The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Object Concept Modeling Based on the Relationship among Appearance, Usage and Functions

Tomoaki Nakamura and Takayuki Nagai
Department of Electronic Engineering, The University of Electro-Communications
1-5-1 Chofugaoka Chofu-shi, Tokyo 182-8585, Japan
naka_t@apple.ee.uec.ac.jp, tnagai@ee.uec.ac.jp

Abstract— **In this paper, a novel object concept model, which encodes the relationship among appearance, functions and usage, is proposed. The essential attribute of an object (artifact) is its function that achieves a particular purpose. Therefore, the function model is constructed through observations from a camera at first. The function is defined as changes in the work object before and after tool use. At the same time, the usage model is constructed from observations of the hand shape, grasping parts, and contact points of the tool. And then, the proposed system learns the object concept that is based on the relationship among appearance, and learnt function and usage models. The object appearance is represented by SIFT (Scale Invariant Feature Transform). Since the proposed models are based on the graphical model, it is possible for the system to stochastically infer unobservable information from observed one. For example, the system can infer usage and/or functions of the tool visually through the proposed model. Some experimental results using the system, in which the proposed model is implemented, are shown to validate the proposed model.**

I. INTRODUCTION

Recognition and understanding of surrounding environments by computers have been an active research area since they are essential functions for intelligent systems such as autonomous mobile robots. Although a vast amount of research on object recognition has been conducted, many of them rely only on visual information of objects [1]–[4]. However, appearance of objects is not sufficient information from the view point of "understanding" of objects. Since each object has its own intended use, it is unavoidable to consider functions and usage of objects.

In the area of developmental psychology, there have been many research on human (infants) object categorization. Researchers argued that function is a critical aspect of categorizing most human artifacts. As [5] concluded, common sense tells us that if we know what an object is we often know what it does. This fact suggests that the object recognition must be carried out through the relationship between form and function. Moreover, such relationship is also important for defining "understanding". In [7], authors have proposed the multi-modal categorization that leads to the definition of "understanding". In that paper, we assume that the robot can understand objects by inferring unobserved properties through the learnt categories.

This paper attempts to extend the idea of understanding to hand tools. More specifically, we define "object understanding" as a prediction of its function and usage from its appearance, which is grabbed by a visual sensor. To this end, we propose an object concept model, which encodes the relationship among appearance, functions and usage, in this paper. The model is based on a graphical model, which represents object concept as a whole, while it can be divided into parts that represent concepts of functions and usage. The function is modeled by Gaussian distribution of visual changes in a work object, which is influenced by the target object. Usage is modeled by a multinomial distribution of features including the hand shape for grasping, grasping parts, and contact points of the tool. Before forming the entire object concept, parameters of these partial concept models, i.e. function and usage, are learnt by Variational Bayesian method [3]. SIFT (Scale Invariant Feature Transform) [4] is used for representing appearance because local features are suitable for visual representation of tools. The parameters of the entire object concept model are learnt by EM-algorithm in conjunction with the pre-learnt concept models. Finally, the learnt model enables inference such as recognizing object, inference of functions and usage from visual information. As we defined earlier, such inference can lead to the understanding of objects.

Related works include visual object categorization and recognition. Recently, generic object recognition has been widely studied in the area of computer vision. The framework makes it possible for computers to recognize objects with their category names. Such idea is closely related to our problem; however, functions and usage of objects are not incorporated in the framework. By contrast, we argue that functions and usage are relevant to recognize objects (at least artifacts).

In the past, there have been some works on the object recognition considering object functions [8]–[12]. In [8], physical properties are employed to recognize objects. For example, a chair is identified based on the physical property that a person can sit on. Although parameters are learnt automatically from examples, knowledge about object categories, which is called category definition tree, must be designed manually. Moreover, it is not straightforward to apply these methods to hand tools. In contrast, physical changes in the work object, which is influenced by a tool, is observed and the concept of function itself is discovered automatically in this paper. Furthermore, learning of usage is needed for understanding of objects. In this paper we define usage as; 1) the hand shape for grasping the object, 2) grasping parts, and 3) parts of the tool that contact with a work object. These are observed while human use of tools and then the concept of usage is formed.

As for usage, [13] proposes object recognition based on the relationship between the object and human motions. In the area of intelligent robotics, robot manipulation of human tools is not an easy problem [14]. We believe that the proposed model contributes to develop a robot that can manipulate everyday objects.

This paper is organized as follows: We propose an object

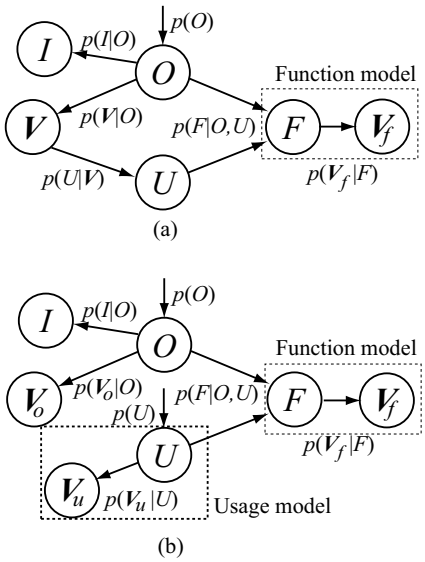

Fig. 1. The object concept model; (a)the object model based on appearance, functions and usage, and (b)the simplified object model.

concept model in the following section. Then, the models of function and usage are discussed in detail in III and the algorithms for learning and inference are explained in IV. The section V is devoted to experiments in order to validate the proposed model and finally, we conclude this paper in VI.

II. Object Concept

In [6], we have proposed a preliminary model of object concept based on function and shape. In this section, we extend the model by adding a usage node to it. Figure 1(a) illustrates the extended graphical model which contains the usage node. In the figure, O, $\boldsymbol{V}$, F, U, I and $\boldsymbol{V}_f$ represent the object category, visual information, functions, usage, object ID and visual features for function (it will be discussed later), respectively. It should be noted that O, F, and U are latent variables. This model represents that usage is determined by visual features of the object and, the object category and usage affect the function of the object.

Here, we assume that visual information $\boldsymbol{V}$ can be divided into appearance of the object $\boldsymbol{V}_o$ and the visual observation of the scene $\boldsymbol{V}_u$, in which the tool is manipulated. This partitioning of $\boldsymbol{V}$ makes the model simpler as in Fig.1(b). The function model which consists of F and $\boldsymbol{V}_f$ is modeled independently and then it is integrated into the entire model of the object concept. The usage model containing U and $\boldsymbol{V}_u$ is also modeled separately. Joint probability of the object concept model can be written as

$$p(O, I, \boldsymbol{V}_o, F, U, \boldsymbol{V}_u, \boldsymbol{V}_f) = p(O)p(I|O)p(\boldsymbol{V}_o|O) \times p(F|O,U)p(U)p(\boldsymbol{V}_u|U)p(\boldsymbol{V}_f|F), \quad (1)$$

where $p(\boldsymbol{V}_u|U)$ and $p(\boldsymbol{V}_f|F)$ are computed independently of the object concept model and fixed when parameters are learnt. Details on learning and inference methods will be discussed later.

III. Appearance, Usage and Functions

We explain each node of the model in Fig.1(b).

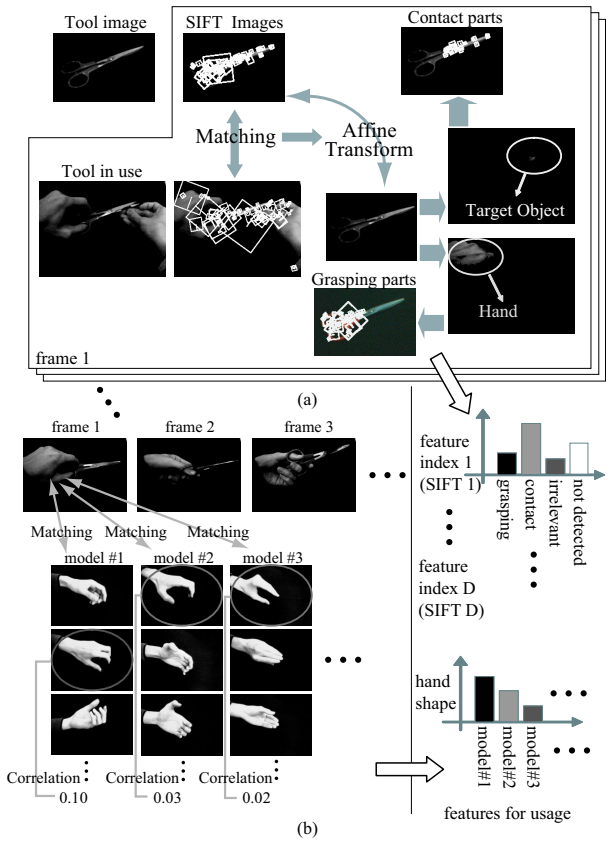

Fig. 2. Image processing for the usage model: (a)detection of grasping and contact parts, and (b)estimation of hand shpae.

A. Object appearance ($\boldsymbol{V}_o$)

SIFT [4] is used for representing visual information of objects. The SIFT detects key points and features are calculated around them. The SIFT descriptor takes a histogram of the gradient features and it is robust over changes in rotation and brightness. However, the number of features varies according to the image contents and it is inconvenient to use as the visual feature of tools. For this reason, we take the Bag of Features approach. The SIFT descriptors are vector quantized using a predefined D dimensional codebook. Therefore each SIFT descriptor is transformed into a feature index ($\in \{1, \ldots, D\}$). Finally we take the histogram $\boldsymbol{V}_o (= \{V_{o,1}, \ldots, V_{o,D}\})$ of the feature index as visual information.

B. Usage model (U)

We define usage of a tool based on the following three aspects; the grasping parts, the contact parts with a work object and the hand shape when it is manipulated. These are extracted from the scene, in which the object is in use.

1) Observations for grasping and contact parts: The image processing for obtaining grasping parts and contact parts with a work object is illustrated in Fig.2 (a). During the tool is in use, SIFT descriptors are calculated in each frame. Corresponding points between a frame of the tool use sequence and an image of the tool, which is captured before the usage, are searched. The affine parameters are calculated using the corresponding points to deform the tool image so that it fits to the tool in the usage sequence. Then the hand and the work object are extracted from each frame of the

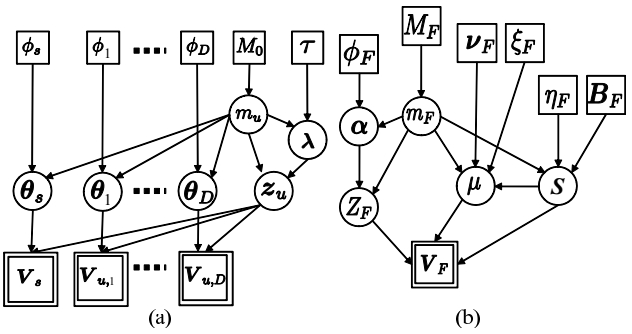

Fig. 3. The graphical model of each concept; (a)the model of usage, and (b)the model of function.

tool use sequence using color information. By overlapping the extracted hand region with the deformed tool image, SIFT descriptors for the grasping part can be detected. SIFT descriptors for the contact points can be also detected in the same way using the extracted work object region. Thus each feature in the tool use images is classified into following three categories; to be grasped, to contact with a work object and irrelevant to usage. Finally, each feature index $V_{o,i}$ has four attributes, that is, the grasping part, the contact part with a work object, irrelevant to usage, and not detected. The occurance frequency of these four attributes are modeled by a multinomial distribution.

2) Observation for the hand shape: The hand shape is estimated by the SIFT-based matching with the hand shape model. The model consists of many hand images, which are taken from various view points in advance. The hand shape is determined by the sum of matching results in all frames. Figure2 (b) illsutrates the method of hand shape estimation. The maximum value in correlations between each hand shape in the database and the input frame is used as correlation between the model and the input frame. Then the sum of correlations of all frames is calculated with respect to each hand model. These are also modeled by a multinomial distribution.

3) Usage model: Figure3 (a) is the graphical model of usage including conjugate prior distributions. $\boldsymbol{\theta}_d$, $\boldsymbol{z}_u$, m_u and $\boldsymbol{\lambda}$ are parameters of the multinomial distribution, latent variable (usage category), the number of models and mixture ration, respectively. M_0, $\boldsymbol{\phi}_*$ and $\boldsymbol{\tau}$ denote hyper parameters. $\boldsymbol{V}_{u,d}$ is the observable variable and represents the frequency of four attributes of the feature index d in a scene. $\boldsymbol{V}_h(=\{V_{h1},\ldots,V_{hN_h}\})$ represent a hand shape information and V_{hi} represents a normalized sum of correlation coefficients for i-th hand shape model.

C. Function model (F)

The concept of function is formed by clustering of changes in a work object. Here, four features are computed considering properties of general hand tools. (1) Color change on the surface of the work object; this change can be captured by computing the correlation coefficient between color histograms of the work object before and after manipulation. (2) Contour change of the work object; to capture this change the correlation coefficient between Fourier descriptors of the work object before and after manipulation is computed. (3) Barycentric position change of the work object; the relative distance between barycentric positions of the work object before and after manipulation is computed. (4) Change in number of the work object; this can be detected by counting the connected components relevant to the work object.

These four parameters are treated as a four-dimensional vector and it is modeled by Gaussian distribution. The function model is illustrated in Fig.3 (b). In this figure, $\boldsymbol{V}_F$ represents the observable feature vector of the work object. m_F, $\boldsymbol{\mu}$, $\boldsymbol{S}$ and $\boldsymbol{\alpha}$ denote the number of functions, mean vector, covariance matrix and mixture ratio, respectively. $\boldsymbol{Z}_F$ is a latent variable, which represents the usage category. The rectangles in the figure are hyper parameters. ϕ_F is a parameter of Dirichlet prior distribution, which is a hyper parameter of $\boldsymbol{\alpha}$. $\boldsymbol{\nu}$ and ξ_F $\boldsymbol{S}$ are the mean vector and covariance matrix of Gaussian distribution which is a prior distribution of $\boldsymbol{\mu}$. η_F and $\boldsymbol{B}_F$ are degree of freedom and the covariance matrix of Wishart distribution which is a prior distribution of $\boldsymbol{S}$.

IV. Learning and Recognition

A. Learning of usage and function

The variational Bayesian method [3] is utilized to learn the parameters of usage and function models. In the variational Bayesian approach, the following marginal likelihood is considered:

$$L(\mathbf{D}) = \log p(\mathbf{D}) = \log \sum_m \sum_{\mathbf{Z}} \int p(\mathbf{D},\mathbf{Z},\boldsymbol{\theta},m)d\boldsymbol{\theta}, \quad (2)$$

where $\mathbf{D}$, $\mathbf{Z}$, $\boldsymbol{\theta}(=\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_I\})$, I, and m represent observations, latent variables, prameters of the model, number of parameters, and the model structure. Now the variational posterior $q(\mathbf{Z},\boldsymbol{\theta},m)$ is introduced to make the problem tractable. Then, $L(\mathbf{D})$ can be written as follows:

$$\begin{aligned} L(\mathbf{D}) &= \log \sum_m \sum_{\mathbf{Z}} \int q(\mathbf{Z},\boldsymbol{\theta},m)\frac{p(\mathbf{D},\mathbf{Z},\boldsymbol{\theta},m)}{q(\mathbf{Z},\boldsymbol{\theta},m)}d\boldsymbol{\theta} \\ &= \sum_m \sum_{\mathbf{Z}} \int q(\mathbf{Z},\boldsymbol{\theta},m)\log\frac{q(\mathbf{Z},\boldsymbol{\theta},m)}{p(\mathbf{Z},\boldsymbol{\theta},m|\mathbf{D})}d\boldsymbol{\theta} \\ &\quad + \sum_m \sum_{\mathbf{Z}} \int q(\mathbf{Z},\boldsymbol{\theta},m)\log\frac{p(\mathbf{D},\mathbf{Z},\boldsymbol{\theta},m)}{q(\mathbf{Z},\boldsymbol{\theta},m)}d\boldsymbol{\theta} \\ &\equiv \mathrm{KL}(q(\mathbf{Z},\boldsymbol{\theta},m),p(\mathbf{Z},\boldsymbol{\theta},m|\mathbf{D})) + F[q], \end{aligned} \quad (3)$$

where $F[q]$ and KL denote free energy and Kullback-Leibler divergence, respectively. Since $L(\mathbf{D})$ does not depend on q, maximization of $F[q]$ with respect to q is equivalent to minimization of Kullback-Leibler divergence between q and p. Therefore, variational posterior q, which maximizes $F[q]$, is the best approximation to the true posterior p. Then the problem becomes the maximization of the free energy $F[q]$ with respect to q:

$$\begin{aligned} F[q] = \sum_m q(m)\Bigg\{ &\left\langle \log\frac{p(\mathbf{D},\mathbf{Z}|\boldsymbol{\theta},m)}{q(\mathbf{Z}|m)} \right\rangle_{q(\mathbf{Z}|m),q(\boldsymbol{\theta}|m)} \\ &+ \sum_{i=1}^{I} \left\langle \log\frac{p(\boldsymbol{\theta}_i|m)}{q(\boldsymbol{\theta}_i|m)} \right\rangle_{q(\boldsymbol{\theta}_i|m)} + \log\frac{p(m)}{q(m)} \Bigg\}. \end{aligned} \quad (4)$$

Finally, we obtain

$$q(\mathbf{Z}|m) \propto \exp\left\langle \log p(\mathbf{D},\mathbf{Z}|\boldsymbol{\theta},m)\right\rangle_{q(\boldsymbol{\theta}|m)}, \quad (5)$$

For $i = 1,\ldots,I$

$$\begin{aligned} q(\boldsymbol{\theta}_i|m) &\propto p(\boldsymbol{\theta}_i|m) \\ &\quad \times \exp\left\langle \log p(\mathbf{D},\mathbf{Z}|\boldsymbol{\theta},m)\right\rangle_{q(\mathbf{Z}|m),q(\boldsymbol{\theta}_{-i}|m)}. \end{aligned} \quad (6)$$

We solve these equations iteratively to obtain the optimal variational posterior q.

The variational posterior of the model structure $q(m)$ can be written as

$$q(m) \propto p(m) \exp\left(F_m\right), \tag{7}$$

where

$$F_m = \left\langle \log \frac{P(\mathbf{D}, \mathbf{Z}|\boldsymbol{\theta}, m)}{q(\mathbf{Z}|m)} \right\rangle_{q(\mathbf{Z}|m), q(\boldsymbol{\theta}|m)} + \sum_{i=1}^{I} \left\langle \log \frac{P(\boldsymbol{\theta}_i|m)}{q(\boldsymbol{\theta}_i|m)} \right\rangle_{q(\boldsymbol{\theta}_i|m)}. \tag{8}$$

F_m does not depend on $q(m)$. Hence, if we assume that $p(m)$ is a uniform distribution, the maximization of $q(m)$ is equivalent to maximization of F_m with respect to $q(\boldsymbol{\theta}_i|m)$ and $q(\mathbf{Z}|m)$. Thus, we can estimate the model structure m by the Variational Bayesian method that maximizes free energy. The models of function and usage can be learnt by computing (5) and (6).

B. Learning of the object concept

Learning of the object concept corresponds to estimation of parameters, which are conditional probabilities of the model. Observable variables are appearance of the object $\boldsymbol{V}_o$, visual information for function $\boldsymbol{V}_f$, and visual information for usage $\boldsymbol{V}_u$. The EM algorithm is utilize to estimate parameters because the model includes the latent variable O, which denotes the object category. The log likelihood of the proposed model can be written as

$$L(\boldsymbol{D}) = \log \sum_U \sum_F \sum_O p(I, \boldsymbol{V}_o, F, U, O|\boldsymbol{\theta}_o) p(\boldsymbol{V}_f|F) p(\boldsymbol{V}_u|U), \tag{9}$$

where $\boldsymbol{D}$ denotes a set of observable variables I, $\boldsymbol{V}_o$, $\boldsymbol{V}_f$ and $\boldsymbol{V}_u$. $p(\boldsymbol{V}_f|F)$ and $p(\boldsymbol{V}_u|U)$ are likelihoods of function F and usage U, respectively. Since function and usage models have been trained, these likelihoods can be calculated directly. In order to simplify the learning algorithm, we use the most likely F and U that provide maximum likelihood values $p(\boldsymbol{V}_f|F)$ and $p(\boldsymbol{V}_u|U)$, respectively. This approximation results in

$$L(\boldsymbol{D}) \approx \log \sum_O p(I, \boldsymbol{V}_o, F, U, O|\boldsymbol{\theta}_o) \times p(\boldsymbol{V}_f|F) p(\boldsymbol{V}_u|U). \tag{10}$$

It is worth noting that F and U are not latent variables at this time. By applying Jensen's inequality, $L(\boldsymbol{D})$ can be written as

$$L(\boldsymbol{D}) \geq F(q(O), \boldsymbol{\theta}_o) = p(\boldsymbol{V}_f|F) p(\boldsymbol{V}_u|U) \times \sum_O q(O|I, \boldsymbol{V}_o, F, U, \hat{\boldsymbol{\theta}}_o) \log \frac{p(I, \boldsymbol{V}_o, F, U|\boldsymbol{\theta}_o)}{q(O|I, \boldsymbol{V}_o, F, U, \hat{\boldsymbol{\theta}}_o)}. \tag{11}$$

Instead of maximizing the log likelihood $L(\boldsymbol{D})$, its lower bound $F(q(O), \boldsymbol{\theta}_o)$ is maximized with respect to $q(O)$ and $\boldsymbol{\theta}_o$ alternately. We assume that $\hat{\boldsymbol{\theta}}_o$ is an estimation of $\boldsymbol{\theta}_o$. The equality of (11) holds true under the following condition:

$$q(O|I, \boldsymbol{V}_o, F, U, \hat{\boldsymbol{\theta}}_o) = p(O|I, \boldsymbol{V}_o, F, U, \boldsymbol{\theta}_o). \tag{12}$$

Hence, maximization of $F(q(O), \boldsymbol{\theta}_o)$ with respect to $q(O)$ can be written as (E-step),

$$p(O|I, \boldsymbol{V}_o, F, U) = \frac{p(O)p(I|O)p(\boldsymbol{V}_o|O)p(F|O, U)}{\sum_O p(O)p(I|O)p(\boldsymbol{V}_o|O)p(F|O, U)}. \tag{13}$$

Maximization of $F(q(O), \theta)$ with respect to θ is to maximize the following Q-function:

$$Q(\boldsymbol{\theta}_o) = \langle \log p(\boldsymbol{V}_f|F) p(\boldsymbol{V}_u|U) \times p(O|I, V_o, F, U, \boldsymbol{\theta}_o) \rangle_{q(O|I, V_o, F, U, \hat{\boldsymbol{\theta}}_o)}. \tag{14}$$

This maximization problem can be solved by Lagrange multiplier method. Finally, we obtain following update equations (M-Step):

$$p(O) \propto \sum_I \sum_i \sum_F \sum_U \{N(V_{o,i}, F, U, I) p(\boldsymbol{V}_f|F) \times p(\boldsymbol{V}_u|U) p(O|I, V_{o,i}, F, U)\}, \tag{15}$$

$$p(I|O) \propto \sum_i \sum_F \sum_U \{N(V_{o,i}, F, U, I) p(\boldsymbol{V}_f|F) \times p(\boldsymbol{V}_u|U) p(O|I, V_{o,i}, F, U)\}, \tag{16}$$

$$p(V_{o,i}|O) \propto \sum_I \sum_F \{\sum_U N(V_{o,i}, F, U, I) p(\boldsymbol{V}_f|F) \times p(\boldsymbol{V}_u|U) p(O|I, V_{o,i}, F, U)\}, \tag{17}$$

$$p(F|O, U) \propto \sum_I \sum_i \{N(V_{o,i}, F, U, I) p(\boldsymbol{V}_f|F) \times p(\boldsymbol{V}_u|U) p(O|I, V_{o,i}, F, U)\}, \tag{18}$$

where $N(\cdot)$ denotes the frequency count of data used for the training.

C. Inference

Various stochastic inference scenarios are possible based on the learnt object concept model. We define such inference among appearance, usage, and functions using the object concept, which is formed through experiences, as understanding of the object. For example, recognition of the object from appearance $\boldsymbol{V}_o$, function F and usage U can be carried out as follows:

$$\underset{O}{\operatorname{argmax}}\, P(O|\hat{I}, \boldsymbol{V}_o, F, U) = \frac{p(O)p(\hat{I}|O)p(\boldsymbol{V}_o|O)p(F|O, U)}{\sum_O p(O)p(\hat{I}|O)p(\boldsymbol{V}_o|O)p(F|O, U)}, \tag{19}$$

where $p(\hat{I}|O)$ is recomputed using the EM algorithm described above to deal with a novel object $\hat{I}$.

Furthermore, inference of function F only from object appearance $\boldsymbol{V}_o$ can be done by

$$\underset{F}{\operatorname{argmax}}\, P(F|\hat{I}, \boldsymbol{V}_o) = \frac{\sum_O \sum_U \left\{p(O)p(\hat{I}|O)p(\boldsymbol{V}_o|O)p(F|O, U)\right.}{\sum_O \sum_U \sum_F \left\{p(O)p(\hat{I}|O)p(\boldsymbol{V}_o|O)p(F|O, U)\right.} \times \frac{p(U)p(\boldsymbol{V}_f|F)p(\boldsymbol{V}_u|U)\}}{p(U)p(\boldsymbol{V}_f|F)p(\boldsymbol{V}_u|U)\}}. \tag{20}$$

Fig. 4. The appearance of the system.

TABLE I
THE TOOLS USED IN THE EXPERIMENT.

Category	Index	# of tools in set A	# of tools in set B	Total
Scissors	T1	7	3	10
Pen	T2	8	3	11
Pliers	T3	2	2	4
Tweezers	T4	3	2	5
Cutter	T5	3	2	5
Stapler	T6	4	2	6
Glue	T7	5	3	8
Scotch tape	T8	4	3	7
Vinyl tape	T9	2	2	4

Inference of usage U only from visual information $\boldsymbol{V}_o$ is carried out by

$$\begin{aligned}\underset{U}{\operatorname{argmax}}\, P(U|\hat{I},\boldsymbol{V}_o) = \\ \frac{\sum_O \sum_F \left\{ p(O)p(\hat{I}|O)p(\boldsymbol{V}_o|O)p(F|O,U) \right.}{\sum_O \sum_U \sum_F \left\{ p(O)p(\hat{I}|O)p(\boldsymbol{V}_o|O)p(F|O,U) \right.} \\ \times \frac{p(U)p(\boldsymbol{V}_f|F)p(\boldsymbol{V}_u|U)\}}{p(U)p(\boldsymbol{V}_f|F)p(\boldsymbol{V}_u|U)\}}. \end{aligned} \quad (21)$$

Other inference is also possible in the same way as above.

V. EXPERIMENTS

A. Experimental setup

A total of 60 hand tools with 9 categories are used in the experiment. They are divided into two data sets. One (set A) is used for learning and the other (set B) is used for recognition. The details of these tools are given in Tab.I. It should be noted that we assign indices (T1, ..., T9) to each category as a matter of convenience. All tools are used 10 times each that results in 600 data in total. A scenery of the experiment is shown in Fig.4.

B. Model of object function

The parameters of the function model are learnt using the data of set A by the Variational Bayesian method. Fig.5 (a) shows the number of discovered functions versus free energy. In this figure, one can see that the optimal number of functions is six. Fig.5 (b) illustrates the classification result of tools in terms of six functions. In the figure, the number of objects, which are classified into each category, is visualized by a gray-scale bar. It should be noted that T1,$\cdots$,T9 in the figure represent indices of the category in Tab.I. From the figure, it can be figured out that functions 1 to 6 are "cut", "coloring", "deformation", "transfer", "bond" and "bond with coloring", respectively.

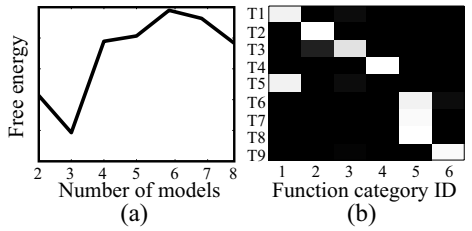

Fig. 5. The model of function: (a)number of functions versus free energy, (b)classification result.

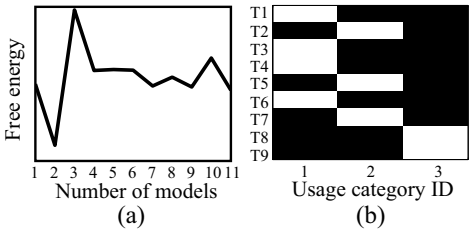

Fig. 6. The model of usage: (a)number of usage versus free energy, and (b)classification result.

C. Model of usage

The parameters of the usage model is learnt using the data of set A by the Variational Bayesian method. Fig.6 (a) shows the relationship between the number of models and free energy. In this case, the optimal number of usage is three. Fig.6 (b) shows the classification of tools based on the learnt usage. It is plausible that pen, utility knife and glue are classified into the same usage group because they are similar in shape. Moreover, they share a similar hand shape when they are grasped. Scotch tape and vinyl tape are classified in the same usage because both of them are grasped on a same position and have no common contact point with the work object. The hand shapes for grasping scissors, pliers, tweezers and stapler are similar to each other and they are classified in the same usage group.

D. Forming object concept

In this experiment, the parameters of the entire object model are trained using data set A. At this time, the models of functions and usage, which have been trained in the foregoing subsections, are used as the basic concepts. The results of classifications with three different information (i.e. appearance only, appearance and functions, and all of three information) are shown in Fig.7. Clearly, classification accuracy is getting better in the order of (a), (b) and (c). In fact, the correct classification rates of (b) and (c) are 89% and 93%, respectively.

E. Inference of functions from appearance

The inference of object functions from appearance is conducted by using the object model, which has been formed in V-D. All tools in set B are used and the most probable functions are estimated only from their appearance. Tab.II shows the result of inference. From the table, it can be confirmed that the inference works as high as 98% accuracy.

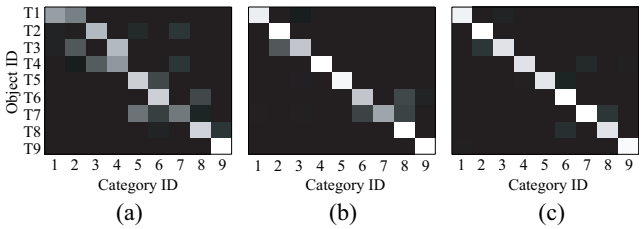

Fig. 7. The result of classification of tools: (a)classification by only appearance, (b)classification by appearance and functions, and (c)classification by appearance, functions and usage.

TABLE II
THE RESULT OF INFERENCE OF THE FUNCTION.

	cut	color	deformation	transfer	bond	bond & color
T1	30	0	0	0	0	0
T2	0	28	2	0	0	0
T3	1	0	19	0	0	0
T4	0	0	0	20	0	0
T5	20	0	0	0	0	0
T6	0	0	0	0	20	0
T7	0	0	0	0	30	0
T8	0	0	0	0	30	0
T9	0	0	0	0	0	20

F. Inference of usage from appearance

Here the grasping parts and contact points with the work object are estimated using the entire object concept model. We use data set B, which is not used in the learning phase. Some examples of inference are shown in Fig.8. The estimated grasping parts and contact points are illustrated by gray circles and white rectangles, respectively. One can see that the system can estimate these positions reasonably well only from its appearance.

G. Extraction of functional visual features

The visual features, which are shared by the tools in a same category, are shown in Fig.9. From the figure it can be seen that distinctive features, which we call functional visual features, for each object are extracted. These results indicate the potential to recognize objects and/or infer functions in the complex natural scenes.

VI. CONCLUSION

In this paper, we have proposed a novel object concept model based on appearance, functions and usage. The model consists of the function model and the usage one. All of these models are based on the graphical models that make it possible to stochastically infer unobservable information from observed one. The experimental results showed that the proposed system could reasonably infer functions and usage of objects only from their appearance. The implementation on a real robot is left for the future work.

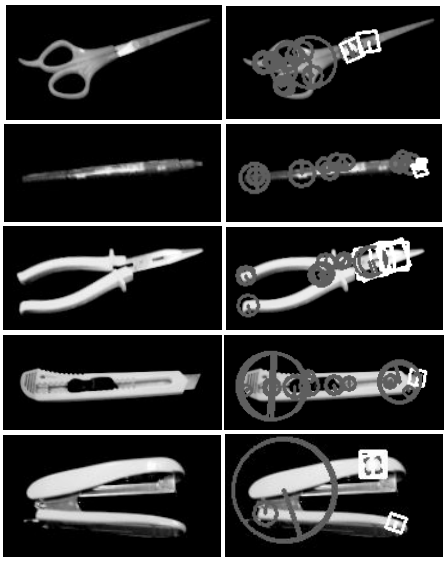

Fig. 8. The result of inference of usage.

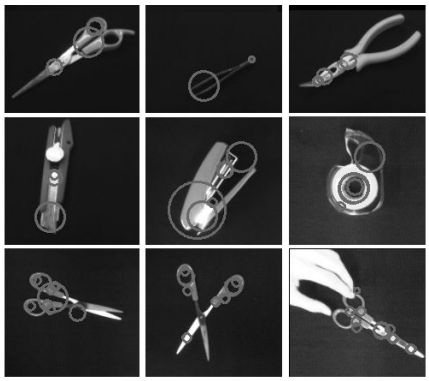

Fig. 9. The functional visual features.

REFERENCES

[1] J.Sivic, B.C.Russell, A.A.Efros, A.Zisserman, and W.T.Freeman, "Discovering objects and their location in images", in Proc. of ICCV2005, vol.1 pp.370–377, 2005.
[2] P.R.Fergus and A.Zisserman, "Object class recognition by unsupervised scale-invariant learning", in Proc. CVPR, pp.264-271, Feb.2003.
[3] H. Attias, "Infering Parameters and Structure of Latent Variable Models by Variational Bayes", in Proc. of Uncertainty in Artificial Intelligence, 1999
[4] David G. Low, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 2004.
[5] B. Landau and L. Smith and S. Jones, "Object Shape, Object Function, and Object Name", Journal of Memory and Language," ML972533, 38, pp.1–27,1998.
[6] Y.Shinchi, Y.Sato and T.Nagai, "Bayesian network model for object concept", in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, Vol.2, pp.473–476, Apr.2007.
[7] T.Nakamura, T.Nagai and N.Iwahashi, "Multimodal object categorization by a robot", in Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.2415–2420, Nov.2007.
[8] A.D.L.Stark and K.Bowyer, "Recognizing object functioin through reasoning about partial shape descriptions and dynamic physical properties", Proceedings of The IEEE, 84(11):1640-1656, Nov.1996.
[9] E.Rivlin, S.J.Dickinson and A.Rosenfeld, "Recognition by functional parts", Computer Vision and Image Understanding : CVIU, 62(2):164-176, 1995.
[10] K.Woods, D.Cook, L.Hall, K.W.Bowyer and L.Stark, "Learning membership functions in a function-based object recognition system", Journal of Artificial Intelligence Research, 3:187-222, 1995.
[11] R.S.Amant and A.B.Wood, "Tool Use for Autonomous Agents", AAAI, pp.184–189, 2005.
[12] I.Shimshoni, E.Rivlin and O.Soldea, "Efficient Search and Verification for Function Based Classification from Real Range Images", in Proc. of International Conference on Pattern Recognition (ICPR'06), 2006.
[13] A.Kojima, M.Higuchi, T.Kitahashi and K.Fukunaga, "Toward a Cooperative Recognition of Human Behaviors and Related Objects", The Second International Workshop on Man-Machine Symbiotic Systems, Kyoto, Japan, Nov.2004.
[14] C.C.Kemp, A.Edsinger and E.Torres-Jara, "Challenges for robot manipulation in human environments", IEEE Robotics & Automation Magazine, vol.14, pp.20–29, Mar.2007.