Takano *ROBOMECH Journal* [redacted]
DOI 10.1186/s40648-014-0022-7

ROBOMECH Journal
a SpringerOpen Journal

RESEARCH ARTICLE Open Access

Learning motion primitives and annotative texts from crowd-sourcing

Wataru Takano

Abstract

Humanoid robots are expected to be integrated into daily life, where a large variety of human actions and language expressions are observed. They need to learn the referential relations between the actions and language, and to understand the actions in the form of language in order to communicate with human partners or to make inference using language. Intensive research on imitation learning of human motions has been performed for the robots that can recognize human activity and synthesize human-like motions, and this research is subsequently extended to integration of motions and language. This research aims at developing robots that understand human actions in the form of natural language. One difficulty comes from handling a large variety of words or sentences used in daily life because it is too time-consuming for researchers to annotate human actions in various expressions. Recent development of information and communication technology gives an efficient process of crowd-sourcing where many users are available to complete a lot of simple tasks. This paper proposes a novel concept of collecting a large training dataset of motions and their descriptive sentences, and of developing an intelligent framework learning relations between the motions and sentences. This framework enables humanoid robots to understand human actions in various forms of sentences. We tested it on recognition of human daily full-body motions, and demonstrated the validity of it.

Keywords: Motion primitives; Natural language; Crowd-sourcing

Background

Robots are able to understand their surroundings by relying on senses supplied by their body, which they can then move to act on the environment. For some time, research has been conducted on imitation learning [1,2], where the bodily motions of humans are projected onto the bodily motions of humanoid robots and recorded as dynamical system [3-6] and statistical model [7-10] parameters while compressing the information. By using these models, it has become possible for robots to recognize human bodily motions and to generate their own natural human-like motions. However, in the motion recognition phase, the motion is classified into its specific model, and in the motion generation phase, a command specifying the model is given to a robot. More specifically, indices of the motion models that are not understood by human partners intervene in the motion recognition and generation. The intermediate codes that can be intuitively understood by the human partners are required. A natural language can be its solution, and facilitate an intuitive interaction between humans and robots. Several approaches extend the motion models to the language expressions, where robots understand human motions as text and can then generate bodily motions from text input [11,12]. Several models for a robot manipulating object via linguistic instructions have been developed using a neural network [13,14]. The Variation of the objects and actions is small. Our daily lives are overflowing with a huge variety of possible motions and expressions for describing them. Therefore, there is a need for humanoid robots to be able to adapt to this diversity.

In this study, I created a training dataset of motions and corresponding texts describing those motions by assigning a variety of text phrases to human bodily motions via crowdsourcing [15]. I then built an intellectual framework that can understand language for expressing movement by learning the correspondence between bodily motions and language expressions via a statistical model. This technology to collect and utilize a massive amount of text

Correspondence: takano@ynl.t.u-tokyo.ac.jp
The Univ. of Tokyo, Bunkyoku Hongo 7-3-1, Tokyo, Japan

expressions as training data is expected to form the foundation for intelligence that can adapt to a diversity of language expressions.

Method

Motion annotations

The full-body motions of humans were measured by optical motion capture or wearable motion sensors. Position data at each point on the body were converted into motions of a computer-generated model character using inverse kinematic calculations. Videos of these motions were made viewable on the Internet. Figure 1 shows examples of frames from the videos.

The task of manually assigning descriptive annotation to each motion video was carried out via crowdsourcing. In the annotation task, a video, a playback time, and a word representing the subject are presented. The user inputs descriptive text in English corresponding to the motion initiated by the given subject at the specified time. Using this task, a training dataset of motions and corresponding descriptive texts can thus be collected. In this study, the annotation task was openly available from our research laboratory's website as shown by Figure 2. The students and researchers from my department are allowed to annotate the motions such that appropriately assigned descriptive texts can be collected efficiently.

The task described above provides descriptive sentences and their corresponding times. This task does not provide a start point and an end point of a motion segment to which the descriptive sentence is assigned. I manually detected the start point and end point for each motion segment after the annotation task, and consequently obtained datasets of the motion segments and their descriptive sentences.

Learning motions and annotations

A human full-body motion is represented by a sequence of angles of all the joints. Each sequence is encoded into an HMM λ. An HMM is a statistical model used to classify input data into an appropriate category. An HMM is defined by a compact notation $\lambda = \{\boldsymbol{Q}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi}\}$, where $\boldsymbol{Q} = \{q_1, q_2, \cdots, q_n\}$ is the set of nodes, $\boldsymbol{A} = \{a_{ij}\}$ is the matrix whose entries a_{ij} are the probability of transitioning from the ith node to the jth node, $\boldsymbol{B}$ is the set of output probability density functions at the nodes, and $\boldsymbol{\Pi} = \{\pi_1, \pi_2, \cdots, \pi_n\}$ is the set of initial node distribution. In this study, the parameters of the HMM are optimized by Baum–Welch algorithm using its corresponding sequence of the joint angles. Baum–Welch algorithm is

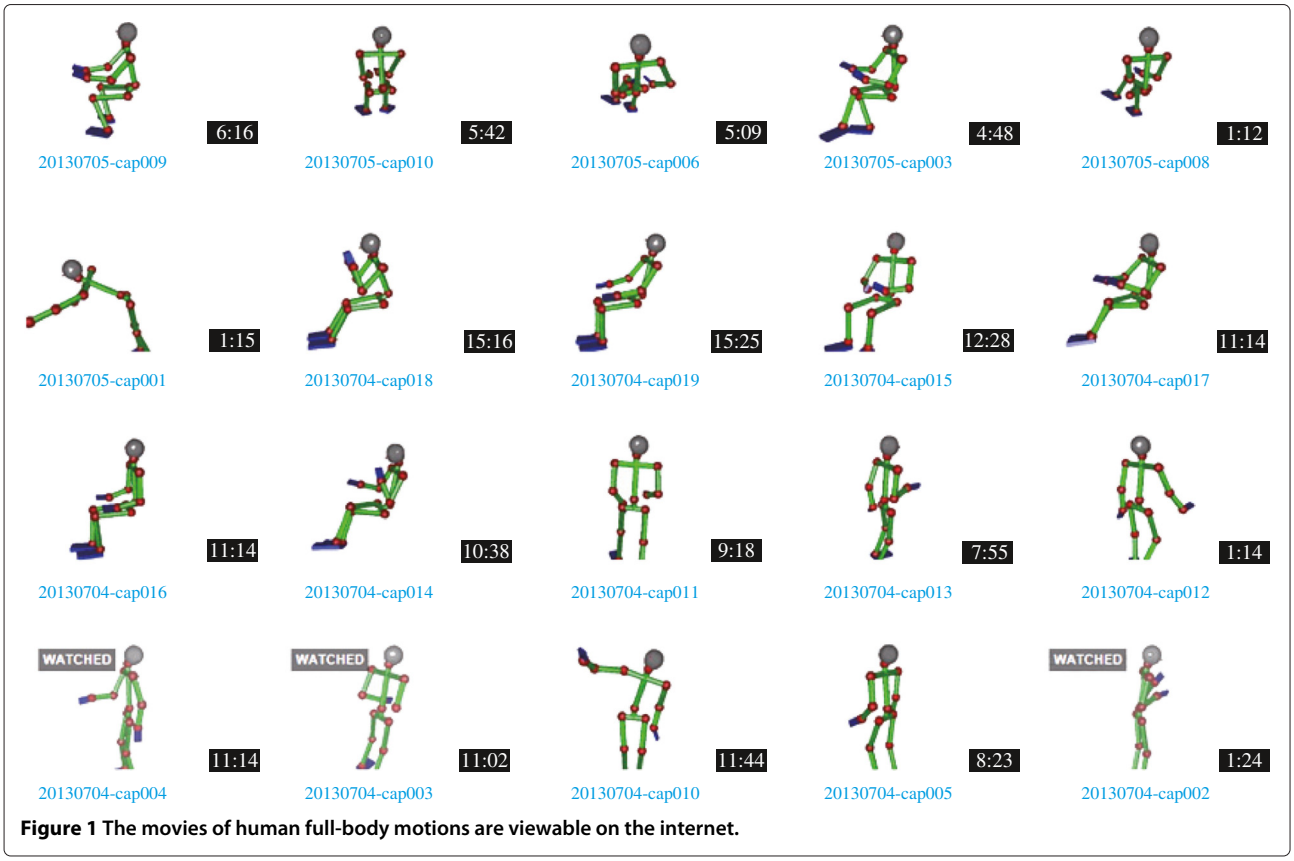

Figure 1 The movies of human full-body motions are viewable on the internet.

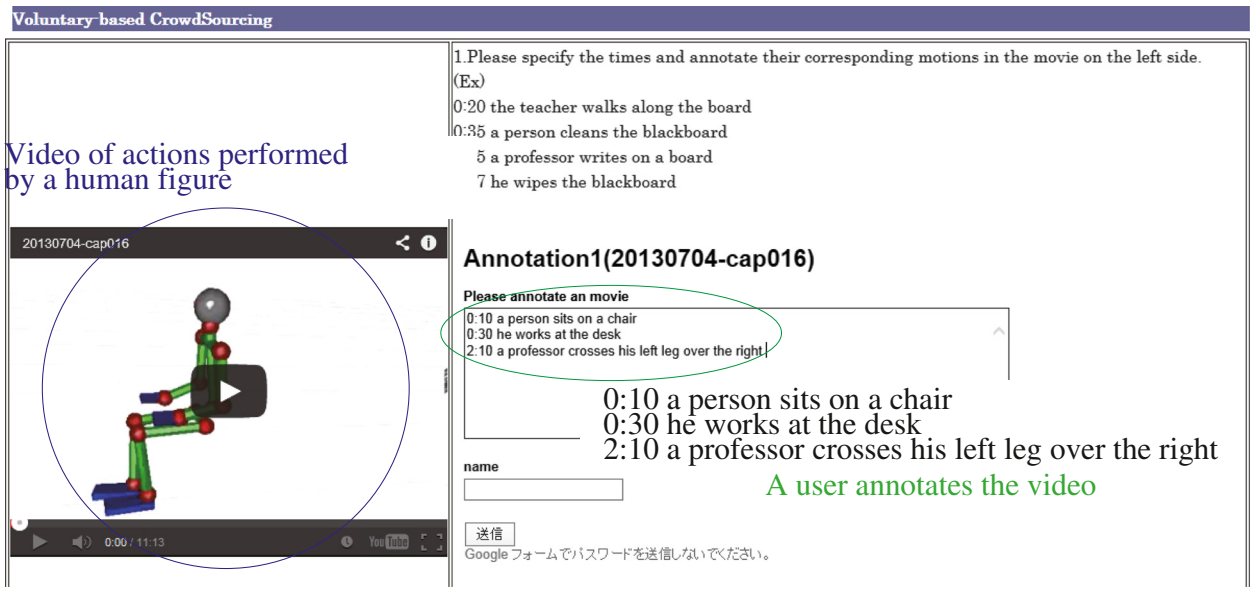

Figure 2 A video containing human activity is openly available. Any user can assign the descriptive annotation to the video on the Internet.

one of the expectation maximization (EM) algorithms [16]. The motion can be classified into its relevant HMM that is the most likely to generate this motion. The motion is expressed by the discrete form of the index of the HMM, and the HMM is hereinafter referred to as a "motion symbol".

In the annotation task, a descriptive annotation is assigned to each motion symbol. Consequently, a training dataset of motion symbols and descriptive texts is collected. More specifically, each training data is a pair of motion symbol λ_k and a descriptive sentence $\boldsymbol{\omega}_k$, where the descriptive sentence is expressed by a sequence of l_k words, $\boldsymbol{\omega}_k = \left\{\omega_1^k, \omega_2^k, \cdots, \omega_{l_k}^k\right\}$. This paper proposes a statistical model that converts the motion symbol to descriptive sentences as shown by Figure 3 [12]. This conversion results in understanding human full-body motion in the forms of sentences. The statistical model consists of two modules. One module learns the probabilistic relations between a motion symbol λ and a word ω. This module is hereinafter referred to as "motion language module". The other module learns the probabilistic relations of transition of two words in a sentence. This module is referred to as "natural language module".

Figure 4 shows an overview of the motion language module that consists of three layers. The top layer includes motion symbols, the middle layer includes latent states, and the bottom layer includes words. A motion symbol generates a latent state, and a latent state generates a word. Association between the motion symbols and the words are represented by a generative model. Probabilistic relation between the motion symbol and word is represented using the probability $P(s|\lambda)$ that the motion symbol λ generates the latent state s, and the probability $P(\omega|s)$ that the latent state s generates the word ω. These probabilities are optimized such that the total probability that motion symbols generate the words in the descriptive sentences in the training dataset is maximized. The logarithm of the total probability is written as

$$\Phi(\theta) = \log \prod_k P\left(\omega_1^k, \omega_2^k, \cdots, \omega_{l_k}^k | \lambda_k\right) \tag{1}$$

$$= \sum_k \log P\left(\omega_1^k, \omega_2^k, \cdots, \omega_{l_k}^k | \lambda_k\right) \tag{2}$$

$$= \sum_{k,i} \log P\left(\omega_i^k | \lambda_k\right) \tag{3}$$

$$= \sum_{k,i} \log \sum_j P\left(\omega_i^k, s_j | \lambda_k\right) \tag{4}$$

where θ is a set of the probabilities $P(s|\lambda)$ and $P(\omega|s)$. I assume that a word is independent of each other, and is dependent on only the motion symbol in the motion language module. Equation (2) can be subsequently rewritten as Equation (3). The dependence relationship between two words is learned by a natural language module. The The optimal θ is derived by the iterative computation. Let $\theta^{[t]}$ be the set θ derived at t-th iteration. The probabilities $P(\omega, s|\lambda)$, $P(s|\lambda)$, and $P(\omega|s)$ derived at t-th iteration are rewritten as $P(\omega, s|\lambda, \theta^{[t]})$, $P(s|\lambda, \theta^{[t]})$, and

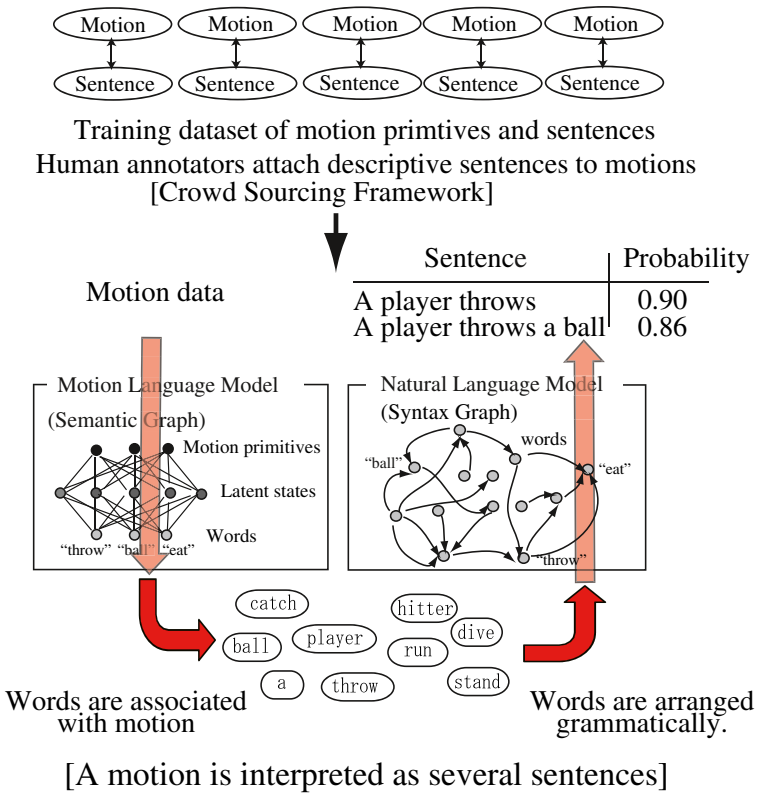

Figure 3 Human annotators gives descriptive sentences to the motions. Training datasets of the motions are sentences are used to model the referential structure between the motions and sentences, to model the sentence structures only using sentences. The models are applied to interpretation of motions in the form of natural language.

$P(\omega|s,\theta^{[t]})$ respectively. Equation (4) at t-th iteration is rewritten as

$$\Phi(\theta^{[t]}) = \sum_{k,i} \log \sum_{j} P\left(\omega_i^k, s_j|\lambda_k, \theta^{[t]}\right) \tag{5}$$

$$= \sum_{k,i} \log \sum_{j} P\left(s_j|\lambda_k, \omega_i^k\right) \frac{P\left(\omega_i^k, s_j|\lambda_k, \theta^{[t]}\right)}{P\left(s_j|\lambda_k, \omega_i^k\right)} \tag{6}$$

$$= \sum_{k,i} \log E_{P\left(s|\lambda_k,\omega_i^k\right)} \left[\frac{P\left(\omega_i^k, s|\lambda_k, \theta^{[t]}\right)}{P\left(s|\lambda_k, \omega_i^k\right)}\right], \tag{7}$$

where $E_P[R]$ denotes the expected value of R given the distribution P. According to Jensen's inequality, Equation (7) satisfies the following relation.

$$\Phi\left(\theta^{[t]}\right) \geq \sum_{k,i} E_{P\left(s|\lambda_k,\omega_i^k\right)} \left[\log \frac{P\left(\omega_i^k, s|\lambda_k, \theta^{[t]}\right)}{P\left(s|\lambda_k, \omega_i^k\right)}\right] \tag{8}$$

Using Equation (3) and Equation (8), the following equations can be derived.

$$\log P\left(\omega_i^k|\lambda_k\right) - E_{P\left(s|\lambda_k,\omega_i^k\right)} \left[\log \frac{P\left(\omega_i^k, s|\lambda_k, \theta^{[t]}\right)}{P\left(s|\lambda_k, \omega_i^k\right)}\right] \tag{9}$$

$$= E_{P\left(s|\lambda_k,\omega_i^k\right)} \left[\log \frac{P\left(s|\lambda_k, \omega_i^k\right)}{P\left(s|\lambda_k, \omega_i^k, \theta^{[t]}\right)}\right] \tag{10}$$

$$= KL\left(P\left(s|\lambda_k, \omega_i^k\right)||P\left(s|\lambda_k, \omega_i^k\theta^{[t]}\right)\right). \tag{11}$$

Equation (11) represents the Kullback Leibler information that measures the dissimilarity between the distributions $P(s|\lambda_k, \omega_i^k)$ and $P(s|\lambda_k, \omega_i^k, \theta^{[t]})$. The Kullback Leibler information becomes zero only when these two distributions are exactly same, and takes a positive value

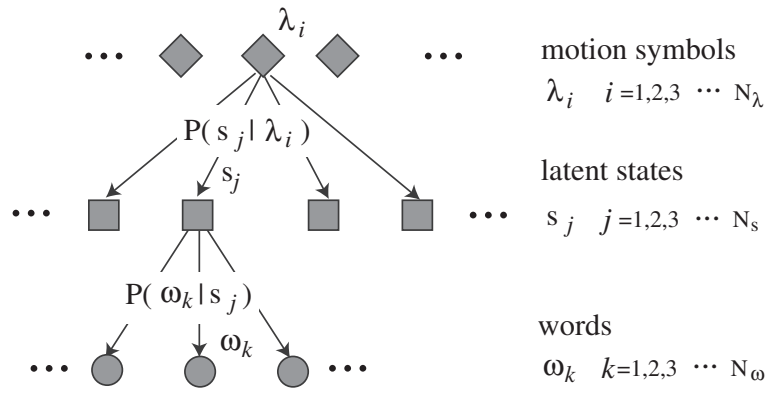

Figure 4 A motion language module learns the probability of a word being generated from a motion symbol. This probability is calculated using the probability of a latent state being generated from the motion symbol and the probability of the word being generated from the latent state.

otherwise. The difference between $\Phi(\theta^{[t+1]})$ and $\Phi(\theta^{[t]})$ is subsequently written as follows:

$$\begin{aligned}\Delta\Phi &= \Phi\left(\theta^{[t+1]}\right) - \Phi\left(\theta^{[t]}\right)\\ &= \sum_{k,i} E_{P\left(s|\lambda_k,\omega_i^k\right)}\left[\log\frac{P\left(\omega_i^k, s|\lambda_k, \theta^{[t+1]}\right)}{P\left(s|\lambda_k,\omega_i^k\right)}\right]\\ &- \sum_{k,i} E_{P\left(s|\lambda_k,\omega_i^k\right)}\left[\log\frac{P\left(\omega_i^k, s|\lambda_k, \theta^{[t]}\right)}{P\left(s|\lambda_k,\omega_i^k\right)}\right]\\ &+ \sum_{k,i} KL\left(P\left(s|\lambda_k,\omega_i^k\right)||P\left(s|\lambda_k,\omega_i^k,\theta^{[t+1]}\right)\right)\\ &- \sum_{k,i} KL\left(P\left(s|\lambda_k,\omega_i^k\right)||P\left(s|\lambda_k,\omega_i^k,\theta^{[t]}\right)\right) \qquad (12)\end{aligned}$$

The distribution $P(s|\lambda_k,\omega_i^k)$ is assumed to be estimated as $P(s|\lambda_k,\omega_i^k,\theta^{[t]})$ based on the motion language model derived at t-th iteration, and the third and fourth terms in Equation (12) take a positive value and zero respectively. Hence, I only have to search for $\theta^{[t+1]}$ such that the first term in Equation (12) becomes greater than the second term because the incremental update of $\theta^{[t+1]}$ increases the total probability Φ of the training data. More specifically, the first term only has to be maximized by $\theta^{[t+1]}$. Using the probabilities $P(s|\lambda,\theta^{[t+1]})$ and $P(\omega|s,\theta^{[t+1]})$, This maximization can be reduced as follows

$$\begin{aligned}&\arg\max_{\theta^{[t+1]}}\sum_{k,i,j} P\left(s_j|\lambda_k,\omega_i^k\right)\left[\log P\left(\omega_i^k, s_j|\lambda_k,\theta^{[t+1]}\right)\right.\\ &\qquad\left.-\log P\left(s_j|\lambda_k,\omega_i^k\right)\right]\\ &= \arg\max_{\theta^{[t+1]}}\sum_{k,i,j} P\left(s_j|\lambda_k,\omega_i^k\right)\log P\left(\omega_i^k, s_j|\lambda_k,\theta^{[t+1]}\right)\\ &= \arg\max_{\theta^{[t+1]}}\sum_{k,i,j} P\left(s_j|\lambda_k,\omega_i^k\right)\left[\log P\left(\omega_i^k, |s_j,\theta^{[t+1]}\right)\right.\\ &\qquad\left.+ P\left(s_j|\lambda_k,\theta^{[t+1]}\right)\right] \qquad (13)\end{aligned}$$

where the terms independent of $\theta^{[t+1]}$ are eliminated. The probabilities $P(s|\lambda,\theta^{[t+1]})$ and $P(\omega|s,\theta^{[t+1]})$ are constrained as follows:

$$\sum_j P\left(s_j|\lambda,\theta^{[t+1]}\right) = 1 \qquad (14)$$

$$\sum_i P\left(\omega_i|s,\theta^{[t+1]}\right) = 1 \qquad (15)$$

By applying the method of Lagrange multiplier to Equation (13), the probabilities $P(s|\lambda,\theta^{[t+1]})$ and $P(\omega|s,\theta^{[t+1]})$ at $t+1$-th iteration can be analytically derived.

$$P\left(s|\lambda_k,\theta^{[t+1]}\right) = \frac{\sum_i P\left(s|\lambda_k,\omega_i^k\right)}{\sum_{i,j} P\left(s_j|\lambda_k,\omega_i^k\right)} \qquad (16)$$

$$P\left(\omega_i|s,\theta^{[t+1]}\right) = \frac{\sum_k n_{k,i} P\left(s|\lambda_k,\omega_i\right)}{\sum_{k,i} n_{k,i} P\left(s|\lambda_k,\omega_i\right)} \qquad (17)$$

where $n_{k,i}$ is the number that the word ω_i appears in the sentence $\boldsymbol{\omega}_k$ assigned to the motion symbol λ_k. Note

that ω_i denotes the i-th word in the set of words, and ω_i^k denotes the word at the i-th position in the sentence assigned to the k-th motion symbol. The processes described above are iterated, and consequently the optimal probabilities $P(s|\lambda)$ and $P(\omega|s)$ can be derived.

Figure 5 shows an overview of the natural language module. This module extracts the probability $\pi(\omega)$ of starting at the word ω and the probability $P(\omega_j|\omega_i)$ of transitioning from the word ω_i to the word ω_j using a training dataset of sentences assigned to the motion symbols. The probabilities $\pi(\omega_i)$ and $P(\omega_j|\omega_i)$ are optimized such that the probability that the natural language module generates the training sentences. The logarithm of this probability is expressed by

$$\Psi(\vartheta) = \sum_k \log P(\boldsymbol{\omega}_k) \tag{18}$$

$$= \sum_k \log \pi\left(\omega_1^k\right) + \sum_{k,i} \log P\left(\omega_{i+1}^k|\omega_i^k\right). \tag{19}$$

where ϑ is a set of probabilities $\pi(\omega)$ and $P(\omega_j|\omega_i)$. The optimal ϑ can be analytically derived as follows.

$$\pi(\omega) = \frac{c(\omega)}{\sum_i c(\omega_i)} \tag{20}$$

$$P\left(\omega_j|\omega_i\right) = \frac{c\left(\omega_i, \omega_j\right)}{\sum_j c\left(\omega_i, \omega_j\right)} \tag{21}$$

where $c(\omega)$ is the frequency of the sentence starting at the word ω, and $c(\omega_i, \omega_j)$ is the frequency of transitions from the word ω_i to the word ω_j.

The conversion from the motion symbol $\lambda_{\mathcal{R}}$ to its descriptive sentences $\boldsymbol{\omega}_{\mathcal{R}}$ can be treated as the problem of searching for the sentences that are most likely to be generated by the motion symbols. This problem is expressed as follows:

$$\omega_{\mathcal{R}} = \arg\max_{\hat{\omega}} P\left(\hat{\omega}|\lambda_{\mathcal{R}}\right) \tag{22}$$

$$= \arg\max_{\hat{\omega}} P\left(\hat{\omega}_1, \hat{\omega}_2, \cdots, \hat{\omega}_l|\lambda_{\mathcal{R}}\right) P\left(\hat{\omega}|\hat{\omega}_1, \hat{\omega}_2, \cdots, \hat{\omega}_l\right) \tag{23}$$

where $P(\hat{\omega}_1, \hat{\omega}_2, \cdots, \hat{\omega}_l|\lambda_{\mathcal{R}})$ is the probability that the motion language module generates a set of words $\left\{\hat{\omega}_1, \hat{\omega}_2, \cdots, \hat{\omega}_l\right\}$ from the motion symbol $\lambda_{\mathcal{R}}$, and $P(\hat{\boldsymbol{\omega}}|\hat{\omega}_1, \hat{\omega}_2, \cdots, \hat{\omega}_l)$ is the probability that the natural language module arranges the set of words $\left\{\hat{\omega}_1, \hat{\omega}_2, \cdots, \hat{\omega}_l\right\}$ into the sentence $\hat{\boldsymbol{\omega}}$. Therefore, these two probabilities can be written using the probabilities defining the motion language module and the natural language module.

$$P\left(\hat{\omega}_1, \hat{\omega}_2, \cdots, \hat{\omega}_l|\lambda_{\mathcal{R}}\right) = \prod_i P\left(\hat{\omega}_i|\lambda_{\mathcal{R}}\right) \tag{24}$$

$$P\left(\hat{\omega}|\hat{\omega}_1, \hat{\omega}_2, \cdots, \hat{\omega}_l\right) = \pi\left(\hat{\omega}_1\right) \prod_i P\left(\hat{\omega}_{i+1}|\hat{\omega}_i\right) \tag{25}$$

where $P(\hat{\omega}_i|\lambda_{\mathcal{R}})$ can be calculated as $\sum_j P(\hat{\omega}_i|s_j)P(s_j|\lambda_{\mathcal{R}})$. Substituting Equation (24) and Equation (25) into Equation (23) and taking the logarithm of it, Equation (23) can be reduced to the following equation.

$$\omega_{\mathcal{R}} = \arg\max_{\hat{\omega}} \left[\sum_i \log P\left(\hat{\omega}_i|\lambda_{\mathcal{R}}\right) + \log \pi\left(\hat{\omega}_1\right) + \sum_i \log P\left(\hat{\omega}_{i+1}|\hat{\omega}_i\right) \right] \tag{26}$$

Equation (26) can be efficiently solved using Dijkstra's algorithm.

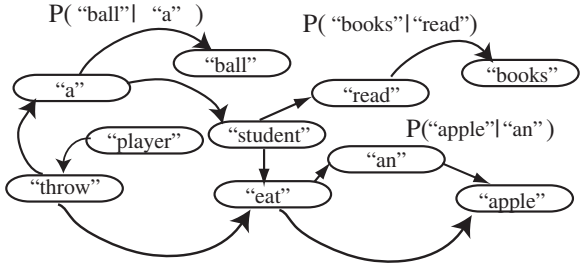

Figure 5 A natural language module learns the probability of transitioning between words in sentences.

Result and discussion

Experiments

An experiment on the conversion from the full-body motions of human to the descriptive sentences was conducted by using our proposed statistical framework. The full-body motions were measured using an inertial motion capture system where 17 IMU sensors were attached to a human performer. This measurement was conducted with the approval of the ethical committee of the University of Tokyo. Positions of 34 selected bodied part in the human full-body in the trunk coordinate system were derived via kinematic computation using a human figure model with 34 degrees of freedom. Each measured motion segment is encoded into an HMM. The HMM consists of 30 nodes, each of which has one Gaussian distribution, and the type of node connection is left-to-right. A descriptive sentence is manually assigned to each HMM via crowdsourcing. In this study the full-body motions of one performer were measured during working at the office or giving a lecture, and 621 motion symbols, each of which

a sentence is assigned to by five users, were subsequently collected. The number of different words used in the descriptive sentences was 419. Table 1 shows sample parts of the training dataset of motions and their descriptive sentences.

After learning the motion language module and the natural language module using the training dataset as shown by Table 1, the proposed framework was tested on 100 different full-body motions of human. Each motion is converted to five descriptive sentences that are most likely to be generated by both the motion language module and the natural language module. Figure 6 shows the experimental result of conversion from a full-body motion to sentences, where a sentence containing less than three words is removed as a candidate sentence. A motion "sitting" is converted into sentences "a person sits", "a person sits down", "a person sits back" and "he sits down". A motion "drinking" is converted into sentences "a person is drinking" and "he is drinking". These sentences were confirmed to correctly represent the full-body motions. A motion "sitting with legs crossed" is correctly converted into sentences "he is sitting", but it is wrongly converted into another sentence "he is sitting with his legs". Additionally, it is correctly converted into a long sentence "he is sitting with his legs crossed", that is ranked lower than the wrong sentence "he is sitting with his legs". A motion "writing on a blackboard" is converted into a correct sentence "he is writing", and wrong sentence "he is writing on" and "he is writing a blackboard" that are close to the correct sentence "he is writing on a blackboard". The several wrong sentences are terminated at the inappropriate words, and longer sentences are unlikely to be generated. The natural language model needs to be extend to word trigrams such that it represents the relations among words that are distant from each other in the sentences, and the conversion from the motion to the sentence, expressed by Equation (26), should be modified to take into account the length of sentences.

I also quantitatively evaluate the conversion from the motions to sentences. Five users assigned a descriptive sentence to each test motion. The performer and users in this test phase are same as those in the learning phase.

Table 1 Motions λ and Annotations ω in the training dataset

λ	ω	λ	ω
1	a person is sitting	2	a person is sitting
3	a performer is sitting	4	a person is working at his desk
5	a performer is working at his desk	6	a person is sitting in a chair
7	a person is reaching out a hand	8	a person sits back
9	a performer sits back	10	a person crosses his right leg over the left
11	he crosses his right leg over the left	12	a performer crosses his right leg over the left
13	a person crosses his right leg over the left	14	a person is operating a computer with his legs crossed
15	a person is sitting in a chair	16	a performer is sitting with his legs crossed
17	a person sits down	18	a professor sits down
19	a person is sitting in a chair	20	a performer is sitting in a chair
21	he scratches his shoulder	22	he is reading
23	he is relaxed	24	he concentrates on reading
25	he concentrates	26	he puts down his book
27	he puts down	28	he is crossing his left leg
29	he is reading	30	a person is sittiing down
31	he is writing on a blackboard	32	he is checking
33	he is walking	34	he is checking his notebook
35	he is writing on a blackboard	36	he is looking at students
37	he is teaching	38	he is writing on a blackboard
39	he is pointing out	40	he is explaining
41	he plants his arm on his chin	42	he plants his arm on a table
43	he is drinking	44	he is drinking
45	he puts down something	46	he is resting
47	he puts his hands on a table	48	he drinks
49	he is studying	50	he is crossing his arms

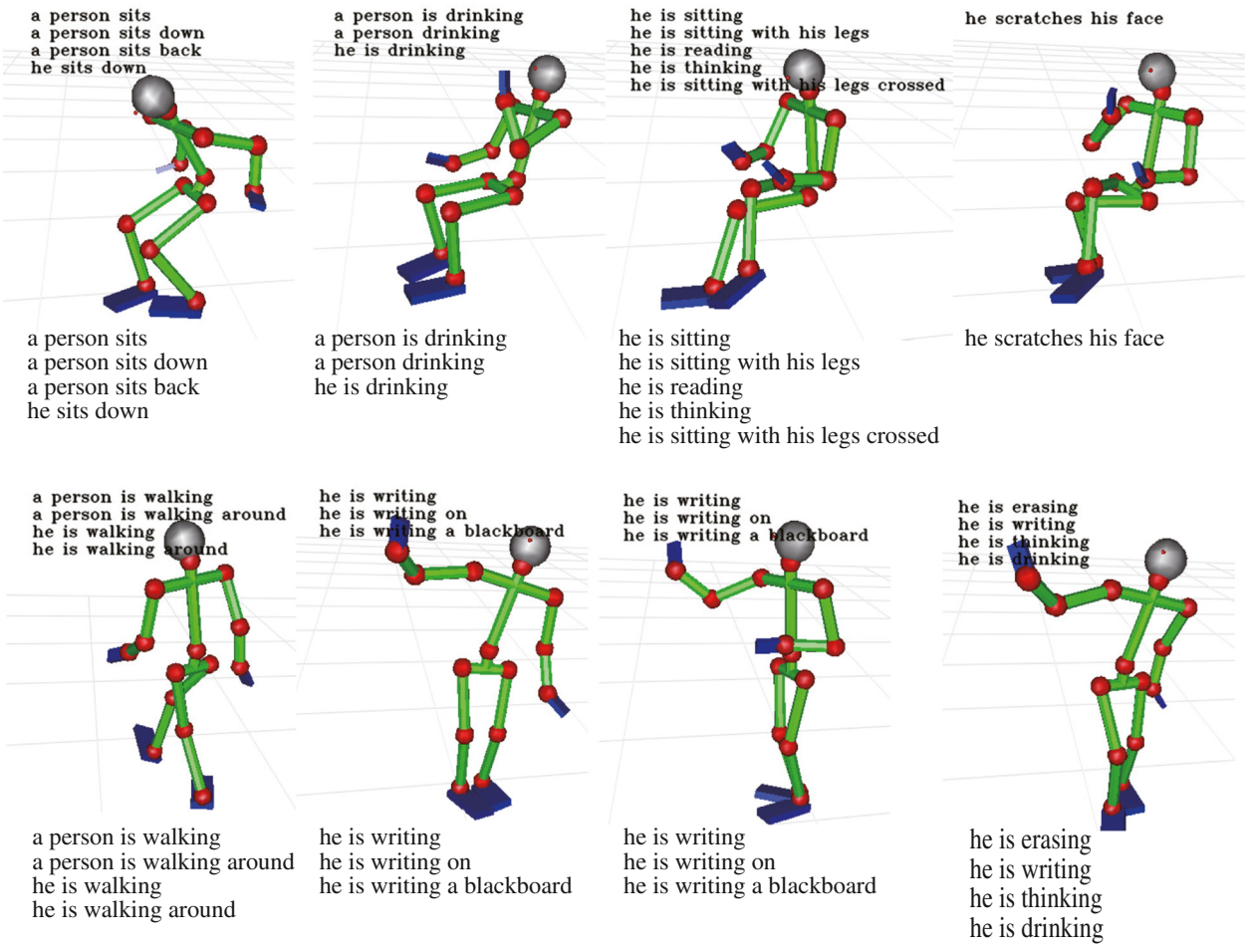

Figure 6 A human full-body motion is converted to multiple descriptive sentences. This conversion make it possible for a humanoid robot to understand observation of the human motion in the forms of sentence.

Each motion that was converted to several candidate sentences, one of which is exactly same as the sentence assigned to this test motion was counted as the correct. The accuracy of the conversion can be computed as a ratio of the correct motions to the test motions. The number of the candidate sentences was varied. In the case that the number of the candidate sentences was set to 1, the accuracy of the conversion was 0.34. The number of the candidate sentences was set to 2, the accuracy of the conversion reached 0.59. Three, four, and five candidate sentences result in the accuracies of 0.64, 0.68 and 0.71 respectively.

Conclusion

The contributions of this paper are summarized as follows.

1. This paper proposes a novel scheme of collecting a training dataset of human full-body motions and their descriptive sentences via crowdsourcing. Videos containing human activity are made viewable on the Internet. The task of assigning the descriptive annotations to the videos is designed. The task is openly available, and can be carried out by any users. Through this simple task, a training data set of motions and corresponding descriptive sentences can be collected. In this study, there are 621 motions and descriptive sentences with 419 different words in the training dataset.
2. This paper proposes a statistical framework to convert a full-body motion to multiple descriptive sentences. This framework consists of two modules : motion language module and natural language module. The motion language module statistically learns association between motions and words, and the natural language module learns transition between two words in the sentences. The integration of these two modules enables a humanoid robot to convert a human full-body motion to its descriptive sentences.

3. The experiment on the conversion from the human full-body motion to the sentences was conducted using dataset of motions and descriptive annotations derived via the crowdsourcing. I varied the number of candidate sentences converted from the motion. The accuracy of the conversion of 0.34, 0.59, 0.64, 0.68 and 0.71 were obtained from one, two, three, four and five candidate sentences respectively. The experiment shows that the full-body motions are converted to correct descriptive sentences, and demonstrates the validity of the proposed statistical framework for the conversion of the motions to the sentences. Additionally I found several limitations that a long sentence is unlikely to generated, and that many sentences are terminated at the wrong words.

Competing interests
The author declares that he has no competing interests.

Acknowledgements
This research was supported by Grant-in-Aid for Young Scientists (A) (26700021), Japan Society for the Promotion of Science.

Received: 18 August 2014 Accepted: 15 November 2014
Published online: 20 January 2015

References

1. Breazeal C, Scassellati B (2002) Robots that imitate humans. Trends Cognitive Sci 6(11):481–487
2. Argall B, Chernova S, Veloso M, Browning B (2009) A survey of robot learning from demonstration. Robot Autonomous Syst 57(5):469–483
3. Okada M, Tatani K, Nakamura Y (2002) Polynomial design of the nonlinear dynamics for the brain-like information processing of whole body motion. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp 1410–1415
4. Ijspeert AJ, Nakanishi J, Shaal S (2003) Learning control policies for movement imitation and movement recognition. Neural Inf Process Syst 15:1547–1554
5. Kadone H, Nakamura Y (2005) Symbolic memory for humanoid robots using hierarchical bifurcations of attractors in nonmonotonic neural networks. In: Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp 2900–2905
6. Ito M, Noda K, Hoshino Y, Tani J (2006) Dynamic and interactive generation of object handing behaviors by a small humanoid robot using a dynamic neural network model. Neural Netw 19(3):323–337
7. Inamura T, Toshima I, Tanie H, Nakamura Y (2004) Embodied symbol emergence based on mimesis theory. Intl J Robot Res 23(4):363–377
8. Asfour T, Gyarfas F, Azad P, Dillmann R (2006) Imitation learning of dual-arm manipulation task in humanoid robots. In: Proceedings of the IEEE-RAS International Conference on Humanoid Robots. pp 40–47
9. Billard A, Calinon S, Guenter F (2006) Discriminative and adaptive imitation in unimanual and bi-manual tasks. Robot Autonomous Syst 54:370–384
10. Kulic D, Takano W, Nakamura Y (2008) Incremental learning, clustering and hierarchy formation of whole body motion patterns using adaptive hidden markov chains. Intl J Robot Res 27(7):761–784
11. Takano W, Yamane K, Nakamura Y (2007) Capture database through symbolization, recognition and generation of motion patterns. In: Proceedings of the IEEE International Conference on Robotics and Automation. pp 3092–3097
12. Takano W, Nakamura Y (2008) Integrating whole body motion primitives and natural language for humanoid robots. In: Proceedings of the IEEE-RAS International Conference on Humanoid Robots. pp 708–713
13. Tuci E, Ferrauto T, Zeschel A, Massera G, Nolfi S (2011) An experiment on behavior generalization and the emergence of linguistic compositionality in evolving robots. IEEE Trans Autonomous Mental Dev 2(2):176–189
14. Tuci E, Ferrauto T, Zeschel A, Massera G, Nolfi S (2010) The facilitatory role of linguistic instructions on developing manipulation skills. IEEE Comput Intell Mag 5(3):33–42
15. Howe J (2006) The Rise of Crowdsourcing. Wired Magazine 14(6)
16. Rabiner L, Juang BH (1993) Fundamentals of speech recognition. In: Prentice Hall Signal Processing Series.