An Efficient Global Point Cloud Descriptor for Object Recognition and Pose Estimation

João Paulo Silva do Monte Lima*†
*Departamento de Estatística e Informática (DEINFO)
Universidade Federal Rural de Pernambuco (UFRPE)
Recife, Brazil
joao.mlima@ufrpe.br

Veronica Teichrieb†
†Voxar Labs - Centro de Informática (CIn)
Universidade Federal de Pernambuco (UFPE)
Recife, Brazil
{jpsml, vt}@cin.ufpe.br

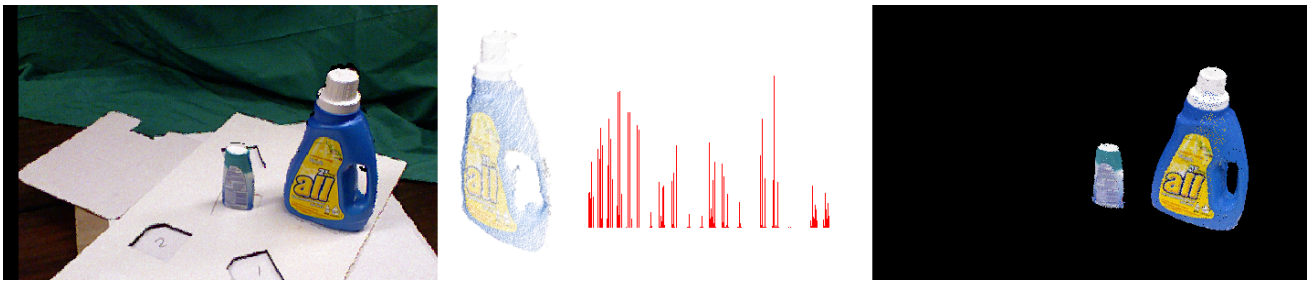


Fig. 1. Input colored point cloud of a given scene (left), partial view of an object and global descriptor computed from it (center) and recognized objects in the given scene rendered with their estimated poses (right).

***Abstract*—This paper presents a global point cloud descriptor to be used for efficient object recognition and pose estimation. The proposed method is based on the estimation of a reference frame for the whole point cloud that represents an object instance, which is used for aligning it with the canonical coordinate system. After that, a descriptor is computed for the aligned point cloud based on how its 3D points are spatially distributed. Such descriptor is also extended with color distribution throughout the aligned point cloud. The global alignment transforms of matched point clouds are used for computing object pose. The proposed approach was evaluated with a publicly available dataset, showing that it outperforms major state of the art global descriptors regarding recognition rate and performance and that it allows precise pose estimation.**

***Keywords*-cloud descriptor; object recognition; pose estimation.**

I. Introduction

Real-time object recognition and pose estimation has applications in many areas, such as augmented reality, robotics and human-machine interaction. 3D point cloud processing can be utilized to perform this task, presenting some advantages. One of them is to offer a practical way of acquiring 3D models of the objects to be detected. It is also possible to automatically determine the real scale of candidate object instances, which may not be possible when other data types are used (e.g. only RGB images). Besides that, the fact that 3D point clouds may provide both geometric (3D coordinates, surface normals, edges from depth discontinuities, high curvature regions, etc.) and photometric (colors, color gradients, edges from color discontinuities, local discriminant features, etc.) information contributes for obtaining superior results. In recent years, RGB-D sensors have become low cost consumer devices accessible to general users. Such sensors can be used to generate colored 3D point clouds of a given scene surface in real-time.

A common way to recognize and estimate the pose of objects consists in matching feature descriptors extracted from the input scene with previously obtained objects models. One key advantage of this approach is scalability with respect to the number of objects in the database, since descriptor matching can be efficiently performed using approximate nearest neighbor search strategies [1].

In this context, this paper presents a novel global descriptor named Globally Aligned Spatial Distribution (GASD), which was designed for efficient object recognition and pose estimation from point clouds. It is based on the concepts of global reference frame and globally aligned shape and color distributions. The proposed method allows object recognition and pose estimation in a fast, accurate and robust manner.

The contributions of this work are: (1) a point cloud descriptor based on global reference frame estimation and globally aligned shape and color distributions that is suitable for object recognition and pose estimation; (2) the use of the global reference frame concept together with existing global descriptors as a way to improve their results; (3) an evaluation regarding object recognition rate and pose estimation accuracy; (4) a performance evaluation of the proposed approach in comparison with existing techniques.

This paper is organized as follows. Section II presents works related to point cloud description. Section III describes

the GASD descriptor. Section IV details the results obtained with the proposed method. Conclusions and future work are discussed in Section V .

II. Related Work

There are mainly two types of descriptors used for object recognition from point clouds: local descriptors, which are used to match localized features of the input point cloud with corresponding features of the objects' models; and global descriptors, which aim to match whole objects or significant parts of them that were previously segmented from the point clouds.

Examples of local point cloud descriptors are Persistent Feature Histograms (PFH) [2], Fast Point Feature Histograms (FPFH) [3], Signature of Histograms of OrienTations (SHOT) [4], Color SHOT (CSHOT) [5], Binary Appearance and Shape Elements (BASE) [6] and Binary Robust Appearance and Normals Descriptor (BRAND) [7]. However, in order to achieve real-time results, it is important to extract in an efficient way a not too large set of local point cloud features that present a high level of repeatability and discriminative power. Existing techniques that perform this task, such as Local Surface Patches (LSP) [8] and Intrinsic Shape Signatures (ISS) [9], still do not reach this goal. A recent evaluation of local 3D feature detection methods available in [10] points out that none of the evaluated detectors was capable of handling a typical 3D point cloud generated by a low cost RGB-D sensor in less than 1 second. In addition, in the evaluation of local 3D feature descriptors described in [11], it was verified that the evaluated descriptors present weak results with data from low cost sensors, suggesting as alternative that research should be directed towards the design of local descriptors that are suitable for data with low resolution and high noise level.

Since global approaches generate less descriptors than local ones, descriptor matching is often faster and less memory resources are commonly needed. Regarding existing global point cloud descriptors, the Viewpoint Feature Histogram (VFH) [12] is composed of a viewpoint component and a surface shape component. The viewpoint component consists in a histogram of the angles between each point normal and the central viewpoint direction. The surface shape component is given by a histogram of relative pan, tilt and yaw angles between each point normal and the object centroid normal. While VFH is truly global, computing a single descriptor for an entire object instance, there are other descriptors that are considered semi-global, since they may compute a few descriptors using clusters extracted from the whole object surface. This is the case of the Clustered VFH (CVFH) descriptor [13], where a smooth region growing algorithm is used for extracting stable clusters from the point cloud. CVFH has the same viewpoint and surface shape components of VFH, but centroid position and normal are computed from each cluster instead of the entire object. It also has an additional shape distribution component that consists in a histogram of normalized distances between each cloud point and the cluster's centroid. The Oriented, Unique and Repeatable CVFH (OUR-CVFH) descriptor [14] shares the same viewpoint and surface shape components of CVFH, but the shape distribution component is replaced by 8 histograms of distances between points and centroid, one for each octant of a reference frame. The computation of such reference frame for the extracted cluster is done using the Semi-Global Unique Reference Frame (SGURF) technique. In [15], OUR-CVFH was extended with color information in the YUV space, also taking into account the cluster's reference frame. Finally, the Ensemble of Shape Functions (ESF) [16] is a truly global descriptor obtained by combining angle, point-distance and area shape functions to the object point cloud.

However, as will be explained in the comparisons detailed in Section IV, the GASD descriptor proposed in this paper presents a better balance between performance and recognition results than the aforementioned global descriptors.

III. Globally Aligned Spatial Distribution

The proposed global description method takes as input a 3D point cloud that represents a partial view of a given object (Fig. 1 center). The first step consists in estimating a reference frame for the point cloud, which allows the computation of a transform that aligns it to the canonical coordinate system, making the descriptor pose invariant. After alignment, a shape descriptor is computed for the point cloud based on the spatial distribution of the 3D points. Color distribution along the point cloud can also be taken into account for obtaining a shape and color descriptor with a higher discriminative power (Fig. 1 center). Object recognition is then performed by matching query and train descriptors of partial views. The pose of each recognized object is also computed from the alignment transforms of matched query and train partial views (Fig. 1 right). All these procedures are detailed in the following subsections.

A. Reference Frame Estimation

The method employed for estimating a reference frame for the object partial view is based on the normal and orientation estimation step of the DARC technique described in [17]. However, DARC takes as input planar contours and aims to rectify them, while the current approach handles free-form surfaces and intends to align them with the canonical coordinate system for later description. The reference frame estimation method also resembles the SGURF technique detailed in [14]. Nevertheless, SGURF extracts smooth point clusters from the input partial view and estimates a reference frame for each cluster, while the current approach estimates a single reference frame for the entire point cloud that represents the partial view. In addition, SGURF uses surface normal information, which is not utilized by the current method.

The reference frame is estimated using a Principal Component Analysis (PCA) approach. Given a set of 3D points $\mathbf{P_i}$ that represents a partial view of an object, with $i \in \{1, ..., n\}$, the first step consists in computing their centroid by

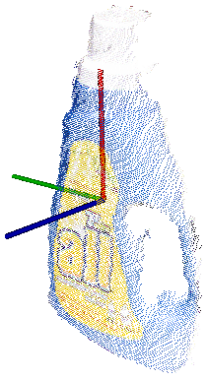


Fig. 2. Reference frame estimated from a 3D point cloud that represents a partial view of a given object: x axis (red), y axis (green) and z axis (blue).

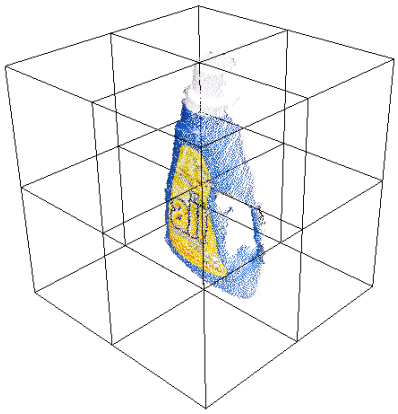


Fig. 3. Example grid with size $m_s = m_c = 2$ used for computing the GASD descriptor for a given point cloud.

$$\overline{\mathbf{P}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{P_i}. \tag{1}$$

The origin of the reference frame is given by $\overline{\mathbf{P}}$. Then a covariance matrix $\mathbf{C}$ is computed from $\mathbf{P_i}$ and $\overline{\mathbf{P}}$ as follows:

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{P_i} - \overline{\mathbf{P}})(\mathbf{P_i} - \overline{\mathbf{P}})^T. \tag{2}$$

After that, the eigenvalues λ_j and corresponding eigenvectors $\mathbf{v_j}$ of C are obtained, with $j \in \{1, 2, 3\}$, such that $\mathbf{C}\mathbf{v_j} = \lambda_j \mathbf{v_j}$. Considering that the eigenvalues are arranged in ascending order, the eigenvector $\mathbf{v_1}$ associated with the minimal eigenvalue is used as the z axis of the reference frame. If the angle between $\mathbf{v_1}$ and the viewing direction is in the $[-90°, 90°]$ range, then $\mathbf{v_1}$ is negated. This ensures that the z axis always points towards the viewer. The x axis of the reference frame is the eigenvector $\mathbf{v_3}$ associated with the maximal eigenvalue. The y axis is given by $\mathbf{v_2} = \mathbf{v_1} \times \mathbf{v_3}$. The reference frame estimated for a given partial view is illustrated in Fig. 2.

From the reference frame, it is possible to compute a transform $[\mathbf{R}|\mathbf{t}]$ that aligns it with the canonical coordinate system. All the points $\mathbf{P_i}$ of the partial view are then transformed with $[\mathbf{R}|\mathbf{t}]$, which is defined as follows:

$$\begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{v_3}^T & -\mathbf{v_3}^T\overline{\mathbf{P}} \\ \mathbf{v_2}^T & -\mathbf{v_2}^T\overline{\mathbf{P}} \\ \mathbf{v_1}^T & -\mathbf{v_1}^T\overline{\mathbf{P}} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{3}$$

B. Shape Description

Once the point cloud is aligned using the reference frame, a pose invariant global shape descriptor can be computed from it. In the proposed approach, a single descriptor is computed for the entire point cloud. It also does not rely on surface normals, which allows a faster computation.

The descriptor is based on the distribution of the 3D points in the cloud. The point cloud axis-aligned bounding cube centered on the origin is divided into an $m_s \times m_s \times m_s$ regular grid, as illustrated in Fig. 3 and Fig. 4. For each grid cell, the number of points that belong to it is stored, forming a histogram.

The contribution of each sample to the histogram is normalized with respect to the total number of points in the cloud. Optionally, trilinear interpolation may be used to distribute the value of each sample into adjacent cells, in an attempt to avoid boundary effects that may cause abrupt changes to the histogram when a sample shifts from being within one cell to another. The descriptor is then obtained by concatenating the computed histograms.

C. Shape and Color Description

Color information can also be incorporated to the descriptor in order to increase its discriminative power. The color component of the descriptor is computed with an $m_c \times m_c \times m_c$ grid similar to the one used for the shape component, but a color histogram is generated for each cell based on the colors of the points that belong to it. Point cloud color is represented in the HSV space and the hue values are accumulated in histograms with l bins. Similarly to shape component computation, normalization with respect to number of points is performed. Additionally, quadrilinear interpolation of histograms samples may also be performed. The shape and color components are concatenated, resulting in the final descriptor.

D. Descriptor Matching and Pose Estimation

Query and train descriptors are matched using a nearest neighbor search approach. After that, for each matched object instance, a coarse pose is computed using the alignment transforms obtained from the reference frames of the respective query and train partial views. Given the transforms $[\mathbf{R_q}|\mathbf{t_q}]$ and $[\mathbf{R_t}|\mathbf{t_t}]$ that align the query and train partial views, respectively, the object coarse pose $[\mathbf{R_c}|\mathbf{t_c}]$ is obtained by

$$\begin{bmatrix} \mathbf{R_c} & \mathbf{t_c} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R_q} & \mathbf{t_q} \\ \mathbf{0} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{R_t} & \mathbf{t_t} \\ \mathbf{0} & 1 \end{bmatrix}. \tag{4}$$

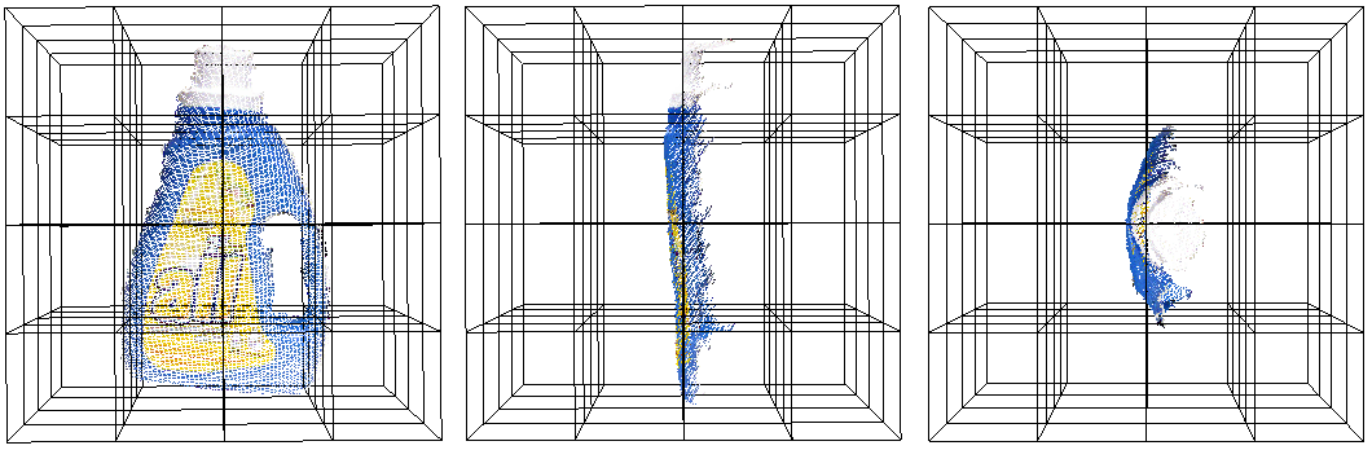


Fig. 4. Frontal view (left), side view (center) and top view (right) of an example grid with size $m_s = m_c = 4$ used to compute the GASD descriptor of a given point cloud.

The coarse pose $[\mathbf{R_c}|\mathbf{t_c}]$ can then be refined using the Iterative Closest Point (ICP) algorithm [18].

IV. Results

The GASD descriptor was evaluated under an object recognition and pose estimation scenario. The hardware used in the evaluations was a laptop with an Intel Core i7-5500U @ 2.40 GHz processor and 16 GB RAM. The publicly available Challenge dataset [1] was used, which contains 35 objects and 176 scenes where one or more of these objects appear, with a total of 434 objects instances. The colored point clouds of the objects' models and the scenes were obtained using a Microsoft Kinect v1 RGB-D sensor. The dataset provides ground truth poses for each object instance in the scenes. It also makes available segmented and registered partial views used to generate the objects models, which cover a loop around each object with a 10° step. The train descriptors were generated from these partial views.

In order to compute descriptors for objects instances in the input scenes, candidate partial views were initially segmented based on the efficient approach detailed in [19]. However, in a few cases it was not possible to correctly segment the object instance, and in such situations the method described in [20] was employed, which is more robust but slower. GASD was compared to the following shape descriptors: VFH [12], CVFH [13], OUR-CVFH [14] and ESF [16]. Similarly to GASD, the ESF descriptor does not rely on normal information, which is used by VFH, CVFH and OUR-CVFH. For the later methods, normals are estimated with the smoothed depth changes approach presented in [21]. The descriptors implementations available in the Point Cloud Library (PCL) [2] with their default parameters' values were utilized in the tests. The only change made was to force ESF to always use the same seed for random number generation, in order to make its results deterministic.

[1]https://repo.acin.tuwien.ac.at/tmp/permanent/ghv_results/dataset_index.php

[2]http://www.pointclouds.org

For each descriptor type, it was chosen the distance metric to be applied in the nearest neighbor search that gave best results. L1 distance was used with GASD and VFH, L2 distance was used with CVFH and OUR-CVFH and χ^2 distance was used with ESF.

A. Recognition Evaluation

First, different configurations of the GASD shape only descriptor were evaluated. The shape grid size m_s was set to different values. In addition, two variants of the GASD shape only descriptor were considered: with and without trilinear interpolation (GASD-SI and GASD-S, respectively). In this experiment, a correct recognition occurs when the nearest neighbor of the query descriptor is a train descriptor that was obtained from the ground truth object. As can be seen in Fig. 5, the configuration that obtained best results was GASD-SI with $m_s = 8$, resulting in a descriptor with $8{\times}8{\times}8 = 512$ elements. It should also be noted that GASD-S with $m_s = 6$ offers a good tradeoff between descriptor length ($6 \times 6 \times 6 = 216$ elements) and recognition rate. These results can be explained by the fact that, while a low number of histogram bins makes the descriptor less discriminative, a high number may cause the descriptor to be more sensitive to distortions.

In the next experiment, GASD-SI with $m_s = 8$ was compared to other shape descriptors by retrieving the k nearest neighbors of each query descriptor. If any of the retrieved train descriptors were computed from the ground truth object, then it was considered to be correctly recognized. Different values of k were tested, ranging from 1 to 15. CVFH and OUR-CVFH may compute several descriptors for a single partial view, therefore if at least one of them is associated with a train descriptor from the correct object, then this is counted as a true positive. Fig. 6 shows that the results obtained with GASD-SI are better than the ones obtained with the other descriptors, especially when using fewer nearest neighbors. This is an interesting property, since it may lead to test fewer hypotheses until the correct object is found, thus improving recognition time. VFH-based methods presented low recognition rates,

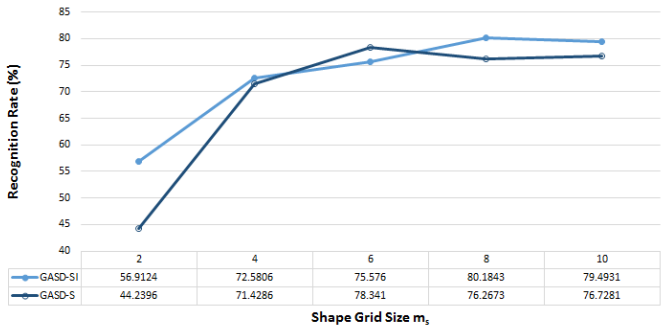


Fig. 5. Evaluation of different shape grid sizes m_s for the GASD shape only descriptor with (GASD-SI) and without (GASD-S) trilinear interpolation.

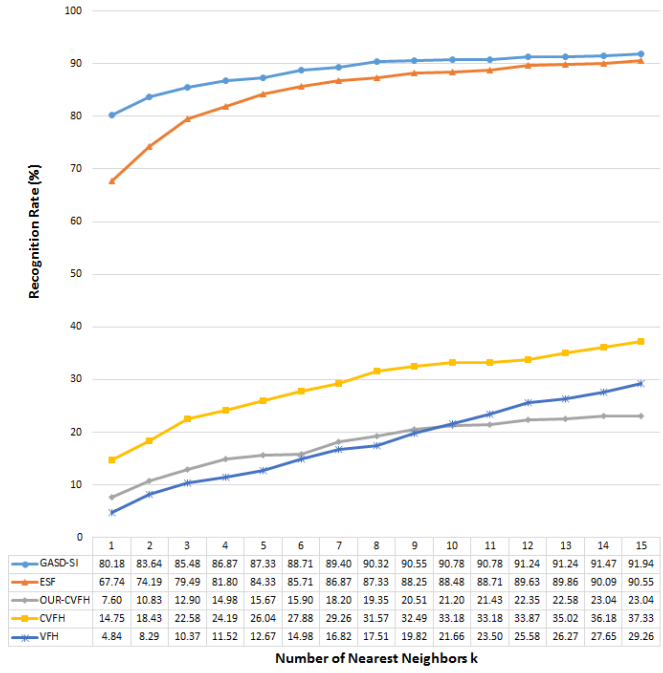


Fig. 6. Recognition rate with respect to number of nearest neighbors of the evaluated approaches with their default configurations.

which indicates that the use of normal information did not bring enough distinctiveness in the evaluated scenario. In the particular cases of CVFH and OUR-CVFH, these results can also be explained by the fact that the smooth clusters extracted from the point cloud were not too much repeatable and discriminative. In Fig. 7, it is shown that, in a given scene, GASD-SI correctly recognizes all 4 objects, while ESF fails to detect the meat can and detergent instances.

It was also evaluated the effect of computing VFH, CVFH and ESF descriptors from the aligned point cloud with respect to the estimated reference frame (VFH + RF, CVFH + RF and ESF + RF, respectively). OUR-CVFH was not considered in this test, since it already uses a reference frame and a prior alignment did not bring any improvement. As depicted in Fig. 8, using a reference frame improved the results of existing descriptors, since very similar partial views are used for computing matching descriptors. By comparing these results with the ones in Fig. 6, it can be noted that the ESF + RF variant obtained equal or slightly better recognition rates when compared to GASD-SI for most values of k. However, as later presented in Subsection IV-C, GASD-SI is much faster than ESF + RF.

Different configurations of (GASD-S, $m_s = 6$) extended with color information were also tested, considering versions with (GASD-SCI) and without (GASD-SC) interpolation and different values of m_c and l. The best configuration was GASD-SC with $m_c = 4$ and $l = 12$, as shown in Fig. 9, resulting in a final descriptor that contains $216+4\times4\times4\times12 = 984$ elements. In the experiments conducted it was seen that using even higher values of m_c would increase recognition rate, but this was not done in order to avoid a high increase in descriptor dimensionality. It is worth noting that using interpolation caused a decrease in recognition rate.

Fig. 10 compares the results obtained with the best configurations of GASD-SI and GASD-SC. It shows that exploiting color information contributes to obtaining better recognition rates. This is also illustrated in the example shown in Fig. 11: while GASD-SI is not able to distinguish the two juice bottles and the correct side of the soy milk box, this is properly done by GASD-SC.

B. Pose Estimation Evaluation

In order to evaluate the pose estimation accuracy of the proposed method, it was calculated the translation error of the pose computed for each object that was correctly recognized by GASD-SC when only the first nearest neighbor is considered. The coarse poses computed from the reference frames alignment were compared to the refined poses obtained with ICP. As can be seen in the error histogram shown in Fig. 12, many of the coarse pose errors ranged from 18 to 21 mm, while several fine pose errors ranged from 15 to 18 mm. The mean and standard deviation in mm of coarse and fine pose errors, respectively, were 19.24 ± 8.51 and 17.80 ± 7.50.

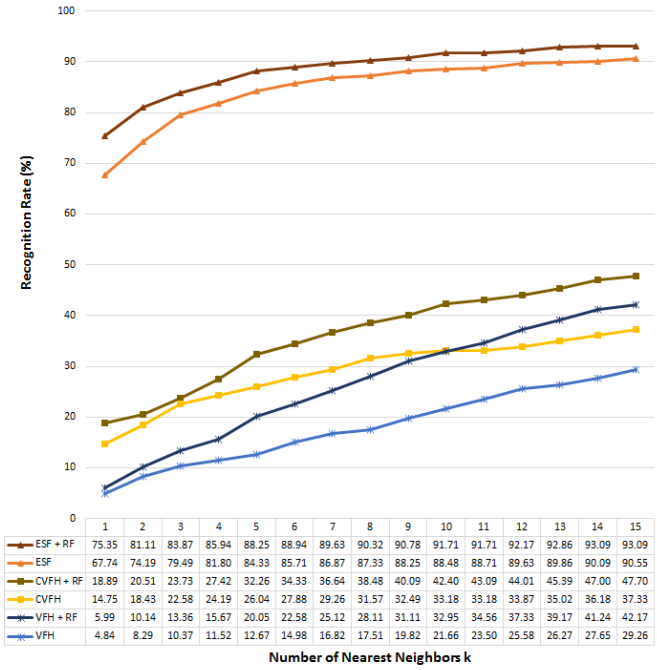
C. Runtime Analysis

Table I presents a performance evaluation of a non-optimized version of the proposed approach in comparison with the other descriptors. GASD-SI together with reference frame estimation take on average 1.03 ms per object instance, being the fastest alternative. GASD-SC + reference frame estimation have a mean execution time of 1.39 ms per object instance. All the other descriptors are slower, with ESF spending on average more than 30 ms per object instance, and both CVFH and OUR-CVFH taking on average more than 120 ms per object instance. Pose refinement with ICP is currently a bottleneck, having a mean execution time of almost 120 ms per object instance. Brute force nearest neighbor search for descriptor matching is executed only once per scene, taking on average between 1.5 and 4.5 ms.

D. Failure Cases

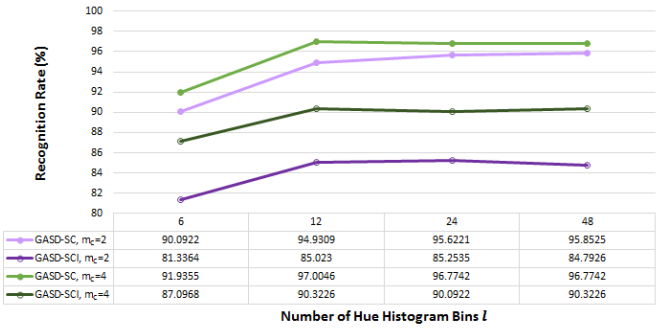
Since the GASD descriptor is scale invariant, in a few cases it may confuse partial views with similar shape and color

Fig. 7. Object recognition and pose estimation example: input scene (left), results obtained using ESF (center) and GASD-SI (right).



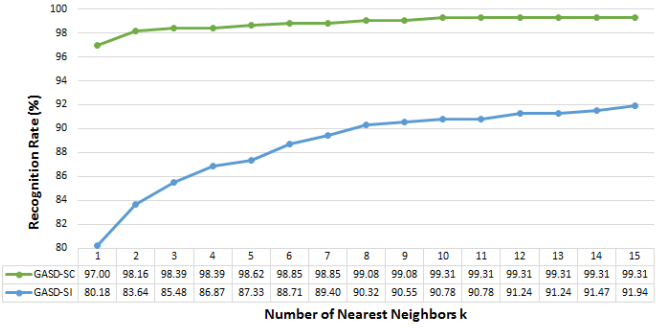
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ESF + RF	75.35	81.11	83.87	85.94	88.25	88.94	89.63	90.32	90.78	91.71	91.71	92.17	92.86	93.09	93.09
ESF	67.74	74.19	79.49	81.80	84.33	85.71	86.87	87.33	88.25	88.48	88.71	89.63	89.86	90.09	90.55
CVFH + RF	18.89	20.51	23.73	27.42	32.26	34.33	36.64	38.48	40.09	42.40	43.09	44.01	45.39	47.00	47.70
CVFH	14.75	18.43	22.58	24.19	26.04	27.88	29.26	31.57	32.49	33.18	33.18	33.87	35.02	36.18	37.33
VFH + RF	5.99	10.14	13.36	15.67	20.05	22.58	25.12	28.11	31.11	32.95	34.56	37.33	39.17	41.24	42.17
VFH	4.84	8.29	10.37	11.52	12.67	14.98	16.82	17.51	19.82	21.66	23.50	25.58	26.27	27.65	29.26

Fig. 8. Recognition rate with respect to number of nearest neighbors of existing descriptors with and without using reference frame alignment.



	6	12	24	48
GASD-SC, m_c=2	90.0922	94.9309	95.6221	95.8525
GASD-SCI, m_c=2	81.3364	85.023	85.2535	84.7926
GASD-SC, m_c=4	91.9355	97.0046	96.7742	96.7742
GASD-SCI, m_c=4	87.0968	90.3226	90.0922	90.3226

Fig. 9. Evaluation of different color grid sizes m_c and number of hue histogram bins l for the GASD shape and color descriptor with (GASD-SCI) and without (GASD-SC) interpolation, using $m_s = 6$.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
GASD-SC	97.00	98.16	98.39	98.39	98.62	98.85	98.85	99.08	99.08	99.31	99.31	99.31	99.31	99.31	99.31
GASD-SI	80.18	83.64	85.48	86.87	87.33	88.71	89.40	90.32	90.55	90.78	90.78	91.24	91.24	91.47	91.94

Fig. 10. Recognition rate with respect to number of nearest neighbors of GASD when only shape information is used (GASD-SI) and both shape and color are employed (GASD-SC).

TABLE I
MEAN AND STANDARD DEVIATION OF TIME SPENT BY EACH PROCEDURE FOR PROCESSING A SINGLE OBJECT INSTANCE.

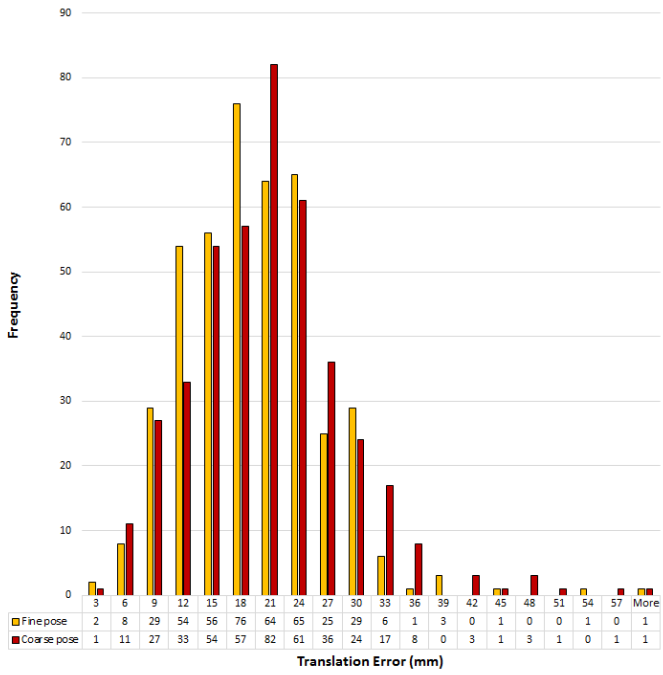
Procedure		Time (ms)
Reference frame estimation		0.33 ± 0.23
Description	GASD-SI	0.70 ± 0.45
	GASD-SC	1.06 ± 0.77
	ESF	31.30 ± 4.96
	OUR-CVFH	128.27 ± 111.00
	CVFH	124.33 ± 108.39
	VFH	3.86 ± 2.76
Pose refinement		119.27 ± 103.04

but significantly different sizes, as can be seen in Fig. 13. However, such problem can be easily avoided by comparing the bounding boxes of the aligned partial views.

Since the proposed descriptor is global, it is sensitive to partial occlusions, which can harm the estimation of the reference frame and the computation of histograms. Such issue is illustrated by the example in Fig. 14, where a reference frame could not be properly estimated for the blue detergent bottle due to the partial occlusion suffered by it. This caused the object instance to be confused with a bag box.

In some situations, especially when dealing with symmetric objects, the proposed descriptor does not present enough distinctiveness for retrieving the correct training view. Therefore,

Fig. 11. Object recognition and pose estimation example: input scene (left), results obtained using GASD-SI (center) and GASD-SC (right).



	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45	48	51	54	57	More
Fine pose	2	8	29	54	56	76	64	65	25	29	6	1	3	0	1	0	0	1	0	1
Coarse pose	1	11	27	33	54	57	82	61	36	24	17	8	0	3	1	3	1	0	1	1

Fig. 12. Histogram of translation errors for correctly recognized objects using GASD-SC with and without ICP pose refinement.

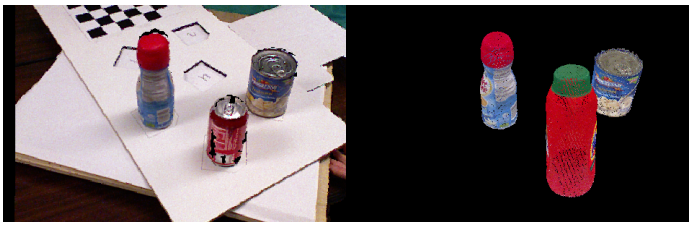


Fig. 13. GASD confuses partial views of a soda can and a detergent bottle.

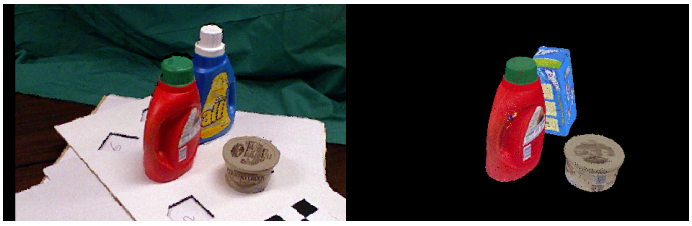


Fig. 14. GASD fails to recognize a partially occluded blue detergent bottle.

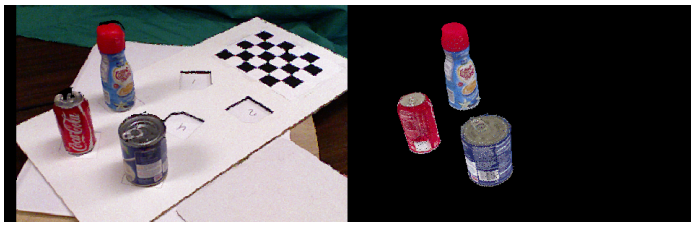


Fig. 15. Error in rotation estimation of the soda and soup cans by GASD.

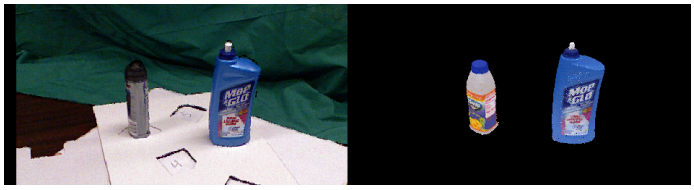


Fig. 16. The shaving cream can, which has desaturated colors, is confused with a juice bottle by GASD.

object rotation is not accurately computed, as can be noted in the soda and soup cans depicted in Fig. 15.

In some cases, the proposed approach may fail to recognize objects that have mainly desaturated colors, as illustrated in Fig. 16. This is due to the fact that only the hue component is used by GASD, and desaturated colors may have the same hue value of other different colors.

When objects have similar shape and color distributions, the proposed descriptor might sometimes not be able to distinguish between them. This is the case with the two different kinds of tomato soup in the example shown in Fig. 17.

Fig. 17. The proposed approach confuses the tomato soups.

V. Conclusion and Future Work

It was presented GASD, which is an efficient approach for global point cloud description that was successfully applied to object recognition and pose estimation. The proposed method exploits a reference frame estimated for the entire point cloud for computing a globally aligned shape distribution that can also be extended with color information. It was able to obtain better recognition rates than some existing global descriptors in a publicly available dataset. It was shown that the reference frame can be used together with existing global descriptors for improving their results and allows accurate pose estimation. In addition, the average processing time taken by the proposed technique for handling a single object instance was 1 or 2 orders of magnitude lower than most of the other evaluated descriptors.

Current limitations of the proposed approach are: the pose refinement step is very time consuming in comparison with the other procedures; it is not robust to partial occlusions; rotation estimation of symmetric objects can be imprecise in some cases; objects with desaturated colors may be incorrectly matched; and it may fail when handling objects with similar shape and color distributions.

As future work, pose refinement speed may be improved with a GPU implementation of ICP, such as in [22]. It will also be investigated the performance and pose estimation accuracy tradeoff of using color information in ICP, as done in [23]. In order to cope with partial occlusions, one possible direction would be to use GASD together with local features in a hybrid manner. It will be studied how descriptor robustness can be further increased, by for example evaluating the use of different color channels/spaces and interpolation strategies (such as the Gaussian interpolation used in [14]). Finally, it is intended to incorporate the proposed descriptor into a real-time scalable object recognition and pose estimation pipeline, including steps such as point cloud acquisition and segmentation.

Acknowledgment

The authors would like to thank Rafael Alves Roberto for meaningful comments and discussion, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (process 456800/2014-0) for partially funding this research.

References

[1] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *International Conference on Computer Vision Theory and Application VISAPP'09)*. INSTICC Press, 2009, pp. 331–340.

[2] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz, "Aligning point cloud views using persistent feature histograms," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2008, pp. 3384–3391.

[3] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, May 2009, pp. 3212–3217.

[4] F. Tombari, S. Salti, and L. Di Stefano, *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ch. Unique Signatures of Histograms for Local Surface Description, pp. 356–369.

[5] F. Tombari, S. Salti, and L. D. Stefano, "A combined texture-shape descriptor for enhanced 3d feature matching," in *2011 18th IEEE International Conference on Image Processing*, Sept 2011, pp. 809–812.

[6] E. R. Nascimento, W. R. Schwartz, G. L. Oliveira, A. W. Veira, M. F. M. Campos, and D. B. Mesquita, "Appearance and geometry fusion for enhanced dense 3d alignment," in *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*, Aug 2012, pp. 47–54.

[7] E. R. Nascimento, G. L. Oliveira, M. F. M. Campos, A. W. Vieira, and W. R. Schwartz, "Brand: A robust appearance and depth descriptor for rgb-d images," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, pp. 1720–1726.

[8] H. Chen and B. Bhanu, "3d free-form object recognition in range images using local surface patches," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1252 – 1262, 2007.

[9] Y. Zhong, "Intrinsic shape signatures: A shape descriptor for 3d object recognition," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, Sept 2009, pp. 689–696.

[10] F. Tombari, S. Salti, and L. Di Stefano, "Performance evaluation of 3d keypoint detectors," *International Journal of Computer Vision*, vol. 102, no. 1, pp. 198–220, 2013.

[11] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok, "A comprehensive performance evaluation of 3d local feature descriptors," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 66–89, 2016.

[12] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3d recognition and pose using the viewpoint feature histogram," in *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, Oct 2010, pp. 2155–2162.

[13] A. Aldoma, M. Vincze, N. Blodow, D. Gossow, S. Gedikli, R. B. Rusu, and G. Bradski, "Cad-model recognition and 6dof pose estimation using 3d cues," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, Nov 2011, pp. 585–592.

[14] A. Aldoma, F. Tombari, R. B. Rusu, and M. Vincze, *Pattern Recognition: Joint 34th DAGM and 36th OAGM Symposium, Graz, Austria, August 28-31, 2012. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. OUR-CVFH – Oriented, Unique and Repeatable Clustered Viewpoint Feature Histogram for Object Recognition and 6DOF Pose Estimation, pp. 113–122.

[15] A. Aldoma, F. Tombari, J. Prankl, A. Richtsfeld, L. D. Stefano, and M. Vincze, "Multimodal cue integration through hypotheses verification for rgb-d object recognition and 6dof pose estimation," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, May 2013, pp. 2104–2111.

[16] W. Wohlkinger and M. Vincze, "Ensemble of shape functions for 3d object classification," in *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, Dec 2011, pp. 2987–2992.

[17] J. P. Lima, F. Simões, H. Uchiyama, V. Teichrieb, and E. Marchand, "Depth-assisted rectification for real-time object detection and pose estimation," *Machine Vision and Applications*, vol. 27, no. 2, pp. 193–219, 2016.

[18] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," *Proc. SPIE*, vol. 1611, pp. 586–606, 1992.

[19] A. J. Trevor, S. Gedikli, R. B. Rusu, and H. I. Christensen, "Efficient organized point cloud segmentation with connected components," *Semantic Perception Mapping and Exploration (SPME)*, 2013.

[20] A. Richtsfeld, T. Mrwald, J. Prankl, M. Zillich, and M. Vincze, "Segmentation of unknown objects in indoor environments," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, pp. 4791–4796.

[21] S. Holzer, R. B. Rusu, M. Dixon, S. Gedikli, and N. Navab, "Adaptive neighborhood selection for real-time surface normal estimation from organized point cloud data using integral images," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct 2012, pp. 2684–2689.

[22] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, Oct 2011, pp. 127–136.

[23] M. Korn, M. Holzkothen, and J. Pauli, "Color supported generalized-icp," in *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, vol. 3, Jan 2014, pp. 592–599.