Abstraction Multimodal Low-Dimensional Representation from High-Dimensional Posture Information and Visual Images

Tatsuya Hirose
Kyoto University

Tadahiro Taniguchi
Ritsumeikan University

Abstract

Imitative learning is an effective method for robots to obtain a novel movement from a person demonstrating many kinds of movement. Many problems need to be solved, however, before a robot can achieve imitative learning. One problem is how to convert visual information on the demonstrator's motion to kinematic posture information for the learner. This is referred to as a correspondence problem and we have focused on this problem in this study. To solve it, we focus on the formation of a low-dimensional representation that integrates sensory information from two different modalities. We propose a computation method for constructing the low-dimensional representation combining posture information and visual images by using kernel canonical correlation analysis (KCCA). Using this method, a robot becomes able to estimate posture information from visual images in a bottom-up way. Using several experiments we show how effective our proposed method is in estimating kinematic information.

Key Words: Kernel Canonical Correlation Analysis, Imitation Learning, Body Schema

1 Introduction

If robots can imitate a demonstrator's movement just by observing it visually, they can obtain novel movements without the supervisor's or designer's specification of the concrete coordinates of individual joint angles for each time.

Many studies have been done on robotic imitative methods, but many difficulties still exist. Children even as young as one year old imitate their parents' motion and simultaneously obtain many kinds of motion. Computational understanding of imitative learning is important in understanding human cognitive abilities and the developmental process supporting imitative learning capability.

Breazeal and Scassellati focused on two fundamental problems regarding robotic imitative learning [1].

The first problem was how the robot knows what to imitate. When a robot observes a person exhibiting a sequence of motions, how does the robot know

which part of this sequence should be imitated? Taniguchi et al. studied an algorithm by which a robot divides demonstrator motion into meaningful parts and learns these [2]. Breazeal et al. studied an attention system that plays an important role in imitative learning using a robot with a face detector, color detector and motion detector [3].

The second problem they investigated was how the robot knows how to imitate. When a robot imitates a target motion, how does the robot convert the observed motion into a series of action commands that the robot can carry out? Alissandrakis, Nehaniv, and Dautenhahn also studied this problem, which they called a *correspondence problem* [5]. This correspondence problem occurs when the robot maps a visual image to its joint angles [4]. Mapping is generally difficult because there are many mapping possibilities. Specifically, when a robot tries to imitate a person, the robot has to deal with differences between the robot and human body parts and kinematics. There are many combinations of robot and human body parts. In addition, the degrees of freedom (DOF) and coordinate system of their body's motor systems also differ. Using a game of chess as the setting, Alissandrakis, Nehaniv, and Dautenhahn studied an algorithm for imitating those that move differently[5]. In a realistic robotic imitative learning problem, a robot has to estimate a demonstrator's postural information and project it onto its body coordinates without knowing the DOF of the demonstrator's body system or the parameters of individual joints. This paper introduces a learning method in which a robot automatically obtains appropriate maps from visual information about a demonstrator's movement to its posture information in a bottom-up way.

1.1 Posture Estimation

Various methods to estimate a posture from an image of a demonstrator have been studied over the years. There are two main approaches.

The first approach is to use a human body model directly. Lee, Cohen, and Jung studied a method for estimating model parameters using a particle filter [6]. Shotton et al. proposed an algorithm to quickly estimate human joint angles from human body images [7]. This algorithm can estimate joint angles in real time. A camera, depth sensor, and human body model work collaboratively and collaborative calculation enables the estimation of joint angles quickly. These methods cannot, however, be applied to creatures that have unknown or different body structures. Robots can have various body structures and kinematics, so each type of robot has to obtain its own mapping from visual image to posture information. In other words, a developmental robot has to obtain or improve its body model. In addition, providing a computational model that describes how human beings are constructed and how they improve themselves is important to understanding how children can obtain a human body model.

The second approach does not use a human body model. Yamane et al. proposed an algorithm for marker-free motion capture using multiple cameras [8]. In this algorithm, the body model is estimated automatically. Agarwal et al. conducted an experiment to estimate human joint angles from the shape

context of the human silhouette using ridge regression, a relevance vector machine, and a support vector machine [9]. Grauman et al. proposed an algorithm for estimating human joint angles from the silhouette of a person using the low-dimensional representation obtained by probabilistic principal component analysis (PCA) [10]. Since only linear mapping is used in probabilistic PCA, PCA cannot deal with nonlinearity in the human body structure. To solve this problem, Henrik et al. proposed an algorithm for estimating posture information using a low-dimensional representation obtained by Gaussian process latent variable models (GPLVM), which is known as a probabilistic nonlinear PCA approach [11]. The Gaussian process (GP) is a Bayesian extension of kernel regression (KR). The relation between input and output data is modeled by the GP by using a Gaussian stochastic process [12]. GPLVM assumes that two or more high-dimensional data sets are generated from one shared low-dimensional data set by using nonlinear mappings. In GPLVM, nonlinear maps and the hidden low-dimensional space base are estimated using the GP. Posture information and visual images are assumed to be generated from a hidden low-dimensional space. Henrik et al. showed that human joint angles can be estimated from a silhouette using information on its probability. Human joint angles are also estimated from a sequence of silhouettes. This algorithm has some problems, however. It takes much computation time to converge and it is difficult to obtain a global optimal solution because a gradient method has to be used to locate inversely mapped observed data in low-dimensional space. This optimization process has huge solution space. To solve this problem, we propose using kernel canonical correlation analysis (KCCA), instead of using GPLVM, to obtain the low-dimensional hidden space. KCCA obtains a global solution for a generalized eigenvalue problem at reasonable computational cost. We explain here how we applied KCCA to high-dimensional human upper body data to obtain a human body model in order to construct a low-dimensional representation of the human body and to construct mapping from visual images to posture information.

1.2 Low-Dimensional Representation

Low-dimensional representation is considered to be useful when a robot observes a demonstrator's motion and then imitate it. One of the reasons for this is that posture information and visual images have much unnecessary information that is not shared mutually between the two modalities. It is sufficient to use only a small amount of information that is shared by two modalities to estimate posture variables from visual images. This mutually shared information is expected to form a low-dimensional space. When a person imitates a demonstrator, it is actually known that such a person uses a low-dimensional structured schema called a body schema [13]. It is also known that the sequence of joint angles obtained when a person walks is efficiently and synergistically compressed by applying PCA [14]. Applying statistical compression techniques to posture information is therefore considered to be effective.

We propose an algorithm for obtaining a low-dimensional representation in

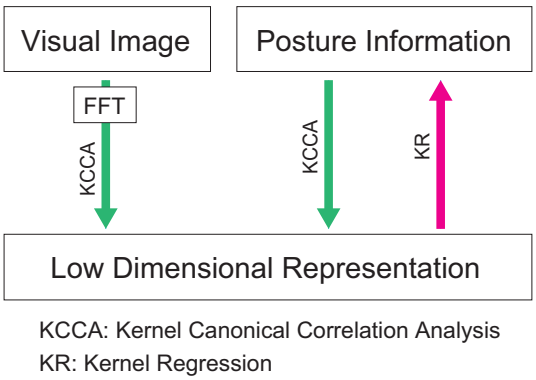

Figure 1: Overview of proposed algorithm

a bottom-up way by using KCCA and a method for estimating posture information from visual images via the low-dimensional representation. The method is superior to previous methods in terms of computation time because the essential dimension of mapping is reduced. A low-dimensional representation made from posture information and visual images is something equivalent to a body model because it can be used to estimate two information sources with several parameters. We discuss here the effectiveness of the estimation via the low-dimensional representation obtained by KCCA using human upper body motion data.

2 Algorithms

2.1 Overview

We propose an algorithm for a robot to estimate its joint angles from images of a human being (Figure 1). To estimate these effectively, the proposed algorithm uses a low-dimensional representation that integrates high-dimensional posture information and visual images and extracts important features shared by the two information sources. Canonical correlation analysis (CCA) is known as a method for obtaining low-dimensional information from two information sources by omitting irrelevant information using linear transformation. In this paper, we propose the use of KCCA to obtain the low-dimensional representation as a substitute for CCA because mapping from multimodal sensor motor information to low-dimensional information usually does not have a linear property. KCCA provides maps to a low-dimensional space from each information source. Mapping from the low-dimensional representation to posture information is necessary, however, for estimating posture information from visual images. For this purpose, we introduce kernel regression (KR) to estimate a map from low-dimensional space to posture information space.

The proposed algorithm is separated into two steps. First, the robot obtains

a low-dimensional representation and mapping from a data set of posture information and visual images using KCCA and KR (Figure 3). Second, the robot estimates joint angles from novel images of a human being (Figure 13). The proposed algorithm has six free parameters which are two hyperparameters of kernel functions, two regularization parameters of KCCA, one hyperparameter of kernel functions and one regularization parameter of KR.

2.2 Kernel Regression

Regression analysis is a method for quantitatively analyzing the relation between independent variables $\mathbf{x}$ and dependent variable y. When a set of independent variables $\mathbf{x}$ and dependent variable y are given as a training data set, regression analysis is applied to estimate the map from $\mathbf{x}$ to y using the data set. If map f is linear mapping, this method is called linear regression. We extend it to nonlinear mapping by assuming that map f is a linear combination of a d number of nonlinear functions (feature functions) $\phi = (\phi_1, \ldots, \phi_d)^{\mathrm{T}}$:

$$f(\mathbf{x}) = \sum_{i=1}^{d} a_i \phi_i(\mathbf{x}). \tag{1}$$

When n number of training data and $\{\mathbf{x}, y\}$ are given, coefficients $\mathbf{a} = (a_1, \ldots, a_d)^{\mathrm{T}}$ are usually calculated by minimizing least square error. Overfitting occurs, however, when the data set has space or includes an outlier. Coefficients $\mathbf{a}$ are therefore calculated by minimizing $R(\mathbf{a})$, which is least square error, added by regularization term $\lambda \parallel \mathbf{a} \parallel^2$:

$$R(\mathbf{a}) = \sum_{i=1}^{n} |y^{(i)} - f(\mathbf{x}^{(i)})|^2 + \lambda \parallel \mathbf{a} \parallel^2, \lambda > 0, \tag{2}$$

where $\mathbf{x}^{(i)}$ and $y^{(i)}$ represent i-th data. By adding a regularization term, it becomes possible to use a kernel method and to represent map f by using a linear combination of kernel functions k, which is the inner product of feature functions [15]:

$$f(\mathbf{x}) \quad = \quad \sum_{i=1}^{n} \alpha_i k(\mathbf{x}^{(i)}, \mathbf{x}). \tag{3}$$

We write $R(\mathbf{a})$ in respect to $\alpha = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$:

$$R(\alpha) = \sum_{i=1}^{n} |y^{(i)} - f(\mathbf{x}^{(i)})|^2 + \lambda \alpha^{\mathrm{T}} K \alpha, \lambda > 0, \tag{4}$$

where K is called a gram matrix and $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. α that minimizes $R(\alpha)$ is :

$$\alpha = (K + \lambda I_n)^{-1} \mathbf{y}, \tag{5}$$

where $\mathbf{y} = (y^{(1)}, \ldots, y^{(n)})^{\mathrm{T}}$.

2.3 Kernel Canonical Correlation Analysis

CCA was proposed by Hotelling [16] and is known as a method for compressing information that two information sources have in common. CCA infers two linear mappings, f and g, that map data $\mathbf{x}$ and $\mathbf{y}$ to one scalar data item and maximize the coefficient of correlation of two given data sets.

$$\rho = \mathrm{Cor}[f(\mathbf{x}), g(\mathbf{y})] = \frac{\mathrm{Cov}[f(\mathbf{x}), g(\mathbf{y})]}{\sqrt{\mathrm{Var}[f(\mathbf{x})]}\sqrt{\mathrm{Var}[g(\mathbf{y})]}}. \tag{6}$$

If two information sources have Gaussian distribution, the low-dimensional representation obtained by CCA has maximum mutual information between the two information sources. CCA is not effective for posture information and visual images, however, because these generally have a nonlinear relation. KCCA is an extended form of CCA whose mapping and linear combinations of feature functions are nonlinear [17]. KCCA assumes that maps f and g are represented by feature functions ψ_x and ψ_y as :

$$f(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\psi_x(\mathbf{x}) \tag{7}$$

$$g(\mathbf{y}) = \mathbf{b}^{\mathrm{T}}\psi_y(\mathbf{y}). \tag{8}$$

Coefficient vectors $\mathbf{a}$ and $\mathbf{b}$ were chosen to maximize the coefficient of correlation added by regularization term $\zeta_x \parallel \mathbf{a} \parallel^2 + \zeta_y \parallel \mathbf{b} \parallel^2$ because increasing the norm of $\mathbf{a}$ and $\mathbf{b}$ causes overfitting. Nonlinear mapping f and g can therefore be represented by a linear combination of kernel functions :

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k_x(\mathbf{x}^{(i)}, \mathbf{x}) \tag{9}$$

$$g(\mathbf{y}) = \sum_{i=1}^{n} \beta_i k_y(\mathbf{y}^{(i)}, \mathbf{y}). \tag{10}$$

Let α be $(\alpha_1 \cdots \alpha_n)^{\mathrm{T}}$ and β be $(\beta_1 \cdots \beta_n)^{\mathrm{T}}$. Finally, α and β, which maximize coefficient of correlation ρ, are obtained by solving the generalized eigenvalue problem as :

$$\begin{pmatrix} 0 & K_x J_n K_y \\ K_y J_n K_x & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \lambda \begin{pmatrix} K_x J_n K_x + \zeta_x K_x & 0 \\ 0 & K_y J_n K_y + \zeta_y K_y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \tag{11}$$

where $(K_x)_{ij} = k_x(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}), (K_y)_{ij} = k_y(\mathbf{y}^{(i)}, \mathbf{y}^{(j)})$, and J_n represents all component of an n dimension unit matrix minus one.

Here, λ is equal to the coefficient of correlation with the regularization term and is called a canonical correlation coefficient. Additionally, the image of data using a corresponding map is called a canonical variable. In this paper, the n-th largest canonical correlation coefficient is called an n-th canonical correlation

Figure 2: Kinect

coefficient and the corresponding canonical variable is called an n-th canonical variate. There are many kinds of kernel functions. In our experiments, we used the Gaussian kernel

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\sigma^2} \parallel \mathbf{x} - \mathbf{x}' \parallel^2\right) \tag{12}$$

as a kernel function for KCCA and KR.

3 Experiments

To evaluate the effectiveness of the proposed method, we conducted a sequence of experiments integrating posture information and visual images into a low-dimensional representation. We used human joint angles as posture information measured by using Kinect Microsoft (Figure 2) . We used upper body images drawn based on measured joint angles as virtual visual images.

First, we obtained the low-dimensional representation by applying KCCA to a data set of joint angle data and upper body images. The low-dimensional representation was a combination of canonical variables in which there were various combinations. The vector combined from the 1st to n-th canonical variable is called n-dimensional low-dimensional representation. In this experiment, we obtained ten low-dimensional representations.

Second, we obtained maps from each low-dimensional representation to joint angles by using KR. Joint angles were estimated from novel body images generated from novel measured joint angles as test data. The accuracy of estimation was evaluated based on mean error. We compared our proposed method to KR, in which joint angles are directly estimated from body image data. Hyperparameters were determined by 12-fold cross-validation. Criteria for cross-validation are the coefficient of correlation for KCCA and mean error for KR.

3.1 Experiment 1

We conducted an experiment to obtain low-dimensional representation that integrates posture information and visual images. Figure 3 shows an overview of the experiment.

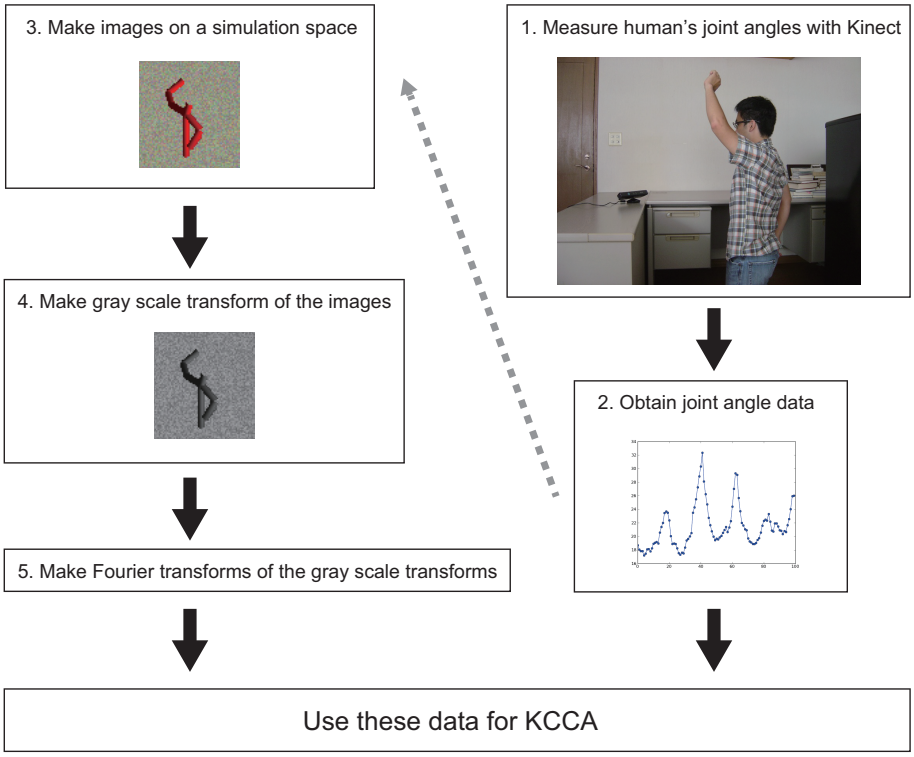

Figure 3: Overview of Experiment 1

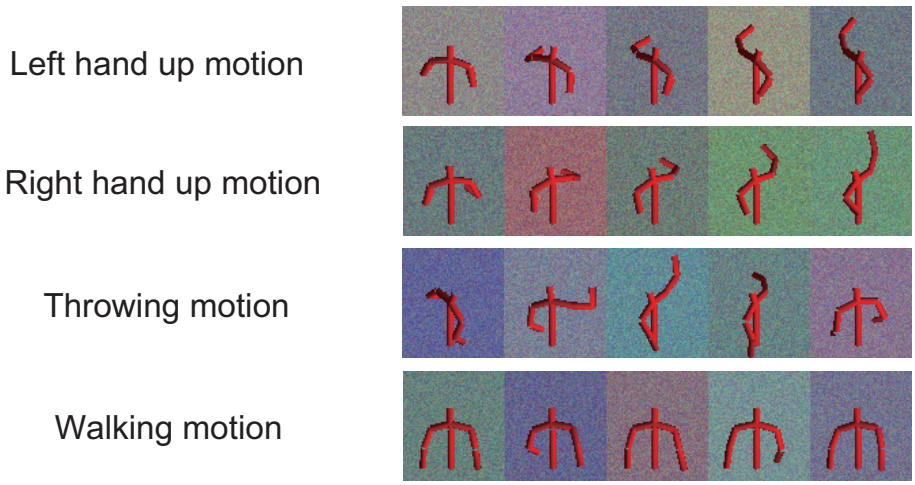

Figure 4: Training Data

Table 1: Estimated parameters of KCCA

σ_{img}	ζ_{img}	σ_{ang}	ζ_{ang}
1.2×10^4	8.0×10^{-2}	3.0×10	4.0×10^{-2}

Table 2: Estimated parameters of KR

Dimension	σ_n	ζ_n
1	2.0×10^{-1}	3.0×10^{-2}
2	4.0×10^{-1}	3.2×10^{-1}
3	7.0×10^{-1}	1.2×10^{-1}
4	8.0×10^{-1}	2.6×10^{-1}
5	8.0×10^{-1}	5.0×10^{-2}
6	8.0×10^{-1}	3.0×10^{-2}
7	8.0×10^{-1}	1.0×10^{-2}
8	9.0×10^{-1}	1.0×10^{-2}
9	1.0	1.0×10^{-2}
10	1.0	1.0×10^{-2}

3.1.1 Experimental Conditions

Measured training data consisted of a left hand up motion, a right hand up motion, a throwing motion, and a walking motion. Joint angle data had 11 dimensions consisting of the neck and both of the upper arms (an Euler angle having 3 degrees of freedom) and lower arms (the y-axis of the Euler angle for each arm). We used only the y-axis of lower arms because other angles contributed little to image drawing of the upper body. Images of the upper body model were made based on measured joint angles (Figure 4). The resolution of images was 75×80 pixels (6,000 pixels). Generated noise was added to images to express changes in background and noise. Background color was sampled at intervals of $[96,\ 160]$. Noise, which was sampled at intervals of $[-32,\ 32]$, was added to each pixel, where 256 colors are expressed in RGB and all noise was sampled from a uniform distribution. Images were then converted to grayscale images, Fourier transformation was applied to grayscale visual images and a training dataset was obtained.

3.1.2 Results

We describe estimated parameters of KCCA in table 1 and KR in table 2. In table 1, σ_{img} and ζ_{img} are hyperparameter of kernel functions and regularization parameter of mapping from images. σ_{ang} and ζ_{ang} are hyperparameter of kernel functions and regularization parameter of mapping from joint angles in KCCA. In table 2, there are parameters of mapping from each low-dimensional representation to joint angles. σ_n and ζ_n are hyperparameter of kernel functions and

regularization parameter of mapping from n-dimensional representation.

Figure 5 plots estimated hyperparameters σ_n. This figure implies that the value of the parameter increases as the number of dimensions increases. The increment in parameter value is greater than the increment in the length of a diagonal line in a hypercube. If the dimension of the low-dimensional representation increases, the low-dimensional representation has information that is irrelevant to regression. The parameter that corresponds to the range in which the kernel function value is large increases in order to reduce this effect. That is why, as is shown later, mean error of estimation from novel visual images increases together with the increment in dimension of the low-dimensional representation.

Figure 6 shows 128 coefficients of correlation obtained from training data by KCCA in descending order. There is sufficient number of high value coefficients of correlation.

To evaluate the characteristics of low-dimensional space generated by KCCA, we compared KCCA with PCA, KPCA, and CCA. The three-dimensional representation obtained from training data by using PCA, KPCA, CCA, and KCCA are illustrated in Figures 7-12 where triangles represent visual images, circles represent joint angles, and each axis represents first canonical variable (D1), second canonical variable (D2), and third canonical variable (D3). In results of PCA derived from visual images (Figure 7), circles are arranged irregularly, but there is a point mass. This point mass corresponds to images of walking motion that have almost no differences. Results of PCA from joint angles (Figure 8) show a cylindrical structure. This reflects a case in which left hand up and right hand up motions are connected without sufficient distance. Results of KPCA from visual images (Figure 9) show a horseshoe-like structure. Nonlinear maps of KPCA seem to derive high variance components. This does not, however, reflect a significant structure. In results of KPCA from joint angles (Figure 10), there is a further highlighted cylindrical structure. The results of CCA (Figure 11) show a point mass of walking motion, but no other structures were reflected. In results of KCCA (Figure 12), we found point masses of walking motion, and left hand up and right hand up motions are well divided and form a branching structure. From this, the best low-dimensional representation was obtained by KCCA. Although the low-dimensional representation obtained by KCCA has the same scale in each dimension, that obtained by other methods has a different length in each dimension. If the kernel function is a radial basis function, the low-dimensional representation obtained by KCCA is therefore best for estimation.

3.2 Experiment 2

Next, we conducted an experiment to examine the generalization performance of the low-dimensional representation for novel data. Figure 13 shows an overview of the experiment.

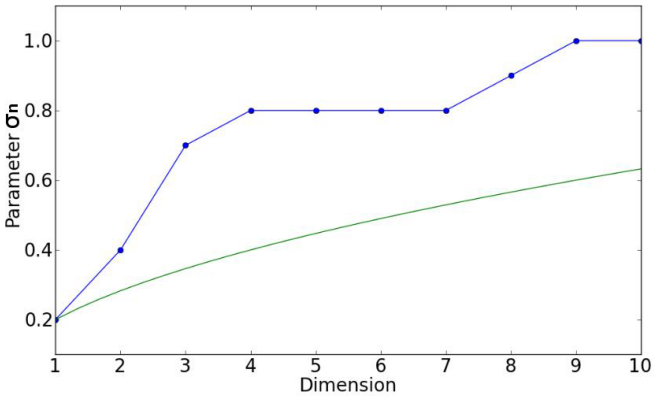

Figure 5: Parameter σ_n: The line with dots represents estimated parameters and the solid line represents $0.2\sqrt{n}$.

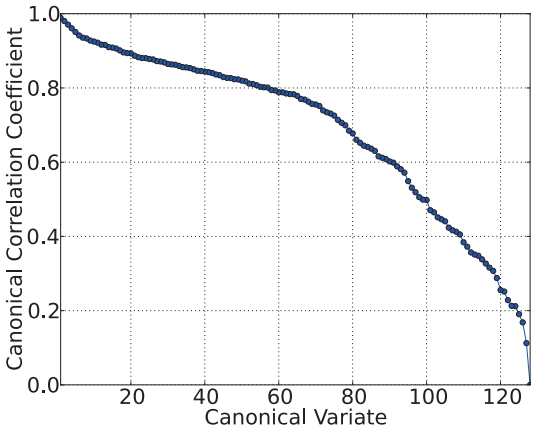

Figure 6: Canonical Correlation Coefficient

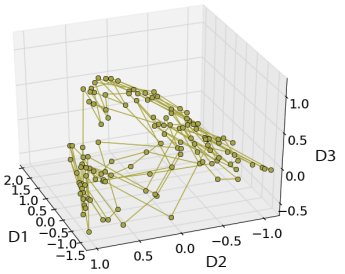

Figure 7: Low-Dimensional Representation of Body Images by PCA

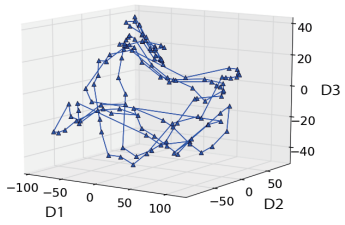

Figure 8: Low-Dimensional Representation of Joint Angles by PCA

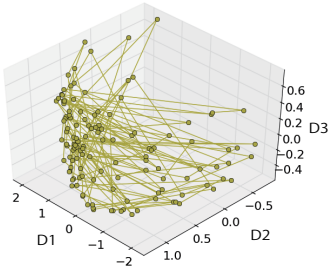

Figure 9: Low-Dimensional Representation of Body Images by KPCA

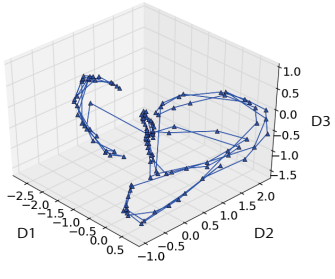

Figure 10: Low-Dimensional Representation of Joint Angles by KPCA

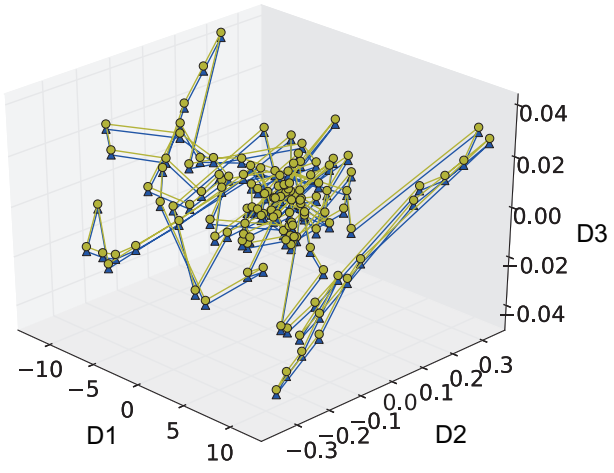

Figure 11: Low-Dimensional Representation of Body Images by CCA

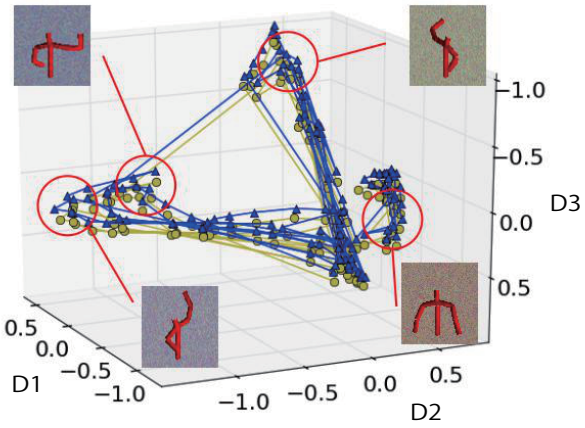

Figure 12: Low-Dimensional Representation (3D): Triangles represent posture information and circles represent visual images. Individual dots are lined up according to a sequence.

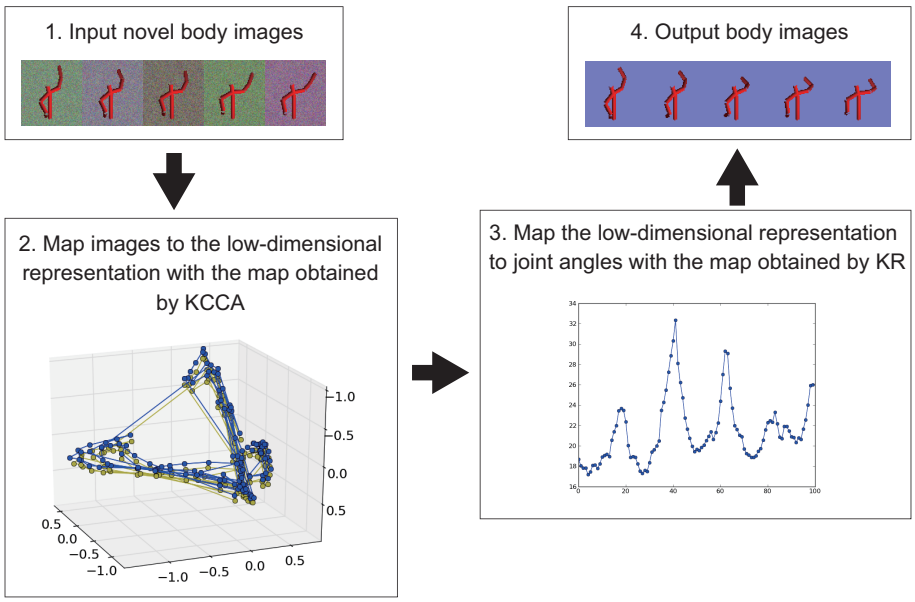

Figure 13: Overview of Experiment 2

3.2.1 Experimental Conditions

To estimate posture information from visual images, a map has to be obtained to posture information from the low-dimensional representation because maps to the low-dimensional representation obtained by KCCA cannot calculate posture information from visual images. KR is therefore used to obtain this map. Test data consisted of a handwaving motion and a reaching motion (Figures 15,16). Posture information and visual images were made in the same way as in experiment 1. Using these data, we estimated joint angles estimated from novel body images via the low-dimensional representation. Additionally, to evaluate estimation via the low-dimensional representation, we estimated joint angles that were estimated from body images directly by KR.

3.2.2 Results

Figure 14 plots mean error of joint angles estimated from novel body images via low-dimensional space. When the dimension of the low-dimensional representation is four in handwaving motion and three in reaching motion, mean error is the lowest and the value is smaller than the value estimated directly. This suggests that estimation performance via the best low-dimensional representation is better than best direct estimation performance.

Computation time for estimation of 50 images took 5.6 [s] in direct estimation on a 1.4 GHz Intel Core 2 Duo with 2GB memory. Computation time for estimation via the low-dimensional representation, however, took just 0.37 [s]. This trend was independent of dimensionality. Computation time for estimation via the low-dimensional representation was clearly much faster than that for direct estimation. Next, Figure 17 shows computation time for estimation from each 5 of 5 to 50 images. Computation time for estimation by both direct projection and projection via the low-dimensional representation is proportional to the number of images, and estimation via the low-dimensional representation was faster than direct estimation. By projecting a high-dimensional visual image onto low-dimensional space, our method made it possible to reduce computation time required to calculate a gram matrix representing visual image data. Figures 15 and 16 show upper body images generated from estimated joint angles via the low-dimensional representation. In Figure 15, the first two estimated images are similar to corresponding test data but the rest are not similar. Mean error of each estimation is lower for the first two estimations and higher for the rest. In Figure 16, the right arm of the upper body in images is not extended, even though in corresponding test data, the right arm was extended to the right. This is because training data did not have data in which the right arm was extended to the right.

It is expected that KCCA can extract the complete hidden structure that holds the informational relationship between posture information and visual images. The low-dimensional space was generated from posture data, however, and visual image data was included in training data set by KCCA. Original high-dimensional space has the same dimensionality as the number of training

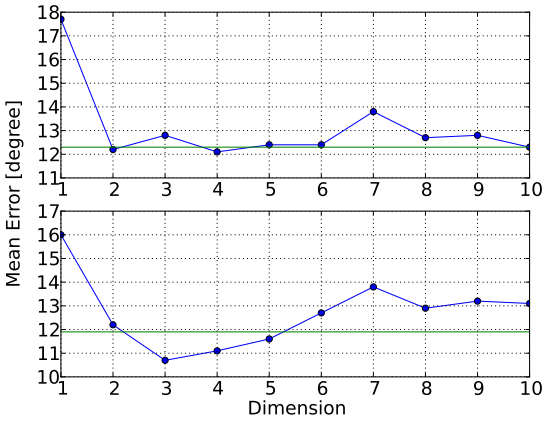

Figure 14: Mean Error (Top: Handwaving Motion, Bottom: Reaching Motion): Lines with dots represent mean error via each low-dimensional representation, the parallel lines represent mean error obtained from direct estimation.

data. Mapping from high-dimensional visual image space to low-dimensional space can therefore properly cover only subspace of visual image space involving observed data.

There are two approaches to avoid this problem. One approach is to use another kernel function. The Gaussian kernel function is not the only candidate, however, and there may be another kernel function that reflects the nature of the human body structure more effectively. The other approach is to use a large amount of training data. This approach has limitations, however, because computation time increases as the amount of data increases.

4 Conclusions

We have proposed a novel method for generating a low-dimensional representation by integrating posture information and visual images using KCCA. Using this method, a robot can estimate posture information from observed visual images of the human body. Results have indicated that joint angles were estimated via the low-dimensional representation as accurately as in estimation done directly from a visual image. Unlike the conventional probabilistic method, the global solution of low-dimensional representation is obtained by using KCCA without iteration. The low-dimensional representation obtained is useful for estimating posture information. Determining the dimensionality of a low-dimensional representation is an important problem in constructing such representation. Finding a suitable method for determining dimensionality is a task for the future. We plan to use a method that applies model selection

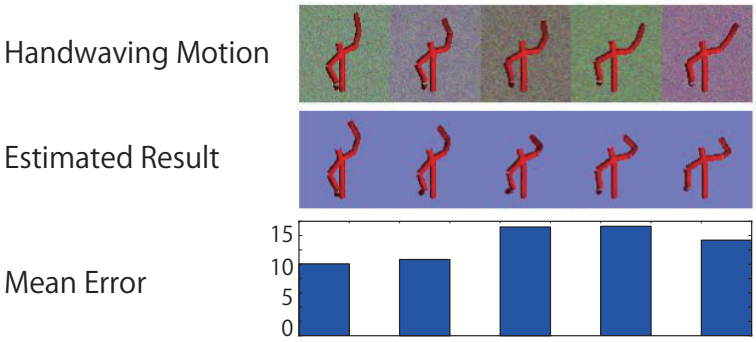

Figure 15: Handwaving Motion in test data, Estimated Images, and Mean Error of each estimation

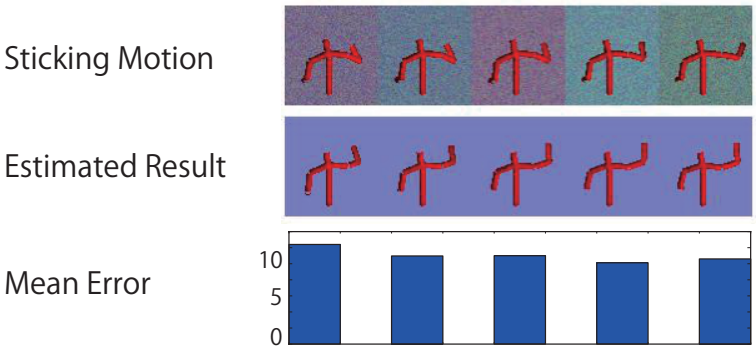

Figure 16: Sticking Motion in test data, Estimated Images, and Mean Error of each estimation

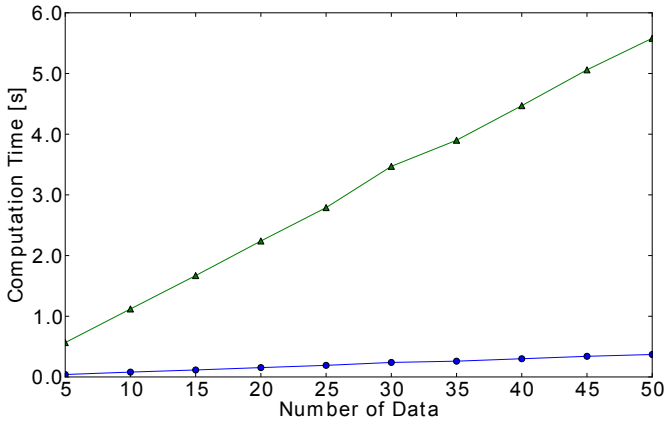

Figure 17: Computation time for estimation: The line with circles represents the computation time for estimation via the low-dimensional representation, and the line with triangles represents the computation time for direct estimation.

or cross-validation [18] and a method that extends CCA stochastically [19] to solve this problem. It is possible to apply such methods to our proposed algorithm. In addition, we did not compare our proposed method to previous complicated probabilistic methods, such as GPLVM, because it is difficult to compare them fairly due to qualitative differences between them. Further analysis of our method and a comparison to related work are also tasks for future work.

We have assumed in this paper that both posture information and visual images are those of an imitator. This in turn assumes a situation in which a robot learns its body scheme by looking in a mirror that reflects its own body image. We assume that an organized body scheme can be applied to the visual images of others whose motion the robot tries to imitate if their body structures are similar.

In experiments, the background of the training dataset and the surface pattern of the body image were almost erased. This was because we focused on the evaluation and analysis of our proposed algorithm in this paper. Similar images can be obtained, however, by a background subtraction method using a stereovision camera or a depth sensor. To apply these methods to real images is also a task for future work. Our proposed algorithm can be applied to any robot having any body structure, for example, a robot that has two or more legs [20, 21], or can be used as a hand pose estimator [22] because we do not assume a particular body model. When the dimension of posture information increases, however, the method requires a larger amount of data. This is a problem of scalability. We will work on this problem in the future.

Acknowledgment

This research is supported by a Grant-in-Aid Creative Scientific Research 2007-2011 (19GS0208) funded by the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

[1] Cynthia Breazeal and Brain Scassellati, "Robots that imitate humans", TRENDS in Cognitive Sciences, Vol.6, pp11, November 2002

[2] Taniguchi T., Iwahashi N., Sugiura K., and Sawaragi T., "Constructive Approach to Role-Reversal Imitation Through Unsegmented Interactions", Journal of Robotics and Mechatronics, Vol.20, No.4, pp567-577, 2008

[3] Cynthia Breazeal and Brain Scassellati, "A context-dependent attention system for a social robot", In Proc. Sixteenth Int. Joint Conf. Artif. Intell. (IJCAI99), pp.1146-1151, 1999

[4] C.L. Nehaniv and K. Dautenhahn, "Imitation in Animals and Artifacts", The MIT Press, pp41-61, 2002

[5] Aris Alissandrakis, Chrystopher L. Nehaniv and Kerstin Dautenhahn, "Imitation With ALICE :Learning to Imitate Corresponding Actions Across Dissimilar Embodiments", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS PART A: SYSTEMS AND HUMANS, Vol.32, pp482-496, 2002

[6] Mun Wai Lee, Isaac Cohen and Soon Ki Jung, "Particle Filter with Analytical Inference for Human Body Tracking", IEEE Workshop on Motion and Video Computing, 2002

[7] Shotton J., Fitzgibbon A., Cook M., Sharp T., Finocchio M., Moore R., Kipman A. and Blake A., "Real-Time Human Pose Recognition in Parts from Single Depth Images", CVPR, 2011

[8] Katsu Yamane, Daisuke Fukuda, and Yoshihiko Nakamura, "Markerless Motion Capture with Structure Estimation Capability", Journal of Robotics and Mechatronics, Vol.20, No.2, pp322-331, 2008

[9] Agarwal A. and Triggs B., "Recovering 3d human pose from monocular images", IEEE Trans. Pattern Anal. Mach. Intell, Vol.28, pp44-58, 2006

[10] Grauman K., Shakhnarovich G. and Darrell T., "Inferring 3d structure with a statistical image-based shape model", ICCV, pp641-648, 2003

[11] Carl Henrik Ek, Philip H. S. Torr and Neil D. Lawrence, "Gaussian Process Latent Variable Models for Human Pose Estimation", MLMI, pp132-143,2007

[12] Rasmussen C.E. and Williams C.K., "Gaussian Processes for Machine Learning", The MIT Press, 2006

[13] A. Maravita, C. Spence and Driver, "Multisensory integration and the body schema: Close to hand and within reach", Current Biology, Vol.13, 2003

[14] N.A. Borghese, L. Bianchi and F. Lacquaniti, "Kinematic determinants of human locomotion", J Physiology, 1996

[15] Christopher M. Bishop, "Pattern Recognition And Machine Learning", Springer-Verlag, 2006

[16] H.Hotelling, "Relations between two sets of variates", Biometrika, Vol.28, pp321-377, 1936

[17] D. R. Hardoon, S. Szedmak and J. Shawe-Taylor, "Canonical correlation analysis: an overview with application to learning methods", Neural Computation, Vol.16, pp2639-2664, 2004

[18] C. Wang, "Variational Bayesian approach to Canonical Correlation Analysis", In IEEE Transactions on Neural Networks, 2007

[19] Piyush Rai and Hal Daume, "Multi-label prediction via sparse infinite CCA", Advances in Neural Information Processing Systems, Vol.22, pp1518-1526, 2009

[20] Kazuo Morita, and Hidenori Ishihara, "Four-Legged Mechanism for Realizing Dynamic Running −Design of Prototype with Drive System that Enables Dynamic Locomotion Change−", Journal of Robotics and Mechatronics, Vol.20, No.2, pp234-240, 2008

[21] Takahiro Doi, Kazunori Miyata, Takamasa Sasagawa, and Kenjiro Tadakuma, "Multi-Leg System for Aerial Vehicles", Journal of Robotics and Mechatronics, Vol.24, No.1, pp174-179, 2012

[22] Kiyoshi Hoshino and Motomasa Tomida, "3D Hand Pose Estimation Using a Single Camera for Unspecified Users", Journal of Robotics and Mechatronics, Vol.21, No.6, pp749-757, 2009