This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

A Global Hypothesis Verification Framework for 3D Object Recognition in Clutter

Aitor Aldoma, Federico Tombari, *Member, IEEE,*
Luigi Di Stefano, *Member, IEEE,* and Markus Vincze, *Member, IEEE*

Abstract—Pipelines to recognize 3D objects despite clutter and occlusions usually end up with a final verification stage whereby recognition hypotheses are validated or dismissed based on how well they explain sensor measurements. Unlike previous work, we propose a Global Hypothesis Verification (GHV) approach which regards all hypotheses jointly so as to account for mutual interactions. GHV provides a principled framework to tackle the complexity of our visual world by leveraging on a plurality of recognition paradigms and cues. Accordingly, we present a 3D object recognition pipeline deploying both global and local 3D features as well as shape and color. Thereby, and facilitated by the robustness of the verification process, diverse object hypotheses can be gathered and weak hypotheses need not be suppressed too early to trade sensitivity for specificity. Experiments demonstrate the effectiveness of our proposal, which significantly improves over the state-of-art and attains ideal performance (no false negatives, no false positives) on three out of the six most relevant and challenging benchmark datasets.

Index Terms—3D Object Recognition, Hypothesis Verification

1 Introduction

Recognizing objects and estimating their 6-degrees-of-freedom (6DOF) pose based on processing 3D data in scenes featuring significant clutter and occlusion gathers ever-increasing attention from the scientific community. Indeed, many industrial robotics scenarios related to manufacturing and logistics require object recognition and 6DOF pose estimation, sought items often featuring texture-less or low-textured surfaces such that they may hardly be dealt with by state-of-the-art image matching methods. Furthermore, 3D vision represents the natural perception module for robotic agents in fields such as service and personal robotics, where autonomous robots are meant to operate and interact within human-populated environments such as houses, hospitals and stations.

The majority of state-of-the-art methods for 3D object recognition in clutter are based on either local or global features [1]–[11], with the former effectively withstanding occlusions, the latter deploying segmentation to enable recognition of smooth objects lacking distinctive surface traits. Whereas most approaches rely solely on features computed from 3D data [1]–[3], [5]–[9], a few recent proposals deploy intensity or color information too [4], [10], [11]. Alternatively to features, template matching [12], [13] accomplishes recognition of each given object by matching a set of templates taken from different vantage points into the current RGB-D image. Regardless of the adopted paradigm, 3D object recognition pipelines end up typically with a final stage, known as *Hypothesis Verification* (HV), whereby each *object hypothesis* determined through previous processing is geometrically verified so to reject false detections.

Unlike previous stages, though, the final HV process has been relatively unexplored thus far, with only a few published techniques explicitly addressing it [1]–[3], [14]. In current literature, the common HV approach relies on considering one hypothesis at a time and thresholding a consensus score related to the amount of scene points explained by transformed model points, thereby totally disregarding the valuable insights associated with interaction between hypotheses. Hence, unless the consensus threshold is kept as low as to hinder specificity, known HV methods tend to fail in detecting highly occluded objects, as these necessarily score low in terms of explained scene points. A recent review of verification and feature-based recognition methods can be found in [15].

Instead of analysing one hypothesis at a time and deciding whether it should be validated or dismissed, the Global Hypothesis Verification (GHV) method proposed in this paper considers simultaneously all available hypotheses and selects a specific subset providing globally the most coherent explanation of the captured scene data. Accordingly, the verification problem is formalised as the minimization of a cost function defined over the set of all available hypotheses and the scene under consideration. The minimization is guided by geometrical and appearance cues enforcing a solution that best explains the scene and is physically plausible. As Fig. 1 highlights, the GHV framework enables synergistic deployment of different paradigms with complementary strengths as well as multiple cues within the proposed object recognition pipeline. Thereby, even weak evidence

- *A. Aldoma and M. Vincze are with the Vision4Robotics group (ACIN - Technical University of Vienna)*
  E-mail: {aldoma,vincze}@acin.tuwien.ac.at
- *F. Tombari and L. Di Stefano are with the CVLAB group (DISI - University of Bologna)*
  E-mail: {federico.tombari,luigi.distefano}@unibo.it

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

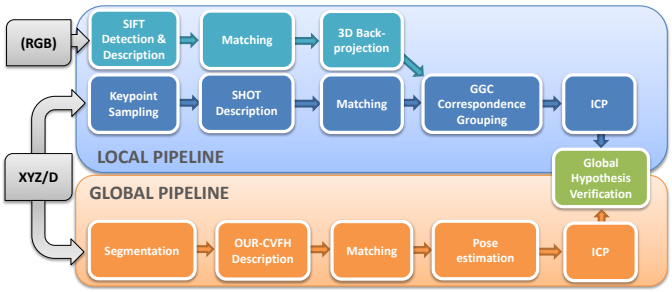


Fig. 1: Our 3D object recognition methodology leverages on both local and global features, with hypotheses gathered from the two pipelines merged into the final Global Hypothesis Verification stage. When available, color information can be deployed within the local pipeline.

based on varied pieces of information can be gathered throughout the pipeline so to let the final verification stage establish a global solution upon all detection hypotheses. As a result, weak evidence about an object can still permit correct detection without deteriorating specificity as long as it turns out coherent with strong cues on the presence of occluders. Global Hypothesis Verification, together the deployment of diverse recognition paradigms and cues, turns out key to performance: the pipeline in Fig. 1 compares favorably to the state-of-the-art on an experimental evaluation encompassing all the most relevant 3D object recognition datasets.

2 Related work

Hereby, we outline the most relevant Hypothesis Verification approaches for 3D object recognition in clutter.

In [16], using as seeds the correspondences supporting a hypothesis, a set of scene points is grown iteratively by including the neighbors which lie closer than a given distance to the transformed model points. If the final set is larger than a pre-defined fraction of the total number of model points (from one fourth to one third), the hypothesis is validated and ICP is run to refine object's pose. One disadvantage of such an approach is that it can not handle occlusion levels higher than 75%.

The HV method proposed in [1] ranks hypotheses based on the quality of supporting correspondences, so that they are verified sequentially starting from the highest rank. To verify each hypothesis, ICP is run and then two terms evaluated: the former, similarly to [16], is the ratio between the number of model points having a correspondent in the scene and the total number of model points, the latter is the product between this ratio and the quality score of supporting correspondences. Two additional checks are then enforced, so as to prune hypotheses based on the number of *outliers* (model points without a correspondent in the scene) as well as on the amount of occlusion generated by the current hypothesis with respect to the remaining scene points. If an hypothesis gets through each of these steps, it is accepted and its associated scene points are eliminated from the scene, so that they will not be taken into account when verifying the next hypothesis.

In [7] the set of hypotheses associated with a model is first pruned by thresholding the number of supporting correspondences. Then, the best hypothesis is chosen based on the overlap area $A(H_{best})$ between the model and the scene, and the initial pose refined by ICP. Finally, the accuracy of the selected hypothesis is given by the ratio $\frac{A(H_{best})}{M_a(H_{best})}$, where $M_a(H_{best})$ is the total visible surface of the model within the bounding box of the scene. The model is said to be present in the scene if its accuracy is above a certain threshold and, upon acceptance, the scene points associated with $A(H_{best})$ are removed.

Papazov and Burschka [3] evaluate how well a model hypothesis fits into the scene by means of an *acceptance function* which takes into account, as a bonus, the number of transformed model points falling within a certain distance from a scene point (*support*) and, as a penalty, the number of model points that would occlude other scene points. A hypothesis is accepted by thresholding its support and occlusion sizes. Given the hypotheses fulfilling the requirements set forth by the acceptance function, a conflict graph is built, wherein forks are created every time two hypotheses share a percentage of scene points above a threshold. Surviving hypotheses are then selected by means of a non-maxima suppression step over the graph and based on the acceptance function. This approach is the most similar to ours, because, thanks to the conflict graph, interactions between hypotheses are taken into account. Nevertheless, their method is only partially global, since the first stage of the verification process still relies on pruning hypotheses one at a time and a *winner-takes-all* strategy is used to handle conflicting hypotheses. Moreover, unlike [3], our approach is amenable to incorporate additional cues into the optimization problem, like color, to ameliorate discriminative power when the sensor delivers 3D data enriched with RGB triplets, or prior information about the presence of specific shapes such as planes.

3 Proposed recognition methodology

This section describes the proposed object recognition methodology, which comprises two separate 3D pipelines providing hypotheses to the final GHV stage based on either local or global features (see Fig. 1).

3.1 Input data

Our approach relies on processing point clouds related to models and scene, without any further assumption concerning the input data. Models can be provided as 3D meshes, point clouds or range maps, either as a collection of views of the same 3D model or as a fully registered 3D model. In case models are provided as fully registered rather than collections of views, during a pre-processing step they are transformed into a set of rendered views by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

placing a virtual camera on each vertex of a tessellated sphere centered at the model's centroid.

As it occurs in most applications, a scene is represented by a range map or a point cloud obtained from a single viewpoint. The proposed approach is able to handle data in the form of RGB-D images.

3.2 Local Pipeline

The local pipeline is based on descriptors computed on a small 3D neighborhood of a set of 3D keypoints. If RGB information is available in an *organized* way - as it is the case of RGB-D images - our local pipeline allows also for computing image descriptors related to 2D neighborhoods on a set of image keypoints.

3.2.1 Keypoint detection, description and matching

Keypoints are extracted at uniformly sampled positions on the surface of models and scene. Then, the SHOT local descriptor [17] is computed at each keypoint using the parameter values originally proposed in [17].

To attain 3D correspondences, scene and model descriptors are matched. To handle recognition of multiple model instances, each scene descriptor is matched via fast approximate indexing (i.e., randomized kd-trees [18]) against all models descriptors. We explicitly avoid using a matching threshold to reject weak correspondences, given the difficulty to choose such thresholds in general settings. Furthermore, we build a single kd-tree on all model descriptors, instead of one kd-tree per model. Although possibly increasing matching ambiguities, this approach reduces the complexity of the algorithm with respect to the number of models from linear to sub-linear. Accordingly, our matching scheme can scale to a high number of models without losing computational efficiency.

As anticipated, in case of RGB-D data an additional set of descriptors can be computed based on the color image associated with 3D points. Specifically, we compute SIFT keypoints and descriptors [19] and back-project them onto the 3D point cloud by means of the available depth information. This yields an additional set of 3D keypoints with associated descriptors. After matching scene and models descriptors, SIFT and SHOT correspondences are merged into a unique set before the correspondence grouping stage, so that the clustering algorithm can determine correspondence subsets by deploying both types of features seamlessly.

3.2.2 Correspondence grouping

As a result of the matching stage, a set of point-to-point correspondences $\mathcal{C} = \{c_1, ..., c_i, ..., c_n\}$, $c_i = \{p_i^m, p_i^s\}$, with p_i^m and p_i^s being a model and a scene 3D keypoint, respectively, is determined by association of model-scene descriptors that lie close in the descriptor space. This set of correspondences typically contains outliers that ought to be discarded. Popular methods for outlier rejection such as RANSAC are not suited to the multi-instance object recognition problem, as a set of correspondences may comprise several consensus sets related to different instances of a given model in the scene. Hence, specific *correspondence grouping* methods have been proposed. Assuming the transformation between the model and one of its instances in the current scene to be rigid, these methods enforce geometric consistency with the goal of clustering the correspondence set into geometrically coherent subsets, referred to as *object hypotheses*, each holding consensus for a specific rotation and translation of a model in the scene. Subsets whose consensus is too small can then be discarded simply by thresholding their cardinality (and under the constraint that at least 3 points are required to estimate a 6DoF transformation).

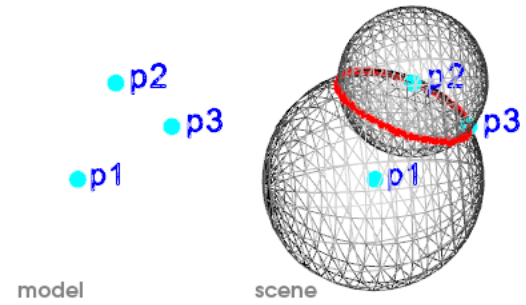


Fig. 2: GC constraint ambiguity: in this toy example, let the 3 points $\{p_1, p_2, p_3\}$ on the model (left-side) be associated with the respective ones on the scene (right-side), thus forming the 3 correspondences $\{c_1, c_2, c_3\}$. If the current consensus set only contains c_1, by evaluating p_2, all points belonging to the sphere centered in p_1 and of radius $||p_2 - p_1||$ will satisfy the GC constraint. If the consensus set contains both c_1 and c_2, when evaluating p_3, all points lying on the intersection of the two spheres centered in p_1 and in p_2 of radius, respectively, $||p_3 - p_1||$ and $||p_3 - p_2||$, (depicted in red) will satisfy the constraint.

An example of correspondence grouping method is the Iterative Geometric Consistency (denoted here as IGC) algorithm in [5]. Starting from an initial seed of correspondences, IGC iteratively builds up clusters of coherent correspondences by enforcing 3D consistency between pairs of correspondences:

$$\left| ||p_i^m - p_j^m||_2 - ||p_i^s - p_j^s||_2 \right| < \varepsilon \qquad (1)$$

with ε representing the maximum allowed difference between the feature distances measured on the model and the scene.

Because of (1), IGC aggregates correspondences holding consensus for a 6DoF transformation by means of a 1-dimensional constraint. On the one hand, this yields to a simple and fast method which has been shown to work satisfactorily in several practical scenarios [5]. On the other hand, though, this brings in an inherent ambiguity (see Fig. 2 for a graphical explanation) which makes the final consensus set highly dependent from the order in which correspondences are iteratively processed. While this is usually not a problem if $\mathcal{C}$ contains

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

enough correct correspondences, under more challenging circumstances featuring few inliers and/or a low inlier-to-outlier ratio, as it often occurs in scenes with significant occlusion and/or clutter, the correspondence processing order might turn out unfavorable so to lead IGC to miss clusters pertaining good object hypotheses.

Thus, we propose here a novel correspondence grouping method based on graph inference aimed at mitigating IGC caveats. This approach, referred to as Graph-based Geometric Consistency (GGC), will be then employed within the proposed object recognition methodology. A first modification to the GC constraint (1) aimed at decreasing ambiguities (see Fig. 2) can be attained by considering surface normals. Let $\{n_i^m, n_i^s\}$ be the surface normals at points $\{p_i^m, p_i^s\}$ associated with c_i and $\{n_j^m, n_j^s\}$ those at $\{p_j^m, p_j^s\}$ associated with c_j. Correspondences c_i, c_j are geometrically consistent if (1) holds and

$$\left| n_i^m \cdot n_j^m - n_i^s \cdot n_j^s \right| < \varepsilon_n \tag{2}$$

holds as well. ε_n represents the maximum angle deviation between normals in the scene and the model so that c_i, c_j are geometrically consistent.

Besides, we propose to formulate Correspondence Grouping as an inference problem on a graph $G_{GC} = (\mathcal{C}, E)$, where the node set consists of all correspondences, while the edge set, E, is created by joining the node pairs c_i, c_j that are consistent according to Equations (1) and (2). This novel formulation allows to solve the correspondence grouping problem optimally by finding all maximal cliques within G_{GC} such that its size is greater than the given consensus threshold τ_{GC}. Unfortunately, finding the maximal cliques in graph is known to be a hard problem, though fast algorithms for *small* graphs have been devised [20]. However, in the object recognition problem the number of correspondences may be quite large in the indeed favorable circumstances of non-occluded (or scarcely occluded) model instances. Therefore, to solve the maximal cliques problem underpinning our GGC formulation we propose the mixed algorithm described in Algorithm 1; relying on maximal cliques whenever it is possible to extract them within a fixed amount of time or alternatively, greedily merging correspondences into consistent clusters as in IGC.

Because the number of cliques can be large even for small graphs, the parameter max_{taken} controls the amount of cliques that are allowed to generate object hypotheses. Specifically, as the sorted cliques are processed and object hypotheses are generated, the algorithm counts how many times a specific correspondence has been used. Once a correspondence reaches the max_{taken} value, the algorithm forbids further use in the upcoming cliques. A value of 5 usually provides a good trade-off between accuracy and number of generated hypotheses and is used throughout the experimental results of this paper. It is worth observing that max_{taken} implicitly takes a value of 1 for IGC.

A major advantage of GGC over IGC is that the sorting stage takes places after all possible consensus sets have been generated. This allows to consider global clique properties (i.e, clique size as well average correspondence distance in both descriptor and Euclidean spaces) instead of single correspondence properties which are in general less resilient to noise. In addition, splitting the problem by means of the graph's connected components allows to properly handle the case of multiple instances of the same object being present in the scene, some of them occluded and cluttered (GGC) while others easy to detect (IGC).

Algorithm 1 GGC

Require: $G_{GC} = (\mathcal{C}, E)$, τ_{GC}, $max_{taken} = 5$
  $\mathcal{H} = \{\emptyset\}$
  $CC_{G_{GC}} = \{cc_1, ..., cc_n\} \leftarrow$ *biconnected_comp*(G_{GC})
  for all $cc_i \in CC_{G_{GC}}$ **do**
    if $\left|cc_i\right| \geq \tau_{GC}$ **then**
      success, cliques $\leftarrow$ *maximal_cliques*(cc_i, 100ms)
      if success **then**
        sort(cliques)
        for all clique $\in$ cliques **do**
          clique $\leftarrow$ *preprocess*(clique, max_{taken})
          clique $\leftarrow$ *RANSAC*(clique)
          if $\left|\right.$ clique $\left|\right. \geq \tau_{GC}$ **then**
            $\mathcal{H} \Leftarrow$ *obj_inst*(clique)
            Increment *taken* counter $\forall c_i \in$ clique.
          end if
        end for
      else
        $\mathcal{H} \Leftarrow IGC(cc_i)$
      end if
    end if
  end for
  return $\mathcal{H}$

3.3 Global Pipeline

Besides local descriptors, the proposed system deploys global descriptors by a second pipeline, as shown in Fig. 1. More in details, an approach similar to [4] is followed here. While global descriptors can be directly computed on each model view, when dealing with scenes comprising clutter and occlusion a segmentation step must be run on the view capturing the scene under analysis so to isolate smooth clusters of connected 3D points. Accordingly, we rely on the segmentation approach described in [21], which exploits the presence of a dominant plane to extract 3D clusters laying on such a plane.

Thus, a global descriptor is computed on each model view and scene cluster. In particular, we employ the OUR-CVFH descriptor [22], a viewpoint-dependent global representation which explicitly associates a Reference Frame to each descriptor and can also incorporate RGB information [4]. Each scene descriptor is matched against all models descriptors (one descriptor for each view related to each model), so as to find the k-NN ($k = 30$) model descriptors, each match yielding an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

object hypothesis. The 6DOF pose associated to each object hypothesis is given by the transformation to align the Reference Frames of the matched scene and model descriptor.

3.4 ICP refinement

The pose hypotheses generated by the recognition pipelines can be further refined by ICP. Similarly to [23], we rely on a fast ICP based on model distance transforms for the nearest neighbour correspondence problem, followed then by a few standard ICP iterations to achieve the final poses.

4 Global Hypothesis Verification

This Section illustrates the proposed GHV approach. After introducing notation, we formulate the cost function and analyze in detail the geometrical cues included therein. We consider a model library consisting of m point clouds, $\mathbf{M} = \{\mathcal{M}_1, \cdots, \mathcal{M}_m\}$, together with a scene point cloud, $\mathcal{S}$. We address the general case of $\mathcal{S}$ containing any number of instances from $\mathbf{M}$, including multiple instances of the same model as well as no model instance at all. The pose, $\mathcal{T}$, which relates a model to its instance in $\mathcal{S}$ is given by a 6DoF rigid body transformation (i.e. a 3D rotation and translation). We assume that the previous stages of the 3D object recognition pipeline provide 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 $\mathcal{M}_{h_i} \in \mathbf{M}$ being the model hypothesis and $\mathcal{T}_{h_i}$ the pose hypothesis which relates $\mathcal{M}_{h_i}$ to $\mathcal{S}$.

The goal of the proposed method is to validate a subset of hypotheses (up to n) belonging to $\mathcal{H}$ in order to maximize the number of correct recognitions (TPs) while minimizing the number of wrong recognitions (FPs). Purposely, we determine and minimize 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 dismissed/validated (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. More precisely, $\mathfrak{F}$ is written as

$$\mathfrak{F}(\mathcal{X}) = f_{\mathcal{S}}(\mathcal{X}) + \lambda \cdot f_{\mathcal{M}}(\mathcal{X}) + \beta \cdot \|\mathcal{X}\|_0 \tag{3}$$

where λ controls the influence of model cues, while $\beta \cdot \|\mathcal{X}\|_0$ regularizes the solution by favoring sparsity so to reduce false positives potentially arising from highly occluded objects hardly distinguishable due to the scarcity of the sensed visual information. The two cost terms $f_{\mathcal{S}}$, $f_{\mathcal{M}}$ account for cues defined over scene and model points, respectively:

$$f_{\mathcal{S}}(\mathcal{X}) = \sum_{p \in \mathcal{S}} \left(\Upsilon_{\mathcal{X}}(p) - \Omega_{\mathcal{X}}(p) + \Lambda_{\mathcal{X}}(p) \right) \tag{4}$$

$$f_{\mathcal{M}}(\mathcal{X}) = \sum_{i=1}^{n} |\Phi_{h_i}| \cdot x_i \tag{5}$$

and $|\cdot|$ represents the set cardinality operator. The four terms appearing in equations (4), (5), namely Ω, Φ, Λ and Υ, are associated with four basic cues, referred to as i)-iv). As it will be discussed in details in the following, i) is aimed at maximizing the number of recognized model instances, while ii), iii) and iv) try to penalize unlikely hypotheses through geometrical constraints, so as to minimize false detections.

4.1 Occlusions

Before defining the terms appearing in the cost function, the concept of *visible model* ought to be introduced. Each model hypothesis $\mathcal{M}_{h_i}$ has certain parts that are not visible in the scene due to self-occlusions or occlusions generated by other scene parts, which should be removed since they cannot provide consensus for h_i. Occlusion detection in 3D can be carried out quite efficiently, by establishing whether each model point is visible or not through the range image associated with the scene point cloud. If the former is not available directly, it can be generated from the point cloud via z-buffering: each point of the cloud is back-projected on a rendered range image by means of the vantage point associated with the current model view, and it is then considered visible if its projection falls on a valid pixel and its depth is smaller than that of the pixel [1], [3]. An analogous reasoning applies for detecting self-occlusions. Hence, each $\mathcal{M}_{h_i}$ is associated with a point cloud, denoted as $\mathcal{M}_{h_i}^v$, obtained by transforming the model according to $\mathcal{T}_{h_i}$ and removing all occluded points as explained above.

4.2 Cues i, ii) - scene fitting and model outliers

After the set $\mathcal{M}_{h_i}^v$ has been obtained, we wish to determine whether each such point has a *good* correspondent in the scene, i.e. the point *explains* a measurement concerning the scene and is thus said a *model inlier*, or it does not provide any consensus for the presence of the object in the scene and is so dubbed a *model outlier*. Classification of visible model points into *inliers* or *outliers*, and of scene points into *explained* or *unexplained*, sets forth the primary elements to reason upon solutions. Indeed, the higher the quantity of explained scene points and the fewer the model outliers, the more likely may be rated a solution.

A similar intuition underpins [1], [3], [5], [7], [16], with model and scene points classified by hard thresholding the Euclidean distance between each model point and its closest neighbor in the scene. Differently, we propose to achieve classification by estimating the likelihood of a model point to be an inlier based on a multivariate Gaussian distribution defined in a suitable feature space and learned from training data. Together with to the Euclidean distance, this novel classification scheme enables deployment of cues such as surface normal and color without introduction of additional threshold parameters that may be difficult to set. However, training data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

are needed to estimate the parameters of the Gaussian distribution: we will show in the experimental section how these can be obtained easily by labeling a few scenes according to correct object hypotheses. It is worthwhile noticing that the adopted generative formulation does not require negative training samples that would be difficult to obtain.

Given a model point $q \in \mathcal{M}_{h_i}^v$ and a small neighborhood in the scene $\mathcal{N}(q, \mathcal{S})$[1], we define a feature vector $\mathbf{f}_q$ aimed at capturing the degree of fit between q and $\mathcal{N}$ through the geometric and color cues listed in Table 1. In particular, depending on the available data modalities, $\mathbf{f}_q$ would either encode geometric fit into a 2-dimensional vector or comprehend also color similarity within a 5-dimensional representation. To classify q we rely on the Mahalanobis distance to the feature distribution learned for the inliers:

$$D_M(\mathbf{f}_q) = \sqrt{(\mathbf{f}_q - \mu)^T \Sigma^{-1} (\mathbf{f}_q - \mu)} \gtrless \rho_e, \tag{6}$$

where ρ_e is a threshold determining whether the point is judged an inlier ($D_M(\mathbf{f}_q) \leq \rho_e$) or outlier ($D_M(\mathbf{f}_q) > \rho_e$) while Σ and μ are the learned covariance and mean. It is worth pointing out that, due to the Gaussian assumption, setting a threshold on the Mahalanobis distance is equivalent to choosing a probability threshold on the likelihood of the observed feature vector for inliers. As such, ρ_e can be considered as a confidence level which abstracts away peculiarities of the data, such as resolution and noise, accounted for by the learned parameters of the Gaussian.

To weigh visible model points within the terms of the global cost function, we conveniently define

$$\delta(q) = \begin{cases} 1 - \frac{D_M(\mathbf{f_q})}{\rho_e}, & D_M(\mathbf{f}_q) \leq \rho_e \\ 0, & otherwise \end{cases}, \tag{7}$$

so that any $q \in \mathcal{M}_{h_i}^v$ is weighed proportionally to the likelihood of the observed features if an inlier, given null weight if an outlier. Moreover, we denote as Φ_{h_i} the set of outliers for hypothesis h_i and as $|\Phi_{h_i}|$ the cardinality of such a set. The amount of outliers associated with the active hypotheses in a solution (i.e. those h_i such that $x_i = 1$) is expressed in the cost function by (5). In the bottom right picture of Fig. 3, visible model points classified as either inliers or outliers are shown in orange and green respectively.

A scene point $p \in \mathcal{S}$ is considered *explained* by hypothesis h_i if it belongs to the neighborhood $\mathcal{N}(q, \mathcal{S})$ of at least one inlier in $\mathcal{M}_{h_i}^v$, *unexplained* otherwise. Accordingly, we define

$$\eta_{h_i}(p) = \begin{cases} 1, & \exists q \in \mathcal{M}_{h_i}^v : p \in \mathcal{N}(q, \mathcal{S}) \wedge \delta(q) > 0 \\ 0, & otherwise \end{cases} \tag{8}$$

The set of all scene points explained by h_i according to (8) will be denoted hereinafter as $\mathcal{S}_{h_i}$. Due to the neighborhood $\mathcal{N}(q, \mathcal{S})$ comprising more than one point, a scene point p may be explained by multiple inliers in $\mathcal{M}_{h_i}^v$. Denoted as $Q \in \mathcal{M}_{h_i}^v$ the set of such points, we introduce a function to express how accurately p gets explained by h_i:

$$\omega_{h_i}(p) = \max_{q \in Q} \delta(\mathbf{f}_q). \tag{9}$$

Generalizing the above definition from a single hypothesis to a solution $\mathcal{X}$, we define a scene point p to be explained by $\mathcal{X}$ if there is at least one model $\mathcal{M}_{h_i}^v$ associated with an active hypothesis in $\mathcal{X}$ that explains p. This is expressed by term $\Omega_{\mathcal{X}}(p)$ in (4), which weighs proportionally to $\omega_{h_i}(p)$ each scene point *explained* by solution $\mathcal{X}$:

$$\Omega_{\mathcal{X}}(p) = \max_{i=1..n} \left(x_i \cdot \omega_{h_i}(p) \right) \tag{10}$$

As anticipated, the amounts of explained scene points and outliers are deployed to score a solution $\mathcal{X}$ within the GHV framework. In particular, as vouched by equations (3), (4) and (5): i) the number of explained scene points should be maximized; and ii) the number of outliers associated with all active hypotheses should be minimized.

Cues	$\mathbf{f}_q$	Description
Geometry	$\min_{i=1..9} \|p_i - q\|_2$	Euclidean Distance
	$\max_{i=1..9} \langle \mathbf{n}_{p_i}, \mathbf{n}_q \rangle$	Surface normal fit
Color	$\min_{i=1..9} \|L_{p_i} - L_q\|$	Color distance (L-channel)
	$\min_{i=1..9} \|A_{p_i} - A_q\|$	Color distance (A-channel)
	$\min_{i=1..9} \|B_{p_i} - B_q\|$	Color distance (B-channel)

TABLE 1: Features included in vector $\mathbf{f}_q$

4.3 Cue iii) - multiple assignments

An important cue highlighting the existence of incoherent hypotheses within a solution deals with a surface patch in the scene being associated with multiple hypotheses. This can be exploited by penalizing scene points explained by two or more hypotheses. Thus, given a solution $\mathcal{X}$ and a scene point p, we define function $\Lambda_{\mathcal{X}}(p)$

$$\Lambda_{\mathcal{X}}(p) = \begin{cases} \sum_{i=1}^{n} x_i \omega_{h_i}(p), & \sum_{i=1}^{n} x_i \eta_{h_i}(p) > 1 \\ 0, & otherwise \end{cases} \tag{11}$$

which soft-weighs the number of conflicting hypotheses with respect to p according to (8) and (9). In the bottom left picture of Fig. 3, scene points explained by multiple hypotheses are colored in green, those explained by a single hypothesis in blue.

Hence, as shown by equations (3) and (4), another cue enforced within the GHV cost function through $\Lambda_{\mathcal{X}(p)}$ is that iii) the number of multiple hypothesis assignments to scene points should be minimized.

1. i.e. the 9 nearest neighbours of q in $\mathcal{S}$, denoted as $p_i, i = 1 \dots 9$ in Table 1

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

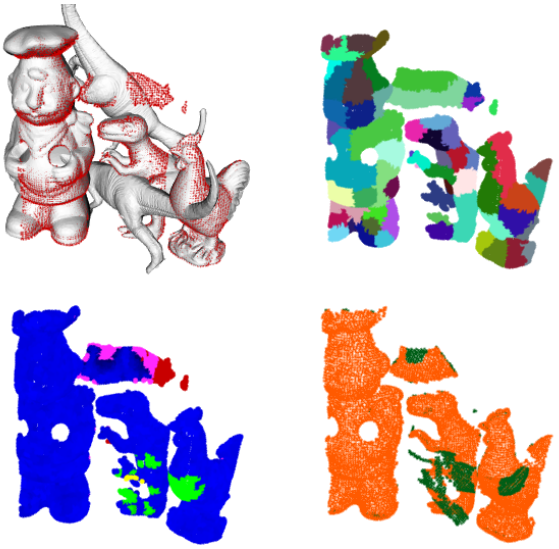


Fig. 3: Cues for GHV. *Top left*: a solution visualized by super-imposing the set of active model hypotheses onto the sensed scene points displayed in red. *Top right*: scene labeling via smooth surface segmentation. *Bottom left*: classification of scene points between explained by a single hypothesis (blue) or multiple hypotheses (green), unexplained (red), penalized by the clutter cue (purple). *Bottom right*: classification of visible model points between inliers (orange) and outliers (green).

4.4 Cue iv) - clutter

In many application scenarios not all sensed shapes can be fitted with some known object model. Exceptions may occur in some controlled industrial settings wherein all the objects making up the scene are known *a priori*. More generally, though, several visible scene parts which do not correspond to any model in the library might locally - and erroneously - fit some model shapes, potentially leading to false detections. Although useful to increase the number of correct recognitions, maximizing the number of explained scene points (i.e. cue i)) favors this circumstance. On the other hand, computing the outliers associated with hypotheses (cue ii)) might not always help, as the parts of the model which do not fit the scene might turn out occluded or outside the field of view of the 3D sensor.

To counterattack the effect of clutter, we devised an approach aimed at penalizing a hypothesis that locally explains some part of the scene but not points belonging to the same smooth surface patch. This is enforced based on over-segmenting the scene into smooth surface patches. Purposely, we deploy the supervoxel extraction strategy proposed in [24]. Then, supervoxels (with their associated normals) are merged together by creating a graph where an edge links a pair of supervoxels in case they are adjacent and the angle between normals is small. The connected components of the graph yield a segmentation of the scene into smooth patches. As a result, each scene point is associated with a unique segment label $l(p)$. The top right picture of Fig. 3 provides an example of scene segmentation into smooth patches.

Hence, given a solution $\mathcal{X}$, we compute a clutter term, $\Upsilon_{\mathcal{X}}(p)$, at each unexplained scene point p, so as to penalize those belonging to the same surface as the points explained by the active hypotheses in $\mathcal{X}$:

$$\Upsilon_{\mathcal{X}}(p) = \sum_{i=1}^{n} x_i \cdot \gamma_{h_i}(p) \tag{12}$$

$\forall p \in \mathcal{S}$ such that $\Omega_{\mathcal{X}}(p) = 0$ and

$$\gamma_{h_i}(p) = \kappa \frac{|\mathcal{S}_{h_i}^{l(p)}|}{|\mathcal{S}^{l(p)}|} \tag{13}$$

where $\frac{|\mathcal{S}_{h_i}^{l(p)}|}{|\mathcal{S}^{l(p)}|}$ is the ratio of explained points by h_i with label equal to $l(p)$ over the total amount of points in $\mathcal{S}$ with label $l(p)$, while κ is a parameter that weighs the penalty associated to the clutter term. The idea behind equation (13) is to control the impact of the clutter term for small under-segmentation artifacts, which might generate a small portion of explained points within nearby segments which do not belong to the object underlying h_i. In such circumstance, the ratio included in (13) effectively alleviates the wrong penalty brought in by the clutter term. Advantageously with respect to the formulation proposed in [5], equation (13) does not rely on any additional parameter to define the radius of influence of the clutter term associated to each point. The bottom left picture of Fig. 3 displays in purple the unexplained scene points that introduce penalties due to the clutter constraint. Thanks to the proposed clutter cue, incorrect active hypotheses, such as the dinosaur in Fig. 3, may penalize significantly the global cost function through term $\Upsilon_{\mathcal{X}}$.

Hence, as shown by equations (3) and (4), the last cue enforced within the cost function is that iv) the amount of unexplained scene points coherent to an active hypothesis according to (13) should be minimized.

5 Extension to Color Data

As previously mentioned, several sensors can acquire 3D data enhanced with color information, either as RGB-D images or in the form of point clouds with associated RGB triplets. In either case, color may enable the GHV stage to penalize hypotheses that explain the 3D shape but not the chromatic structure of the scene. The proposed framework is as flexible as to allow incorporating color information within the cost function. The introduced definitions of explained scene points and model inliers can already take into account color information when available (see (7) and Table 1).

Generalization to color, however, inherently exposes the GHV process to varying illumination conditions and color distortions between models and scene. We propose to mitigate color discrepancies due to these nuisances through a specific *tonal registration* step. Purposely, we exploit the alignment between an object hypothesis and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

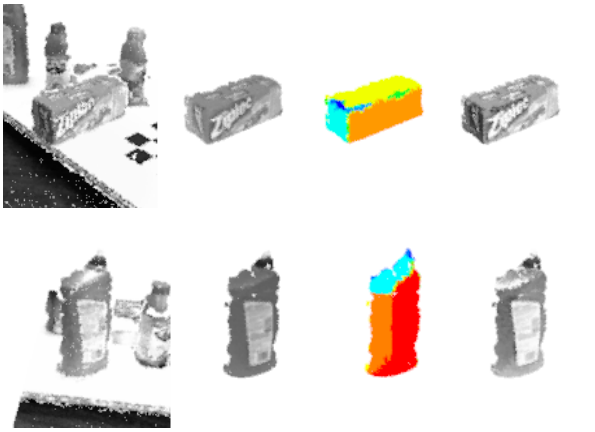


Fig. 4: Tonal registration across object's smooth faces. The L-channel is displayed as a grayscale image. From left to right: relevant part of the scene, object hypothesis, visible smooth faces, object hypothesis after independent tonal registration of each smooth face. In both rows, thanks to independent re-mapping of each smooth face, model points get tonally registered to either darker of brighter scene surfaces.

the scene to tonally re-map the model's color distribution so to match that of *potentially explained* scene points. More precisely, the L-channel of the model is tonally registered by means of the Histogram Specification technique [25] to those scene points in the neighborhood of the hypotheses visible points (see Section 4.2). Then, the mapping achieved by Histogram Specification is used to adapt the luminance values (i.e. L-channel) of model points before the evaluation of (7).

As illustrated in Fig. 4), a more effective re-mapping may be obtained by independently specifying the L-channel for each smooth face of the model, so as to possibly compensate for diverse lighting conditions across different parts of the object. Hence, during the off-line stage, models are analyzed to determine their smooth faces. In particular, we rely on the same strategy proposed in Section 4.4 used to extract smooth surface patches in the scene. During the verification stage, the L-channel histogram of each smooth model face is specified toward that of the corresponding *potentially explained* scene points, so that luminance values can be re-mapped accordingly prior to assessment of (7).

6 Planar Hypotheses

A common trait of most object recognition scenarios is represented by the presence in the scene of planar surfaces such as tables, ground floor, walls, etc.. Although not belonging to the database of sought models, more often than not these items would easily account for the majority of points in the scene. Thus, correctly recognizing the planar surfaces in the scene holds the potential to diminish false detections significantly, as all object hypotheses associated with such planes may be dismissed. Furthermore, the interaction between object and planar hypotheses can be exploited to penalize solutions featuring physically implausible configurations (i.e. an object hypothesis intersecting a plane).

Based on the above motivations, in this Section we describe how to extend the GHV framework so as to handle explicitly the presence of *planar hypotheses* within a solution. We would also like to point out that the proposed extension inherently delivers additional knowledge that may be deployed in a variety of applications (e.g. path-planning, high-level human-user interaction, etc.), calling for more comprehensive scene understanding.

6.1 Hypotheses generation

To generate planar hypotheses we rely on two alternatives approaches aimed at plane segmentation, which depend on the characteristics of the data representing the scene. If data is acquired by means of RGB-D sensors providing an *organized* structure of the point cloud, we make use of the multi-plane segmentation approach described in [26]. If the point cloud is *unorganized*, we follow a simple iterative plane fitting approach based on RANSAC whereby, after each iteration, the points associated to the dominant plane are removed from the scene. With both approaches, each element p_i of the set of generated planar hypotheses $\mathcal{P} = \{p_1, \cdots, p_m\}$ is given by the pair $(\mathcal{M}_{p_i}, \boldsymbol{p}_i)$, where $\mathcal{M}_{p_i}$ is the set of model points and $\boldsymbol{p}_i = \{n_{x_i}, n_{y_i}, n_{z_i}, d_i\}$ encodes plane coefficients. $\mathcal{M}_{p_i}$ is obtained by projecting the scene points holding consensus for plane $\boldsymbol{p}_i$, hereinafter denoted by set $\mathcal{S}_{p_i}$, onto plane $\boldsymbol{p}_i$.

6.2 GHV framework with planar hypotheses

Injection of planar hypotheses into the GHV framework does increase the dimensionality of the boolean vector representing the solution from n to $n + m$, each solution $\mathcal{X} = \{x_0, \cdots, x_n, x_{n+1}, \cdots, x_{n+m}\}$ encoding now a specific configuration of dismissed/validated object and planar hypotheses. Then, given the above definition of the members $p_i = (\mathcal{M}_{p_i}, \boldsymbol{p}_i)$ of set $\mathcal{P}$, planar hypotheses can be deployed seamlessly within the GHV framework so as to contribute to the computation of the four terms Ω, Φ, Λ, Υ in equations (4), (5) according to the procedures described in Sections 4.2-4.4.

Eventually, when planar hypotheses are handled explicitly within GHV, the global cost function $\mathfrak{F}$ is modified so to penalize physically implausible configurations. This is achieved by adding into the right-hand side of equation (3) a term, $f_{\mathcal{P}}$, which takes into account interaction between planar and object hypotheses:

$$f_{\mathcal{P}}\left(\mathcal{X}\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \Pi(p_j, h_i) \cdot x_i \cdot x_{n+j} \qquad (14)$$

$\Pi(p_j, h_i)$ in (14) represents the penalty due to interaction

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

between planar hypothesis p_j and object hypothesis h_i:

$$\Pi(p_j, h_i) = \begin{cases} 0, & \mathcal{S}_{p_j} \cap \mathcal{S}_{h_i} = \oslash \\ min\left(\Pi^+\left(p_j, h_i\right), \Pi^-\left(p_j, h_i\right)\right), & otherwise \end{cases} \tag{15}$$

with

$$\Pi^+(p_j, h_i) = \sum_{q \in \mathcal{M}_{h_i}} \left\langle \boldsymbol{p}_j, (q_x, q_y, q_z, 1) \right\rangle \geq \epsilon \tag{16}$$

$$\Pi^-(p_j, h_i) = \sum_{q \in \mathcal{M}_{h_i}} \left\langle \boldsymbol{p}_j, (q_x, q_y, q_z, 1) \right\rangle \leq -\epsilon \tag{17}$$

In words, the penalty brought in by $\Pi(p_j, h_i)$ is zero if the two hypotheses p_j and h_i do not share scene points, otherwise it equals the less populous between the two subsets of points belonging to model $\mathcal{M}_{h_i}$ lying on either side of plane $\boldsymbol{p}_j$. It is worth pointing out that, as $\Pi(p_j, h_i)$ does not depend on the specific solution $\mathcal{X}$, it can be pre-computed at the beginning of the optimization stage, just after planar hypotheses have been extracted for the current scene.

7 Optimization Strategies

To solve the optimization problem in (3), a solution $\tilde{\mathcal{X}}$ minimizing function $\mathfrak{F}(\mathcal{X})$ over the solution space $\mathbb{B}^n$ has to be determined:

$$\tilde{\mathcal{X}} = \underset{\mathcal{X} \in \mathbb{B}^n}{\operatorname{argmin}} \left(\mathfrak{F}(\mathcal{X}) \right) \tag{18}$$

As the cardinality of the solution space is 2^n, even with a relatively small number of recognition hypotheses (e.g. in the order of tens) exhaustive enumeration becomes prohibitive. To reach an approximate solution within a feasible computation time, a solver for the class of pseudo-boolean optimization problems has to be employed. Please note that in the proposed formulation where all hypotheses might have an effect in the acceptance of any other hypothesis (particularly due to the clutter term), finding the global optimum is NP-hard. Purposely, we have considered and evaluated three popular meta-heuristic techniques: Local Search (LS), Simulated Annealing [27] and Tabu Search [28][2].

Local Search (LS) is a monotonic optimization method subject to local minima, where transitions are accepted only when the cost associated with the new state is lower than the current one. There are two variants of LS: (i) the first improving move is used to transition to a new state (Hill Climbing, hereinafter LS_hc) or (ii) the best improving move in the current neighborhood is selected (Gradient Descent, hereinafter LS_gd). Differently, Simulated Annealing (SA) [27] includes a mechanism to avoid local minima by allowing transitions to solutions with high costs during the initial stage of the optimization procedure (*high temperature* stage). Termination criteria are either the minimum temperature T_{min} has been reached, or no improvement has been found during the last N_{max} moves. Finally, Tabu Search (TS) [28] employs a *tabu list* containing previously visited solutions, so that a solution is accepted for further exploration only if not included in this list. The algorithm terminates either when all solutions in the neighborhood of the current one are in the *tabu list* or no improvement has been achieved during the last N_{max} moves.

2. We have used the implementations available in the METslib library *projects.coin-or.org/metslib*

7.1 Local neighborhood (Moves)

A key component for determining an appropriate solution by means of meta-heuristic techniques deals with the definition of the *neighborhood* of a specific solution, $\mathcal{N}(\mathcal{X})$. To this end, we define efficient moves to transition between $\mathcal{X}$ and $\mathcal{X}' \in \mathcal{N}(\mathcal{X})$ in order to explore the solution space of the hypothesis verification problem. Because the costs associated with the solutions in $\mathcal{N}(\mathcal{X})$ are required to guide the optimization process, it is crucial that their computation is efficient to render the overall optimization computationally feasible. In particular, moves should be designed in such a way that the cost $\mathfrak{F}(\mathcal{X}')$ can be computed incrementally from $\mathfrak{F}(\mathcal{X})$, i.e., allowing recycling computations associated with common elements between $\mathcal{X}$ and $\mathcal{X}'$. Two moves have been designed for the hypothesis verification problem: (i) *switch state* and (ii) *replace active hypothesis*.

7.1.1 Switch state

Given the current solution $\mathcal{X} = \{x_1, ..., x_n\}$ with $x_i \in \{0, 1\}$, a *switch state* move applied on the i-th hypothesis will switch the boolean value associated with h_i such that $x'_i = \neg x_i$ where $x_i \in \mathcal{X}$ and $x'_i \in \mathcal{X}'$.

A *switch state* move enable efficient computation of the cost $\mathfrak{F}(\mathcal{X}')$ based on the pre-transition cost $\mathfrak{F}(\mathcal{X})$ as well as on scene points influenced by the i-th hypothesis together with its model outliers Φ_{h_i}. As an example, consider the model term $f_{\mathcal{M}}(\mathcal{X})$, it can be incrementally computed as follows:

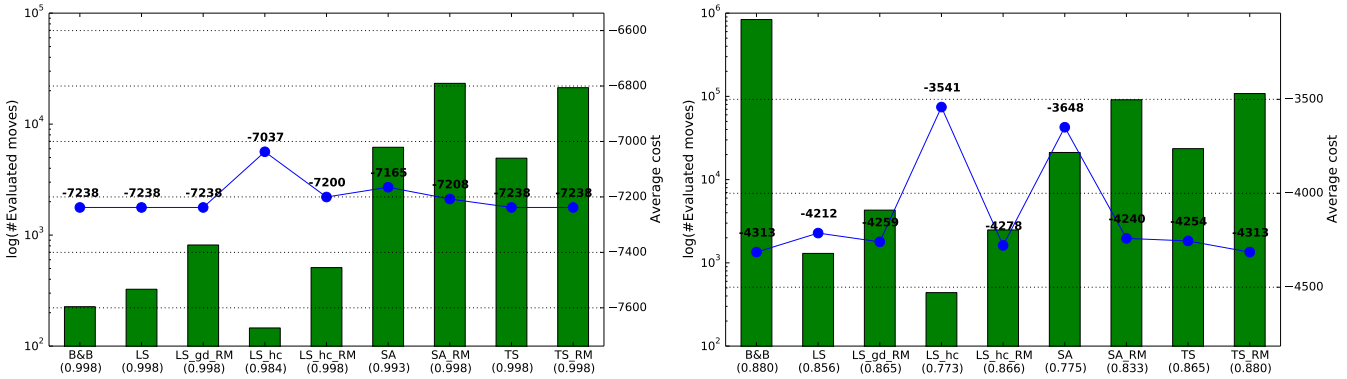
$$f_{\mathcal{M}}(\mathcal{X}') = \begin{cases} f_{\mathcal{M}}(\mathcal{X}) - \Phi_{h_i}, & x'_i = 0 \\ f_{\mathcal{M}}(\mathcal{X}) + \Phi_{h_i}, & x'_i = 1 \end{cases} \tag{19}$$

where h_i is the hypothesis involved in the move. Given their structure, all of the terms in $\mathfrak{F}$ can be incrementally computed from previous moves.

7.1.2 Replace active hypothesis

While *switch state* moves deal with a single variable of the current solution, *replace active hypothesis* moves are applied on pairs of variables, so to enlarge the explored neighborhood of a solution. Specifically, this move can be regarded as a combination of two *switch state* moves and be applied only on a pair of variables including one active and one non-active variable, i.e. $x_i, x_j \in \mathcal{X} s.t. x_i = 1, x_j = 0$. Upon application, $x'_i = 0$ and $x'_j = 1$. To increase the amount of explored solutions in a meaningful way to the GHV algorithm, we propose to evaluate *replace active hypothesis* moves only if h_i and h_j, i.e. the two hypotheses associated with variables x_i, x_j, *do interact* each other: $\mathcal{S}_{h_i} \cap \mathcal{S}_{h_j} \neq \oslash$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Laser Scanner¶ (b) Clutter

Fig. 5: Results by the different optimization algorithms on datasets (a), (b). Suffix _RM indicates the use of *replace active hypothesis* moves in addition to *switch state* moves. SA and TS are both configured with $N_{max} = 100$. The plot displays the number of evaluated moves until convergence (green bars, left vertical axis in logarithmic scale) as well as the average cost (blue markers, right vertical axis) over the whole dataset. The F-score is also reported between parenthesis below each optimization method, so to highlight the impact on recognition results.

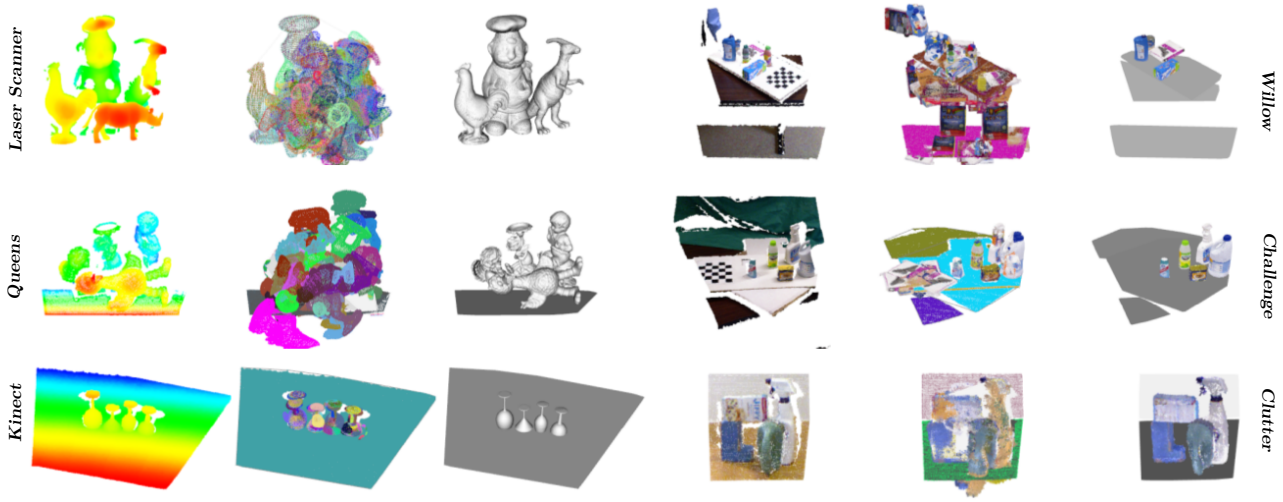


Fig. 6: From left to right: scene, object and plane hypotheses, verified hypotheses for the 6 considered datasets.

Dataset	Properties			
	#M./#S.	#Inst.	M. data	S. data
[1] *L. Scanner*	4 / 50	188	3D mesh	XYZ
[1] *L. Scanner*¶	5 / 50	217	3D mesh	XYZ
[6] *Queen's*	5 / 80	240	3D mesh	XYZ
[5] *Kinect*	35 / 50	176	3D mesh	XYZ image
Challenge	35 / 176	434	XYZRGB	RGB-D
Willow	35 / 353	1628	XYZRGB	RGB-D
[9] *Clutter*	18 / 30	120	XYZ	XYZ image

TABLE 2: Datasets properties. *Laser Scanner*¶ denotes the dataset with the *rhino* model. #M. and #S. represent the number of models and scenes, respectively. XYZRGB and XYZ stand for point clouds with and without associated RGB triplets, respectively.

7.2 Evaluation of optimization strategies

To determine the best algorithm for the GHV problem, we have compared the analyzed meta-heuristics using as input the hypotheses yielded by the pipeline proposed in Section 3 on two benchmark datasets (see Section 8). Unlike in [5], the initial solution for all algorithms consists in all hypotheses being deactivated ($x_i = 0, \forall x_i \in \mathcal{X}$), as this tends to yield faster convergence: indeed, typical working conditions are characterized by a high number of hypotheses fed to GHV, the majority being false positives to be switched off in the final solution.

As reported in Fig. 5, in our comparison we have considered the average number of evaluated solutions as well as the average final value of the cost function. In addition to a relative performance comparison among different meta-heuristics, we have implemented a Branch&Bound (B&B) approach in order to evaluate the approximate solutions found by the different optimization algorithms with respect to the global optimum configuration attained by B&B. The supplementary material contains additional information about this method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

While of limited practical interest due to its high computational complexity (see Fig. 5-(b)), B&B provides a reference to evaluate the absolute performance of the different meta-heuristics.

Regarding the number of evaluated solutions, we can observe that methods based on Local Search (LS_hc and LS_gd) tend to converge faster than the more complex SA and TS meta-heuristics. Their performance in terms of average minimum cost — in particular with the availability of *Replace active hypothesis* moves — is surprisingly good. However, in the *Clutter* dataset, the performance of TS turns out slightly superior (see Fig. 5-(d)). This can be ascribed to monotonic methods being unable to explore region of the solution space requiring activation of two or more hypotheses with none of them yielding a cost improvement by itself. As TS always explores the solution space given by the best move (regardless of cost improvements), the case where two or more hypotheses are required for a positive cost contribution is potentially explored. Through this analysis we have realized that the requirement of activating two or more hypotheses for a positive global cost contribution is due to the clutter term in combination with an under-segmentation of the scene in the smooth surface patches extraction presented in Section 4.4 (e.g. Fig. 6-Clutter where both objects on the left make a single smooth surface due to a rather unfortunate alignment for what concerns the clutter cue within GHV). On one hand, this aspect motivates even further the main idea of this work, namely to solve the verification of object hypotheses aiming for a globally coherent interpretation of the scene. On the other hand, it highlights some limitations of the simplest meta-heuristic strategies. SA appears to be the worst method in terms of convergence ratio and quality of the attained solution. This can be ascribed to a poor algorithm parametrization as well as to the structure of the GHV problem (requiring greedy moves at the beginning to reach a good solution and consecutively explore areas of the solution space associated with a temporal cost increase in order to escape local minima).

Overall, this comparison makes us lean toward TS, that with *Replace active hypothesis* moves can attain on both datasets the globally optimal configuration found by B&B, as the most suited approach to the GHV problem, due to the favorable trade-off between object recognition accuracy and average number of evaluated solutions. Nevertheless, optimization of the cost function by LS methods (especially with replace moves) should be considered in situations where computational efficiency is key and scene configurations are unlikely to include strong interactions among objects (e.g. scenarios where individual objects are far away from each other).

Finally, it is worth highlighting how in Fig. 5 the solutions associated with lower costs results in better -or equivalent- F-scores, this indicating the ability of the cost function to capture effectively the degree of fit between a given set of model instances and the sensed scene data.

8 Experimental Evaluation

8.1 Datasets

Thanks to its ability to handle diverse data types and scene configurations, to validate the system proposed in this paper we can consider all the main benchmark datasets for 3D object recognition in clutter. These, indeed, are heterogeneous as regards both the traits of the sought objects as well as the type of model and scene data, as depicted in Fig. 6 and detailed in Table 2. While some datasets include point clouds and 3D meshes of highly distinctive shapes acquired through laser scanners (*Laser Scanner*, *Queens*), others have been sensed by consumer depth cameras and address robotic settings dealing with manipulation of typical household objects (*Kinect*, *Challenge*, *Willow*, *Clutter*). Since all the considered datasets have been extensively used in literature, direct performance comparison between our proposal and state-of-the-art approaches can be attained straightforwardly (see Section 8.4). All six datasets and their ground-truth information can be downloaded at http://goo.gl/liCpfQ.

As described in Section 4.2, the definition of inlier model points and explained scene points depends on the Mahalanobis distance appearing in equation (6). To learn the parameters needed to compute the distance, we deploy ground-truth true hypotheses within an off-line learning stage. Precisely, for each dataset we learn Σ by picking randomly 10% of the scenes and using the associated ground-truth hypotheses, the only exception being Challenge where we use the same covariance as learned from Willow to demonstrate effective transfer of learned parameters between similar application settings. As regards the mean, we simply set $\mu = (0, 1)$ or $\mu = (0, 1, 0, 0, 0)$ depending on the feature space being either 2 or 5-dimensional (see Table 1) due to the dataset providing only shape or also color information, respectively.

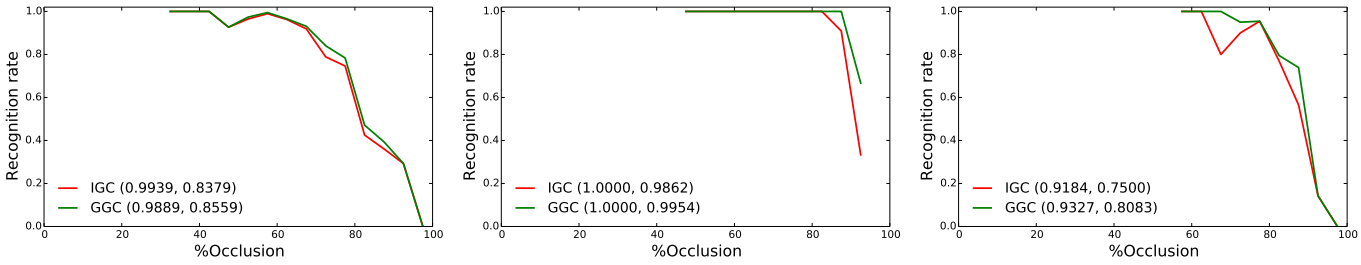
Throughout all experiments and regardless of the dataset, we set the 4 parameters of the GHV method to $\rho_e = 4.5$, $\lambda = 4$, $\beta = 100$ and $\kappa = 5$.

8.2 Correspondence grouping

First, we compare the GGC correspondence grouping method proposed in this paper to the IGC approach deployed in [5]. To this purpose, we rely on the correspondences generated by the local recognition pipeline (see Fig. 1) and use either GGC or IGC to form the hypotheses which then get validated by GHV. Fig. 7 highlights the advantages brought in by GGC compared to IGC through the Recognition Rate vs. Occlusion Rate charts obtained on three datasets, with Precision and Recall also reported between brackets[3]. In particular, we can observe a higher recognition rate as the occlusion rate increases, the latter causing correct correspondences

3. The totally occluded objects present in the Willow dataset are not considered here to compute Precision and Recall.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Willow (b) Laser Scanner¶ (c) Clutter

Fig. 7: GGC vs IGC with the local pipeline (SHOT and, when applicable, SIFT).

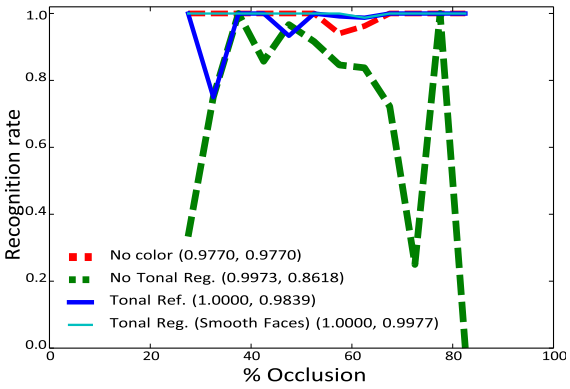


Fig. 8: Impact of color strategies on the *Challenge* dataset.

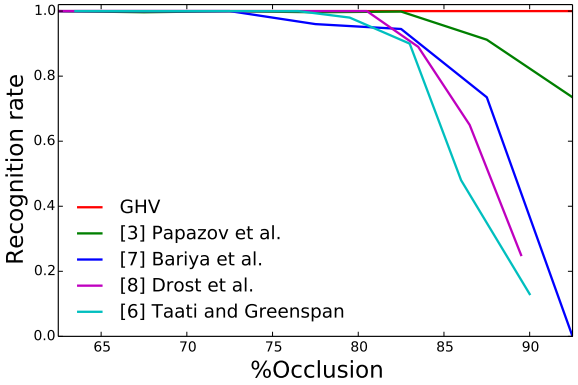


Fig. 9: Recognition Rate vs Occlusion Rate on the *Laser Scanner* dataset (without *rhino*).

to become smaller in number and thus exposing IGC's fragility with respect to the selection order of seed correspondences.

8.3 Color and tonal registration

To assess the impact of deploying color within GHV (Section 5), we have compared the results obtained by four different GHV configurations: no color, color without tonal registration, color with tonal registration using all scene points explained by an hypothesis and color with independent tonal registration at each model smooth face. Fig. 8 reports the results on the *Challenge* dataset, showing how color enhances the capability of GHV to tell apart similarly shaped objects, as vouched by the higher precision provided by all the configurations deploying color. Moreover, the Figure shows that tonal registration is key to recall as otherwise correct hypotheses may be rejected due to color inconsistencies between scene and model data. Finally, independent tonal registration at each model face yields definitely the best performance thanks to the ability to handle illumination changes that do not turn out consistent across the whole object: all 434 instances but one can be correctly detected without any false positive.

8.4 Comparative performance analysis

Based on the six benchmark datasets introduced in Section 7.1.1, we compare here the pipeline featuring the GHV stage proposed in this paper (Fig. 1) to published state-of-the-art 3D object recognition algorithms. Table 3 reports Precision, Recall and F-score obtained by our method (denoted as GHV) as well as published results from competing proposals on the *Clutter*, *Kinect*, *Willow* and *Challenge* datasets, wherein performance are customarily evaluated according to these figures of merit [4], [9]–[11]. As for *Laser Scanner* and *Queen's*, following the standard evaluation procedure on these datasets [3], [6]–[8], in Figs. 9, 10 we report the Recognition Rate vs. Occlusion Rate charts.

As vouched by the Table and the Figures, GHV outperforms neatly the other published methods on all datasets used in the evaluation with the exception of *Challenge* and *Willow*, where its performance turns out either exactly equivalent or comparable to that of [11]. It is worth noting, though, that the hypotheses generation and model selection strategy proposed in [11] assume objects to be singled out from clutter, this reducing the applicability of [11] to scenarios where segmentation is feasible (e.g. table-top settings). In contrast, our method can be applied successfully in a wider range of settings, as vouched by the diverse datasets used in this evaluation, while performing comparably to [11] in table-top scenarios. It is also worth highlighting that the proposed pipeline yields ideal performance (Precision=1, Recall=1) on the *Queen's* benchmark datasets, a remarkable result which, to the best of our knowledge, has not been

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

Dataset	Method	Prec.	Recall	F-score
Laser Scanner	GHV	0.9890	1.0000	0.9944
Laser Scanner¶	GHV	1.0000	0.9954	0.9976
Queen's	GHV	1.0000	1.0000	1.0000
Kinect	GHV	0.9420	0.9200	0.9308
	Glover [9]	0.8940	0.8640	0.8788
	Aldoma [5]	0.9090	0.7950	0.8481
Challenge	GHV	1.0000	0.9977	0.9988
	Tang [10]	0.9873	0.9023	0.9429
	Xie [11]	1.0000	0.9977	0.9988
	Aldoma [4]	0.9977	0.9977	0.9976
Willow	GHV	0.9728	0.8563	0.9108
	Xie [11]	0.9828	0.8778	0.9273
	Aldoma [4]	0.9430	0.7086	0.8091
	Tang [10]	0.8875	0.6479	0.7490
Clutter	GHV	0.9612	0.8250	0.8878
	Glover [9]	0.8380	0.7330	0.7819
	Aldoma [5]	0.8290	0.6420	0.7236

TABLE 3: Precision, Recall and F-score on the six benchmark datasets.

achieved by any published method yet.

Finally, our pipeline allows to accurately estimate the 6DoF pose of recognized models. Indeed, the average translation and rotation errors on *Laser Scanner* and *Queen's* are below $1mm$ and $2°$, while the average translation error on the datasets acquired by RGB-D sensors is less than $8mm$ (the average rotation error could not be evaluated due to the presence of highly symmetrical objects).

8.5 Sensitivity analysis

Table 4 allows for analyzing the sensitivity of GHV with respect to the threshold on the Mahalanobis distance, ρ_e. As expected, we observe generally higher Recall and lower Precision as the parameter is increased with respect to the default setting $\rho_e = 4.5$. This is due to a larger quantity of model points being judged as inliers which causes more scene points getting explained by hypotheses. Yet, in datasets where the whole scene can be explained by the model library (i.e. Laser Scanner¶, Queens, Challenge) we may notice how the increase in Recall achievable by a larger ρ_e does not imply a significant decrease in Precision. On the other hand, in datasets without appearance information and characterized by the presence of very similar models (e.g. Kinect), lower values of ρ_e may provide slightly better performance. It is worth highlighting that with a proper tuning of ρ_e, the proposed pipeline is able to achieve ideal performance (Precision=1, Recall=1) also on two other benchmark datasets, i.e. *Laser Scanner* and *Challenge*, another result that, to the best of our knowledge, has not been attained by any published method yet.

9 Final Remarks

One key finding of our work is that hypothesis verification can boost 3D object recognition performance by seeking for a globally coherent explanation of sensor measurements. Moreover, Global Hypothesis Verification provides a framework to merge seamlessly multiple recognition cues so to deal with diverse settings and objects without changing the algorithms across the processing chain nor modifying the parameters therein. As shown in Fig. 11, the two main steps of GHV require execution times in the order of a few seconds and exhibit linear dependency with respect to the total number of visible points. As the workload consists mainly of

Dataset	ρ_e 3	3.5	4	4.5	5
L. Scanner	**1.000/1.000**	**1.000/1.000**	0.995/1.000	0.989/1.000	0.989/1.000
L. Scanner¶	**1.000/0.995**	**1.000/0.995**	**1.000/0.995**	**1.000/0.995**	**1.000/0.995**
Queens	1.000/0.992	1.000/0.996	1.000/0.996	**1.000/1.000**	**1.000/1.000**
Kinect	**0.970/0.915**	0.958/0.909	0.953/0.915	0.942/0.920	0.936/0.920
Challenge	1.000/0.530	1.000/0.906	1.000/0.972	1.000/0.998	**1.000/1.000**
Willow	0.999/0.621	0.994/0.787	0.983/0.840	**0.973/0.856**	0.955/0.867
Clutter	0.962/0.625	0.967/0.733	0.969/0.792	**0.961/0.825**	0.951/0.817

TABLE 4: Precision/Recall for different values of parameter ρ_e. The setting yielding the highest F-score is highlighted in boldface.

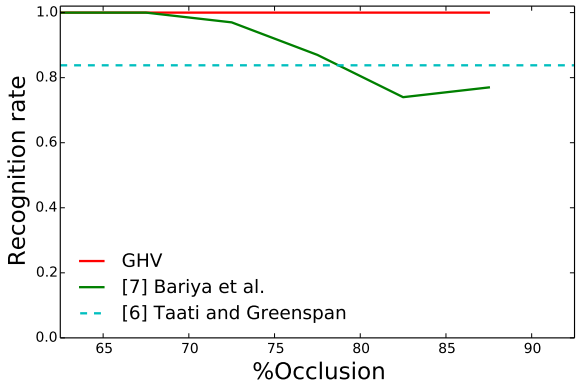


Fig. 10: Recognition Rate vs Occlusion Rate on the *Queen's* dataset (all 80 scenes). For Taati and Greenspan, the average recognition rate rather than the full chart is plotted, according to the data reported in [6].

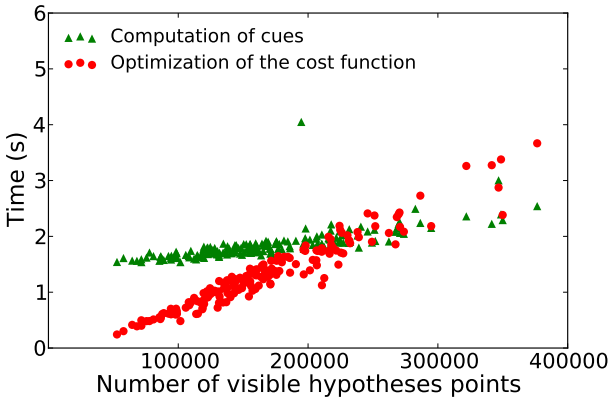


Fig. 11: Execution times on the Challenge dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2015.2491940, IEEE Transactions on Pattern Analysis and Machine Intelligence

point-wise independent calculations, GHV turns out a pleasingly parallel problem amenable to acceleration on modern GPUs.

REFERENCES

[1] A. Mian, M. Bennamoun, and R. Owens, "3d model-based object recognition and segmentation in cluttered scenes," *TPAMI*, 2006.

[2] P. Bariya and K. Nishino, "Scale-hierarchical 3d object recognition in cluttered scenes," in *CVPR*, 2010.

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

[4] 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 *ICRA*, 2013.

[5] A. Aldoma, F. Tombari, L. Di Stefano, and M. Vincze, "A global hypothesis verification method for 3d object recognition," in *ECCV*, 2012.

[6] B. Taati and M. Greenspan, "Local shape descriptor selection for object recognition in range data," *CVIU*, 2011.

[7] P. Bariya, J. Novatnack, G. Schwartz, and K. Nishino, "3d geometric scale variability in range images: Features and descriptors," *IJVC*, 2012.

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

[9] J. Glover and S. Popovic, "Bingham procrustean alignment for object detection in clutter," *IROS*, 2013.

[10] J. Tang, S. Miller, A. Singh, and P. Abbeel, "A textured object recognition pipeline for color and depth image data," in *ICRA*, 2012.

[11] Z. Xie, A. Singh, J. Uang, K. S. Narayan, and P. Abbeel, "Multimodal blending for high-accuracy instance recognition," in *IROS*, 2013.

[12] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit, "Gradient response maps for real-time detection of texture-less objects," *TPAMI*, 2012.

[13] M. Ulrich, C. Wiedemann, and C. Steger, "Combining scale-space and similarity-based aspect graphs for fast 3d object recognition," *TPAMI*, 2012.

[14] A. E. Johnson and M. Hebert, "Surface matching for object recognition in complex three-dimensional scenes," *IVC*, 1998.

[15] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan, "3d object recognition in cluttered scenes with local surface features: A survey," *TPAMI*, 2014.

[16] A. E. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *TPAMI*, 1999.

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

[18] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration," in *VISAPP*, 2009.

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[20] E. Tomita, A. Tanaka, and H. Takahashi, "The worst-case time complexity for generating all maximal cliques and computational experiments," *Theoretical Computer Science*, 2006.

[21] A. Richtsfeld, T. Morwald, J. Prankl, M. Zillich, and M. Vincze, "Segmentation of unknown objects in indoor environments," in *IROS*, 2012.

[22] A. Aldoma, F. Tombari, R. B. 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.

[23] 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 *ACCV*, 2013.

[24] J. Papon, A. Abramov, M. Schoeler, and F. Wörgötter, "Voxel cloud connectivity segmentation - supervoxels for point clouds," in *CVPR*, 2013.

[25] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., 2006.

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

[27] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, 1983.

[28] F. Glover and C. McMillan, "The general employee scheduling problem. an integration of ms and ai," *Computers & operations research*, 1986.

Aitor Aldoma received his diploma in computer science from the Polytechnical University of Catalunya (FIB-UPC) in 2008. In 2014, he received his Ph.D from the Technical University Wien (TUW). He has specialized in object recognition and scene understanding for robotic applications publishing more than 10 papers in these fields. In 2011, he has been a visiting scholar at Willow Garage and since then, a developer for the Point Cloud Library.

Federico Tombari holds an appointment as Assistant Professor at the University of Bologna, after obtaining from the same institution a Ph.D in 2009. His current research activity concerns computer and robot vision, and it encompasses co-authoring more than 50 refereed papers. In 2004 he has been visiting student at University of Technology, Sydney, while in 2008 he has been an intern at Willow Garage, California. He is a Senior Scientist volunteer for the Open Perception Foundation and a developer for the Point Cloud Library. He is member of IEEE and IAPR-GIRPR.

Luigi Di Stefano received a Ph.D in Electronic Engineering and Computer Science from the University of Bologna in 1994. He is currently associate professor at the Department of Computer Science and Engineering, University of Bologna. His research interests include computer vision, image processing and computer architecture. He is the author of more than 150 papers and five patents. He is a member of the IEEE Computer Society and the IAPR-IC. He is a member of the Scientific Council of T3LAB, a public institution devoted to technology transfer, as well as of the Scientific Advisory Board of Datalogic Group.

Markus Vincze received his Ph.D in Electronic Engineering at TUW in 1993. With a grant from the Austrian Academy of Sciences he worked at HelpMate Robotics Inc. and at the Vision Laboratory of Gregory Hager at Yale University. In 2004, he obtained his habilitation in robotics. Presently he leads a group of researchers in the Vision for Robotics laboratory at TUW. With Gregory Hager he edited a book on Robust Vision for IEEE and is (co)-author of over 250 papers. Markus' special interests are computer vision techniques for robotics solutions situated in real-world environments and especially homes.