RGBD Object Pose Recognition using Local-Global Multi-Kernel Regression

Tarek El-Gaaly, Marwan Torki, Ahmed Elgammal
Rutgers University
{*tgaaly,mtorki,elgammal*}*@cs.rutgers.edu*

Maneesh Singh
Siemens Corporate Research
maneesh.singh@siemens.com

Abstract

The advent of inexpensive depth augmented color (RGBD) sensors has brought about a large advancement in the perceptual capability of vision systems and mobile robots. Challenging vision problems like object category, instance and pose recognition have all benefited from this recent technological advancement. In this paper we address the challenging problem of pose recognition using simultaneous color and depth information. For this purpose, we extend a state-of-the-art regression framework by using a multi-kernel approach to incorporate depth information to perform more effective pose recognition on table-top objects. We do extensive experiments on a large publicly available dataset to validate our approach. We show significant performance improvements (more than 20%) over published results.

1. Introduction

The availability of inexpensive (color+) depth sensors (e.g. Microsoft Xbox Kinect [1]) has brought about a large advancement in 3D perception over the last couple of years. This has had an immediate impact in mobile robotics where robots are being fitted with Kinects to enhance their perceptive capability. A challenging problem for the robot is to automatically identify the object (category and instance) as well as its pose. We feel that in this space, while the problems of object category and instance recognition have received a lot of attention, pose recognition has not been addressed adequately. Pose recognition is an important problem than can help the robot answer important questions, for example, *what is the arrangement and orientation of various objects or people?, Where is the handle of the mug?*...etc. Addressing these questions is important for a variety of tasks like scene understanding, activity recognition, object manipulation in mobile robotics as well as for the wider areas of computer vision.

3D object pose recognition is a rich field of research [13, 5, 11, 3, 9]. However, the cheap availability of scene depth information synchronized with photometric (RGB/grayscale) is only a recent phenomenon. As a result, approaches that use both modalities of information are rather sparse. It is expected that this new modality has the potential to provide considerable boost in the accuracy of pose recognition systems.

Figure 1. Ground truth and estimated poses overlaid for sample images. The red line signifies the ground-truth pose, green represents the estimated pose using visual + depth and blue is the estimated pose using visual local features only. [1]

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. [8, 7]. In [8], the authors show significant performance improvements for simultaneous recognition of object category, instance and pose recognition. They use machine learning to build an object-pose tree model from RGBD images and perform hierarchical inference on it. Although the performance improvements on category and instance recognition are impressive, object pose recognition performance is modest. The main reasons for this is that they use a classification strategy for pose recognition (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 [12], we have addressed these two issues. Since the pose space is continuous, we developed a *regression* framework for estimating object pose (using only color images). Moreover, along

with the appearance properties, our framework explicitly represents and models spatial distributions of salient features on the object surfaces to get a (continuous) pose estimate. This framework was shown to achieve state-of-the-art results on pose estimation from 2D images and significantly improved estimation on a variety of difficult datasets.

This paper makes the following contributions: firstly, we use a multi-kernel learning (MKL) approach to extend our framework for pose estimation to use multimodal RGBD data and show the benefits of doing this. This confirms that 3D shape information captured in the form of depth-maps provides valuable additional information about object pose as well as the fact that the proposed framework is able to exploit this information. Using depth is very useful for objects which lack visual features and/or suffer from partial occlusion (e.g. second pitcher in fig. 1) which severely limits pose estimation using visual cues alone. Furthermore, depth features also have the advantage of being illumination-invariant.

Secondly, we do extensive experimentation on a large, publicly available RGBD dataset [7] specially suited for this purpose. The dataset consists of RGBD data for 300 objects. The objects were put on a turntable and the Kinect sensor was used to gather data from 3 different sensor heights and 250 different object poses at each sensor height[2]. We evaluate our algorithm on this dataset and demonstrate that we achieve significant performance improvements over best published results.

In section 2, we describe how we extend our regression framework to use multiple kernels to incorporate depth information. In section 2.1 we describe the method of building the visual and depth kernels and lastly in section 2.2 we describe the various methods we have explored to learn using multiple kernels (MKL).

2. RGBD Multi-Kernel Regression

Let an image X^k be represented by its features – $X^k = \{x_i^k \in \mathbb{R}^2, f_i^k \in \mathbb{R}^{\mathbb{F}}, d_i^k \in \mathbb{R}^{\mathbb{D}}\}$ where $i = 1, ..., N_k$. Let x_i^k denote the i^{th} spatial location, and f_i^k and d_i^k the respective RGB and depth feature descriptors at that location. N_k is the number of local features (variant across images). Each image is also associated with a pose $v^k \in \mathbb{R}^{\mathbb{V}}$. The goal of regression is to learn a regularized mapping function $\hat{g}$ from input image features to the pose space from (paired) training data: (X^k, v^k). As shown in [12], the regression can be reduced to:

$$v = \hat{g}(X) = \sum_j b_j K(X, X^j) \quad (1)$$

[2]Other RGBD datasets are also available which are smaller and less suitable to validate pose-estimation performance – for a comprehensive description refer to [6].

$K(\cdot, \cdot)$, a p.d. kernel, measures the similarity between images. We extend this approach to use multiple kernels. We define two kernels – K for representing visual similarity and $G(\cdot, \cdot)$ for depth similarity. Multi-kernel regression is similarly reduced to:

$$v = \hat{g}(X) = \sum_j (b_j K(X, X^j) + c_j G(X, X^j)) \quad (2)$$

where G measures depth similarity between images and K measures similarity of visual local features within and between images.

We depict the proposed multi-kernel regression pipeline in fig. 2. It consists of two steps: feature embedding (section 2.1) and pose regression (section 2.2).

2.1. Feature Embedding

Our pose regression framework uses the feature embedding concept defined in our prior work [12] (please refer for details). This section summarizes the key ideas of this work focusing on the extensions to use depth information. The goal of feature embedding is to enforce regularity constraints: inter-image feature affinity, intra-image spatial affinity and prior manifold structure (topology, 1D...etc). At the same time, we want the embedded feature space to be independent of the variant number of features in input images. In this work, we learn two feature embeddings: one for the RGB features and another for the depth features. This requires the specification of corresponding affinity matrices.

RGB feature embedding follows [12]: The spatial affinity within each image and the feature affinity between images is computed using kernels resulting in two weight matrices. We use Geometric Blur (GB) as the RGB feature. A regularized objective function is then optimized (eqn 7, [12]) from training images to obtain a Laplacian Eigenmap embedding. Depth feature embedding is done in a slightly different (albeit simpler) way. Kinect depth data has many missing holes (see fig. 4) due to infrared absorption by some materials. We empirically established that a single global depth feature (per image) is more reliable than locally anchored multiple depth features. We use HOG applied to the depth images (dHOG) as a single depth descriptor per image. We now obtain a Laplacian embedding for the depth features also.

Consequently, each training image is mapped to two embedded feature representations corresponding to the RGB and depth data. Embedding the features from an unseen test image is done by solving an out-of-sample problem which estimates the mapping between the input space and embedding space.

2.2. Pose Regression

Once the feature embedding is done, pose regression is performed to learn the model for predicting pose from

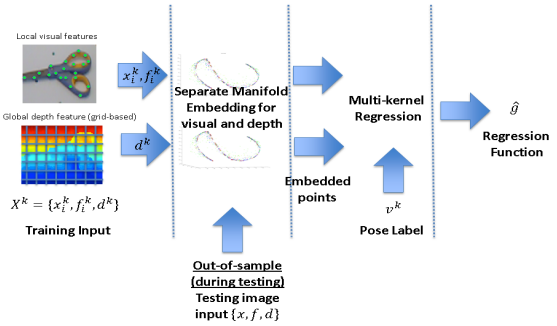

Figure 2. Pose Regression Using Feature Embedding. Model is learnt using training data. Estimation is done for test data by treating them as out-of-sample cases.

embedded features (see fig. 2). We experimented with two simple approaches.

In the first approach, called Multiple Kernel Regression (MKR), we estimated coefficients b_j and c_j in eq. 2 from the training data using Euclidean error norm on the estimated pose. Note that the X in eq. 2 now corresponds to embedded features computed in the last section. The advantage of this approach is that since weights for both RGB and depth kernels are learned simultaneously for each training example, example-specific weights are learnt than can achieve a good tradeoff between the relative quality of information (or lack thereof) in depth or RGB data. Thus, objects with good discriminative visual features can be robust to noisy or non-discriminative depth features and vice versa. The resulting pose estimator can be represented as follows:

$$v = \hat{g}(X) = \begin{bmatrix} K|G \end{bmatrix} \begin{bmatrix} b \\ - \\ c \end{bmatrix} \tag{3}$$

The second approach we tried is Multi-Kernel Learning (MKL) [2] where effectively a new kernel is represented as a weighted sum of several kernels (as in eq. 4). This method is analogous to a system with multiple experts. Weights must be assigned to each kernel according to its discriminative power in the regression problem. For our scenario, this approach has an effect of learning an apriori relative bias for/against the RGB information vis-a-vis the depth information.

$$K = \sum_{j}^{M} \eta_j K_j \tag{4}$$

M is the number of kernels and η are the scalar weights corresponding to each individual kernel. There are multiple ways of learning the discriminative power of each kernel. A common way is to use a correlation measure between individual kernel regressor and the ground truth values in the training data. This paper uses two heuristics: F-measure and M-measure [10]. The former measures the kernel alignment using the Frobenius norm between the kernel regresser and the inner product of the pose labels. The higher this value is for a kernel, the more it contributes to the combined kernel. The latter measure is based on the mean square error when performing regression using each kernel on training data. The smaller the MSE per kernel is, the less contribution that kernel has to the combined kernel. These two measures give us the weights η_j corresponding to each kernel K^j. The weighted sum kernel is now used to learn the pose regression from (1). We call the two pose regressors MKL-F and MKL-M respectively.

3. Experiments

We evaluate the presented approach on a large, publicly available RGBD dataset [7] specially suited for pose recognition. The Kinect sensor was used to gather data from 3 different sensor heights and poses were densely captured per object using a turntable. We followed the same experimental setup and loss function as in [8] for a fair comparison with this state-of-the-art approach. Thus, images captured from elevation angles of 30°and 60°were used for training while those from 45°were used for testing.

Geometric Blur (GB) features were computed on RGB images to represent photometric information while HOG features [4] were computed on the depth images to represent depth information (dHOG). While 45 most-significant GB features (local) were used, we used a single global dHOG per image. The depth data has large areas of missing depth (see fig. 4) due to infrared absorption by some materials and a single global (dHOG) was found to be more reliable than multiple local features. We also empirically analyzed the eigenvalues of the Laplacian Eigenmap embedding on the training data and found the best embedding dimensions to be 100 for GB (204-dimensional) and 75 for the 9x9-grid dHOG. We used Radial-Basis Function (RBF) kernels with Euclidean distances between feature vectors. These settings were used for all experiments.

The experimental results are presented in Table 1. The baseline approach [8] is presented in the last row. The first two rows represent our regression approach using only depth features (dHOG) and only RGB features (gsGB [12]), respectively. MKR represents our multi-kernel regression approach by concatenating the visual features kernel and global depth features kernel. MKL-F and MKL-M represent our regression approach when the kernel weights were computed using the F-measure and M-measure, respectively.

The dHOG approach shows worst performance, significantly lower than gsGB. This is due to two reasons: the depth feature is more noisy (large missing holes)

than the RGB data. Secondly, for robustness, we use a single global depth feature while gsGB uses multiple locally anchored visual features. The relative spatial arrangement of these features is very informative about the object pose and is effectively exploited by our visual-only algorithm. It can be seen that the proposed MKR and MKL approaches are able to use information from both color and depth features to provide a statistically significant improvement over regression using only depth or only color features. Note that we see a larger jump in the median performance than in the mean performance. The reason for this is that we get a significant performance boost for most of the object classes. A few object categories do not perform as well which can be attributed to being objects that have uniform non-discriminative pose-invariant shape, such as ball and bowl categories. In fact, asking the question of which pose these objects are in is an ill-posed question. Note that we get a relative improvement of approximately 11% and 4% over our own color-only baseline implementation.

Lastly, we compare the performance of the proposed regression framework with the baseline algorithm in [8]. The proposed MKR and MKL approaches easily outperform the baseline. In fact, MKL-M achieves a relative performance improvement of approx. 21% and 32% in the median and average pose accuracy, respectively. The median performance using MKL-M is shown for a representative subset of objects in fig. 3.

4. Conclusion

In this paper, we presented a novel MKL based regression framework that automatically combines the color and depth information present in the multi-modal RGBD image data for effective object pose estimation. We extensively evaluated the performance of the presented framework on a large dataset and showed that we are able to significantly outperform best published results in this area, thus validating the efficacy of the presented framework.

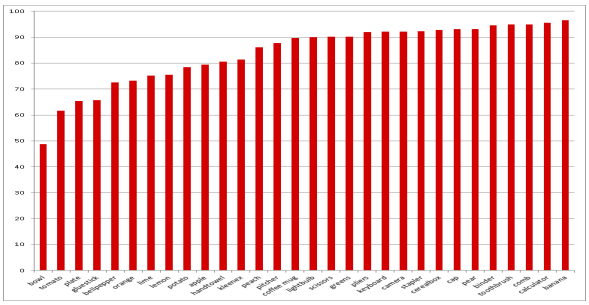

Figure 3. Median % accuracy for a subset of objects from RGBD-dataset

[1]Pose annotations indicate the assigned pose to the object image w.r.t an arbitrary reference. The marker orientation is not reflective of the surface normal to which the marker may seem to be anchored (anchoring is at the center of the image). Also, we show the yaw angle on a unit circle.

Table 1. Pose Recognition Performance over all 51 classes of RGBD-dataset

Method	Median Pose	Avg Pose	St. Dev.
dHOG (Depth)	51.25	50.62	5.16
gsGB (RGB) [12]	77.8	72.06	14.39
MKR (RGB+D)	**85.0**	**74.58**	13.69
MKL-F (RGB+D)	**86.3**	**75.50**	12.71
MKL-M (RGB+D)	**86.7**	**74.76**	14.21
[8]	71.40	56.80	-

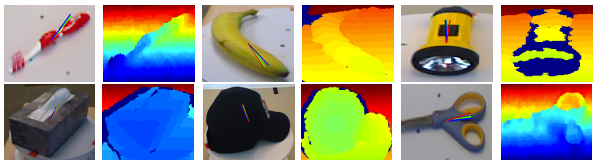

Figure 4. Pose estimates with RGB and depth images. Note missing depth on parts of the flashlight depth image (dark blue regions) [1]

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] L. Bo, X. Ren, and D. Fox. Kernel descriptors for visual recognition. In *NIPS*, 2010.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010.

[6] A. Janoch, S. Karayev, Y. Jia, J. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3-d object dataset: Putting the kinect to work. 2011.

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

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

[9] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[10] 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.

[11] S. Savarese and F.-F. Li. 3d generic object categorization, localization and pose estimation. In *ICCV*, 2007.

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

[13] M. Torres, A. Romea, and S. Srinivasa. Moped: A scalable and low latency object recognition and pose estimation system. In *ICRA*, 2010.