Object Purpose Based Grasping

Song Cao, Jijie Zhao

Abstract—**Objects often have multiple purposes, and the way humans grasp a certain object may vary based on the different intended purposes. To enable robots to acquire similar ability, instead of inferring grasping point for objects only based on their appearance, a new method, object purpose based grasping point inference, is proposed. First, shape context descriptor and SVMs are applied for detecting purposes of objects. Then generalized linear model and multi-scale vision features are used to predict objects' grasping point based on their appearance. Experimental results for both purpose detection and grasping point inference are provided, which show that our method gives reasonably good performance in terms of accuracy in prediction.**

I. Introduction

THE notion of affordance was first introduced by J.J. Gibson [3] in 1979 to characterize the properties or possibilities of object to interact with other objects within the environment. An important part of affordance is the perception of objects. For example, a hammer, to a human, may be perceived as an object with a heavy end and a light end. Based on its appearance to human, a hammer's affordance is hammering when grasped at the light end of it. In the past few years, affordance has been an interesting topic in various fields including perceptual psychology, cognitive psychology, human-computer interaction (HCI), and robotics. Although the meaning of affordance may vary in different areas, the basic concerns of object affordance mainly involve the object appearance, shape, texture, and object contexts, e.g., objects can be used differently in different environment.

In robotics, much work has been done in robot manipulation and robot interaction with environment, for example, the handling of dishwashers, doors, books, etc. While robot manipulation spans areas such as object detection, localization, and action selection, grasping point is an important problem when we consider the object affordance. This is because first, certain object requires certain grasping point for use, for example, we can only grasp the light side of a hammer when we want to do hammering. Second, different grasping points can result in different usage for the same object, for example, a screwdriver can be used as a screwdriver when we hold the thick side; however, it may also be used as a hammer when held by the thin side. The fact that different grasping points imply different purposes impels us to explore the use of purposes in grasping point inference.

Past work mainly focused on the object and environment model building and designated manipulation tasks. A. Saxena [1] presented the problem of grasping novel objects that a robot perceives for the first time in vision. Instead of building 3D models of objects, they use machine learning algorithms to directly predict the grasping point using images of objects. Inspired by their work, we further introduce object purposes to novel object grasping problem. We aim to build a generalized learning system that can predict the objects' purposes and corresponding grasping points for each purpose.

Our motivation lies in two aspects: First, different purposes lead to different grasping points. In the case that an object has multiple purposes, predicting the grasping points based on each of the purposes would give us multiple grasping points. So users are able to choose from these grasping points according to their needs. Second, when faced with novel objects, even though we cannot accurately classify the object into categories, we are still able to infer their purposes based on their vision features and grasping points for each of their purposes.

In our project, we propose a two-phase learning algorithm for object purpose based grasping. In the first phase, we detect object purposes using machine learning algorithms with shape context descriptors [5]. In the second phase, for each object purpose detected, we use probabilistic model to infer the corresponding grasping point. Figure 1 shows the flow chart of our system. In the second part, some related work is introduced. Then in the second and third part, we'll introduce our methods for purpose detection and purpose based grasping point inference respectively, followed by experimental results in the fourth part.

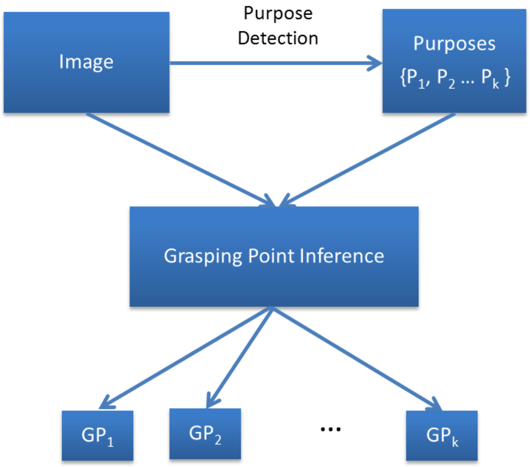

Fig. 1. Flow chart of our system.

II. Related Work

S. Belongie *et al.*[5] proposed shape context feature, which uses the position relationships between points to represent the shape of image. In [5], they have applied shape context in shape matching and object recognition work. Here we use shape context as a feature to help us classify the object purposes.

For classification, various multi-class classification algorithms based on Support Vector Machine (SVM) are provided [6]. Among them, OvA SVM classifier is used in our project. The OvA SVM classifier method constructs *N* SVM classifiers in an *N*-classes classification problem.

A. Saxena *et al.*[1, 2] provided a learning algorithm to detect the grasping point for novel object using vision. The main approach of their work is to use supervised learning to find set of grasping points, and then calculate the 3D location of them. Our project is mainly based on A. Saxena's work, and part of our data set is from theirs.

B. Ridge *et al.*[4] proposed an affordance learning algorithm which uses mapping from feature space to affordance space. They also used SVM training to determine the affordance class given a feature vector. This work inspires us with the relationship between affordance and object handling.

III. Purpose Detection

For purpose detection problem, we intend to classify objects into different purpose categories. As one object could have multiple purposes, we use "OvA" (one against all) SVMs to perform the classification [6].

By observation, there exists a logical connection between objects' shapes and their purposes. For example, if an object has an uneven distribution of thickness along its axis in the way that is similar to a hammer, then it's mostly likely that it has the purpose of hammering. Under this assumption, we use shape information as the descriptor for purpose detection. More specifically, shape context descriptor [5], which has been applied to shape matching and object recognition, is used here to concisely and representatively describe objects' shapes. Then, "OvA" (one against all) SVMs are trained to determine purposes of any given object. A flow chart describing this process is shown in Figure 2.

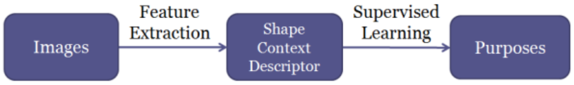

Fig. 2. Process of purpose detection.

A. *Shape Context Descriptor*

Different from the problem setting of shape matching and object recognition, where the descriptor of each point on the edge is important distinguishing feature, our problem doesn't necessarily require the computation of the descriptor for each point, since the objects are the set of entities that needs to be classified, descriptive details of each point in every object would not be efficient in terms of representing the overall shape of the object. Instead, we compute one descriptor for each object, using the following steps.

First, we perform canny edge detection algorithm [7] on input images.

Second, we randomly sample some certain number of points from this set of points on the edges detected, on which further computations are based on these sample points. Design choices must be made here to ensure the sampling will reduce the computation complexity while maintaining the shape information contained in the positions of sampled points.

Third, we compute the centroid of the sample points. The step gives us the point of origin of a polar coordinate system which is utilized later for embedding orientation invariance into our shape context descriptor.

Fourth, the vector from centroid to the point with largest distance from centroid is used as the polar axis of the polar coordinate system, with which we could finally compute the polar coordinates of each sample points.

Fifth, we construct the shape context descriptor of the centroid using these sample points. Shape context is essentially a log-polar histogram of the polar coordinates of the sample points. In this project, we have used 5 and 12 bins for $\log r$ and θ, respectively. Thus, we have in total 60-dimension vector as our descriptor. In each dimension, the value represents the proportion of sampled points in the corresponding bin of $\log r$ and θ.

B. *One-against-all (OvA) SVMs*

The basic idea is to build *N* two-class SVM classifiers, where N is the number of classes. When the i-th SVM is being trained, we label the samples in the i-th class to be positive, and all the rest samples to be negative. In the testing phrase, we run the example in each SVM classifiers to separately determine the belonging relationship to each class.

IV. Grasp Point Inference

Based on the purpose detected, every object may have different ways to grasp. We use different categories of objects with the same kind of purpose to build a probabilistic model used for grasping point inference for given objects and their corresponding set of purposes.

A. *Vision Feature Extraction*

A. Saxena [1] has used image-based features to infer grasping points of novel objects. In our work, similar methods are applied to the problem of finding grasping points for novel objects. Specifically, 51-dimension vision features are computed with multiple scales to deal with scale differences among real images of objects.

B. *Purpose Based Grasping Point Inference*

What's specifically challenging in our problem is that it's highly probable that many categories of objects share a common purpose. Therefore, it is critical to design an inference system that is able to extract the common characteristics of objects that share a common purpose. In other words, the relationship between an object's appearance (vision features) and its purpose must be accurately extracted.

To accomplish this goal, an underlying balance of effectiveness of finding grasping points of any particular category of objects and the ability of dealing with different categories of objects with common purposes must be achieved, as shown in Figure 3. For the second goal, individual inference models are generated for each purpose, so that relevant inference models could detect novel objects with potential similarity in terms of purpose based on their appearance. As for the effectiveness mentioned above, certain parameters involving weights on different categories of objects and portion of subsets used in training phase must be adjusted based on the relationship between their categories and purposes. More specifically, for a particular purpose, it's reasonable to differentiate the objects that belong to the categories with the same intrinsic designed purpose and the ones that do not in the training phase. Experiments in the following section show that the system demonstrates both desired properties discussed above.

After proper selection of training datasets, generalized linear models are learned based on them. For each purpose, one such model is learned and then specifically applied to those images classified into this purpose category. Therefore, for each purpose of an object, a specific grasping strategy will be generated.

To reduce the computation cost while maintaining the quality of our inference, we divide each image into 64 by 48 grids. For each grid, a 51-dimensional vision feature is calculated. Therefore, the original dimension of linear model is significantly reduced, thus the grasping point inference could be done much faster.

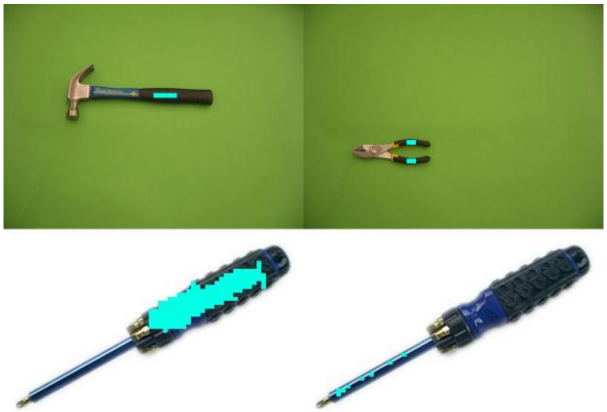

Fig. 3. Two Goals of object purpose based grasping: the upper 2 images demonstrate the goal of effectively infer grasping points of objects in any category; the lower 2 images demonstrate the goal of being able to infer different grasping points based on different purposes of the same object: the grasping point on the left is of a screwdriver used as a screwdriver, while the grasping point on the right is of a screwdriver used as a hammer.

V. Experiments

A. Data Description

Images of objects in 7 categories are used in our experiment. An overview of image data used is shown in Table 1. And a few samples of data images are shown in Figure 4. Note: data sets are acquired from Internet and [1].

For each image in our data set, its correct grasping points for each purpose are also marked and later used as training data and ground truth for calculating accuracy. For each grid in the image (mentioned in the last part), we decide whether it's grasping point or not by its similarity to the corresponding marking colors. Some sample images are shown in Figure 5.

TABLE I
Description of Experimental Data

Category	Number of Images	Set of Purposes
Hammers	20	Hammers
Screwdrivers	20	Screwdrivers, Hammers
Scissors	16	Scissors
Forks	21	Forks
Pliers	14	Pliers
Rulers	21	Rulers, Knives
Knives	16	Knives

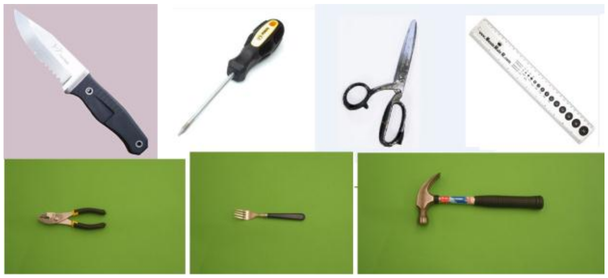

Fig. 4. Samples of our data set.

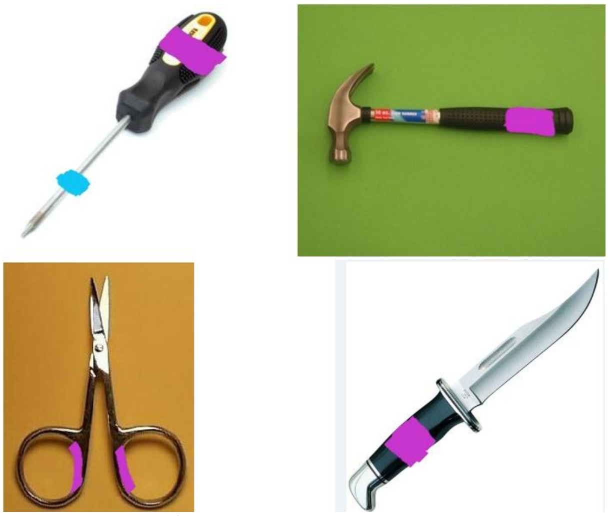

Fig. 5. Samples of marked images. For multi-purpose objects like screwdriver, all the grasping points for different purposes are marked with different colors.

B. Purpose Detection

Images in each category are divided into 2 parts, one part used as training data, and the other used as testing data. After training the SVMs with training data, we use them to predict the purposes of objects in the test data and measure the accuracy. By changing the subset of images selected as training data and testing data, our system is tested in 4 rounds. For each round, the testing data and training data are of roughly the same amount and chosen randomly from our dataset.

The results of our purpose detection experiments are shown in Table 2. The average error rate of purpose detection is 13.83%.

TABLE II
ERROR RATE OF PURPOSE DETECTION (%)

Category	Round 1	Round 2	Round 3	Error Rate 4	Average
Hammers	22.22	26.98	28.57	17.46	23.81
Screwdrivers	22.22	15.87	19.05	15.87	18.25
Scissors	6.35	6.35	6.35	12.70	7.94
Forks	6.35	9.52	11.11	3.17	7.54
Pliers	12.70	14.29	17.46	6.35	12.70
Rulers	12.70	7.94	9.52	12.70	10.72
Knives	14.29	19.05	17.46	12.70	15.88
Average	13.83	14.29	15.65	11.56	13.83

C. Grasping Point Inference

For evaluating the performance of our grasping point inference algorithm, we also divide images in each category are divided into training data and testing data. Then for each testing image, we calculate the false positive rate and false negative rate when compared to the ground truth.

After the training phase, each purpose has an individual grasping point inference model. Thus, for each testing image, we calculate one grasping point for each of its purposes. The grasping point is essentially a map of probabilities of every point to be the grasping point, calculated by normalizing the output of generalized linear model. A few examples are shown in Figure 6.

With the probabilistic map, we threshold the values on each grid to decide whether it is the grasping point or not. Then a comparison to the ground truth is made to calculate false positive rate and false negative rate. The experimental results of FP rates and FN rates are shown in Table 3.

As could be seen from the results, the average accuracy of our inference system is reasonably high. The average FP rate is 11.00% and average FN rate is 10.47%.

VI. CONCLUSIONS

A new perspective of grasping point inference is proposed in our project, in which grasping point of objects are not solely determined by their appearance and type information, but also by their purposes. Depending on different purposes of an object, there might be multiple ways of grasping it, which is not possible to deal with without considering object purpose based grasping point inference.

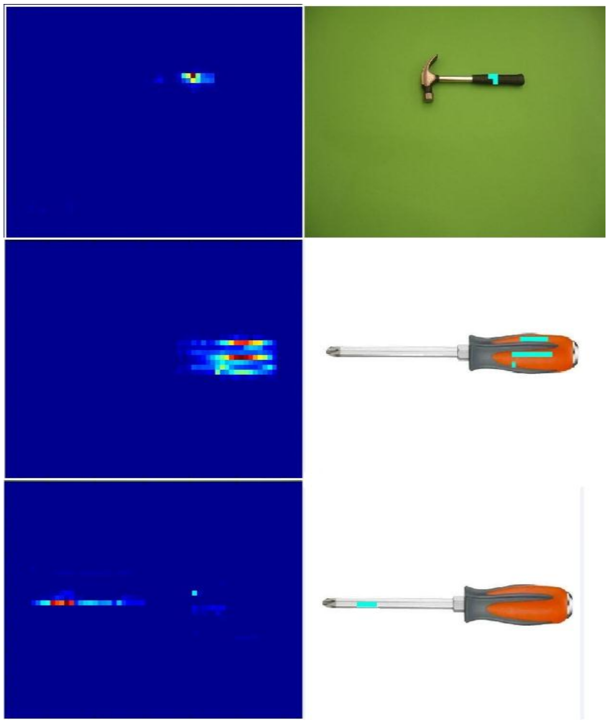

Fig. 6. Examples of grasping point inference. The 2 images on the top are the grasping point map and the corresponding original image of a hammer. The middle 2 images are similar images of a screwdriver used as a screwdriver. The 2 images at the bottom are similar images of a screwdriver used as a **hammer**.

TABLE III
FP AND FN RATES OF GRASPING POINT INFERENCE (%)

Category	Average FP Rate	Average FN Rate
Hammers	24.98	22.59
Screwdrivers	3.17	4.76
Scissors	0.66	12.11
Forks	4.76	7.09
Pliers	41.43	15.90
Rulers	1.59	5.02
Knives	0.44	5.86
Average	11.00	10.47

Vision feature extraction is used in our project to generate descriptive and compact representation of objects' appearance. And supervised learning algorithms such as SVMs and generalized linear model are used to model the relationship between objects' appearance and the underlying purposes, and further, the connection between these two elements and the grasping points.

Future work includes improving both purpose detection and grasping point inference. It's possible there's other ways to incorporate objects' purposes into decision on the grasping points, and there're better vision features to describe the appearance of objects more concisely and representatively. However, the incorporation of purpose information into decision-making process of grasping is a promising direction in improving the intelligence of current robots.

ACKNOWLEDGMENT

We would like to thank Prof. Ashutosh Saxena for his support and encouragement and Sajeev Krishnan and Avtar Khalsa for their technical assistance.

References

[1] A. Saxena, J. Driemeyer, J. Kearns, and A. Y. Ng, “Robotic grasping of novel objects,” In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems Conference (NIPS)*, Vancouver, Canada, 2006.

[2] A. Saxena, J. Driemeyer, Justin Kearns, C. Osondu, Andrew Y. Ng. “Robot grasping of novel objects,” In *10th International Symposium on Experimental Robotics (ISER)*, 2006.

[3] J.J. Gibson. *The Ecological Approach to Visual Perception*, Lawrence Erlbaum Associates, 1986.

[4] B. Ridge, D. Skocaj, A. Leonardis, “Unsupervised learning of basic object affordances from object properties,” *Computer Vision Winter Workshop*, Eibiswald, Austria, February 2009.

[5] Serge Belongie, Jitendra Malik and Jan Puzicha, “Shape Matching and Object Recognition Using Shape Contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, April 2002 (vol. 24 no. 4), pp. 509-522.

[6] Gjorgji Madzarov, Dejan Gjorgjevikj and Ivan Chorbev. “A Multi-class SVM Classifier Utilizing Binary Decision Tree,” *Informatica* 33 (2009) 233-241.

[7] J. Canny, “A Computational Approach To Edge Detection”, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 8(6):679–714, 1986.