OUR-CVFH – Oriented, Unique and Repeatable Clustered Viewpoint Feature Histogram for Object Recognition and 6DOF Pose Estimation

Aitor Aldoma[1], Federico Tombari[2], Radu Bogdan Rusu[3], and Markus Vincze[1]

[1] Vision4Robotics Group, ACIN, Vienna University of Technology
[2] Computer Vision Lab., DEIS - ARCES, University of Bologna
[3] Open Perception Inc.

Abstract. We propose a novel method to estimate a unique and repeatable reference frame in the context of 3D object recognition from a single viewpoint based on global descriptors. We show that the ability of defining a robust reference frame on both model and scene views allows creating descriptive global representations of the object view, with the beneficial effect of enhancing the spatial descriptiveness of the feature and its ability to recognize objects by means of a simple nearest neighbor classifier computed on the descriptor space. Moreover, the definition of repeatable directions can be deployed to efficiently retrieve the 6DOF pose of the objects in a scene. We experimentally demonstrate the effectiveness of the proposed method on a dataset including 23 scenes acquired with the Microsoft Kinect sensor and 25 full-3D models by comparing the proposed approach with state-of-the-art global descriptors. A substantial improvement is presented regarding accuracy in recognition and 6DOF pose estimation, as well as in terms of computational performance.

1 Introduction and Related Work

Recognizing free-form shapes in clutter and occlusion is currently one of the most ambitious and challenging task in the field of 3D computer vision, given the typical distortions which 3D data undergoes due to noisy sensors, viewpoint changes and point density variations. This task, often recalled as *3D object recognition*, is usually carried out together with 3D pose estimation, which requires to compute the 6 degree-of-freedom (DOF) transformation between the current model being recognized and its instance in the scene under analysis. The main advantages of performing object recognition in the 3D space rather than on the image plane are the higher discriminative capabilities towards objects characterized by low informative content in terms of appearance (e.g. low texture), as well as the possibility of directly recovering the full 6DOF pose of the object. 3D object recognition is now a key step for several application scenarios, such as robot grasping and manipulation, scene understanding and place recognition, human-robot interaction.

Recently, research in the field of 3D object recognition has been fostered not only by the development of the aforementioned scenarios, but also by the availability of low-cost, real-time 3D sensors such as the Microsoft Kinect and the

A. Pinz et al. (Eds.): DAGM/OAGM 2012, LNCS 7476, pp. 113–122, 2012.

Asus Xtion. Several algorithms have been proposed in literature in the past few years, which can be divided between *local* [5, 6, 8] and *global* approaches [1, 3, 9]. On one side, local algorithms extract repeatable *keypoints* on the 3D surface of models and scene [5], then associating each keypoint with a *description* of its local neighborhood [4, 5, 8], so that, by means of descriptor matching, reliable scene-to-model point correspondences can be determined. This set of correspondences is then usually clustered [5,6] by enforcing geometrical constraints derived from the assumption of rigid transformations between the model and the scene, each correspondence cluster defining a model *hypothesis*, i.e. a subset of correspondences holding consensus for a 6DOF pose of a specific model within the library.

Global approaches compute one single descriptor for each object encompassing the whole object surface. Examples of global descriptors are the Clustered Viewpoint Feature Histogram (CVFH) [1], the Viewpoint Feature Histogram (VFH) [7], Ensemble of Shape Functions (ESF) [9] and the Global Radius-based Surface Descriptors (GRSD) [3]. Obviously, global descriptors can not be directly applied on cluttered scenes, which need to undergo a proper 3D pre-segmentation stage aimed at localizing possible object instances. By means of descriptor matching, each 3D segment extracted on the scene is then associated to a model of the library, yielding model hypotheses. Hence, although less effective in presence of partial object occlusions, the global approach is characterized by a smaller complexity in the description and matching stage with respect to local methods, since each surface is characterized by one single (or a few for multivariate semi-global features [1]) descriptor. Furthermore, this is beneficial also in terms of memory footprint, since a notably reduced amount of information needs to be stored to represent the model library. These properties make global pipelines appealing in scenarios where segmentation is feasible, objects do not present high levels of occlusions and efficiency represents a relevant constraint. Since point-to-point correspondences are not explicitly determined by global algorithms, in order to reconstruct the full 6DOF poses associated with each hypothesis particular expedients have to be deployed. For example, and most notably, in [1] each scene segment is matched to a specific model view, each view retaining its associated yaw and pitch angles of the camera viewpoint. Then, the remaining angle, i.e. the camera roll angle, is retrieved through a specific algorithm known as the Camera Roll Histogram (CRH) [1]. Finally, it is worth noting that both the local and the global pipelines undergo a final stage, known as *Hypothesis Verification, HV*, aimed at pruning inconsistent model hypotheses [6], [4], [5].

In this paper, we aim at 3D object recognition based on the global pipeline, which builds on the proposal in [1] with the aim of improving its capabilities in terms of recognition and pose estimation. The main idea behind this work is to deploy the definition of a *semi-global Reference Frames*, i.e. a repeatable Reference Frame combining local and global aspects of the segmented surface $\mathcal{S}$ being recognized, to improve the performance of global descriptors. Note that, conversely to *local Reference Frames* [8], the *semi-global Reference Frame* does not rely on a pre-defined support size, instead its support inherently adapts to the geometric characteristics of $\mathcal{S}$ driven by smoothness and continuity.

The contribution of this paper is three-fold. i) a method to estimate Semi-Global Unique Reference Frames (SGURF) computed on the object surface as seen from a single viewpoint. The definition of such reference frame allows avoiding the ambiguity over the camera roll angle, thus eliminating the need for the CRH stage. ii) An efficient 3D semi-global descriptor based on SGURF and CVFH, dubbed OUR-CVFH (Oriented, Unique and Repeatable CVFH), which exploits the orientation provided by the reference frame to efficiently encode the geometrical properties of an object surface. iii) A complete global object recognition pipeline to recognize and estimate the 6DOF pose of objects in scenes obtained with a depth sensor. A main contribution in this aspect is a greedy HV method aimed at selecting the best hypothesis among those directly obtained from the descriptor matching stage. Through an experimental comparison we demonstrate that the proposed reference frame and descriptor allows outperforming the state of the art in terms of recognition and pose estimation while reducing the computational burden.

2 SGURF and OUR-CVFH

In this section, we first describe in detail the CVFH [1] descriptor, then we define the SGURF proposal and show how it can be computed on the visible surface ($\mathcal{S}$) of an object seen from a single viewpoint. Upon the definition of SGURF, we introduce the novel OUR-CVFH descriptor, which relies on SGURF to yield a descriptive and distinctive spatial distribution of $\mathcal{S}$, and finally the complete recognition pipeline is presented.

CVFH. In [1], Aldoma et al. proposed the CVFH descriptor as an extension to the VFH descriptor [7] in order to estimate a more robust coordinate frame that could deal with the different data properties of the models (views obtained by virtually rendering accurate 3D meshes from different viewpoints) and scenes (Kinect data with missing parts due to noise and sensor and segmentation artifacts). The basic idea is to identify smooth and continuous regions $\mathcal{C}_i$ – also called CVFH clusters – on the surface $\mathcal{S}$ to be described and use only the points in $\mathcal{C}_i$ to build a coordinate system while still using all points in $\mathcal{S}$ to describe its geometry. Depending on the structure of $\mathcal{S}$, it might be composed of several $\mathcal{C}_i$ from which a different coordinate system is obtained and therefore a different CVFH histogram, each one describing the same surface but encoding it differently. Each $\mathcal{C}_i$ is paired with a (c_i,n_i), respectively representing the centroid and the average of the normals of $\mathcal{C}_i$. Each pair (c_i,n_i) is then independently deployed as one of the axis of a pointwise reference frame (depending also on $p_j \in \mathcal{S}$) from which three angular distributions (each made out of 45 bins) of the normal n_j can be computed and finally added in the corresponding histogram bin. CVFH includes as well a fourth and fifth component (45 and 128 bins respectively) into the histogram, the fourth being based on the $L1$-distribution obtained from c_i

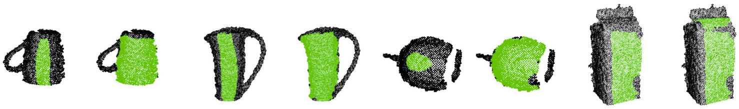


Fig. 1. SGURF and CVFH clusters for different surfaces, left and right respectively. Cloud resolution (r) is 3mm, t_n is 0.15, t_c is 0.015 and $t_d = 2.5 * r$.

and each $p_j \in \mathcal{S}$ and the fifth resulting from yet another angular distribution obtained from each n_j and the central view direction. The total size of a CVFH histogram is 308. CVFH differs as well from other approaches because it explicitly encodes the size of $\mathcal{S}$ by avoiding normalization over the histograms. The assumption is that the spatial sampling resolution of $\mathcal{S}$ in both training and recognition surfaces is equal and therefore the total amount of points in $\mathcal{S}$ is a useful information concerning its size.

CVFH has been shown to deliver good results in the context of 3D recognition as shown in [1]. However, CVFH has two major drawbacks: (i) there is no notion of an aligned Euclidean space causing the feature to miss a proper spatial description and (ii) it is invariant to rotations about the camera's roll angle, addressed in [1] with the CRH as aforementioned in order to yield a full 6DOF pose.

SGURF aims at addressing the limitations of CVFH by defining multiple repeatable coordinate systems on $\mathcal{S}$. These coordinate systems allow, on the one hand, to increase the spatial descriptiveness of the descriptor, on the other to avoid CRH computation and matching by directly obtaining the 6DOF from the alignment of the reference frames.

The first step consists in estimating smooth and continuous clusters $C_i \in \mathcal{S}$ similarly to what CVFH does. First, points whose curvature is higher than a certain t_c threshold are removed from $\mathcal{S}$, yielding $\mathcal{S}^f$. Afterwards, each new cluster is initialized with a random point in $\mathcal{S}^f$ which has not been yet assigned to any cluster. A point p_k with normal n_k is added to a cluster C_i if the cluster contains a point p_j with normal n_j in the direct neighborhood of p_k with a similar normal, i.e. the following constraint is fulfilled:

$$\exists p_j \in C_i : ||p_h - p_j|| < t_d \wedge n_h \cdot n_j > t_n \tag{1}$$

In plain words, the surface $\mathcal{S}^f$ is clustered into smooth and continuous regions, smoothness being controlled by the dot product between the normals of neighboring points while continuity by their Euclidean distance. Differently to CVFH, the points $p_k \in C_i$ are filtered once more by the angle between n_k and n_i (the average normal of the points in C_i). Figure 1 shows the clusters C_i of different surfaces before and after the filtering stage resulting in better shaped clusters for a more robust estimation of the reference frame directions. Each C_i is associated with a pair (c_i, n_i) representing its centroid and average normal. For a specific C_i, the computation of the associated SGURF is as follows:

(i) Compute the eigenvectors of the weighted scatter matrix of the points in C_i, similar to [8]:

$$\mathbf{M} = \frac{1}{\sum\limits_{k \in C_i} (R-d_k)} \sum_{k \in C_i} (R - d_k)(\mathbf{p}_k - \mathbf{c}_i)(\mathbf{p}_k - \mathbf{c}_i)^T \tag{2}$$

where $d_k = \|\mathbf{p}_k - \mathbf{c}_i\|_2$ and R is the maximum euclidean distance between any point in C_i and c_i.

(ii) The sign of the eigenvector related to the smallest eigenvalues, $\mathbf{v_3}$, is disambiguated, differently from [8], by taking the direction yielding a positive dot product with n_i and will represent the z-axis of SGURF. Because the normals of a surface are oriented towards the position of the camera and $\mathbf{v_3}$ is often nearly orthogonal to the surface, the sign disambiguation for this axis is robust.

(iii) At this point, the sign of one axis among the remaining eigenvectors ($\mathbf{v_1}$,$\mathbf{v_2}$) needs to be disambiguated. Let us recall as $\mathbf{v_1}^-$ and $\mathbf{v_2}^-$ as the opposite vectors to ($\mathbf{v_1}$,$\mathbf{v_2}$). Disambiguation is carried out by evaluating the difference of point density between the two hemispheres defined by each eigenvector as in [8]. Conversely to [8], though, the disambiguation deploys the whole surface S (and not just those points used for computing the eigenvectors – this characterizing the *global* aspects of SGURF) and weights each point k according to their distance to c_i. For example, the sign of $\mathbf{v_1}$ is established as follows (analogously for $\mathbf{v_2}$):

$$S^+_{\mathbf{v_1}} = \sum_{k \in S} \|(\mathbf{p}_k - \mathbf{c}_i) \ \cdot \mathbf{v_1}\| \cdot ((\mathbf{p}_k - \mathbf{c}_i) \cdot \mathbf{v_1} \geq 0) \tag{3}$$

$$S^-_{\mathbf{v_1}} = \sum_{k \in S} \|(\mathbf{p}_k - \mathbf{c}_i) \cdot \mathbf{v_1}\| \cdot \big((\mathbf{p}_k - \mathbf{c}_i) \cdot \mathbf{v_1}^- > 0\big) \tag{4}$$

$$\mathbf{v_1} = \begin{cases} \mathbf{v_1}, & |S^+_{\mathbf{v_1}}| \geq |S^-_{\mathbf{v_1}}| \\ \mathbf{v_1}^-, & \text{otherwise} \end{cases} \tag{5}$$

For each of the two eigenvectors, we also compute a *disambiguation factor* f_1, f_2:

$$f_i = \frac{min(|S^-_{\mathbf{v_i}}| \ , \ |S^+_{\mathbf{v_i}}|)}{max(|S^-_{\mathbf{v_i}}| \ , \ |S^+_{\mathbf{v_i}}|)}, \ i = 1, 2 \tag{6}$$

This factor ranges in $[0, 1]$, 0 representing perfect disambiguation while 1 representing complete ambiguity.

(iv) Among $\mathbf{v_1}$,$\mathbf{v_2}$, the one with lower disambiguation factor (f_1,f_2) is chosen as the x-axis of SGURF, since the lower this factor, the less ambiguous the choice of the sign of the eigenvector.

(v) The final y-axis is obtained as $x \times z$.

Unfortunately, in some specific situations the disambiguation is not robust. For example, when both eigenvectors report a similar disambiguation factor, we need to generate two RFs, one using $\mathbf{v_1}$ as the x-axis and the other using $\mathbf{v_2}$. The most challenging case occurs when f_1 and f_2 are similar and both close to

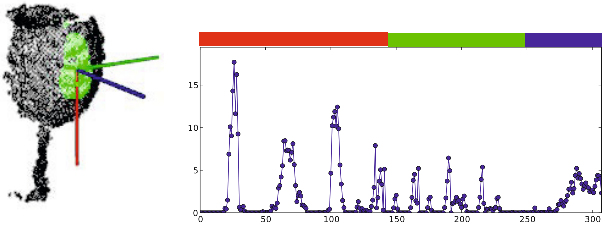


Fig. 2. Left: Point cloud (black) of a wine glass with associated C_i (green) and the SGURF reference frame. Right: The resulting OUR-CVFH histogram. Red and blue bins represent the normal distributions (145 bins) and viewpoint component of CVFH (64 bins). Green bins are the 8 spatial distributions obtained from the points in each octant (104 bins) and the centroid of C_i.

1. In this case, four different reference frames ought to be generated, including both eigenvectors, each encompassing both signs. Figure 2-(a) shows the SGURF associated with a glass of wine. Observe that the x-axis (red) direction is selected along the stem, this helping disambiguation.

The OUR-CVFH Descriptor. So far, for a specific surface $\mathcal{S}$ we have computed N triplets (c_i, n_i, RF_i) obtained from the smooth clustering and the SGURF computation. For the surface description we extend CVFH in the following way: first, c_i and n_i are used to compute the first three components of CVFH and the viewpoint component as presented in [1]. The viewpoint component is however encoded using 64 bins instead of the original 128. Since normals are always pointing towards the sensor position, their dot product with the central view direction is ensured to be in the range $[0, 1]$ and therefore there is no need to reserve histogram space for the rest of the range.

The fourth component of CVFH is completely removed and instead the surface $\mathcal{S}$ is spatially described by means of the computed RF_i. To perform this, $\mathcal{S}$ is rotated and translated so that the RF_i is aligned with the x, y, z axes of the original coordinate system of $\mathcal{S}$ and centered in c_i. For future use in Section 2.1, let us refer to such transformation as $\mathcal{T}$. After the transformation, the points in $\mathcal{S}$ can be easily divided into the 8 octants naturally defined by the signed axes (x^-, y^-, z^-) ... (x^+, y^-, z^-) ... (x^+, y^+, z^+). Additionally, in order to account for perturbations on RF_i due to noise or partially missing parts, interpolation is performed between neighboring octants by associating to each point p_k eight weights, each referred to one octant. The weights are computed by placing three 1-dimensional Gaussian functions over each axis centered at c_i and with $\sigma =$ 1cm, which are combined by means of weight multiplication. Finally, the weights associated with p_k are added to all 8 histograms, its index in each histogram being selected as $\frac{c_i}{R}$, where R is the maximum distance between any point in $\mathcal{S}$ and c_i. The total size of the descriptor is $45 * 3 + 8 * 13 + 64 = 303$ bins. In Figure 2-(b) a OUR-CVFH histogram of a wine glass is reported.

2.1 Global Recognition Pipeline

The recognition pipeline presented in Aldoma et al. [1] consists of four steps: (i) segment the scene using a dominant plane assumption and flood-filling to yield possible model hypotheses therein, (ii) describe each model hypothesis using CVFH and retrieve the best N candidates views from the training views (obtained by virtually rendering the training 3D models) (iii) the candidate 6DOF pose is estimated by means of CRH and successively refined via Iterative Closest Point (ICP) (iv) finally, the best hypothesis is selected by counting the number of inliers that a specific candidate presents in the scene. Steps (ii)-(iv) are repeated as many times as model hypotheses are found by step (i).

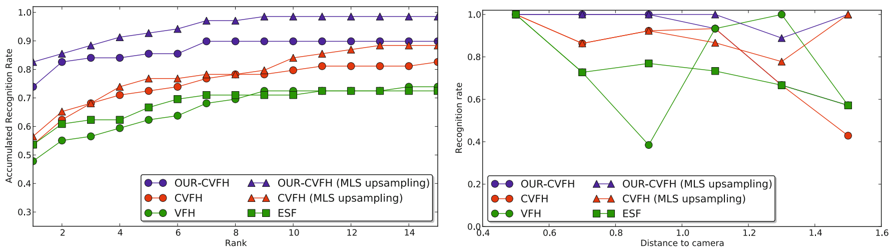
The proposed object recognition pipeline follows the same guidelines, however, step (iii) is replaced by SGURF alignment and the metric in step (iv) is modified in order to consider not only inliers, but outliers as well. Specifically for step (iii), a surface $\mathcal{S}$ – segmented object to be recognized – is matched against a list of candidates ($\mathcal{O}_1$, ... ,$\mathcal{O}_N$) from the model library. $\mathcal{S}$ is associated with a SGURF transformation $\mathcal{T}_S$ (see Section 2) and symmetrically, the candidates are associated with ($\mathcal{T}_{\mathcal{O}_1}$, ... ,$\mathcal{T}_{\mathcal{O}_N}$). Therefore, the 6DOF pose ($\mathcal{P}_{\mathcal{O}_1}$) of $\mathcal{O}_1$ in the scene is given by $\mathcal{P}_{\mathcal{O}_1} = \mathcal{T}_S^{-1}\mathcal{T}_{\mathcal{O}_1}$.

In step (iv), after the pose of all candidates for each segmented object has been computed and refined by ICP ($\mathcal{P}_{\mathcal{O}_i} = \mathcal{P}_{\mathcal{O}_i}\mathcal{P}_{icp}$), we need to select the best candidate, i.e, the one best *explaining* each segmented object. To do so, we compute for each candidate the number of *inliers* and the number of *outliers* based on a distance threshold (t_i). A model point p_j is considered an inlier if, after transformation, its distance to the closest point in S is $\leq t_i$, otherwise it is an outlier. In order for occluded model points not to be considered in the outliers count, a reasoning about occlusions oughts to be made. A model point p_j is considered to be visible if its back-projection on the depth map falls onto a valid pixel (u, v) and its depth d is $\leq$ than the depth at (u, v), meaning that the point lies between the sensor and an actual object. Otherwise, p_j is considered to be occluded and not taken into account for the inliers/outliers count. For each candidate $\mathcal{O}_i$, we compute a metric $\mathcal{M}_{\mathcal{O}_i}$ as follows:

$$\mathcal{M}_{\mathcal{O}_i} = \#inliers - \lambda\#outliers \tag{7}$$

where λ is a weight for the outliers count. The best candidate is determined as the one maximizing $\mathcal{M}_{\mathcal{O}_i}$.

Moving Least Squares (MLS) Upsampling. We experimentally observed (see Figure 3-(b)) that both CVFH and OUR-CVFH recognition capabilities tends to decrease rapidly as the distance from the camera of the object to be recognized increases. CVFH and OUR-CVFH rely on a common resolution between models and scene data to incorporate the object size in the descriptor. Far away from the camera, the Kinect resolution is lower than 3mm (which is the models' resolution), this violating the aforementioned assumption regarding a common resolution between models and scene. To overcome this issue, we add a preprocessing step during recognition where the segmented object surface is



(a) (b)

Fig. 3. (a) Accumulated Recognition Rate for all scenes in the dataset. (b) Recognition rate relative to sensor distance (computed as the distance from the camera to the centroid of the segmented object). Best viewed in color.

upsampled by means of uniformly sampling the MLS plane computed at each original point [2]. This increases the point density of the surface which afterwards is downsampled to the desired 3mm resolution.

3 Experimental Evaluation

In this Section we demonstrate the applicability of SGURF and OUR-CVFH in the context of 3D object recognition. Specifically, we experimentally evaluate the proposed reference frame, surface description and global pipeline on a test dataset composed of 23 scenes obtained with the Kinect Sensor[1]. The scenes contain a total of 69 instances of 25 models. All scenes have been annotated with the object identifiers composing the scene and their respective 6DOF pose.

Experiment 1. First, we evaluate the performance of the different descriptors regarding object recognition and ignore pose estimation. This experiment allows us to evaluate the distinctiveness of each descriptor independently from the other pipeline stages. One single run is performed over the whole dataset retrieving the first 15 nearest neighbors in the descriptor space. An object is considered to be correctly recognized if the selected id matches that of the ground truth. The rank where the correct id is found is saved and results are presented in Figure 3-(a) in form of accumulated recognition rate vs rank. For CVFH and OUR-CVFH variants, histograms were compared by means of the distance metric presented in [1]. For VFH [7]) and ESF [9] we evaluated different metrics — $L1$, $L2$ and χ^2. VFH performed the best with χ^2 and ESF with $L2$ (both depicted in Figure 3). Figure 3-(a) highlights the importance of an oriented reference frame for a distinctive description of the objects as OUR-CVFH clearly outperforms the compared descriptors (especially when a low number of candidates is retrieved).

Experiment 2. Moreover, we compare the 6DOF pose estimation capabilities of SGURF and CRH within the proposed object recognition pipeline. To this aim,

[1] http://users.acin.tuwien.ac.at/aaldoma/datasets/DAGM.zip

Table 1. Results yielded by the proposed pipeline and OUR-CVFH with MLS upsampling at different ICP iterations (0,10,30), comparing pose estimation yielded by SGURF and by CRH

	#correct_id			**#correct_pose**			**time (s)**		
ICP iterations:	0	10	30	0	10	30	0	10	30
SGURF	62	64	66	57	61	63	28.1	48.5	79.2
CRH	49	61	61	35	53	57	42.0	61.3	129.0
Difference:	+13	+3	+5	+22	+8	+6	-13.9	-12.8	-49.8

Table 2. Results in terms of recognition and 6DOF pose estimation comparing OUR-CVFH and the global pipeline with the SHOT [8] descriptor and the local pipeline, both using MLS upsampling

	#correct_id			**#correct_pose**			**time (s)**		
ICP iterations:	0	10	30	0	10	30	0	10	30
OUR-CVFH	62	64	66	57	61	63	28.1	48.5	79.2
SHOT	51	61	62	46	60	61	94.4	148.9	229.8
Difference:	+11	+3	+4	+11	+1	+2	-66.3	-100.4	-150.0

we select the best performing descriptor from Figure 3, i.e, *OUR-CVFH MLS upsampling*. The first 10 candidates are retrieved and their pose independently estimated with SGURF and CRH. If an object in the scene has no recognition candidates left after the HV stage ($max(\mathcal{M}_{\mathcal{O}_i}) < 0$), the object is considered to be not recognized. Instead, if the best candidate after HV has the same id as in the annotated data and the RMSE error between ground truth and the aligned model is ≤ 0.005, the object is considered to be correctly recognized and its pose correctly estimated. Results are presented in Table 1 where the candidates' pose is refined with 0, 10 and 30 ICP iterations. Table 1 clearly shows the superiority of SGURF over CRH, this being even more notable when ICP refinement is not performed. SGURF is also superior in terms of efficiency. As explained in [1], CRH is not always resolutive and several roll hypotheses might need to be post-processed in order to select the best one (those within 0.8 of the biggest cross-correlation peak as suggested in [1]). This impacts recognition time since a higher number of hypotheses are generated with respect to SGURF.

Experiment 3. This experiment is similar to experiment 2, but uses as comparison a standard local pipeline (see Section 1) where point-to-point correspondences are established by means of the Signature of Histograms of OrienTations (SHOT [8]) descriptor. Results are presented in Table 2 showing that OUR-CVFH is superior in terms of accuracy (especially when ICP is not deployed) and its computational burden its substantially smaller. For a fair comparison, the local pipeline processes only the segmented surfaces yielded by the initial segmentation stage, and deploys the same HV stage as the global pipeline.

4 Conclusion

We have presented a novel approach to estimate semi-global unique and repeatable reference frames (SGURF) on object surfaces. By combining local and global properties of the surface, SGURF is robust to common artifacts and distortions present in 3D data. SGURF is useful to (i) design a semi-global descriptor, i.e. OUR-CVFH, with spatial awareness resulting in increased distinctiveness and (ii) efficiently estimate the 6DOF pose of an object. The proposed OUR-CVFH descriptor also improves CVFH thanks to an interpolation step conferring improved robustness. Overall, the proposed descriptor and object recognition pipeline are able to correctly recognize and accurately estimate the 6DOF pose of 57 objects out of 69 objects (82%), requiring on the average 0.4s per object. With ICP refinement, the recognition rate increases up to 91% requiring an average of 1s per object. This highlights the importance of faster pose refinement in order to achieve real-time recognition and 6DOF pose estimation.

References

1. Aldoma, A., Blodow, N., Gossow, D., Gedikli, S., Rusu, R.B., Vincze, M., Bradski, G.: CAD-Model Recognition and 6DOF Pose Estimation Using 3D Cues. In: 3DRR Workshop, ICCV (2011)
2. Alexa, M., Behr, J., Cohen-or, D., Fleishman, S., Levin, D., Silva, C.T.: Computing and rendering point set surfaces. IEEE Transactions on VCG 9, 3–15 (2003)
3. Marton, Z., Pangercic, D., Blodow, N., Beetz, M.: Combined 2D-3D categorization and classification for multimodal perception systems. IJRR (2011)
4. Mian, A., Bennamoun, M., Owens, R.: 3D model-based object recognition and segmentation in cluttered scenes. IEEE Trans. PAMI (10) (2006)
5. Mian, A., Bennamoun, M., Owens, R.: On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes. IJCV (2010)
6. Papazov, C., Burschka, D.: An Efficient RANSAC for 3D Object Recognition in Noisy and Occluded Scenes. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part I. LNCS, vol. 6492, pp. 135–148. Springer, Heidelberg (2011)
7. Rusu, R.B., Bradski, G., Thibaux, R., Hsu, J.: Fast 3D recognition and pose using the viewpoint feature histogram. In: IROS, Taipei, Taiwan (October 2010)
8. Tombari, F., Salti, S., Di Stefano, L.: Unique Signatures of Histograms for Local Surface Description. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part III. LNCS, vol. 6313, pp. 356–369. Springer, Heidelberg (2010)
9. Wohlkinger, W., Vincze, M.: Ensemble of shape functions for 3D object classification. In: ROBIO (2011)