Multimodal Cue Integration through Hypotheses Verification for RGB-D Object Recognition and 6DOF Pose Estimation

A. Aldoma[1] and F. Tombari[2] and J. Prankl[1] and A. Richtsfeld[1] and L. Di Stefano[2] and M. Vincze[1]

***Abstract*— This paper proposes an effective algorithm for recognizing objects and accurately estimating their 6DOF pose in scenes acquired by a RGB-D sensor. The proposed method is based on a combination of different recognition pipelines, each exploiting the data in a diverse manner and generating object hypotheses that are ultimately fused together in an Hypothesis Verification stage that globally enforces geometrical consistency between model hypotheses and the scene. Such a scheme boosts the overall recognition performance as it enhances the strength of the different recognition pipelines while diminishing the impact of their specific weaknesses. The proposed method outperforms the state-of-the-art on two challenging benchmark datasets for object recognition comprising 35 object models and, respectively, 176 and 353 scenes.**

I. Introduction

Objects in a domestic environment come in all sorts of shapes, sizes and colors. Some objects are denoted by a particular shape that make them highly distinguishable, while others can be singled out based on their texture. Humans exploit such properties in order to recognize and localize objects to carry out specific tasks, and can do this in an extremely efficient way. The ability of recognizing specific object instances is also key to autonomous robots that need to operate in domestic environments such as our homes.

The recent advent of sensing devices providing dense 3D reconstruction - even on untextured surfaces - enhanced with color information (*RGB-D* data) allows robots to deploy cues similar to those used by humans for the task of identifying objects. Moreover, the availability of 3D data allows both recognizing object instances in a scene together as well as estimating their 6 Degree-Of-Freedom (6DOF) pose (position and orientation), thus potentially enabling precise manipulation of objects.

However, despite the benefits of newly developed sensing technology, there are still several challenges that need to be taken into account when designing a recognition system. The possibility of processing multimodal (range and color) data at high frame rates represents an important advantage, which is though only partially exploited by current recognition methods due to, on the one side, the relatively low efficiency of most methods and, on the other, the typical approach of most methods of separately working either on color or on geometry. In particular, specializing the recognition skills of one method on a specific modality or data characteristic (e.g. shape, color or texture) tends to reduce the generalization capabilities of such systems to different environments. For instance, algorithms based on *local* image descriptors perform well on textured objects but tend to fail on low-textured ones. 3D local descriptors perform well on objects with locally rich geometrical structure but do not handle effectively simple objects that present repetitive structures. On the other hand, global and semi-global descriptors [1]–[3] make use of the global properties of the surface shape but require segmentation and their performance decreases notably when objects undergo occlusions. These examples show how difficult it becomes to design a system general enough to deal with many diverse scenarios, especially under the constraint of limited computational resources.

In this paper, we propose an object recognition and pose estimation algorithm that synergically exploits the information provided by different data cues. Specifically, several recognition pipelines based on 2D local features and 3D semi-global and local features are run in parallel and independently, so as to generate object hypotheses (identity and 6DOF pose). These hypotheses then undergo a common verification stage, which aims at retaining those that are consistent with the observations while removing those that do not explain the scene accurately enough (see Figure 1). Furthermore, novel contributions with respect to previous work related to specific stages of the proposed algorithm are also proposed throughout the paper. In this respect, the main contributions are two-fold. On the one side, we improve the robustness and recognition capabilities of the semi-global pipeline based on the OUR-CVFH descriptor proposed in [3] by exploiting RGB information and multiple clustering stages. On the other, we improve the effectiveness of the Hypothesis Verification stage proposed in [4] by exploiting additional cues that leverage on color information and environment constraints. The proposed recognition system outperforms the state of the art on two challenging benchmark datasets proposed for the ICRA 2011 Solutions in Perception challenge.

II. Related work

A variety of methods do exist in literature concerning the problem of object recognition and pose estimation. They can be organized in image based methods [6]–[9], 3D surface based methods (global and semi-global methods like [1]–[3] or local methods such as [4], [10], [11]) and multimodal approaches using color and surface cues [12], [13].

Due to the recently published results on the public benchmark datasets related to the ICRA 2011 Solutions in Percep-

[1]A. Aldoma, J. Prankl, A. Richtsfeld and M. Vincze are with Vision4Robotics - ACIN, Technical University of Vienna, Austria aldoma, prankl, arichtsfeld, vincze at acin.tuwien.ac.at

[2]F. Tombari and L. Di Stefano are with DISI, University of Bologna, Italy federico.tombari, luigi.distefano at unibo.it

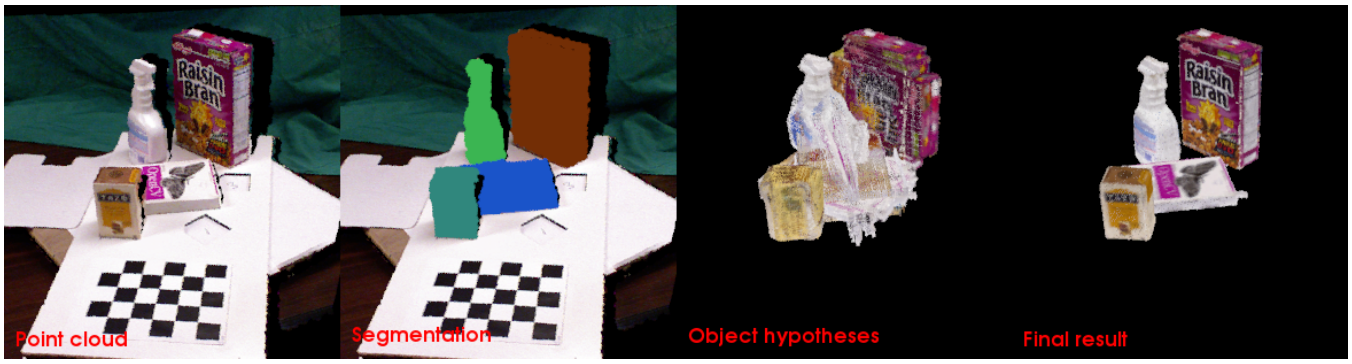


Fig. 1: Different stages of the recognition pipeline. From left to right: the point cloud obtained from the Kinect sensor, the segmentation results obtained with the method proposed by Richtsfeld et al in [5], the object hypotheses generated by the proposed recognition pipelines and the final objects selected by the hypothesis verification stage. The final result is a consistent representation of the scene with correct identification objects and accurate 6DOF pose estimation.

tion challenge [14], the method by Tang et al [6] is particularly relevant to this work as it offers a direct performance comparison. The authors present a RGB-D recognition and 6DOF pose estimation pipeline for textured objects based on global color histograms (to trim the possible identities of the object) and SIFT image features, backprojected to the 3D coordinate system of the model, to ultimately detect the object and estimate its 6DOF pose. However, unlike our method, solely SIFT correspondences are deployed to estimate the pose of the recognized object instances. The method proposed in this paper can estimate the 6DOF pose based on any of the deployed features, therefore holding the potential to recognize poorly textured objects as well as to yield accurate pose estimations in a wider range of scenarios.

III. System overview

As previously mentioned, our recognition system is based on three different pipelines that take advantage of the multimodality of the data:

- A semi-global 3D descriptor representing an extension of the OUR-CVFH approach [3] based on the color, shape and object size cues. Regarding the segmentation stage required by the semi-global pipeline, we propose the use of two different strategies recently proposed in [5], [15].
- A 2D local descriptor (SIFT [8]) which is able to generate object hypotheses with associated 6DOF pose by back-projection of the 2D keypoint locations into the 3D space.
- A 3D local descriptor (SHOT [10]) aimed at establishing correspondences between model and scene surface patches.

Figure 2 sketches the proposed algorithm by showing the various stages therein and the way the three different pipelines are merged together, ending up into a final Hypothesis Verification stage which is in common with all pipelines. As usual for recognition systems, our system consists of a training stage, where models of the objects to be recognized are learned (outlined in the next section) and of an online stage dealing with the identification and pose estimation of objects in the scene (described in Sections V and VI).

IV. Offline stage: training

In order to deploy the aforementioned pipelines, we first need to gather some information (*object model*) about the objects we would like the system to recognize. The easiest way to gather such information is to *look* at the objects from different perspectives to obtain evidence of the appearance and shape of the object of interest as seen from those perspectives. In our case, with the use of recent sensing devices like the Kinect, such process results in a set of RGB-D images covering a 360° angle around the object.

For our recognition system to function properly, we need to process the set of RGB-D images to obtain (i) a full 3D point cloud with RGB information - $\mathcal{M}_i$ - fusing the partial surface contained in each RGB-D frame and (ii) create a compact representation of the appearance of the objects, in terms of visual and shape features as well as their location relative to the coordinate system on which $\mathcal{M}_i$ is embedded. Repeating this process for all objects results in a model library $\mathcal{M}$.

To reconstruct a full 3D cloud, $\mathcal{M}_i$, out of a set of ordered views, a tracking approach is used, where features are extracted and matched along the frames in order to report the 3D points of each frame to a common reference system. Motivated by *KinectFusion* [16], tracked frames are aligned with a global model using Iterative Closest Point (ICP). The global model where the surface points are accumulated during reconstruction is represented as a voxel grid with a 3mm voxel size. As reconstruction proceeds and new views are processed, the weight at each voxel increases if new surface measurements vote for that voxel, while in case a view ray passes a voxel the weight is decreased indicating that previous surface measurements in that voxel might be incorrect. Hence, correct measurements are accumulated and filtered by their mean and wrong points are deleted.

An initial guess for the ICP alignment is provided by tracking the last segmented view of the object to the current

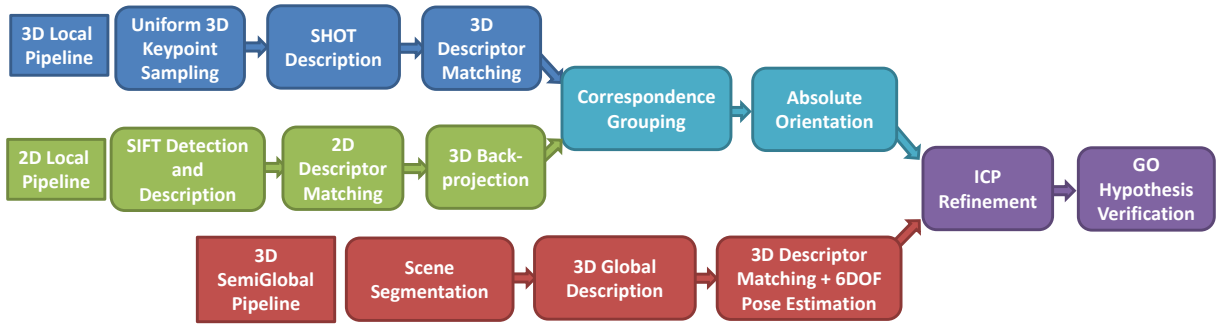


Fig. 2: The proposed 3D Object Recognition algorithm is based on 3 different recognition pipelines which are then merged together at the Hypothesis Verification stage. In particular, local correspondences coming from the 2D and 3D local pipeline are merged together at the Correspondence Grouping stage to try increasing the desired consensus between scene and model.

frame using SIFT keypoints and a rigid transformation estimated from SIFT correspondences by means of RANSAC, which results to be useful when training data is sparse. However, the *Challenge* and *Willow* datasets include a small number of RGB-D frames per object (one frame every $10°$). Furthermore, some object views contain as few as 100 points and very low texture (side view of flat and small objects). This makes pose estimation between consecutive frames via SIFT keypoints not feasible. Hence, in this case we exploit the presence of a checkerboard pattern (visible in all frames of the training dataset) to compute the initial guess pose transformation for the ICP refinement with the partial reconstruction of $\mathcal{M}_i$. Repeating this process for all available data, results in a full 3D cloud, each point associated with RGB information (see Figure 3).

Once this cloud is computed, views from uniformly sampled viewpoints are successively rendered in order to satisfy the requirement of global pipelines to deal with viewpoint-dependent, uniformly sampled model views [3] on which the semi-global descriptors (Section V-B) can be directly computed. Local pipelines (Section VI), can be trained directly on the original views, due to their stronger invariance to viewpoint changes. To do so, at each new frame we evaluate whether a large percentage of the local descriptors (SIFT and SHOT in our case) learned so far can be matched in the current frame. If this is not the case, then the frame is marked as a *keyframe*, meaning that it provides valuable unseen information, and the position of the features (extracted on the current view of the object) are transformed to the coordinate system of $\mathcal{M}_i$.

To terminate the training stage, after the selection of the keyframes, the descriptors sparsely representing the object are clustered using the reciprocal nearest neighbor (RNN) algorithm proposed by Leibe [17] resulting in two separate codebooks (one for SIFT and one for SHOT) that will be used during the online recognition stage to establish correspondences between the models and the scene. Each codebook entry represents a set of occurrences associated with a representative descriptor, each occurrence associated to the 3D position of the descriptor on $\mathcal{M}_i$'s surface.

V. RGB-D SEMI-GLOBAL PIPELINE

A. Segmentation

To segment the objects in the scene in order to deploy the semi-global pipeline, our recognition framework is equipped with two alternative segmentation methods:

The first, based on [15], is a simple but highly efficient two step strategy: (i) multi-plane segmentation of the scene and (ii) connected component clustering of points above any detected plane[1]. To efficiently compute planar regions in a scene, it uses a connected components strategy where neighboring pixels are considered to be in the same component (planar region in this case) if the dot product of their normals and the euclidean distance between the points are within a certain range. The found planar regions are further analyzed to merge regions that share the same planar model and were not detected during the first stage due to the constrained 4 neighborhood search. The second step performs similarly to the first, and groups points (without taking into consideration the points belonging to the detected planes) in the same component if their euclidean distance is smaller than τ. The resulting components form the object hypotheses provided

[1]Only planes with a certain amount of inliers (i.e, 10000) are considered to provide enough support.

Fig. 3: 10 of the 35 object models reconstructed with proposed method. The models are quite accurate but, due to the sparsity of the training data, they have only a resolution of 3mm and present some artifacts on thin areas.

to the recognition pipeline. Such a segmentation strategy assumes that the objects to be recognized will lie on a planar surface and that points belonging to different objects are at least two pixel away in a Manhattan world or farther away than τ. For future reference, we will refer to this method as *MPS*.[2]

The second one, introduced recently by Richtsfeld et al. in [5], is a generic segmentation method for unknown objects. This method pre-segments RGB-D data using a recursive normal clustering approach to extract continuous surface patches before planes and B-spline surfaces are fitted, generating parametric models of the patches. Model selection with Minimum Description Length (MDL) chooses, in a merging procedure, whether a plane or a B-spline model fits better to the patches and delivers the best model representation for a given point cloud. Relations between parametric models can be found by taking into account the principles of perceptual organization. Support vector machines (SVM) are learning this principles during a training period that avoids the reduction of the segmentation framework to model matching. Finally a graph is built, consisting of surface models as nodes and predictions from the SVM's as edges, and a globally optimal segmentation solution can be found even if single predictions are wrong.[3] The perceptual grouping rules are generic for different datasets allowing to use the previous learned rules from the *Object Segmentation Database*[4].

B. OUR-CVFH

The Oriented, Unique Repeatable Clustered Viewpoint Feature Histogram (OUR-CVFH) was recently introduced in [3] as an alternative and improvement to CVFH [2] focusing on two aspects: (i) the discriminative power of the CVFH descriptor and (ii) a method to directly estimate the 6DOF pose of the objects simultaneously to descriptor matching, without the need for additional stages. The basic idea behind OUR-CVFH is to define several repeatable Reference Frames (RFs), one for each smooth patch of the object surface, in order to spatially orient the description of the object surface relatively to each reference frame. To estimate the 6DOF pose of an object, the RFs associated to the descriptor of a scene segment and that of a training view are aligned one to another, this directly allowing the retrieval of the 6DOF pose of the model in the scene. In [3] we show how these modifications clearly improve the discriminative power of the feature as well as the pose estimation performance. An interesting property of these RFs regards their *semi-global* approach: the principal directions are computed relatively to just a small surface patch, while the signs and the relative ordering of the three unit vectors are selected via a disambiguation stage that takes into account the whole object surface. It is worth noting that a smooth patch refers here to a subset of the object surface, and smooth represents continuity in both the point coordinate and the normal domains (see [2], [3] for details on the smooth clustering strategy). Figure 4 depicts the smooth patches and the associated RFs obtained on a surface.

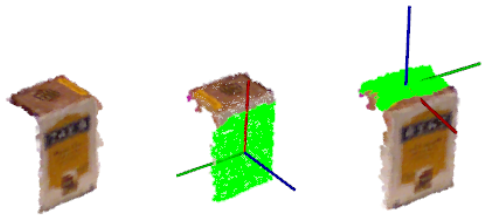


Fig. 4: Left: segmented cluster from a scene. Middle and Right: smooth patches (green) and the associated Reference Frames used to compute color and shape distributions; in this specific example, two OUR-CVFH descriptors are computed for the object on the left.

Our first proposed contribution in this aspect is related to the use of the RFs in order to describe the color properties of the surface. Hence, likewise the shape distributions in OUR-CVFH, 8 color distributions are computed. The points used to compute each color distribution are obtained by the natural division defined by the octants of the RF. Each color distribution is obtained from the YUV values associated with each point and binned into a $2\times8\times8$ grid. A coarser binning for the Y channel with respect to U and V is desired in order to increase robustness with respect to illumination changes. Similar to the L1-shape distributions in [3], we apply a trilinear interpolation on the color distributions to account for small perturbations in the RF. The 8 color distributions are appended at the end of the OUR-CVFH histogram resulting in a feature dimensionality of $303 + 8 \times 128 = 1327$.

A second proposed contribution to OUR-CVFH is related to the estimation of the smooth patches on the surface of an object which are the basis for the RF estimation. OUR-CVFH provides an accurate description and pose estimation thanks to the repeatable RFs computed on both model views and scene segments. Unfortunately, the repeatability of the RF might be compromised due to noisy or missing parts that are often present in data acquired by RGB-D sensors. In order to increase the repeatability of the RF, we propose to smooth the recognition surface with an adaptive Moving Least Squares (MLS) algorithm, similarly to [3] which smooths and upsamples the resolution of the data to ensure that model views and scenes share the same resolution. Unlike [3], we use here a multi-parametric smooth clustering stage, whereby different clustering instances are run on the same data, each with a different parameter set[5]. Figure 5 shows the effect of different parametrization for the smooth clustering

[2]In the experiments, all objects to be recognized are found on a single table-top plane. In such situations to reduce the computational time of the subsequent recognition stages, only the points above the highest plane in the direction of the normal of the largest plane are considered for the second step of *MPS*.

[3]During the experiments and for the sake of efficiency, the segmentation was applied to the points above the table-plane selected by *MPS*.

[4]http://www.acin.tuwien.ac.at/?id=289

[5]Concretely, we run 3 instances of the smooth clustering varying the maximum curvature (t_c in the notation from [3]) accepted at a point to be considered part of a smooth cluster; ($t_c \in (0.015, 0.02, 0.035)$. During training, t_c is fixed to 0.015.

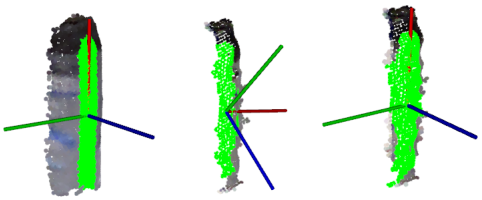


Fig. 5: Left: model view of object "19" and its associated RF. Middle and Right: scene segment relative to the same object with two associated RFs, yielded by two different clustering parameterizations. Despite the amount of noise and missing points, the RF on the right is repeatable enough to provide a correct match.

stage. Please note, that each clustering instance might yield a different set of smooth regions, this in turn resulting in a different set of RFs and descriptors. This results in a higher number of descriptors representing the same object surface but encoding it differently; e.g., a surface made up by 2 smooth patches might end up being associated with 16 descriptors due to different clustering parametrization as well as ambiguities in the disambiguation stage [3].

C. Recognition and 6DoF pose estimation

Once we have computed several semi-global features for a segmented object in the scene, the resulting descriptors are independently matched against the descriptors representing our training data using the metric proposed in [2]. In order to avoid an explosion on the number of hypotheses being generated and to remove some wrong hypotheses early in the pipeline, we filter after matching the hypotheses associated with descriptors whose distances to the segmented scene object's descriptor are smaller than 0.85 relative to the best matching model descriptor, as well as duplicated hypotheses for object surfaces that result in very similar RFs under different clustering parameterizations[6]. The RFs of remaining hypotheses after this preliminary filtering stage are used to estimate a 6DoF pose by aligning the RF pair associated with the model and segmented object descriptors.

VI. 2D and 3D local pipelines

The use of *local* - i.e. whose support is limited to a small neighborhood around the keypoint - descriptors is motivated mainly by the need to deal with the presence of clutter and occluded objects. The two descriptors included in the proposed algorithm are: (i) SIFT [8], aimed at texture-rich image patches, and (i) SHOT [10], for objects with distinctive 3D shapes. In both cases, standard recognition pipelines are deployed based on keypoint detection, description and matching. Once correspondences are determined in both the 2D and 3D domains, they are merged together (2D keypoint coordinates on the image plane are backprojected to the 3D space by means of depth information); this super-set is then fed to a unique Correspondence Grouping algorithm based on geometric consistency between pairs of correspondences [4], whose goal is to cluster correspondence subsets providing consensus for a specific object hypothesis in the scene, while discarding outliers (i.e. isolated correspondences).

Differently from [4], and as introduced in Section IV, during the training stage model descriptors are clustered together in order to form a descriptor codebook (one for SIFT features, and one for SHOT features). This codebook is then used during the recognition stage to associate to each scene descriptor its nearest-neighbor entry in both codebooks and, in turn, all model descriptors that were associated with the codeword, this operation substituting the standard descriptor matching stage between model and scene descriptors. The main advantage of this approach is computational efficiency, due to the need of searching in a size-limited codebook rather than over the set of all model descriptors, as well as the possibility of better determining correspondences in the case of symmetrical structures and repetitive surface patches.

VII. Hypotheses Verification

The Hypothesis Verification (HV) stage aims at analyzing object hypotheses previously generated along the recognition pipeline so as to reject false detections by enforcing geometrical constraints between models and scenes [4], [11], [18]. Recently, a HV method has been proposed [4] - referred to hereinafter as *Global Optimization for HV* (GO) - which, unlike other approaches, is based on an optimization framework that simultaneously takes into account all object hypotheses in order to handle interactions between them, yielding a solution globally consistent with the scene. GO has shown a peculiar ability to detect "weak" (i.e. supported by a small number of points, such as the case of highly occluded objects) correct hypotheses while filtering out a high number of false positives, thus moving the operating point of the recognition system toward a higher recall without sacrificing precision [4].

According to the notation used in [4], the proposed recognition pipelines generate a set of n recognition hypotheses $\mathcal{H} = \{h_1, \cdots, h_n\}$, each hypothesis h_i given by the pair $(\mathcal{M}_{h_i}, \mathcal{T}_{h_i})$, with $\mathcal{M}_{h_i}$ being the model associated to h_i and $\mathcal{T}_{h_i}$ being the transformation which relates $\mathcal{M}_{h_i}$ to $\mathcal{S}$, $\mathcal{S}$ being the point cloud representation of the scene. Hence, the goal of the HV stage is to choose an arbitrary (up to n) number of elements belonging to $\mathcal{H}$ in order to maximize the number of correct recognitions (TPs) while minimizing the number of false positives (FPs). The GO algorithm relies on minimizing a suitable *cost* function defined over the solution space of the HV problem. In particular, we denote a solution as a set of boolean variables $\mathcal{X} = \{x_0, \cdots, x_n\}$ having the same cardinality as $\mathcal{H}$, with each $x_i \in \mathbb{B} = \{0, 1\}$ indicating whether the corresponding hypothesis $h_i \in \mathcal{H}$ is discarded/accepted (i.e. $x_i = 0/1$). Hence, the *cost* function can be expressed as $\mathfrak{F}(\mathcal{X}) : \mathbb{B}^n \rightarrow \mathbb{R}$, $\mathbb{B}^n$ being the solution space, of cardinality 2^n. A polynomial-time resolution of the

[6]Equal clusters resulting from different parameterizations might as well be filtered based on the similarity of the clusters itself in order to avoid repetitive computations of the descriptor. Nevertheless, the computation of the descriptor is fast enough so we did not consider this in the scope of the paper

optimization problem is provided by means of Simulated Annealing. The cost function includes four different cues: i) scene fitting (how well a hypothesis is supported by scene points, term $\Omega_{\mathcal{X}}(p)$; ii) model outliers (how many model points are left unexplained, term $f_{\mathcal{M}}(\mathcal{X})$; iii) multiple assignment (how many scene points are simultaneously associated to different hypotheses, term $\Lambda_{\mathcal{X}}(p)$; iv) clutter (how well the hypothesis fits to neighboring scene regions, term $\Upsilon_{\mathcal{X}}(p)$). For more details, we refer the reader to [4].

As previously introduced, we proposed to merge and optimize in the HV stage the hypotheses generated by the three pipelines employed by the proposed approach. Moreover, given the multimodal nature of the available data, and in line with the contributions proposed for the semiglobal descriptor, we also provide an extension of the GO algorithm to color cues. The optimization framework proposed in [4] is particularly flexible to handle additional cues for the global cost function. For this reason, we propose to add a fifth cue which measures how well each scene point *explains* its corresponding model point - according to a certain hypothesis - within the color domain. This novel term is thus inherently related to the inlier weighting cue, but only taking into account color similarity. Following the notation in [4], for each hypothesis h_i and each associated model point p we thus compute a weight $\omega_{h_i}^{C}(p, \mathcal{N}(p))$ defined as follows

$$\omega_{h_i}^{C}(p, \mathcal{N}(p)) = exp\left(-\frac{\|\kappa(p) - \kappa(\mathcal{N}(p))\|_2}{2\sigma_C^2}\right) \quad (1)$$

where $\mathcal{N}(p)$ is the nearest-neighbor of p on the scene, and $\kappa(p)$ is the 3D vector representing the color triplet associated with point p in the YUV space. In order to increase robustness to illumination changes, the weight of the Y channel is reduced by 2. The final cost function is then given by

$$\mathfrak{F}(\mathcal{X}) = f_{\mathcal{S}}(\mathcal{X}) + \lambda \cdot f_{\mathcal{M}}(\mathcal{X}) + f_{\mathcal{C}}(\mathcal{X}) + f_{\mathcal{E}}(\mathcal{X}) \quad (2)$$

where λ is a regularizer aimed at penalizing model outliers, and $f_{\mathcal{S}}$, $f_{\mathcal{M}}$ account, respectively, for geometrical cues defined on scene points and model points:

$$f_{\mathcal{S}}(\mathcal{X}) = \sum_{p \in \mathcal{S}} \left(\Lambda_{\mathcal{X}}(p) + \Upsilon_{\mathcal{X}}(p) - \Omega_{\mathcal{X}}(p)\right) \quad (3)$$

$$f_{\mathcal{M}}(\mathcal{X}) = \sum_{i=1}^{n} (|\Phi_{h_i}| \cdot x_i) \quad (4)$$

Φ_{h_i} representing the set of *outliers* (model points without a counterpart in the scene — see [4] for further details) in the i-th hypothesis. $f_{\mathcal{C}}$ evaluates the color registration between model and scene:

$$f_{\mathcal{C}}(\mathcal{X}) = \sum_{i=1}^{n} (|\Psi_{h_i}| \cdot x_i) \quad (5)$$

where Ψ_{h_i} represents the color registration quality between the i-th hypothesis and the scene and is defined as follows:

$$\Psi_{h_i} = \sum_{p=1}^{m} \left(1 - \omega_{h_i}^{C}(p, \mathcal{N}(p))\right) \quad (6)$$

where m is the number of points of the model associated with h_i currently being *explained* by the scene.

Finally, $f_{\mathcal{E}}$ considers physical constraints:

$$f_{\mathcal{E}}(\mathcal{X}) = \sum_{i=1}^{n} u(h_i) \cdot x_i + w_o \cdot f_o(h_i) \cdot x_i \quad (7)$$

where $u(h_i)$ is the sum of model points associated with h_i being under the table plane, w_o is a penalization term and $f_o(h_i)$ is a boolean function indicating whether the i-th hypothesis is in contact with the table or not. The optimization framework allows to add a pool of such physical constraints that improve robustness based on application and environment knowledge that might prove to be particularly useful in robotic applications as well as industrial applications under controlled situations.

VIII. Experimental results and discussion

In order to validate the effectiveness of the system, we present experimental results on two large benchmark RGB-D object recognition datasets proposed for the ICRA 2011 *Solutions in Perception* challenge organized by Willow Garage. Following the naming conventions introduced in [6], we refer to the first test dataset as *Challenge* and to the second one as *Willow*. Additionally, we present 6DOF pose estimation results on the *Challenge* dataset for which the ground truth concerning poses is available. To give the reader a better intuition about the different modules of the system, we present also results on the *Challenge* dataset provided by each of the individual pipeline deployed by the proposed system.

A. Datasets

The *Challenge* test dataset is composed of 434 object instances organized in 39 scene sequences. Each scene contains from 1 up to 5 object instances and does not contain any object instance outside of 35 models used to train the system. It was carefully created to avoid including objects undergoing strong occlusions. Differently, the *Willow* dataset is significantly more complex. It shares the same model library as the *Challenge* dataset, but it includes distracting objects with shapes and colors similar to those in the 35 objects to be recognized as well as several occluded objects. In this case, some sequences were recorded under saturated illumination causing some objects — especially those with metallic parts — to present also several artifacts in the point cloud acquired with the Kinect sensor. This dataset contains a total of approximately 1500 object instances to be recognized.

B. Results

Table I presents precision and recall results on the *Willow* and *Challenge* datasets obtained by the method proposed in this paper, as well as those reported by Tang et al. in [6]. Our method performs better in both datasets, with a remarkable accuracy on the *Challenge* dataset where just a single object out of the 434 instances was confused with a similar one (see Figure 7-(a)) and overall yielding 1 FP

	Precision	Recall
Willow (Proposed system)	**94.30**%	**70.86**%
Willow (Tang et al. [6])	88.75%	64.79%
Challenge (Proposed system)	**99.77**%	**99.77**%
Challenge (Tang et al. [6])	98.73%	90.23%
Challenge (Aldoma et al. [3] + Richtsfeld)	92.79 %	85.94%

TABLE I: Precision and recall results for the *Willow* and *Challenge* datasets.

	Precision	Recall	Time
Proposed system	99.77%	**99.77**%	6.44 [s]
RGB/3D global + Richtsfeld	99.77%	99.31%	5.23 [s]
RGB/3D global + MPS	99.77%	98.39%	3.88 [s]
2D/3D local (table plane)	98.47%	88.92%	3.45 [s]
2D local (table plane)	**100.00**%	71.43%	**1.72** [s]

TABLE II: Precision and recall results for the *Challenge* dataset with different combinations of the pipelines of the proposed system. Reported results are obtained with 10 ICP iterations, $sigma_s = 1cm$ (inlier threshold), $\sigma_C = 35$(color sigma), $\lambda = 1$(model outliers weight) and $\kappa = 1$(clutter weight). Average execution time per scene is also reported.

and 1 FN. The last row of Table I reports also a comparison with the 3D-only semi-global pipeline proposed in [3] which reports a worse performance, thus validating the usefulness of the contributions proposed in this paper.

The lack of strong occlusions in the *Challenge* dataset allows leveraging on the power of the RGB/3D semi-global pipeline as outlined in Table II where such a pipeline performs almost as good as the combination of the three recognition pipelines. Because the objects in this dataset are easy to segment, the *MPS* segmentation delivers similar results to that of Richtsfeld et al. and is slightly faster (only in the scene shown in Figure 1 the MPS segmentation can not separate the book from the spray bottle). Table II reports also the average execution time required by the different pipelines per scene. In this aspect, the 2D local pipeline is the fastest due to the GPU implementation used for SIFT keypoint detection, descriptor computation and histogram matching as well as the fact that this pipeline results in a much smaller number of hypotheses, compared to the semi-global pipeline. The 2D/3D local pipeline performs relatively well with an improvement of almost 20% over the 2D pipeline alone.

The improvement with respect to [6] on the *Willow* dataset is approximately 6% in terms of both Precision and Recall. Due to the high number of occluded object instances, the performance of the semi-global pipeline is not as good as on the *Challenge* dataset and local pipelines do not turn out as effective as expected (see Table III) due to the noisy point clouds, different illumination conditions and some object instances with no more than a few points being visible. Due to the distracting objects, we used stricter HV parameters[7] that might remove correct hypotheses with an

[7]We increase penalization factors in the HV stage for model outliers and clutter points ($\lambda = 1.5$ and $\kappa = 2$) and reduce the inlier threshold ($sigma_s = 7.5mm$). The rest of the parameters remain unchanged.

	Precision	Recall
Proposed system	94.30%	**70.86**%
RGB/3D global	93.70%	58.92%
2D/3D local	**98.4**%	56.20%

TABLE III: Precision and recall results for the *Willow* dataset motivating the fusion of different pipelines to handle challenging scenarios.

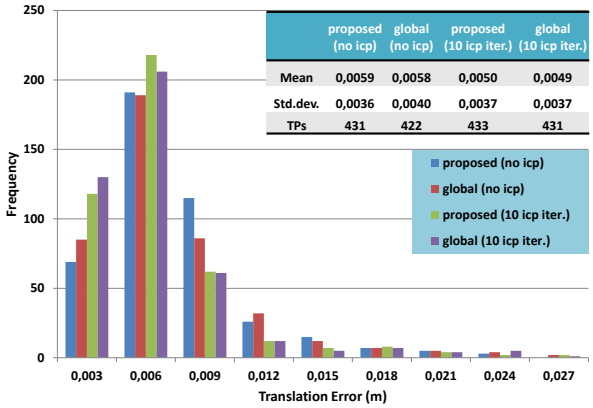


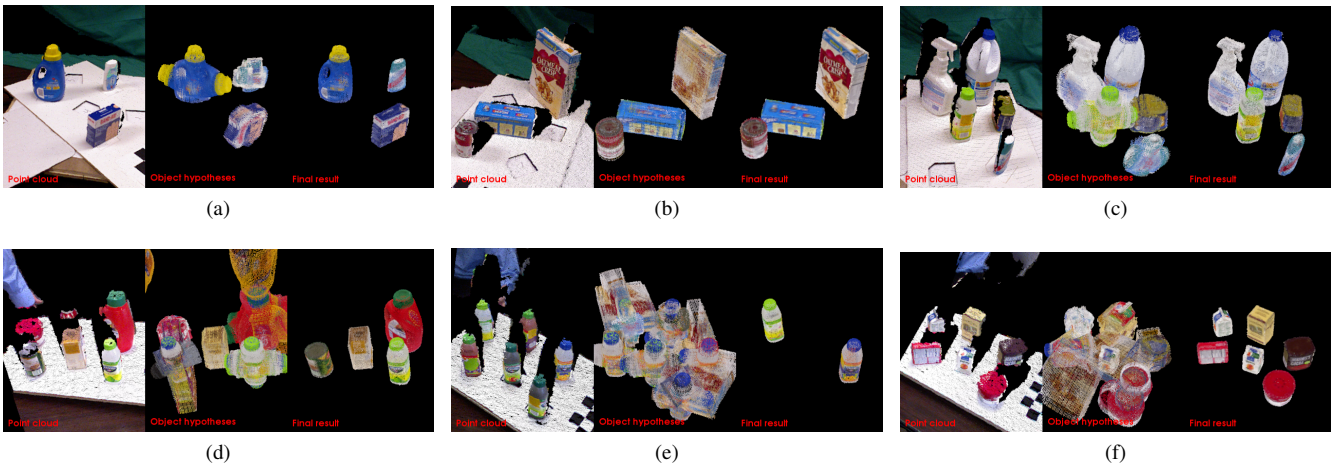
Fig. 6: Histogram of translation errors for the proposed system as well as for the semiglobal pipeline. Results reported with both 0 and 10 ICP iterations.

inaccurate registration. An interesting additional experiment would consist in reporting recognition results with respect to occlusion level. Such a measure is commonly used in the 3D community [4], [18], but it requires an accurate ground truth pose, that was not available for the *Willow* dataset, in order to estimate the percentages of occlusion for each object.

Figure 6 summarizes the results regarding 6DoF pose estimation as well as the behavior of the semi-global pipeline and the proposed system with a different number of ICP iterations. Observe how the proposed system without the pose refinement stage, performs slightly better than the semi-global pipeline alone with 431 and 422 TPs respectively. Overall, the translation errors between the model centroids of the recognized pose and the groundtruth centroids are notably low, being mostly between 0 and $0.01m$ and always less than $0.03m$ - e.g. they favorably compare to those reported in [6], where translation errors are mostly between 0 and $0.05m$ and can get up to $0.2m$. We also computed two additional values regarding pose, the RMS error for shape and color, with respectively a mean and standard deviation of $0.003[m] \pm 0.0017$ and $36.12 \pm 13,74$ (these last values referred to the RGB space normalized between [0, 255]).

IX. Conclusions

We have presented a modular recognition and 6DoF pose estimation system exploiting three different recognition pipelines that take advantage of the multimodal nature of the data provided by recent RGB-D sensors. We have shown how a hypothesis verification stage offers a good opportunity to fuse results from the several pipelines and proposed several



(a) (b) (c)

(d) (e) (f)

Fig. 7: Recognition results obtained with the proposed method concerning the *Challenge* dataset (left column) and the *Willow* dataset (right column). (a) The single False Positive yielded by our method: a very similar object with respect to the correct one is recognized. Observe in (e), how our method is able to correctly recognize the two Odwalla bottles, while the four distractors, having same shape but different texture, yield object hypotheses that are not consistent in terms of color and are thus discarded.

formulations of the HV stage to exploit color information as well as environment constraints. Additionally, we have presented a semi-global pipeline based on OUR-CVFH to exploit color and shape information simultaneously. The experimental evaluation on two challenging benchmark datasets demonstrates the practical applicability of the proposed approach and brings in significant improvements over the state-of-the-art.

REFERENCES

[1] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3d recognition and pose using the viewpoint feature histogram," in *Proceedings of the 23rd IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, 10/2010 2010.

[2] A. Aldoma, N. Blodow, D. Gossow, S. Gedikli, R. B. Rusu, M. Vincze, and G. Bradski, "CAD-Model Recognition and 6DOF Pose Estimation Using 3D Cues," in *Workshop: 3rd IEEE Workshop on 3D Representation and Recognition, ICCV*, 2011.

[3] A. Aldoma, F. Tombari, R. Rusu, and M. Vincze, "Our-cvfh: Oriented, unique and repeatable clustered viewpoint feature histogram for object recognition and 6dof pose estimation," in *Joint DAGM-OAGM Pattern Recognition Symposium*, 2012.

[4] A. Aldoma, F. Tombari, L. Di Stefano, and M. Vincze, "A global hypothesis verification method for 3d object recognition," in *European Conference on Computer Vision (ECCV)*, 2012.

[5] A. Richtsfeld, T. Mörwald, J. Prankl, M. Zillich, and M. Vincze, "Segmentation of Unknown Objects in Indoor Environments," in *IROS*, 2012.

[6] J. Tang, S. Miller, A. Singh, and P. Abbeel, "A textured object recognition pipeline for color and depth image data," in *In the proceedings of the International Conference on Robotics and Automation (ICRA)*, 2012.

[7] A. Collet, D. Berenson, S. S. Srinivasa, and D. Ferguson, "Object recognition and full pose registration from a single image for robotic manipulation," in *in IEEE ICRA. Kobe*, 2009, pp. 48–55.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[9] I. Gordon and D. G. Lowe, "What and where: 3d object recognition with accurate pose," 2006.

[10] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of Histograms for local surface description," in *Proc. 11th European Conference on Computer Vision (ECCV 10)*, 2010.

[11] C. Papazov and D. Burschka, "An efficient ransac for 3d object recognition in noisy and occluded scenes," in *Proc. 10th ACCV*, 2010.

[12] M. Muja, R. B. Rusu, G. Bradski, and D. Lowe, "REIN - A Fast, Robust, Scalable REcognition Infrastructure," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[13] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *ICCV'11*, 2011, pp. 858–865.

[14] "Willow Garage: Solutions In Perception Challenge, ICRA 2011," http://opencv.willowgarage.com/wiki/SolutionsInPerceptionChallenge, 2011, [Accessed on Sept. 2012.].

[15] D. Holz, A. J. B. Trevor, M. Dixon, S. Gedikli, and R. B. Rusu, "Fast segmentation of rgb-d images for semantic scene understanding."

[16] R. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, David andKim, A. Davison, P. Kohli, J. Shotton, and A. Hodges, Steveand Fitzgibbon, "KinectFusion: Real-time dense surface mapping and tracking," in *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE InternationalSymposium on*. IEEE, 2011, pp. 127–136.

[17] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization andsegmentation," *Int. J. Comput. Vision*, vol. 77, no. 1-3, pp. 259–289, 2008.

[18] A. Mian, M. Bennamoun, and R. Owens, "3d model-based object recognition and segmentation in cluttered scenes," *IEEE Trans. PAMI*, no. 10, 2006.