On the Automation of "Ground Truth" Annotation for RGB-D Object Instance Recognition Datasets

Aitor Aldoma, Thomas Fäulhammer and Markus Vincze

***Abstract*— This paper studies the problem of automatically annotating "ground-truth" for object instance recognition in RGB-D datasets. To this end, we propose to leverage the ability of sensing a scene from multiple vantage points combined with recent single-view object recognition techniques in order to create a rich and integrated representation of the environment, in the form of a 3D reconstructed scene as well as object hypotheses therein. We argue that such a representation facilitates improved recognition to an extent that the recovered results, obtained by means of a suitable 3D hypotheses verification stage, closely resemble the ground truth of the scene under consideration. The results obtained by our approach on three large datasets – with 97.9%, 99.1% and 93.2% correct annotations – support our claim that this method can effectively reduce the labour intensity of ground truth annotation.**

I. INTRODUCTION

Object instance recognition and pose estimation is a well-researched problem in 2D and 3D Computer Vision [1]–[4]. With the advent of affordable RGB-D sensors, it has received increased attention in robotics perception. Despite recent advances [5]–[8], some challenges still need to be solved in order to reliably deploy recognition in integrated robotic systems: changing lighting conditions, complex scene layouts, sensor nuisances as well as objects undergoing occlusions from a certain viewpoint, being not easy segmentable and/or not presenting unique features.

In order to evaluate progress within a certain field, the availability of challenging and varied datasets is a key element to foster research in the correct direction. While a few RGB-D datasets are available for object class [9] and object instance recognition [6], [10], more datasets are required to cover the whole spectrum of challenges. A major issue holding back the proliferation of datasets is related to their annotation being time consuming and tedious; in particular, when accurate poses for object instances are required.

While it is possible to *automate* the process by means of fiducial patterns, using such techniques results in unnatural scenes and imposes restrictions on the scene layout (e.g. table-top scenarios). For instance, in the datasets proposed for the ICRA11 Perception Challenge[1], objects are placed using fixtures on a planar surface with a checker-board pattern. Since the fixtures' position and orientation relative to the pattern are known, the pose of the objects located at each fixture can be estimated up to the accuracy of the pattern detection and fixture-pattern relative measurements. Note that this still requires a human operator to manually provide object-fixture correspondences.

Fig. 1: Annotation examples on *TUW* and *Willow* datasets using the proposed method. The exploitation of multiple vantage points facilitates accurate annotations of objects undergoing strong occlusions in complex scene layouts.

Aiming at reducing the aforementioned burden and current limitations, this paper tackles the problem of *automating* "ground-truth" annotation for RGB-D object instance recognition datasets avoiding the use of fiducial patterns. Specifically, we consider datasets composed of sequences of RGB-D frames, each frame resulting in additional viewpoints of the scene under consideration. To simplify the original recognition problem, the main idea is to exploit the supplementary information provided by multiple vantage points to build a richer and integrated representation of the scene as well as the objects therein. Under a small set of assumptions stated in Section III, we in fact claim that recognition results obtained on such a representation are close to the actual ground truth of the data. With this in mind, the main contributions in this work are related to the questions:

(i) *How to build such a representation?* We do this by deploying *single-view recognition* on each frame and by *reconstructing* a 3D representation of the sequence. We show how single-view detections in combination with visual features, provide good initial pose estimates between pairs of frames and thus result helpful for the reconstruction stage.

(ii) *How to use it in order to solve the multi-view recog-*

Aitor Aldoma, Thomas Fäulhammer and Markus Vincze are with the Vision4Robotics group (ACIN - Vienna University of Technology), Austria {aldoma, faeulhammer, vincze}@acin.tuwien.ac.at

[1]Even though the original website is no longer available, the dataset has been extensively used in the literature to evaluate recognition methods, see [5], [7], [8].

nition problem? By projecting single-view detections (*object hypotheses*) into the reconstructed scene, the problem boils down to selecting a subset of hypotheses that *best* explain the reconstructed sequence, attained in our proposal by means of a multi-view hypothesis verification stage.

While multiple viewpoints increase the probability of seeing the object from an advantageous perspective (i.e, the object becomes fully visible or a distinctive part is revealed), the integrated representation provides a stronger evidence of an hypothesis being actually present in the scene and thus, facilitates the removal of spurious single-view detections.

We used the proposed method to automatically annotate more than 95% of the 3500 object instances in two large datasets totalling 516 RGB-D frames, including many frames where some objects were largely occluded (see Fig. 1). Thus, in combination with a final manual stage to verify and extend automatic annotations, the method results useful to accurately annotate large amounts of data with a significant reduction in the amount of manual intervention.

II. RELATED WORK

Many single-view approaches towards object instance recognition and pose estimation have been proposed in the literature. Focusing on recent methods deployed on RGB-D data without aiming at an extensive review of available approaches, Xie et al. [8] proposed a dense SIFT feature extraction scheme combined with a RANSAC pose estimation stage. Each generated hypotheses is verified by means of a multimodal (color, shape, and gradients) scoring scheme. They obtain excellent performance on the datasets proposed for the ICRA11 Perception Challenge. However, the method is computationally expensive and has several assumptions on the layout of the scene as well as on the objects in it (i.e, textured objects). Targeting recognition of textureless objects, Hinterstoisser et al. propose in [6] a multimodal template (color gradients and surface normals) based matching approach. The method presents excellent real-time performance but suffers when objects become partially occluded.

Regarding the exploitation of multiple vantage points, Collet and Srinivasa [11] propose an *introspective multi-view* method to efficiently recover the identity and 6DoF poses of objects from three 2D cameras. It relies on a single-view algorithm to provide an initial estimate of objects in each camera view which get clustered and verified on a second stage simulatenously considering the results from the individual images. Contrary to our method, the relation between multiple viewpoints of the scene is obtained from a static 2D camera rig with known extrinsic parameters.

Based on single-view detections like our method but aiming at semantic labeling of 3D scenes is the work of Lai et al. [12]. The method consists of four stages: i) reconstruction of the 3D scene [13], ii) detect possible objects in each RGB-D frame, iii) project the single-view scores into the reconstructed scene and iv) enforce label consistency through a voxel-based MRF. In comparison, our method enforces global consistency by a suitable 3D hypothesis verification stage and uses shared single-view recognition results among different frames to aid during reconstruction. The advantage of object detection while mapping an environment has been recently shown by Fioraio and Di Stefano [14] within a joint detection, tracking and mapping framework.

Finally, since 3D models of the objects are available for the problem at hand, an alternative to single-view based methods is represented by directly exploiting the reconstructed 3D scene to match it against the model library by means of suitable 3D object matching techniques [2], [3], [15], [16] which might be extended to use color information [17].

III. PROPOSED APPROACH

Provided with a set of models with m point clouds $\mathcal{M} = \{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_m\}$ and a set of n RGB-D frames belonging to a sequence $\mathcal{S} = \{\boldsymbol{S}_1 \ldots \boldsymbol{S}_n\}$, the goal of the proposed method is to detect in each frame all objects known to the system together with their pose. The overall structure of the method is depicted in Fig. 2.

A. Single-view recognition

The single-view recognition generates for each scene point cloud $\boldsymbol{S}_k \in \mathcal{S}$ a set of hypotheses $\mathcal{H}_k = \left\{h_1^k, h_2^k, \ldots h_p^k\right\}$, where

$$h_j^k = \left\{o_j^k, \boldsymbol{P}_j^k\right\}, \quad 1 \leq j \leq p \tag{1}$$

describes a single hypothesis with the object identity $o_j^k \in \mathcal{M}$ and a 4×4 transformation matrix $\boldsymbol{P}_j^k$ defining the 6DoF object pose with respect to the reference frame of $\boldsymbol{S}_k$.

To this end, we deploy the recognition system proposed in our previous work [7]. The method is based on a combination of 2D and 3D (*local* and *global*) recognition pipelines aiming at exploiting the different strengths of the individual algorithms. The results gathered from the different pipelines are merged in a hypotheses verification stage aiming at finding a combination of hypotheses that best represent the scene under consideration. Thanks to the different pipelines, the algorithm does not make assumptions on the scene layout or objects in it and is thus deployable in a wide range of recognition problems. The rest of the method is independent of this stage and other single-view approaches might be deployed, provided that they retrieve a set of objects with their poses.

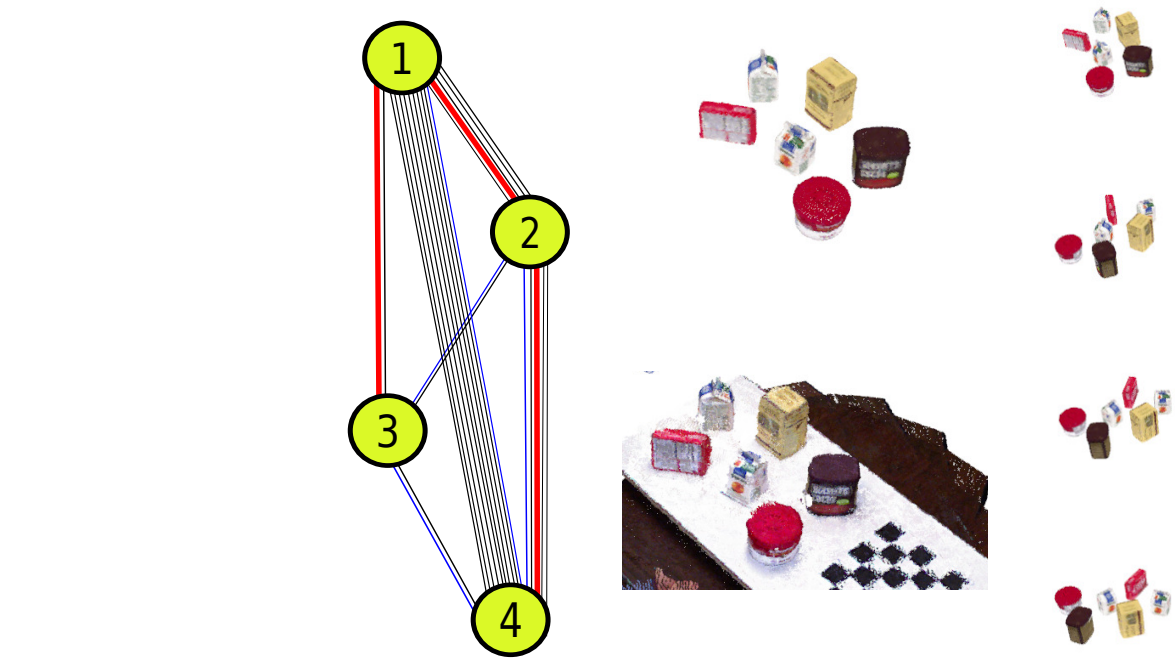
B. Multi-view graph representation

The multi-view stage starts by creating a set of vertices $\mathcal{V} = \{\mathcal{V}_1 \ldots \mathcal{V}_n\}$, where each vertex contains single-view hypotheses with their respective scene point cloud

$$\mathcal{V}_i = \{\boldsymbol{S}_i, \mathcal{H}_i\}, \quad 1 \leq i \leq n. \tag{2}$$

By iteratively comparing vertex pairs with respect to their hypotheses sets, vertices sharing a hypothesis with the same model identity o are connected by an edge

$$\mathcal{E}_{ij}^l = \left\{o_{ij}^l, \boldsymbol{T}_{ij}^l, \vartheta_{ij}^l, i, j\right\} \tag{3}$$
$$\forall i, j \mid \left(o_{ij}^l \in \mathcal{H}_i\right) \wedge \left(o_{ij}^l \in \mathcal{H}_j\right), 0 \leq l \leq n_{ij},$$



(a) (b) (c) (d) (e)

Fig. 2: Workflow of proposed ground-truth annotation generator for an RGB-D sequence of 4 frames a) input RGB-D frames; b) single-view recognition results; c) graph representation of multiple views. If the same object was recognized in two views or the views can be registered by visual features (blue edges), an edge is added to the graph connecting the views with an associated transformation and an appropriate weight. The subsequently calculated Minimum Spanning Tree is shown by thick red edges; d) integrated scene representation and verified hypotheses in common coordinate system; e) verified hypotheses back-projected to original frames, generating "ground-truth" annotations.

with an edge weight ϑ_{ij}^{l} resulting from a certain quality criteria such as described below. The number of shared hypotheses between vertices $\mathcal{V}_i$ and $\mathcal{V}_j$ is represented by the variable n_{ij}, while the relative pose between view $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$ is described by the 4×4 transformation matrix $\boldsymbol{T}_{ij}^{l}$. Given the model identity o_{ij}^{l} is shared amongst both views by hypotheses h_f^i and h_g^j, the transformation is estimated by

$$\boldsymbol{T}_{ij}^{l} = \boldsymbol{P}_f^i \left(\boldsymbol{P}_g^j\right)^{-1}, \tag{4}$$

and similarly for the transformation matrix corresponding to edge $\mathcal{E}_{ji}^{l}$,

$$T_{ji}^{l} = (\boldsymbol{T}_{ij}^{l})^{-1}. \tag{5}$$

If each vertex has a common object hypothesis with any other vertex, a fully-connected multi-view graph $\mathcal{G}$ can be described by

$$\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}, \tag{6}$$

where $\mathcal{E}$ is the set containing all edges from Equation (3).

In order to avoid isolated vertices in $\mathcal{G}$ (e.g. no recognised object) or to possibly obtain a better pairwise transformation in case of weak object pose estimates for a pair of vertices, additional edges are created by means of visual features (*scene to scene* edges). In particular, for each pair of vertices $\{\mathcal{V}_i, \mathcal{V}_j\}$, their respective SIFT features [18] are matched using a 1-NN (*first nearest neighbour*) strategy yielding a correspondence set between both frames, which is posteriorly processed by means of a correspondence grouping stage [4]. The output of the grouping stage is a set of geometrically consistent correspondences from which a rigid transformation is estimated. In our implementation, all consensus sets with more than 15 correspondences are kept and used to create an edge between $\{\mathcal{V}_i, \mathcal{V}_j\}$ effectively extending $\mathcal{E}$. In order to experimentally motivate the creation of edges based on visual features, a small experiment has been conducted evaluating the frequency of edges in the Minimum Spanning Tree originating from this source. In particular, on the *Willow* and *TUW* datasets, *scene to scene* edges were selected 33.9% and 55.4% of the times, respectively. These results indicate that *scene to scene* edges play an important role during the reconstruction stage.

In its most general form, our method does not require the order of the sequence to be provided. However, if the order is known, significant speed ups can be obtained by avoiding creating edges in the graph between frames that are too far away. In the work presented in this paper, we did not deploy any filtering scheme.

C. Edge weight and pairwise registration refinement

In order to ensure that the MST includes edges $\mathcal{E}$ representing a correct and accurate pairwise transformation, the edge weights used in Equation 3 need to be robust and representative for the quality of the estimated transformation.

According to Equation (3), $\boldsymbol{T}_{ij}^{l}$ represents the transformation aligning $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$. To accommodate for small inaccuracies of the single-view pose estimation, $\boldsymbol{T}_{ij}^{l}$ is refined by means of projective ICP prior to the computation of the weight associated to the edge. The refinement stage uses all points in $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$.

To assess the registration quality, a quality measure ω is proposed for the refined transformation. To evaluate registration of two point clouds originating from sensors with a single point of projection (such as the recent RGB-D sensors considered in this work), Huber and Hebert [19] introduced *visibility consistency* measures. For example, a *free space violation* (FSV) occurs when a point in $\boldsymbol{T}_{ij}^{l}\boldsymbol{S}_i$ blocks the visibility of another point in $\boldsymbol{S}_j$ from the sensor's origin of $\boldsymbol{S}_j$. Checking FSV for all points in $\boldsymbol{S}_i$, the FSV fraction becomes

$$f_{ij}^{l} = \frac{\left|X_{FSV}\left(\boldsymbol{T}_{ij}^{l}\boldsymbol{S}_i, \boldsymbol{S}_j\right)\right|}{\left|X_{FSV}\left(\boldsymbol{T}_{ij}^{l}\boldsymbol{S}_i, \boldsymbol{S}_j\right)\right| + \left|X_{SS}\left(\boldsymbol{T}_{ij}^{l}\boldsymbol{S}_i, \boldsymbol{S}_j\right)\right|}, \quad (7)$$

where the number of points in $\boldsymbol{T}_{ij}^{l}\boldsymbol{S}_i$ with a free space violation and points on the same surface with respect to $\boldsymbol{S}_j$ are given by $\left|X_{FSV}\left(\boldsymbol{T}_{ij}^{l}\boldsymbol{S}_i, \boldsymbol{S}_j\right)\right|$ and $\left|X_{SS}\left(\boldsymbol{T}_{ij}^{l}\boldsymbol{S}_i, \boldsymbol{S}_j\right)\right|$, respectively. Intuitively, the lower f_{ij}^{l}, the better is the registration. For an in-depth discussion regarding the FSV fraction, please see [19].

Additionally to the FSV fraction, the computation of ω accounts for the amount of overlap as well as the angle between the normals of each corresponding point pair. In general, transformation estimations of clouds with high overlap are more stable and should therefore be included more often in the MST. While the absolute amount of overlap can be approximated by $\left|X_{SS}\left(\boldsymbol{T}_{ij}^{l}\boldsymbol{S}_i, \boldsymbol{S}_j\right)\right|$, the relative overlap ζ is defined in the following by

$$\zeta_{ij}^{l} = \min\left(\frac{\left|X_{SS}\left(\boldsymbol{T}_{ij}^{l}\boldsymbol{S}_i, \boldsymbol{S}_j\right)\right|}{|\boldsymbol{S}_i|}, \zeta_{\text{max}}\right), \quad (8)$$

where the parameter ζ_{max} indicates the desired amount of overlap between clouds (0.75 in our experiments). The normals' similarity is defined by

$$\psi_{ij}^{l} = \frac{\sum_{k=1}^{|\boldsymbol{S}_i|} \boldsymbol{n}\left(\boldsymbol{p}_i^k\right) \cdot \boldsymbol{n}\left(\Gamma\left(\boldsymbol{T}_{ij}^{l}\boldsymbol{p}_i^k, \boldsymbol{S}_j\right)\right)}{|\boldsymbol{S}_i|}, \quad (9)$$

where $\boldsymbol{p}_i^k$ is the k-th point of point cloud $\boldsymbol{S}_i$, $\boldsymbol{n}(\boldsymbol{p})$ represents the normal vector of point $\boldsymbol{p}$ and $\Gamma\left(\boldsymbol{p}, \boldsymbol{S}_j\right)$ is the NN of point $\boldsymbol{p}$ in point cloud $\boldsymbol{S}_j$ – efficiently computed using projective geometry.

Combining the previous equations, ω is computed by

$$\omega_{ij}^{l} = f_{ij}^{l}\zeta_{ij}^{l}\psi_{ij}^{l}, \quad 0 \leq \omega_{ij}^{l} \leq \zeta_{\text{max}}. \quad (10)$$

Finally, the edge weight is

$$\vartheta_{ij}^{l} = \zeta_{\text{max}} - \min\left(\omega_{ij}^{l}, \omega_{ji}^{l}\right). \quad (11)$$

Fig. 3: *Left*: Screenshot of the reconstructed scene without filtering; several artifacts are observable due to axial and lateral noise. *Right*: Artifacts are removed after filtering points by means of a suitable noise model, providing a better representation for the verification stage.

D. Hypotheses extension and sequence registration

Given the graph $\mathcal{G}$ with the edge weights assigned in Subsection III-C, a subgraph $\mathcal{G}'$ is created that connects all vertices $\mathcal{V}$ without cycles and with the lowest total cost in terms of the Prim's Minimum Spanning Tree (MST) algorithm [20]

$$\mathcal{G}' = \{\mathcal{V}, \mathcal{E}'\}, \quad \mathcal{E}' \subset \mathcal{E}. \quad (12)$$

Note that since Prim's Minimum Spanning Tree algorithm is only defined for undirected edges, the set of edges defined by Equation (3) needs to be adjusted accordingly.

Describing the root of the MST by $\mathcal{V}_{\text{root}} \in \mathcal{V}$, a world coordinate system is set to the camera coordinate system of $\mathcal{V}_{\text{root}}$. Starting from $\mathcal{V}_{\text{root}}$ and traversing through $\mathcal{G}'$, the hypotheses set $\mathcal{H}_{\text{root}}$ is augmented by all hypotheses in the graph

$$\mathcal{H}_{\text{root}} \rightarrow \{\mathcal{H}'_k\}, \quad 1 \leq k \leq n, \quad (13)$$

where $\mathcal{H}'_k$ is the set of hypotheses $\mathcal{H}_k$ with pose matrices multiplied iteratively by all the edge transformation matrices from node $\mathcal{V}_k$ to the root. Note that the choice of $\mathcal{V}_{\text{root}}$ is irrelevant.

Applying a similar procedure to all n point clouds $\boldsymbol{S}_i$ in the sequence, a registered point cloud $\boldsymbol{S}$ is obtained

$$\boldsymbol{S} = \{\boldsymbol{T}_i\boldsymbol{S}_i\}, \quad 1 \leq i \leq n \quad (14)$$

where $\boldsymbol{T}_i$ denotes the transformation bringing the i-th frame to the world coordinate system. Even though, the pairwise registration is in general accurate, small errors get accumulated after concatenating a few transformations. To reduce the overall registration error, these errors can be corrected by means of a global registration stage that simultaneously optimizes the poses of all overlapping views. We used the method proposed by Fantoni et al. [21] to refine the transformations. Since distance transforms for large volumes result in a large memory footprint, finite differences are computed using appropriate nearest neighbour searches in an Octree structure. To speed up this process, the refinement is deployed solely with the 3D positions of the visual feature keypoints extracted before.

Fig. 4: 3D+RGB Hypothesis Verification; Left: registered point cloud $\boldsymbol{S}$, Middle: extended hypothesis set $\mathcal{H}_{\text{root}}$, Right: selected subset $\mathcal{H}_{\text{verified}} \subset \mathcal{H}_{\text{root}}$ after verification. Note that the unrecognized bottles are not in the training set.

E. 3D+RGB hypothesis verification

The previous stages result in a set of hypotheses $\mathcal{H}_{\text{root}}$ (obtained by transforming hypotheses generated in single frames to a global coordinate system) and a reconstructed scene point cloud $\boldsymbol{S}$ (obtained by registering the different frames in the sequence). Since $\mathcal{H}_{\text{root}}$ might contain wrong or redundant hypotheses, the following stage aims at selecting a subset of $\mathcal{H}_{\text{root}}$ consistent with $\boldsymbol{S}$ (see Fig. 4). To obtain the best hypothesis subset, the single-view verification method presented in [4] is extended to handle scene clouds seen from several vantage points as well as to consider color information. Because RGB-D sensors present several artifacts that become evident once several clouds are merged together, we apply the RGB-Dnoise model of Nguyen et al. [22] in order to improve the reconstructed scene $\boldsymbol{S}$ (see Fig. 3) before hypothesis verification.

The algorithm of [4] relies on minimizing a suitable *cost* function defined over the solution space $\mathbb{B}^n$ of the hypothesis verification problem. In particular, a solution is denoted by a set of boolean variables $\mathcal{X} = \{x_0, \cdots, x_n\}$ having the same cardinality as $\mathcal{H}$. Each $x_i \in \mathbb{B} = \{0, 1\}$ indicates whether the corresponding hypothesis $h_i \in \mathcal{H}$ is discarded/accepted (i.e. $x_i = 0/1$) so that the *cost* function can be expressed as $\mathfrak{F}(\mathcal{X}) : \mathbb{B}^n \rightarrow \mathbb{R}$. The cost function includes four different cues:

1) scene fitting term $\Omega_{\mathcal{X}}(\boldsymbol{p})$ – how well the scene points are *supported* by the hypotheses,
2) model outliers term $f_{\mathcal{M}}(\mathcal{X})$ – how many *visible* model points are left unexplained,
3) multiple assignment term $\Lambda_{\mathcal{X}}(\boldsymbol{p})$ – how many scene points are simultaneously associated to different hypotheses,
4) clutter term $\Upsilon_{\mathcal{X}}(\boldsymbol{p})$ – how well the hypothesis fits to neighboring scene regions.

The cost function $\mathfrak{F}(\mathcal{X})$ is then defined by

$$\mathfrak{F}(\mathcal{X}) = f_{\boldsymbol{S}}(\mathcal{X}) + \lambda f_{\mathcal{M}}(\mathcal{X}), \tag{15}$$

where λ is a regularization term, and $f_{\boldsymbol{S}}$, $f_{\mathcal{M}}$ account, respectively, for cues defined on scene points and model points. Defining $|\Phi_{h_i}|$ as the number of visible model outliers for hypothesis h_i, these terms are calculated by

$$f_{\boldsymbol{S}}(\mathcal{X}) = \sum_{\boldsymbol{p} \in \boldsymbol{S}} \left[\Lambda_{\mathcal{X}}(\boldsymbol{p}) + \Upsilon_{\mathcal{X}}(\boldsymbol{p}) - \Omega_{\mathcal{X}}(\boldsymbol{p})\right], \tag{16}$$

$$f_{\mathcal{M}}(\mathcal{X}) = \sum_{i=1}^{n} x_i |\Phi_{h_i}|. \tag{17}$$

Thus, the cost function $\mathfrak{F}(\mathcal{X})$ aims at maximizing the amount of supported points in the scene while enforcing geometrical constraints to reject inconsistent hypotheses. For more details, we refer the reader to [4].

Since the verification stage was originally designed to be deployed on 3D data and does not make use of the grid structure present in RGB-D data (except for reasoning about visible and occluded model points), the multi-view extension turns out to be straightforward. In particular, we only need to change the definition of visible model points. Thus, for the multi-view case, a model point $\boldsymbol{q}$ is considered *visible* if it is *visible* in at least one of the original frames used to reconstruct $\boldsymbol{S}$. Let $\boldsymbol{S}_i$ be a frame in the sequence and $\boldsymbol{T}_i$ the transformation bringing $\boldsymbol{S}_i$ to the world coordinate system. Given f, c_x, c_y (focal length and central projection points of the camera), the visibility $V(q, \boldsymbol{S}_i)$ of q in $\boldsymbol{S}_i$ is assessed by:

$$V(q, \boldsymbol{S}_i) = \begin{cases} 1, & \text{if} \quad (q_z - \delta) \leq \boldsymbol{S}_i \left(\frac{f q_x}{q_z} + c_x, \frac{f q_y}{q_z} + c_y \right)_z \\ 0, & \text{elsewhere} \end{cases} \tag{18}$$

assuming that $\boldsymbol{q}$ and $\boldsymbol{S}_i$ are in the same coordinate system, which is obtained by transforming the model point in the coordinate system of $\boldsymbol{S}$ with $\boldsymbol{T}_i^{-1}$). δ is a small threshold (3 millimeters) representing the inaccuracy of the data and $\boldsymbol{S}_i(u, v)$ the point located at (u, v) in the grid structure of the original frame $\boldsymbol{S}_i$.

Finally, color information is incorporated into the verification stage by modifying when a scene point is *explained* or *supported* by an hypothesis. In [4], a scene point $\boldsymbol{p}$ is said to be *explained* by an hypothesis h_i if there exists a model point $\boldsymbol{q}$ such that $\|\boldsymbol{p} - \boldsymbol{q}\|_2 \leq \rho_e$ (ρ_e is an inlier threshold). When color is available, $\boldsymbol{p}$ is supported by h_i according to the original definition and if $\boldsymbol{q}$ simultaneously fulfils

$$e^{-\frac{1}{2}\left[\frac{(q_L - p_L)^2}{\sigma_L{}^2} + \frac{(q_A - p_A)^2}{\sigma_{AB}{}^2} + \frac{(q_B - p_B)^2}{\sigma_{AB}{}^2}\right]} \geq \rho_{color}, \tag{19}$$

where $\rho_{color} \in [0, 1]$ is a user-defined parameter indicating the desired color similarity between scene and model points and $\sigma_{AB,L}$ represent the expected amount of color variance. Symmetrically, any visible model point that does not simultaneously fulfil both equations for any point $\boldsymbol{p} \in \boldsymbol{S}$ is considered to be a model outlier. To increase robustness to illumination changes, the normalized LAB color space ($-1 \leq L \leq 1$ and $0 \leq \{A, B\} \leq 1$) is used for both scene and model points. In all our experiments, $\sigma_{AB} = \sigma_L = 0.35$ and $\rho_{color} = 0.8$.

Note that by changing the definition of *explained* points in the scene as well as by taking into account model *outliers* including not only distance but also color constraints, all

terms included in Equation (15) are affected. This subtle change enables the verification stage to use the powerful mechanisms within the verification framework. For example, imagine an object hypothesis aligned to a part of the scene where an impostor object (with same shape as the model associated with the object hypothesis but partially different color properties) is located. By changing the definition of *explained* points and model *outliers*, the points in the scene with different color will become *unexplained* by the hypothesis while those with similar color will still be *explained*. The activation of the hypothesis will result in a significant increase of the clutter related term and thus effectively reject the hypothesis.

F. Ground truth annotation: Back-projection to each view

The verification stage presented above results in a verified hypotheses set $\mathcal{H}_{\text{verified}} \subset \mathcal{H}_{\text{root}}$. By means of the respective transformation, these hypotheses can be transferred to the single-view sequence frames and thereby efficiently generate "ground-truth" annotations for each of the specific frames. For instance, the pose of $h_k^{\text{root}} \in \mathcal{H}_{\text{verified}}$ in the i-th frame is given by $\boldsymbol{T}_i^{-1}\boldsymbol{P}_k^{\text{root}}$, where $\boldsymbol{P}_k^{\text{root}}$ represents the pose of the object associated with h_k^{root} in the global coordinate system.

G. Assumptions

In order for the generated annotations to be complete (all frames annotated) and meaningful (objects annotated with a correct pose), the following assumptions need to hold for the sequence under consideration:

1) The multi-view graph $\mathcal{G}$ contains a single connected component and all edges included in the Minimum Spanning Tree provide an accurate pairwise alignment.
2) Each object (from those in our model library) in the sequence needs to be recognized with the correct pose in at least one frame.

IV. RESULTS

To demonstrate the performance of the proposed method on real scenarios, we have performed several experiments on three RGB-D datasets.

A. Datasets

The first two datasets, *Willow* and *Challenge*, respectively contain 24 and 39 sequences of RGB-D frames of a turn-table with several object instances (as well as impostors for *Willow*) on top of it. The training set is composed of 35 models including common textured household objects. Test sequences on *Willow* contain between 11 and 19 frames inducing strong occlusions for some object instances. On the other hand, the objects in the *Challenge* sequences are in general not occluded and the number of frames ranges between 3 and 6. Because of the turn-table setup, the frames in these datasets were processed by first removing any point farther away than 1.5 meters with respect to the camera as well as points below the highest detected plane (i.e, the turn-table). This effectively allowing the algorithm to focus on the part of the data (objects on the table) we are interested in. Notice that even after such a pre-processing stage, some inconsistent data (moving differently than the table) remains unfiltered and thus, motivate the deployment of ζ_{max} to quantify pairwise registration quality.

In order to show the performance of the method in more realistic scenarios (objects on top of each other, multiple supporting surfaces in form of tables or cabinets, high amounts of clutter, etc.), a third dataset, *TUW*, was acquired in our lab using the STRANDS robot *Werner*[2]. This training set is composed of 17 objects with different recognition relevant properties, e.g., textured and textureless objects and geometrically common or unique. Instead of a turn-table setup, this test set is obtained by moving the robot around a static scenario. Statistics on the different datasets are summarized in Table I.

TABLE I: Datasets properties: number of sequences, number of objects instances showcased in all sequences, number of frames and number of object instances over all frames are reported.

Dataset	Sequences	Objects	Frames	Instances
Challenge	39	97	176	434
Willow	24	110	353	1628
TUW	15	162	163	1911

B. Occlusion percentage

Having the pose of each object available for every frame allows us to calculate the amount of occlusion, which is an important factor for the performance of many recognition methods. Given a scene $\boldsymbol{S}$, the occlusion percentage for a specific ground-truth instance is computed from the ratio between *scene-supported* and total number of points. Let $\boldsymbol{M}$ be a point cloud representing a model in the training set and $\boldsymbol{P}$ the ground-truth transformation aligning $\boldsymbol{M}$ to $\boldsymbol{S}$. A point $q \in \boldsymbol{M}$ is *scene-supported* if there exists a point $p \in \boldsymbol{S}$ such that $\|p - \boldsymbol{P}q\|_2 \leq \rho_{\text{occlusion}}$. To accommodate for sensor noise and minor pose estimate errors, we set $\rho_{\text{occlusion}}$ to 3 millimeters in our experiments.[3] Figure 7 shows how the object instances are distributed on the three datasets with regard to their occlusion percentages.

C. Evaluation of the generated "ground truth"

For the *Willow* and *Challenge* datasets, the method was able to detect all objects in the respectively 24 and 39 sequences and did not incur in a single false positive. Regarding pose accuracy, the method had as well an outstanding performance with only 3 sequence-wise inaccurate estimates. All inaccurate poses were related to infamous *object_19* (a specular, almost textureless and symmetric object) and occurred due to the inability of the single-view recognition

[2] `http://strands.acin.tuwien.ac.at/`

[3] Low occlusion ratios (i.e. below 50%) are explained by the fact that $\boldsymbol{M}$ was acquired by placing the object up-right on a turn-table. Due to self-occlusions, such a configuration causes $\boldsymbol{M}$ to miss certain parts (e.g. bottom surface) of the actual object, which in some cases represent a large part of it (e.g. a book modeled while lying on a table, see `goo.gl/9X0qN3`).

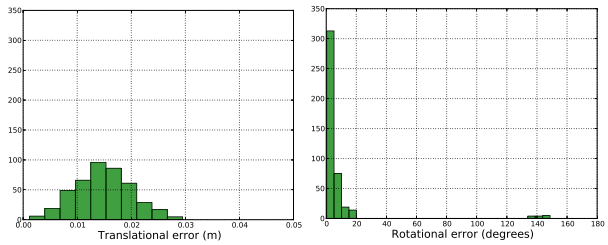


Fig. 5: Translational and rotational errors for the Challenge Dataset, original annotations by means of fixtures and checkerboard detection versus our automatic annotations. Large rotational errors ($> 20°$) occur due to wrong pose estimates of the proposed method in two sequences where one of the assumptions is not fulfilled.

Fig. 6: Inaccurate poses on the manual ground-truth annotations for the Willow Challenge dataset. Left: original annotations by means of fixtures and checkerboard detection; Right: annotations obtained with our method.

to estimate an accurate pose in any of the frames composing the three sequences. While the rotation around the symmetry axis was not properly retrieved, the translation of the object was correctly estimated.

Since ground-truth annotations were originally provided for the *Challenge* dataset, we performed a quantitative evaluation to compare the annotations provided by the proposed method and the original ground truth (we used the corrected annotations provided by [8])[4]. Figure 5 reports the results. Since errors were relatively large for visually pleasant annotations, we carefully analyzed the original ground-truth data to discover that the original poses were in some scenes significantly wrong, especially the translational component (for an example see Figure 6). Pose inaccuracies on the dataset were already reported by [8]. Even though such errors significantly reduce the value of the provided evaluation, we can still observe that the estimated annotations are *close* to those obtained by means of fiducial methods given the method assumptions hold. The errors and inaccuracies on the original annotations motivates even further the need for automating the process.

Regarding the more challenging *TUW* dataset, the method reported 1763 TPs, 0 FPs and 148 FNs, resulting in 100% precision and 92.26% recall. Sequence-wise, 11 objects out of 162 where not detected, resulting in 93.2% recall. Actual ground truth for this dataset was obtained by using the procedure presented in the upcoming section. Errors were mostly caused due to the inability of the single-view recognizer to detect the objects in any frame (assumption 2). Individual frame registration (i.e. accurate camera pose estimation) was attained for all sequences and thus, assumption 1 held for all sequences.

To visualize the annotation results for the three datasets obtained with our method, please checkout the supplementary material at `goo.gl/qXkBOU`. Ground truth annotations and training and test data are available at the same site.

D. Manual verification and correction

In order to provide valuable data to the community, we have designed a small graphical tool to correct and extend the automatic annotations provided by our method. The tool is able to load the reconstructed scene $\boldsymbol{S}$ and the verified hypotheses $\mathcal{H}_{\text{verified}} \subset \mathcal{H}_{\text{root}}$. A set of mechanisms is available within the tool to efficiently remove false positives, correct erroneous object poses, and add missing hypotheses. Once the operator has finished, the corrected annotations are back-projected to the single frames as in Section III-F. By means of automatic annotations and directly interacting with the reconstructed 3D scene, the process is greatly simplified.

E. Single-view recognition

To provide a baseline for single-view recognition methods as well as to further motivate the advantages of multi-view recognition frameworks, we have conducted an experiment to evaluate the performance of a simple single-view recognition method based on image features. In particular, we deploy the "2D Local Pipeline" [7] followed by single-view hypothesis verification stage. These results are reported in Fig. 7. As expected, performance decreases as occlusion percentages increase.

V. CONCLUSIONS

This paper presented a recognition method which exploits the ability of sensing a scene from multiple perspectives. We have shown how it can effectively aid in the creation of annotated RGB-D datasets for object instance recognition and how the availability of multiple vantage points significantly improves recognition results compared to single-view methods. Similarly, we have seen that the assumptions of the method hold in most practical scenarios, which validates our contributions and indicates that the method assumptions are realistic.

With the current set of tools in place, we plan on extending the *TUW* dataset by including a larger set of objects and sequences; aiming at the creation of a benchmark dataset to effectively evaluate the performance of existing and upcoming single-view recognition methods. We expect as well that such a dataset fosters research in new multi-view strategies to be deployed in online operation modus, ideally in combination with other related areas such as best-view planning, key-frame selection, scene reconstruction and object search at larger scales. Together with an appropriate amount of engineering to render the methods scalable, a

[4] `http://rll.berkeley.edu/2013_IROS_ODP/`

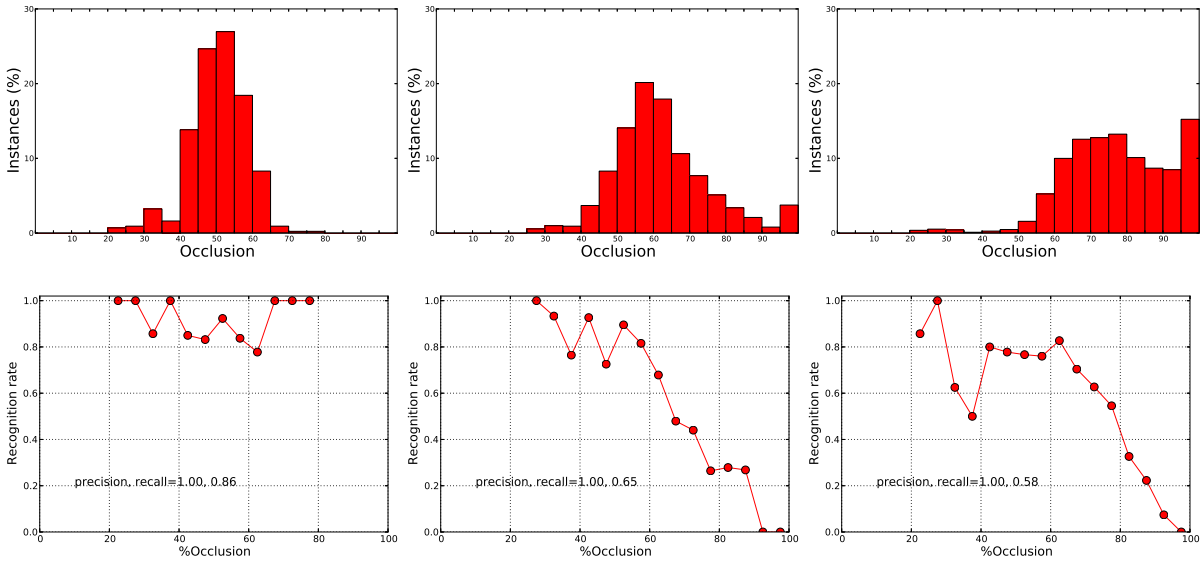


Fig. 7: *First row:* Occlusion distribution for the object instances in the Challenge, Willow, and TUW datasets. *Second row:* Single view recognition performance using a standard SIFT-based pipeline. Precision and recall values were computed for all object instances except those with $> 95\%$ of occlusion.

significant boost in the recognition capabilities of robots is possible.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Communitys Seventh Framework Programme FP7/2007-2013 under grant agreement No. 600623, STRANDS and the Austrian Science Foundation (FWF) under grant agreement No. I513-N23, vision@home.

REFERENCES

[1] I. Gordon and D. Lowe, "What and where: 3d object recognition with accurate pose," in *Toward Category-Level Object Recognition*, ser. Lecture Notes in Computer Science, 2006, vol. 4170, pp. 67–82.

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

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

[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] 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.

[6] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Computer Vision ACCV 2012*, 2013, vol. 7724, pp. 548–562.

[7] A. Aldoma, F. Tombari, J. Prankl, A. Richtsfeld, L. Di 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*, 2013, pp. 2104–2111.

[8] Z. Xie, A. Singh, J. Uang, K. S. Narayan, and P. Abbeel, "Multimodal blending for high-accuracy instance recognition," in *Proceedings of the 26th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013.

[9] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *IEEE International Conference on on Robotics and Automation*, 2011.

[10] J. Glover and S. Popovic, "Bingham procrustean alignment for object detection in clutter," *CoRR*, vol. abs/1304.7399, 2013.

[11] A. Collet and S. S. Srinivasa, "Efficient multi-view object recognition and full pose estimation," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2050–2055.

[12] K. Lai, L. Bo, X. Ren, and D. Fox, "Detection-based object labeling in 3d scenes," in *IEEE International Conference on on Robotics and Automation*, 2012.

[13] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "Rgbd mapping: Using depth cameras for dense 3d modeling of indoor environments," in *In RGB-D: Advanced Reasoning with Depth Cameras Workshop in conjunction with RSS*, 2010.

[14] N. Fioraio and L. D. Stefano, "Joint detection, tracking and mapping by semantic bundle adjustment," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 1538–1545, 2013.

[15] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *Proc. CVPR*, 2010.

[16] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of Histograms for local surface description," in *Proc. 11th ECCV*, 2010.

[17] F. Tombari, S. Salti, and L. Di Stefano, "A combined texture-shape descriptor for enhanced 3d feature matching," in *IEEE International Conference on Image Processing (ICIP), 2011*, 2011, pp. 809–812.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[19] D. Huber and M. Hebert, "Fully automatic registration of multiple 3d data sets," in *IEEE Computer Society Workshop on Computer Vision Beyond the Visible Spectrum(CVBVS 2001)*, December 2001.

[20] R. C. Prim, "Shortest connection networks and some generalizations," *Bell system technical journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[21] S. Fantoni, U. Castellani, and A. Fusiello, "Accurate and automatic alignment of range surfaces," in *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT)*, 2012, pp. 73–80.

[22] C. V. Nguyen, S. Izadi, and D. Lovell, "Modeling kinect sensor noise for improved 3d reconstruction and tracking." in *3DIMPVT*. IEEE, 2012, pp. 524–530.