

# Depth-Encoded Hough Voting for Joint Object Detection and Shape Recovery

Min Sun<sup>1</sup>

Gary Bradski<sup>2</sup>

Bing-Xin Xu<sup>1</sup>

Silvio Savarese<sup>1</sup>

<sup>1</sup>Electrical and Computer Engineering  
University of Michigan, Ann Arbor, USA  
{sunmin,xbx}@umich.edu {silvo}@eecs.umich.edu

<sup>2</sup>Willow Garage, Menlo Park, CA, USA  
{bradski}@willowgarage.com

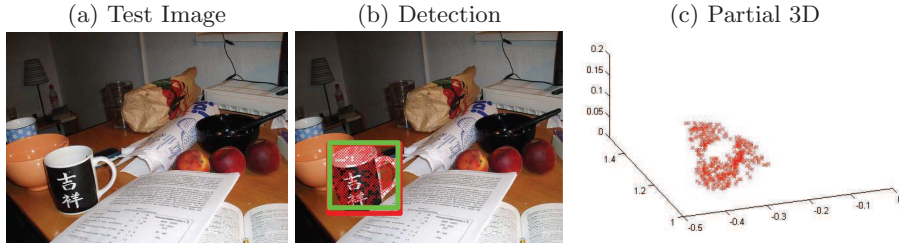
**Abstract.** Detecting objects, estimating their pose and recovering 3D shape information are critical problems in many vision and robotics applications. This paper addresses the above needs by proposing a new method called DEHV - Depth-Encoded Hough Voting detection scheme. Inspired by the Hough voting scheme introduced in [13], DEHV incorporates depth information into the process of learning distributions of image features (patches) representing an object category. DEHV takes advantage of the interplay between the scale of each object patch in the image and its distance (depth) from the corresponding physical patch attached to the 3D object. DEHV jointly detects objects, infers their categories, estimates their pose, and infers/decodes objects depth maps from either a single image (when no depth maps are available in testing) or a single image augmented with depth map (when this is available in testing). Extensive quantitative and qualitative experimental analysis on existing datasets [6,9,22] and a newly proposed 3D table-top object category dataset shows that our DEHV scheme obtains competitive detection and pose estimation results as well as convincing 3D shape reconstruction from just one single uncalibrated image. Finally, we demonstrate that our technique can be successfully employed as a key building block in two application scenarios (highly accurate 6 degrees of freedom (6 DOF) pose estimation and 3D object modeling).

## 1 Introduction

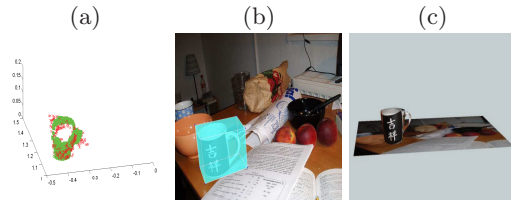
Detecting objects and estimating their geometric properties are crucial problems in many application domains such as robotics, autonomous navigation, high-level visual scene understanding, activity recognition, and object modeling. For instance, if one wants to design a robotic system for grasping and manipulating objects, it is of paramount importance to encode the ability to accurately estimate object orientation (pose) from the camera view point as well as recover structural properties such as its 3D shape. This information will help the robotic arm grasp the object at the right location and successfully interact with it.

This paper addresses the above needs, and tackles the following challenges:

- i) Learn models of object categories by combining view specific depth maps



**Fig. 1.** Illustration of key steps in our method. Given a single (previously) unseen testing image (panel a), our DEHV (Depth-Encoded Hough Voting-based) scheme is used to detect objects (panel b). Ground truth bounding box is shown in red. Our detection is shown in green. The centers of the image patches which cast votes for the object location are shown in red crosses. During detection, our method simultaneously infers object depth maps of the detected object (panel c). This allows the estimation of the partial 3D shape of the object from a single image!



**Fig. 2.** Point clouds (green) from a 3D model is registered to the inferred partial 3D point cloud (red) by DEHV (a). This allows us to achieve an accurate 6 DOF pose estimation (b) and realistic 3D object modeling (c).

along with the associated 2D image of objects in the same class from different vantage points. We demonstrate that combining imagery with 3D information helps build richer models of object categories that can in turn make detection and pose estimation more accurate. ii) Design a coherent and principled scheme for detecting objects and estimating their pose from either just a single image (when no depth maps are available in testing) (Fig. 1(b)), or a single image augmented with depth maps (when these are available in testing). In the latter case, 3D information can be conveniently used by the detection scheme to make detection and pose estimation more robust than in the single image case. iii) Have our detection scheme recover the 3D structure of the object from just a single uncalibrated image (when no 3D depth maps are available in testing) (Fig. 1(c)) and without having seen the object instance during training.

Inspired by implicit shape model (ISM) [13], our method is based on a new generalized Hough voting-based scheme [2] that incorporates depth information into the process of learning distributions of object image patches that are compatible with the underlying object location (shape) in the image plane. We call our scheme *DEHV - Depth-Encoded Hough Voting scheme* (Sec. 3). DEHV addresses the intrinsic weaknesses of existing Hough voting schemes [13,10,16,17] where errors in estimating the scale of each image object patch directly affects the ability of the algorithm to cast consistent votes for the object existence. To

resolve this ambiguity, we take advantage of the interplay between the scale of each object patch in the image and its distance (depth) from the corresponding physical patch attached to the 3D object, and specifically use the fact that objects (or object parts) that are closer to the camera result in image patches with larger scales. Depth is encoded in training by using available depth maps of the object from a number of view points. At recognition time, DEHV is applied to detect objects (Fig. 1(b)) and simultaneously infer/decode depths given hypotheses of detected objects (Fig. 1(c)). This process allows the reinforcement of the existence of an object even if a depth map is not available in testing. If depth maps are available in testing, the additional information can be used to further validate if a given detection hypothesis is correct or not. As a by-product of the ability of DEHV to infer/decode depth at recognition time, we can estimate the location in 3D of each image patch involved in the voting, and thus recover the partial 3D shape of the object. Critically, depth decoding can be achieved even if just a single test image is provided. Extensive experimental analysis on a number of public datasets (including car Pascal VOC07 [6], mug ETHZ Shape [9], mouse and stapler 3D object dataset [21]) as well as a newly created in-house dataset (comprising 3 object categories) are used to validate our claims (Sec. 4). Experiments with the in-house dataset demonstrate that our DEHV scheme: i) achieves better detection rates (compared to the traditional Hough voting scheme); further improvement is observed when depth maps are available in testing; ii) produces convincing 3D reconstructions from single images; the accuracy of such reconstructions have been qualitatively assessed with respect to ground truth depth maps. Experiments with public datasets demonstrate that our DEHV successfully scales to different types of categories and works in challenging conditions (severe background clutter, occlusions). DEHV achieves state of the art detection results on several categories in [6,9], and competitive pose estimation results on [21]). Finally, we show anecdotal results demonstrating that DEHV is capable to produce convincing 3D reconstructions from single uncalibrated images from [6,9,21] in Fig. 12.

We demonstrated the utility of DEHV in two applications (Sec. 4.3): i) Robot object manipulation: we show that DEHV enables accurate 6 DOF pose estimation (Fig. 2(b)); ii) 3D object modeling: we show that DEHV enables the design of a system for obtaining eye catching 3D objects models from just one single image (Fig. 2(c));

## 2 Previous Work

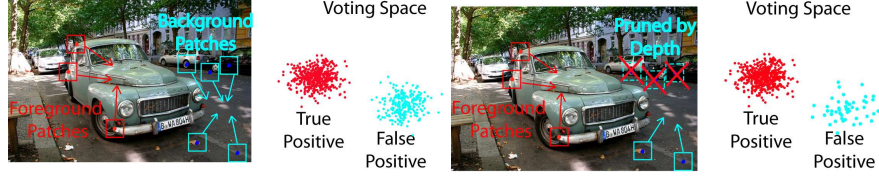
In the last decade, the vision community has made substantial progress addressing the problem of object categorization from 2D images. While most of the work has focussed on representing objects as 2D models [4,13,8] or collections of 2D models [23], very few methods have tried to combine in a principled way the appearance information that is captured by images and the intrinsic 3D structure representative of an object category. Works by [25,21,22] have proposed solutions for modeling the way how 2D local object features (or parts) and their

relationship vary in the image as the camera view point changes. Other works [11,27,14,1] propose hybrid models where reconstructed 3D object models are augmented with features or parts capturing diagnostic appearance. Few of them (except [26] for objects) have demonstrated and evaluated the ability to recover 3D shape information from a single query image. However, instead of using image patches to transfer meta-data (like depth) to the testing instance as in [26], 3D information is directly encoded into our model during training. Other works propose to address the problem of detecting and estimating geometrical properties of single object instances [12,19,18,15]; while accurate pose estimation and 3D object reconstruction are demonstrated, these methods cannot be easily extended to incorporate intra-class variability so as to detect and reconstruct object categories. Unlike our work, these techniques also require that the objects have significant interior texture to carry out geometric registration. Other approaches assume that additional information about the object is available in both training and testing (videos, 3D range data) [20,5]. Besides relying on more expensive hardware platforms, these approaches tend to achieve high detection accuracy and pose estimation, but fail when the additional 3D data is either partially or completely unavailable.

### 3 Depth-Encoded Hough Voting

In recognition techniques based on hough voting [2] the main idea is to represent the object as a collection of parts (patches) and have each part to cast votes in a discrete voting-space. Each vote corresponds to a hypothesis of object location  $x$  and class  $O$ . The object is identified by the conglomeration of votes in the voting space  $V(O, x)$ .  $V(O, x)$  is typically defined as the sum of independent votes  $p(O, x, f_j, s_j, l_j)$  from each part  $j$ , where  $l_j$  is the location of the part,  $s_j$  is the scale of the part, and  $f_j$  is the part appearance.

Previously proposed methods [13,10,16,17] differ mainly by the mechanism for selecting good parts. For example, parts may be either selected by an interest point detector [13,16], or densely sampled across many scales and locations [10]; and the quality of the part can be learned by estimating the probability [13] that the part is good or discriminatively trained using different types of classifiers [16,10]. In this paper, we propose a novel method that uses 3D depth information to guide the part selection process. As a result, our constructed voting space  $V(O, x|D)$ , which accumulates votes for different object classes  $O$  at location  $x$ , depends on the corresponding depth information  $D$  of the image. Intuitively, any confusing part that is selected at a wrong scale can be pruned out by using depth information. This allows us to select parts which are consistent with the object physical scale. It is clear that depending on whether object is closer or further, or depending on the actual 3D object shape, the way how each patch votes will change (Fig. 3).



**Fig. 3.** Left panel shows that patches associated to the actual object parts (red boxes) will vote for the correct object hypothesis (red dots) in the voting space on the right. However, parts from the background or other instances (cyan boxes) will cast confusing votes and create a false object hypothesis (green dots) in the voting space. Right panel shows that given depth information, the patches selected in a wrong scale can be easily pruned. As a result, the false positive hypothesis will be supported by less votes.

In detail, we define  $V(O, x|D)$  as the sum of individual probabilities over all observed images patches at location  $l_j$  and for all possible scales  $s_j$ , i.e,

$$\begin{aligned} V(O, x|D) &= \sum_j \int p(O, x, f_j, s_j, l_j | d_j) ds_j \\ &= \sum_j \int p(O, x | f_j, s_j, l_j, d_j) p(f_j | s_j, l_j, d_j) p(s_j | l_j, d_j) P(l_j | d_j) ds_j \end{aligned} \quad (1)$$

where the summation over  $j$  aggregates the evidence from individual patch location, and the integral over  $s_j$  marginalizes out the uncertainty in scale for each image patch. Since  $f_j$  is calculated deterministically from observation at location  $l_j$  with scale  $s_j$ , and we assume  $p(l_j | d_j)$  is uniformly distributed given depth, we obtain:

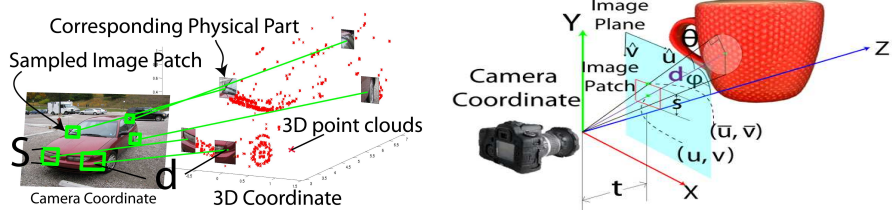
$$\begin{aligned} V(O, x|D) &\propto \sum_j \int p(O, x | f_j, s_j, l_j, d_j) p(s_j | l_j, d_j) ds_j \\ &= \sum_{j,i} \int p(O, x | C_i, s_j, l_j, d_j) p(C_i | f_j) p(s_j | l_j, d_j) ds_j \end{aligned} \quad (2)$$

Here we introduce codebook entry  $C_j$ , matched by feature  $f_j$ , into the framework, so that the quality of a patch selected will be related to which codeword it is matched to. Noting that  $C_j$  is calculated only using  $f_j$  and not the location  $l_j$ , scale  $s_j$ , and depth  $d_j$ , we simplify  $p(C_j | f_j, s_j, l_j, d_j)$  into  $p(C_j | f_j)$ . And by assuming that  $p(O, x | \cdot)$  does not depend on  $f_j$  given  $C_j$ , we simplify  $p(O, x | C_j, f_j, s_j, l_j, d_j)$  into  $p(O, x | C_j, s_j, l_j, d_j)$ .

Finally, we decompose  $p(O, x | \cdot)$  into  $p(O | \cdot)$  and  $p(x | \cdot)$  as follows:

$$V(O, x|D) \propto \sum_{j,i} \int p(x | O, C_i, s_j, l_j, d_j) p(O | C_i, s_j, l_j, d_j) p(C_i | f_j) p(s_j | l_j, d_j) ds_j$$

**Scale to depth mapping.** We design our method so as to specifically select image patches that tightly enclose a sphere with a fix radius  $r$  in 3D during



**Fig. 4.** Illustration of depth to scale mapping. Right panel illustrates the concept of depth to scale mapping. Training under the assumption that an image patch (green box) tightly encloses the physical 3D part with a fix size, our method deterministically selects patches given the patch center  $l$ , 3D information of the image, and focal length  $t$ . During testing, given the selected image patches on the object, our method directly infers the location of the corresponding physical parts and obtains the 3D shape of the object. Left Panel illustrates the physical interpretation of Eq. 3. Under the assumption that image patch (red bounding box) tightly encloses the 3D sphere with radius  $r$ , the patch scale  $s$  is directly related to the depth  $d$  given camera focal length  $t$  and the center  $l = (u, v)$  of the image patch. Notice that this is a simplified illustration where the patch center is on the  $yz$  plane. This figure is best viewed in color.

training. As a result, our model enforces a 1-to-1 mapping  $m$  between scale  $s$  and depth  $d$ . This way, given the 3D information, our method deterministically select the scale of the patch at each location  $l$ , and given the selected patches, our method can infer the underlying 3D information (Fig.4). In detail, given the camera focal length  $t$ , the corresponding scale  $s$  at location  $l = (u, v)$  can be computed as  $s = m(d, l)$  and the depth  $d$  can be inferred from  $d = m^{-1}(s, l)$ . The mapping  $m$  obeys the following relations:

$$s = 2(\bar{v} - v); \quad \bar{v} = \tan(\theta + \phi)t; \quad \theta = \arcsin\left(\frac{r}{d_{yz}}\right); \quad \phi = \arctan\left(\frac{v}{t}\right)$$

$$d_{yz} = \frac{d\sqrt{t^2 + v^2}}{\sqrt{u^2 + v^2 + t^2}} : \text{d projected onto } yz \text{ plane} \quad (3)$$

Hence,  $p(s|l, d) = \delta(s - m(d, l))$ . Moreover, using the fact that there is a 1-to-1 mapping between  $s$  and  $d$ , probabilities  $p(x|.)$  and  $p(O|.)$  are independent to  $d$  given  $s$ . As a result, only scale  $s$  is directly influenced by depth.

In the case when depth is unknown,  $p(s|l, d)$  becomes a uniform distribution over all possible scales. Our model needs to search through the scale space to find patches with correct scales. This will be used to detect the object and simultaneously infer the depth  $d = m^{-1}(s, l)$ . Hence, the underlying 3D shape of the object will be recovered.

**Random forest codebook.** In order to utilize dense depth map or infer dense reconstruction of an object, we use random forest to efficiently map features  $f$  into codeword  $C$  (similar to [10]) so that we can evaluate patches densely distributed over the object. Moreover, random forest is discriminatively trained to select salient parts. Since feature  $f$  deterministically maps to  $C^i$  given the  $i_{th}$

random tree, the voting score  $V(O.x|D)$  becomes:

$$V(O, x|D) \propto \sum_{j,i} \int p(x|O, C^i(f_j), s_j, l_j) p(O|C^i(f_j)) p(s_j|l_j, d_j) ds_j \quad (4)$$

where the summation over  $i$  aggregates the discriminative strength of different trees. In section 3.1, we describe how the distributions of  $p(x|O, C^i(f_j), s_j, l_j)$  and  $p(O|C^i(f_j))$  are learned given training data, so that each patch  $j$  knows where to cast votes during recognition.

### 3.1 Training the model

We assume that for a number of training object instances, the 3D reconstruction  $D$  of the object is available. This corresponds to having available the distance (depth) of each image object patch from its physical location in 3D. Our goal is to learn the distributions of location  $p(x|.)$  and object class  $p(O|.)$ , and the mapping of  $C^i(f)$ . Here we define location  $x$  of an object as a bounding box with center position  $q$ , height  $h$ , and aspect ratio  $a$ . We sample each image patch centered at location  $l$  and select the scale  $s = m(l, d)$ . Then the feature  $f$  is extracted from the patch  $(l, s)$ . When the image patch comes from a foreground object, we cache: 1) the information of the relative voting direction  $b$  as  $\frac{q-l}{s}$ ; 2) the relative object-height/patch-scale ratio  $w$  as  $\frac{h}{s}$ ; 3) the object aspect ratio  $a$ . Then, we use both the foreground patches (positive examples) and background patches (negative examples) to train a random forest to obtain the mapping  $C^i(f)$ .  $p(O|C)$  is estimated by counting the frequency that patches of  $O$  falls in the codebook entry  $C$ .  $p(x|O, C, s, l)$  can be evaluated given the cached information  $\{v, w, a\}$  as follows:

$$p(x|O, C, s, l) \propto \sum_{j \in g(O, C)} \delta(q - b_j \cdot s + l, h - w_j \cdot s, a - a_j) \quad (5)$$

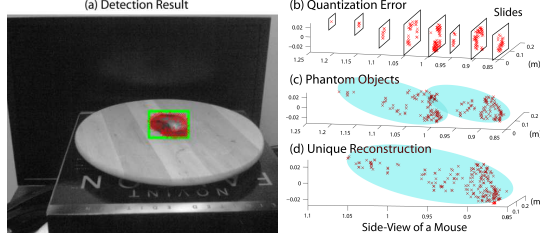
where  $g(O, C)$  is a set of patches from  $O$  mapped to codebook entry  $C$ .

### 3.2 Recognition and 3D reconstruction

**Recognition when depth is available.** It is straightforward to use the model when 3D information is observed during recognition. Since the uncertainty of scale is removed, Eq. 4 becomes

$$V(O, x|D) \propto \sum_{j,i} p(x|O, C^i(f_j), m(l_j, d_j), l_j) p(O|C^i(f_j)) \quad (6)$$

Since  $s_j = m(l_j, d_j)$  is a single value at each location  $j$ , the system can detect objects more efficiently by computing less features and counting less votes. Moreover, patches selected using local appearance at a wrong scale can be pruned out to reduce hallucination of objects (Fig. 3).



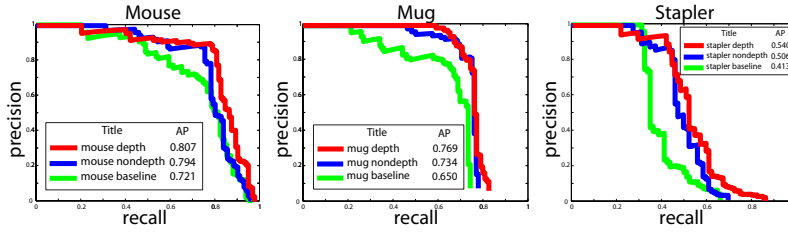
**Fig. 5.** A typical detection result in (a) shows object hypothesis bounding box (green box) and patches (red crosses) vote for the hypothesis. A naive reconstruction suffers from quantization error (b) and phantom objects (c). Our algorithm overcomes these issues and obtains (d)

**Recognition when depth is not available.** When no 3D information is available during recognition,  $p(s_j|l_j, d_j)$  becomes a uniform distribution over the entire scale space. Since there is no closed form solution of integral over  $s_j$ , we propose to discretize the space into a finite number of scales  $S$  so that Eq. 4 can be approximated by  $V(O, x|D) \propto \sum_{j,i} \sum_{s_j \in S} p(x|O, C^i(f_j), s_j, l_j) p(O|C^i(f_j))$

**Decoding 3D information.** Once we obtain a detection hypothesis  $(x, O)$  (Green box in Fig. 5(a)) corresponding to a peak in the voting space  $V$ , the patches that have cast votes for a given hypothesis can be identified (Red cross in Fig. 5(a)). Since the depth information is encoded by the scale  $s$  and position  $l$  of each image patch, we apply Eq 3 in a reverse fashion to infer/decode depths from scales. The reconstruction, however, is affected by a number of issues: i) **Quantization error:** The fact that scale space is discretized into a finite set of scales, implies that the depths  $d$  that we obtained are also discretized. As a result, we observe the reconstructed point clouds as slices of the true object (See Fig. 5(b)). We propose to use the height of the object hypothesis  $h$  and the specific object-height/patch-scale ratio  $w$  to recover the continuous scale  $\hat{s} = h/w$ . Notice that since  $w$  is not discretized,  $\hat{s}$  is also not discretized. Hence, we recover the reconstruction of an object as a continuum of 3D points (See Fig. 5(c)). ii) **Phantom objects:** The strength and robustness of our voting-based method comes from the ability to aggregate pieces of information from different training instances. As a result, the reconstruction may contain multiple phantom objects since image patches could resemble those coming from different training instances with slightly different intrinsic scales. Notice that the phantom objects phenomenon reflects the uncertainty of the scale of the object in an object categorical model. In order to construct a unique shape of the detected object instance, we calculate the relative object height in 3D with respect to a selected reference instance to normalize the inferred depth. Using this method, we recover a unique 3D shape of the detected object.

## 4 Evaluation

We evaluated our DEHV algorithm on several datasets. The training settings were as follows. For each training image, we randomly sample 100 image patches from object instances and 500 image patches from background regions. The



**Fig. 6.** Object localization results are shown as precision recall curves evaluated using PASCAL VOC protocol. (Green curve) Result using standard ISM model (baseline). (Blue curve) Result using DEHV with no depth information during testing. (Red curve) Result using DEHV with partial depth information during testing. Notice the consistent improvement of average precision (AP) compared to the baseline hough voting.

scale of the patch size from the corresponding object instance is determined by its (known) depth (Fig. 4). At the end, 10 random trees (Sec. 3.1) are trained using the sampled foreground and background patches for each dataset. For all experiment, we use a Hog-like feature introduced in [10]. During detection, our method treats each discrete viewpoint as a different class  $O$ .

#### 4.1 Exp.I: System analysis on a novel 3D table-top object dataset

Due to the lack of datasets comprising both images and 3D depth maps of set of generic object categories, we propose a new 3D table-top object category dataset collected on a robot platform. The dataset contains three common table-top object categories: mice, mugs, and staplers, each with 10 object instances. We arrange these objects in two different sets for the purpose of object localization and pose estimation evaluation. The object localization dataset (Table-Top-Local) contains 200 images with the number of object ranging from 2 to 6 object instances per image in a clutter office environment. The object pose estimation dataset (Table-Top-Pose) contains 480 images where each object instance is captured under 16 different poses (8 angles and 2 heights). For both settings, each image comes with depth information collected using a structure-light stereo camera. Please see the author’s project page (<http://www.eecs.umich.edu/~sunmin>) for more information about the dataset.

We evaluate our method under 3 different training and testing conditions, which are 1) standard ISM model trained and tested without depth, 2) DEHV trained with depth but tested without depth, and 3) DEHV trained and tested with depth. We show that the knowledge of 3D information helps in terms of object localization (Fig. 6), and pose estimation (Fig. 7). Moreover, we evaluate our method’s ability to infer depth from just a single 2D image. Given the ground truth focal length of the camera, we evaluate the absolute depth error for the inferred partial point clouds in table. 1-Left Column. Notice that our

|             | (a) Standard ISM<br>Average Perf.=49.4% | (b) DEHV w/o depth<br>Average Perf.=61.0% | (c) DEHV w/ depth<br>Average Perf.=63.0% |
|-------------|-----------------------------------------|-------------------------------------------|------------------------------------------|
| front       | .00 .27 .00 .00 .09 .09 .05 .09         | f .52 .00 .00 .08 .24 .00 .16 .00         | f .52 .00 .00 .05 .43 .00 .00 .00        |
| front-left  | .00 .06 .15 .04 .00 .15 .15 .04         | fl .00 .63 .00 .11 .00 .05 .21 .00        | fl .11 .02 .11 .00 .32 .05 .00 .00       |
| left        | .00 .00 .64 .00 .00 .04 .28 .04         | l .00 .08 .79 .08 .00 .00 .04 .00         | l .00 .04 .27 .00 .12 .00 .08 .00        |
| left-back   | .00 .07 .07 .63 .00 .00 .21 .04         | lb .00 .04 .12 .77 .00 .00 .08 .00        | lb .00 .00 .08 .71 .12 .00 .00 .08       |
| back        | .00 .25 .00 .00 .00 .00 .04 .08         | b .16 .04 .00 .12 .84 .00 .12 .00         | b .09 .00 .05 .05 .82 .00 .00 .00        |
| back-right  | .00 .09 .00 .00 .00 .00 .57 .05 .04     | br .00 .08 .00 .12 .00 .62 .15 .04        | br .04 .04 .04 .00 .19 .65 .04 .04       |
| right       | .00 .00 .17 .00 .00 .00 .00 .79 .04     | r .00 .04 .12 .12 .00 .00 .72 .00         | r .00 .00 .14 .07 .14 .00 .61 .04        |
| right-front | .00 .11 .00 .07 .00 .00 .00 .48 .33     | rf .00 .00 .10 .85 .00 .00 .25 .20        | rf .04 .00 .00 .25 .21 .00 .04 .46       |
|             | f fl l lb b br r rf                     | f fl l lb b br r rf                       | f fl l lb b br r rf                      |

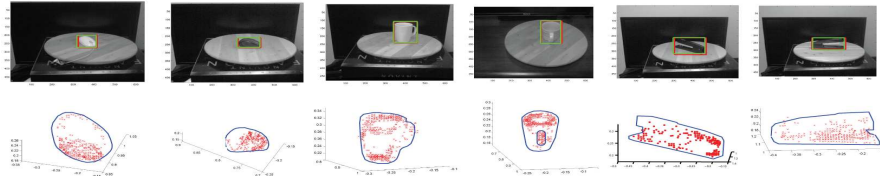
**Fig. 7.** Pose estimation results averaged across three categories. The average accuracy increases when more 3D information is available. And knowing depths in both training and testing sets gives the best performance.

errors are always lower than the baseline errors<sup>1</sup>. We also evaluate the relative depth errors<sup>2</sup> reported in table. 1-Right Column when the exact focal length is unknown. Object detection examples and inferred 3D point clouds are shown in Fig. 8.

| Table 1 | Abs. Depth in (m)<br>(known focal length) | Rel. Depth<br>(unknown focal length) |
|---------|-------------------------------------------|--------------------------------------|
|         | Sparse/Baseline                           | Sparse/Baseline                      |
| Mouse   | 0.0145/0.0255                             | 0.0173/0.0308                        |
| Mug     | 0.0176/0.0228                             | 0.0201/0.0263                        |
| Stapler | 0.0094/0.0240                             | 0.0114/0.0298                        |

Table 2. pose estimation performance on 3D object dataset[21]

| DEHV stapler | DEHV mouse | Savarese et al. '08 [22] | Farhadi et al. '09 [7] |
|--------------|------------|--------------------------|------------------------|
| 75.0         | 73.5       | 64.78                    | 78.16                  |



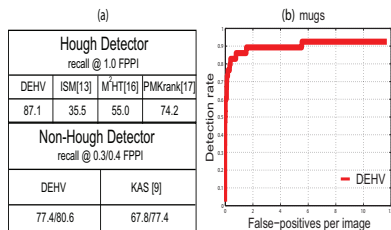
**Fig. 8.** Example of object detections (Top) and inferred 3D point clouds (Bottom). The inferred point clouds preserve the detailed structure of the objects, like the handle of the mug. Object contours are overlaid on top of the image to improve the readers understanding. Please refer to the author’s project page for a better visualization.

## 4.2 Exp.II:Comparison on three challenging datasets

In order to demonstrate that DEHV generalizes well on other publicly available datasets, we compare our results with state-of-the-art object detectors on a sub-

<sup>1</sup> It is computed assuming each depth is equal to the median of the depths of the inferred partial point clouds

<sup>2</sup>  $\frac{\|d - \hat{d}\|}{d}$  where  $d$  is the ground truth depth, and  $\hat{d}$  is the estimated depth. And  $\hat{d}$  is scaled so that  $d$  and  $\hat{d}$  have the same median.



**Fig. 9.** Performance on the mug category of ETHZ shape dataset [9]. (a-Top) Performance comparison with other pure Hough voting methods (M<sup>2</sup>HT) [16] and (PMK rank) [17]. (a-Bottom) Performance comparison between state-of-the-art non-hough voting methods [9]. (b) Detection Rate vs. FPPI of DEHV.

set of object categories from the ETHZ shape dataset, 3D object dataset, and Pascal 2007 dataset. Notice that all of these datasets contain 2D images only. Therefore, training of DEHV is performed using the 2D images from these public available dataset and the depth maps available from the 3D table-top dataset and our own set of 3D reconstruction of cars<sup>3</sup>.

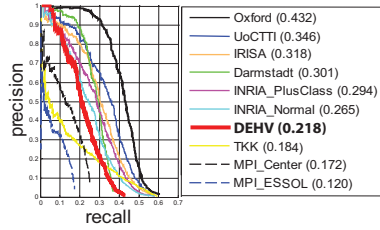
**ETHZ Shape Dataset.** We test our method on the Mug category of the ETHZ Shape dataset. It contains 48 positive images with mugs and 207 negative images with a mixture of apple logos, bottles, giraffes, mugs, and swans. Following the experiment setup in [9], we use 24 positive images and an equal number of negative images for training. We further match the 24 mugs with the mugs in 3D table-top object dataset to transfer the depth maps to the matched object instances so that we obtain augmented depth for positive training images. All the remaining 207 images in the ETHZ Shape dataset are used for testing.

The table in Fig. 9(a)-top shows the comparison of our method with the standard ISM and two state-of-the-art pure voting-based methods at 1.0 False-Positive-Per-Image (FPPI). Our DEHV method (recall 83.0 at 1 FPPI) significantly outperforms Max-Margin Hough Voting (M<sup>2</sup>HT) [16] (recall 55 at 1 FPPI) and pyramid match kernel ranking (PMK ranking) [17] (recall 74.2 at 1 FPPI). The table in Fig. 9(a)-bottom shows that our method is comparable to state-of-the-art non-voting-based method KAS [9]. Note that these results are not including a second stage verification step which would naturally boost up performance. The recall vs (FPPI) curve of our method is shown in Fig. 9(b).

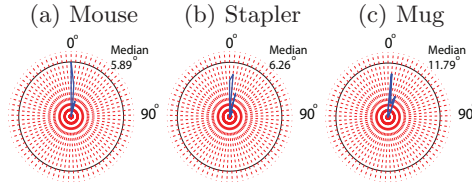
**3D object dataset.** We test our method on the mouse and stapler categories of the 3D object dataset [21,22], where each category contains 10 object instances observed under 8 angles, 3 heights, and 2 scales. We adapt the same experimental settings as [21,22] with additional depth information from the first 5 instances of the 3D table-top object dataset to train our DEHV models. The pose estimation performance of our method is shown in table.2. It is superior than [22] and comparable to [7] (which primarily focuses on pose estimation only).

**Pascal VOC 2007 Dataset.** We tested our method on the car category of the Pascal VOC 2007 challenge dataset [6], and report the localization performance. Unfortunately PASCAL does not contain depth maps. Thus, in order to train DEHV with 3D information, we collect a 3D car dataset containing 5 car instances observed from 8 viewpoints, and use Bundler [24] to obtain its 3D

<sup>3</sup> Notice that only depth is used from our own dataset.



**Fig. 10.** Object Localization result using PASCAL VOC07 dataset. The precision-recall generated by our method (red) is compared with the results of 2007 challenge [6]-Oxford, [6]-UoCTTI, [6]-IRISA, [6]-Darmstadt, [6]-INRIAPlusClass, [6]-INRIANormal, [6]-TKK, [6]-MPICenter, [6]-MPIESSOL.



**Fig. 11.** Circular histograms of 6 DOF error in degree, where 6 DOF error is defined as the angle (in degree) between the ground truth and estimated object orientation (ex: a normalized 3D vector pointing from the center to the front of an object.)

reconstruction. We match 254 car instances<sup>4</sup> in the training set of Pascal 2007 dataset to the instances in 3D car dataset and associate depth maps to these 254 Pascal training images. This way the 254 positive images can be associated to a rough depth value. Finally, both 254 positive Pascal training images and the remaining 4250 negative images are used to train our DEHV detector. We obtain reasonably good detection performance (Average Precision 0.218) even though we trained with fewer positive images (Fig. 10). Detection examples and inferred objects 3D shape are shown in Fig. 12.

### 4.3 Applications: 6 DOF pose estimation and 3D object modeling

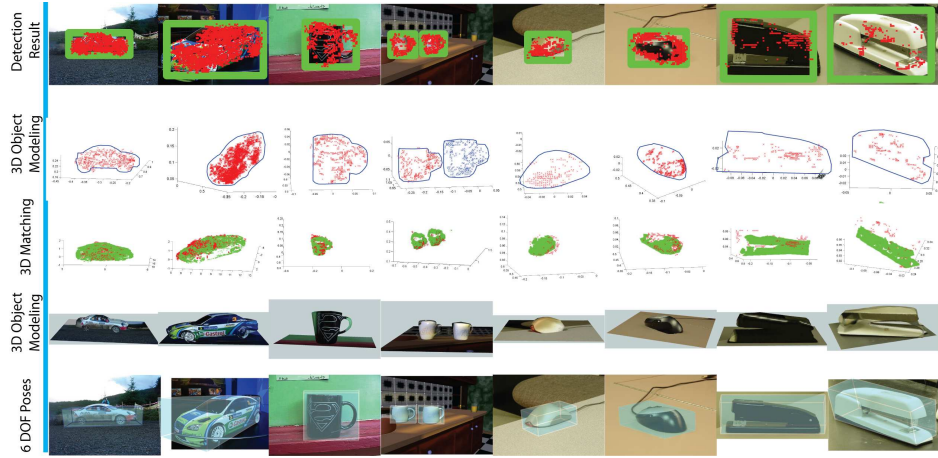
DEHV detects object classes, estimates a rough pose, and infers a partial reconstruction of the detected object. In order to robustly recover the accurate 6 DOF pose and the complete 3D shape of the object, we propose to register the inferred partial 3D point cloud (Fig. 1(c)) to a set of complete 3D CAD models<sup>5</sup>. Having estimated pose during detection allows us to highly reduce the complexity of this registration process. A modified ICP algorithm [3] is used for registration. Quantitative evaluation of 6 DOF pose estimation are shown in Fig. 11. We also obtain a full 3D object model by texture mapping the 2D image onto the 3D CAD model. Anecdotal results are reported in the 5<sub>th</sub> row of figure 12.

## 5 Conclusion

We proposed a new detection scheme called DEHV which can successfully detect objects, estimate their pose from either a single 2D image or a 2D image com-

<sup>4</sup> 254 cars is a subset of the 1261 positive images in the PASCAL training set. The subset is selected if they are easy to match with the 3D car dataset.

<sup>5</sup> The models are collected only for registration usage



**Fig. 12.** Examples of the complete 3D object inference process using the testing images from Pascal VOC07 [6], ETHZ Shape [9], and 3D object dataset [21]. This figure should be viewed in color. **Row 1** Detection results (green box) overlaid with image patch centers (red cross) which cast the votes. **Row 2** Inferred 3D point clouds (red dots), given the detection results. **Row 3** 3D registration results, where red indicates the inferred partial point clouds and green indicates the visible parts of the 3D CAD model. **Row 4** 3D Object modeling using the 3D CAD models and estimated 3D pose of the objects. Notice that the supporting plane in 3D object modeling are manually added. **Row 5** Visualizations of the estimated 6 DOF poses. (See author’s project page for 3D visualization.)

binning with depth information. Most importantly, we demonstrated that DEHV is capable of recovering the 3D shape of object categories from just one single uncalibrated image.

**Acknowledgments** We acknowledge the support of NSF (Grant CNS 0931474) and the Gigascale Systems Research Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation Entity, and Willow Garage, Inc. for collecting the 3D table-top object category dataset.

## References

1. Arie-Nachimson, M., Basri, R.: Constructing implicit 3d shape models for pose estimation. In: ICCV (2009)
2. Ballard, D.H.: Generalizing the hough transform to detect arbitrary shapes. Pattern Recognition (1981)
3. Besl, P.J., Mckay, H.D.: A method for registration of 3-d shapes. IEEE Trans. PAMI 14(2), 239–256 (1992)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)

5. Deselaers, T., Criminisi, A., Winn, J., Agarwal, A.: Incorporating on-demand stereo for real time recognition. In: CVPR (2007)
6. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results
7. Farhadi, A., Tabrizi, M.K., Endres, I., Forsyth, D.: A latent model of discriminative aspect. In: ICCV (2009)
8. Fergus, R., Perona, P., Zisserman, A.: A sparse object category model for efficient learning and exhaustive recognition. In: CVPR (2005)
9. Ferrari, V., Fevrier, L., Jurie, F., Schmid, C.: Groups of adjacent contour segments for object detection. IEEE Trans. PAMI 30(1), 36–51 (2008)
10. Gall, J., Lempitsky, V.: Class-specific hough forests for object detection. In: CVPR (2009)
11. Hoeim, D., Rother, C., Winn, J.: 3d layoutcrf for multi-view object class recognition and segmentation. In: CVPR (2007)
12. Huttenlocher, D.P., Ullman, S.: Recognizing solid objects by alignment with an image. IJCV 5(2), 195–212 (1990)
13. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: ECCV workshop on statistical learning in computer vision (2004)
14. Liebelt, J., Schmid, C., Schertler, K.: Viewpoint-independent object class detection using 3d feature maps. In: CVPR (2008)
15. Lowe, D.G.: Local feature view clustering for 3d object recognition. In: CVPR (2001)
16. Maji, S., Malik, J.: Object detection using a max-margin hough tranform. In: CVPR (2009)
17. Ommer, B., Malik, J.: Multi-scale object detection by clustering lines. In: ICCV (2009)
18. Romea, A.C., Berenson, D., Srinivasa, S., Ferguson, D.: Object recognition and full pose registration from a single image for robotic manipulation. In: ICRA (2009)
19. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3D object modeling and recognition using affine-invariant patches and multi-view spatial constraints. In: CVPR (2003)
20. Rusu, R.B., Blodow, N., Marton, Z.C., Beetz, M.: Close-range scene segmentation and reconstruction of 3d point cloud maps for mobile manipulation in human environments. In: IROS (2009)
21. Savarese, S., Fei-Fei, L.: 3D generic object categorization, localization and pose estimation. In: ICCV (2007)
22. Savarese, S., Fei-Fei, L.: View synthesis for recognizing unseen poses of object classes. In: ECCV (2008)
23. Schneiderman, H., Kanade, T.: A statistical approach to 3D object detection applied to faces and cars. In: CVPR (2000)
24. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: SIGGRAPH (2006)
25. Su, H., Sun, M., Fei-Fei, L., Savarese, S.: Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In: ICCV (2009)
26. Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Van Gool, L.: Using multi-view recognition and meta-data annotation to guide a robot’s attention. Int. J. Rob. Res. (2009)
27. Yan, P., Khan, D., Shah, M.: 3d model based object class detection in an arbitrary view. In: ICCV (2007)