Multi-Modal RGBD Sensors for Object Grasping and Manipulation

Tarek El-Gaaly, Marwan Torki, Ahmed Elgammal and Maneesh Singh

Extended Abstract

***Abstract*— RGBD sensors, such as the Microsoft Xbox Kinect [1] are types of multi-modal perceptual sensors that have appeared in recent years. RGBD sensors have become standard perceptual tools for robots as they provide a unique multi-modal approach to perception. A vital pre-cursing challenge in object grasping and manipulation is object pose recognition. A robot must identify the pose (*i.e.* orientation) of an object in order to perform grasping/manipulation tasks accurately. In this work we focus on combining multi-modal RGBD data to reduce uncertainty and hence solve the problem of object pose recognition more accurately. We experiment on an RGBD dataset and show that our approach has a significant improvement of more than 20% over state-of-the-art.**

I. INTRODUCTION

A key pre-cursing challenge in robotic grasping and manipulation is identifying an object and its pose (i.e. orientation). Object pose estimation helps answer critical questions, *e.g. where is the handle of the mug?*, that a robot must answer before attempting to interact with the environment. The more successful a robot can estimate the pose of an object, the more capable it will be at identifying how it should grasp the object and identifying the object's orientation during manipulation (closing the feedback loop).

Accurate object manipulation requires a robot to overcome uncertainties in the environment. One method of doing this is by combining multiple modes of data. In this work we present a method of combining modes of information to deal with uncertainty. We take an in-depth look at a specific instance of multi-modal data: RGBD, and see how to solve the problem of object pose recognition.

The most relevant work in the area of pose (along with category and instance) recognition using synchronized multimodal photometric and depth data (*i.e.* RGBD) is by Lai et al. [5], [4]. In [5], the authors show significant performance for category and instance recognition. Despite this, performance of object pose recognition is less significant. The main reasons for this is that they use a classification strategy (resulting in coarse pose estimates) and do not fully utilize the information present in the spatial distribution of features on object surfaces.

In our previous work [3], we have addressed these issues. Since the pose space is continuous, we developed a *regression* framework for estimating object pose from color images. This work extends this approach by making the following contributions: 1. We use a multi-kernel learning (MKL) approach to extend our framework for pose estimation to use multimodal RGBD data. 2. Extensive experiments on a large available RGBD dataset [4]. We demonstrate significant performance improvements over best published results.

II. RGBD Multi-Kernel Regression/Learning

Building on [3], we perform regression using multiple similarity kernels. Two kernels are built from training images; one for local Geometric Blur feature similarity K and one for global depth HOG feature similarity G. K is built over a Laplacian Eigenmap embedding which enforces inter-image visually similar feature points and intra-image spatially close feature points to lie in close proximity. It also enforces a manifold constraint on the pose labels v (pose = yaw angle of object). In the case of G, this amounts to an image embedding in a lower dimensional space without the local feature constraints, but also with a manifold constraint keeping global feature points from images at similar viewpoints close to each other. The kernels are used to perform regression over training image poses by solving a system of linear equations to recover coefficients b_j and c_j.

$$v = \hat{g}(X) = \sum_j (b_j K(X, X^j) + c_j G(X, X^j)) \tag{1}$$

An out-of-sample problem is then solved for each kernel to find the similarity between training and testing images. This kernel similarity is then used to infer the poses of test images using eq 1. Refer to [3] for details. This constitutes Multi-Kernel Regression (MKR). MKL on the other hand performs a weighted sum of the similarity kernels [2]. To obtain the weights, we use heuristics: F-measure and M-measure [6]. Once the kernels are summed up, the pose can be estimated using eq 1 but with one similarity kernel in this case.

Fig. 1. Pose estimates with RGB and depth images. Color code: ground-truth pose (red), visual only regression (blue), dual-modality using visual + depth regression (green). Zoom in for best view

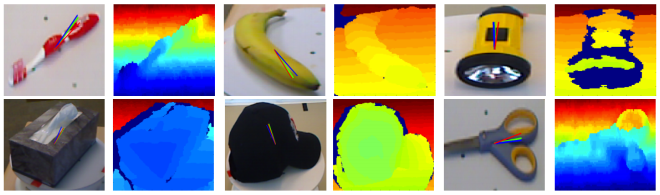

TABLE I
Pose Recognition Performance on RGBD-dataset

Method	Median Pose	Avg Pose	St. Dev.
dHOG (Depth only)	51.25	50.62	5.16
gsGB (RGB - visual only) [3]	77.8	72.06	14.39
MKR (RGB+D)	**85.0**	**74.58**	13.69
MKL-F (RGB+D using F-measure)	**86.3**	**75.50**	12.71
MKL-M (RGB+D using M-measure)	**86.7**	**74.76**	14.21
[5]	71.40	56.80	-

The results of our approaches and a comparison with [5] are shown in table I. The first two rows are single-modalities *i.e.* using single similarity kernels. The three bottom rows show the improvement using both visual and depth modes *i.e.* combining multiple kernels through MKR and MKL.

III. Conclusion

In this work we built a framework for MKR and MKL which enables us to utilize multiple modalities of perception. We have shown improved results when combining multiple modalities (RGB+D). This work shows the advantages of using Machine Learning to overcome uncertainty and the potential to improve the accuracy of object grasping and manipulation tasks by using inexpensive multi-modal sensors, such as the Xbox Kinect [1].

REFERENCES

[1] Microsoft Kinect. http://www.xbox.com/en-us/kinect.

[2] Alpaydin and Ethem. Introduction to Machine Learning. MIT Press, Cambridge, MA, 2. edition, 2010.

[3] M. Torki and A. Elgammal. Regression from local features for viewpoint and pose estimation. In ICCV, 2011.

[4] K.Lai, L.Bo, X.Ren and D.Fox. A large-scale hierarchical multi-view rgbd object dataset. In ICRA, 2011.

[5] K. Lai, L. Bo, X. Ren, and D. Fox. A scalable tree-based approach for joint object and pose recognition. In AAAI, 2011.

[6] S. Qiu and T. Lane. A framework for multiple kernel support vector regression and its applications to sirna efficacy prediction. IEEE/ACM Trans. Comput. Biology Bioinform, 2009.