Iterative Hough Forest with Histogram of Control Points for 6 DoF Object Registration from Depth Images

Caner Sahin, Rigas Kouskouridas and Tae-Kyun Kim *

Abstract

State-of-the-art techniques for 6D object pose recovery depend on occlusion-free point clouds to accurately register objects in the 3D space. To reduce this dependency, we introduce a novel architecture called *Iterative Hough forest with Histogram of Control Points* that is capable of estimating occluded and cluttered objects' 6D pose given a candidate 2D bounding box. Our *iterative Hough forest* is learnt using patches extracted only from the positive samples. These patches are represented with *Histogram of Control Points (HoCP)*, a "scale-variant" implicit volumetric description, which we derive from recently introduced Implicit B-Splines (IBS). The rich discriminative information provided by this scale-variance is leveraged during inference, where the initial pose estimation of the object is iteratively refined based on more discriminative control points by using our *iterative Hough forest*. We conduct experiments on several test objects of a publicly available dataset to test our architecture and to compare with the state-of-the-art.

1 Introduction

Object registration is an important task in computer vision that determines the translation and the rotation of an object with respect to a reference coordinate frame. By utilizing such a task, one can propose promising solutions for various problems related to robotics [Kouskouridas et al., 2014], scene understanding, augmented reality *etc*. Recent developments on visual depth sensors and their increasing ubiquity have allowed researchers to make use of the information acquired from these devices to facilitate the registration.

When the target point cloud is cleanly segmented, Iterative Closest Point (ICP) algorithm [Besl and McKay, 1992], point-to-model based methods [Rouhani and Sappa, 2011, Rouhani and Domingo Sappa, 2013, Doumanoglou et al., 2015, Kouskouridas et al., 2016, 2013] and point-to-point techniques [Rusu et al., 2009, Kim and Medioni, 2011] demonstrate good results. However, the performance of these approaches is severely degraded by the challenges such as heavy occlusion and clutter, and similar looking distractors. In order to address these challenges, several learning based methods formulate occlusion aware features Bonde et al. [2014], derive patch-based (local) descriptors [Tejani et al., 2014] or encode the contextual information of the objects with simple depth pixels [Brachmann et al., 2014] and integrate into random forests. Particularly, iterative random forest algorithms such as Latent-Class Hough forest (LCHF) [Tejani et al., 2014] and iterative Multi-Output Random forest (iMORF) [Zhao et al., 2014] obtain the state-of-the-art accuracy on pose estimation. On the other hand, these methods rely on the scale-invariant features and the exploitation of the rich discriminative information that is inherently embedded into the scale-variability is one

*All authors are with the Imperial Computer Vision and Learning Lab (ICVL), at the Department of Electrical and Electronic Engineering, Imperial College London, UK, {c.sahin14, r.kouskouridas, tk.kim}@imperial.ac.uk

Submitted to IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016), Daejeon, Korea.

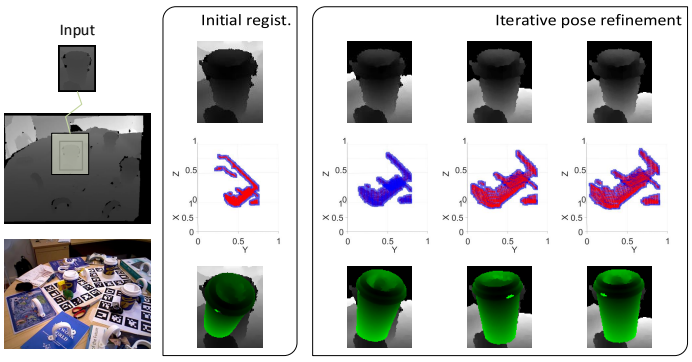


Figure 1: Sample result of our system: the *initial registration* roughly aligns the test object and the *iterative pose refinement* further refines this alignment (The RGB correspondence is for visualization purposes).

important point overlooked.

Unlike the aforementioned learning-based methods, Novatnack *et al.* [Novatnack and Nishino, 2008, Bariya and Nishino, 2010] utilize the detailed information of the scale variation in order to register the range images in a coarse-to-fine fashion. They extract and match conventional salient 3D key points. However, real depth sensors have several imperfections such as missing depth values, noisy measurements, foreground fattening, *etc*. Salient feature points tend to be located on these deficient parts of the depth images, and hence, they are rather unstable [Zach et al., 2015]. In such a scenario, 3D reconstruction methods that provide more reliable shape information can be utilized [Bonde et al., 2014]. Implicit B-Splines (IBS) [Rouhani and Domingo Sappa, 2013, Rouhani et al., 2015] are yet other techniques that can provide shape descriptors through their zero-sets and reconstruct surfaces. These techniques are based on the locally controlled functions that are combined via their control points and this local control allows patch-based object representation.

Our architecture is originated from these observations. We integrate the coarse-to-fine registration approach presented in [Novatnack and Nishino, 2008] into the random forests [Tejani et al., 2014, Zhao et al., 2014] using the Histogram of the Control Points (HoCP) that we adapt from recently introduced IBSs [Rouhani et al., 2015]. We train our forest only from positive samples and learn the detailed information of the scale-variability during training. We normalize every training point cloud into a unit cube and then generate a set of scale-space images, each of which is separated by a constant factor. The patches extracted from the images in this set are represented with the scale-variant HoCP features. During inference, the patches centred on the pixels that belong to the background and foreground clutters are removed iteratively using the most confident hypotheses and the test image is updated. Since this removal process decreases the standard deviation of the test point cloud, subsequent normalization applied to the updated test image increases the relative scale of the object (foreground pixels) in the unit cube. More discriminative descriptors (control points) are computed in higher scales and this ensures further refinement of the object pose. Note that we employ a compositional approach, that is, we concurrently detect the object in the target region and estimate its pose by aligning the patches in order to increase robustness across clutter. Figure 1 depicts a sample result of our architecture. To summarize, our main contributions are as follows:

- To the best of our knowledge this is the first time we adapt an implicit object representation, Implicit B-Spline, into a "scale-variant" patch descriptor and associate with the random forests.
- We introduce a novel iterative algorithm for the Hough forests: it finds out an initial hypothesis and improves its confidence iteratively by extracting more discriminative "scale-variant" descriptors due to the elimination of the background/foreground clutter.

2 Related Work

A large number of studies have been proposed for the object registration, ranging from the point-wise correspondence based methods to the learning based approaches. Iterative Closest Point (ICP)

algorithm, originally presented in [Besl and McKay, 1992], requires a good initialization in order not to be trapped in a local minimum during fine tuning. This requirement is reduced in [Yang et al., 2013] providing globally optimal registration by the integration of a global branch-and-bound (BnB) optimization into the local ICP. The point-wise correspondence problem is converted into a point-to-model registration in [Rouhani and Sappa, 2011, Rouhani and Domingo Sappa, 2013]. The object model is represented with implicit polynomials (IP) and the distance between the test point set and this model is minimized via the Levenberg-Marquardt algorithm (LMA). The study [Zheng et al., 2008] that utilizes 3D IPs for 6 DoF pose estimation on ultrasound images is further extended in [Zheng et al., 2013] by a coarse-to-fine IP-driven registration strategy. The point-to-point techniques build point-pair features for sparse representations of the test and the model point sets [Choi et al., 2012]. Rusu *et al.* align two noisy point clouds of real scenes by finding correct point-to-point correspondences between the Point Feature Histograms (PFH) and feed this alignment to an ICP algorithm for fine tuning [Rusu et al., 2008]. The votes of the matching features are accumulated in [Drost et al., 2010] to hypothesise the poses of the cluttered and partially occluded objects. Choi *et al.* [Choi and Christensen, 2012] propose point-pair features for both RGB and depth channels and they are conducted in a voting scheme to hypothesise the rotation and translation parameters of the objects in the cluttered scenes. Despite achieving good registration results, these techniques underperform when the scenes are under heavy occlusion and clutter, and the target objects' geometry are indistinguishable from background clutter.

Learning-based methods have good generalization across severe occlusion and clutter. The state-of-the-art accuracy on registration is acquired by the iterative random forest algorithms, particularly [Tejani et al., 2014] and [Zhao et al., 2014], which form a basis for our *iterative Hough forest* architecture. Tejani's patch-based strategy [Tejani et al., 2014] refines the initially hypothesised object pose by iteratively updating the object class distributions in the leaf nodes during testing. Iterative Multi Output Random forest (iMORF) [Zhao et al., 2014] jointly predicts the head pose, the facial expression and the landmark positions. The relations between these tasks are modelled so that their performances are iteratively improved with the extraction of more informative features. Whilst these approaches rely on the scale-invariant features to improve the confidence of a pose hypothesis, inspired by [Novatnack and Nishino, 2008], we design scale-variant features getting more discriminative with the increase in the scale. Novatnack et al. [Novatnack and Nishino, 2008, Bariya and Nishino, 2010] introduce a framework that registers the range images in a coarse-to-fine fashion by utilizing the detailed information provided by the scale variation. The shape descriptors with the coarsest scale are matched initially and a rough alignment is achieved since fewer features are extracted in coarser scales. The descriptor matching at higher scales results improved predictions of the pose.

3 Our Registration Approach

In this section we detail our registration approach by firstly describing the computation procedure of the HoCP features. We then present how to encode the discriminative information of these scale-variant features into the forest. Finally, we demonstrate how to exploit the learnt shape information in a coarse-to-fine fashion to refine the pose hypotheses.

3.1 Histogram of Control Points (HoCP)

We demonstrate the computation procedure of the HoCP features over a positive depth image selected from the training dataset. It is initially normalized into a unit cube and then new point clouds at different scales are sampled as follows:

$$\{\mathbf{X}_N\}_i = \frac{\mathbf{X}_{n\times 3} - \bar{\mathbf{X}}_{n\times 3}}{s_i * \alpha} + 0.5, \quad i = 0, 1, 2, ..., m \tag{1}$$

with

$$\alpha = \max \left\{ \begin{array}{c} \max(X)\text{-}\min(X) \\ \max(Y)\text{-}\min(Y) \\ \max(Z)\text{-}\min(Z) \end{array} \right\}, h_i = \max(Z_{N_i}) - \min(Z_{N_i}) \tag{2}$$

where $\mathbf{X} = [XYZ]$ is the world coordinate vector of the original foreground point cloud, $\bar{\mathbf{X}}$ is the mean of $\mathbf{X}$, $\mathbf{X}_N = [X_N Y_N Z_N]$ is the normalized foreground pixels, m is the number of the scales, α

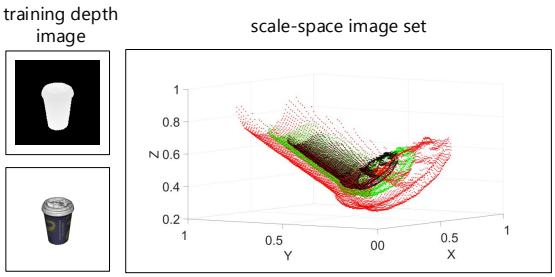


Figure 2: The initial normalization ($s_0 = 1$) of the training depth image is the outmost (red) point cloud and the inner ones (green and black) are sampled by different s_i values. The point clouds in this set are used in Fig. 3 to further explain how to compute the descriptors.

is the scale factor and h is the scale. The constant s_i takes real numbers to generate the point clouds at different scales, starting from $s_0 = 1$ that corresponds to the initial normalization. A training image and its samples at different scales are shown in Fig. 2.

Once we generate a set of scale-space images (Fig. 2), we represent these point clouds with the control point descriptors first globally. The descriptor computation procedure is the same as presented in [Rouhani et al., 2015]. The unit cube is split into an $N \times N \times N$ voxel grid where N is the IBS resolution. Each descriptor Γ is defined with an index-weight pair: the index number indicates the vertex of this grid at which the related control point is located. The weight informs the descriptor significance about the control of the geometry to be represented. The scale-space images in Fig. 2 are globally represented in Fig. 3 (a). We partition the global representation at each scale into patches. We express the patch size g in image pixels and it is a constant that depicts the ratio between the sizes of the extracted patch and the bounding box of the global point cloud. A window with the specified patch size is traversed in the unit cube of each scale-space image and the patches are extracted around non-zero pixels. Each patch has its own implicit volumetric representation, formed by the closest control points to the patch center, the ones lying inside the window along depth direction. The patches sampled at different scales in Fig. 3 (b) represent the same shape. However, their volumetric descriptions (blue) are getting more discriminative as the scale increases, since the greater number of descriptors are computed at higher scales. We encode this discriminative information into histograms in spherical coordinates. Each of the patch centres is coincided with the center of a sphere. The control points of the patch are described by the log of the radius t_r, the cosine of the inclination t_θ and the azimuth t_ϕ. Then, the sphere is divided into the bins and the relation between the bin numbers h_r, h_θ, h_ϕ and the histogram coordinates t_r, t_θ, t_ϕ is given as follows [Kroon, 2011]:

$$
\begin{aligned}
t_r &= \frac{h_r}{log(\frac{r_{max}}{r_{min}})} log(\frac{r}{r_{min}}) \\
t_\theta &= h_\theta \frac{z}{r} \\
t_\phi &= \frac{h_\phi}{2\pi} tan^{-1}(\frac{y}{x})
\end{aligned}
\tag{3}
$$

where r_{min} and r_{max} are the radii of the nested spheres with the minimum and the maximum volumes, x, y, z are the Cartesian coordinates of each descriptor with radius r. r_{max} equals to the distance between the patch center and the farthest descriptor of the related patch. The numbers of the control points in each bin are counted and stored in a $d = h_r * h_\theta * h_\phi$ dimensional feature vector $\mathbf{f}$. The volumetric descriptions in Fig. 3 (b) are shown with their related histograms in Fig. 3 (c). Thus, the sample shape (patch) is represented with the scale-variant HoCP features.

3.2 The Combination of HoCP and iterative Hough Forest

The proposed iterative Hough forest is the combination of randomized binary decision trees. It is trained only on foreground synthetically rendered depth images of the object of interest. We generate a set of scale-space images from each training point cloud and sample a set of annotated patches $\{\cup_{i=1}^{p} P_i\}$ as follows [Tejani et al., 2014]:

$$
\mathcal{P} = \{\cup_{i=1}^{p} P_i\} = \{\cup_{i=1}^{p} (\mathbf{c}_i, \Delta\mathbf{x}_i, \theta_i, \mathbf{f}_i, D_i)\} \tag{4}
$$

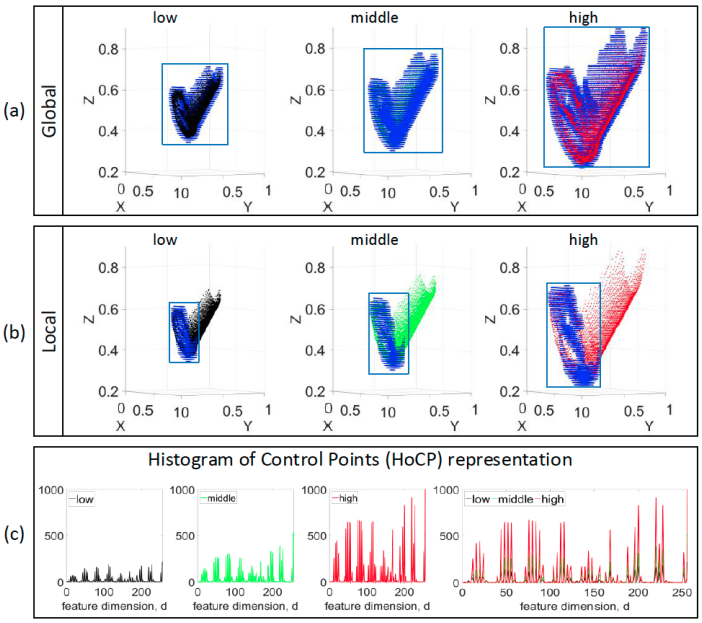


Figure 3: *Histogram of Control Points:* the scale-space images are globally represented in (a) and the same shape (patch) is extracted at each scale in (b). The HoCP representation of this shape is shown in (c). The scales are encoded as low, middle and high.

where $\mathbf{c}_i = (c_{x_i}, c_{y_i})$ is the patch centre in pixels, $\Delta\mathbf{x}_i = (\Delta x_i, \Delta y_i, \Delta z_i)$ is the 3D offset between the centres of the patch and the object, $\theta_i = (\theta_{r_i}, \theta_{p_i}, \theta_{y_i})$ is the rotation parameters of the point cloud from which the patch P_i is extracted and D_i is the depth map of the patch.
Each tree is constructed by using a subset $\mathcal{S}$ of the annotated training patches $\mathcal{S} \subset \mathcal{P}$. We randomly select a template patch T from $\mathcal{S}$ and assign it to the root node. We measure the similarity between T and each patch S_i in $\mathcal{S}$ as follows:

- **Depth check:** The depth values of the descriptors S_{i_Γ} and T_Γ that represent the patches S_i and T are checked, and the spatially inconsistent ones in S_{i_Γ} are removed as in [Tejani et al., 2014], generating Ω that includes the spatially consistent descriptors of the patch S_i.
- **Similarity measure:** Using Ω, the feature vector $\mathbf{f}_\Omega$ is generated and the $\mathcal{L}_2$ norm between this vector and $\mathbf{f}_T$ is measured:

$$\mathcal{F}(S_i, T) = \parallel \mathbf{f}_\Omega - \mathbf{f}_T \parallel_2 \tag{5}$$

- **Similarity score comparison:** Each patch is passed either to the left or the right child nodes according to the split function that compares the score of the similarity measure $\mathcal{F}(S_i, T)$ and a randomly chosen threshold τ.

The main reason why we apply a depth check to the patches is to remove the structural perturbations, due to occlusion, clutter [Tejani et al., 2014]. These perturbations most likely occur on the patches extracted along depth discontinuities such as the contours of the objects of interest. They cause to diverge a test patch (occluded/cluttered) from its positive correspondence by changing its representation, r_{max} of the sphere, and the histogram coordinates consequently.

A group of candidate split functions are produced at each node by using a set of randomly assigned patches $\{T_i\}$ and thresholds $\{\tau_i\}$. The one that best optimize the offset and pose regression entropy [Fanelli et al., 2011] is selected as the split function. Each tree is grown by repeating this process recursively until the forest termination criteria are satisfied. When the termination conditions are met, the leaf nodes are formed and they store votes for both the object center $\Delta\mathbf{x} = (\Delta x, \Delta y, \Delta z)$ and the object rotation $\theta = (\theta_r, \theta_p, \theta_y)$.

3.3 Initial Registration and Iterative Pose Refinement

The proposed architecture registers the object of interest in two steps: the *initial registration* and the *iterative pose refinement*. The *initial registration* roughly aligns the test object and this alignment is

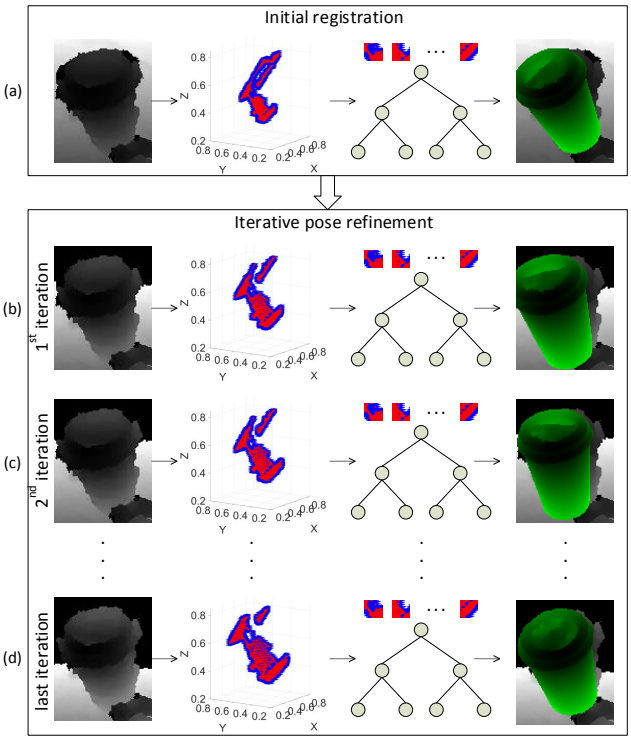


Figure 4: The initial pose estimation of the test object is iteratively refined based on the more discriminative control points that are extracted due to the elimination of background/foreground clutter.

further improved by the *iterative pose refinement*.
Consider an object that was detected by a coarse bounding box, I_b, as shown in the leftmost image of Fig. 4 (a). At an iteration instant k, the following quantities are defined:

- $\Delta\mathbf{x}^{0:k} = \{\Delta\mathbf{x}^0, \Delta\mathbf{x}^1, ..., \Delta\mathbf{x}^k\} = \{\Delta\mathbf{x}^0, \Delta\mathbf{x}^{1:k}\}$: the history of the object position.
- $\theta^{0:k} = \{\theta^0, \theta^1, ..., \theta^k\} = \{\theta^0, \theta^{1:k}\}$: the history of the object rotation.
- $V^{1:k} = \{v^1, v^2, ..., v^k\}$: the history of the inputs (noise removals) applied to the test image.
- $m^{0:k} = \{m^0, m^1, ..., m^k\} = \{m^0, m^{1:k}\}$: the history of the set of the feature vectors where $m^k = \{\cup_{i=1}^{n} \mathbf{f}_i\}$.
- h^k: the object scale (the scale of the foreground pixels) in the unit cube at iteration k (see Eq. 2).

We formulate the *initial registration* as follows:

$$(\Delta\mathbf{x}^0, \theta^0) = \arg\max_{\Delta\mathbf{x}^0, \theta^0} p(\Delta\mathbf{x}^0, \theta^0 | I_b, m^0, h^0). \tag{6}$$

We find the best parameters that maximize the joint posterior density of the initial object position $\Delta\mathbf{x}^0$ and the initial object rotation θ^0. This initial registration process is illustrated in Fig. 4 (a). The test image is firstly normalized into a unit cube. Unlike training, this is a "single" scale normalization that corresponds to $s_0 = 1$ (see Eq. 1). The patches extracted from the globally represented point cloud are described with the HoCP features and passed down all the trees. We determine the effect that all patches have on the object pose by accumulating the votes stored in the leaf nodes as in [Tejani et al., 2014] and approximate the initial registration given in Eq. 6. Once the initial hypothesis $\mathbf{x}^0 = (\Delta\mathbf{x}^0, \theta^0)$ is obtained, the pixels that belong to the background/foreground clutter $\{\cup_{j=0}^{f} p_j\}$ are removed from I_b according to the following criterion:

$$v^k = \begin{cases} I_b(p_j) = \mathcal{D}_{I_b}(p_j), & \gamma\psi_1 < \mathcal{D}_{I_b}(p_j^k) < \beta\psi_2 \\ I_b(p_j) = 0, & otherwise \end{cases} \tag{7}$$

with

$$\gamma = \min(\mathcal{D}_H^k), \quad \beta = \max(\mathcal{D}_H^k) \tag{8}$$

where $\mathcal{D}_H^k$ and $\mathcal{D}_{I_b}$ are the depth maps of the hypothesis H at iteration k, and of the I_b, ψ_1 and ψ_2 are the scaling coefficients. The efficacy of v^k is illustrated in Fig. 4. In the rightmost image of Fig. 4 (a), the test image and the initial hypothesis are overlaid. This hypothesis is exploited and the test image is updated by v^1 as in Eq. 7. This updated image is shown in Fig. 4 (b) and assigned as input for the 1^{st} iteration. It is normalized and represented globally. Note how the object "scale" (h^1) in the unit cube is relatively increased and more discriminative descriptors m^1 are extracted (compare with the initial registration). This is mainly because of that the standard deviation of the input image is decreased since we removed foreground/background clutter. The resultant hypothesis of the 1^{st} iteration is shown on the right. The extraction of more discriminative descriptors and the noise removal process result more accurate and confident hypothesis. This pose refinement process is iteratively performed until the maximum iteration is reached (see Fig. 4 (d)):

$$(\Delta\mathbf{x}^k, \theta^k) = \arg\max_{\Delta\mathbf{x}^k, \theta^k} p(\Delta\mathbf{x}^k, \theta^k \mid m^{1:k}, V^{1:k}, \mathbf{x}^0, h^k) \tag{9}$$

We approximate the registration hypothesis at each iteration by using the stored information in the leaf nodes as we do in the initial registration.

4 Experimental Results

We have analysed the ICVL dataset [Tejani et al., 2014] and have found that the "coffee cup" and the "camera" are some of the best demonstrable objects to test and compare our registration architecture with the state-of-the-art methods since they are located in highly occluded and cluttered scenes. We further process the test images of these objects to generate a new test dataset according to the following criteria:

- Since the HoCP features are scale-variant, the depth values of the training and the test images should be close to each other up to a certain degree. In this study, we train the forests at a single depth value, 750 mm, and test with the images at the range of $[750 \mp 35]$ mm.
- The test object instances located at the range of $[750 \mp 35]$ mm are assumed as detected by coarse bounding boxes (see Fig. 1). The image regions included in these bounding boxes are cropped (see Fig. 4) and the new test dataset is generated (276 "coffee cup" and 360 "camera" RGBD test images).

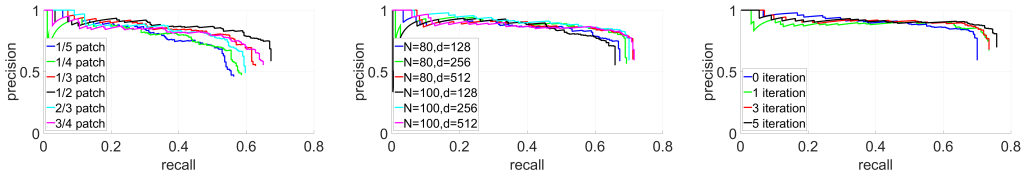
The maximum depth is 25 and the number of the maximum samples at each leaf node is 15 for each tree. Every forest is the ensemble of 3 trees with these termination criteria. Our experiments are two folds: *intraclass* and *interclass*. Both experiments use the metric proposed in [Hinterstoisser et al., 2012] to determine whether a registration hypothesis is correct. This metric outputs a score ω that calculates the distance between the ground truth and estimated poses of the test object. The registration hypothesis that ensures the following inequality is considered as correct:

$$\omega \leq z_\omega \Phi \tag{10}$$

where Φ is the diameter of the 3D model of the test object and z_ω is a constant that determines the coarseness of an hypothesis that is assigned as correct. We set z_ω to 0.08 in the intraclass and interclass experiments.

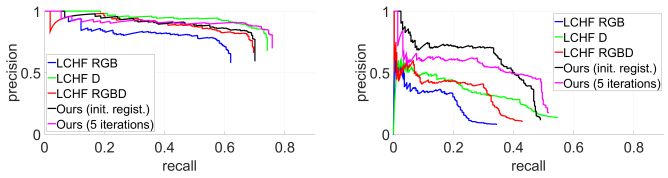
4.1 Intraclass Experiments

These experiments are performed on the "coffee cup" dataset to determine the optimal parameters of the proposed approach. The effect of the patch size g is firstly examined by setting the IBS resolution N to 80, the HoCP feature dimension d to 128 in addition to the previously defined forest parameters. We test the patch sizes $g = \{0.20, 0.25, 0.33, 0.50, 0.66, 0.75\}$. The resultant Precision-Recall (PR) curves are shown in Fig. 5 (a). When we increase the patch size until it is 0.5 times of the bounding box, the registration performance is improved since the greater patches can encode more discriminative shapes. We continue to extend the patch size till it is 0.75 times of the bounding box and observe that the performance is degraded since these patches tend to contain the noisy parts of the scene. According to this figure and their corresponding F1 scores (see Table 1), we choose $\frac{1}{2}$ as



(a) (b) (c)

Figure 5: PR curves obtained from the intraclass experiments. According to these results we choose $\frac{1}{2}$ patch size and set $N = 100$, $d = 256$. For the corresponding F1 scores see the Table 1.



(a) (b)

Figure 6: PR curves of the "coffee cup" (left) and the "camera" (right) dataset obtained from the interclass experiments: Each image compares our method (initial registration and iterative pose refinement) with the LCHFs trained separately on the RGB, D and RGB-D channels. The F1 scores are presented in Table 2.

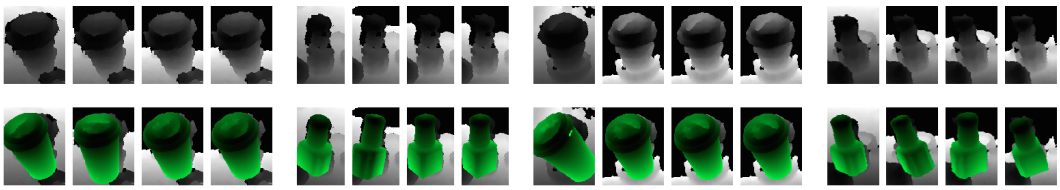


Figure 7: Some qualitative results. For each octonary: the 1^{st} column illustrates the test image and the initial hypothesis (*initial registration*) and the remaining columns demonstrate the 1^{st}, the 3^{rd} and the 4^{th} iterations (*iterative pose refinement*). The test images are updated by removing the background/foreground clutter.

the optimal patch size.

Using the selected patch size, we next tune the IBS resolution N and the HoCP feature dimension d. We test the combinations of $N = \{80, 100\}$ and $d = \{128, 256, 512\}$, the ones that are the most applicable $N - d$ pairs to represent $\frac{1}{2}$ patch size. The PR curves of these combinations are depicted in Fig. 5 (b) and the corresponding F1 scores are illustrated in Table 1. We take into account both the memory consumption and the accuracy, and agree on the values of $N = 100$ & $d = 256$.The last parameter we test in the intraclass experiments is the iteration number. We test several *iterative Hough forests with Histogram of Control Points* each of which has $k = 0, 1, 3$, and 5 iterations, respectively. Their PR curves are shown in Fig. 5 (c). As expected, the forests that use greater number of iterations show better performances (see Table 1) since more discriminative features are extracted thanks to the noise removal process.

4.2 Interclass Experiments

These experiments are conducted on the "coffee cup" and the "camera" datasets to compare our approach with the state-of-the-art methods including the Latent-Class Hough forests (LCHF) [Tejani et al., 2014] trained separately on the color gradient (LCHF-RGB), the surface normal (LCHF-Depth) and the color gradient + the surface normal (LCHF-RGBD) features. In order to make a fair comparison between methods, we train and test these versions of the LCHF by using the authors' software. The forest parameters are the same as our own approach.

According to the F1 scores in Table 2, we observe that the LCHF trained on the color gradient features underperforms other methods. The main reason of this underperformance is the distortion along the object borders arising from the occlusion and the clutter, that is, the distortion of the color gradient information in the test process. When we train the LCHF by only using the

Table 1: F1 scores determined for different patch sizes, IBS resolution (N) & feature dimension (d) and number of iteration

Patch Size	F1 Score	N & d	F1 Score
$\frac{1}{5}$	0.5966	80 & 128	0.7068
$\frac{1}{4}$	0.6096	80 & 256	0.7368
$\frac{1}{3}$	0.6532	80 & 512	0.7425
$\frac{1}{2}$	**0.7068**	100 & 128	0.6870
$\frac{2}{3}$	0.6341	**100 & 256**	**0.7510**
$\frac{3}{4}$	0.6539	100 & 512	0.7438

# iter	F1 Score
0	0.7510
1	0.7742
3	0.7745
5	**0.7932**

Table 2: F1 scores of the "coffee cup" and the "camera" datasets are shown. In both datasets our approach with iterative pose refinement outperforms.

Approach	Coffee Cup	Camera
LCHF-RGB	0.6595	0.2478
LCHF-Depth	0.7860	0.3386
LCHF-RGBD	0.7390	0.3456
Ours (init. reg.)	0.7510	0.4534
Ours (iter. pose ref.)	**0.7932**	**0.4693**

depth information, we infer that the surface normals outperform the color gradients. The combined utilization of the color gradients and the surface normals in the LCHF produces approximately the same results as the LCHF-Depth. Our approach with the iterative pose refinement outperform other methods. Regarding the 'camera' object, we observe that the registration performances of all methods are relatively decreased. This is mainly because of that this object has large amount of missing depth pixels in addition to severe occlusion and clutter. Figure 7 illustrates several qualitative results of our approach on the camera and the coffee cup objects.

5 Conclusion

In this study, we have proposed a novel architecture, *iterative Hough forest with Histogram of Control Points*, for 6 DoF object registration from depth images. We have introduced the Histogram of the Control Points, a scale-variant patch representation, and have encoded their rich discriminative information into the random forests. We train our forest using only the positive samples. During testing, we first roughly align the object and then iteratively refine this alignment. The experimental results report that our approach show better registration performance than the state-of-the-art methods. In the future, we plan to engineer a variable patch size approach and integrate it into the proposed iterative Hough forest architecture for further exploitation of the rich discriminative information provided by the HoCP features.

References

Prabin Bariya and Ko Nishino. Scale-hierarchical 3d object recognition in cluttered scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1657–1664. IEEE, 2010.

Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

Ujwal Bonde, Vijay Badrinarayanan, and Roberto Cipolla. Robust instance recognition in presence of occlusion and clutter. In *Computer Vision–ECCV 2014*, pages 520–535. Springer, 2014.

Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision–ECCV 2014*, pages 536–551. Springer, 2014.

Changhyun Choi and Henrik I Christensen. 3d pose estimation of daily objects using an rgb-d camera. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 3342–3349. IEEE, 2012.

Changhyun Choi, Yuichi Taguchi, Oncel Tuzel, Ming-Yu Liu, and Srikumar Ramalingam. Voting-based pose estimation for robotic assembly using a 3d sensor. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1724–1731. IEEE, 2012.

Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. 6d object detection and next-best-view prediction in the crowd. *arXiv preprint arXiv:1512.07506*, 2015.

Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010.

Gabriele Fanelli, Juergen Gall, and Luc Van Gool. Real time head pose estimation with random regression forests. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 617–624. IEEE, 2011.

Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Computer Vision–ACCV 2012*, pages 548–562. Springer, 2012.

Eunyoung Kim and Gerard Medioni. 3d object recognition in range images using visibility context. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 3800–3807. IEEE, 2011.

Rigas Kouskouridas, Antonios Gasteratos, and Christos Emmanouilidis. Efficient representation and feature extraction for neural network-based 3d object pose estimation. *Neurocomputing*, 120: 90–100, 2013.

Rigas Kouskouridas, Kostantinos Charalampous, and Antonios Gasteratos. Sparse pose manifolds. *Autonomous Robots*, 37(2):191–207, 2014.

Rigas Kouskouridas, Alykhan Tejani, Andreas Doumanoglou, Danhang Tang, and Tae-Kyun Kim. Latent-class hough forests for 6 dof object pose estimation. *arXiv preprint arXiv:1602.01464*, 2016.

Dirk-Jan Kroon. *Segmentation of the mandibular canal in cone-beam CT data*. Citeseer, 2011.

John Novatnack and Ko Nishino. Scale-dependent/invariant local 3d shape descriptors for fully automatic registration of multiple sets of range images. In *Computer Vision–ECCV 2008*, pages 440–453. Springer, 2008.

Mohammad Rouhani and Angel Domingo Sappa. The richer representation the better registration. *Image Processing, IEEE Transactions on*, 22(12):5036–5049, 2013.

Mohammad Rouhani and Angel D Sappa. Correspondence free registration through a point-to-model distance minimization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2150–2157. IEEE, 2011.

Mohammad Rouhani, Angel D Sappa, and Edmond Boyer. Implicit b-spline surface reconstruction. *Image Processing, IEEE Transactions on*, 24(1):22–32, 2015.

Radu Bogdan Rusu, Nico Blodow, Zoltan Csaba Marton, and Michael Beetz. Aligning point cloud views using persistent feature histograms. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3384–3391. IEEE, 2008.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. IEEE, 2009.

Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class hough forests for 3d object detection and pose estimation. In *Computer Vision–ECCV 2014*, pages 462–477. Springer, 2014.

Jiaolong Yang, Hongdong Li, and Yunde Jia. Go-icp: Solving 3d registration efficiently and globally optimally. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1457–1464, 2013.

Christopher Zach, Adrian Penate-Sanchez, and Minh-Tri Pham. A dynamic programming approach for fast and robust object pose recognition from range images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 196–203, 2015.

Xiaowei Zhao, Tae-Kyun Kim, and Wenhan Luo. Unified face analysis by iterative multi-output random forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1772, 2014.

Bo Zheng, Ryo Ishikawa, Takeshi Oishi, Jun Takamatsu, and Katsushi Ikeuchi. 6-dof pose estimation from single ultrasound image using 3d ip models. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

Bo Zheng, Ryo Ishikawa, Jun Takamatsu, Takeshi Oishi, and Katsushi Ikeuchi. A coarse-to-fine ip-driven registration for pose estimation from single ultrasound image. *Computer Vision and Image Understanding*, 117(12):1647–1658, 2013.