PAPER

Construction of Latent Descriptor Space and Inference Model of Hand–Object Interactions

Tadashi MATSUO[†a)] *and* Nobutaka SHIMADA[†], *Members*

SUMMARY Appearance-based generic object recognition is a challenging problem because all possible appearances of objects cannot be registered, especially as new objects are produced every day. *Function of objects*, however, has a comparatively small number of prototypes. Therefore, function-based classification of new objects could be a valuable tool for generic object recognition. Object functions are closely related to hand–object interactions during handling of a functional object; i.e., how the hand approaches the object, which parts of the object and contact the hand, and the shape of the hand during interaction. Hand–object interactions are helpful for modeling object functions. However, it is difficult to assign discrete labels to interactions because an object shape and grasping hand–postures intrinsically have continuous variations. To describe these interactions, we propose the *interaction descriptor space* which is acquired from unlabeled appearances of human hand–object interactions. By using interaction descriptors, we can numerically describe the relation between an object's appearance and its possible interaction with the hand. The model infers the quantitative state of the interaction from the object image alone. It also identifies the parts of objects designed for hand interactions such as grips and handles. We demonstrate that the proposed method can unsupervisedly generate interaction descriptors that make clusters corresponding to interaction types. And also we demonstrate that the model can infer possible hand–object interactions.

key words: *feature extraction, unsupervised machine learning, object classification*

1. Introduction

Appearance-based generic object recognition is a challenging problem because all possible appearance of new objects, which are produced every day cannot be completely registered. In contrast, *function of objects* is common to many objects regularly handled by humans and has a comparatively small number of prototypes. Therefore, a function-based classification of new objects could be valuable for generic object recognition. The effectiveness of object functions in generic object recognition has been already discussed and indicated [1], [2]. However, in the papers, each function is manually defined for each object category. It is desirable that information specifying functions can be extracted without manually assigning function labels to many objects.

The object function is closely related to the interactions between the functional object and human hand. Specifically, it embodies the approach of the hand to the object, the parts of the object contacted by the hand, and the hand shape activated by the interaction [3]. The interaction types specified by such factors have been precisely analyzed in the literature [4]. Hand–object interactions are therefore promising for function-based classification in image-based recognition. Some parts of objects, such as grips, bottoms, and brims, are handled in typical ways (Fig. 1). Such interactions with specific parts are called *perceived affordance* [5].

Assuming that a hand–object interaction can be represented by a descriptor, the descriptor can be considered as a latent attribute of the object itself. Such descriptors are available for training samples but not for the test samples. In the context of machine learning, training with *hidden information* (such as latent attributes) can improve the classification accuracy [6]–[9]. The hidden information contains additional records of each training sample: for example, age, gender, or race in facial recognition algorithms. The classifier is trained to recognize facial image patterns by calculating the similarity metric of the hidden information such as age, gender, or race, which provides the error costs. Although the hidden information is not available for test samples, considering the hidden information on training brings the classification boundaries with no over-fitting and good inference performance. A similar framework is potentially applicable to recognition based on hand–object interactions.

Alessandro Pieropan et al. have proposed an method estimating an object function from a sequence of interactions [10]. They focus on defining a function by a sequence of descriptions of predefined actions (comparatively large motion), not including shapes of hands and objects, which are important for interaction between a hand and a tool. Dan Song et al. have proposed an method estimating a human intention from an image sequence of an interaction [11]. In the method, relationship between intention and appearances is learned supervisedly. The interaction type is discretely

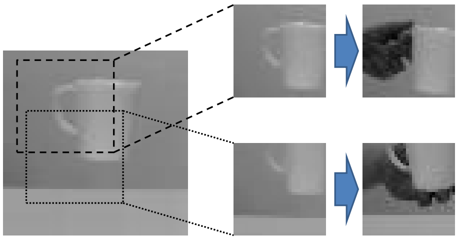


Fig. 1 Local appearances and typical hand–object interactions of a cup

Manuscript received September 27, 2016.
Manuscript revised February 1, 2017.
Manuscript publicized March 13, 2017.

[†]The authors are with the College of Information Science and Engineering, Ritsumeikan University, Kusatsu-shi, 525–8577 Japan.
a) E-mail: matsuo@i.ci.ritsumei.ac.jp

DOI: 10.1587/transinf.2016EDP7410

defined and required to be given manually before modeling for each action samples. Since an object shape and grasping hand–postures intrinsically have continuous variations, descriptions of interactions should reflect such variation continuously.

We propose a system that can embed a hand–object interaction as a "interaction descriptor" vector in a small dimensional space. The interaction descriptor represents hand–object interactions continuously in contrast to discrete label sets, which only discriminate several predefined objects ("cup", "pen", ...) or function classes (for "drink", "write", ...). The proposed method can achieve the embedding in unsupervised way. Assuming that object function is closely related to hand-object interactions, the interaction descriptor is helpful for modeling object functions.

For a numerical representation of hand–object interactions, we introduce the *interaction descriptor space*. This space is unsupervisedly constructed by a convolutional autoencoder (CAE) [12], an unsupervised feature extraction method. When training the model, we introduce a sparseness term in the evaluation function that clusters similar interactions in the descriptor space. The training is based on the appearances of hand–object interactions of typical functional objects such as scissors, cutters, pens, and cups. The latent attributes in the training are the *interaction images*, comprising the appearance itself and its segmentation images of the hand and object. The descriptor space can quantitatively discriminate among an infinite number of functional object types.

Employing the convolutional neural network (CNN) [13], we then model the relation between an object's appearance and its corresponding hand–object interaction state in the interaction descriptor space. In this way, the model can infer the interaction state from the object image alone.

We demonstrate that the descriptor space and the proposed framework successfully encode the hand–object interaction state from a single object image.

2. Interaction Descriptor Space

The object function is closely related to the type of hand–object interaction during handling of a functional object, such as grasping the object, picking it up, the direction of approach of the hand, and other characteristic motions. Therefore, these hand–object interactions are potentially useful for describing the object function. Since an object shape and grasping hand–postures intrinsically have continuous variations, interaction descriptors should continuously reflect such variations. We represent an interaction descriptor as a vector in a continuous vector space, "interaction descriptor space". We generate an interaction descriptor vector by encoding an appearance of a hand–object interaction because the appearance reflects the type of the interaction.

The problem is how to generate the mapping from an appearance to an interaction descriptor. The mapping should satisfy the following conditions:

A. The mapping extracts only the essential information of the interaction. The detailed shape or texture of the object, which are not relevant to the interaction, should be ignored.
B. The mapping can be learned with a set of unlabeled appearances and therefore can be generated without manually classifying the interactions beforehand.
C. Images corresponding to different interactions are mapped to distantly spaced descriptors. The difference between two interactions should be reflected in the numerical distance between their corresponding descriptors.
D. Images corresponding to similar interactions are mapped to closely spaced descriptors, even when the objects differ in size or shape and are slightly displaced from each other in the images.
E. Certain spatially local features, such as edges and grips, are common to multiple interactions and are effective for distinguishing among interactions. Such useful features should be automatically found from a set of appearances.

The essential information can be extracted by the autoencoder method [14], [15], which employs an encoder and a decoder. The encoder converts an input to a code with lower dimensionality, and the decoder approximately restores the original input from the code. Both elements are trained such that the combination restores the input as correctly as possible for a certain set of vectors. Under this constraint, the encoder generates a numerical representation of the principal components required for input restoration. In addition, the encoder and decoder can be trained with unlabeled vectors (satisfying condition B, mentioned above).

If we can restore an interaction appearance from a descriptor, then the descriptor contains information of the interaction. The mapping that satisfies conditions A and B is generated by autoencoder method.

To satisfy condition C, we concentrate the descriptors corresponding to certain types of interaction appearances and isolate them from descriptors corresponding to other types of interactions. If a label specifying an interaction type can be assigned to each appearance, we can further constrain the mapping so that the descriptors of two different interaction types are distantly spaced. However, to satisfy B, the important components that specify an interaction must be found from unlabeled appearances. In previous studies, the important components among a set of unlabeled vectors have been found by *sparse coding* methods [16]–[20]. However, these methods require additional inequality or equality constraint.

To resolve this problem, we introduce a sparseness constraint to the autoencoder. Although sparse autoencoders have been previously proposed [15], [21], the method in [21] requires an inequality constraint when training the model. The method in [15] requires the scheduling of the sparsity level.

We introduce the sparseness constraint that does not

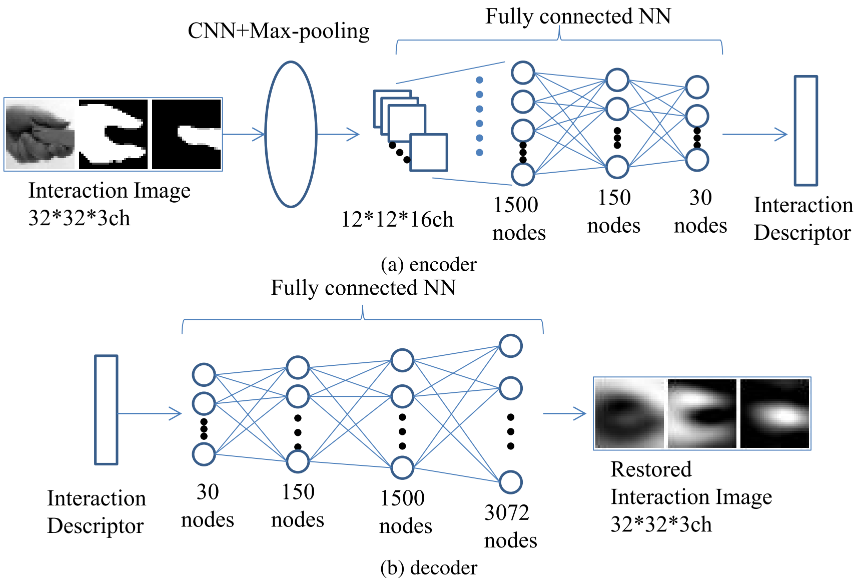


Fig. 2 Network structure of the encoder and decoder

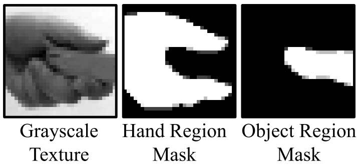


Fig. 3 Components of an interaction image

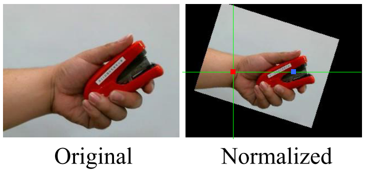


Fig. 4 Normalization of a training image

require equality or inequality constraint. It is applicable to a general CNN-based autoencoder that is followed by fully connected neural networks with non-linear activation functions. As a CNN consists of convolutional filters that uniformly extract spatially local features from an image, the extracted local features are position-independent (satisfying conditions D and E). In addition, the CNN filters can be trained by an unsupervised learning method (satisfying B).

3. Interaction Image

Before generating a descriptor from an interaction appearance, we need to define the *appearance* which contains sufficient information to distinguish among interactions.

The appearance is derived from an *interaction image* (Fig. 3), a 3-channel (32 × 32) pixel normalized image consisting of a total appearance, a hand region mask and an object region mask.

To focus on the hand–object interaction, we approximately normalize the positions and directions of the training images for each type of hand–object interaction. The training images can be automatically normalized in the wrist–object coordinate system [22], as shown in Fig. 4.

4. Autoencoder for Generating Descriptors

We simultaneously constructed the interaction descriptor space and the mapping using a sparse convolutional autoencoder (CAE).

4.1 Network Structure

To automatically extract the local features that effectively identify an interaction, we place the CNN as the first layer of the encoder. For learning the nonlinear relation between the local features and interactions, the first layer is followed by a three-layer fully connected neural network in which each layer precedes a nonlinear activation function (Fig. 2 (a)). Similarly, the decoder is a fully connected neural network with a nonlinear activation function for representing the nonlinear relation (Fig. 2 (b)).

4.2 Cost Function

Generally, an autoencoder is trained such that the encoder–decoder combination approximately restores an input in a certain input set. It is formulated as

$$\underset{D,E}{\operatorname{argmin}} \sum_{I \in S} \|I - D(E(I))\|_2^2, \tag{1}$$

where I, S, $D(\cdot)$, $E(\cdot)$, and $\|\cdot\|_p$ denote the input to be reconstructed, a set of inputs, the encoder, the decoder, and the ℓ^p norm, respectively. In our problem, I and $E(I)$ denote the interaction image and its corresponding descriptor, respectively.

In the above objective function, the encoder should pre-

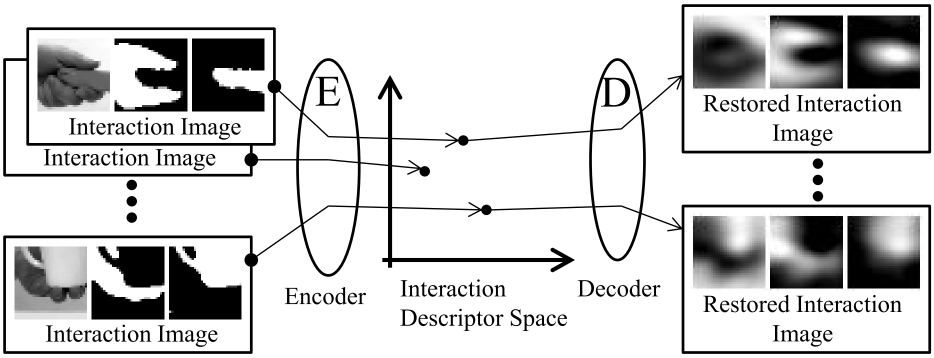


Fig. 5 Encoder and decoder outputs

serve the information of an input among a certain set of inputs. Our problem requests that the encoder can extract the essential components common to the interaction appearances of similar type. According to the basis pursuit concept [23] or sparse coding method, this can be achieved by constraining the ℓ^1 norm of the encoder's output. Simply, the ℓ^1 constraint can be imposed on the objective function as follows:

$$\beta \sum_{I \in S} \|I - D(E(I))\|_2^2 + \lambda \sum_{I \in S} \|E(I)\|_1 . \tag{2}$$

However, simply adding the ℓ^1 norm term is ineffective because the additional term can be rendered arbitrarily small by scalar multiplication of the encoder output and the decoder input. Basis pursuit avoids this problem by adding a constraint of the magnitude of the encoder matrix, as given by Eq. (1) in [21]:

$$\begin{aligned} &\underset{\mathbf{D},\mathbf{z}}{\operatorname{argmin}} \frac{1}{2} \sum_{n=1}^{N} \|\mathbf{x}_n - \mathbf{D}\mathbf{z}_n\|_2^2 + \beta \|\mathbf{z}_n\|_1 \\ &\text{subject to} \mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K] \\ &\qquad \|\mathbf{d}_k\|_2^2 \leq 1 \;\; for \;\; k = 1, \ldots, K, \end{aligned} \tag{3}$$

where $\mathbf{x}_n$ and $\mathbf{D}$ mean an input vector and an decoder matrix, respectively, and $\mathbf{z}_n$ denotes the code corresponding to $\mathbf{x}_n$. However, such a constraint is not easily imposed on an NN-based encoder. Instead, we introduce a constraint term C_{sparse} that simultaneously limits the ℓ^1 norm and the magnitude of the encoder's output as follows:

$$C_{err} = \sum_{I \in S_I} \|I - D(E(I))\|_2^2 , \tag{4}$$

$$C_{sparse} = \sum_{I \in S_I} \left(\frac{\|E(I)\|_1}{\|E(I)\|_2} \right)^2 , \tag{5}$$

$$C = \beta C_{err} + \lambda C_{sparse} . \tag{6}$$

The additional term C_{sparse} is the ratio of the ℓ^1 norm to the ℓ^2 norm of the descriptor $E(I)$. C_{sparse} is smaller if the descriptor vector $E(I)$ is more sparse [24], [25]. The autoencoder is trained to minimize the total cost function C.

For a d-dimensional vector $\mathbf{v}$, $(\|\mathbf{v}\|_1 / \|\mathbf{v}\|_2)^2$ is minimized at 1 only when the vector $\mathbf{v}$ has a single non-zero component and all other components are zero (the sparsest case). Conversely, it is maximized when all components of $\mathbf{v}$ have a common absolute value. The lower the C_{sparse}, the fewer basis vectors required in the weighted sum that approximates the decoder outputs.

This decoder obtains the shapes of a hand and an object and their spatial relation from a point on the interaction descriptor space (Fig. 5).

5. Inference Model

With the numerical interaction descriptor, the CNN can learn the relation between an object appearance and a possible interaction (Fig. 6). We can then infer an instance of the interaction descriptor from the appearance of an object.

Each interaction descriptor does not represent an absolute direction of an interaction in an image because it is based on interaction images and they are normalized by the wrist–object coordinate system, as shown in Fig. 4. An interaction descriptor represents shapes of a hand and an object, their relative position and their relative direction. So, when pairing an object-only appearance with an interaction descriptor for training the inference model, any rotated versions of the object-only appearance may be paired with the interaction descriptor. Although it is possible to train the inference model with all pairs of any rotated object-only appearance and an interaction descriptor, a convolutional neural network (CNN) cannot effectively extract common components from many rotated variations [26]. To learn common shapes of objects with common possible interactions by a CNN effectively, it is desirable that directions of objects with common interactions are standardized. Since an object-only appearance does not have an obvious standard direction, we have normalized a direction of an object-only appearance so that the object has a direction similar to that of an object in the paired interaction image.

Due to the normalization, when inferring an interaction descriptor from an object-only appearance, we have to rotate the appearance so that its direction matches to that of a similar object used in training. Since appropriate rotation for an object is unknown generally, we have to infer interaction descriptors from all possible versions of the rotated appearance.

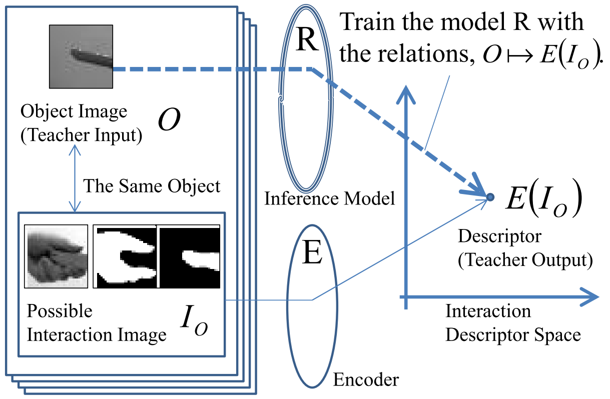


Fig. 6 Training of inference model

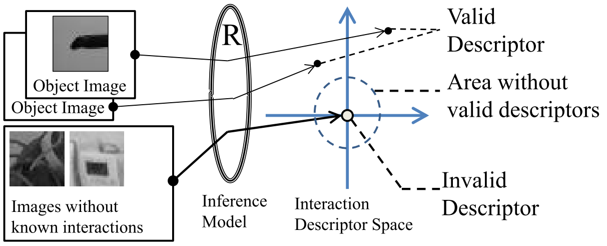


Fig. 7 Training with invalid descriptor

However, if objects have some typical poses in images due to gravity or other reason, the inference model should be trained with pairs of an appearances of an object in such a typical pose and a corresponding interaction descriptor. The inference model trained in that way can infer an interaction descriptor from an object-only appearance itself because it matches to one of typical poses.

We also introduce an *invalid descriptor* that discriminates between images with and without known interactions. For an input without known interactions, the model is trained to output the descriptor closest to the invalid descriptor. The invalid descriptor is defined as a zero vector. The model is trained with two types of teacher samples (Fig. 7); pairs of an image with a known interaction and its descriptor (positive samples), and pairs of an image without known interactions and an invalid descriptor (negative samples).

After training the model, we estimate the probability distribution of the L$_2$-norms of the inferred descriptors for samples of each teacher type. These samples differ from the training samples of the inference model R. Figure 8 shows the estimated distributions of $P\left(\|R(O)\|_{\mathrm{L2}} \,\middle|\, O \in I_{pos}\right)$ and $P\left(\|R(O)\|_{\mathrm{L2}} \,\middle|\, O \in I_{neg}\right)$, where R denotes the inference model, O denotes an input image, and I_{pos} and I_{neg} are the sets of teacher images with and without known interactions, respectively. As shown in the figure, a high norm of an inferred descriptor indicates large likelihood that the input has a known interaction. The likelihood f that an input image O has a known interaction is given by

$$f(O) = \frac{g(\|R(O)\|_{\mathrm{L2}})}{g(\|R(O)\|_{\mathrm{L2}}) + h(\|R(O)\|_{\mathrm{L2}})}, \tag{7}$$

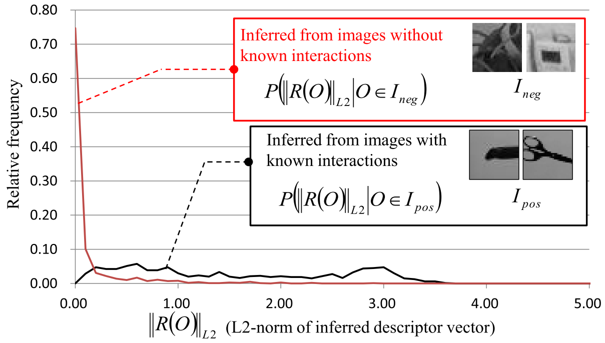


Fig. 8 Distribution of L$_2$-norms of inferred descriptors

where

$$\begin{aligned} g(d) &= P\left(d = \|R(O')\|_{\mathrm{L2}} \,\middle|\, O' \in I_{pos}\right), \\ h(d) &= P\left(d = \|R(O')\|_{\mathrm{L2}} \,\middle|\, O' \in I_{neg}\right). \end{aligned} \tag{8}$$

By connecting the inference model R and the decoder D, our system infers a possible interaction image from the appearance of an object (Fig. 9). The inference model R is the CNN shown in Fig. 10.

6. Experiment

The encoder and decoder were trained with interaction images generated from 1,680 scenes showing 12 types of interactions (Fig. 11). Each interaction image was generated from a (32 × 32)[pixel] sub-image randomly located in the scene image. We generated multiple instances of interaction images from each scene by randomly extracting subsquares with sufficient area of a hand region. The total variation exceeds 500,000. Masks in interaction images were generated by skin color extraction and background subtraction. The encoder and decoder were trained by minimizing C using stochastic gradient descent (SGD) [13].

6.1 Distribution of Descriptors

To demonstrate the effect of the sparseness cost, we define the diameter $\mathcal{D}$ of a set of descriptors as follows:

$$\mathrm{dia}\,\mathcal{D} \overset{\mathrm{def}}{=} \max\{\|x - y\| \mid x, y \in \mathcal{D}\} \tag{9}$$

We also define $\mathcal{D}_k$ as a set of descriptors of the k-th interaction type, and denote μ_{dia} as the mean of $\mathrm{dia}\,\mathcal{D}_k$ for k. If μ_{dia} is small, the descriptors corresponding to a similar interaction are closely placed. This is a desirable property because similarity of interactions should be reflected in closeness of their descriptors.

Figure 12 shows the relation between the sparseness cost C_{sparse} and the mean diameter μ_{dia} of the descriptor sets. Each point corresponds to a pair of C_{sparse} and μ_{dia} after the training process for each weight λ in the cost function (4). The case $\lambda = 0$ is equivalent to the case without a sparseness cost. The figure shows that larger λ brings smaller

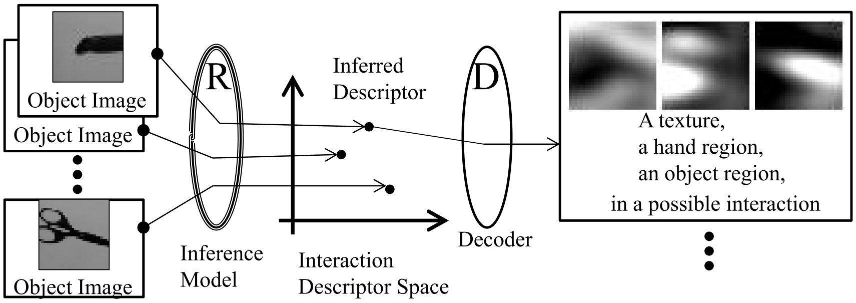


Fig. 9 Operation of the inference system

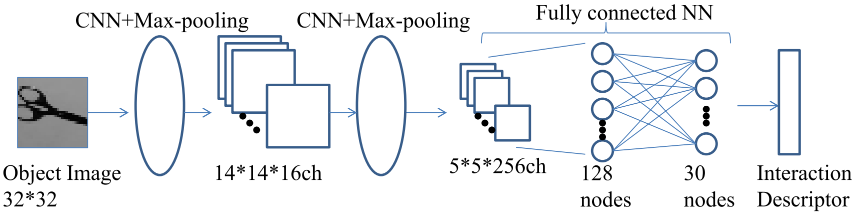


Fig. 10 Network structure of the inference model

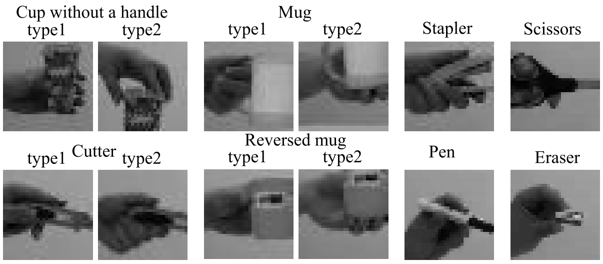


Fig. 11 Interaction types in the encoder–decoder training step

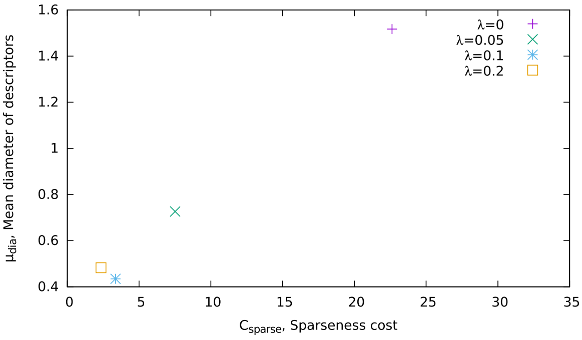


Fig. 12 Mean diameter versus sparseness cost of the descriptor groups (each symbol denotes a different λ)

C_{sparse} and smaller C_{sparse} brings smaller μ_{dia}. Smaller μ_{dia} means that descriptors corresponding to a similar interaction are more aggregated. This result shows that the proposed method can unsupervisedly generate interaction descriptors that make aggregates corresponding to hand–object interactions.

Table 1 Purity of clusters of descriptors

Autoencoder	Purity
Without the sparseness cost	0.16
With the sparseness cost (the proposed method)	0.99

To compare distributions of descriptors by autoencoders trained with/without the sparseness cost C_{sparse} in (4), we calculated the purity, an evaluation measure of clustering quality, The purity is defined by how many samples in a cluster belong to the most frequent class label (correct label given manually) in the cluster as follows;

$$(\text{PURITY}) \overset{\text{def}}{=} \sum_{c:\text{cluster index}} \frac{1}{n_c} \max_{i:\text{interaction type}} n_{c,i}, \tag{10}$$

where n_c means the number of samples assigned to the c-th cluster, and $n_{c,i}$ means the number of samples from the i-th type interaction assigned to the c-th cluster. If the purity of clusters is close to 1, almost all descriptors of each cluster belong to a common interaction type. This means that the clusters generated without information of interaction types approximately form subdivision of interaction types. We calculated descriptors from 800 hand–object interaction images not used in the training for each autoencoder, and applied mean-shift clustering to the descriptors. Table 1 shows the purities for autoencoders trained with/without the sparseness cost C_{sparse}. The purity for the autoencoder trained with the sparseness cost (the proposed method) was 0.99, while that for that without sparseness cost was 0.16. 99 percent of descriptors in a cluster by the proposed method belong to a common interaction type. This implies that the descriptors for the same interaction type are more separately embedded by the autoencoder trained with sparseness cost than without sparseness cost. It is important that these clusters with the similar interaction was unsupervisedly generated from only image signals.

Figure 13 shows the distribution of various object images within 11th and 18th dimensions of the interaction de-

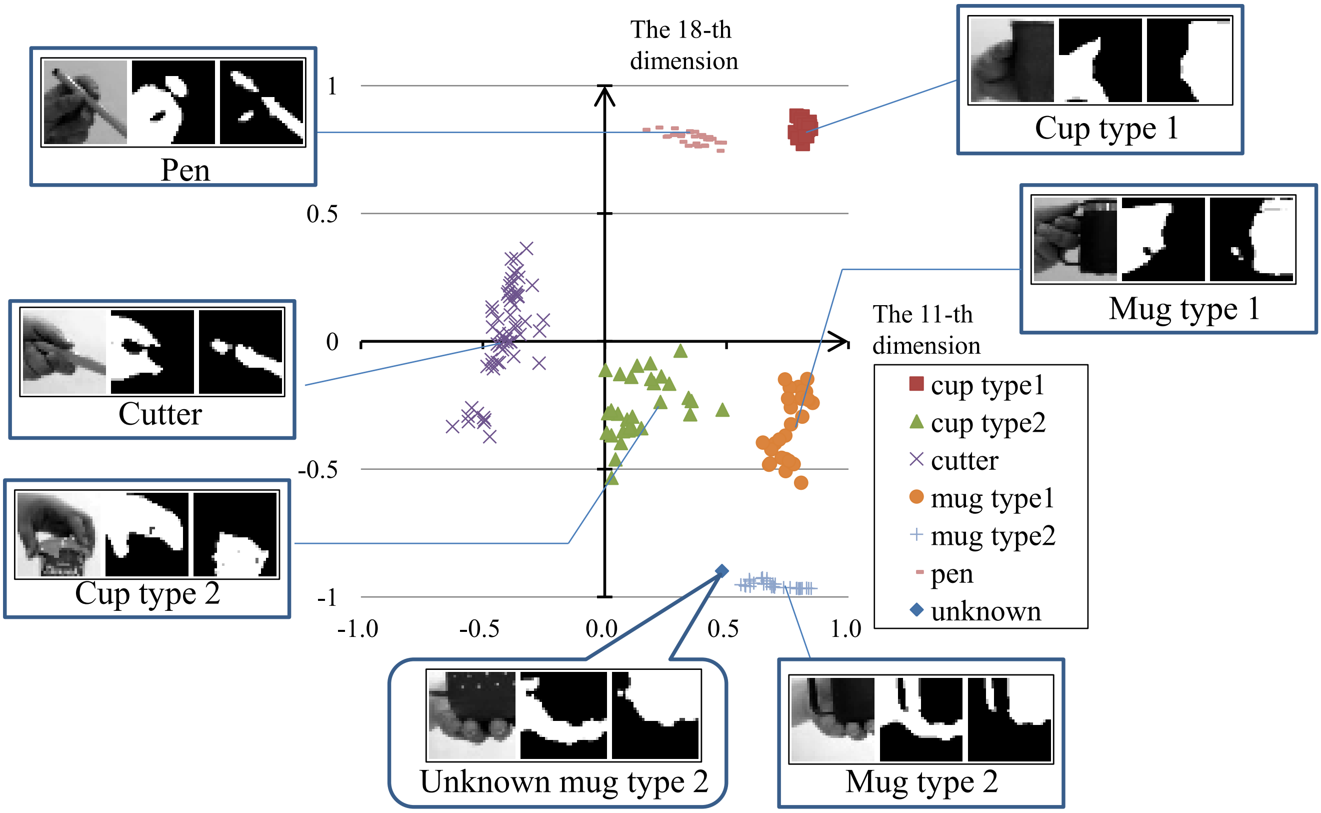


Fig. 13 One plane of the interaction descriptor space

scriptor space. “Mug type 1” and “Mug type 2” are different hand–mug interactions. In the former interaction, the hand grips the mug’s handle; in the latter, the hand holds the mug from the bottom. As shown in Fig. 13, these interactions form two separate clusters in the descriptor space. In addition, the interaction image “Mug type 2”, which was not used in training, maps to a descriptor near those of the mug–hand interaction images used in training. As a group of similar interactions composes a cluster in the interaction descriptor space, that space well characterizes the types of hand–object interactions.

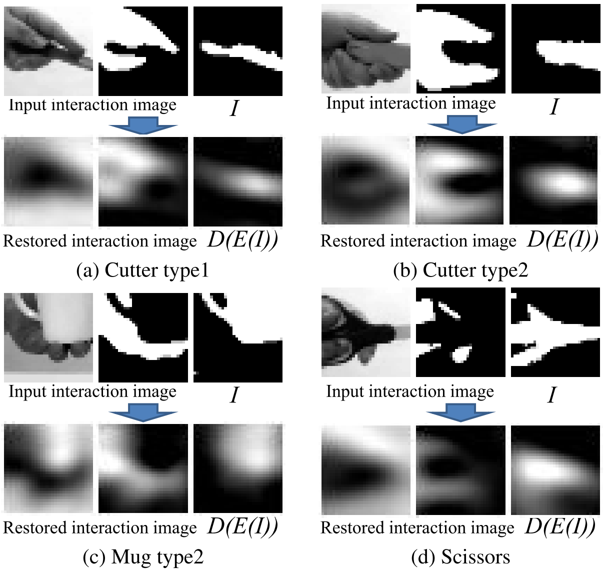
6.2 Restoration by the Decoder

Figure 14 shows examples of the decoder restorations. The encoder abstracts the rough shapes and positions of image features, ignoring their specific textures.

6.3 Inference of an Interaction

Figure 15 shows interaction images inferred from appearances of objects. In these examples, the pairs of masks of the inferred hand and object show their positional relations in the possible interaction. The hand–object interactions were successfully inferred from the single object images.

Figure 16 shows the region of interaction between scissors and a human hand. The colored regions mark the center of a window in which the likelihood f exceeds 0.9. High and



(a) Cutter type1

(b) Cutter type2

(c) Mug type2

(d) Scissors

Fig. 14 Restoration by autoencoder

low likelihoods are inferred around the grip and edges of the scissors, respectively, reflecting the human interactions with the grip of scissors and avoidance of the edge in teacher images. The inference model can therefore infer the interaction regions of human hands and scissors.

Figure 17 is a color representation of the inferred interaction descriptors for parts of a cup. The color of each

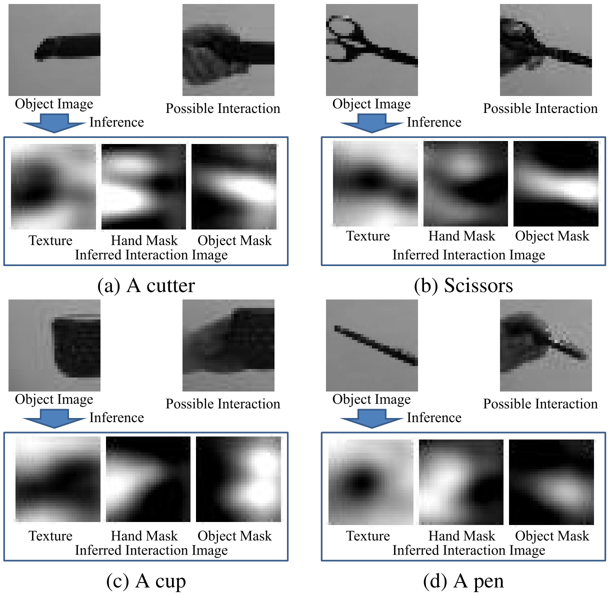


Fig. 15 Interaction images inferred from unknown objects

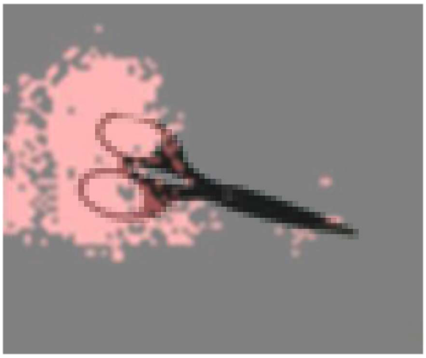


Fig. 16 Regions (pink) of possible interaction between human hand and scissors

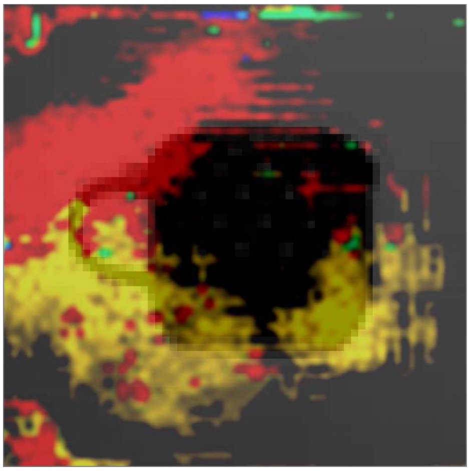


Fig. 17 Inferred descriptors for interactions with handle (red) and bottom (yellow) of a cup

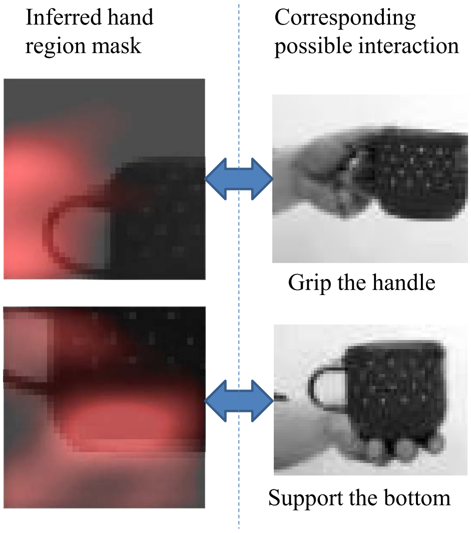


Fig. 18 Inferred hand-region masks and their possible interaction

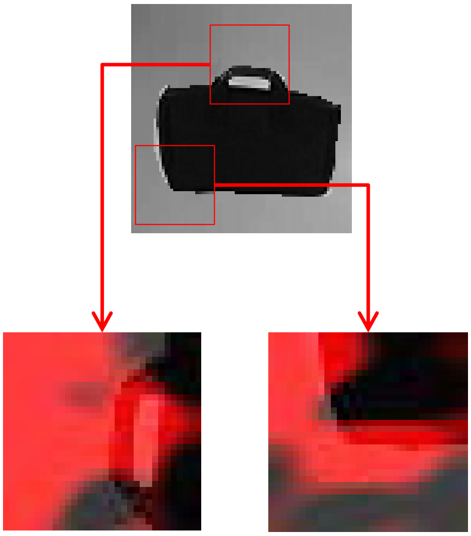


Fig. 19 Hand-region masks inferred from an object in unknown category

position is determined by mean shift clustering of the vectors containing the inferred descriptor and the position. This figure reveals the different types of interactions inferred on the grip and the bottom of the cup. Figure 18 shows the hand-region masks inferred in these two interaction types. The model can infer an interaction descriptor corresponding to possible interaction at a particular position.

To demonstrate that the proposed method can infer a possible interaction from an appearance of an unknown object in an unknown category, we inferred interaction descriptors from an appearance of a bag shown in the top image in Fig. 19. No bags are not used in training, but a grip of the bag has an appearance similar to that of a mug used in training. The bottom left image in Fig. 19 is a hand region mask inferred from an image around a grip of the bag. It shows that the inference model can infer a hand region mask like handling the grip of the bag from an image rotated so that the direction of the grip is close to that of a mug used in training the inference model. This means that the inference model learns the relation between a grip-like shape and an interaction for handling it. And also, as shown in the bottom right image in Fig. 19, a hand region like supporting the bottom is inferred from the other part of the bag, which is similar a bottom of a mug.

This indicates that the inference model can infer a possible interaction from a partial appearance of an unknown object in an unknown category if the model is trained with similar partial appearances included in other objects.

To evaluate the proposed inference model, we compared an interaction image inferred from an appearance of an object with a real instance of an interaction image occurred with the same object. To compare the two interaction images quantitatively, we calculated mean peak signal to noise ratios (PSNRs) between them for each channel (total appearance, hand region mask and object region mask), which is defined as below.

$$\frac{1}{N}\sum_{(I,I_{\mathrm{obj}})} 10\log_{10}\frac{V_c^2}{\frac{1}{M}\left\|[I]_c-\left[D\left(R\left(I_{\mathrm{obj}}\right)\right)\right]_c\right\|_{\mathrm{L2}}^2}, \tag{11}$$

where

$$\begin{aligned}
[I]_c &= (\text{the } c\text{-th channel of the interaction image } I),\\
I_{\mathrm{obj}} &= (\text{an object appearance}),\\
I &= (\text{the interaction image to } I_{\mathrm{obj}}).\\
N &= (\text{the number of samples}),\\
M &= (\text{the number of pixels in a channel}),\\
V_c &= (\text{the possible maximum pixel value}\\
&\quad \text{of the } c\text{-th channel}).
\end{aligned} \tag{12}$$

The calculated values of PSNRs are shown in Table 2. The values of PSNRs for training samples are from 8[dB] to 10[dB]. They are lower than 20[dB] that indicates unacceptable image quality in image compression [27]. This is because the autoencoder extracts essential components common to some appearances of interactions instead of encoding detail of each interaction image. However, Fig. 15 shows that the proposed method can infer rough shapes of a hand and an object. From an appearance of a cutter, a hand mask region like grasping the cutter is inferred. An object hand mask region inferred from the cutter indicates a narrow and long region and it matches rough shape of the grip of the cutter. From an appearance of a cup, a hand mask region like grasping the cup is inferred. Although the value of PSNRs are not high, the proposed method can roughly infer a possible interaction.

Table 2 Mean PSNR of the inference model

	Mean PSNR [dB] for each channel		
	Total appearance	Hand region	Object region
for training samples	8.80	10.33	10.94
for test samples	3.66	6.53	7.22

7. Conclusion

We proposed the *interaction descriptor space* for describing the hand–object interactions of functional objects. The space is automatically constructed from sets of object-handling images of typical functional objects such as mugs, scissors, cutters. We demonstrated that a descriptor corresponds to a quantitative interaction state and descriptors make clusters consistent with interaction types. We also proposed an inference model that infers a possible interaction from an object image alone. Given an object image, the model successfully inferred an interaction descriptor corresponding to a possible interaction at each position of the image.

The interaction descriptor space can characterize hand–object interactions and it can be used to model the relations between an object and its possible interactions. The proposed approach is a potentially valuable tool in function-based classification.

References

[1] L. Stark and K. Bowyer, "Achieving generalized object recognition through reasoning about association of function to structure," IEEE Trans. Pattern Anal. Mach. Intell., vol.13, no.10, pp.1097–1104, Oct. 1991.

[2] K. Bowyer, M. Sutton, and L. Stark, "Object recognition through reasoning about functionality: A survey of related work," Object Categorization: Computer and Human Vision Perspectives, p.129, 2009.

[3] D. Bub and M. Masson, "Gestural knowledge evoked by objects as part of conceptual representations," Aphasiology, vol.20, no.9, pp.1112–1124, 2006.

[4] J.R. Napier, "The prehensile movements of the human hand," J. Bone and Joint Surgery, vol.38, no.4, pp.902–913, 1956.

[5] D.A. Norman, "Affordance, conventions, and design," Interactions, vol.6, no.3, pp.38–43, May 1999.

[6] Z. Wang, X. Wang, and Q. Ji, "Learning with hidden information," Pattern Recognition (ICPR), 2014 22nd International Conference on, pp.238–243, Aug. 2014.

[7] Z. Wang, T. Gao, and Q. Ji, "Learning with hidden information using a max-margin latent variable model," Pattern Recognition (ICPR), 2014 22nd International Conference on, pp.1389–1394, Aug. 2014.

[8] X. Sun, T. Matsuzaki, and W. Li, "Latent structured perceptrons for large-scale learning with hidden information," Knowledge and Data Engineering, IEEE Transactions on, vol.25, no.9, pp.2063–2075, Sept. 2013.

[9] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," Neural Networks, vol.22, no.5-6, pp.544–557, 2009. Advances in Neural Networks Research: {IJCNN20092009} International Joint Conference on Neural Networks.

[10] A. Pieropan, C.H. Ek, and H. Kjellström, "Functional object descriptors for human activity modeling," Robotics and Automation (ICRA), 2013 IEEE International Conference on, pp.1282–1289, May 2013.

[11] D. Song, N. Kyriazis, I. Oikonomidis, C. Papazov, A. Argyros, D. Burschka, and D. Kragic, "Predicting human intention in visual observations of hand/object interactions," Robotics and Automation (ICRA), 2013 IEEE International Conference on, pp.1608–1615,

May 2013.

[12] J. Masci, U. Meier, D. Cirean, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in Artificial Neural Networks and Machine Learning ICANN 2011, ed. T. Honkela, W. Duch, M. Girolami, and S. Kaski, Lecture Notes in Computer Science, vol.6791, pp.52–59, Springer Berlin Heidelberg, 2011.

[13] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol.86, no.11, pp.2278–2324, Nov. 1998.

[14] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima," Neural Networks, vol.2, no.1, pp.53–58, 1989.

[15] A. Makhzani and B.J. Frey, "k-sparse autoencoders," CoRR, vol.abs/1312.5663, 2013.

[16] I. Ramirez, P. Sprechmann, and G. Sapiro, "Classification and clustering via dictionary learning with structured incoherence and shared features," Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp.3501–3508, June 2010.

[17] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, New York, NY, USA, pp.689–696, ACM, 2009.

[18] H. Lee, C. Ekanadham, and A.Y. Ng, "Sparse deep belief net model for visual area v2," in Advances in Neural Information Processing Systems 20, ed. J. Platt, D. Koller, Y. Singer, and S. Roweis, pp.873–880, Curran Associates, 2008.

[19] D.L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," Proceedings of the National Academy of Sciences, vol.100, no.5, pp.2197–2202, 2003.

[20] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE Conference on, pp.1794–1801, June 2009.

[21] H. Bristow, A. Eriksson, and S. Lucey, "Fast convolutional sparse coding," Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pp.391–398, June 2013.

[22] S. Morioka, T. Matsuo, Y. Hiramoto, N. Shimada, and Y. Shirai, "Automatic image collection of objects with similar function by learning human grasping forms," in Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction, ed. F. Schwenker, S. Scherer, and L.P. Morency, Lecture Notes in Computer Science, vol.8869, pp.3–14, Springer International Publishing, 2015.

[23] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," SIAM Review, vol.43, no.1, pp.129–159, 2001.

[24] P. Yin, E. Esser, and J. Xin, "Ratio and difference of l1 and l2 norms and sparse representation with coherent dictionaries," Commun. Inform. Systems, vol.14, no.2, pp.87–109, 2014.

[25] N. Hurley and S. Rickard, "Comparing measures of sparsity," CoRR, vol.abs/0811.4706, 2008.

[26] G. Cheng, P. Zhou, and J. Han, "Rifd-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[27] N. Thomos, N.V. Boulgouris, and M.G. Strintzis, "Optimized transmission of jpeg2000 streams over wireless channels," IEEE Trans. Image Process., vol.15, no.1, pp.54–67, Jan. 2006.

Tadashi Matsuo received the B.E., M.E. and Ph.D. degrees from Kyoto Institute of Technology in 2001, 2003 and 2006, respectively. He was a researcher in Research Organization of Science and Engineering, Ritsumeikan University from 2006 to 2011. He was an assistant in College of Information Science and Engineering, Ritsumeikan University in 2012, and is currently a TOKUNIN assistant professor. His research interest includes image recognition and machine learning. He is a member of IEICE.

Nobutaka Shimada received the B.E., M.E. and Ph.D. degrees from Osaka University in 1992, 1994 and 1997, respectively. He was an assistant in Graduate School of Engineering, Osaka University from 1997 to 2002. In 2003, He was an associate professor in Graduate School of Engineering, Osaka University. He was an associate professor in College of Information Science and Engineering, Ritsumeikan University from 2004 to 2011, and is currently a professor. In 2007, he was engaged in research as a visiting associate professor in the Robotics Institute, Carnegie Mellon University. His research interest includes computer vision, gesture interface and interactive robots. He is a member of IEEE, IEICE, IPSJ and JSAI.