Autonomous Robots (2019) 43:913–925
https://doi.org/10.1007/s10514-018-9762-1

Linking human motions and objects to language for synthesizing action sentences

Wataru Takano[1] · **Yoshihiko Yamada**[1] · **Yoshihiko Nakamura**[1]

Received: 6 March 2017 / Accepted: 18 April 2018 / Published online: 2 May 2018

Abstract
This paper proposes a novel framework for generating action descriptions from human whole body motions and objects to be manipulated. This generation is based on three modules: the first module categorizes human motions and objects; the second module associates the motion and object categories with words; and the third module extracts a sentence structure as word sequences. Human motions and objects to be manipulated are classified into categories in the first module, then words highly relevant to the motion and object categories are generated from the second module, and finally the words are converted into sentences in the form of word sequences by the third module. The motions and objects along with the relations among the motions, objects, and words are parametrized stochastically by the first and second modules. The sentence structures are parametrized from a dataset of word sequences in a dynamical system by the third module. The link of the stochastic representation of the motions, objects, and words with the dynamical representation of the sentences allows for synthesizing sentences descriptive to human actions. We tested our proposed method on synthesizing action descriptions for a human action dataset captured by an RGB-D sensor, and demonstrated its validity.

Keywords Motion classification · Object classification · Sentence generation

1 Introduction

The demographic trend in advanced countries is that the percentage of elderly people is increasing, even as the total population is shrinking. The availability of comprehensive nursing services to provide living support for elderly people to improve per-capita productivity by covering for labor force shortages are important problems. The use of humanoid robots is expected to address these problems. Because humanoid robots are similar in shape to humans, they can perform human-like actions, and there is no need to adapt the environment from what suits humans. This allows humanoid robots to be used as a replacement labor force for humans for everyday tasks.

Research and development into humanoid robots has been actively pursued in a variety of fields in recent years (Sugano and Kato 1987; Kuroki et al. 2003; Kaneko et al. 2002; Cheng et al. 2007). Although this research has focused on increasing the integration density and accuracy of hardware technology, other elements are essential to constructing intelligent humanoid robots: software for obtaining external information corresponding to the five human senses (sight, sound, touch, taste, and smell), perceiving by using the obtained information, and controlling the motion of the robot. Developments in these areas are expected to bring value for replacement labor force, communication with humans, and information processing systems that exceed the capabilities of humans. Development of not only the hardware but also the software (here, considered as the intelligence of the robots) is an important element that is expected to create new value in robots.

Research into artificial intelligence that is able to perform advanced information processing like a human involves a variety of academic topics beyond robot engineering, including also linguistics, semiotics, anthropology, psychology,

This research was partially supported by the Strategic Information and Communications R&D Promotion Program (No. 142103011) of the Ministry of Internal Affairs and Communications.

✉ Wataru Takano
takano@ynl.t.u-tokyo.ac.jp

Yoshihiko Yamada
yamada@ynl.t.u-tokyo.ac.jp

Yoshihiko Nakamura
nakamura@ynl.t.u-tokyo.ac.jp

[1] Bunkyoku, Hongo, Tokyo, Japan

neuroscience, brain science, and sociology. The key difference between human intelligence and that of other animals is said to be the ability to use language that incorporates advanced symbols. Humans have acquired language through evolutionary processes. For example, in the phrase gread a book,h the concept of a book is expressed by the word gbook.h, and the concept of a particular bodily action is expressed by the word gread.h Phenomena in the real world can be expressed in this way through language. Human intelligence is built on a symbolic system. Because of this, we are able to think specifically about actions, understand abstract concepts, and share the ideas of other people by using symbols. The information processing of symbols and language forms the foundation of the advanced intelligence expressed in the computations of humans.

Ferdinand de Saussure described the composition of language as using "signe" (sign), "signifiant" (signifier), and "signifie" (signified) (Saussure 1966). The signifier is the symbolic representation of the specified item, the signified is the content of the sign that represents the specified concept, and the relation between the signifier and signified gives rise to signs. In the example above, the word "book" is the signifier, and this word signifies the book that exists in the real world. The book itself is the signified. The ability to manipulate signs and real world phenomena is improved and language is developed by arbitrarily forming associations between signs and the real world.

In brain science, Rizzolatti et al. (2001) discovered the existence of a set of neurons (mirror neurons) in the brains of Macaque monkeys that fire when the behavior of another is observed and when movement is performed by oneself. Mirror neurons are related to the generation and recognition of motion, but have also been found in the Brocafs area, which is responsible for the language processing in humans. This implies a relation among the mirror neuron system, the generation and recognition of motions, and the language.

Based on this knowledge, imitation learning models have been proposed in which the robot learns new actions by imitating the actions of humans (Kuniyoshi et al. 1994; Morimoto and Doya 2001; Mataric 2000). Research into constructing intelligence based on encoding bodily motions into symbols has been conducted. Haruno et al. (2001) proposed the module selection and identification for control (MOSAIC) system, which performs environment recognition and action generation in the framework of reinforcement-based learning of multiple learning modules that store different action primitives. Tani and Ito (2003) proposed the recurrent neural network with parametric bias (RNNPB) method, in which multiple action primitives are encoded into bias parameters to be added to a recurrent neural network in which the parameters switch the action primitives. Inamura et al. (2004) proposed a model of encoding motion patterns into hidden Markov models (HMMs). Furthermore, Takano and Nakamura (2015a, b) proposed a model that combines motion symbols characterized by HMMs with natural language, and developed a computation method for creating sentences that represent motions.

However, these motion recognition systems use only bodily motion information such as the three-dimensional position of each part of the body or the time-series data of joint angles, and it is anticipated that these systems will be extended to handle environment (a) for understanding actions in which meaning is imparted to human motion by interactions with the environment, and (b) for generating actions such as manipulation of objects in the environment. For example, in the case of the action "moving hand towards mouth," there is the problem of not being able to understand whether something is held in the hand and, if so, what that object is. This is because object type and position information are not used. Information about object manipulations associated with human actions is important for associating meaning with actions, and the intelligence to understand this is mandatory for humanoid robots that operate in living environments.

In the field of computer vision, the importance of using information from human motion and objects has been noted for understanding the actions that accompany objects (manipulation targets) in everyday life, such as human–object interactions (HOI) (Gupta et al. 2009; Yao and Fei-Fei 2012). For this, it has been noted that the pose of the human is important when detecting and recognizing objects in images. At the same time, manipulated object information is important when recognizing human motions, and the performance of both is expected to be improved by using both object and human-pose information. Information about body motion and information about target object motion are important elements in understanding human actions, so it is expected that performance can be improved by using both types of information. Recently, there have been extensive works on linking natural language description to static images or videos (Li 2011; Kojima et al. 2002). Krishnamoorthy et al (2013). presented an approach to recognizing visual objects and activities, and generating triples of subject, verb and object in two probabilistic vision and natural language scores. Kulkarni et al. (2011) also proposed a probabilistic approach to generating image descriptions (Kulkarni et al. 2011). Objects in an image are detected and their regions are processed for the positional relationship. A conditional random field predicts labels for the image description by incorporating the image potentials and natural language potential. Deep neural networks has demonstrated the rich representation for the image classification, and the neural networks has been actively applied to the image encoder and the description decoder (Vinyals et al. 2015; Karpathy and Fei-Fei 2015). These methods focus on generating sentences describing the images. They don't handle the positions of the performer

and manipulated object in the three dimensional environment, and therefore cannot be directly reused to generate the activities from the description. Action recognition that includes motion recognition, object recognition, and generation of sentences that represent the action can be approached by extending the statistical information processing system proposed by Takano and Nakamura (2015b), which uses three dimensional body motion and language, and developing an information processing mechanism that includes a function for flexible handling of manipulation target (i.e., object information).

This paper proposes a link of human whole body motions, manipulation target objects and language for synthesizing sentence describing human actions. Human motions and the spatial relations between the object and body parts are classified into motion categories, and objects to be manipulated are classified into object categories by motion recognition and manipulation target object recognition, respectively. Including the spatial relation between an object and body parts as part of the action information allows identification of motions that could not be identified from body motion information alone. This ability extends beyond that of conventional methods. The spatial relation is defined by the distances and relative positions of nodes, in a manner similar to an interaction mesh (Ho et al. 2010; Ho and Shum 2013). The mesh is used in motion synthesis with adaption for objects and obstacles in the environment and to detect objects in the environment. Object segmentation is derived from color and depth information in point cloud data, and object identification is subsequently performed using the extracted image information. We additionally construct a statistical model (the motion object language model) that learns the relations that connect motions, objects, and a sentence. The sentence represents the action by a recurrent neural network (natural language model) that learns the order of words in the sentence as a dynamical system. Words to reflect the body motion and object information are identified by applying motion recognition, object recognition, and the motion object language model, after which sentences are generated by reordering the words according to the natural language model. This is done with the aim of more correctly understanding human actions by using multimodal information comprising body motion information, such as three-dimensional position information of each body part and time-series data of joint angles, and the positions and types of objects in the environment with descriptive sentences representing the action.

2 Motion and object primitives

A human action consists of a human whole body motion and an object to be manipulated. The classifications of the human whole body motion and the object are required to generate sentences describing the human action. This section describes the representations of the human motion and the object, and their classifiers.

2.1 Human whole body primitives

Action recognition is a highly competitive field, and many approaches that handle human body motions and objects to be manipulated have been reported. Wang et al. (2012) used a feature called "local occupancy pattern" in which elements represent the area occupied by an object around each joint, and they defined the feature for action recognition by combining the local occupancy pattern with the positional relations between two joints over the set of all joint pairs. The temporal sequences of these features were converted to a Fourier temporal pyramid and classified into the relevant action category by using a support vector machine (SVM). The method proposed by Yu et al. (2014) is similar to that of Wang et al., but they used the distances between two joints in the whole body as features. SVM-based methods adopt a discriminative approach to classification that generally outperforms the generative approach typified by hidden Markov models (HMMs), but they cannot recover the human motions, such as a sequence of joint positions. In this paper, we use HMMs to encode the human motions to create motion primitives because HMMs can be used for action recognition and generation of human-like motion for robots.

Feature $\boldsymbol{x}$ of human motion (Fig. 1) consists of two elements: position $\boldsymbol{p}_i$ of the ith joint in the trunk coordinate system, and distance d_i between the ith joint and the manipulated object. $\boldsymbol{p}_i$ and d_i are concatenated into $\boldsymbol{x}$ over all joints. The human motion is expressed by a sequence of these features, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$.

Human motions are encoded into a set of parameters to characterize HMMs. HMMs are generative models optimized such that the likelihood of the human motion $\boldsymbol{x}$ being

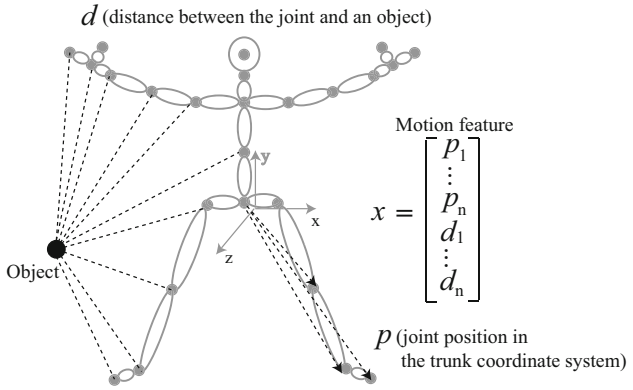


Fig. 1 The motion feature is expressed by the vector $\boldsymbol{x}^T$ whose elements are joint positions in the trunk coordinate system or distances between the joints and the manipulated object

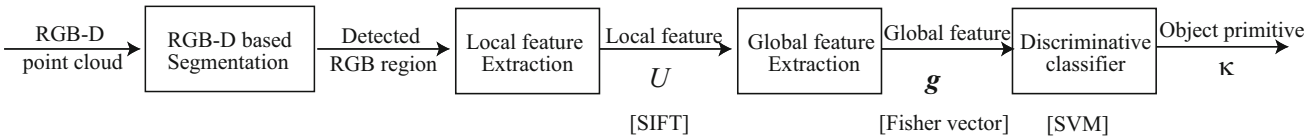


Fig. 2 Point cloud data is segmented into object regions, local features are computed in each object region, the global feature is extracted from the computed local features, and the global feature is then classified into an object primitive

generated from HMM λ is maximized. The parameters to be optimized are a vector $\boldsymbol{\Pi}$ whose entries π_i are (for each i) the probability of starting at the ith node, the matrix $\boldsymbol{A}$ whose entries a_{ij} are the probabilities of transition from the ith node to the jth node, and the output distribution $\boldsymbol{B}(\boldsymbol{x})$ whose entries $b_i(\boldsymbol{x})$ are the probabilities of $\boldsymbol{x}$ being generated from the ith node. The Baum–Welch algorithm can optimize these parameters (Rabiner 1989). Moreover, HMMs can be used to classify human motions $\boldsymbol{x}$ into the specific HMM $\lambda_{\mathcal{R}}$ that is the most likely to generate $\boldsymbol{x}$.

$$\lambda_{\mathcal{R}} = \arg\max_{\lambda} P(\boldsymbol{x}|\lambda) \tag{1}$$

2.2 Object primitives

An object to be manipulated is captured by an RGB-D camera. The image and depth data are segmented into an object region, scale-invariant feature transform (SIFT) descriptors are computed for the local feature in the object region, and the Fisher vector descriptor is extracted from the computed local features as the global feature, which is classified into an object primitive. Figure 2 shows the pipeline to convert captured RGB-D data into the corresponding object primitive.

The method of region growing and region merging partitions RGB-D data into object regions (Zhan et al. 2009). The region growing process randomly selects an ungrouped point, and then groups it with all ungrouped points closer than a manually given threshold. This process is iterated until all points are grouped into one of the regions. The region merging process finds two regions that are close to each other and aggregates them into one region. The merging process results in the object region.

The segmented object region is processed to extra the local features of the objects contained in the region. SIFT descriptors are adopted for the local features because they are colored and scale-invariant (Lowe 1999; Abdel-Kalim et al. 2006). The SIFT descriptors represent only local patches in the segmented region. The Fisher vector is introduced as a global feature to represent the entire region. The derivation of the Fisher vector is described in the "Appendix". The Fisher vector is classified into its relevant object primitive by using the SVM technique (Cortes and Vapnik 1995). These processes together convert the captured point cloud data into object primitives.

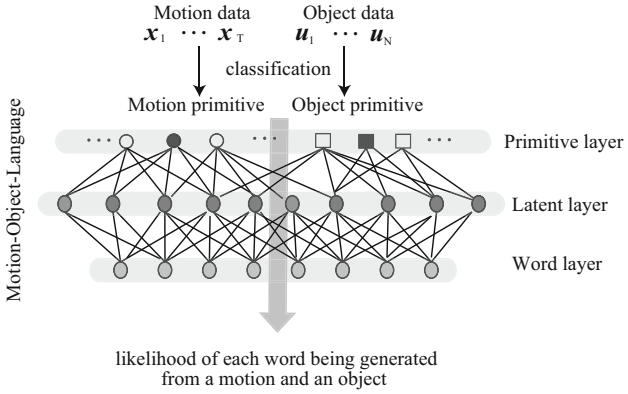


Fig. 3 Human whole body motion and RGB-D images are classified into the motion primitive and object primitive. These primitives are connected to their relevant words stochastically. The probabilities of the latent node being generated from the motion and object primitives and the probabilities of the word from the latent node are optimized such that the words related to the action are the most likely to be generated from the motion and object to be manipulated

3 Connection between human actions and description

Our framework to generate the sentences from human actions consists of two modules. The first module combines the human whole body motions and the manipulated objects with their relevant words. The second module models sequences of the words in the sentences. This section describes these modules in details.

3.1 Stochastic model of words from motions and objects

The motion primitives and object primitives, which are derived by classifying human whole body motions and images containing an object to be manipulated, are connected to their relevant words stochastically. Figure 3 shows the stochastic model for the connections. This stochastic model is made of three layers: the top layer contains the primitives, the bottom layer contains the words, and the middle layer contains the latent nodes. The latent nodes connect the motions and the objects to the words. The connectivities are characterized by the conditional probability $P(s|\lambda, \kappa)$ of latent node s being generated from motion primitive λ and object primitive κ and the probability $P(\omega|s)$ of word ω being gen-

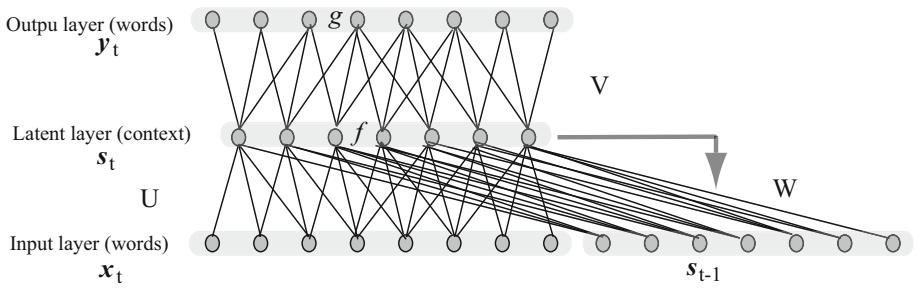


Fig. 4 A recurrent neural network consists of the input, latent, and output layers. The latent layer retains the dynamics of word sequences in sentences, and this neural network can predict a word following the input word sequence

erated from latent node s. These probabilities are optimized by expectation maximization, iterating the E-step and M-step so that the training dataset of the motions, objects, and words is the most likely to be generated from this model. The training dataset $\left\{\lambda^{(i)}, \kappa^{(i)}, \omega_1^{(i)}, \ldots, \omega_{n_i}^{(i)}\right\}$ is given; in this, $\omega_j^{(i)}$ is the jth word in the word sequence (sentence) that is manually attached to the ith action whose whole body motion is classified into the motion primitive $\lambda^{(i)}$, and the manipulated object is classified into the object primitive $\kappa^{(i)}$. The E-step estimates the distribution for latent node s conditioned on motion primitive λ, object primitive κ, and word ω as

$$P(s|\lambda, \kappa, \omega) = \frac{P(\omega|s)P(s|\lambda, \kappa)}{\sum_k P(\omega|s_k)P(s_k|\lambda, \kappa).} \tag{2}$$

The M-step updates probabilities $P(s|\lambda, \kappa)$ and $P(\omega|s)$ to

$$P(\omega_i|s) = \frac{\sum_{i,j} n(\lambda_i, \kappa_j, \omega)P(s|\lambda_i, \kappa_j, \omega)}{\sum_{i,j,k} n(\lambda_i, \kappa_j, \omega_k)P(s|\lambda_i, \kappa_j, \omega_k)} \tag{3}$$

$$P(s|\lambda, \kappa) = \frac{\sum_i n(\lambda, \kappa, \omega_i)P(s|\lambda, \kappa, \omega_i)}{\sum_i n(\lambda, \kappa, \omega_i),} \tag{4}$$

where $n(\lambda, \kappa, \omega)$ is a function that counts the number of words ω attached to the actions for which the whole body motions are classified into motion primitive λ and the objects to be manipulated are classified into object primitive κ. Alternating the E-step and M-step results in the optimal probabilities for $P(s|\lambda, \kappa)$ and $P(\omega|s)$. The deviation of the EM algorithm is described in the "Appendix".

3.2 Recurrent neural network for action descriptions

Neural networks have been widely used for modeling sentences (Bengio et al. 2006), and have been extended to recurrent neural networks to handle the dynamics of word sequences in sentences (Mikolov et al. 2010). Recurrent neural networks predict words that follow the input words via latent layers that can handle context in sentences. Figure 4 shows a recurrent neural network that consists of input, latent, and output layers. The input and output layers comprise word nodes. The number of nodes in the input and output layers is the same for each layer as the number of words that can appear in the sentences. The input layer is connected to the output layer through latent nodes, which represent the current state and retain the previous state. Specifically, the input vector is $\boldsymbol{x}_t \in R^{N_\omega}$ and the output vector is $\boldsymbol{y}_t \in R^{N_\omega}$, where N_ω is the number of distinct words. The activities of the latent node are expressed by $\boldsymbol{z}_t \in R^{N_z}$ for current activities and $\boldsymbol{z}_{t-1} \in R^{N_z}$ for past activities, where N_z is the number of nodes in the latent layer. If the kth word is given for the input, $\boldsymbol{x}_t$ is set to the binary vector

$$x_i = \begin{cases} 1 & if \ \ i = k \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where x_i is the ith element in $\boldsymbol{x}_t$. $\boldsymbol{z}_t$ is computed from $\boldsymbol{x}_t$ as

$$\tilde{\boldsymbol{z}}_t = \boldsymbol{U}\boldsymbol{x}_t + \boldsymbol{W}\boldsymbol{z}_{t-1} \tag{6}$$

$$z_i = f(\tilde{z}_i), \tag{7}$$

where z_i and $\tilde{z}_i$ are the ith elements in $\boldsymbol{z}_t$ and $\tilde{\boldsymbol{z}}_t$, respectively, $\boldsymbol{U} \in R^{N_z \times N_\omega}$ and $\boldsymbol{W} \in R^{N_z \times N_z}$ are weight matrices, and $f(z)$ is a sigmoid function. $\boldsymbol{y}_t$ is computed from $\boldsymbol{z}_t$ in a similar manner,

$$\tilde{\boldsymbol{y}}_t = \boldsymbol{V}\boldsymbol{z}_t \tag{8}$$

$$y_i = g_i(\tilde{\boldsymbol{y}}_t), \tag{9}$$

where y_i and $\tilde{y}_i$ are the ith elements in $\boldsymbol{y}_t$ and $\tilde{\boldsymbol{y}}_t$, respectively, $\boldsymbol{V} \in R^{N_\omega \times N_z}$ is a weight matrix, and $g_i(\tilde{\boldsymbol{y}}_t)$ is the following function. In this function, y_i represents the probability of the ith word being generated from the input word sequence.

$$g_i(\tilde{\boldsymbol{y}}_t) = \frac{\exp(\tilde{y}_i)}{\sum_k \exp(\tilde{y}_k)} \tag{10}$$

The weight matrices, $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{W}$, are trained by back propagation through time; this method incrementally updates the weight parameters to reduce the errors between the output vectors $\boldsymbol{y}_t$ and the correct vectors $\boldsymbol{d}_t$. Weight matrix $\boldsymbol{V}$ is tuned as

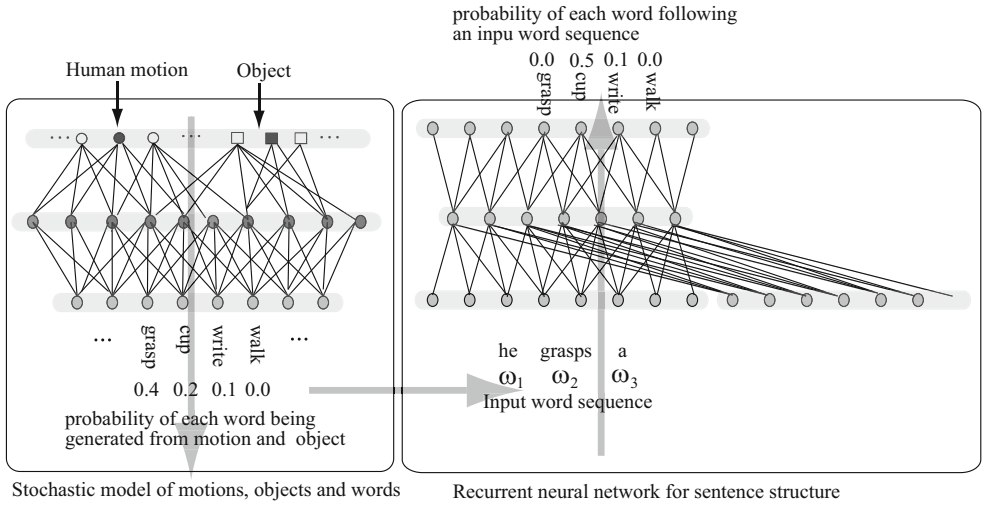


Fig. 5 A module on the left panel stochastically extracts the relation among human motions, objects, and words. Another module on the right panel extracts the dynamics of word sequences. The link between these two modules allows the synthesis of sentences describing human actions

$$\boldsymbol{e}_t = \boldsymbol{d}_t - \boldsymbol{y}_t \tag{11}$$

$$\boldsymbol{V}_{t+1} = \boldsymbol{V}_t + \alpha \boldsymbol{s}_t \boldsymbol{e}_t{}^T. \tag{12}$$

The errors $\tilde{\boldsymbol{e}}$ are propagated from the output layer to the latent layer.

$$\tilde{e}_{it} = h_i(\boldsymbol{e}_t{}^T \boldsymbol{V}, t) \tag{13}$$

$$h_i(x, t) = x s_{it}(1 - s_{it}) \tag{14}$$

$\tilde{e}_{it}$ is the ith element of $\tilde{\boldsymbol{e}}_t$, and $s_i t$ is the ith element of $\boldsymbol{s}_t$. The weight matrices $\boldsymbol{U}$ and $\boldsymbol{W}$ are updated by using the error $\tilde{\boldsymbol{e}}$.

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t + \beta \boldsymbol{x}_t \tilde{\boldsymbol{e}}_t^T \tag{15}$$

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t + \gamma \boldsymbol{s}_{t-1} \tilde{\boldsymbol{e}}_t^T \tag{16}$$

α, β, and γ are learning rates; these have been set to decrease monotonically, following Bergstra and Bengio (2012).

3.3 Generation of action descriptions from motions and objects

Integrating the two modules described above allows the generation of sentences describing human actions. Figure 5 shows an overview of this integration. An observation containing a human motion and an object is classified into a motion primitive and an object primitive. The words relevant to the motion and object are associated by the stochastic model, and these words are arranged into a sentence according to the recurrent neural network. Specifically, given motion primitive $\lambda_{\mathcal{R}}$ and object primitive $\kappa_{\mathcal{R}}$, the pair of primitives is converted to a sentence that is the most likely to be generated from these primitives. The sentence can be formed by searching for a sequence of words according to the probability of the sentence being generated from two modules given the motion primitive and the object primitive as

$$P(\boldsymbol{\omega}|\lambda_{\mathcal{R}}, \kappa_{\mathcal{R}}) = \prod_{i=1}^{l} P(\omega_i|\lambda_{\mathcal{R}}, \kappa_{\mathcal{R}}) \prod_{i=1}^{l-1} P(\omega_{i+1}|\omega_1, \ldots, \omega_i). \tag{17}$$

Here, sentence $\boldsymbol{\omega}$ is expressed by word sequence $\omega_1, \omega_2, \ldots, \omega_l$. We assume that a set of words contained in the sentence depends on only the motion primitive and the object primitive, and that a sequence of words depends on only the set of words. The probability of each word being generated from the motion primitive and object primitive can be computed by using Eqs. 3 and 4.

$$P(\omega|\lambda_{\mathcal{R}}, \kappa_{\mathcal{R}}) = \sum_{s} P(\omega|s) P(s|\lambda_{\mathcal{R}}, \kappa_{\mathcal{R}}) \tag{18}$$

The probability of word sequence $\omega_1, \omega_2, \ldots, \omega_i$ being followed by word ω_{i+1} can be computed by the recurrent neural network. Taking the logarithm of $P(\boldsymbol{\omega}|\lambda_{\mathcal{R}}, \kappa_{\mathcal{R}})$ and using Dijkstra's algorithm, we can search for the sentence that is the most likely to be generated from the motion and object.

4 Experiments

Our proposed approach is tested to see how well it generates descriptions from observations of human actions. The observation data are collected by using an RGB-D sensor (Kinect, Microsoft Corporation) and contain human whole body motions and objects to be manipulated. The RGB-D sensor measures the positions of 25 joints in the whole

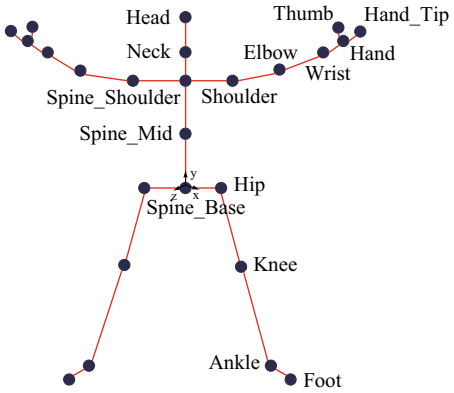


Fig. 6 The positions of 25 joints are measured by the RGB data. These data are fitted to a human character with 34 degrees of freedom, to which 35 markers are attached. The motion by the human character is encoded into the motion primitive

body, as shown in Fig. 6. These positions are converted to the positions of 34 markers that are attached to a human character with 34 degrees of freedom. The attachment follows the Helen Hayes marker set placement (Kadaba et al. 1990). The positions of the 34 markers in the character's trunk coordinate system and the distances between these markers and an object to be manipulated together express the human whole body motion to be used for motion primitives. An image from the RGB-D sensor is segmented into an object region. The local features are extracted from the object region, and then global features are computed for the object primitives. We measure actions by three performers, and 320 observations are collected for each of these three performers, giving 960 observations in total. Motion and object data contained in these datasets can be grouped into 24 motion primitives and 30 object primitives. Additionally, five students attached one sentence descriptive of each action. Figure 7 shows several samples of the action and the manually attached descriptions. The dataset contains 960 sentences, with 335 different words. The dataset is grouped into a training dataset containing 576 actions and a test dataset containing 384 actions.

Figure 8 qualitatively shows the experimental results. Three sentences that are most likely to be generated from each observation are displayed. The observation of "blowing the nose" is described by three sentences: "a person blows their nose with a tissue", "a person blows their nose with tissue paper", and "a person is blowing their nose". The observation of "sweeping" is expressed by sentences "a person is sweeping the floor with a broom", "they are cleaning the floor" and "they clean the floor with a broom". The observation of "picking" can be described as sentences: "a person picks up the box on the bottle" "they pick up the box on the bottle" and "a person picks up a box". The first sentence is same as the training sentence as shown in Fig. 7. The observation of "drinking" generates the sentences "a person drinks a bottle of tea", "a person is drinking a bottle of tea" and "a person drinks out of a bottle"; these are similar to sentences attached to the action of "drinking", as shown in Fig. 7. Other observations are also described by qualitatively correct sentences.

We also quantitatively test the sentence generation. In the first test, up to five sentences are generated from each test observation. When the generated sentence is the same as the sentence attached to the test observation, this sentence is counted as correct. This is the 5-best condition; more generally, for the m-best condition, the number of generated sentences is set to m, and if any of the generated sentences is correct, the generation is counted as correct. The correct ratios are 0.71, 0.84, 0.89, 0.91, and 0.92 for the 1-best, 2-best, 3-best, 4-best, and 5-best conditions, respectively.

It is important to evaluate the improvement by adding the object to be manipulated for the sentence generation. We removes the layer of the object primitives from the module as shown in Fig. 3. More specifically, we tested the sentence generation only from the human motion. The correct ratios of the sentence generation are 0.54, 0.54, 0.55, 0.59, and 0.59 for the 1-best, 2-best, 3-best, 4-best, and 5-best conditions, respectively. The comparison of these correct ratios with those derived in our proposed framework demonstrates that the information of the objects is effectively used to generate sentences from the action observations. Additionally, 9944 sentences with 1174 different words attached to the human action are crowdsourced. After training 576 human actions and 5984 sentences attached to these actions, we tested the sentence generation. Multiple sentences are attached to each human action. When the generated sentence is the same as one of the sentences attached to the test observation, this

Motion primitive : type
Object primitive : laptop
Description:
a person types on a laptop.
he types on a computer.
he uses a laptop.
a man types on a keyboard
a person is typing.

Motion primitive : drink
Object primitive : mug
Description:
a person drinks from a cup.
he drinks a cup of water all in one go.
a man drinks a cup of tea.
a man is drinking a mug of coffee.
a person is drinking a mug of coffee.

Motion primitive : pick up
Object primitive : box
Description:
a person picks up a box on the table.
he takes a box in his hand.
a man takes a sweets box.
a man is picking up a box.
he is picking up a blue box on the table.

Fig. 7 The datasets contain human whole body motions, objects and sentences describing human actions. The actions in the left panel consist of the motion primitive "type", the motion primitive "laptop" and descriptions "a person types on a laptop"

[1] a person blows his nose with a tissue
[2] a person blows his nose with a tissue paper
[3] a person is blowing his nose

[1] a person is sweeping the floor with a broom
[2] he is cleaning the floor
[3] he cleans a floor with a broom

[1] a person drinks a bottle of tea
[2] a person is drinking a bottle of tea
[3] a person drinks out of a bottle

[1] a person eats something with holding a cup
[2] a person eats
[3] a person eats something with holding a black bowl

[1] a person is holding a bucket
[2] he is holding a bucket
[3] he is standing with a bucket on his right hand

[1] he is fanning himself
[2] he is waving a fan
[3] he fans himself

[1] he opens a gum bottle and picks a gum
[2] he opens a gum bottle and picks up a gum
[3] a person opens a gum bottle and picks a gum

[1] a student opens a pot noodle
[2] he opens a cover of a pot noodle
[3] a person opens a cover of cup noodle

[1] a person picks up a box on the table
[2] he picks up a box on the table
[3] a person picks up a box

[1] a person pours a dollop of shampoo into his hand
[2] he takes a push of shampoo
[3] a person pours a dollop of shampoo into his left hand

[1] a person wears a cap
[2] a person pulls a cap over his head
[3] a student puts on a cap

[1] he puts on brown slippers
[2] a student puts on the slippers
[3] he puts on his slippers

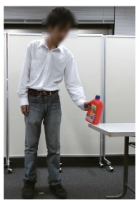
[1] a person puts a coin into a box
[2] a person puts a coin in a piggy bank
[3] he puts a coin nto a box

[1] a student puts a white box on the table
[2] he puts a white box on the table
[3] a person puts a white box on the table

[1] he is reading a yellow book
[2] a person is reading a yellow book
[3] he is reading a book

[1] he is shaking the maracas
[2] a person plays the maracas
[3] he is playing maracas

[1] he opens a sauce bottle and pours sauce
[2] a person opens sauce bottle and pours sauce
[3] a person pours sauce



[1] he picks up a detergent container
[2] a person picks up a detergent container
[3] he is picking up a detergent

[1] he is spraying deodrant
[2] a person sprays air freshener
[3] a man is spraying air refreshener

[1] he picks up trash with tongs
[2] he picks up a trash and put it in the trash can
[3] he picks up a wastepaper and puts into a dust bin

[1] he types at a keyboard
[2] he is typing the keyboard on the desk
[3] he types at the keyboard

[1] he is using a tablet terminal
[2] a person is using a tablet terminal
[3] a student is using a tablet terminal

[1] a person is wahsing a dish
[2] a person is washing a bowl
[3] a person is washing a dish with a sponge

[1] he is writing on a blue notebook with a red marker
[2] he is writing on a blue notebook with a red pen
[3] a student is writing on a blue notebook with a red marker

Fig. 8 Observation is classified into its relevant motion primitive and object primitive, and a pair of these primitives is converted to sentences describing the observation. The three mostly likely sentences from each observation are displayed. For example, the observation "blowing the nose" is described by the sentences "a person blows their nose with a tissue", "a person blows their nose with tissue paper", "a person is blowing their nose"

generation is counted as correct. The correct ratios of the sentence generation are 0.67, 0.69, 0.78, 0.84, and 0.90 for the 1-best, 2-best, 3-best, 4-best, and 5-best conditions, respectively.

5 Conclusions

This research is summarized as follows.

1. We proposed a framework for linking human actions (consisting of human whole body motions and objects to be manipulated) with sentences describing the actions. For this, the human whole body motion and positional relation between the body and the object are encoded into a motion primitive; also, an object feature is extracted from an object region in a captured image and is then encoded into an object primitive. A pair of motion primitive and object primitive is stochastically connected to words relevant to the action. Additionally, the dynamics of word sequences in sentences descriptive of the actions is trained by a recurrent neural network, which can predict that word that is likely to follow a sequence of words.
2. We linked two modules: a stochastic model between motions, objects, and words; and a recurrent neural network for the sentence structure. The link makes it possible to search for the sentences that are most likely to be generated from the observation of human action. Specifically, the recurrent neural network efficiently generates the sentence whose words are most likely to be generated from a given motion primitive and object primitive in the stochastic model. This link implies that the observation of the human action can be interpreted by considering corresponding descriptive sentences.
3. We constructed a stochastic model to describe the relations between motions, objects, and words, and a recurrent neural network to describe the sentence structures. Each was trained against a dataset containing 576 pieces of data for triples comprising a human motion, an object, and a (manually chosen) sentence. We conducted an experiment with the stochastic model and the neural network, generating sentences for 384 test observations. We qualitatively and quantitatively confirmed that our proposed method can generate correct sentences from observations of human actions.

Acknowledgements This research was partially supported by a Grant-in-Aid for Young Scientists (A) (No. 26700021) and Challenging Research (Exploratory) (No. 17K20000) from the Japan Society for the Promotion of Science.

Appendix A

The distribution of the local features in the training dataset is assumed to be expressed by a Gaussian mixture model (GMM) κ. The GMM consists of three kinds of parameters: mean vector $\boldsymbol{\mu}_i$, covariance matrix $\boldsymbol{\Sigma}_i$, and weight parameter w_i for the ith Gaussian distribution. The likelihood of local feature $\boldsymbol{u}$ being generated from GMM κ is written as

$$P(\boldsymbol{u}|\kappa) = \sum_{i=1}^{K} P(\boldsymbol{u}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, w_i) \tag{19}$$

$$P(\boldsymbol{u}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, w_i) = \frac{w_i}{\sqrt{(2\pi)^d ||\boldsymbol{\Sigma}_i||}} \times \exp\left\{-\frac{1}{2}\left(\boldsymbol{u}-\boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{u}-\boldsymbol{\mu}_i\right)\right\}, \tag{20}$$

where d is the number of dimensions of $\boldsymbol{u}$ and the K is the number of Gaussian distributions. Note that w_1 can be removed from the parameter set because of the constraint

$$\sum_{i=1}^{K} w_i = 1 \tag{21}$$

When N local descriptors, $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N$, are found in an object region, the likelihood of these local descriptors being generated from the GMM is calculated as

$$P(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N|\kappa) = \sum_{i=1}^{N} P(\boldsymbol{u}|\kappa). \tag{22}$$

The Fisher vector, $\boldsymbol{v}$, is defined as the gradient vector of the log-likelihood of $P(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N|\kappa)$:

$$\boldsymbol{v} = \boldsymbol{F}^{-\frac{1}{2}} \nabla_\kappa P(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N|\kappa), \tag{23}$$

where the elements of the gradient vector are

$$\frac{\partial P(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N|\kappa)}{\partial w_k} = \sum_{i=1}^{N} \left(\frac{\gamma_k(\boldsymbol{u}_i)}{w_k} - \frac{\gamma_k(\boldsymbol{u}_1)}{w_1} \right) \quad (24)$$

$$\frac{\partial P(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N|\kappa)}{\partial \boldsymbol{\mu}_k} = \sum_{i=1}^{N} \gamma_k(\boldsymbol{u}_i) \left(\boldsymbol{u}_i - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \quad (25)$$

$$\frac{\partial P(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N|\kappa)}{\partial \boldsymbol{\Sigma}_k}$$

$$= \sum_{i=1}^{N} \frac{1}{2} \gamma_k(\boldsymbol{u}_i) \left[\mathrm{Tr} \left\{ \left((\boldsymbol{u}_i - (\boldsymbol{\mu}_k) \left((\boldsymbol{u}_i - (\boldsymbol{\mu}_k \right)^T \right\} \times \boldsymbol{\Sigma}_k^{-2} - \boldsymbol{\Sigma}_k^{-1} \right]. \quad (26)$$

Note that $\boldsymbol{\Sigma}_k$ is assumed to be a diagonal matrix, and that $\gamma_k(\boldsymbol{u})$ is written as

$$\gamma_k(\boldsymbol{u}) = \frac{w_k P(\boldsymbol{u}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, w_k)}{P(\boldsymbol{u}|\kappa).} \quad (27)$$

Here, $\boldsymbol{F}$ is the Fisher information matrix, and is defined as

$$\boldsymbol{F} = E_{P(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N|\kappa)} \times \left[\nabla_\kappa P(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N|\kappa) \nabla_\kappa P(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N|\kappa)^T \right], \quad (28)$$

where $E[*]$ is the expectation value.

Appendix B

Human actions contain data about human whole body motions and objects. The whole body motion, including the relative position of the object to be manipulated, is encoded into motion primitive λ, and the image containing the object is encoded into object primitive κ. The sentence describing the human action is manually assigned to the action. The sentence is expressed by a sequence of words, ω. Let the training dataset be $\left\{ \lambda^{(i)}, \kappa^{(i)}, \omega_1^{(i)}, \ldots, \omega_{n_i}^{(i)} \right\}$, consisting of motion primitives, object primitives, and sentences. Then, the logarithm of the probability of the words $\omega_1^{(i)}, \ldots, \omega_{n_i}^{(i)}$ in the sentence being generated from motion primitive $\lambda^{(i)}$ and object primitive $\kappa^{(i)}$ over the training dataset is written as

$$\Phi = \sum_i \ln P\left(\omega_1^{(i)}, \ldots, \omega_{n_i}^{(i)}|\lambda^{(i)}, \kappa^{(i)}\right) \quad (29)$$

$$= \sum_{i,j} \ln P\left(\omega_j^{(i)}|\lambda^{(i)}, \kappa^{(i)}\right), \quad (30)$$

where we assume that the word depends on only the motion and the object. According to the marginal distribution of the latent node, Eq. 29 is rewritten as

$$\Phi = \sum_{i,j} \ln \sum_k P\left(\omega_j^{(i)}, s_k|\lambda^{(i)}, \kappa^{(i)}\right). \quad (31)$$

This equation can be written as the expectation

$$\Phi = \sum_{i,j} \ln \sum_k \tilde{P}\left(s_k|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right) \frac{P\left(\omega_j^{(i)}, s_k|\lambda^{(i)}, \kappa^{(i)}\right)}{\tilde{P}\left(s_k|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \quad (32)$$

$$= \sum_{i,j} \ln E_{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \left[\frac{P\left(\omega_j^{(i)}, s|\lambda^{(i)}, \kappa^{(i)}\right)}{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \right]. \quad (33)$$

The lower limit of this expectation, Φ_L, is given by the Jensen inequality.

$$\Phi_L = \sum_{i,j} E_{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \left[\ln \frac{P\left(\omega_j^{(i)}, s|\lambda^{(i)}, \kappa^{(i)}\right)}{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \right] \quad (34)$$

From Φ and Φ_L, the following equations are derived.

$$\Phi - \Phi_L = \sum_{i,j} \left\{ \ln P\left(\omega_j^{(i)}|\lambda^{(i)}, \kappa^{(i)}\right) - E_{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \left[\ln \frac{P\left(\omega_j^{(i)}, s|\lambda^{(i)}, \kappa^{(i)}\right)}{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \right] \right\}$$

$$= \sum_{i,j} \left\{ \ln P\left(\omega_j^{(i)}|\lambda^{(i)}, \kappa^{(i)}\right) - E_{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \times \left[\ln \frac{P\left(\omega_j^{(i)}|\lambda^{(i)}, \kappa^{(i)}\right) P\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)}{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \right] \right\}$$

$$= \sum_{i,j} E_{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \left[\ln \frac{\tilde{P}\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)}{P\left(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\right)} \right] \quad (35)$$

Here, $P(s|\lambda, \kappa, \omega)$ is the estimated distribution of latent node s based on the model, and $\tilde{P}(s|\lambda, \kappa, \omega)$ is the true distribution of s. Equation 34 implies the Kullback–Leibler information $KL\left(\tilde{P}(s|\lambda, \kappa, \omega)||P(s|\lambda, \kappa, \omega)\right)$ between these two distributions. The Kullback–Leibler information is nonnegative, and is zero only when these two distributions are the same. Therefore, estimating the true distribution, $\tilde{P}(s|\lambda, \kappa, \omega)$, as the model-based distribution, $P(s|\lambda, \kappa, \omega)$, yields zero for the Kullback–Leibler information. The E-step estimates the distribution $\tilde{P}(s|\lambda, \kappa, \omega)$ of the latent node such that the Kullback–Leibler information becomes zero.

After the E-step, the model parameters are iteratively optimized such that Φ increases incrementally. Let $\Phi^{[t+1]}$ and $\Phi^{[t]}$ be the objective functions given at the $(t+1)$th and tth

iteration steps, respectively. The relation between $\Phi^{[t+1]}$ and $\Phi^{[t]}$ is

$$\begin{aligned}\Phi^{[t+1]} - \Phi^{[t]} &= \Phi_L^{[t+1]} + \sum_{i,j} KL\big(\tilde{P}\big(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\big)||P^{[t+1]} \\ &\big(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\big)\big) - \Phi_L^{[t]} \\ &- \sum i, jKL\big(\tilde{P}\big(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\big)||P^{[t]}\big(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\big)\big),\end{aligned} \tag{36}$$

where $\Phi_L^{[t+1]}$ and $\Phi_L^{[t]}$ are the lower limits of $\Phi^{[t+1]}$ and $\Phi^{[t]}$, respectively, and $P^{[t+1]}(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)})$, and $P^{[t]}(s|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)})$ are the estimated distributions of the latent node based on the models derived at the $(t+1)$th and tth iteration steps, again respectively. Estimation of the true distribution $\tilde{P}(s|\lambda, \kappa, \omega)$ of the latent node as the distribution $P^{[t]}(s|\lambda, \kappa, \omega)$ based on the model derived at the tth iteration step leads to the relation

$$\Phi^{[t+1]} - \Phi^{[t]} \geq \Phi_L^{[t+1]} - \Phi_L^{[t]} \tag{37}$$

because the second and fourth terms in Eq.36 take positive and zero values, respectively. The search for the parameters that maximize $\Phi_L^{[t+1]}$ at the $(t+1)$th iteration step results in $\Phi^{[t+1]}$ becoming larger than $\Phi^{[t]}$. By assigning the distribution $P^{[t]}(s|\lambda, \kappa, \omega)$ to $\tilde{P}(s|\lambda, \kappa, \omega)$ in Eq. 34, $\Phi_L^{[t+1]}$ is rewritten as

$$\Phi_L^{[t+1]} = \sum_{i,j,k} P^{[t]}\big(s_k|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\big) \ln \frac{P\big(\omega_j^{(i)}, s_k|\lambda^{(i)}, \kappa^{(i)}\big)}{P^{[t]}\big(s_k|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\big)}. \tag{38}$$

Ignoring terms that do not depend on the model parameters, the function $\hat{Phi}^{[t+1]}$ to be maximized can be written as

$$\begin{aligned}\hat{Phi}^{[t+1]} &= \sum_{i,j,k} P^{[t]}\big(s_k|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\big) \\ &\quad \ln P\big(\omega_j^{(i)}, s_k|\lambda^{(i)}, \kappa^{(i)}\big) \\ &= \sum_{i,j,k} P^{[t]}\big(s_k|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\big) \\ &\quad \Big[\ln P\big(\omega_j^{(i)|s_k}\big) + \ln P\big(s_k|\lambda^{(i)}, \kappa^{(i)}\big)\Big].\end{aligned} \tag{39}$$

The model parameters $P(\omega|s)$ and $P(s|\lambda, \kappa)$ to be optimized must satisfy the following constraints

$$\sum_i P(\omega_i|s) = 1 \tag{40}$$

$$\sum_k P(s_k|\lambda, \kappa) = 1. \tag{41}$$

The Lagrange function is obtained as

$$\begin{aligned}\mathcal{L} = &\sum_{i,j,k} P^{[t]}\big(s_k|\lambda^{(i)}, \kappa^{(i)}, \omega_j^{(i)}\big) \\ &\Big[\ln P\big(\omega_j^{(i)|s_k}\big) + \ln P\big(s_k|\lambda^{(i)}, \kappa^{(i)}\big)\Big] \\ &- \sum_k \alpha_k \left[\sum_i P(\omega_i|s_k) - 1\right] \\ &- \sum_{i,j} \beta_{ij} \left[\sum_k P(s_k|\lambda_i, \kappa_j) - 1\right].\end{aligned} \tag{42}$$

The derivative of the Lagrange function with respect to $P(\omega_i|s)$ or $P(s|\lambda, \kappa)$ is zero at the optimal parameter, which is derived as

$$P^{[t+1]}(\omega_i|s) = \frac{\sum_{i,j} n(\lambda_i, \kappa_j, \omega) P^{[t]}(s|\lambda_i, \kappa_j, \omega)}{\sum_{i,j,k} n(\lambda_i, \kappa_j, \omega_k) P^{[t]}(s|\lambda_i, \kappa_j, \omega_k)} \tag{43}$$

$$P^{[t+1]}(s|\lambda, \kappa) = \frac{\sum_i n(\lambda, \kappa, \omega_i) P^{[t]}(s|\lambda, \kappa, \omega_i)}{\sum_i n(\lambda, \kappa, \omega_i)} \tag{44}$$

The M-step searches for the optimal parameters in this manner, and the EM algorithm alternates the E-step and the M-step.

References

Abdel-Kalim, A. E., & Farag, A. A. (2006). Asift: A sift descriptor with color invariant characteristics. In *Proceedings of 2006 IEEE Computer society conference on computer vision and pattern recognition*, Vol. 2, pp. 1978–1983.

Bengio, Y., Schwenk, H., Senécal, J., Morin, F., & Gauvain, J. (2006). Neural probabilistic language models. In *Innovations in machine learning*, pp. 137–186.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, *13*(1), 281–305.

Cheng, G., Hyon, S. H., Morimoto, J., Ude, A., Hale, J. G., Colvin, G., et al. (2007). CB: A humanoid research platform for exploring neuroscience. *Advanced Robotics*, *21*(10), 1097–1114.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Gupta, A., Kembhavi, A., & Davis, L. S. (2009). Observing human–object interactions: Using spatial and functional compatibility for

recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(10), 1775–1789.
Haruno, M., Wolpert, D., & Kawato, M. (2001). Mosaic model for sensorimotor learning and control. *Neural Computation*, *13*, 2201–2220.
Ho, E. S., & Shum, H. (2013). Motion adaptation for humanoid robots in constrained environments. In *Proceedings of IEEE international conference on advanced robotics*, pp. 3813–3818.
Ho, E. S., Komura, T., & Tai, C. (2010). Spatial relationship preserving character motion adaptation. In *ACM transactions on graphics*, Vol. 29, p. 33.
Inamura, T., Toshima, I., Tanie, H., & Nakamura, Y. (2004). Embodied symbol emergence based on mimesis theory. *International Journal of Robotics Research*, *23*(4), 363–377.
Kadaba, M. P., Ramakrishnan, H. K., & Wootten, M. E. (1990). Measurement of lower extremity kinematics during level walking. *Journal of Orthopaedic Research*, *8*(3), 383–392.
Kaneko, K., Kanehiro, F., Kajita, S., Yokoyama, K., Akachi, K., Kawasaki, T., et al. (2002). Design of prototype humanoid robotics platform for HRP. In *Proceedings of the 2002 IEEE/RSJ international conference on intelligent robots and systems*, Vol. 3, pp. 2431–2436.
Karpathy, A., & Fei-Fei, L. (2015). Deep visual semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 664–676.
Kojima, A., Tamura, T., & Fukunaga, K. (2002). Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, *50*(2), 171–184.
Krishnamoorthy, N., Malkarnenkar, G., Mooney, R., Saenko, K., & Guadarrama, S. (2013). Generating natural-language video descriptions using text-mined knowledge. In *Proceedings of the 27th AAAI conference on artificial intelligence*, pp. 541–547.
Kulkarni, G., Premraj, V., Dhar, S., & Li, S. (2011). Baby talk: Understanding and generating simple image descriptions. In *Proceedings of IEEE conference on computer vision and pattern recognition*, pp. 1601–1608.
Kuniyoshi, Y., Inaba, M., & Inoue, H. (1994). Learning by watching: Extracting reusable task knowledge from visual observation of human performance. *IEEE Transactions on Robotics and Automation*, *10*(6), 799–822.
Kuroki, Y., Fujita, M., Ishida, T., Nagasaka, K., & Yamaguchi, J. (2003). A small biped entertainment robot exploring attractive applications. In *Proceedings of the IEEE international conference on robotics and automation*, Vol. 1, pp. 471–476.
Li, S., Kulkarni, G., Berg, T. L. , Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. In *Proceedings of the 15th conference on computational natural language learning*, pp. 220–228.
Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on Computer vision*, Vol. 2, pp. 1150–1157.
Mataric, M. J. (2000). Getting humanoids to move and imitate. *IEEE Intelligent Systems*, *15*(4), 18–24.
Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *INTERSPEECH*, pp. 1045–1048.
Morimoto, J., & Doya, K. (2001). Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. *Robotics and Autonomous Systems*, *36*(1), 37–51.
Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, Vol. 77, pp. 257–286.
Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews*, *2*, 661–670.
Saussure, F. D. (1966). *Course in general linguistics*. New York: McGraw-Hill Book Company.
Sugano, S., & Kato, I. (1987). Wabot-2: Autonomous robot with dexterous finger-arm–finger-arm coordination control in keyboard performance. In *Proceedings of 1987 IEEE international conference on robotics and automation*, Vol. 4, pp. 90–97.
Takano, W., & Nakamura, Y. (2015). Symbolically structured database for human whole body motions based on association between motion symbols and motion words. *Robotics and Autonomous Systems*, *66*, 75–85.
Takano, W., & Nakamura, Y. (2015). Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. *International Journal of Robotics Research*, *34*(10), 1314–1328.
Tani, J., & Ito, M. (2003). Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment. *IEEE Transactions on Systems, Man and Cybernetics Part A: Systems and Humans*, *33*(4), 481–488.
Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of IEEE conference on computer vision and pattern recognition*, pp. 3156–3164 .
Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *IEEE conference on computer vision and pattern recognition*, pp. 1290–1297.
Yao, B., & Fei-Fei, L. (2012). Recognizing human–object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(9), 1691–1703.
Yu, G., Liu, Z., & Yuan, J. (2014). Discriminative orderlet mining for real-time recognition of human–object interaction. In *In Proceedings of Asian conference on computer vision*, pp. 50–65.
Zhan, Q., Liang, Y., & Xiao, Y. (2009). Color-based segmentation of point clouds. In *Proceedings of ISPRS laser scanning workshop*, Vol. 38, pp. 248–252.

Wataru Takano is an Associate Professor at at Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo. He was born in Kyoto, Japan, in 1976. He received the B.S. and M.S. degrees from Kyoto University, Japan, in precision engineering in 1999 and 2001, Ph.D. degree from Mechano-Informatics, the University of Tokyo, Japan, in 2006. He was an Project Assistant Professor at the University of Tokyo from 2006 to 2007, and a Researcher on Project of Information Environment and Humans, Presto, Japan Science and Technology Agency from 2010. His field of research includes kinematics, dynamics, artificial intelligence of humanoid robots, and intelligent vehicles. He is a member of IEEE, Robotics Society of Japan, and Information Processing Society of Japan. He has been the chair of Technical Committee of Robot Learning, IEEE RAS.

Yoshihiko Yamada received the B.S. degree and M.S. degree from the University of Tokyo, Japan, in 2013 and 2015 respectively. His research interests are statistical modeling of human actions.

Yoshihiko Nakamura is a Professor at Department of Mechano-Informatics, School of Information Science and Technology, University of Tokyo. He was born in Osaka, Japan, in 1954. He received the B.S., M.S., and Ph.D. degrees from Kyoto University, Japan, in precision engineering in 1977, 1978, and 1985, respectively. He was an Assistant Professor at the Automation Research Laboratory, Kyoto University, from 1982 to 1987. He joined the Department of Mechanical and Environmental Engineering, University of California, Santa Barbara, in 1987 as an Assistant Professor, and became an Associate Professor in 1990. He was also a co-director of the Center for Robotic Systems and Manufacturing at UCSB. He moved to University of Tokyo as an Associate Professor of Department of Mechano-Informatics, University of Tokyo, Japan, in 1991. His fields of research include the kinematics, dynamics, control and intelligence of robots? particularly, robots with non-holonomic constraints, computational brain information processing, humanoid robots, human-figure kinetics, and surgical robots. He is a member of IEEE, ASME, SICE, Robotics Society of Japan, the Institute of Systems, Control, and Information Engineers, and the Japan Society of Computer Aided Surgery. He was honored with a fellowship from the Japan Society of Mechanical Engineers. Since 2005, he has been the president of Japan IFToMM Congress. He is a foreign member of the Academy of Engineering in Serbia and Montenegro.