ADAPTIVE FUSION-BASED 3D KEYPOINT DETECTION FOR RGB POINT CLOUDS

Muhammad Zafar Iqbal⋆*, Dmytro Bobkov, and Eckehard Steinbach*

Chair of Media Technology, Department of Electrical and Computer Engineering
Technical University of Munich (TUM), Germany

ABSTRACT

We propose a novel keypoint detector for 3D RGB Point Clouds (PCs). The proposed keypoint detector exploits both the 3D structure and the RGB information of the PC data. Keypoint candidates are generated by computing the eigenvalues of the covariance matrix of the PC structure information. Additionally, from the RGB information, we estimate the salient points by an efficient adaptive difference of Gaussian-based operator. Finally, we fuse the resulting two sets of salient points to improve the repeatability of the 3D keypoint detector. The proposed algorithm is compared against the state-of-the-art algorithms on two benchmark datasets. The experimental results show that the proposed scheme outperforms the best existing method by 5.35% and 60.98 points on the SHOT-Kinect dataset and by 5.45% and 145.54 points on the SHOT-SpaceTime dataset in terms of relative and absolute repeatability, respectively.

Index Terms— 3d keypoint detector, difference of Gaussian, point cloud, salient point

1. INTRODUCTION

In the past years, 3D data processing has received significant attention from the research community fuelled by the availability of low-cost 3D sensory devices like the Microsoft Kinect. The detection of distinctive and repeatable 3D keypoints is an important aspect for 3D data processing. A keypoint detection scheme not only reduces the computational complexity by decreasing the number of processed points used for shape description and matching, but also improves the performance of various applications like objection recognition and classification. Therefore, in the last decade, 3D keypoint detection schemes have gained much attention, and several algorithms have been proposed (see [1, 2, 3] for an overview). Despite significant research effort, the repeatability or distinctiveness of the existing algorithms still needs to be further improved. Therefore, 3D keypoint detection is still an open topic and demands for more robust detection schemes with high repeatability and distinctiveness.

A straightforward and computationally efficient keypoint detection approach is based on sparse surface sampling and mesh decimation. These techniques are, however, inefficient in terms of repeatability and distinctiveness [2]. Existing 3D keypoint detectors either use a predefined/fixed scale space or adaptively select the scale space for the detection of salient points. Therefore, the keypoint detectors can be broadly classified into two main categories, namely fixed-scale detectors and adaptive-scale or scale-invariant detectors [2]. Examples of state-of-the-art fixed scale detectors are the Local Surface Patches (LSP) [4], Heat Kernel Signature (HKS) [5], Intrinsic Shape Signatures (ISS) [6], Key-Point Quality (KPQ) [7], Harris 3D [8], Harris 6D [9] and Histogram of Normal Orientation (HoNO) [10]. The adaptive-scale keypoint detection methods are motivated by the SIFT keypoint detector [11] for 2D images. In the adaptive-scale category, Laplace-Beltrami Scale Space (LBSS) [12], Salient Points (SP) [13], Mesh Difference of Gaussians (Mesh-DoG) [14], and Key-Point Quality Adaptive Scale (KPQ-AS) [7] are examples of state-of-the-art techniques. Due to space limitations, details of these techniques are omitted. For a comprehensive review of these techniques we refer the reader to [2, 3].

Recently, salient points have been estimated using learning-based techniques [15, 16, 17]. These techniques have two major issues. The first is that they require a large amount of training data for learning and the second issue is that the keypoints are learned for a specific descriptor, therefore, performance inconsistency is a major concern for descriptor replacement.

3D keypoint detection techniques available in the literature use either the Point Clouds (PCs) directly or convert the PCs into meshes for further processing. The conversion of PCs into meshes increases the computational cost especially for large PC datasets which typically consist of millions of points. These techniques utilize either the geometry described by PC data or alternatively the photometric appearance (RGB information) to detect the keypoints. Therefore, these techniques do not employ all accessible information (3D PC structure and RGB information) which limits the performance of these techniques. To address this issue, we propose a novel

The work of M. Z. Iqbal was supported by a Ph.D. scholarship provided by the Higher Education Commission (HEC) of Pakistan in collaboration with DAAD Germany.

Muhammad Zafar Iqbal, Student Member, IEEE, Dmytro Bobkov, Student Member, IEEE, and Eckehard Steinbach, Fellow, IEEE are with the Chair of Media Technology, Technical University of Munich, Munich 80333, Germany, (email: mzafar.iqbal@tum.de, dmytro.bobkov@tum.de, eckehard.steinbach@tum.de) ⋆ Corresponding author

fusion-based adaptive scale-space keypoint detector for RGB PCs. The proposed method detects salient points by leveraging; a) the local surface geometry described by the PC using EigenValue Decomposition (EVD) of the covariance matrix of the neighbourhood geometry and b) the intensity information of the photometric appearance by applying the Difference of Gaussian (DoG) operator. The proposed algorithm was tested on two benchmark datasets. The experimental results demonstrate the superiority of the proposed algorithm against the state-of-the-art algorithms in terms of absolute and relative repeatability and distinctiveness. In the literature, the distinctiveness and repeatability (absolute and relative) of keypoints are considered as an efficient and valuable evaluation metric for 3D detectors [2].

2. PROPOSED TECHNIQUE

We use the input PC, $P = \{\mathbf{p_i} \in \mathbb{R}^6 : i = 1, \cdots, N\}$ which contains both point coordinate and RGB information. The first three elements of $\mathbf{p_i}$ represent the point coordinates and the remaining three elements represent the RGB information, which are denoted as $\mathbf{p_i^{cor}}$ and $\mathbf{p_i^{rgb}}$, respectively in the rest of the manuscript. N is the number of points in P and i represents the point index. The proposed keypoint selection process on the input PC consists of three stages. In the first stage we process the geometrical information ($\mathbf{p_i^{cor}}$) of the PC by Adaptive EVD (AEVD) of the covariance matrix. In the second stage, we apply the adaptive DoG operator to the available RGB information ($\mathbf{p_i^{rgb}}$) and in the final stage, we fuse the two sets of candidate keypoints. In the following section, we discuss all these stages.

2.1. Adaptive EVD of the Covariance Matrix

To compute the saliency measure from the geometrical structure ($\mathbf{p_i^{cor}}$) of the PC we follow the approach presented in [14]. We calculate the EVD of the covariance matrix of all non-boundary points (boundary points are estimated by the technique presented in [18]) of the PC, for a predefined set of neighbourhood support radii $R = \{r^a \in \mathbb{R}^+ : a = 1, \cdots, A\}$ as:

$$\begin{aligned} \mathbf{Cov_i^a} &= \frac{1}{M}\sum_{j=1}^{M}(\mathbf{p_j^{cor}} - \mathbf{p_i^{cor}})(\mathbf{p_j^{cor}} - \mathbf{p_i^{cor}})^T, \quad \text{with} \\ NN_i^a &= \{\mathbf{p_j^{cor}} : |\mathbf{p_i^{cor}} - \mathbf{p_j^{cor}}| < r^a\}, \end{aligned} \tag{1}$$

where M represents the number of neighbouring points $\mathbf{p_j}$ in NN_i^a at support radius r^a of the point $\mathbf{p_i}$ and A represents the number of selected support radii (three different support radii were used for the evaluations). After that, the EVD is performed on $\mathbf{Cov_i^a}$ and the eigenvalues are recorded in descending order of magnitudes as $\lambda_{i1}^a, \lambda_{i2}^a, \lambda_{i3}^a$. We compute the highest eigenvalues across all selected scales as:

$$e_{it} = \max(\{\lambda_{it}^a : a = 1, \cdots, A\}), \quad \text{where} \quad t = 1, 2, 3 \tag{2}$$

To avoid the detection of points which have similar spread along the principal directions, we prune the salient points by the ratio of two successive eigenvalues against two threshold values Th_{12} and Th_{23} as follows:

$$prune = \frac{e_{i2}}{e_{i1}} > Th_{12} \wedge \frac{e_{i3}}{e_{i2}} > Th_{23}$$

Among the remaining points, the saliency is determined by the smallest eigenvalue e_{i3}, and only the points with large variations (greater than a threshold value Th_3) along the principal direction are considered. To further prune the salient points and to avoid multiple points from the same region we perform Non-Maxima Suppression (NMS) at support radius r^{nms}. The point with the largest e_{i3} value among the other points in support radius r^{nms} is the only selected keypoint.

2.2. Adaptive DoG

In parallel, we apply the DoG operator at multiple scales on the intensity values which are computed from the photometric appearance (RGB channel) of the PC $\mathbf{p_i^{rgb}}$. The intensity is computed from RGB values according to ITU-R specification/recommendation BT.601 as follows:

$$I_i = 0.299 * R_i + 0.587 * G_i + 0.114 * B_i, \tag{3}$$

where R_i, G_i and B_i represent the red, green and blue colour channel values at point $\mathbf{p_i}$. After that, we produce the scale space $S = \{s^b \in \mathbb{R}^+ : b = 1, \cdots, B\}$ and generate a Gaussian filter bank $g_{ik}(s^b)$ with these scales. We convolve the intensity values with the determined filter bank to obtain h_i^b as:

$$\begin{aligned} g_{ik}(s^b) &= exp(-\frac{|\mathbf{p_k^{cor}} - \mathbf{p_i^{cor}}|^2}{2(s^b)^2}), \\ h_i^b &= \frac{\sum_k I_k * g_{ik}(s^b)}{\sum_k g_{ik}(s^b)}, \quad \text{where} \\ s^b &= s_m * 2^{\left(\frac{b-1}{B}\right)}, \quad \text{and} \\ NM_i^b &= \{\mathbf{p_k^{cor}} : |\mathbf{p_i^{cor}} - \mathbf{p_k^{cor}}| < r^b\}. \end{aligned} \tag{4}$$

Here $\mathbf{p_k}$ represents the points within a predefined neighbourhood support radius $r^b = 3 * s^b$ of point $\mathbf{p_i}$, B is the number of scales and s_m represents the minimum scale. Thus, the saliency value is computed as:

$$d_i^b = h_i^b - h_i^{b-1}. \tag{5}$$

After the saliency (DoG) estimation, instead of comparing salient values in the current and the adjacent scales within a predefined neighbourhood as proposed in [11, 14], we calculate the maximum or minimum value of the DoG across all scales. This step not only reduces the concentration of the salient points in specified regions but also improves the detection performance, as shown in the experimental evaluation (Section 3). In particular, it is formally expressed as follows:

$$\begin{aligned} dd_i^{min} &= \min(d_i^b : b = 1, \cdots, B), \\ dd_i^{max} &= \max(d_i^b : b = 1, \cdots, B). \end{aligned} \tag{6}$$

We perform the NMS on dd_i^{min} and dd_i^{max} using a predefined neighbourhood support radius r^{nms} and recognize the point with the local minimum or maximum as the salient point. Further pruning steps are performed on the salient points using

two criteria. The first criterion is that the minimum contrast value ($|dd_i^{min}|$ or $|dd_i^{max}|$) should be greater than a threshold value Tr_{cn} to retain the salient point. The second criterion is based on the eigenvalues computed by EVD of the covariance matrix of the point $\mathbf{p_i^{rgb}}$ (RGB information) at a predefined neighbourhood support radius r^{ccv}. We compare the ratio of the adjacent eigenvalues (v_{i1}, v_{i2}, v_{i3} in descending order of magnitudes) with threshold values Tr_{12} and Tr_{23} and the smallest eigenvalue with the threshold value Tr_3 for pruning the salient points as:

$$retain = \frac{r_2}{r_1} > Tr_{12} \wedge \frac{r_3}{r_2} > Tr_{23} \wedge r_3 > Tr_3$$

2.3. Fusion of Salient Points

In the next step we fuse the salient points estimated by EVD (see Section 2.1) with the salient points detected by applying the DoG operator (see Section 2.2). The straightforward approach is to merge both sets of salient points, but experimentally it is observed that while directly adding the points increases the repeatability of the keypoint detector, selecting points in the neighbourhood of other points increases the computational complexity of the descriptor computation without improving the performance of various applications. Hence, we consider all the salient points detected from the geometric structure as keypoints while we take into account only the salient points from the DoG which are at a predefined distance Tr_d from previously selected points.

3. EXPERIMENTAL EVALUATION

The proposed algorithm is implemented in C++ using existing functions available in the PCL [9]. The results for the existing methods [2], the code for evaluation and the benchmark datasets are publicly available at [19]. Hence, we evaluate the proposed detection algorithm on the given code without any parameter tuning for a fair comparison. Fig. 1 shows the even distribution of detected keypoints by the proposed detector on Mario of the Shot-Kinect dataset [19].

3.1. Datasets

Our experiments were performed on two benchmark datasets (PCs with photometric information) for the evaluation of the proposed detector against the state-of-the-art algorithms. These benchmark databases (SHOT-SpaceTime and SHOT-Kinect) have been adopted for evaluation in [2] and are publicly available on the SHOT website [19]. For more details about these databases, we refer the reader to [2, 19].

3.2. Metrics

The most significant attribute of a keypoint detector is its repeatability, which describes the capability of the detector to find the corresponding set of keypoints on various instances of a given model with change in viewpoint, partial occlusion, clutter, and corruption with noise or combination of these impairments [1]. For the repeatability (absolute, relative) computation we use the metrics described in [2].

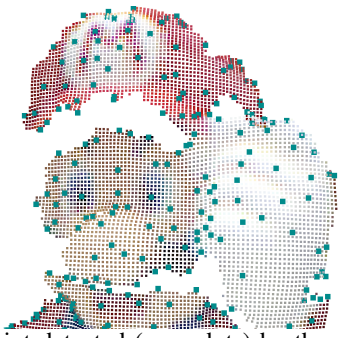

Fig. 1: Keypoint detected (cyan dots) by the proposed keypoint detector on Mario of the Shot-Kinect dataset. (Best viewed on screen.)

3.3. Parameters

We express the metric parameters like radii, distances, and NMS in Mesh Resolution (MR) as the average distance between neighbouring points. The parameter settings were provided in [2] for various state-of-the-art techniques. We only discuss the parameters which were used for the evaluation of the proposed algorithm. For experimental evaluation, the support radii for adaptive EVD of the covariance matrix were selected as $R = \{6MR, 8MR, 10MR\}$. The search radius for the boundary points was set at $1.2MR$. The threshold values were set as $Th_{12} = Th_{23} = 0.95$ and $Th_3 = 0.0005$ and NMS support radius was $r^{nms} = 4MR$.

For adaptive DoG estimation, we used the minimum scale as $s_m = 1MR$ while scale spaces $B = 3$, and NMS support radius was selected as $r^{nms} = 4MR$. Moreover, for pruning, the threshold value was $Tr_{12} = Tr_{23} = 0.5$ and $Tr_3 = 0.05$. We adjusted the threshold value of the minimum contrast at $Tr_{cn} = 0.1$, support radius (for RGB information covariance matrix calculation) $r^{ccv} = 6MR$ and fusion search radius was fixed at $Tr_d = 2MR$.

3.4. Repeatability Results

The proposed method is an unsupervised scheme, so we do not include the learning-based algorithms in this comparison [15, 16], as these techniques are associated with training and work for a specific descriptor only. In Table 1 the absolute and relative repeatability of the proposed algorithm is presented against the state-of-the-art algorithms for the SHOT-Kinect and SHOT-SpaceTime dataset. For better comparison, we consider both the fixed scale and adaptive scale detectors available in the literature. For the fixed scale detector we only consider the scale which gives the best results. Table 1 reveals that the proposed detector outperforms the related work in terms of both absolute and relative repeatability.

The experimental results in Table 2 exhibit the absolute and relative repeatability of the AEVD of the covariance matrix on the geometric structure of the PC, DoG [14], the pro-

Table 1: Absolute and relative repeatability of the proposed and the state-of-the-art algorithms on the SHOT-Kinect and SHOT-SpaceTime datasets [19]. (*Represents the proposed schemes.)

Repeatability	Detector	SHOT-Kinect	SHOT-SpaceTime
Absolute Repeatability	LSP [4]	18.6	63.75
	HKS [5]	3.43	12.13
	ISS [6]	7.47	47.42
	KPQ [7]	31.27	112.96
	Harris 3D [8]	1.35	7.46
	Harris 6D [9]	17.06	86.13
	HoNO [10]	49.9	257.21
	LBSS [12]	4.53	17.63
	SP [13]	39.73	96.92
	MeshDoG [14]	102.86	243.88
	KPQ-AS [7]	89.06	257.03
	PDoG + AEVD*	**163.84**	**402.75**
Relative Repeatability	LSP [4]	0.1678	0.2657
	HKS [5]	0.0873	0.1234
	ISS [6]	0.1243	0.4275
	KPQ [7]	0.2510	0.3940
	Harris 3D [8]	0.0804	0.2060
	Harris 6D [9]	0.2824	0.5849
	HoNO [10]	0.3133	0.5886
	LBSS [12]	0.0882	0.1578
	SP [13]	0.2817	0.3580
	MeshDoG [14]	0.4534	0.5611
	KPQ-AS [7]	0.3918	0.4639
	PDoG + AEVD*	**0.5069**	**0.6431**

posed DoG (PDoG) on RGB information, and the proposed fusion (PDoG + AEVD) method. During the experiments, all parameters remained the same as discussed previously in Section 3.3 except for DoG the number of scale space is five (without pyramid). These results indicate that even the PDoG, where the optimized (minima or maxima) value of DoG at scale space was used, outperforms the state-of-the-art algorithms. The absolute repeatability of PDoG on the SHOT-Kinect dataset is slightly less than for DoG, while the relative repeatability is much better. Moreover, the fusion of salient points enhances the absolute and relative repeatability of the SHOT-Kinect dataset with a minor decrease in relative repeatability and a noticeable improvement in absolute repeatability for the SHOT-SpaceTime dataset.

The PDoG and DoG are evaluated for different numbers of scales as shown in Table 3. One can observe from the table that the repeatability of the PDoG is noticeably unaffected by the number of scales, while the DoG repeatability drastically changes with the variation of the number of scales in the scale space. Hence, it can be argued that the PDoG is computationally efficient and scale invariant as compared to DoG [14].

The above given discussion and observations reveal that

Table 2: Performance evaluation of AEVD, DoG [14], PDoG, DoG + AEVD and PDoG + AEVD detectors on the SHOT-Kinect and the SHOT-SpaceTime datasets [19]. (*Represents the proposed schemes.)

Repeatability	Detector	SHOT-Kinect	SHOT-SpaceTime
Absolute Repeatability	AEVD	8.88	52.04
	DoG [14]	142.65	257.67
	PDoG*	134.27	337.63
	DoG + AEVD	**171.65**	326.13
	PDoG + AEVD*	163.84	**402.75**
Relative Repeatability	AEVD	0.1694	0.3879
	DoG [14]	0.4650	0.5548
	PDoG*	0.4772	**0.6492**
	DoG + AEVD	0.4989	0.5702
	PDoG + AEVD*	**0.5069**	0.6431

Table 3: Repeatability of DoG [14] and PDoG at different numbers of scale space on the SHOT-Kinect and the SHOT-SpaceTime datasets [19]. (*Represents the proposed schemes.)

Repeatability	Detector	Scale Space	SHOT-Kinect	SHOT-SpaceTime
Absolute Repeatability	DoG [14]	3	88.69	193.33
		5	142.65	257.67
	PDoG*	3	134.27	337.63
		5	136.94	333.83
Relative Repeatability	DoG [14]	3	0.3953	0.5113
		5	0.4650	0.5548
	PDoG*	3	0.4772	0.6492
		5	0.4792	0.6486

the proposed technique (PDoG + AEVD) exhibits superior performance in terms of both absolute and relative repeatability.

4. CONCLUSION

To the best of our knowledge, the proposed technique is the first keypoint detection scheme which uses both the geometrical structure of the PC as well as the RGB information to detect keypoints. We have observed that DoG with maximum and minimum value is more robust and scale invariant for the PC than the method presented in [14]. The fusion of salient points detected by PDoG from RGB information and salient points detected by AEVD from the PC structure improves the absolute and relative repeatability and distinctiveness of the proposed keypoint detector. Furthermore, we used the proposed detector in combination with the SHOT descriptor [20] for object instance recognition as presented in [21]. We have experimentally observed that the proposed keypoint detector when paired with the SHOT descriptor for object recognition tasks shows promising performance.

References

[1] S. Salti, F. Tombari, and L. D. Stefano, "A performance evaluation of 3d keypoint detectors," in *IEEE International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, Hangzhou, China, pp. 236–243, May 2011.

[2] F. Tombari, S. Salti, and L. D. Stefano, "Performance evaluation of 3d keypoint detectors," *International Journal of Computer Vision*, vol. 102, no. 1-3, pp. 198–220, March 2013.

[3] V. K. Ghorpade, P. Checchin, L. Malaterre, and L. Trassoudaine, "Performance evaluation of 3d keypoint detectors for time-of-flight depth data," in *14th International Conference on Control, Automation, Robotics and Vision*, Phuket, Thailand, pp. 1–6, November 2016.

[4] H. Chen and B. Bhanu, "3d free-form object recognition in range images using local surface patches," *Pattern Recognition Letters*, vol. 28, no. 10, pp. 1252–1262, July 2007.

[5] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *Computer Graphics Forum*, vol. 28, pp. 1383–1392, November 2009.

[6] Y. Zhong, "Intrinsic shape signatures: A shape descriptor for 3d object recognition," in *IEEE 12th International Conference on Computer Vision Workshops*, Kyoto, Japan, pp. 689–696, September 2009.

[7] A. Mian, M. Bennamoun, and R. Owens, "On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes," *International Journal of Computer Vision*, vol. 89, no. 2, pp. 348–361, September 2010.

[8] Ivan Sipiran and Benjamin Bustos, "Harris 3D: a robust extension of the harris operator for interest point detection on 3d meshes," *The Visual Computer*, vol. 27, no. 11, pp. 963–976, July 2011.

[9] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *IEEE International Conference on Robotics and Automation*, Shanghai, China, pp. 1–4, May 2011.

[10] S. M. Prakhya, B. Liu, and W. Lin, "Detecting keypoint sets on 3d point clouds via histogram of normal orientations," *Pattern Recognition Letters*, vol. 83, no. 1, pp. 42–48, November 2016.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[12] R. Unnikrishnan and M. Hebert, "Multi-scale interest regions from unorganized point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Alaska, USA, pp. 1–8, June 2008.

[13] U. Castellani, M. Cristani, S. Fantoni, and V. Murino, "Sparse points matching by combining 3d mesh saliency with statistical descriptors," *Computer Graphics Forum*, vol. 27, no. 2, pp. 643–652, April 2008.

[14] A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud, "Surface feature detection and description with applications to mesh matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, Florida, USA, pp. 373–380, June 2009.

[15] S. Salti, F. Tombari, R. Spezialetti, and L. D. Stefano, "Learning a descriptor-specific 3d keypoint detector," in *IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 2318–2326, December 2015.

[16] A. Tonioni, S. Salti, F. Tombari, R. Spezialetti, and L. D. Stefano, "Learning to detect good 3d keypoints," *International Journal of Computer Vision*, vol. 126, no. 1, pp. 1–20, January 2018.

[17] Georgios Georgakis, Srikrishna Karanam, Ziyan Wu, Jan Ernst, and Jana Košecká, "End-to-end learning of keypoint detector and descriptor for pose invariant 3d matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1965–1973, June 2018.

[18] R. B. Rusu, *Semantic 3D object maps for everyday manipulation in human living environments*, Ph.D. thesis, Department of Electrical and Computer Engineering, Technical University of Munich, Munich, Germany, 2009.

[19] F. Tombari and S. Salti, "3d keypoint detection benchmark," Availalbe: `https://vision.deis.unibo.it/keypoints3d/`.

[20] F. Tombari, S. Salti, and L. D. Stefano, "Unique signatures of histograms for local surface description," in *11th European Conference on Computer Vision*, Heraklion, Crete, Greece, pp. 356–369, September 2010.

[21] F. Tombari and L. D. Stefano, "Object recognition in 3d scenes with occlusions and clutter by hough voting," in *Fourth Pacific-Rim Symposium on Image and Video Technology*, Singapore, pp. 349–355, November 2010.