16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Recognition and 6D Pose Estimation of Large-scale Objects using 3D Semi-Global Descriptors

David Nospes
TH Mittelhessen
Giessen, Germany
david.nospes@ei.thm.de

Kirill Safronov
KUKA Deutschland
Augsburg, Germany
kirill.safronov@kuka.com

Sarah Gillet
KUKA Deutschland
Augsburg, Germany
sarah.gillet@kuka.com

Klaus Brillowski
TH Mittelhessen
Friedberg, Germany
klaus.brillowski@m.thm.de

Uwe E. Zimmermann
KUKA Deutschland
Augsburg, Germany
uwe.zimmermann@kuka.com

Abstract

While the main focus of 3D object recognition is on small human manipulable objects, we face the problem of recognizing large-scale objects. These objects have a huge impact on mobile robot manipulation and navigation tasks, especially if the object is only partially visible and due to its size is far away from the camera and robot. In our work, we propose a framework capable of recognition and pose estimation for large-scale objects. We propose the use of semi-global descriptors for scene segments and model views in combination with up-sampling and segment label merging techniques. To achieve high accuracy, the initially estimated pose is first refined and afterwards verified. A performance comparison between different model descriptors shows that the chosen semi-global descriptor gives most promising results. By using simultaneous reconstruction, segmentation and recognition, we have built a framework which recognizes large-scale objects and estimates their 6D poses.

1 Introduction and related work

With an increasing number of mobile robots being deployed in both factory floors and service environments, the need for understanding the 3D environment arises. Object recognition and pose estimation of large-scale objects are important aspects of 3D environments since they are often only partially visible in the sensory system of the mobile robot. Through additional information about pose and object classification, semantic information, e.g. possible directions of movement, can not only increase the performance of obstacle avoidance but also support approaching large objects.

The problem of scene exploration using visual sensors is well-known as *visual SLAM*. The recent advances in this field allow us to get a dense 3d map of the environment with millimeter-accuracy. An example for these SLAM systems is presented in [1]. This system is supposed to be used with RGB-D, stereo and monocular cameras. It can estimate the camera trajectory using ORB and applying a loop-closing technique to reconstruct the environment. If the camera position is lost the relocalization based on searching for the correspondences in the bag-of-visual-words database is carried out. Other SLAM systems for RGB-D cameras are proposed in [2] and [3]. A very promising system that can reconstruct the environment, calculate the camera trajectory and simultaneously carry out segmentation of the generated 3d point cloud is presented in [4]. It analyzes the depth images in order to provide real-time segmentation of the reconstructed 3d map. The next step is to detect and classify objects as well as to estimate their 6DOF poses. Therefore 3d surface based methods are suitable and more effective in our case. They can be divided into following groups: methods based on using local, semi-global and global descriptors. Methods of the first group use local descriptors calculated for the single keypoints and their neighborhood in the 3d point cloud such as NARF [5], PPF [6], RSD [7], etc. Local methods are preferably used in cluttered and occluded scenes. The global approaches such as VFH [8], GRSD [9], ESF [10] etc. use a histogram that can describe the complete surface represented by a single point cloud. This can help decreasing the computational complexity of the matching process. Methods based on OUR-CVFH descriptors matching [11] can be assigned to the group of semi-global approaches. Such methods combine both advantages of the global and local approaches. Addressing 3D object detection and pose estimation, recent work utilizes these local handcrafted features when working with RGB-D input data. C. Tsai et al. present an approach using CSHOT for enhanced description of the local features with color information leading to higher accuracy in the matching state followed with a hough voting scheme for pose extraction[12]. J. Vidal et. al. propose a similar approach using a variation of the local PPF for matching and and a hough-like voting for pose estimation.

2 Recognition and pose estimation framework

In this paper, we propose a framework for simultaneous recognition and pose estimation of large-scale objects. This framework consists of two modules:

- reconstruction and segmentation
- object recognition and pose estimation

that are described in the following sections. The complete object recognition and pose estimation pipeline of the proposed method is presented in figure 1.

2.1 Reconstruction and Segmentation

The reconstruction and segmentation module is implemented according to the work of Tateno et al. [13]. The method used in this paper combines the point based fusion method described in [14] and the segmentation method presented in [15]. The reconstruction method uses depth images and surface normals and is similar to Kinect Fusion technique proposed in [16]. But instead of creating a voxel-based representation of the environment the proposed SLAM algorithm uses a point-based representation that contains normal information for local regions (so-called *surfel*). This leads to lower memory consumption. After estimating the camera position the current depth image can be merged into the *global environment model* [4]. The segmentation method is an adapted version of the method described in [15] that uses depth images. Each depth image is transformed into the binary *edge map* and the *normal edge analysis* is carried out. That way the *global segmentation map* is calculated during the environment reconstruction. From the single depth images extracted segments are grouped into clusters. Each cluster gets an unique label. Furthermore some clusters can be merged into one. A scene reconstructed and segmented using the *InSeg* method proposed in [4] is presented in figure 2.

2.2 Object recognition and pose estimation

In this subsection we present our approach to recognition and pose estimation of large-scale objects based on 3d semi-global pipeline presented in [17].

Instead of matching the complete 3D scene segment against the full 3D model of the object like [13], our method is based on matching single viewpoint data with partial 3D model views, namely descriptors for classification and point cloud data for pose estimation. We achieve this by extracting 3D scene segments from one single depth image and utilizing semi-global descriptors for every model view. Therefore we expect that most large-scale models will be too big to be captured with one view. The usage of segment smooth surfaces describing semi-global descriptors for object recognition for our often not fully visible large-scale objects comes from the idea, that describing their surfaces leads to correct recognition, even in cases were only a few of them are visible. Furthermore large scale objects are rarely seen from close up. That leads to sparse point clouds which need to be upsampled to achieve good classification results. Additionally, unlike [17] we are computing our descriptors only on real world data and do not use virtual views of 3D CAD models. Therefore we can extend our database with views of new models quickly. In contrast to [12] we rely on point clouds without color information for feature computation because it is more stable against illumination changes which typically show up while scanning large objects in industrial environment. Additionally the pose estimation step is performed using the computed semi-global feature properties and ICP rather than using Hough-like voting scheme with local features like in [12] and [18].

Object Modeling The object model in our framework is represented by the set of different 3d views of an object with the calculated semi-global descriptors for each view. These views can be captured manually or can be generated from CAD data including a reference. Each view is represented in the world coordinate system.

Descriptors After the segmentation step, each segmented scene cluster on is matched with the model views using semi-global descriptors that describe partial, smooth segments of the point cloud (i.e. OUR-CVFH, etc.). We discovered that semi-global descriptors are more suitable for our purposes since the processed scene is segmented and each point cloud segment has to be classified before pose estimation. So the matching between two semi-global descriptors represented by 2d histograms is faster than the matching of many local ones.

Our approach is based on using OUR-CVFH (Oriented Unique Repeatable Clustered Viewpoint Feature Histogram) descriptors presented in [19]. This descriptor compensates the disadvantages of local and global descriptors using semi-global features and provides an initial 6DOF pose estimation. In this way, the OUR-CVFH descriptor is calculated for each model view and stored together with the SGURF (Semi-global Unique Reference Frame) - the repeatable reference frame [11]. A scene segment and its corresponding OUR-CVFH histogram are shown in figure 3.

Object Recognition After the reconstruction and segmentation step, each depth map cluster of the scene gets a unique label and is considered as a separated point cloud. 3d points that do not correspond to any cluster are discarded.

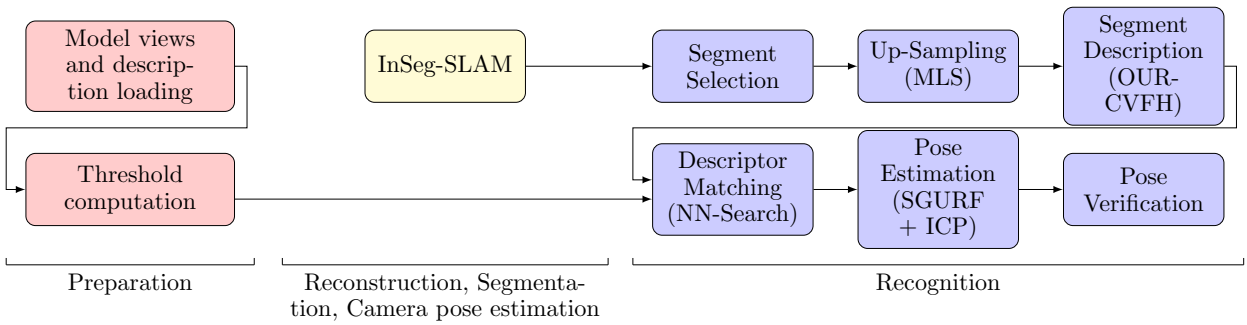


Figure 1: Object recognition and pose estimation pipeline

Figure 2: Scene reconstructed and segmented with *InSeg*

Before object classification can be carried out these extracted clusters must be preprocessed with the following steps:

1. removing outliers
2. ignoring small clusters: clusters that are interpreted as small using the metric computed in the *threshold computation* stage (i.e. minimal volume, size of the bounding box, etc.) are ignored
3. up-sampling only of clusters of far away objects with MLS: point clouds that represent far away objects have very small density

After the preprocessing step, an OUR-CVFH histogram is calculated for each cluster and matched with the histogram of every view of every model in the database using a nearest-neighbor search. The best match is the model with the smallest euclidean distance between the histograms and the highest confidence level. This confidence level is computed by the number of matching histograms from the same model.

Pose Estimation Rough pose estimation is defined using SGURFs, which are calculated for the corresponding scene cluster and model view. If it is not possible to calculate the SGURF for a certain point cloud, the transformation between the centroids of both clouds is used.

The pose refinement is carried out using a normal based ICP algorithm with a point-to-plane error metric.

Hypotheses Verification and Classification Check The hypotheses verification used in our approach is based on the method proposed in [11]. It rejects false recognitions by analyzing geometrical constraints of hypotheses and comparing them with scene constraints.

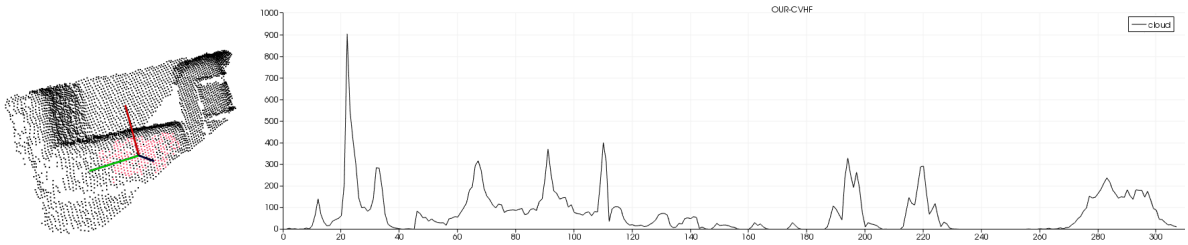
After matching the descriptors and estimating the pose, we obtain one hypothesis for each selected segment. Depending on the number of unrecognized segments and found correspondences in the current frame, multiple hypotheses are verified.

Furthermore since we also have complex large objects in our database, segmentation may fail to group the depth map pixels correctly. Thus we check the area around every classified segment for segments with the same classification for merging.

3 Experimental results

To evaluate our proposed attempt to recognize large-scale objects for mobile robotics, we collected a dataset of objects in industrial and business environments. An extract of the dataset is shown in figure 4. For visualization purposes in this image the point cloud data from different views has been merged though object recognition is carried out by single view matching. The chosen objects are of different dimensions: from small boxes and stools to big couches and robots. Additionally, the generated models are noisy due to their acquisition with structured light cameras.

Figure 5 shows a reconstructed scene of a business environment containing large-scale objects such as armchairs, tables, office chairs etc. All objects are placed to represent a realistic working area for mobile service robots.



(a) Scene segment and SGURF

(b) OUR-CVFH histogram. The x-axis spans the histogram bins, the y-axis the histogram values.

Figure 3: Visualization of the semi-global descriptor

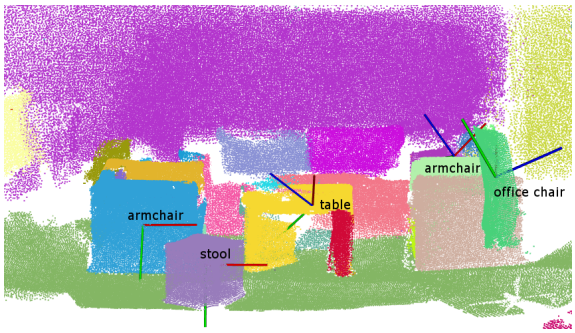


Figure 6: Segmented scene with recognized objects

Figure 7: Original scene

Figure 4: Stool, Office Chair, VALERI-System [20]

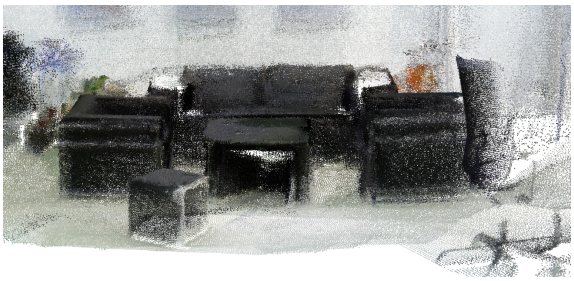


Figure 5: Reconstructed scene

The segmented scene with all recognized objects and their coordinate systems in respect to the world coordinate system is shown in figure 6. An extract of the original scene is shown in figure 7. As we are identifying large-scale items, each object may contain multiple unique classified labels. When those labels are merged or deleted during the segmentation phase, their classification is updated to avoid multiple or invalid classifications. Additionally, if differently classified labels are merged reclassification with corresponding models is performed.

For testing the pipeline performance, a computer with Intel Core i7 2,4 GHz CPU and 16 Gb RAM was used. The implementation is running on CPU without using GPU acceleration. The average execution time is presented in table 1. To evaluate the approach of using semi-global descriptors to recognize large-scale objects, we compare the semi-global descriptor with the global descriptors ESF (Ensemble of Shape Functions) [10] and GRSD (Global Radius-based Surface Descriptor). ESF is capable to deal with partial point clouds due to its robustness to noise and holes. GRSD on the other hand is a fast global descriptor that can also use point clouds with low resolution. Therefore, it could be useful for navigation through industrial and business environment in real-time.

Figure 8 shows the performance difference between OUR-CVFH (with upsampling using MLS), ESF and GRSD estimation on scene segments over a range of 300 frames. ESF uses a combination of three different shape functions for every point in a voxelized

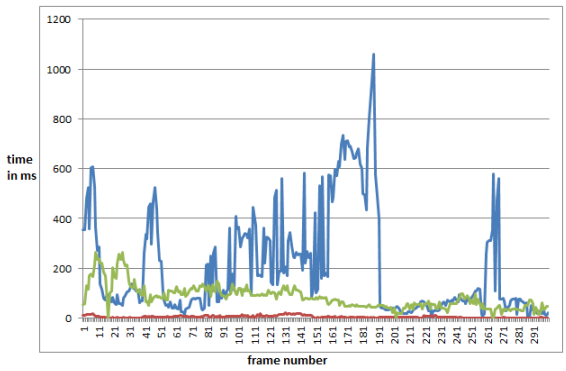


Figure 8: Descriptor estimation time of OUR-CVFH with upsampling (blue), ESF (green) and GRSD (red) in ms

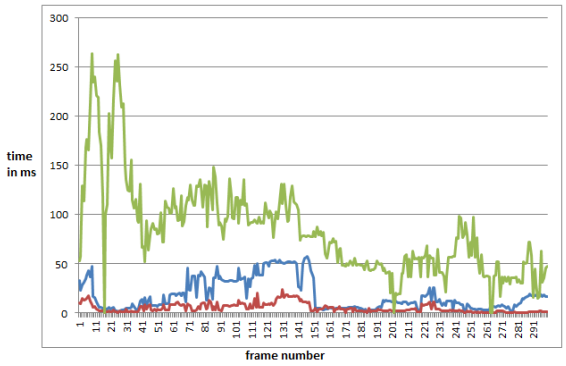


Figure 9: Descriptor estimation time of OUR-CVFH without upsampling (blue), ESF (green) and GRSD (red) in ms

	InSeg SLAM
Reconstruction And Segmentation	51.77
Object Recognition	
Descriptor Matching	2.85
Pose Estimation	19.93
Hypothesis Verification	14.85
Classification Correction	2.66

Table 1: Mean of the measured execution times in ms over 1000 frames of a recorded scene with a depth map resolution of 640 × 480.

point cloud to describe the distances, angles and regions. GRSD first estimates labels of the cloud and classifies them. Then the whole cloud is handled as a plane, cylinder, edge, rim or sphere and the surface is described using the relationships between keypoints and their nearest neighbors. While GRSD provides the fastest computation with a range from 0.2 to 18 ms, it does not need a high cloud density because of its internal use of a voxelization with the radius of 2.5 cm [9]. ESF also uses a voxel grid approach internally, but takes up to 14 times the computation time of GRSD. In comparison, the computation of OUR-CVFH for a point cloud with a resolution lower than the original model cloud is computationally intensive due to the upsampling. The success of the matching process with OUR-CVFH depends on the similarity of model and segment resolution. The performance difference of the OUR-CVFH estimation without upsampling compared to the computation of ESF and GRSD on scene segments over a range of 300 frames is shown in figure 9. The computational burden of OUR-CVFH is significantly smaller than of upsampling. The maximum required time in this case is three times the maximum of GRSD computation. With the possibility to use four threads to estimate the descriptors of multiple segments simultaneously, the maximum time could be decreased by another 30 percent. Descriptor estimation times for the keyframes are presented in table 2.

In our large scale objects dataset tests we achieved at best a true positive recognition rate of 83% in the scene for OUR-CVFH. By comparison, using the global descriptor ESF only 66% were found, while GRSD, having by far the smallest computational burden, provided even less recognitions. Over longer distances the resolution of the cloud decreases, reducing the recognition rate of OUR-CVFH. To counter this reduction, MLS increases the amount of points in these clusters and leads to better results. The down-side though is prevention of fast frame processing. In contrast GRSD provides fast recognition results independent of the cloud resolution.

4 CONCLUSIONS AND FUTURE WORK

The experimental results show that the proposed framework is able to recognize large objects and estimate their pose. We show that the combination of OUR-CVFH descriptors is most suitable for recognizing large-scale objects. To be able to perform promising 3D obstacle avoidance, the objects, in addition, have to be detected as early as possible. The introduction of up-sampling for sparse segments allows us to recognize objects which are further away.

Future work includes performance tests of the framework on a mobile platform moving in a real environment. Especially in industrial environments, large-scale objects can be harmful to the robot if not the full knowledge is used to avoid or approach these objects. Based on the performance and upcoming scenarios, one could imagine new motion planner which can deal with semantic information of objects.

Acknowledgement This work has been partially funded by the European Union through the project

Frame	OUR-CVFH (upsampling)	OUR-CVFH	ESF	GRSD
25	56,51	1,98	235,62	0,93
50	230,26	7,99	85,98	2,61
75	79,49	37,52	109,22	2,28
100	357,96	32,51	95,99	8,81
125	514,39	52,91	77,01	7,76
150	422,63	38,27	77,97	2,01
175	640,49	5,96	50,46	4,95
200	39,04	12,58	20,27	2,97
225	29,93	25,51	55,87	3,68
250	73,62	7,67	56,57	1,77

Table 2: Descriptor estimation times at specific frames on a recorded scene with large scale objects in ms.

RobDREAM (grant 64540) and by the German Federal Ministry of Education and Research (BMBF) through the project Hybr-iT (grant 01IS16026A).

References

[1] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras," IEEE Trans. Robot., vol. 33, no. 5, pp. 1255–1262, 2017.

[2] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-d mapping with an RGB-D camera," IEEE Trans. Robot., vol. 30, no. 1, pp. 177–187, 2014.

[3] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," CoRR, vol. abs/1607.02565, 2016.

[4] K. Tateno, F. Tombari, and N. Navab, "Real-time and scalable incremental segmentation on dense SLAM," in 2015 IEEE/RSJ Int. Conf. on Intell. Robots and Systems, IROS 2015, Hamburg, Germany, pp. 4465–4472, 2015.

[5] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "Point feature extraction on 3d range scans taking into account object boundaries," in IEEE Int. Conf. on Robot. and Aut., ICRA 2011, Shanghai, China, pp. 2601–2608, 2011.

[6] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in 23rd IEEE Conf. on Comp. Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, pp. 998–1005, 2010.

[7] Z. Marton, D. Pangercic, N. Blodow, J. Kleinehellefort, and M. Beetz, "General 3d modelling of novel objects from a single view," in 2010 IEEE/RSJ Int. Conf. on Intell. Robots and Systems, IROS 2010, Taipei, Taiwan, pp. 3700–3705, 2010.

[8] R. B. Rusu, G. R. Bradski, R. Thibaux, and J. M. Hsu, "Fast 3d recognition and pose using the viewpoint feature histogram," in 2010 IEEE/RSJ Int. Conf. on Intell. Robots and Systems, IROS 2010, Taipei, Taiwan, pp. 2155–2162, 2010.

[9] Z. C. Marton, D. Pangercic, N. Blodow, and M. Beetz, "Combined 2D-3D Categorization and Classification for Multimodal Perception Systems," The Int. Jour. of Robotic Research, vol. 30, pp. 1378–1402, September 2011.

[10] W. Wohlkinger and M. Vincze, "Ensemble of shape functions for 3d object classification," in 2011 IEEE Int. Conf. on Robot. and Biomimetics, ROBIO 2011, Karon Beach, Thailand, pp. 2987–2992, 2011.

[11] 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 Pattern Recognition - Joint 34th DAGM and 36th OAGM Symposium, Graz, Austria, pp. 113–122, 2012.

[12] C. Tsai and S. Tsai, "Simultaneous 3d object recognition and pose estimation based on RGB-D images," IEEE Access, vol. 6, pp. 28859–28869, 2018.

[13] K. Tateno, F. Tombari, and N. Navab, "When 2.5d is not enough: Simultaneous reconstruction, segmentation and recognition on dense SLAM," in 2016 IEEE Int. Conf. on Robot. and Aut., ICRA 2016, Stockholm, Sweden, pp. 2295–2302, 2016.

[14] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, and A. Kolb, "Real-time 3d reconstruction in dynamic scenes using point-based fusion," in 2013 Int. Conf. on 3D Vision, 3DV 2013, Seattle, Washington, USA, pp. 1–8, IEEE Computer Society, 2013.

[15] A. Ückermann, C. Elbrechter, R. Haschke, and H. J. Ritter, "3d scene segmentation for autonomous robot grasping," in 2012 IEEE/RSJ Int. Conf. on Intell. Rob. and Syst., IROS 2012, Vilamoura, Algarve, Portugal, pp. 1734–1740, IEEE, 2012.

[16] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. W. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," in 10th IEEE Int. Symp. on Mixed and Augm. Reality, ISMAR 2011, Basel, Switzerland, pp. 127–136, IEEE, 2011.

[17] 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 2013 IEEE Int. Conf. on Robot. and Aut., Karlsruhe, ICRA 2013, Germany, pp. 2104–2111, IEEE, 2013.

[18] J. Vidal, C. Lin, and R. Martí, "6d pose estimation using an improved method based on point pair features," CoRR, vol. abs/1802.08516, 2018.

[19] A. Aldoma, F. Tombari, L. di Stefano, and M. Vincze, "A global hypotheses verification method for 3d object recognition," in Computer Vision - ECCV 2012 - 12th Eur. Conf. on Computer Vision, Florence, Italy, Proceedings, Part III (A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), vol. 7574 of Lecture Notes in Computer Science, pp. 511–524, Springer, 2012.

[20] "Valeri project, http://www.valeri-project.eu/," 2012.