motions from the results of motion classification. We created the dataset of motions and their descriptions in the same manner as the preliminary experiment. 919 sentences were manually attached to 21 motion patterns. 334 different words were used for these sentences. Figure 8 shows 10 descriptions with high probability for each motion. When the male participant

TABLE IV
RESULTS ON THE CROWDSOURCING-BASED EVALUATION OF SENTENCE GENERATION. N_{true} AND N_{false} DENOTE THE NUMBERS OF REPLIES OF "TRUE" AND "FALSE" RESPECTIVELY

	Rank									
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
N_{true}	1753	1347	1349	1551	1113	974	1133	1195	1071	976
N_{false}	734	1105	1142	982	1379	1536	1387	1303	1448	1522
Accuracy rate	0.705	0.549	0.542	0.612	0.447	0.388	0.450	0.478	0.425	0.391

was lying in bed, the description "he lies in bed" was the first generated candidate. The second generated candidate was "she stays asleep." This study did not include calculations for estimating the viewed subject, leading to the incorrect "she" for the grammatical subject, but the remainder of the description was accurate. The description "she rolls a wheelchair" was generated when the female participant rolls in a wheelchair. These examples qualitatively confirm that appropriate descriptions are generated for movements by older adults.

We additionally evaluated the generation of sentences from the motions by the BLEU score. Table III shows average BLEU scores over the generated sentences from all the human motions to be tested. The average BLEU scores of generated sentences ranked in top 10 ranged from 0.71 to 0.93. The high scores were achieved during generating sentences from the motions. More reference sentences were attached to each motion pattern than in the preliminary experiment. When a generated sentence was very close to at least one of the reference sentences, this sentence generation gained a high BLEU score.

We also quantitatively evaluated the motion generation through a crowdsourcing framework. We created a video containing participant's behavior and generated sentences, and uploaded it on Youtube. We asked a Yahoo user to watch this movie and to make a reply of "true" when the sentence was judged as correct, or to make a reply of "false" otherwise. We could effectively collected many replies. Table IV shows the statistics of the collected replies. The numbers of replies "true" and "false" to the generated sentences ranked at the first were 1753 and 734 respectively, and the accuracy rate was 0.705. The accuracy rates for generated sentences ranked at the second, third and fourth ranged from 0.542 to 0.612. The generated sentences ranked from fifth to tenth were given low evaluation. These evaluations on the BLEU score and the crowdsourcing demonstrated that correct sentences could be generated from human motions captured in a camera.

V. CONCLUSION

The conclusions we obtained from this research are as follows:

1) We designed a feature of human full-body motion captured by a monocular camera. This feature is defined in a coordinate system with the midpoint between the left hip and right hip being the origin, a line connecting the origin and the left hip being the x-axis, and a line connecting the origin and the neck being the y-axis. An element of the feature vector is a unit vector from the origin in the direction of each key point detected in the camera image. This feature is invariant to translational position of a performer, and his/her body shape.
2) We created a classifier of human motion in a camera image. A sequence of the feature for human full-body motion is encoded into HMM. An observation of human full-body motion is classified into an HMM that is the most likely to generate the observation. It implies that the HMM can be used as a motion classifier.
3) We presented a probabilistic framework to link human full-body motions to language. The framework is established on a semantic function module and a syntactic function module. The semantic function module connects a classified motion to its relevant words. The syntactic function module represents a transition among words in a sentence. The integration of these two modules makes it possible to convert a human full-body motion into a sentence.
4) We tested the sentence generation from a human full-body motion on two datasets. One dataset was created by recording actions performed by a subject in a laboratory with 155 motion types. The other dataset was created by recording actions performed by older adults in care facilities with 21 movement types. Experiments confirmed that approximately 90% and 70% of motions in the laboratory and the care facilities were correctly classified. Additionally, they qualitatively and quantitatively demonstrated that appropriate descriptions of these motions were generated.

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 Interact. 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] I. Mordatch, K. Lowrey, G. Andrew, Z. Popovic, and E. Todorov, "Interactive control of diverse complex characters with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3132–3140.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] 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.
[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
[25] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] Z. Cao, T. Simon, S. E.Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 7291–7299.
[32] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.
[33] 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. Linguist.*, 2002, pp. 311–318.