Fine Pose Estimation of Known Objects in Cluttered Scene Images

Sudipto Banerjee Sanchit Aggarwal Anoop M. Namboodiri
CVIT, International Institute of Information Technology, Hyderabad, India
{sudipto.banerjee@research., sanchit.aggarwal@research., anoop@}iiit.ac.in

Abstract

Understanding the precise 3D structure of an environment is one of the fundamental goals of computer vision and is challenging due to a variety of factors such as appearance variation, illumination, pose, noise, occlusion and scene clutter. A generic solution to the problem is ill-posed due to the loss of depth information during imaging. In this paper, we consider a specific but common situation, where the scene contains known objects. Given 3D models of a set of known objects and a cluttered scene image, we try to detect these objects in the image, and align 3D models to their images to find their exact pose. We develop an approach that poses this as a 3D-to-2D alignment problem. We also deal with pose estimation of 3D articulated objects in images. We evaluate our proposed method on BigBird dataset and our own tabletop dataset, and present experimental comparisons with state-of-the-art methods.

1. Introduction

The problem of detection, segmentation and recognition of objects in natural and indoor scenes are fundamental to computer vision and has received a significant amount of attention over the past decade [1, 4, 5, 7, 8]. However, various factors like illumination, background clutter and object-object interaction (*e.g.* occlusion), add to the complexity of the problem in unconstrained environments. Apart from object detection and recognition, estimating the position and orientation of objects in an image is of significant interest in tasks like 3D scene understanding and reconstruction.

Approaches to segmentation of indoor scenes include modeling the appearance of objects from images as well as their structure in RGBD data. Richtsfeld *et al*. [16], presegment the input image based on surface normals of objects, and perform segmentation on a graph over the estimated surface patches from NURBS. Hariharan *et al*. [8], proposed a CNN based architecture that classifies region proposals for simultaneous detection and segmentation of all instances of a category in an image. Aggarwal *et al*. [1] proposes a method specific to estimating floor regions from

Figure 1: We attempt to align 3D CAD models of objects to a single image. As observed, our proposed method is invariant to shape, size and texture of objects, and deals with scenes of varying complexity in terms of occlusion and clutter. (Best viewed in color)

appearance and geometric cues. Recent works have also explored the task of pose estimation of known objects in scene images. Zhang *et al*. [20] proposed a robust pose estimation method from line correspondences. In [11], Lim *et al*. have estimated the pose of furnitures in common indoor scene images, using correspondences between the test image and candidate poses obtained from training data. Zhu *et al*. [21] attempt to estimate the pose of objects in outdoor scene images, for applications in robotics grasping. The approach of aligning CAD models of objects to their images to estimate object pose has been explored in the context of generic furnitures and chairs was attempted by Lim *et al*. [10] and Aubry *et al*. [2] respectively. They used either geomtric features of 3D models or appearance of rendered exemplars to carry out the alignment. These are closest to our work in their approach, and we compare our results with the latter.

Our goal differs from the above primarily in its assumptions about the objects and their models. We try to find the pose of common tabletop objects that could potentially be low-lying and the models are assumed to be without any texture, making the alignment problem hard. Our major contributions include: 1) An ensemble of shape features that work well for aligning textureless 3D models, 2) A two-stage alignment scheme that is efficient and accurate and 3) An extension of the proposed approach to handle articulated

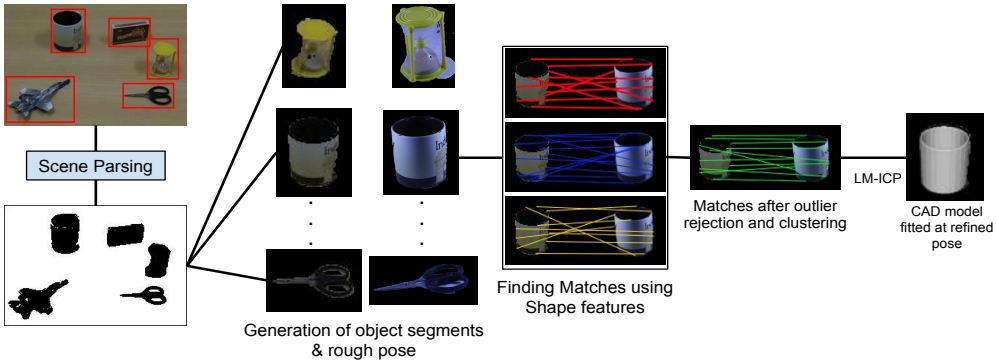

Figure 2: Method Outline: Input scene image is parsed to detect and segment objects using appearance and geometric cues. Initial pose is estimated from the closest matching examplar. An ensemble of shape features is used on object contour points, followed by rejection of spurious marches to obtain correspondences. Pose is refined by iteratively minimizing the reprojection error between the model and object. (Best viewed in color)

objects. We demonstrate our results on a variety of tabletop objects including transparent ones and scene images with occlusion and background clutter. Note that textureless 3D models are used to generalize our proposed method to objects that are very similar in shape and size, but vary in texture. Experimental results show that the proposed method outperforms the state-of-the-art method for pose estimation.

2. Method Overview

Figure 2 gives an outline of our proposed method. We work with images of known objects of different shape, size, texture and poses. We parse a given indoor scene image into constituent entities of individual objects present in the scene, using a rough segmentation algorithm. The retrieved entities in the scene with their rough boundaries are used to get the respective matched codebook from the learned exemplars of individual objects. The azimuth and elevation angle captured from the matched image is then used as an initial estimate of the pose. To further refine the estimate, we use an ensemble of shape features to establish matches between object and model silhouettes. We align a 3D model of the object, by iteratively minimizing the reprojection error.

2.1. Estimation of Entites and their rough Contours

For each of the object pose image in the training data, an exemplar model is learned based on appearance and geometric cues. Appearance cues capture the local shape of the object, while geometric cues are significant in cluttered scenes, where the object's color and texture are not easily computed, due to occlusions. We learn the appearance cues from the superpixel segmented image of the individual objects using simple linear iterative clustering (SLIC) superpixels. Similar to Aggarwal *et al*. [1], we use color, texture and shape cues of the superpixels and learn a GMM for separating object superpixels from that of background pixels. We learn exemplar SVM as proposed in [13] to learn the global shape and geometric cues extracted, using the HoG and CNN features.

Assuming we have N poses for each of the M models from our dataset, we learn $M \times N$ pose level GMMs and $M \times N$ exemplar SVMs. Using these discriminative classifiers, the probable objects in the test scene image are classified, and the closest possible match from the training data of all possible poses is found. GMMs give the probability estimates which acts as the unary potential, and a standard Potts model is used to get the pairwise potential. A standard implementation of MRF Grabcut [9] is used to iteratively refine the object pose estimation.

2.2. Fine Pose Estimation

We define a pose by two parameters, *azimuth angle* (θ), and *elevation angle* (ϕ), which are part of the spherical coordinate system. Given the initial estimate and the object class, our goal is to find the exact pose of the object in the image. We render an image of the corresponding CAD model from the initial estimate, using Back Face Culling. Model silhouette for a pose $[R|t]$ can be obtained by finding points X which satisfy the equation:

$$(R.N)^T.(R.X + t) = N^T.(X + R^T t) = 0 \qquad (1)$$

where R and t are the $Rotation$ matrix and $Translation$ vector respectively. The model and the image contours are scaled, and correspondences are obtained using feature matching. We now explain the approach in detail.

2.2.1 Local shape information using a Buffer of HOG features

In a cluttered environment, it is difficult to obtain precise segments from the image. To make our method robust, we create a buffer of Histogram of Oriented Gradients(HOG) features of K neighbouring points on either side, along the contour. Doing so captures the local shape information with respect to a point. This step prevents redundant matches due to lack of neighbouring information. Assuming we

have N object sample points, we learn N binary classifiers [11, 2], taking one feature as positive set and others as negative set. Each model point is matched to one of the classes which minimizes the classification score. More formally, the matching score between p-th object point and q-th model point is defined as $S(p,q) = \min_p w_p^T x_q$, where x_q is feature vector of the q-th model point, w_p is the weight vector learned by LDA classifier for object point p. The set of HOG features H is normalized such that $\sum_{i=1}^{|H|} h_i = 1$, where $h_i \in H$, and $|H|$ is the number of points in the contour. Value of K is fixed at 10 in our experiments.

2.2.2 Shape context and Chamfer matching

We use a combination of using shape context features [14] and chamfer matching [12, 17] to obtain correspondences. The shape context of p_i is defined as a normalized k-bin log-polar histogram of the relative coordinates, defined by vector $(q - p_i)$ of the remaining points q with respect to p_i. The key idea of this feature is to capture the distribution of all points with respect to a reference point. Cost of matching a model point to an object point is given by:

$$C = C_S + \beta C_F + (1 - \beta) C_C \tag{2}$$

where C_S is the χ^2 distance between the shape contexts of the point pairs, C_F and C_C are the figural continuity cost and curvature cost [19] respectively. Here, $0 \leq C_S, C_F, C_C \leq 1$, and β is the weighting parameter. Modelling this as a bipartite matching problem, we solve this problem as

$$\text{minimize} \sum_{p_i \in O} \sum_{q_j \in M} C(i,j) x_{ij} \tag{3}$$

subject to constraints,

$$\sum_{q_j \in M} x_{ij} = 1 \text{ for } p_i \in O \qquad \sum_{p_i \in O} x_{ij} = 1 \text{ for } q_j \in M$$

where $x_{ij} \geq 0$ for $p_i, q_j \in O, M$. We use Jonker-Volgenant algorithm to solve this linear assignment problem, because of its robustness and speed. $C(i,j)$ is pairwise cost matrix. Variable x_{ij} is 1 when sample point p_i from model contour M corresponds to sample point q_j from object contour O and 0 otherwise. Chamfer matching uses chamfer distance as the metric for establishing matches. Individually, they both are prone to occlusion and background clutter. However chamfer matching used by Thayananthan *et al.* [19] is not robust to intense background clutter and occlusion. Hence, following the work of Nguyen *et al.* [15], we formulate the matching task as a maximum of a posteriori (MAP) problem. In other words, for each model point m, finding the corresponding object point o is equivalent to solving the following:

$$p(o|m) = \kappa \max_o e^{-\alpha d(m,o)} \tag{4}$$

where κ and α are positive parameters. $p(o|m)$ is the probability of having an object point o corresponding to model point m and, $d(m,o)$ denotes the chamfer distance between m and o with edge orientation error between them also taken into account. Similar to [15], Variational Mean Field approach is used to solve the MAP problem. Finally, we obtain a set of point-to-point matches based on shape context features and chamfer distance metric.

2.2.3 Ensemble of shape features

The set of correspondences obtained from each of the shape matching methods, might contain outliers. The problem intensifies when an object is occluded. Using segmented masks of each object, we obtain a partial contour for an occluded object by subtracting the part of the contour common to both the object in front and the occluded one. We use a faster variant of RANSAC: Progressive Sample Consensus (PROSAC) [3], to remove the outliers. The use of PROSAC drastically reduces the computation time for obtaining the inliers. We use an ensemble of these shape features to get a reduced set of *pure* matches, which contain the least number of outliers. We partition the sampled points on the object contour into K clusters. We find that, clustering using k-means algorithm yields satisfactory results. The idea is that, for a model point, if the shape features are good enough to capture its local appearance, they should all map to object points which are *close* to each other. More precisely, for a model point, we assert that object point matches obtained from the shape features, are similar if they belong to the same cluster. We finally obtain the *common* 3D-to-2D correspondence, where the model point is matched with the object point which has the minimum reprojection error with that model vertex. As shown in figure 3(a), using the reduced set of matches, our algorithm converges with significantly less number of iterations with minimum error. We plot the error function with respect to azimuth and elevation angle for an example, and find that it is much smoother, when using ensemble of features. Hence, a lesser likelihood of getting stuck at a local minima.[1] Finally, given a set of pure 3D-to-2D correspondences, we use Iterative Closest Point algorithm using Levenberg-Marquardt optimization (LM-ICP) [6], using the initial estimate obtained before, to obtain the transformation matrix which aligns the model to the object. Table 2 shows the reprojection error, RMSE and time taken at each step of our pipeline. While calculating the reprojection error and RMSE, the points have been normalized such that $0 \leq p, q \leq 1$, where p and q are model and object points respectively.

[1] The plots could not be shown due to limitations in space.

	Transparent Bottle		Sand Clock		Pen Stand		Scissors		Spectacles Case	
	Accuracy(%)	MPE	Accuracy(%)	MPE	Accuracy(%)	MPE	Accuracy(%)	MPE	Accuracy(%)	MPE
S3DC	31.11	0.147	83.33	0.086	**91.11**	0.093	6.67	0.4	44.44	**0.016**
Ours	**83.52**	**0.0016**	**86.81**	**0.033**	83.52	**0.002**	**73.63**	**0.086**	**54.65**	0.121

Table 1: We compare the classification accuracy and mean pose error (MPE) with S3DC, for some of the objects from our dataset with varying complexity in terms of shape, size and material.

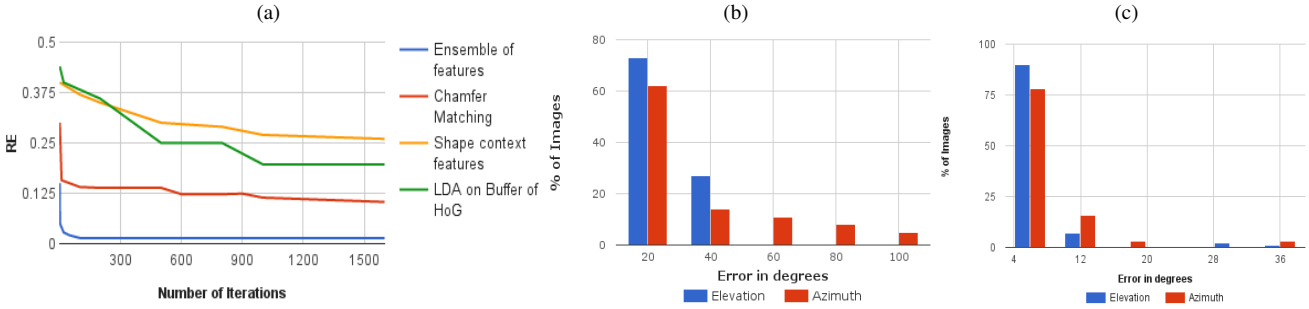

Figure 3: (a) Plot showing Reprojection Error (RE) vs number of iterations required to align a model to an object. Note that, using correspondences obtained from ensemble of features, RE reaches convergence with lower number of iterations. (b) Error in rough estimate. (c) Error in pose after fine alignment. Fine alignment error is shown those examples whose the initial estimate is within 20°of groundtruth. (Best viewed in color)

	RE	RMSE	Time(in sec)
LDA-BHOG	0.197	0.243	12.87
Shape Context	0.271	0.336	5.49
Chamfer Distance	0.089	0.097	2.46
Ensemble of features	**0.015**	**0.026**	**23.56**

Table 2: Table showing the Reprojection Error(RE), Root Mean Square Error(RMSE), and time taken for estimating pose of object with each of the shape features individually, and using ensemble of features, given an initial estimate. The Chamfer Distance algorithm was implemented in C++.

3. Experimental Results and Analysis

We evaluate our proposed method on the BigBird dataset [18] as well as our own dataset of tabletop objects (referred to as TableTop). We experiment with 15 objects from BigBird, consisting of 600 images per object, and 50 objects from TableTop, each having 180 images. We also run our approach on cluttered tabletop scene images from our dataset. The outputs of our solution are given in Figure 4. We also compare our results with Seeing 3D Chairs [2] (S3DC).

We learn the exemplar model on our training data, and obtain a learned classifier for each object sample. A test image is classified into the object class corresponding to the sample classifier yielding the maximum score. Figure 3(b) and 3(c) summarizes the quantitative results of our method. As observed, we are able to provide an initial rough estimate within 20° of groundtruth for 76% of the correctly classified examples. As evident from figure 3(c), we further refine error till 6° of groundtruth for 83% of our examples. Given the groundtruth pose vector, (c_x, c_y, c_z) and predicted pose vector, (c'_x, c'_y, c'_z), the pose error is taken as the euclidean distance between them. Table 1 gives results for some of the objects from our dataset. We compare the classification accuracy and mean pose errors with S3DC in table 3. Figure 6 show the qualitative comparisons. We also evaluate our approach on tabletop scene images. Using the pose of objects from a single RGB scene image, we fit their corresponding CAD models. The method fails when the initial estimate is far away from groundtruth. Failure cases are shown in figure 6.

We also run our experiments on articulated objects. Objects such as *scissors* have articulation points which influence certain model vertices to rotate or translate with respect to them. Given the articulation points, and their transformation matrices, we modify the LM-ICP framework to account for the error due to local deformations in the object. Figure 5 shows the results on articulated objects.

	BigBird		**TableTop**	
	Accuracy (%)	MPE	Accuracy (%)	MPE
S3DC	34.5	0.013	45.7	0.044
Ours	**49.7**	**0.008**	**67.3**	**0.021**

Table 3: Quantitative comparison of S3DC with our proposed method on BigBird and TableTop dataset. Note that our method outperforms S3DC with respect to classification accuracy and mean pose error. Mean Pose errors are compared for correctly classified examples only.

Figure 4: We show results of our proposed approach. Top row shows the scene images from our dataset. In bottom row, we superimpose the CAD models onto the image in their exact pose. Note that occluded objects are also aligned with minimal error. (Best viewed in color)

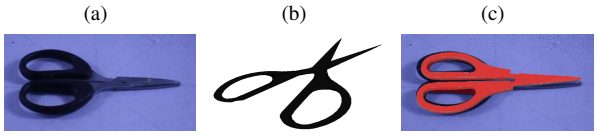

Figure 5: Demonstration of our method on articulated object. (a) Input test image (b) Original model (c) Deformed model superimposed on image. (Best viewed in color)

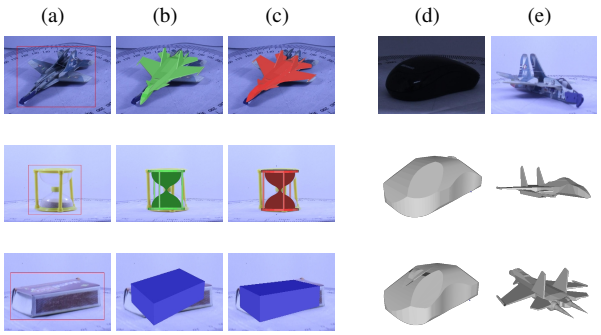

Figure 6: (a) Input test images (b) Fine poses obtained from S3DC (c) Fine poses obtained from the proposed approach. Although, S3DC correctly classifies objects, they fail to align the CAD model properly. (d-e) Failure cases. Top row shows input images. Middle and bottom rows show groundtruth and obtained pose respectively. (Best viewed in color)

4. Conclusions

We have developed an ensemble of shape features for the purpose of fine alignment of textureless 3D models of objects to 2D images of real-world objects. We are able achieve this in spite of the clutter and occlusion in the scene as well as the lack of texture on the 3D models. The method is shown to be very effective on a dataset of tabletop objects and robust against partial occlusion and comparative results with state of the art are presented.

References

[1] S. Aggarwal, A. M. Namboodiri, and C. Jawahar. Estimating floor regions in cluttered indoor scenes from first person camera view. In *ICPR, 2014*. 1, 2

[2] M. Aubry, D. Maturana, A. A. Efros, B. C. Russell, and J. Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR, 2014*. 1, 3, 4

[3] O. Chum and J. Matas. Matching with prosac-progressive sample consensus. In *CVPR, 2005*. 3

[4] J. Dong, Q. Chen, S. Yan, and A. Yuille. Towards unified object detection and semantic segmentation. In *ECCV, 2014*. 1

[5] A. Faktor and M. Irani. Co-segmentation by composition. In *ICCV, 2013*. 1

[6] A. W. Fitzgibbon. Robust registration of 2d and 3d point sets. *Image and Vision Computing*, 2003. 3

[7] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. *IJCV, 2014*. 1

[8] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV, 2014*. 1

[9] V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *PAMI*, 2006. 2

[10] J. J. Lim, A. Khosla, and A. Torralba. Fpm: Fine pose parts-based model with 3d cad models. In *ECCV, 2014*. 1

[11] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In *ICCV, 2013*. 1, 3

[12] M.-Y. Liu, O. Tuzel, A. Veeraraghavan, and R. Chellappa. Fast directional chamfer matching. In *CVPR, 2010*. 3

[13] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 2

[14] G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. *TPAMI, 2005*. 3

[15] D. T. Nguyen. A novel chamfer template matching method using variational mean field. In *CVPR, 2014*. 3

[16] A. Richtsfeld, T. Morwald, J. Prankl, M. Zillich, and M. Vincze. Segmentation of unknown objects in indoor environments. In *IROS, 2012*. 1

[17] J. Shotton, A. Blake, and R. Cipolla. Multiscale categorical object recognition using contour fragments. *PAMI*, 2008. 3

[18] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel. Bigbird: A large-scale 3d database of object instances. In *ICRA, 2014*. 4

[19] A. Thayananthan, B. Stenger, P. H. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *CVPR, 2003*. 3

[20] L. Zhang, C. Xu, K.-M. Lee, and R. Koch. Robust and efficient pose estimation from line correspondences. In *ACCV, 2012*. 1

[21] M. Zhu, K. G. Derpanis, Y. Yang, S. Brahmbhatt, M. Zhang, C. Phillips, M. Lecce, and K. Daniilidis. Single image 3d object detection and pose estimation for grasping. In *ICRA, 2014*. 1