Robotics and Autonomous Systems 66 (2015) 35–43

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Correlated space formation for human whole-body motion primitives and descriptive word labels

Wataru Takano *, Seiya Hamano, Yoshihiko Nakamura

Mechano-Informatics, University of Tokyo, 7-3-1, Hongo, Bunkyoku, Tokyo, 113-8656, Japan

HIGHLIGHTS

- We construct correlated spaces of human motions and word labels.
- The correlated space can be applied to searching for motions from word queries.
- The motions can be retrieved, even if the queries are not assigned to the motions.
- This technology can be helpful for reusing the motion data.

ARTICLE INFO

Article history:
Received 17 April 2014
Received in revised form
22 September 2014
Accepted 1 November 2014
Available online 13 January 2015

Keywords:
Motion primitive
Word label
Canonical correlation analysis

ABSTRACT

The motion capture technology has been improved, and widely used for motion analysis and synthesis in various fields, such as robotics, animation, rehabilitation, and sports engineering. A massive amount of captured human data has already been collected. These prerecorded motion data should be reused in order to make the motion analysis and synthesis more efficient. The retrieval of a specified motion data is a fundamental technique for the reuse. Imitation learning frameworks have been developed in robotics, where motion primitive data is encoded into parameters in stochastic models or dynamical systems. We have also been making research on encoding motion primitive data into Hidden Markov Models, which are referred to as "motion symbol", and aiming at integrating the motion symbols with language. The relations between motions and words in natural language will be versatile and powerful to provide a useful interface for reusing motion data. In this paper, we construct a space of motion symbols for human whole body movements and a space of word labels assigned to those movements. Through canonical correlation analysis, these spaces are reconstructed such that a strong correlation is formed between movements and word labels. These spaces lead to a method for searching for movement data from a query of word labels. We tested our proposed approach on captured human whole body motion data, and its validity was demonstrated. Our approach serves as a fundamental technique for extracting the necessary movements from a database and reusing them.

1. Introduction

To understand real-world phenomena that are not clearly separated, humans segment those phenomena and perceive them as symbols. Rather than being based on physical properties, this segmentation is arbitrary and depends on the society to which the person belongs [1]. However, these symbols have been refined by recording the correspondence between the arbitrarily segmented world and the symbols used for denoting it, as well as the relation between the different symbols, and through the evolutionary process of the cumulative utilization of these symbols. This immense system of intricately intertwined symbols sublimated into language, allowing humans to communicate efficiently with one another and to perform high-order reasoning. It can be said without exaggeration that the high-order cognitive capabilities of humans are a product of language.

For humanoid robots to coexist with humans, they will have to be able to use the same symbols and language systems as humans. Research on robot body motion has focused on imitating learning methods that optimize the parameters of mathematical models based on various motion patterns [2,3]. In this framework, time-series data (e.g., data about the joint angles representing motions) are memorized as symbols represented discretely by parameters of a statistical model [4,5] or of a dynamical system [6–10]. This sort

* Corresponding author. Tel.: +81 3 5841 6378; fax: +81 3 3818 0835.
E-mail addresses: takano@ynl.t.u-tokyo.ac.jp (W. Takano), hamano@ynl.t.u-tokyo.ac.jp (S. Hamano), nakamura@ynl.t.u-tokyo.ac.jp (Y. Nakamura).

http://dx.doi.org/10.1016/j.robot.2014.11.020

of robot intelligence is the ability not only to understand human behavior by comparing human motions with previously memorized motions but also to generate continuous motions from memorized symbolic representations of motions to apply them to the real world. The expansion of the range of fields where humanoid robots are used is creating an increasingly strong demand for a framework to memorize many motion symbols [11].

With the improvement and spread of optical motion capture technologies, data about human body motions is getting applied not only in robotics but also in various other fields, such as animation, sports engineering and rehabilitation. A massive amount of data about human body motion has already been collected. However, when reusing memorized motion data in order to synthesize new motions for animated characters or to perform motion analysis by comparing motions with previous ones, efficient search techniques should be used to find the necessary motions in the collected data. Currently, searching for and reusing the necessary motion data is based on labels such as measurement time or motion description. An environment where it is possible to search for only labels that match the input data places a large burden on operators that reuse the motion data by requiring them to memorize the exact measurement time or motion description.

The ability of a robot to perform intelligent information processing by encoding and categorizing large amounts of body motion data and linking that data to linguistic representations forms the basis of the robot's comprehension of language and body motion. Also, this is closely related to technology for searching and presenting motion data related to simple linguistic input. This ability would substantially improve the reusability of motion data in motion analysis or motion generation for CG characters. Frameworks proposed thus far have been based on arrays of motion symbols representing body motion data learned through a Hidden Markov Model (HMM) [12] and arrays of verb labels attached to those motions. In those frameworks, emphasis is placed on learning the correspondence between motions and verbs by restricting the linguistic representations to verbs and considering the context of the symbol and verb arrays, however, without taking into account the interrelation between motions or other linguistic units, such as nouns or adverbs [13]. One proposed method for expressing the interrelation between motion symbols involves calculating the distances between individual motion symbols and constructing a multidimensional motion symbol space and assigning motion symbols to points such that those distances are preserved [14]. Furthermore, there are language processing techniques in which sentences composed of verbs, nouns and other elements are represented as points in a vector space based on the presence, absence or frequency of the constituent words [15–17]. Thus, it may be possible to construct a computational model connecting motions to word labels by using a common representation of both motions and word labels as points in some spaces. In this paper, we construct a space of motion symbols learned by applying an HMM to body motion data and a word label space consisting of verbs, nouns and other words assigned to those motions. Next, through canonical correlation analysis [18], these spaces are reconstructed such that a strong correlation is formed between motion symbols and word labels. Using these spaces, we propose a method for searching for motion data based on word labels. This serves as a fundamental technique for extracting the necessary motions from a database and reusing them.

2. Mapping between motions and word labels

Research on intelligent robots through conversion of bodily senses or movements into symbols is being conducted in robotics. These approaches encode the continuous spatio-temporal data of motions into the low dimensional parameters of motion primitives, and these parameters allow robots to classify the motions into the motion primitives. However, the motion primitives represented in the parameters cannot be intuitively understood by humans. Humans have acquired language through the process of evolution. We can understand motions in same expression that others can do by using the language. The mapping between the motion parameters and words is crucial to establishing communication between robots and humans. This section describes an approach to extract the mapping between motions and word labels. The motion data is encoded into a Hidden Markov Model (HMM), which is referred to as "motion symbol". The motion data is also given word labels by human annotators. Relation between the motions and the word labels is extracted from the training pairs of the motion symbols and the word labels as shown in Fig. 1.

2.1. Extracting correlation between motions and words

Fig. 2 shows the overview of mapping between human whole body motions and word labels. The human motion primitive data are encoded in Hidden Markov Models (HMMs). Each HMM is referred to as "motion symbol" since it represents spatio-temporal features of its corresponding motion primitive. Dissimilarity between each motion symbol can be calculated by using the Kullback–Leibler information.

$$d(\lambda_i, \lambda_j) = \sum_{\hat{\mathbf{o}}_i^{(k)}:\, k=1,2,3,\ldots,N} \frac{1}{N} \left\{ \ln P(\hat{\mathbf{o}}_i^{(k)}|\lambda_i) - \ln P(\hat{\mathbf{o}}_i^{(k)}|\lambda_j) \right\} \quad (1)$$

$d(\lambda_i, \lambda_j)$ is the Kullback–Leibler information from motion symbol λ_i to motion symbol λ_j. $\hat{\mathbf{o}}_i^{(k)}$ is the kth motion data that the motion symbol λ_i generates by the Monte Carlo method. $P(\hat{\mathbf{o}}_i^{(k)}|\lambda_j)$ is the likelihood that motion symbol λ_j generates the motion data $\hat{\mathbf{o}}_i^{(k)}$. The Kullback–Leibler information does not necessarily satisfy the symmetry. In Eq. (2), $d(\lambda_i, \lambda_j)$ and $d(\lambda_j, \lambda_i)$ are summed to obtain $D(\lambda_i \parallel \lambda_j)$, which satisfies the symmetry.

$$D(\lambda_i \parallel \lambda_j) = \frac{d(\lambda_i, \lambda_j) + d(\lambda_j, \lambda_i)}{2}. \quad (2)$$

This is defined as the distance between motion symbol λ_i and motion symbol λ_j. All of the motion symbols are arranged as points on a multidimensional space such that the distance between all the motion symbols is satisfied. The coordinates of the point in the multidimensional space corresponding to motion symbol λ_i are taken as $\boldsymbol{x}_i$, and this position is found such that the following error function is minimized.

$$T = \sum_{\forall i,j} \frac{\left(D(\lambda_i \parallel \lambda_j)^2 - d_{ij}^2\right)^2}{4D(\lambda_i \parallel \lambda_j)^2} \quad (3)$$

$$d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|. \quad (4)$$

Here, the multidimensional scaling proposed by Takane et al. [19] is used. The error function T is represented by a fourth-order polynomial in coordinate $\boldsymbol{x}_i$ of motion symbol λ_i. The optimal position of the motion symbol, which minimizes the error function T, can be found by the Newton–Raphson method. This process constructs the motion symbol space based on dissimilarity between the motion symbols by the multidimensional scaling.

Multiple word labels are manually assigned to the same motion primitive data that the HMM encodes into the motion symbol. The word labels are descriptive of the motion primitive. A set of the word labels is represented by a feature vector with binary elements taking value 1 if the corresponding word label is present in the set

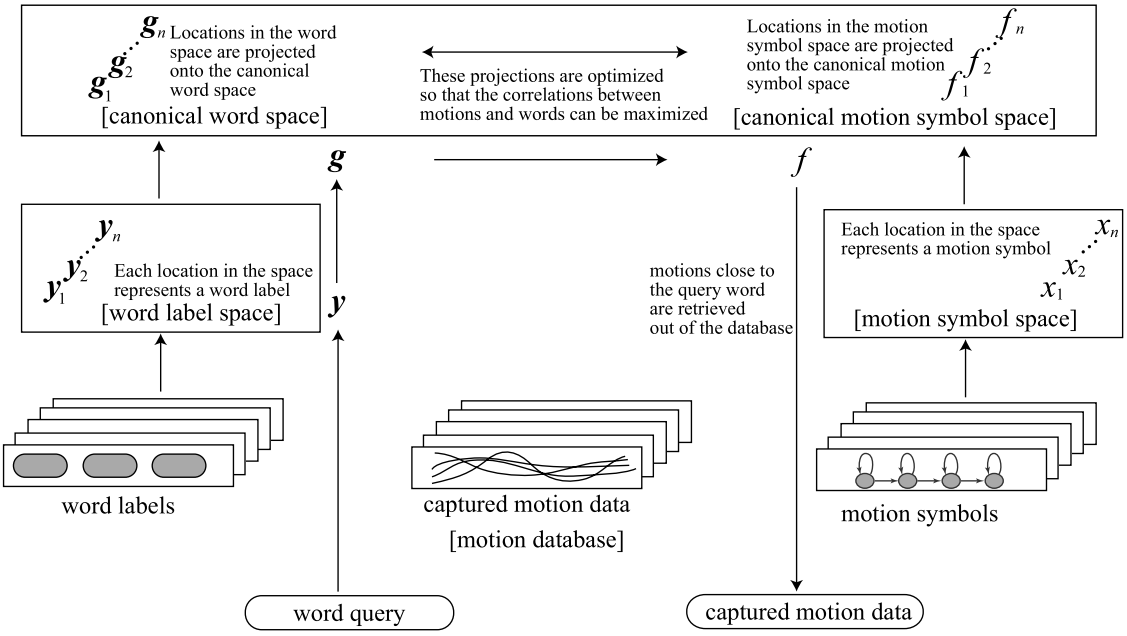

Fig. 1. The overview of motion database and its application to retrieve captured motion data from word label queries. The motion data is encoded into an Hidden Markov Model, which is referred to as "motion symbol". Human annotators attach multiple word labels to this motion data. The motion symbol and the word labels are converted to feature vectors in the motion symbol space and word label space. These two feature vectors are linearly transformed into new vectors so that correlation between the transformed vectors can be maximized. These derived vectors can be used to retrieve motion data corresponding to word label queries. The word label queries are represented by the feature vector. This feature vector is converted to the feature in the motion symbol space. The motion symbols, whose feature vectors are close to converted feature vector, are searched for in the motion database. The motion data, which are encoded into these closest motion symbols, can be retrieved.

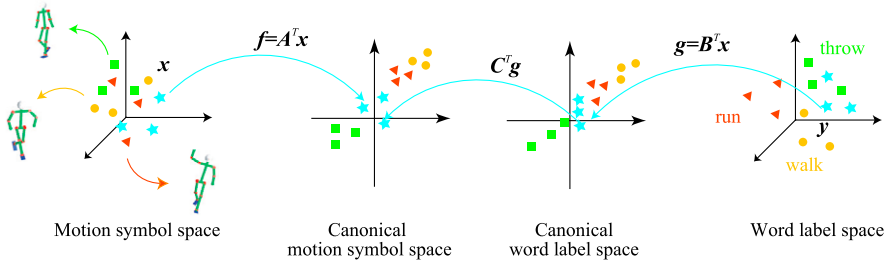

Fig. 2. Motion patterns are symbolized by HMMs. Motion symbols form "motion symbol space" based on dissimilarities among the symbols. The motion patterns are also given motion words, which represent the motions. The motion words form "word space". Canonical correlation analysis establishes the canonical spaces for the motions and the words. These spaces make it possible to retrieve motions from word queries.

of the word label, and 0 if the word label is not present. The number of dimensions of the feature vector is equal to the number of the different word labels. For example, four word labels "run", "walk", "throw" and "ball" can be used, and two word labels, "throw" and "ball" are assigned to motion primitive data. In this case, feature vector $\boldsymbol{y} = (0, 0, 1, 1)^T$ corresponding to the motion primitive is derived. The position of the word labels in the word label space can be defined as the feature vector $\boldsymbol{y}$.

The motion primitive data $\{\mathbf{O}_i\}_{i=1}^n$ can be represented not only by the position $\{\boldsymbol{x}_i\}_{i=1}^n$ in the motion symbol space but also by the position $\{\boldsymbol{y}_i\}_{i=1}^n$ in the word label space. n is the number of the motion primitive data. By using training dataset of the positions, $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$, these positions are linearly projected onto the basis vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ by Canonical Correlation Analysis (CCA) such that the correlation between the resulting positions is maximized. The matrices, $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be derived by arranging the vectors of the motion symbols and the word labels along the column direction respectively:

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^T \tag{5}$$

$$\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]^T. \tag{6}$$

The correlation r_{fg} between the projections of $f_i = \boldsymbol{x}_i^T \boldsymbol{a}$ and $g_i = \boldsymbol{y}_i^T \boldsymbol{b}$ can be given by

$$r_{fg} = \frac{\boldsymbol{a}^T \Sigma_{XY} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^T \Sigma_X \boldsymbol{a}} \sqrt{\boldsymbol{b}^T \Sigma_Y \boldsymbol{b}}} \tag{7}$$

$$\Sigma_X = \frac{1}{n-1} \boldsymbol{X}^T \boldsymbol{X} \tag{8}$$

$$\Sigma_Y = \frac{1}{n-1} \boldsymbol{Y}^T \boldsymbol{Y} \tag{9}$$

$$\Sigma_{XY} = \frac{1}{n-1} \boldsymbol{X}^T \boldsymbol{Y}. \tag{10}$$

The CCA seeks to find the optimal basis vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ to maximize the correlation r_{fg} subject to following constraints that the covariances of the resulting projections take value 1.

$$\boldsymbol{a}^T \Sigma_X \boldsymbol{a} = 1 \tag{11}$$

$$\boldsymbol{b}^T \Sigma_Y \boldsymbol{b} = 1. \tag{12}$$

The CCA optimization problems can be simplified to the following eigenvalue problems.

$$\Sigma_{XY} \Sigma_Y^{-1} \Sigma_{YX} \boldsymbol{a} = \alpha^2 \Sigma_X \boldsymbol{a} \tag{13}$$

$$\Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}\boldsymbol{b} = \alpha^2\Sigma_Y\boldsymbol{b} \tag{14}$$

$\{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k\}$ and $\{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_k\}$ denote the eigenvectors corresponding to the largest eigenvectors $\alpha_1^2 \geq \alpha_2^2 \geq \cdots \geq \alpha_k^2$. The projections $\boldsymbol{x}^T\boldsymbol{a}_i$ and $\boldsymbol{y}^T\boldsymbol{b}_i$ are ith dimensional coordinates of a motion symbol and word labels in the canonical spaces respectively. The correlation between these two coordinates becomes α_i. The matrices $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_k]$ and $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_k]$ project the motion symbols and word labels onto k-dimensional canonical spaces.

2.2. Retrieval of motions from word labels

Reuse technologies of prerecorded motion data improve analysis of human motions, edition of character animation, and interaction between a human and a robot. The reuse of the motion data archive has remained challenging. The reuse technologies need the effective search for specific motion data in the archive. The conventional search has been depending on recording date labels, or names of motion data files. Users have to know the date or names of their necessary motion data. The relation between motion symbols and word labels through the CCA leads to a useful search for motion data from queries of word labels. The users only have to input word labels in order to find prerecorded motion data.

We assume that the relation between the project $\boldsymbol{f} = \boldsymbol{A}^T\boldsymbol{x}$ and the project $\boldsymbol{g} = \boldsymbol{B}^T\boldsymbol{y}$ can be represented as the linear transformation, $\boldsymbol{f} = \boldsymbol{C}^T\boldsymbol{g}$, where $\boldsymbol{C}$ is a transformation matrix. The transformation matrix can be found such that it can minimize the error S between the project $\boldsymbol{A}^T\boldsymbol{x}$ and the project estimated by the transformation $\boldsymbol{C}^T\boldsymbol{g}$.

$$S = \sum_{i=1}^{n} \frac{1}{2}(\boldsymbol{f}_i - \boldsymbol{C}^T\boldsymbol{g}_i)^T(\boldsymbol{f}_i - \boldsymbol{C}^T\boldsymbol{g}_i). \tag{15}$$

The optimal transformation matrix can be represented as

$$\boldsymbol{C} = \Sigma_{FG} \tag{16}$$

$$\Sigma_{FG} = \frac{1}{n-1}\boldsymbol{F}^T\boldsymbol{G} \tag{17}$$

$$\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n]^T \tag{18}$$

$$\boldsymbol{G} = [\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_n]^T. \tag{19}$$

Word labels are given as a search query, and converted to a binary feature vector $\boldsymbol{y}$. The projection $\boldsymbol{g} = \boldsymbol{B}^T\boldsymbol{y}$ can be derived by projecting the binary feature onto the canonical space. This projection is transformed into the feature vector $\boldsymbol{C}^T\boldsymbol{g}$ in the canonical space of motion symbols. Motion symbols whose feature vectors are close to the transformed feature can be retrieved from the database. This operation makes it possible to search for motion data corresponding to a query of word labels.

In addition to the motion retrieval, the relation between motion symbols and word labels through the CCA is also used for motion recognition. The relation between the project $\boldsymbol{f} = \boldsymbol{A}^T\boldsymbol{x}$ and the project $\boldsymbol{g} = \boldsymbol{B}^T\boldsymbol{y}$ can be represented as the linear transformation, $\boldsymbol{g} = \boldsymbol{D}^T\boldsymbol{f}$, where $\boldsymbol{D}$ is a transformation matrix. The transformation matrix can be found such that it can minimize the error S between the project $\boldsymbol{B}^T\boldsymbol{y}$ and the project estimated by the transformation $\boldsymbol{D}^T\boldsymbol{f}$. The optimal transformation matrix can be derived by a similar fashion described above. Observed motion is recognized as a motion symbol that is most likely to generate the observation. The feature vector $\boldsymbol{x}$ representing the position of this motion symbol in the motion symbol space can be derived. This feature vector is projected to $\boldsymbol{f} = \boldsymbol{A}^T\boldsymbol{x}$. This projection is further transformed into the feature vector $\boldsymbol{D}^T\boldsymbol{f}$ in the canonical space of word labels. Words whose feature vectors are close to the transformed feature can be found out of the database. The observed motion can be consequently recognized as the words.

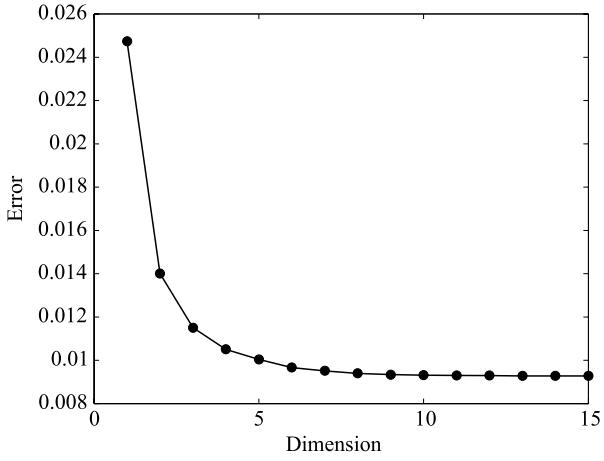

Fig. 3. Relationship between the number of dimensions of the motion symbol space and error between the distance calculated from the Kullback–Leibler and the distance measured in the motion symbol space. The error converges around $d = 10$ as the number of dimensions in the motion symbol space is incremented.

3. Experiments

3.1. Dataset of motions and descriptive word labels

The positions of 34 markers attached to a performer were measured by using an optical motion capture system. The sampling rate of the motion capture system was set to 30 Hz. Joint angles, the height in the vertical direction, speed in the horizontal direction, roll, pitch and yaw of the body trunk were calculated by inverse kinematics based on a human body model with 20 degrees of freedom. Human motion primitive data is defined as a segment of a time series of 26-dimensional feature vector of these joint angles and trunk data. The trunk data include its vertical position p_z, roll angle θ_x, pitch angle θ_y in the global coordinate system, and its horizontal speeds v_x, v_y and yaw angular velocity ω_z in its local coordinate system. The motion primitive data were encoded into HMMs. The HMMs are referred to as motion symbols. The number of nodes in the HMM was set to 30, and 618 motion primitives were measured. This means that 618 motion symbols were derived.

The Kullback–Leibler distances between all the motion symbols were calculated. The motion symbol space was constructed by locating the motion symbols in a multidimensional space such that the error between the Kullback–Leibler distances and the distances measured in this space is minimized. Fig. 3 shows the relationship between the number d of dimensions of the motion symbol space and the error T_d in the Kullback–Leibler distance and the distance between the motion symbols located in the multidimensional space. The error decreases as the number of dimensions of the motion symbol space increases. The error ratio, $\gamma = \frac{T_{d+1}-T_d}{T_d}$, is 0.001 at the 10 dimensions. The error was found to converge at the 10 dimensions. Therefore, we set the number of dimensions of the motion symbol space to 10. The 10 dimensional motion symbol space was used hereafter. The position $\boldsymbol{x}$ of the motion symbol in the motion symbol space represents each motion primitive.

Human annotators assigned descriptive word labels to the motion primitive data. The annotators did not select word labels from the prepared set of word labels, but found the words of noun or verb relevant to the motion data and attached them to the motion. The number of different word labels attached to all the motion primitive data was 253. 253 dimensional binary feature vector $\boldsymbol{y}$ represents the word labels assigned to each motion primitive. Note that several word labels were assigned to each motion primitive. For examples, two word labels, "tennis" and "swing", were assigned to the action of swinging a tennis racket. Moreover,

Table 1
Examples of motion and attached word labels.

Motion	Word labels	Motion	Word labels	Motion	Word labels
jump1	"jump"	tennis swing2	"tennis" "swing"	march walk2	"march" "walk"
jump2	"jump"	tennis swing3	"swing"	march walk3	"march"
jump3	"leap"	tennis swing4	"tennis" "swing"	left highkick1	"left leg" "high kick" "kick"
jump4	"leap"	tennis swing5	"tennis" "swing"	left highkick2	"left leg" "high kick"
jump forward1	"leap"	tennis smash1	"tennis" "smash"	left highkick3	"left leg" "kick"
jump forward2	"forward" "jump"	tennis smash2	"tennis" "smash"	right lowkick1	"right leg" "low kick" "kick"
jump forward3	"forward" "leap"	tennis smash3	"tennis" "swing"	right lowkick2	"right leg" "low kick"
jump down1	"jump down"	bow1	"bow"	right lowkick3	"right leg" "kick"
jump down2	"leap" "jump down"	bow2	"bow"	sweep bloom1	"bloom" "sweep"
jump down3	"leap" "drop"	bow3	"salute"	sweep bloom2	"bloom" "sweep"
play guitar1	"guitar" "play"	bow deeply1	"bow" "apologize"	sweep bloom3	"bloom" "sweep" "clean"
play guitar2	"guitar" "play"	bow deeply2	"bow"	clean1	"cleaner" "vacuum"
play guitar3	"guitar" "perform"	bow deeply3	"salute"	clean2	"cleaner" "vacuum"
play violin1	"violin" "play"	walk1	"walk"	clean3	"cleaner" "clean"
play violin2	"violin" "play"	walk2	"walk"	read book1	"book" "read"
play violin3	"violin" "perform"	walk3	"walk"	read book2	"book" "read"
tennis swing1	"tennis" "swing"	march walk1	"march" "walk"	read book3	"read"

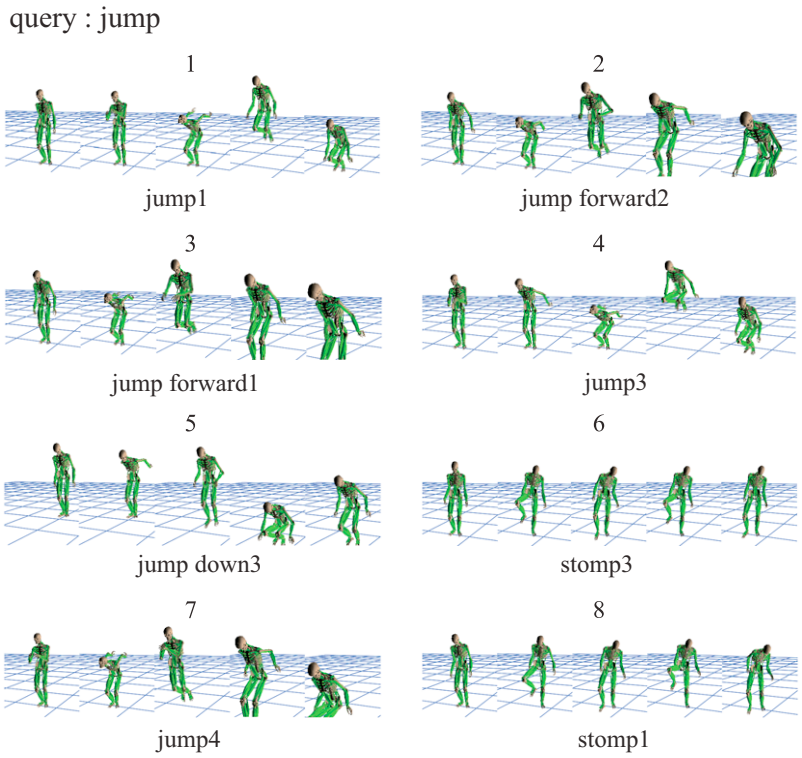

Fig. 4. Eight motions are retrieved from the word query "jump" out of motion dataset. The motion, which is given same word as the query, can be retrieved. These experiment results mean that queried word labels do not correspond to only one motion, but that they can refer to multiple motion symbols in the proposed framework.

same word labels were not necessarily assigned to similar motion primitives. One word label "jump" was assigned to the action of jumping, and different word label "leap" was assigned to another action of jumping. Table 1 shows instances of motion primitives and word labels attached to them. Two word labels "book" and "read" are attached to the motion "read book1" referring to the action of reading a book. Word labels "left leg", "high kick", and "kick" are attached to the motion "left highkick1" referring to the action of high-kicking.

The position $\boldsymbol{x}$ of the motion symbol in the motion symbol space and the binary feature $\boldsymbol{y}$ of the word labels were derived. The CCA forms the canonical spaces of the motion symbols and the word labels by using the pairs $(\boldsymbol{x}, \boldsymbol{y})$ as training data. The transformation matrix $\boldsymbol{C}$ can be also derived by using these data. The transformation matrix allows for projection from the canonical space of the motion words onto the canonical space of the motion symbols.

The relationship between the motion symbols and the word labels through the CCA can be applied to the retrieval of motion data from a query of word labels. This is implemented by representing the query of the word labels by the binary feature vector $\boldsymbol{y}$, projecting the feature vector onto the canonical space of the word labels to derive a feature vector $\boldsymbol{g} = \boldsymbol{B}^T\boldsymbol{y}$, transforming the project $\boldsymbol{g}$ to the feature vector $\boldsymbol{C}^T\boldsymbol{g}$ in the canonical space of the motion symbols by using the transformation matrix, and finding motion symbols which are close to $\boldsymbol{C}^T\boldsymbol{g}$. This implementation retrieves motion data corresponding to a query of word labels.

3.2. Qualitative experimental results

Figs. 4 and 5 show the experimental results of the motion retrieval. The displayed motions are ranked in ascending order of distance between the query and the motion. When a query of

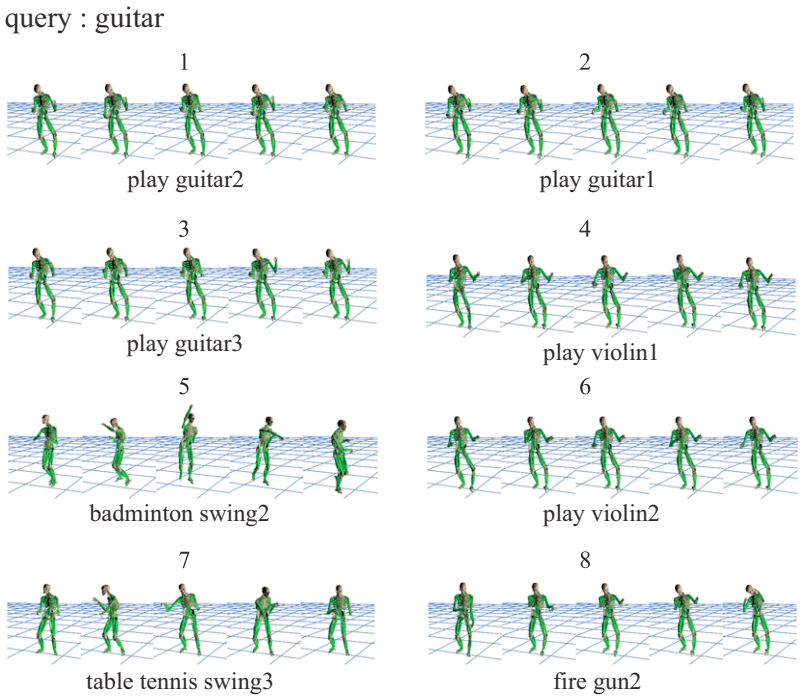

Fig. 5. Eight motions are retrieved from the word query "guitar" out of motion dataset. The motion, which is given same word as the query, can be retrieved. These experiment results mean that queried word labels do not correspond to only one motion, but that they can refer to multiple motion symbols in the proposed framework.

Table 2
Search for motions corresponding to words "march", "cleaner", "book read" and "high kick", respectively.

march		cleaner		book, read		high, kick	
Result	Similarity	Result	Similarity	Result	Similarity	Result	Similarity
march walk3	1.000	clean1	1.000	readbook1	1.000	left highkick2	1.000
march walk1	1.087	clean3	1.274	readbook2	1.000	right highkick1	1.069
march walk2	1.431	clean2	1.393	readbook3	1.195	left highkick1	1.070
run2	1.488	sweep broom3	5.873	PC3	1.674	left highkick3	1.097
run1	1.534	crouch2	6.144	PC2	1.889	left lowkick3	1.111
walk3	1.595	sweep broom1	6.226	PC1	1.891	left lowkick2	1.122
run3	1.610	sit down1	6.263	sit down1	2.459	right highkick3	1.158
crap3	1.626	bow2	6.301	do dish2	2.699	left lowkick1	1.188
shakehand2	1.631	crouch1	6.321	catch3	2.714	run3	1.196
welcome2	1.646	do dish1	6.357	sitdown2	2.725	do dish1	1.198

the word label "jump" was given, the motion "jump1" and the motion "jump forward2", both of which the word label "jump" were attached to, were retrieved. The motion "jump forward1", the motion "jump3", the motion "jump down3" were also retrieved. The word label "jump" was not attached to these motions, but the word label "leap" was attached. These motions can be categorized as the motion of jumping, and they are relevant to the word label "jump". Fig. 6 shows generated trajectories of the trunk height, left knee joint angle, and right knee joint angle, which are generated by HMMs corresponding to "jump1", "jump forward2" and "jump forward1". The trajectory distributions and the average trajectories are displayed in shaded graphs and in solid graphs respectively. The figure illustrates that these motions are very similar to each other. The retrieval results look reasonable. When a query of the word label "guitar" was given, the motion "play guitar2", the motion "play guitar1", and the motion "play guitar3" were retrieved. The motion word "guitar" was attached to these retrieved motions. The motion "play violin1" with the word labels "violin" and "play" attached was retrieved at the fourth rank. The motion of playing the guitar and the motion of playing the violin look similar. The motion "play violin1" was relevant to the word label "guitar". The majority of motions retrieved at the top ranks look good. This experiment clarified that motions relevant to the word label can be retrieved by using our proposed framework.

Another experimental results of motion retrieval are shown in Table 2, where ten motions were retrieved given word labels "march", "cleaner", "book, read" and "high, kick" as retrieval queries, and the similarities between the motion and the query were calculated as the distances divided by the shortest distance such that the shortest distance becomes one. Note that the distances were measured in the canonical space of the motion symbols. The motion "march walk3", the motion "march walk1" and "march walk2" were retrieved at top ranks given the word label "march" as a query. The word label "march" was attached to these three motions. The motion "run2", the motion "run1" and the motion "walk3" were retrieved at the following ranks. Although these three motions were not given the word label "march", they look similar to the motion of marching. The averages of the similarity of motion "march walk" and "run" were 1.17 and 1.54 respectively. The motion "clean1", the motion "clean3" and the motion "clean2", which the word label "cleaner" was attached to, were retrieved given the word label "cleaner". The motion "sweep broom3" was retrieved at the fourth rank. The word label "cleaner" was not attached to this motion. The similarities for the motion "clean2" and for the motion "sweep broom3" were 1.39 and 5.87 respectively. The averaged similarities of the motion "clean" and the motion "sweep broom" were 1.22 and 6.16. The motion "sweep broom3"

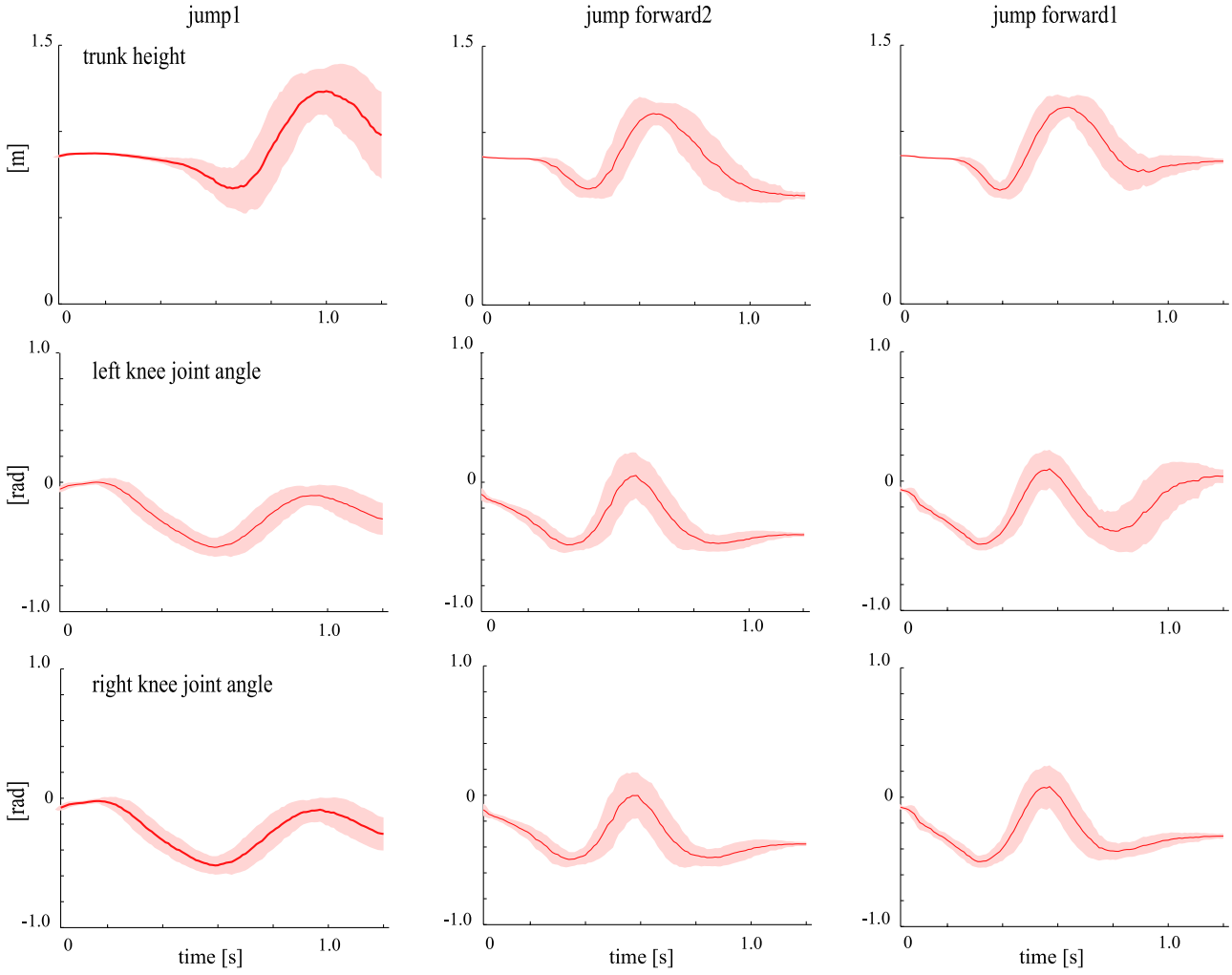

Fig. 6. Motions are generated by sampling-based method from HMMs, which are retrieved from the query of the word label"jump". The average and distribution of the trajectories of the trunk height, the left knee joint angles, and the right knee joint angles are displayed.

is much more irrelevant to the word label "cleaner". However, the word label "clean" was attached to both the motion "clean3" and "sweep broom3". The attachment of these words establishes the relationship between the word label "cleaner" and the motion "sweep broom3", and then makes it possible to retrieve the motion which the queried word label was not attached to. Given two word labels "book" and "read" as a query, the motions "read book1" and "read book2" were retrieved at the first and second ranks. Both the word label "book" and the word label "read" were attached to the retrieved motions. The motion "read book3", which only one word label "read" was attached to, was retrieved at the third rank. The motions retrieved at the following ranks, "PC3", "PC2", "PC1", and "sit down1" look similar to the motion of reading a book, since a performer sat at the desk during recording these motions. Note that the motion "read book" captured full body movement of a performer reading a book during sitting on a chair, and the motion "PC" captured whole body movement of a performer typing at a keyboard during sitting on a chair. The averaged similarities of the motion "readbook" and the motion "PC" were 1.07 and 1.82 respectively. The motions, which the word labels "high" and "kick" were attached to, were retrieved given the word label "high" and the word label "kick" as queries. The motions retrieved at the top ranks look quite relevant to the queried motion words. The proposed retrieval can find the motions relevant to the queried word labels even if the queries are not attached to them.

3.3. Quantitative experimental results

We evaluated the results of retrieving motions given queried word labels. We adopted *Recall (R)* and *Precision (P)* as the evaluation measures. The motion retrieval was tested on all the word labels, which were attached to captured motion primitives. This means that the number of test queries was 618. Note that some overlapped queries were permitted. The retrieved motion, which the queried word labels are attached to, is counted as *True Positive (TP)*. The Recall measures the percentage of TP in all the motions in the database that the queried word labels are assigned to. The Precision measures the percentage of TP in all the retrieved motions. The binary feature for queried word labels is projected onto the canonical space of the motion symbols. The motion whose distance to the project in the canonical space is below the threshold is retrieved. The threshold was increased from 0.2 to 5.0 in increments of 0.1. Fig. 7 shows the relationship between the Recall and the Precision. The F measure, which is the weighted harmonic mean of the Recall and the Precision, and can be simplified to $F = \frac{2PR}{P+R}$, was 0.598 when the Recall was equal to the Precision.

We evaluated the retrieval results by using another measure called *Mean Average Precision (MAP)*. MAP is the average of Precision values obtained after each motion with the query attached is retrieved. When the motion retrieved at the rth rank is counted as TP, Precision $P(r)$ is obtained from tops r retrieved

motions. The retrievals are iterated until all the motions with the query attached are found. MAP is defined as average of all the Precision $P(r)$. $rel(r)$ returns 1 if the motion retrieved at the rth rank is counted as TP, otherwise 0.

$$AP = \frac{1}{N_q} \sum_{r=1}^{N_r} rel(r) PR(r), \tag{20}$$

where N_q is the number of motions that the queried word labels are attached to. All the motions with the query attached are found until the retrievals are iterated until the N_rth rank. All the motions that the query is assigned to are found from until the higher ranks, MAP becomes larger. MAP was 0.822 in this retrieval experiment.

Additionally, we tested our framework on recognizing observed motions as word labels. We chose 617 pairs of motion symbol and the its descriptive word labels as training data in order to construct the motion symbol space and the word label space. One remaining pair was used as a test data. This motion symbol is supposed to be recognized as three word labels. More specifically, three word labels that is the most related to the motion symbol are founded. The found set of word labels including one of the word labels attached to the motion symbol is counted as correct. The average recognition rate was 0.72. Our framework is expected to associate multiple motion symbols and word labels, and to recognize a motion as its relevant word labels that may be attached to this motion. Therefore, we selected three word labels as motion recognition result.

In this paper, 618 motion symbols are projected onto a motion symbol space, and each of them is expressed by 10 dimensional feature vector. However, 253 different word labels are used, and each of them is expressed by 253 dimensional binary feature vector. The more different word labels are assigned to the motion primitive data, the larger size of the binary feature vector is required. To deal with this problem, the feature vector of a word needs to be expressed by a low dimensional feature vector applying dimensionality reduction to the words based on the their co-occurrence frequencies in the documents [16,17]. We should recorded more motion data and their descriptive word labels and test the scalability of our proposed framework.

4. Conclusion

The contribution of this study can be summarized as follows.

1. We proposed a framework for correlating motion symbols learned by applying an HMM to full-body motion data with word labels assigned to those motions. We constructed a motion symbol space based on the dissimilarity between motion symbols, and represented the motion symbols as points in the symbol space. Word labels assigned to those motions were represented as binary vectors composed of elements representing the presence or absence of different word labels. Furthermore, canonical correlation analysis was used to create a linear mapping of the vectors such that the correlation is maximized between the feature vectors representing the positions of symbols in the motion symbol space and the binary vectors for word labels. In this way, canonical spaces could be constructed consisting of points representing the projections of motion symbols and word labels.
2. We proposed a method for searching for motions corresponding to word labels given as input. The input word labels are projected onto the canonical space of motion symbols, and the motion symbols located near the projection point can be extracted as search results. A single word label can extract multiple motions based on their distance in the canonical space.

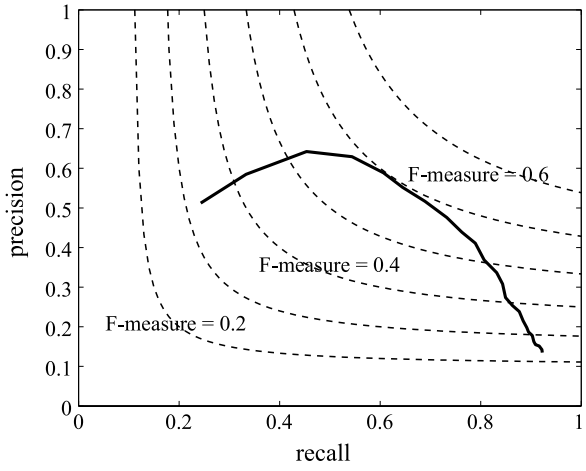

Fig. 7. Recall and precision curve. The retrieval depends on the distance threshold. The motion, whose distance to the word query is smaller than this threshold, is retrieved out of the motion database. This curve can be derived by plotting recall and precision as this distance threshold is incremented.

3. Using training data consisting of 618 symbols for full-body motions measured with a motion capturing system and 253 types of word labels assigned to those motions, we constructed canonical spaces for motion symbols and word labels. Furthermore, we derived a transformation matrix for mapping points from the canonical space of word labels to the canonical space of motion symbols. We verified that given a word label as input, the search extracts not only the motions to which this label is assigned, but also the motions that can be derived by association with that word label. The F-measure calculated from the Recall and Precision of the search was 0.598, and Mean Average Precision for all word labels was 0.822.

Acknowledgments

This research was supported by PREST "Information Environment and Humans", Japan Science and Technology Agency, and Grant-in-Aid for Young Scientists (A) (26700021), Japan Society for the Promotion of Science.

References

[1] F.D. Saussure, Course in General Linguistics, McGraw-Hill Book Company, 1966.
[2] C. Breazeal, B. Scassellati, Robots that imitate humans, Trends Cogn. Sci. 6 (11) (2002) 481–487.
[3] B. Argall, S. Chernova, M. Veloso, B. Browning, A survey of robot learning from demonstration, Robot. Auton. Syst. 57 (5) (2009) 469–483.
[4] T. Inamura, I. Toshima, H. Tanie, Y. Nakamura, Embodied symbol emergence based on mimesis theory, Int. J. Robot. Res. 23 (4) (2004) 363–377.
[5] A. Billard, S. Calinon, F. Guenter, Discriminative and adaptive imitation in uni-manual and bi-manual tasks, Robot. Auton. Syst. 54 (2006) 370–384.
[6] M. Haruno, D. Wolpert, M. Kawato, MOSAIC model for sensorimotor learning and control, Neural Comput. 13 (2001) 2201–2220.
[7] M. Okada, K. Tatani, Y. Nakamura, Polynomial design of the nonlinear dynamics for the brain-like information processing of whole body motion, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2002, pp. 1410–1415.
[8] J. Tani, M. Ito, Self-organization of behavioral primitives as multiple attractor dynamics: A robot experiment, IEEE Trans. Syst. Man Cybern. A 33 (4) (2003) 481–488.
[9] A.J. Ijspeert, J. Nakanishi, S. Shaal, Learning control policies for movement imitation and movement recognition, Neural Inf. Process. Syst. 15 (2003) 1547–1554.
[10] H. Kadone, Y. Nakamura, 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, 2005, pp. 2900–2905.
[11] W. Takano, H. Imagawa, D. Kulic, Y. Nakamura, Organization of behavioral knowledge from extraction of temporal-spatial features of human whole body motions, in: Proceedings of the IEEE International Conference on Biomedical Robotics and Biomechatronics, 2010, pp. 52–57.

[12] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–286.
[13] W. Takano, K. Yamane, Y. Nakamura, Capture database through symbolization, recognition and generation of motion patterns, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2007, pp. 3092–3097.
[14] T. Inamura, H. Tanie, Y. Nakamura, From stochastic motion generation and recognition to geometric symbol development and manipulation, in: Proceedings of IEEE International Conference on Humanoid Robots, 2003, 1b–02.
[15] G. Salton, A. Wong, C.S. Yang, A vector space model for automatic indexing, Commun. ACM 18 (11) (1975) 613–620.
[16] S. Deerwester, S. Dumais, G.W. Furnas, T.K. Landuaer, R. Harchman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (3) (1990) 391–407.
[17] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Mach. Learn. 42 (2001) 177–196.
[18] H. Hotelling, Relation between two sets of variates, Biometrica (1936).
[19] Y. Takane, F.W. Young, J.D. Leeuw, Nonmetric individual differences multidimensional scaling: An alternating least squres method with optimal scaling features, Psychometrika 42 (1977) 7–67.

Wataru Takano is an Assistant Professor 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 a 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 in 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.

Seiya Hamano received the B.S. degree and M.S. degree from Mechano-Informatics, the University of Tokyo, Japan, in 2010 and 2012 respectively. He currently works for Yahoo Japan. His research interests are natural language processing, robot intelligence and human–robot interaction.

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.