IPSJ Transactions on Computer Vision and Applications Vol.4 20–29 (Mar. 2012)

[DOI: 10.2197/ipsjtcva.4.20]

Research Paper

Hough Voting for 3D Object Recognition under Occlusion and Clutter

Federico Tombari[1,a)] Luigi Di Stefano[1]

Received: April 15, 2011, Accepted: October 4, 2011, Released: March 28, 2012

Abstract: This work proposes a novel approach for the detection of free-form shapes in a 3D space. The proposed method matches 3D features through their descriptions to attain correspondences, then accumulates evidence of the presence of the object(s) being sought by verifying the consensus of correspondences within a 3D Hough space. Our approach is capable of recognizing 3D shapes under significant degree of occlusion and clutter and can deal with multiple instances of the shape to be recognized. We validate our proposal by means of a quantitative experimental comparison to the state of the art over two datasets acquired with different sensors (a laser scanner and a stereo camera) and characterized by high degrees of clutter and occlusion. In addition, we propose an extension of the approach to RGB-D (i.e., color and depth) data together with results concerning 3D object recognition from RGB-D data acquired by a Microsoft *Kinect* sensor.

Keywords: 3D object recognition, correspondence grouping, Hough voting, descriptor matching

1. Introduction

Increasing availability of low-cost 3D sensors promotes research toward intelligent processing of 3D information. In this scenario, a major research topic concerns 3D object recognition, that aims at detecting the presence and estimating the pose of objects, represented in the form of 3D models, within the 3D data acquired by a sensor such as a laser scanner, a Time-of-Flight camera, a stereo camera. Though a lot of effort has been devoted to the design of robust and discriminative 3D features aimed at reliably determining correspondences between 3D point sets [5], [7], [9], [14], [22], [23], [24], [25], [28], [35], with scenes characterized by clutter and occlusions relatively few approaches are available for the task of detecting an object and estimating its pose based on feature correspondences. Indeed, many approaches for 3D object recognition are aimed at object retrieval in model databases and hence can not deal with clutter and occlusion. Besides, 3D object retrieval methods usually do not estimate the 3D pose of the object nor can deal with the presence of multiple instances of a given model. This is the case of Bag-of-3D Features methods [19], [20], [26], approaches based on the *Representative descriptor method* [9] and probabilistic techniques such as e.g., Ref. [11] (see Ref. [29] for a survey). On the other hand, the well-known Geometric Hashing technique can in principle be generalized seamlessly to handle 3D data [17], although it hardly withstands a significant degree of clutter [10], [18].

Instead, methods specifically designed for the task of object recognition in 3D scenes with clutter and occlusions typically include a specific stage, usually referred to as *Geometric Validation*, aimed at discarding wrong feature correspondences determined during the feature matching stage that are caused by such nuisances, so as to determine a subset that can be reliably deployed for detecting specific object instances and estimating their pose in the current scene. State-of-the-art approaches to perform this stage are mainly two. On one side, starting from a seed feature correspondence, correspondence grouping is carried out by iteratively aggregating those correspondences that satisfy geometric consistency constraints [6], [15]. The other main approach relies on clustering pose hypotheses in a 6-dimensional pose space, each correspondence providing a pose hypothesis (i.e., rotation and translation) based on the local Reference Frame (RF) associated with the two corresponding features [23], [35]. Once reliable feature correspondences are selected by either enforcement of geometric consistency or clustering, a final processing stage based on Absolute Orientation [12] and/or Iterative Closest Point (ICP) [34], can be performed to further validate the selected subset of correspondences and refine pose estimation.

This work proposes a novel Geometric Validation approach aimed at object recognition in 3D scenes that can withstand sever degrees of cluttered background and occlusions. The proposed approach is general and can handle the recognition of multiple instances of the model to be found simultaneously present in the scene. It is worth pointing out, however, that our method is not meant for recognition of generic object categories (e.g., tables, cups,...) and therefore treats shapes having different sizes as different objects. The proposed approach relies on 3D feature detection, description and matching to compute a set of correspondences between the 3D model currently being recognized and the current scene. In addition, each feature point is associated with its relative position with respect to the centroid of the model, so that each corresponding scene feature can cast a vote in a 3D Hough space to accumulate evidence for possible centroid position(s) in

[1] DEIS, University of Bologna, Bologna, Italy
[a)] federico.tombari@unibo.it

the current scene. This enables simultaneous voting of all feature correspondences within a single tiny 3-dimensional Hough space. To correctly cast votes according to the actual pose(s) of the object(s) being sought, we rely on the local RFs associated with each pair of corresponding features. In addition, we propose an extension of the proposed Hough voting scheme to RGB-D (i.e., color and depth) data, which exploits the color cue to yield an additional set of 3D votes that can render the recognition more robust. We then present experiments by comparing our proposal with the state of the art over two challenging datasets acquired with different sensors (a laser scanner and a stereo camera) and characterized by high degrees of clutter and occlusion. The comparison demonstrates how the proposed approach can be usefully deployed to perform 3D object recognition and allow to assess in quantitative terms how it is significantly more discriminative and robust compared to the main existing methods. Results concerning the proposed extension to RGB-D data are also proposed, demonstrating successful 3D object recognition in clutter and occlusion based on 3D data gathered by a Microsoft *Kinect* sensor.

Next we briefly review the the state-of-the-art concerning the Hough Transform, while in Section 3 we describe the proposed 3D Hough voting approach. Section 4 illustrates the extension to RGB-D data, while Section 5 presents experimental results.

2. Hough Voting

The Hough Transform (HT) [13] is a popular computer vision technique originally introduced to detect lines in 2D images. Successive modifications allowed the HT to detect analytical shapes such as circles and ellipses. Overall, the key idea is to perform a voting of the image *features* (such as edges and corners) in the parameter space of the shape to be detected. Votes are accumulated into an array whose dimensionality equals the number of unknown parameters of the considered shape class. For this reason, although general in theory, this technique can not be applied in practice to shapes characterized by too many parameters, since this would cause a sparse, high-dimensional accumulator array leading to poor computational efficiency and very demanding memory requirements. By means of a matching threshold, peaks in the accumulator highlight the presence of a particular shape in the image. The Generalized Hough Transform (GHT) [2] extends the HT to detection of objects with arbitrary shapes, with each feature voting for a specific position, orientation and scale factor of the shape sought for. To reduce complexity, the gradient direction is usually computed at each feature position to quickly index the accumulator.

The extension of the original HT formulation to 3D data is quite straightforward and allows detection of planes within 3D point clouds. Similarly to the 2D case, also the 3D HT has been modified to deal with additional 3D analytical shapes characterized by a small number of parameters, such as spheres [32] and cylinders [27]. A slightly more general class of objects, i.e., polyhedra, is considered in Ref. [31], with a Hough Voting method in two separate 3D spaces accounting for rotation and translation allowing the detection of objects based on correspondences between vertex points established by matching straight edges. Unlike our proposal, though, this method cannot provide a unique pose hypothesis for each correspondences and, more importantly, cannot be applied to generic free-form objects.

More recently, an extension to the 3D domain of the GHT where gradient directions are substituted by point normals has been proposed [16]. Yet, as pointed out in the paper, this technique has several disadvantages that hardly allow its direct application. In particular, to deal with generic rotations and translations in a 3D space the Hough space becomes 6-dimensional, leading to a high computational cost of the voting process (i.e., $O(M \cdot N^3)$, M being the number of 3D points and N the number of quantization intervals) as well as to high memory requirements. Also, the resulting array would apparently be particularly sparse. Conversely, as described in the next section, with our approach the complexity of the voting process is $O(M_f)$ (M_f being number of feature points) and the Hough space is only 3-dimensional.

Another approach is represented by deployment of Hough voting for the sake of hypothesis verification in 3D object recognition [1]. Unlike the previously discussed 3D extensions, this approach relies on feature correspondences established between the object model and the current scene. Correspondences are grouped in pairs and in triplets in order to vote, respectively, into two distinct 3-dimensional Hough spaces, one meant to parametrize rotation and the other to account for translation. Since otherwise the number of groups would grow prohibitively large, only a fraction of the feature correspondences is deployed in each of the two voting processes. Then, peaks in the Hough spaces indicate the presence of the sought object. Differently, with our approach only a single 3D Hough space is needed and, due to the deployment of the local RFs attached to features, each correspondence can cast its own vote, without any need for grouping correspondences. The latter difference renders our approach intrinsically more robust with respect to wrong correspondences caused by clutter and also allows for deployment of all of the available information (i.e., correspondences) withing the voting process. Finally it is also worth pointing out that, due to the grouping stage, the method in Ref. [1] shares significant similarities with the geometric consistency approaches mentioned in previous section.

3. The Proposed 3D Hough Voting Algorithm

Let us suppose we have an object model that we want to recognize in a scene, both in the form of 3D meshes. The flow of the proposed object recognition approach is sketched in **Fig. 1**. At first, interest points (*features*) are extracted from both the model and the scene, either by choosing them randomly or by means of a suitable feature detector [5], [23], [24], [35]. They are represented as blue circles in the toy example shown in **Fig. 2**. Then, each feature point is enhanced with a piece of information representing a *description* of its local neighborhood, i.e., a 3D feature descriptor [7], [9], [14], [22], [25], [28], [35]. Typically, detecting and describing features of the model(s) can be performed once and for all off-line. Then, given this set of *described* features extracted from the model and the scene, a set of feature correspondences (green arrows in Fig. 2) can be determined. This can be done, e.g., by finding the Nearest-Neighbor of each scene feature within all model features by applying a metric (e.g., the Euclidean distance) between their descriptors, then using a matching thresh-

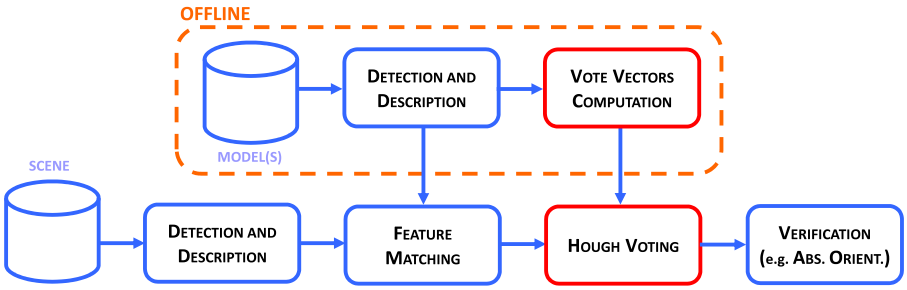

Fig. 1 Use of the proposed Hough voting scheme (red blocks) in a typical 3D object recognition pipeline.

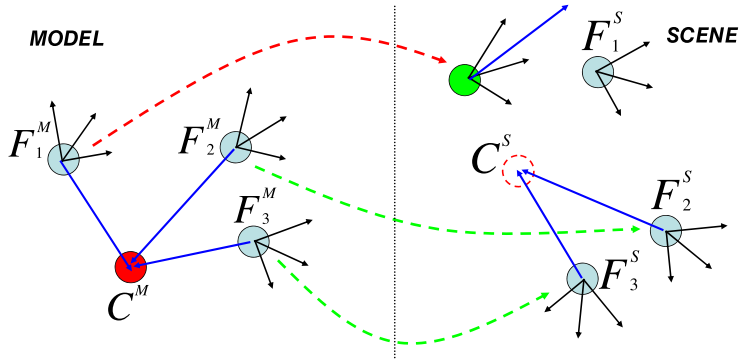

Fig. 2 Example of 3D Hough voting based on local RFs.

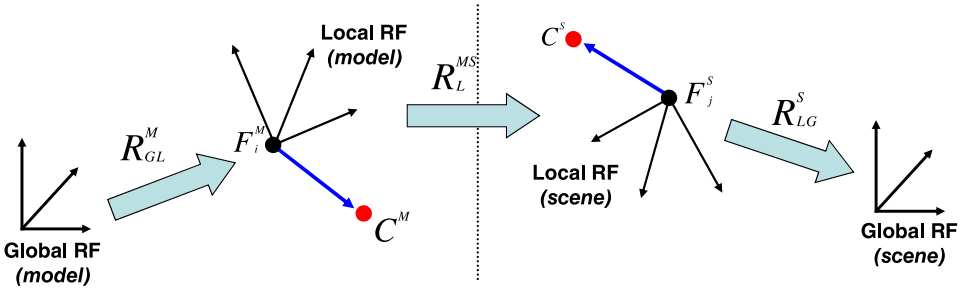

Fig. 3 Transformations induced by the use of local RFs.

old to discard unreliable correspondences. Due to the presence of nuisances such as noise, clutter and partial occlusions of the object, and since the comparison only involves feature descriptors (not taking into account the spatial position of each feature) typically this set includes also wrong correspondences (red arrow in Fig. 2), as locally similar shapes, although belonging to different objects, can potentially yield high degree of similarity.

The use of the proposed Hough voting scheme for Geometric Validation aims at accumulating evidence for the presence of the model in the scene. If enough features vote for the presence of the object in a given position within the 3D space, then the object is detected and its pose is estimated based on the established correspondences. In particular, at initialization time (i.e., off-line) a unique *reference point*, C^M, is computed for the model (red circle in Fig. 2). In our experiments, we have selected the centroid of the model, though this particular choice does not affect the performance of the algorithm. Then, still at initialization time, the vector between each feature and the centroid is computed and stored (blue arrows in Fig. 2). Since we want our method to be rotation and translation invariant, we can not store these vectors in the coordinates of the global RF (Reference Frame) since this would render them dependent on the specific RF of the current 3D mesh. Hence, and as sketched in **Fig. 3**, we need to compute an invariant RF for each feature extracted (i.e., a *local RF*) both in the model and in the scene. In particular, the local RF has to be efficiently computable (since we need to compute one RF for each feature) and very robust to disturbance factors (to hold on its invariance properties).

Several proposals for local RF frames for 3D meshes are present in literature [7], [22], [25], [28], [35]. In our approach, we use the fully unambiguous local RF method proposed in Ref. [30] given its robustness as discussed in Ref. [30]. Given a feature F, this method first computes the EigenValue Decomposition (EVD) of the covariance matrix $\mathbf{\Sigma}$ of the neighboring points P_i falling within the support of radius r around F. In particular, the computation of the covariance matrix is distance-weighted, assigning distant points smaller weights:

$$\mathbf{\Sigma} = \frac{1}{\sum_{i:d_i \le r} (r - d_i)} \sum_{i:d_i \le r} (r - d_i)(P_i - F)(P_i - F)^T \qquad (1)$$

where $d_i = |P_i - F|_2$. Additionally, in order to render the local RF unique and unambiguous, the sign of the smallest and the biggest eigenvectors is disambiguated by having each vector point towards the direction at higher density. Finally, the sign of the remaining eigenvector is obtained via cross product from the

other two.

Hence, in our approach, we perform an additional off-line step (see Fig. 1), that represents the initialization of the Hough accumulator. Assuming that all point coordinates of the 3D model are given in the same global RF, for each model feature point F_i^M we compute first the vector between C^M and F_i^M:

$$V_{i,G}^M = C^M - F_i^M \tag{2}$$

Then, to render this representation rotation and translation invariant, each vector $V_{i,G}^M$ has to be transformed into the coordinates given by the corresponding local RF (i.e., that computed on F_i^M, see Fig. 3) by means of the following transformation:

$$V_{i,L}^M = R_{GL}^M \cdot V_{i,G}^M \tag{3}$$

where $\cdot$ denotes matrix multiplication and R_{GL}^M is the rotation matrix where each line is a unit vector of the local RF of the feature F_i^M:

$$R_{GL}^M = \left[L_{i,x}^M L_{i,y}^M L_{i,z}^M\right]^T \tag{4}$$

The offline stage ends by associating to each feature F_i^M its vector $V_{i,L}^M$.

In the online stage, once correspondences between the model and the scene have been obtained, each scene feature F_j^S for which a correspondence has been found ($F_j^S \leftrightarrow F_i^M$) casts a vote for the position of the reference point in the scene. Since the computation of the local RF for F_j^S is rotation invariant, this allows to determine the transformation shown in Fig. 3 as R_L^{MS}. More precisely, given the aforementioned rotation invariance of the deployed local RF, R_L^{MS} boils down to the identity matrix, so that $V_{i,L}^S = V_{i,L}^M$.

Finally, we have to transform $V_{i,L}^S$ into the global RF of the scene, by means of the following relationship:

$$V_{i,G}^S = R_{LG}^S \cdot V_{i,L}^S + F_j^S \tag{5}$$

where R_{LG}^S is the rotation matrix obtained by lining up by columns the unit vectors of the local RF of the feature F_j^S:

$$R_{LG}^S = \left[L_{j,x}^S L_{j,y}^S L_{j,z}^S\right] \tag{6}$$

It is worth highlighting that, in Eq. (5), vector $V_{i,G}^S$ represents the vector of the $i-th$ vote coordinates in the global Reference Frame while vector $V_{i,L}^S$ is the vector of the $i-th$ vote coordinates in the Reference Frame of the $j-th$ feature point. Hence, as shown in the equation, to transform the vote coordinates from the local Reference Frame to the global Reference Frame, we need to add a displacement to vector $V_{i,L}^S$, which is equal to the feature point coordinates themselves, i.e., F_j^S.

Thanks to these transformations, the feature F_j^S can cast a vote in a tiny 3D Hough space by means of vector $V_{i,G}^S$ (see **Fig. 4**). Evidence for the presence of a particular model can then be evaluated by thresholding the peaks of the Hough space. Seamlessly, multiple peaks in the Hough space highlight the presence of multiple instances of the sought object. In the specific case that only one instance of the object is searched in the scene, the bin in the Hough space having the maximum number of votes is used to identify the presence of the objects. We have developed two slight variants to this scheme, that for each object instance take into account also the neighboring bins rather than only the one over-threshold. They will be described in detail in Section 3.1.

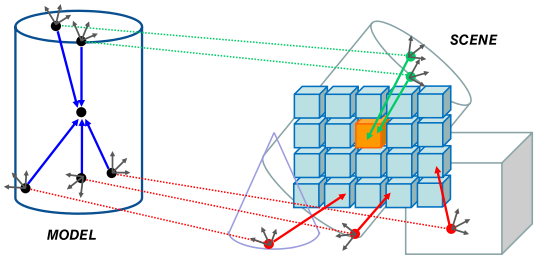

Fig. 4 Toy example showing the proposed 3D Hough voting scheme.

After identifying the bin relative to a model instance, a subset of coherent correspondences providing evidence of the presence of the model is selected as those that voted for that bin. As a successive step, the 3D pose of the model in the scene can be estimated, e.g., by means of the Absolute Orientation algorithm [12]. Repeating this same procedure for different models allows for deciding upon the presence of different objects (i.e., those belonging to a reference database).

3.1 Variants to the Hough Voting Process

The previously described approach for identifying the evidence of a model out of the Hough space is to select the bin yielding the maximum of the Hough space (single instance case) or to threshold the Hough space selecting all over-threshold bins (multiple instance case). By exploiting also the information contained in the neighboring bins of those indicating the presence of an object, we believe we can strengthen the approach by rendering the peak selection process more robust to quantization effects, although this may lead to a less accurate detection of the final correspondence subset. Hence, we have developed two variants to the standard Hough voting scheme.

With the first variant, referred as *Hough N-N* (N stands for *Neighbors*), when searching for the maximum (or analogously, when thresholding the Hough space in the multiple instance case) the value of each bin is added to that of its 6 neighboring bins. This can be motivated by the fact that, in presence of significant noise, correct correspondences can easily fall into neighboring bins, leading to a weakened evidence (i.e., a smaller value) of the maximum bin. Successively, all votes falling in these 7 bins are accumulated to yield the final subset of correspondences identifying the object which will be used for pose estimation.

Since the use of correspondences belonging to neighboring bins may also lead to inaccuracies when they are deployed for the sake of pose estimation, another variant, denoted hereinafter as *Hough N-C* (C stands for *Central*), uses the aggregated value of each bin with its neighbors when looking for the maximum in the Hough space (or analogously, when thresholding the Hough space in the multiple instance case). Then, when selecting the subset of correspondences to be used for pose estimation, only those concerning the central bin are accounted for. Given this notation, the standard Hough scheme (i.e., that relative only to the central bin) will be hereinafter referred to as *Hough C-C*.

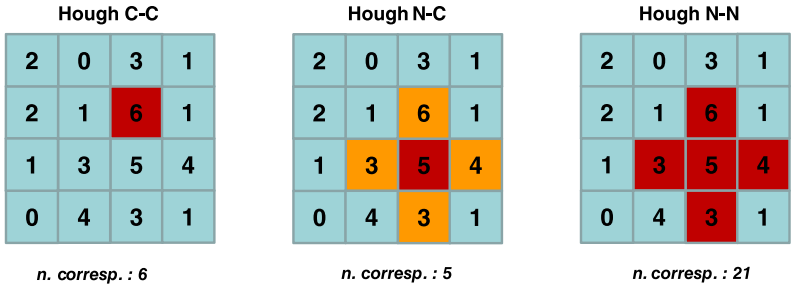

Fig. 5 The three proposed Hough voting schemes in a 2D toy example. A red square indicates the global maximum found within the Hough space, while the number in each bin represents the accumulated votes for that bin. While *Hough C-C* only takes into account individual bins when searching for the global maximum, both *Hough N-C* and *Hough N-N* take into account also the votes of neighboring bins (depicted in orange): the latter also includes these additional votes in the final set of correspondences.

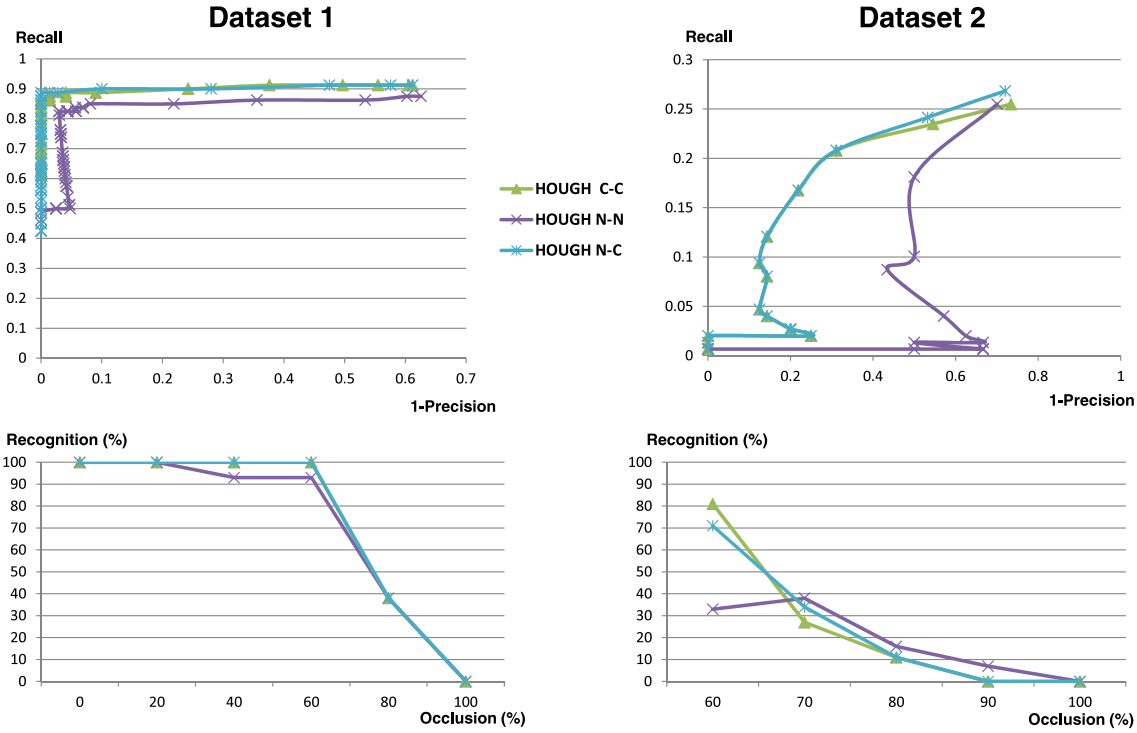

Fig. 6 Results in terms of *Precision vs. Recall* curves (above) and *Recognition vs. Occlusion* curves of the 3 proposed variants of the Hough recognition approaches. Left: 188 object tests on Dataset 1. Right: 200 tests on Dataset 2.

Figure 5 reports a toy example highlighting the differences between the 3 proposed Hough voting schemes. For ease of representation, it refers to the 2D case although the same characteristics hold in the 3D case. In the Figure, a red square indicates the global maximum within the Hough space, while the number in each bin represents the accumulated votes for that bin.

Figure 6 shows the results yielded by the 3 proposed variants of the Hough voting scheme on two different 3D datasets. Results are shown in terms of *Precision vs. Recall* curves (above) and *Recognition vs. Occlusion* curves. For more details on the two datasets, as well as about the adopted curves, please refer to Section 5. As it can be seen from the Figures, *Hough C-C* and *Hough N-C* performs significantly better than *Hough N-N* in terms of *Precision vs. Recall*. They also perform better in terms of *Recognition vs. Occlusion* on Dataset 1, while method *Hough N-N* is able to perform slightly better than the other two variants on highly occluded scenes belonging to Dataset 2. Overall, it seems that using only those votes that accumulate in the bin yielding the maximum of the Hough space improves the results, since evidently deploying also those votes that fall within neighboring bins might deteriorate the pose estimation carried out in the successive stage. As for the two best performing methods, *Hough C-C* and *Hough N-C*, the latter performs slightly better both in terms of *Precision vs. Recall* (on both datasets) as well as in terms of *Recognition vs. Occlusion* (on Dataset 2, while on Dataset 1 performance is equivalent), demonstrating the usefulness of estimating the maximum of the Hough space by deploying also the neighboring bins. Hence, *Hough N-C* is the variant that will be used for the experiments in Section 5.

4. Deployment of RGB-D Data

So far, the described method accumulates evidence for geometrically consistent correspondences obtained from matching 3D features, i.e., features computed from 3D data. In princi-

ple, though, the proposed method can work with features being extracted from other data sources, given that correspondences among these features are computed and that each feature is associated with a 3D local RF. Hence, in this section, we propose an extension of our Hough voting scheme to features that are computed from the texture associated with a 3D point cloud. This is motivated by the fact that 3D sensors that associate color information to depth are increasingly common and cheap. This is the case of, e.g., stereo cameras as well as the recently introduced Microsoft *Kinect* sensor: both devices acquire an intensity image associated with the depth map (RGB-D data). Deploying the additional information brought in by this intensity image can improve the robustness of the recognition, especially when matching objects whose 3D shape is not particularly distinctive.

Thus, in this proposed extension that works with RGB-D data, when matching a model to a scene, features are extracted and described by exploiting both the range map and the intensity image of the model and the scene. More specifically, two different sets of features are computed: one from the intensity image (i.e., 2D features) by means of a feature descriptor that works on images [3], [21], and another one from the range map (i.e., 3D features), by means of a feature descriptor that takes into account the 3D mesh computed out of the range map [7], [9], [14], [25], [28], [30], [35]. Subsequently, two sets of correspondences are established separately during the matching stage, one obtained by matching 3D descriptors, the other by matching 2D descriptors. The 2D and 3D model features that "survived" the matching stage are merged together into a final 3D "super-set," by back-projecting all survived 2D features in the 3D space based on the information provided by the range map. A similar procedure allows for attaining a scene features "super-set." Also, all features belonging to these two super-sets need to be associated with a local RF to perform the subsequent Hough voting stage: to this purpose we use the state-of-the-art technique described in Ref. [30].

Once these two super-sets (one for the model, the other for the scene) have been computed, the 3D Hough voting algorithm described in Section 3 is applied exactly as it was originally proposed. Thus, in this case correspondences yielded by matched 2D features will also contribute to accumulation of evidence upon the presence of the objects being sought by casting votes in the 3D Hough space, strengthening the robustness of the object recognition and improving the accuracy of the successive 3D pose estimation. Qualitative experimental results concerning this extended approach are shown in Section 5.

5. Experimental Evaluation

This Section presents experimental results concerning the proposed Hough voting scheme for 3D object recognition. More specifically, we carry out the experimental evaluation based on two different experiments. Experiment 1 aims at quantitatively compare our proposal to the state-of-the-art in order to evaluate its effectiveness on datasets including clutter and occlusions. Then, in Experiment 2, we evaluate qualitatively the proposed extension to RGB-D data (see Section 4), showing efficient and effective 3D Object Recognition from RGB-D data acquired by means of a *Kinect* sensor.

5.1 Experiment 1

We compare quantitatively the proposed Hough voting method to the state-of-the-art approaches for geometric validation of 3D correspondences. More specifically, we compare our method to the algorithm presented in Ref. [23], as a representative of the approaches relying on clustering in the pose space, and to that described in Ref. [6], as a representative of methods based on geometric consistency. It is worth pointing out that, although the latter method is quite less recent than the former, it arguably still represents the most popular approach for 3D object recognition. Hereinafter, we will refer to these two methods as, respectively, *Cluster* and *GC*.

5.1.1 Methodology

As for the methodology to carry out the experiment, given a set of models to be found in a scene (not containing all models but only a subset of them), the outcome of each single object recognition experiment can be a True Positive (TP), if the model sought for is present in the scene and correctly detected and localized, or a False Positive (FP), either if a model present in the image is detected but not correctly localized or the model is not present in the image but detected by the method. Similarly, True Negatives (TN) and False Negatives (FN) can be scored. Hence, denoting with P the total number of models to be found (*positives*), quantitative results are shown in terms of *Recall vs. 1-Precision* curves, with Recall and Precision defined as follows:

$$Recall = \frac{TP}{P} \tag{7}$$

$$Precision = \frac{TP}{TP + FP}. \tag{8}$$

To evaluate the localization, we first compute the RMSE between ground-truth feature positions (i.e., features mapped according to ground-truth poses) and those obtained by applying the pose transformation estimated by the algorithm, then we assume a correct localization if the RMSE is lower than a fixed threshold (set to 5 times the average model mesh resolution in our experiments). In addition, we also plot *Recognition vs. Occlusion* curves, that show the value of the Recognition Rate as a function of increasing values of the occlusion rate of the scene. It is worth noting that, since the latter type of curves focuses on the ability of methods of identifying objects in occluded scenes, that is they account for Recall only while providing no indication whatsoever of the corresponding values of FPs, they will be plotted for small values of the detection threshold, so as to maximize Recall.

5.1.2 Datasets

We have evaluated the considered methods on two different datasets, which are also those used for the results shown in Fig. 6. The first dataset (*Dataset 1*) has been acquired in our lab by means of the Spacetime Stereo (STS) technique [8], [33], a recent technique for obtaining accurate range maps out of stereo sequences. The dataset is composed of 4 models (referred to as *Rabbit*, *Bust*, *Mario*, *Dino* and shown in **Fig. 7**, top) and 80 scenes. In each scene, one of the 4 models appears at different levels of occlusion and clutter. Acquisition was performed trying

Fig. 7 The models (top) and some scenes at increasing level of occlusions (bottom) belonging to *Dataset 1*. Occlusion levels for the 4 scenes are, from left to right, 30%, 37%, 59% and 94%.

to cover several degrees of occlusions of the model (from 5% to 95% of partial occlusion created by means of other objects). In particular, out of these 80 scenes, for 36 none of the remaining 3 models was used to create occlusion: in these scenes, all the 4 models will be searched. As for the remaining 44 scenes, one or more models are present as part of the occlusion: in these scenes, only the occluded model will be searched for. Overall, the tested dataset includes a total of 188 object recognition experiments (80 positives, 108 negatives). Some scenes at increasing level of occlusions are shown in Fig. 7 (bottom). In addition, we also test the evaluated methods on a public dataset [*1], referred to as *Dataset 2*, which includes 5 models and 50 scenes. As for this dataset, we test the algorithms on the first 40 scenes (leaving the remaining 10 for parameter tuning), for a total of 200 object recognition instances (all models are sought for in each scene). Both datasets are quite challenging in terms of both clutter as well as occlusions.

5.1.3 3D Pipeline

For fairness of comparison, we fed the evaluated geometric validation algorithms with exactly the same correspondences, which are determined as follows. As for detection of keypoints, a fixed number of 3D points is randomly extracted from each model (i.e., 1,000) and each scene (i.e., 3,000). We use a random sampling for fairness of comparison, so as to avoid any possible bias towards some geometric validation methods that might benefit of specific features found by specific 3D detectors. It is also worth pointing out that, overall, the use of a random detector increases the difficulty of the object recognition task, hence reducing the overall expected performance. The descriptor and the feature matcher are also the same for all geometric validation methods, and they are ran with the same parameter values. More precisely, we use the recent hybrid signature-histogram descriptor proposed in Ref. [30], while for feature matching we rely on the euclidean distance and deploy a well-know efficient indexing technique (i.e., *Kd-tree* [4]) to speed-up the computations. Instead, a specific tuning has been performed for the parameters of the geometric validation techniques, i.e., the Hough space bin dimension for the proposed approach, the geometrical consistency threshold for *GC* and the number of k-means clusters for *Cluster*. The number of k-means iterations for *Cluster* has been set to 100 since we noted that the algorithm performance was not sensible to this parameter. Tuning for *Dataset 1* was performed over a tuning dataset composed of the same 4 models included in *Dataset 1* but of 26 different tuning scenes (referred hereinafter as *Tuning* dataset), by selecting the parameters yielding the best performance in terms of *Precision vs. Recall* curves. Tuning for *Dataset 2* was performed on the last 10 scenes of the dataset, using the first 40 ones for testing.

[*1] Available at: www.csse.uwa.edu.au/~ajmal/recognition.html

5.1.4 Results

Figure 8 reports the results concerning Experiment 1. As it can be seen, on both datasets the proposed approach yields notably improved recognition capabilities with respect to *Cluster* and *GC*. In particular, on both datasets the proposed Hough-based recognition approach always yields higher Recall at the same level of Precision. It is also worth pointing out that, in our experiments on Dataset 1, the *GC* approach, unlike the other two, turned out to be particularly sensitive to the threshold parameter. In particular, the selected correspondence subset is always characterized by a smaller cardinality compared to the other two approaches, this denoting a worse capability of consensus grouping for *GC*. Moreover, also in terms of *Recognition vs. Occlusion* curves the proposed approach yields improved results compared to the other two. As for efficiency, with these parameters and in our experiments the proposed approach and *GC* are overall much more efficient than *Cluster* (i.e., they run more than one order of magnitude faster).

5.2 Experiment 2

In this experiment, we evaluate quantitatively the performance of the proposed algorithm extended to RGB-D data as proposed in Section 4. The evaluation relies on data acquired by the recently introduced Microsoft *Kinect* sensor, which delivers texture-augmented range maps in real-time. First of all, in order to better handle Point-of-View (PoV) variations of the models of the database, during the offline training stage we include in our database different views of each model. Then, we extract repeatable features from the RGB image of each model view by means of the SURF [3] detector and then describe them with the

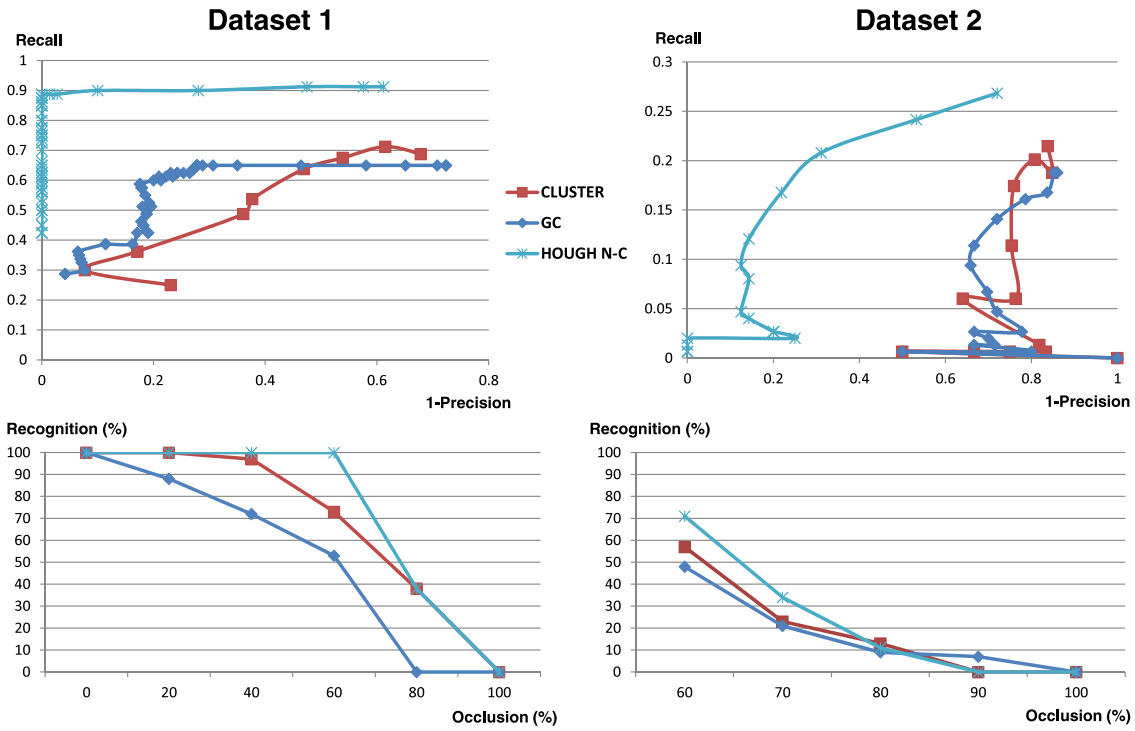

Fig. 8 Results in terms of *Precision vs. Recall* (above) and *Recognition vs. Occlusion* of the considered geometric validation methods. Left: 188 object tests on Dataset 1. Right: 200 tests on Dataset 2.

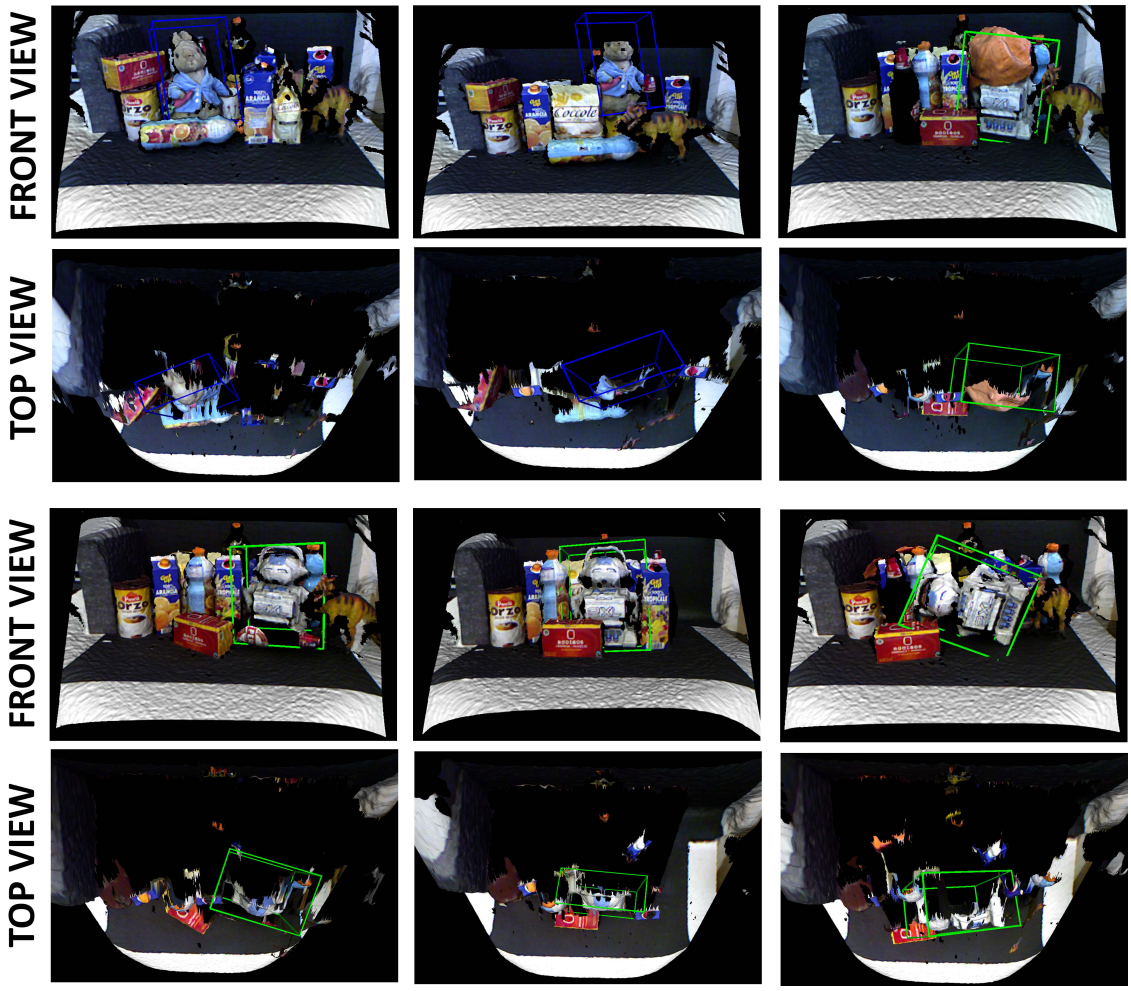

Fig. 9 Qualitative results of the proposed method on RGB-D data acquired by a *Kinect* device with a database including two models (*Bunny, Robot*). Here, only one model is present in each scene.

Fig. 10 Qualitative results of the proposed method applied on RGB-D data acquired by a *Kinect* device with a database including two models (*Bunny, Robot*). Here, both models are present in each scene.

SURF descriptor [3]. This set of extracted features is also back-projected in the 3D space by means of the range map and described by means of the SHOT descriptor [30]. After the matching stage, whereby we match 2D and 3D feature descriptions of the scene against those of each model view, the 3D feature "superset" yielded by point-to-point correspondences on each model view is fed to our Hough voting object recognition algorithm. For each model, the view with the highest number of correspondences that "survive" the Hough-based geometric validation stage is chosen as the best one: if this number is higher than a pre-defined recognition threshold, the model is detected in the scene. Also, in case of recognition, its 3D pose is determined by applying a final RANSAC-based Absolute Orientation stage [12] on the remaining correspondences, so as to yield a Rotation matrix and Translation vector that align the chosen model view to the scene. This algorithm is repeated for each model of the database.

Figures 9, **10** show some qualitative results concerning the RGB-D object recognition experiment. As for this experiment, we are trying to recognize two models: *Bunny*, characterized by a green bounding box, and *Robot*, characterized by a blue bounding box. For each model, 4 views at different angles are stored offline. Then, different scenes are built (12 of them are shown in the figures), each characterized by a significant degree of clutter and occlusions. For each scene, both figures show a front view and a top view of the output of the algorithm, represented by the acquired RGB-D scene, with a superimposed colored bounding box in case a model is detected. More specifically, Fig. 9 shows scenes where only one model is present in each scene, while Fig. 10 includes scenes where both models are present in each scene. All parameters of all stages of the proposed algorithm are kept constant throughout the whole experiment. As it can be seen, the proposed algorithm is able to perform accurate and robust object recognition and 3D pose estimation in the challenging scenario represented by this experiment. It is worth noting that when only one of the two models is present in the scene, the other one is, correctly, not found. Our implementation runs at approximatively $4 \sim 5$ seconds per scene, including all stages, from 3D data acquisition to 3D pose estimation and visualization, on a Intel *i*795 processor. No optimization is present except for parallelization of the code concerning the 3D feature description stage relying on the SHOT algorithm.

6. Conclusion

We have proposed a novel approach based on a 3D Hough voting process for detection and localization of free-form objects in range images, such as those provided by e.g., laser scanners and stereo vision sensors. Quantitative experiments show that our method outperforms clearly the algorithms chosen as represen-

tative of the two main existing approaches, i.e., those relying on geometric consistency and pose space clustering. We have also provided results concerning the extension of our method to RGB-D data acquired by a Microsoft *Kinect* sensor, which demonstrate that our approach is effective in performing object recognition in complex scenes characterized by a significant degree of clutter and occlusion.

Reference

[1] Ashbrook, A., Fisher, R., Robertson, C. and Werghi, N.: Finding surface correspondence for object recognition and registration using pairwise geometric histograms, *Proc. European Conference on Computer Vision*, pp.674–686 (1998).

[2] Ballard, D.: Generalizing the Hough transform to detect arbitrary shapes, *SPIE Proc. Vision Geometry X*, Vol.13, No.2, pp.111–122 (1981).

[3] Bay, H., Ess, A., Tuytelaars, T. and Gool, L.V.: SURF: Speeded Up Robust Features, *CVIU*, Vol.110, No.3, pp.346–359 (2008).

[4] Beis, J. and Lowe, D.: Shape indexing using approximate nearest-neighbour search in high dimensional spaces, *Proc. CVPR*, pp.1000–1006 (1997).

[5] Chen, H. and Bhanu, B.: 3D free-form object recognition in range images using local surface patches, *Pattern Recogn. Lett.*, Vol.28, No.10, pp.1252–1262 (2007).

[6] Chen, H. and Bhanu, B.: 3D free-form object recognition in range images using local surface patches, *J. Pattern Recogn. Lett.*, Vol.28, pp.1252–1262 (2007).

[7] Chua, C.S. and Jarvis, R.: Point Signatures: A New Representation for 3D Object Recognition, *IJCV*, Vol.25, No.1, pp.63–85 (1997).

[8] Davis, J., Nehab, D., Ramamoothi, R. and Rusinkiewicz, S.: Spacetime Stereo: A Unifying Framework for Depth from Triangulation, *PAMI*, Vol.27, No.2, pp.1615–1630 (2005).

[9] Frome, A., Huber, D., Kolluri, R., Bülow, T. and Malik, J.: Recognizing Objects in Range Data Using Regional Point Descriptors, *ECCV*, pp.224–237 (2004).

[10] Grimson, W. and Huttenlocher, D.: On the sensitivity of geometric hashing, *Proc. Int. Conf. Computer Vision*, pp.334–338 (1990).

[11] Hetzel, G., Leibe, B., Levi, P. and Schiele, B.: 3D Object Recognition from Range Images using Local Feature Histograms, *Proc. CVPR*, pp.394–399 (2001).

[12] Horn, B.: Closed-form solution of absolute orientation using unit quaternions, *J. Optical Society of America A*, Vol.4, No.4, pp.629–642 (1987).

[13] Hough, P.: Methods and means for recognizing complex patterns, US Patent 3069654 (1962).

[14] Johnson, A. and Hebert, M.: Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes, *PAMI*, Vol.21, No.5, pp.433–449 (1999).

[15] Johnson, A.E. and Hebert, M.: Surface Matching for Object Recognition in Complex 3-D Scenes, *Image and Vision Computing*, Vol.16, pp.635–651 (1998).

[16] Khoshelham, K.: Extending generalized Hough Transform to detect 3D objects in laser range data, *Proc. ISPRS Ws. Laser Scanning*, pp.206–210 (2007).

[17] Lamdan, Y. and Wolfson, H.J.: Geometric Hashing: A General And Efficient Model-based Recognition Scheme, *Proc. Int. Conf. Computer Vision*, pp.238–249 (1988).

[18] Lamdan, Y. and Wolfson, H.: On the error analysis of 'geometric hashing', *Proc. Conf. Computer Vision and Pattern Recognition*, pp.22–27 (1991).

[19] Li, X., Godil, A. and Wagan, A.: Spatially Enhanced Bags of Words for 3D Shape Retrieval, *Proc. ISVC*, pp.349–358 (2008).

[20] Liu, Y., Zha, H. and Qin, H.: Shape topics: A compact representation and new algorithms for 3d partial shape retrieval, *Proc. CVPR* (2006).

[21] Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints, *IJCV*, Vol.60, pp.91–110 (2004).

[22] Mian, A., Bennamoun, M. and Owens, R.: A Novel Representation and Feature Matching Algorithm for Automatic Pairwise Registration of Range Images, *IJCV*, Vol.66, No.1, pp.19–40 (2006).

[23] Mian, A., Bennamoun, M. and Owens, R.: On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes, *Int. J. Computer Vision*, Vol.89, No.2, pp.348–361 (2009).

[24] Novatnack, J. and Nishino, K.: Scale-Dependent 3D Geometric Features, *Proc. Int. Conf. Computer Vision*, pp.1–8 (2007).

[25] Novatnack, J. and Nishino, K.: Scale-Dependent/Invariant Local 3D Shape Descriptors for Fully Automatic Registration of Multiple Sets of Range Images, *ECCV*, pp.440–453 (2008).

[26] Ohbuchi, R., Osada, K., Furuya, T. and Banno, T.: Salient local visual features for shape-based 3D model retrieval, *Proc. Int. Conf. Shape Modeling and Applications*, pp.93–102 (2008).

[27] Rabbani, T. and Van Den Heuvel, F.: Efficient Hough Transform for automatic detection of cylinders in point clouds, *ISPRS Ws. Laser Scanning*, pp.60–65 (2005).

[28] Stein, F. and Medioni, G.: Structural Indexing: Efficient 3-D Object Recognition, *PAMI*, Vol.14, No.2, pp.125–145 (1992).

[29] Tangelder, J.W.H. and Veltkamp, R.C.: A survey of content based 3d shape retrieval methods, *Proc. Shape Modeling International*, pp.145–156 (2004).

[30] Tombari, F., Salti, S. and Di Stefano, L.: Unique Signatures of Histograms for local surface description, *Proc. ECCV 2010* (2010).

[31] Tsui, H. and Chan, C.: Hough technique for 3D object recognition, *IEE Proceedings*, Vol.136, No.6, pp.565–568 (1989).

[32] Vosselman, G., Gorte, B., Sithole, G. and Rabbani, T.: Recognising structure in laser scanner point cloud, *Int. Arch. of Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol.46, pp.33–38 (2004).

[33] Zhang, L., Curless, B. and Seitz, S.: Spacetime Stereo: Shape Recovery for Dynamic Scenes, *Proc. CVPR* (2003).

[34] Zhang, Z.: Iterative point matching for registration of free-form curves and surfaces, *Int. J. Computer Vision*, Vol.13, No.2, pp.119–152 (1994).

[35] Zhong, Y.: Intrinsic Shape Signatures: A Shape Descriptor for 3D Object Recognition, *Proc. 3DRR Workshop* (*in conj. with ICCV*) (2009).

Federico Tombari received his B.Eng and M.Eng from University of Bologna respectively in 2003 and 2005. From the same institution he received a Ph.D. in Computer Science Engineering in 2009. Currently he is a Senior Post-Doc at Computer Vision Lab, DEIS (Department of Electronics, Computer Science and Systems), University of Bologna. His research interests focus on computer vision and pattern recognition, in particular they include robot and stereo vision, 2D/3D object recognition, algorithms for video-surveillance. He has coauthored more than 40 refereed papers on peer-reviewed international conferences and journals and he is a member of IEEE and IAPR-GIRPR.

Luigi Di Stefano received a degree in Electronic Engineering from University of Bologna, Italy, in 1989 and a Ph.D. degree in electronic engineering and computer science from Department of Electronics, Computer Science and Systems (DEIS) at University of Bologna in 1994. In 1995, he spent six months at Trinity College Dublin as a postdoctoral fellow. He is currently an associate professor at DEIS. In 2009 he joined the Board of Directors of Datalogic SpA as an Independent Director. His research interests include computer vision, image processing, and computer architecture. He is the author of more than 100 papers and five patents. He is a member of IEEE Computer Society and IAPR-IC.

(Communicated by *Han Wang*)