Understanding Hand-Object Manipulation with Grasp Types and Object Attributes

Minjie Cai
Institute of Industrial Science
The University of Tokyo, Japan
cai-mj@iis.u-tokyo.ac.jp

Kris M. Kitani
Robotics Institute
Carnegie Mellon University, USA
kkitani@cs.cmu.edu

Yoichi Sato
Institute of Industrial Science
The University of Tokyo, Japan
ysato@iis.u-tokyo.ac.jp

***Abstract*—Our goal is to automate the understanding of natural hand-object manipulation by developing computer vision-based techniques. Our hypothesis is that it is necessary to model the grasp types of hands and the attributes of manipulated objects in order to accurately recognize manipulation actions. Specifically, we focus on recognizing hand grasp types, object attributes and actions from a single image within an unified model. First, we explore the contextual relationship between grasp types and object attributes, and show how that context can be used to boost the recognition of both grasp types and object attributes. Second, we propose to model actions with grasp types and object attributes based on the hypothesis that grasp types and object attributes contain complementary information for characterizing different actions. Our proposed action model outperforms traditional appearance-based models which are not designed to take into account semantic constraints such as grasp types or object attributes. Experiment results on public egocentric activities datasets strongly support our hypothesis.**

I. Introduction

This work aims to automate the understanding of natural hand-object manipulation in daily tasks using a wearable camera. In particular, we focus on recognizing (1) grasp types, (2) object attributes and (3) actions from image appearance under first-person vision paradigm. These terms are defined as follows: *Grasp types* are a discrete set of canonical hand poses often used in robotics to describe various strategies for holding objects stably in hand. For example, the use of all fingers around a curved object like a cup is called a medium wrap. *Object attributes* characterize physical properties of the objects such as rigidity or shape. *Actions* in this work refer to different patterns of hand-object interactions such as open or pour.

The ability to understand hand-object manipulation automatically from visual sensing is important for the robotics community with potential applications such as robotic hand design and robotic action planning. In robotic hand design, the study of hand grasping behavior in daily manipulation tasks provides critical information about hand functions that can be used for robotic hand development [5, 35, 1, 6]. It can also facilitate robotic task planning by studying the relationship between different components (grasps, objects and actions) in performing a manipulation task [11, 33]. Wearable cameras enable continuous recording of unconstrained natural hand-object interactions at a large scale, both in time and space, and provides an ideal first-person point-of-view under which hands and objects are often visible up-close in the visual field. In this work, we develop automatic egocentric (first-person) vision techniques that can be used as a tool to promote the studies of hand manipulation in real-life settings.

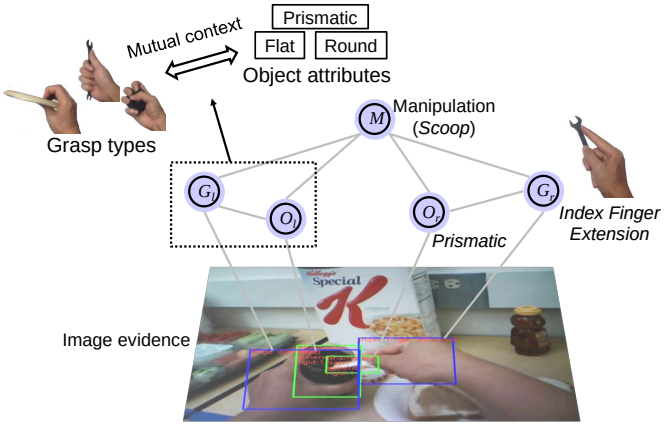

Fig. 1. Relationship between grasp types, object attributes, and manipulation actions. Grasp types and object attributes at both hands are learned from image evidence. Mutual context between grasp types and object attributes is explored. Manipulation actions are modeled based on grasp types and object attributes.

However, the recognition task for understanding daily manipulation tasks from monocular images is also very challenging. There are many occlusions of a hand, especially fingers, during hand-object interactions, making it hard to observe and recognize hand grasps. It is also challenging to reliably detect the manipulated object and infer attributes since the object is also often occluded by the hand. This suggests that visual information about hands and objects need to be reasoned about jointly by taking into account this mutual context.

We propose a vision-based approach to detect the grasped part of the object during manipulation by exploring spatial hand-object configurations. Attribute information is then extracted from the manipulated object. Furthermore, we propose to enhance the recognition of grasp types and object attributes by their mutual context (contextual relationship between two components that by knowing one component facilitates the recognition of the other). Object attributes (*e.g.* , thick or long shape of a bottle) have strong constraints on the selection of hand grasp types (*e.g.* , *Large Wrap*). Thus, with the

knowledge of object attributes, we are able to predict a large percentage of grasp types. On the other hand, humans use the same or similar grasp types for certain types of objects, thus the grasp type used reveals attributes of the object being grasped. We formulate a Bayesian model to encode the mutual context between grasp types and object attributes in which recognizing one facilitates the recognition of the other.

Based on the visual recognition of semantic information of grasp types and object attributes, we provide a semantic action model as illustrated in Figure 1. Specifically, we train discriminative classifiers for different actions based on the probabilistic estimation (belief distribution) of grasp types and object attributes.

There are several advantages for jointly modeling actions in this way: (1) Grasp type helps describe the functionality of an action, whether it requires more power, or more flexible finger coordination; (2) Object attributes provide a general description about the manipulated object and indicates possible interaction patterns; (3) Semantic information of grasp types and object attributes enable the model to encode high-level constraints (*e.g.* , medium wrap can only be used for cylindrical objects) and as a result, the learned action model is immediately interpretable.

The contributions of this work are as follows: (1) We propose a novel method for extracting attribute information of the grasped objects by exploring spatial hand-object configurations; (2) We explore the mutual context of grasp types and object attributes to boost the recognition of both; (3) We propose a semantic action model based on grasp types and object attributes which achieves state-of-the-art recognition performance.

A. Related Work

Hand grasp has been studied for decades to better understand the use of human hands [23, 27, 26, 2, 13]. Grasp taxonomies have also been proposed to facilitate hand grasp analysis [5, 17, 10]. Cai et al. [3] first developed techniques to recognize hand grasp types in everyday hand manipulation tasks recorded with a wearable RGB camera and provided promising results with appearance-based features. Yang et al. [32] utilized a convolutional neural network to classify hand grasp types on unstructured public dataset and presented the usefulness of grasp types for predicting action intention. Saran et al. [28] used detected hand parts as intermediate representation to recognize fine-grained grasp types. However, the recognition performance is still not good enough for practical usage in real-world environments. In this paper we explore object contextual information to improve the grasp recognition performance.

Visual attributes (physical properties inferred from image appearance) are often used as intermediate representation for many applications, such as object recognition [7, 21, 30], facial verification [20], image retrieval and tagging [29, 24, 34]. Lampert et al. [21] performs object detection based on a human-specified high-level description of the target classes for which no training examples are available. The description consists of attributes like shape, color or even geographic information. Parikh and Graumn [24] explored the relative strength of attributes by learning a rank function for each attribute which can be used to generate richer textual descriptions. In this work, we extract visual attribute information from the manipulated object and use it as semantic information for modeling manipulation actions.

The relations between object attributes and hand grasps are widely studied for decades. It has been shown that humans use the same or similar grasp types for certain types of objects, and the shape of the object has a large influence on the applied grasp [18, 12]. Recently, Feix et al. [11] investigated the relationship between grasp types and object attributes in a large real-world human grasping dateset. However, behavioral studies in previous work do not scale to massive dataset. In this work, we use a Bayesian network to model the relations between grasp types and object attributes to boost the recognition of both.

Past researches on recognizing manipulation actions focused on using first-person vision since it provides an ideal viewing perspective for recording and analyzing hand-object interactions. Fathi et al. [8, 9] used appearance around the regions of hand-object interactions to recognize egocentric actions. The work in [25] has shown that recognizing handled objects helps to infer daily hand activities. In [14], hand appearance is combined with dense trajectories to recognize hand-object interactions. However, most of previous work are learning actions directly from image appearance, thus the action models learned are easily overfit to image appearance. There are small number of works which aim to reason beyond appearance models [31, 16, 32]. In [16] a hierarchical model is built to identify persuasive intent of images based on syntactical attributes, such as "smiling" and "waving hand". The work of [32] is very related to our work which seeks to infer action intent from hand grasp types. However, the action model in [32] is relatively simple with only three categories to be learned. In this work, we aim to model manipulation actions by jointly considering grasp types together with object attributes.

II. Our approach

We proposed an unified model to recognize grasp types, object attributes and actions from a single image. The approach is mainly composed by three components: 1) A visual recognition layer which recognizes hand grasp types and attributes of the manipulated objects. 2) A Bayesian network which models the mutual context of grasp types and object attributes to boost the recognition of both. 3) An action modeling layer which learns actions based on the belief distribution of grasp types and object attributes (output of the visual recognition layer).

A. Visual recognition of grasp types and object attributes

The visual recognition layer consists of two recognition modules, one for grasp types and the other for object attributes. Grasp types and object attributes are important for understanding hand manipulation. Grasp types determine the patterns of how a hand grasps an object, while object attributes

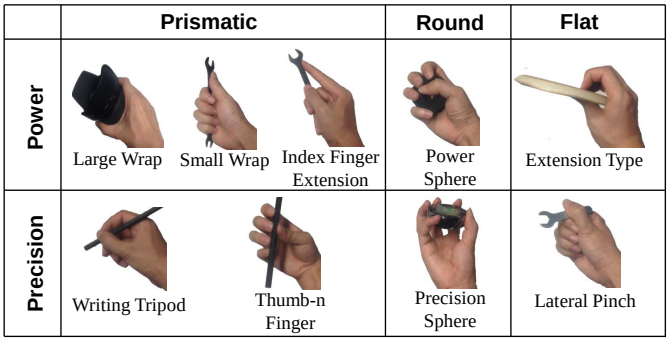

Fig. 2. The list of nine grasp types selected from [10], grouped by functionality (*Power* and *Precision*) and object shape (*Prismatic*, *Round* and *Flat*).

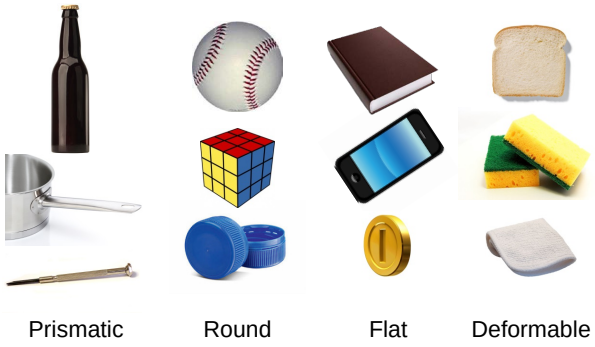

Fig. 3. Object examples with four different attributes: *Prismatic, Round, Flat,* and *Deformable*.

indicate possible hand motion of the interactions. Furthermore, grasp types together with object attributes provide consistent characterization of the manipulation actions.

1) Grasp types: Hand grasp types are important for understanding hand manipulation since they characterize how hands hold the objects during manipulation. A number of work have investigated the categorization of grasps into a discrete set of types [5][10] to facilitate the study of hand grasps. We train classifiers for recognizing nine different grasp types selected from a widely used grasp taxonomy proposed by Feix et al. [10]. The grasp types as shown in Figure 2 are selected to cover different standard classification criterion based on functionality [23], object shape, and finger articulation. We also abstract some grasp types in original taxonomy which are ambiguous from appearance into single grasp type (e.g. Thumb-n Finger). Furthermore, all the nine grasp types have a high frequency of daily usage based on the work of [2]. Thus the selected grasp types can be used to analyze large amount of manipulation tasks and meanwhile are possible for automatic recognition from image appearance.

Hand patches are needed to train grasp classifiers. Following [22], we train a multi-model hand detector composed by a collection of skin pixel classifiers which can adapt to different imaging conditions often faced by a wearable camera. For each test image, a pixel-level hand probability map is generated from the hand detector, and hand patches are then segmented with a bounding box. In detail, candidate hand regions are first selected by binarizing the probability map with a threshold. Regions under a certain area proportion are discarded and at most two regions are retained. Ellipse parameters (length of long/short axis, angle) are fitted to the hand region and the arm part is approximately removed by shortening the length of long axis to 1.5 times of the length of short axis. Then the remaining region is cropped with a bounding box. Linear SVM classifiers are trained for each grasp type using feature vectors extracted from hand patches. As the recognition output, belief distribution of grasp types (or posterior probability of grasp types given image evidence denoted as $P(G|f_G)$) as well as the predicted grasp type with highest probabilistic score are obtained.

Recognition of grasp types provide information about how the hands are holding the objects during manipulation. However, The grasp type alone is not enough to identify fine-grained actions without information from the manipulated objects. In the next section, we present the method to recognize object attributes.

2) Object attributes: Attribute information of the grasped object part is important for understanding hand manipulation since it indicates possible hand motion in hand-object interactions. For example, the body part of a bottle with long and thick shape indicates a motion of "holding", while the bottle cap with small and round shape probably indicates a motion of "screwing". While objects can be assessed by a wide range of attributes (shape, weight, surface smoothness, etc.), we only focus on attributes that are relevant to grasping and are also possible to be learned from image appearance. Figure 3 illustrates the attributes studied in this work, three of which are related to object shape and the fourth is related to object rigidity. We identify three different shape classes based on the criterion in Table I. The fourth attribute of *Deformable* identifies the object that deforms under normal grasping forces. Examples are a sponge or a rag. In this work, we aim to extract the above four object attributes from each grasped object part.

TABLE I
CLASSIFICATION CRITERION OF THREE SHAPE CLASSES. LENGTH OF OBJECT ALONG THREE OBJECT DIMENSIONS (MAJOR AXES OF THE OBJECT) ARE DENOTED AS A, B, AND C, WHERE $A \geq B \geq C$.

Shape classes	Object dimensions
Prismatic	$A > 2B$
Round	$B \leq A < 2B,\ C \leq A < 2C$
Flat	$B > 2C$

Similar to grasp type recognition, appearance-based features are extracted from object patches to train object attribute classifiers. However, object detection is a challenging task in computer vision, particularly unreliable when there are occlusions during manipulation. We observe that hand appearance provides important hint about the relative location and size of the grasped part of the object (not the whole object, and we will refer to "grasped part of the object" simply as "object"

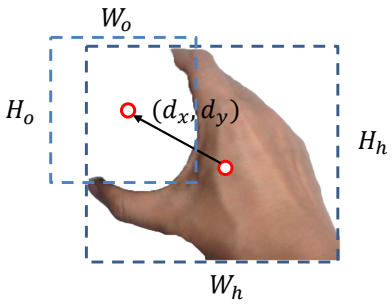

Fig. 4. Illustration of relative location and scale of the hand and the manipulated object.

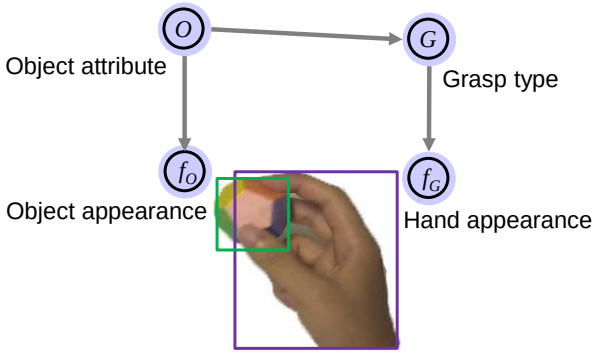

Fig. 5. A Bayesian network modeling the relationship between object attributes and grasp types.

in the following paper) from which attribute information is extracted. As illustrated in Figure 4, relative location (d_x, d_y) from the center of hand to the center of object is consistent to the hand orientation, and the object scale (W_o, H_o) is related to the size of hand opening. Therefore, we propose to train a target regressor for predicting the relative location and scale of the grasped object based on hand appearance. Specifically, we do regression on three quantities: normalized relative location of (N_x, N_y) and relative scale of N_s specified as follows:

$$\begin{cases} N_x = \dfrac{d_x}{W_h} \\ N_y = \dfrac{d_y}{H_h} \\ N_s = \sqrt{\dfrac{W_o \times H_o}{W_h \times H_h}} \end{cases} \quad (1)$$

Here are the steps of how to extract object attribute information: First, linear SVM regressors for object detection are pre-trained based on hand appearance features and manually annotated object bounding boxes. Note that when annotating object bounding box, we are not labeling the whole object but the grasped part of the object. Then, object patches are segmented with bounding boxes estimated based on the regressed quantities defined in Equation 1. Finally, linear SVM classifiers for object attribute classification are trained based on object appearance features extracted from segmented object patches and manually annotated attribute labels. As output, belief distribution of object attributes (or posterior probability of object attributes given image evidence denoted as $P(O|f_O)$) as well as the predicted attributes are obtained.

Visual recognition of grasp types and object attributes are challenging tasks as there are many occlusions during manipulation. In the next section, we present how to boost the recognition performance by mutual context.

B. Mutual context of grasp types and object attributes

There is strong causal relations between object attributes and grasp types. Object attributes such as geometric shape and rigidity have a large impact on the selection of grasp types. To grasp an object with thin prismatic shape (*e.g.* , the mug cup handle), grasp type of *Small Wrap* is often selected, while to grasp an object with small round shape (*e.g.* , a bottle cap), grasp type of *Precision Sphere* is preferred. On the other hand, knowing the grasp types used helps to infer the attributes of the grasped object. Therefore, mutual context between grasp types and object attributes can be explored that knowing the information of one side facilitates the recognition of the other.

We use a Bayesian Network to model the context information between grasp types and object attributes as illustrated in Figure 5. There is a directional connection from object attributes O to grasp types G, encoding the causal relation between object attributes of O and the grasp types of G. f_O and f_G denote the image evidence from the detected object regions and hand regions respectively. Based on this model, the posterior probability of object attributes and grasp types given the image evidence can be computed as:

$$\begin{aligned} P(O, G|f_O, f_G) &= \frac{P(O)P(G|O)P(f_O|O)P(f_G|G)}{P(f_O)P(f_G)} \\ &= \frac{P(G|O)P(f_O, O)P(f_G, G)}{P(f_O)P(f_G)P(G)} \\ &\propto P(G|O)P(G|f_G)P(O|f_O) \end{aligned} \quad (2)$$

Thus, we can jointly infer object attributes O^* and grasp types G^* by maximizing a posterior (MAP) as:

$$\begin{aligned} (O^*, G^*) &= \arg\max_{O,G} P(O, G|f_O, f_G) \\ &= \arg\max_{O,G} P(G|O)P(G|f_G)P(O|f_O) \end{aligned} \quad (3)$$

The optimal inference is obtained by searching the joint space of object attributes and grasp types that maximizes the multiplication of three components. The first component $P(G|O)$ is the conditional probability of grasp types given object attributes and has been learned in advance from occurrence frequencies of the training data. The last two components $P(G|f_G), P(O|f_O)$ are posterior probability of grasp types and object attributes given image evidence, and can be estimated from the probabilistic output of grasp classifiers and object attribute classifiers (belief distribution of grasp types and object attributes) respectively. Note that grasp classifiers and object attribute classifiers are learned from the training data as introduced in previous section II-A.

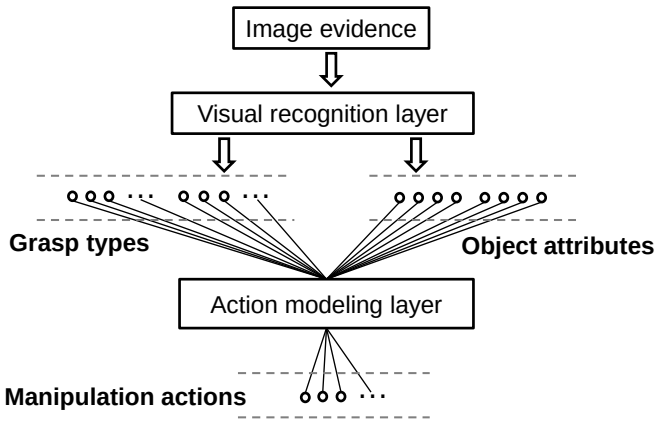

Fig. 6. Hierarchical action model based on belief distribution of grasp types and object attributes from the visual recognition layer.

C. Action modeling

Previous studies [23] showed that action functionality is an important factor that affects human grasp selection. Thus it is possible to infer action functionality from grasp types. In this work, we take a further step to model manipulation actions based on the grasp types of hands as well as the attributes of manipulated objects. Our hypothesis is that grasp types together with object attributes provide complementary information for characterizing the manipulation action.

We propose a semantic action model that builds on the semantic information of grasp types and object attributes. The diagram of the action model is shown in Figure 6. The hierarchical model separates the action modeling part from the low-level visual recognition part, thus the action learned is independent of image appearance which often changes under different scenes. The visual recognition layer is introduced in Section II-A which takes as input the image appearance and output belief distribution of grasp types and object attributes. At action modeling layer, a linear mapping function is learned for each action which models its semantic relationship with grasp types and object attributes denoted as:

$$P_{Action} = f(P_{Gl}, P_{Gr}, P_{Ol}, P_{Or}|\theta) \tag{4}$$

where P_{Action} is the probabilistic estimation of manipulation actions, P_{Gl}, P_{Gr} are belief distribution of grasp types for both hands, P_{Ol}, P_{Or} are belief distribution of object attributes, and θ is a set of parameters that measure the relationship between each action and its semantic components.

More specifically, for each image, the visual recognition layer is applied to extract semantic information represented by a 25-dimensional feature vector, of which 17 dimension is composed by belief distribution of grasp types for two hands (*Writing Tripod* is never used by the left hand) and 8 dimension is composed by belief distribution of object attributes of two grasped objects. Linear SVM classifiers are trained for different actions based on the obtained 25-dimensional feature vectors.

III. Experiments

In this section, we present four sets of results to validate different components of our approach: (1) grasp type recognition, (2) target regression and object attribute recognition, (3) improved recognition by mutual context of object attributes and grasp types, (4) action recognition.

We evaluate our approach on a public dataset (GTEA Gaze Dataset [9]) of daily activities recorded by a head-worn wearable camera. This dataset consists of 17 sequences of cooking activities performed by 14 different subjects. The action verb and object categories with beginning and ending frame are annotated. We also use another public dataset (GTEA Gaze+ Dataset [9]) to test the generality of action models. This dataset consists of seven cooking activities, each performed by 10 subjects. Similarly, action labels are provided. The main difference between these two datasets is that in the former dataset activities are performed near a table while in the second dataset activities are performed in a natural setting. The details of evaluation for each component are introduced in following sections.

A. Grasp type recognition

To train grasp classifiers for grasp type recognition from egocentric video, we annotate grasp types for 1000 hand images from GTEA Gaze Dataset. Previous work on grasp type recognition used Histogram of Oriented Gradient (HoG) [3] and Convolutional Neural Network (CNN) [32] for classifying different grasp types from monocular images. In this work we choose HoG as baseline feature and compare it with two different CNN-based features. Since the number of annotated grasp images is not sufficient for training CNN with large number of parameters, we perform fine-tuning to existing pre-trained CNN model and extract mid-layer features from it. In particular, we combine a large pre-trained CNN model proposed by Krizhevsky et al. [19] with domain-specific fine-tuning on our annotated hand images using the open source Caffe library [15]. We replace the original CNN's 1000-way classification layer with a randomly initialized 9-way classification layer (for the 9 grasp types) and perform stochastic gradient descent (SGD) training at a learning rate of 0.001. Feature vectors are extracted from two different layers (*CNN-pool5* and *CNN-fc6*) of the CNN separately. Compared to *CNN-pool5* which is the max-pooled output of the CNN's final convolutional layer, *CNN-fc6* adds one fully connected layer to *CNN-fc6*. Based on the extracted features, linear SVMs are trained for 9 grasp types. 5-fold cross-validation is used for evaluation.

TABLE II
Classification accuracy for nine grasp types on GTEA Gaze Dataset.

	HoG	CNN-pool5	CNN-fc6
Accuracy	50%	61.2%	56.9%

Grasp recognition performance of different features is

shown in Table II. Highest classification accuracy of of 61.2% is achieved by *CNN-pool5*. It can be seen that CNN-based feature outperforms hand-crafted feature HoG, also validated by the work of [32]. However, our work shows the feasibility of adapting pre-trained CNN model to grasp recognition problem with scarce training data.

B. Object attribute recognition

To train target regressors for predicting object location and scale, we annotated object bounding boxes for 1000 images with well detected hand patches from GTEA Gaze Dataset. The bounding box is annotated to include the object part being grasped. To train attribute classifiers, we also annotate attribute information for regions enclosed by annotated object bounding boxes. Linear SVM target regressors are trained based on annotated object bounding boxes and features extracted from hand patches. Linear SVM attribute classifiers are trained based on annotated object attributes and features extracted from annotated object patches. We use the libSVM library [4] for implementation. Since this is the first work by far as we know on recognizing object attributes from hand-object manipulation, we use same features as in Section III-A.

Table III shows quantitative results of target regression evaluated by Intersection of Union (IoU) which measures the overlap ratio of ground-truth object bounding box and the predicted object bounding box. The predicted object bounding box with equal width and height are determined based on the regressed quantities defined in Equation 1. *CNN-pool5* and *CNN-fc6* obtain similar performance but work much better than HoG. Figure 7 demonstrates qualitative results of target regression. It can be seen that the predicted object bounding boxes match well with ground-truth object bounding boxes, although the background is cluttered and objects are partially occluded by hands. The results indicate that it is possible to detect the grasped object parts simply from hand appearance.

TABLE III
QUANTITATIVE RESULTS OF TARGET REGRESSION EVALUATED BY INTERSECTION OF UNION (IOU).

	HoG	CNN-pool5	CNN-fc6
IoU	0.471	0.739	0.736

TABLE IV
PERFORMANCE OF ATTRIBUTE CLASSIFICATION ON GTEA GAZE DATASET.

Object Attribute	HoG	CNN-pool5	CNN-fc6
Prismatic	80.2%	87.9%	84.5%
Round	94.0%	94.0%	95.7%
Flat	81.0%	85.3%	87.1%
Deformable	88.8%	92.2%	91.4%
Combined	60.3%	72.4%	71.9%

Table IV shows the classification results for four binary object attributes. Accuracy is evaluated for four binary attributes separately as well as combined. When evaluating combined attributes, a prediction is considered as accurate if all the attributes are correctly classified. Accuracy of over 80% is achieved for all binary attributes and the advantage of CNN-based features over hand-crafted features is verified. For combined attributes, CNN-pool5 achieves best accuracy of 72.4% which means the percentage of cases that all binary features are correctly classified is over 72.4%. The results demonstrate the potential of learning physical properties of the manipulated objects from monocular images.

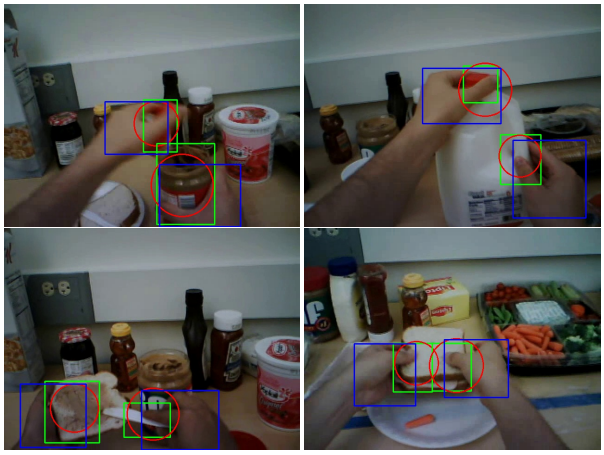

Fig. 7. Qualitative results of target regression. Blue and green bounding boxes show the detected hand regions and ground-truth object regions respectively. Red circles show the predicted object regions with center of circle indicating object location and radius indicating object scale.

C. Better recognition by mutual context

In this section, we show that the recognition of grasp types and object attributes can be improved by mutual context. We estimate the probability of grasp types conditioned on object attributes as prior information by occurrence frequencies from training data. Figure 8 shows the estimated conditional probability. It can be seen that different kinds of objects have very different distribution over applied grasp types. *Rigid-Prismatic* object such as a bottle is often held with *Large Wrap* or *Index Finger Extension*, while *Rigid-Round* object such as a bottle cap is often held with *Precision Sphere*.

We compare the recognition performance of with and without context information. For recognition without context information, grasp types and object attributes are inferred independently by simply selecting the category which outputs best score from classifiers. For recognition with context information, grasp types and object attributes are jointly inferred from Equation 3. Features of *CNN-pool5* are used in both two cases. The results in Table V and Table VI show that visual recognition of grasp types and object attributes are significantly improved by using context information. For grasp types, overall classification accuracy is improved by 12.9%. Performance of most grasp types are improved by object context, except for *Power Sphere* and *Precision Sphere*. We believe the performance deterioration of the two grasp types

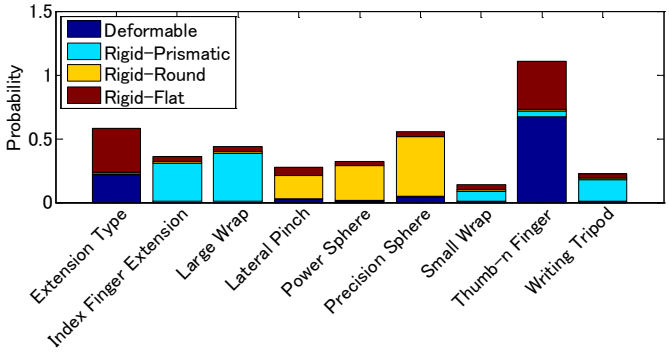

Fig. 8. Probability of grasp types given object attributes estimated by occurrence frequencies from training data.

is due to some false classification of the attribute *Round*. For object attributes, classification accuracy for combined attributes is improved by 9.5%. Experiment results strongly support the use of contextual information for improving visual recognition performance.

TABLE V
PERFORMANCE IMPROVEMENT FOR GRASP TYPE RECOGNITION BY MUTUAL CONTEXT (EVALUATED BY ACCURACY).

Grasp Category	CNN	CNN+Context
Extension Type	16.6%	20%
Index Finger Extension	66.6%	94.9%
Large Wrap	71.1%	81.8%
Lateral Pinch	87.5%	90.3%
Power Sphere	57.1%	33.3%
Precision Sphere	74.9%	66.6%
Small Wrap	52.6%	100%
Thumb-n Finger	55%	59%
Writing Tripod	73.3%	80%
Overall	61.2%	74.1%

TABLE VI
PERFORMANCE IMPROVEMENT FOR OBJECT ATTRIBUTE RECOGNITION BY MUTUAL CONTEXT (EVALUATED BY ACCURACY).

Object Attributes	CNN	CNN+Context
Prismatic	87.9%	88.8%
Round	94.0%	95.7%
Flat	85.3%	88.8%
Deformable	92.2%	92.2%
Combined	72.4%	81.9%

D. Action recognition

In this section, we demonstrate the effectiveness of modeling manipulation actions based on semantic information of grasp types and object attributes. The verb part of original action labels in GTEA Gaze Dataset are used as action labels in this work. For example, "Open a jam bottle"" and "Open a peanut bottle" are considered as the same action "Open". We focus on actions which require two-hand coordination. Seven action classes are learned in this work and are illustrated in Figure 9.

To evaluate the effectiveness of the proposed semantic action model, action recognition performance is compared with four baseline methods. In the first two baseline methods, action classifiers are trained based on appearance features extracted from two hand regions and two object regions. In particular, HoG features (HoG-4, "4" stands for the feature concatenation from four regions) and CNN-based features (CNN-pool5-4, "4" has same meaning as in HoG-4) are extracted. Since gaze information is not available in our proposed system, we didnt directly compare with the features used in [9]. Instead, we extract appearance features (HoG and CNN) from the detected hand and object regions which are used as approximation of the region of gaze. In the third and forth baseline method, grasp types (GpT) and object attributes (OA) are used as intermediate features for training action classifiers respectively. Note that GpT is also used in the method of [32] for recognition of three different action intentions. In the proposed method, grasp types and object attributes are jointly used (GpT+OA). In practice, linear SVMs are used for training action classifiers. Performance is evaluated using 5-fold cross validation based on labeled images from GTEA Gaze Dataset.

The action recognition performance and feature dimension of different methods are shown in Table VII. The proposed GpT+OA outperforms GpT and OA, which indicates the combination of grasp types and object attributes provides better action modeling than using the two components separately. Accuracy of 79.3% achieved by GpT+OA is also comparative to state-of-the-art CNN-based feature representation. Note that our method uses much lower feature dimension of 25 compared to 36864 used in CNN-pool5-4.

TABLE VII
PERFORMANCE COMPARISON FOR RECOGNIZING SEVEN ACTION CLASSES ON GTEA GAZE DATASET.

	Accuracy	Dimension
HoG-4	65.5%	11664
CNN-pool5-4	81.0%	36864
GpT [32]	69.0%	17
OA	70.7%	8
GpT+OA	79.3%	25

To demonstrate the correlation between each action and its semantic components of grasp types and object attributes, we compute model parameters from support vectors learned by each linear SVM classifier. Model parameters indicate the correlation between action and its 25 semantic components. Visualization of model parameters is illustrated in Figure 9. It can be seen that each action has strong correlation to different grasp types and object attributes.

To compare the generality of the proposed semantic action model with appearance-based action model, we performed cross-dataset action recognition by training and testing on two different datasets. While all the training procedure is done on GTEA Gaze Dataset, we test action recognition on

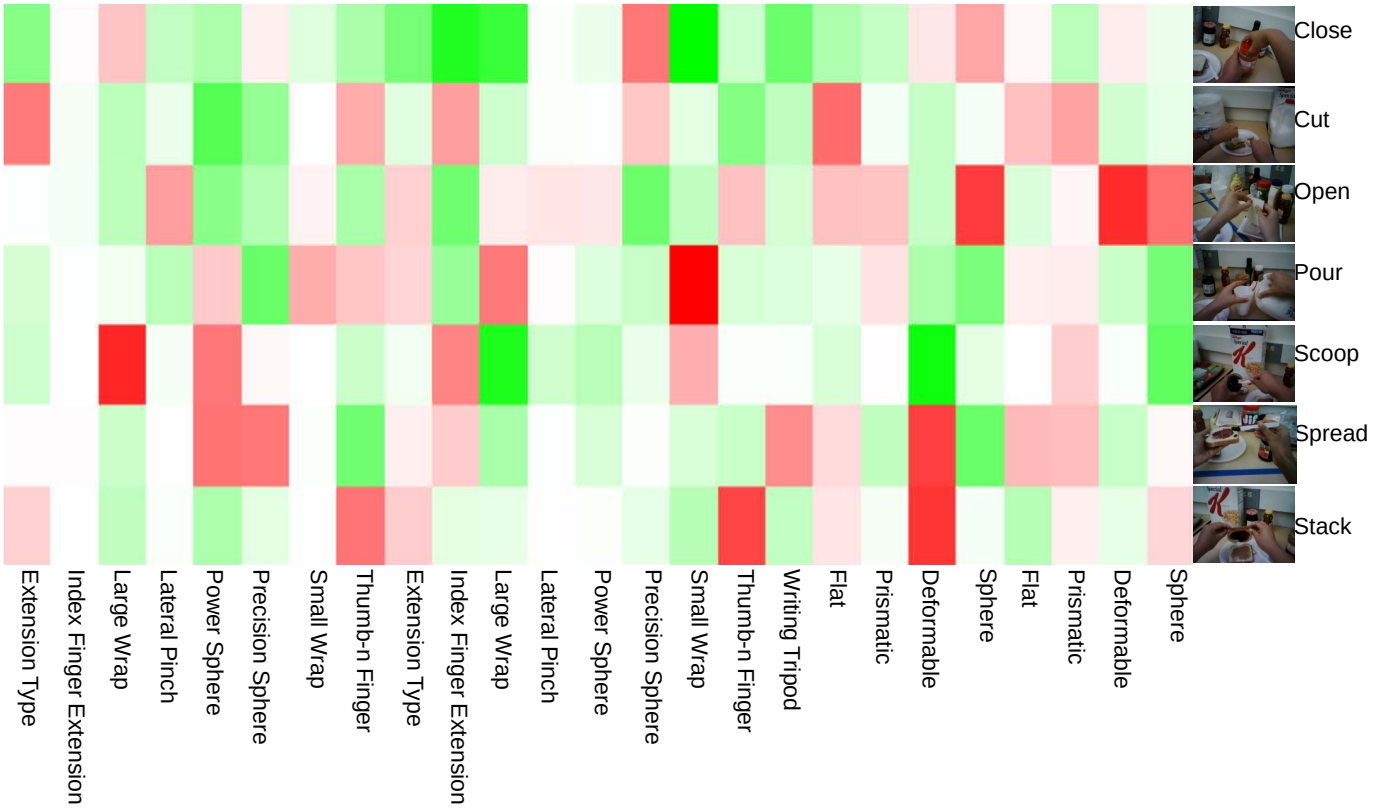

Fig. 9. Visualization of model parameters for seven action classes. The saturation of red color indicates positive correlation while the saturation of green color indicates negative correlation. White color indicates no correlation.

GTEA Gaze+ Dataset recorded in different environments. We selected 100 images for each action category and a total of 700 images from GTEA Gaze+ Dataset are used for testing. Appearance-based model is trained based on CNN-pool5-4, while the proposed hierarchical model is trained based on GpT+OA. Classification accuracy is shown in Table VIII. The proposed semantic model outperforms the appearance-based model by nearly 10%. The experiment verifies the generality of the proposed method which takes into account semantic constraints of grasp types and object attributes, and therefore is more robust to overfitting.

TABLE VIII
GENERALITY EVALUATION OF ACTION MODELS BY TRAINING ON GTEA GAZE DATASET AND TESTING ON GTEA GAZE+ DATASET.

	CNN-pool5-4	GpT+OA
Accuracy	40.7%	50.4%

IV. CONCLUSION

We proposed an unified model for understanding hand-object manipulation with a wearable camera. From a single image, grasp types are recognized from the detected hand regions and attribute information are extracted from the detected object parts. Furthermore, mutual context is explored to boost the recognition of both grasp types and object attributes. Finally, actions are recognized based on belief distribution of grasp types and object attributes.

Experiments were conducted to verify our proposal: (1) We achieved average accuracy of 61.2% for grasp type recognition and 72.4% for object attribute classification. (2) By mutual context, recognition performance is improved by 12.9% for grasp types and by 9.5% for object attributes. (3) Our proposed semantic action model achieved classification accuracy of 79.3% which is comparative to state-of-the-art feature representation with much lower feature dimension. Evaluation results for model generality support our hypothesis that grasp types and object attributes contain consistent information for characterizing different actions. We believe our work of studying the relationship between grasp types, object attributes and actions points out an important direction for understanding hand manipulation from vision.

In future work, we wish to extend the current single-image framework to temporal dimension in order to study the temporal evolution of hand grasping dynamics in different manipulation tasks. Another direction we wish to step on is to extract more complex object attributes (such as 3D shape) in studying grasp-object relationship based on depth information by using wearable RGB-D camera.

ACKNOWLEDGMENTS

This research was funded in part by the JST CREST grant.

References

[1] I.M. Bullock, T. Feix, and A.M. Dollar. Finding small, versatile sets of human grasps to span common objects. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 1068–1075. IEEE, 2013.

[2] I.M. Bullock, J.Z. Zheng, S. Rosa, C. Guertler, and A.M. Dollar. Grasp frequency and usage in daily household and machine shop tasks. *Haptics, IEEE Transactions on*, 6(3):296–308, 2013.

[3] M. Cai, K.M. Kitani, and Y. Sato. A scalable approach for understanding the visual structures of hand grasps. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1360–1366. IEEE, 2015.

[4] C. Chang and C. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[5] M.R. Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *Robotics and Automation, IEEE Transactions on*, 5(3):269–279, 1989.

[6] A.M. Dollar. Classifying human hand use and the activities of daily living. In *The Human Hand as an Inspiration for Robot Hand Development*, pages 201–216. Springer, 2014.

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

[8] A. Fathi, A. Farhadi, and J.M. Rehg. Understanding egocentric activities. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 407–414. IEEE, 2011.

[9] A. Fathi, Y. Li, and J.M. Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012*, pages 314–327. Springer, 2012.

[10] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, and D. Kragic. A comprehensive grasp taxonomy. In *Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, pages 2–3, 2009.

[11] T. Feix, I.M. Bullock, and A.M. Dollar. Analysis of human grasping behavior: Object characteristics and grasp type. *Haptics, IEEE Transactions on*, 7(3):311–323, 2014.

[12] R. Gilster, C. Hesse, and H. Deubel. Contact points during multidigit grasping of geometric objects. *Experimental brain research*, 217(1):137–151, 2012.

[13] D. Huang, M. Ma, W. Ma, and K.M. Kitani. How do we use our hands? discovering a diverse set of common grasps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 666–675, 2015.

[14] T. Ishihara, K.M. Kitani, W. Ma, H. Takagi, and C. Asakawa. Recognizing hand-object interactions in wearable camera videos. In *IEEE International Conference on Image Processing (ICIP)*, 2015.

[15] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[16] J. Joo, W. Li, F. F Steen, and S. Zhu. Visual persuasion: Inferring communicative intents of images. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 216–223. IEEE, 2014.

[17] S. Kang and K. Ikeuchi. Toward automatic robot instruction from perception-recognizing a grasp from observation. *Robotics and Automation, IEEE Transactions on*, 9 (4):432–443, 1993.

[18] R. L Klatzky, B. McCloskey, S. Doherty, J. Pellegrino, and T. Smith. Knowledge about hand shaping and knowledge about objects. *Journal of motor behavior*, 19(2):187–213, 1987.

[19] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[20] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[21] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

[22] C. Li and K.M. Kitani. Pixel-level hand detection in ego-centric videos. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3570–3577. IEEE, 2013.

[23] J.R. Napier. The prehensile movements of the human hand. *Journal of bone and Joint surgery*, 38(4):902–913, 1956.

[24] D. Parikh and K. Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.

[25] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2847–2854. IEEE, 2012.

[26] J. Romero, H. Kjellström, C.H. Ek, and D. Kragic. Non-parametric hand pose estimation with object context. *Image and Vision Computing*, 31(8):555–564, 2013.

[27] M. Santello, M. Flanders, and J.F. Soechting. Postural hand synergies for tool use. *The Journal of Neuroscience*, 18(23):10105–10115, 1998.

[28] A. Saran, D. Teney, and K.M. Kitani. Hand parsing for fine-grained recognition of human grasps in monocular images. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015.

[29] B. Siddiquie, R.S. Feris, and L.S. Davis. Image ranking and retrieval based on multi-attribute queries. In *Com-*

puter Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 801–808. IEEE, 2011.

[30] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M.B. Blaschko, and D. Weiss. Understanding objects in detail with fine-grained attributes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3622–3629. IEEE, 2014.

[31] Y. Yang, C. Fermuller, and Y. Aloimonos. Detection of manipulation action consequences (mac). In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2563–2570. IEEE, 2013.

[32] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos. Grasp type revisited: A modern perspective of a classical feature for vision. In *Computer Vision and Pattern Recognition, 2015 IEEE Conference on*, 2015.

[33] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[34] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1637–1644. IEEE, 2014.

[35] J.Z. Zheng, S. De La Rosa, and A.M. Dollar. An investigation of grasp type and frequency in daily household and machine shop tasks. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 4169–4175. IEEE, 2011.