A Biological Motivated Multi-Scale Keypoint Detector for local 3D Descriptors

Sílvio Filipe and Luís A. Alexandre

Department of Computer Science
IT - Instituto de Telecomunicações
University of Beira Interior, 6200-001 Covilhã, Portugal.
{sfilipe,lfbaa}@ubi.pt

Abstract. Most object recognition algorithms use a large number of descriptors extracted in a dense grid, so they have a very high computational cost, preventing real-time processing. The use of keypoint detectors allows the reduction of the processing time and the amount of redundancy in the data. Local descriptors extracted from images have been extensively reported in the computer vision literature. In this paper, we present a keypoint detector inspired by the behavior of the early visual system. Our method is a color extension of the BIMP keypoint detector, where we include both color and intensity channels of an image. The color information is included in a biological plausible way and reproduces the color information in the retina. Multi-scale image features are combined into a single keypoints map. Our detector is compared against state-of-the-art detectors and is particularly well-suited for tasks such as category and object recognition. The evaluation allowed us to obtain the best pair keypoint detector/descriptor on a RGB-D object dataset. Using our keypoint detector and the SHOTCOLOR descriptor we obtain a good category recognition rate and for object recognition it is with the PFHRGB descriptor that we obtain the best results.

1 Introduction

Keypoint detection has been an area which has attracted a lot of attention in the computer vision community, developing a series of methods which are stable under a wide range of transformations [1]. Some of them developed based on general features [2], specific [3,4,1] or a mixture of them [5]. Given the number of keypoint detectors, it is surprising that many of the best recognition systems do not use these detectors. Instead, they process the entire image, either by preprocessing entire image to obtain feature vectors [6], by sampling descriptors on a dense grid [7] or by processing entire images hierarchically and detecting salient features in the process [8] These approaches provide a lot of data which helps classification, but also introduce a lot of redundancy [9] or high computacional cost [7]. Typically, the largest computational cost of these systems is in the stage of computing the features (called descriptors in 3D), so, it makes sense to use only a non redundant subset of points from the input image or point cloud.

In this paper, we present a new 2D keypoint detector. Our method is a biologically motivated multi-scale keypoint detector, which uses color and intensity channels of an image. As the basis of our method we use the Biologically Inspired keyPoints (BIMP) [1], which is a fast keypoint detector based on the biology of the human visual cortex. We present an extension of this method by introducing the color analysis, similar to what is done in the human retina. A comparative evaluation is made on a large public RGB-D Object Dataset [10], which is composed by 300 real objects from 51 categories. The evaluation of our and the state-of-art keypoint detectors is based on category and object recognition using 3D descriptors. This dataset contains the location of each point in the 2D space, which allows us to use 2D keypoint detector methods on the point clouds.

In [11], the author makes a comparison of the descriptors available the Point Cloud Library (PCL) [12] and explains how they work. The same author studies the accuracy of the distances both for objects and category recognition and finds that simple distances give competitive results [13]. Our work will compare the keypoint detectors using 3D descriptors that present better results in [11] and we will do it using the best distance measure with the best accuracy presented in [13].

The interest on using depth information in computer vision applications has been growing recently due to the decreasing prices of 3D cameras, such as Kinect and Xtion. With this type of cameras we can make a 2D and 3D analysis of the captured objects. The cameras can return directly the 2D image and the corresponding cloud point, which is composed by the RGB and depth information. Depth information improves object perception, as it allows for the determination of its shape or geometry.

The structure of the paper is as follows: the next section presents a description of the methods that we evaluate; in section 3, we will describe our keypoint detector; the section 4 discusses the recognition pipeline used in this paper; the last two sections will discuss the results obtained and present the conclusions.

2 Evaluated Methods

In this section, we will describe the 2D keypoint detectors and 3D descriptors evaluated in this work. The keypoints are used in this work to reduce the overall computational time since instead of finding descriptors on a dense grid, we are extracting them only on a smaller set of locations that correspond to the detected keypoints. In [14], a complete list of 2D keypoint detectors is presented and in [15] the focus is on biologically plausible keypoint detectors. The descriptors are extracted based on a given point location and we use keypoint detectors to provide this information to the descriptors. As mentioned earlier, it is possible to perform at the same time an analysis in 2D and 3D space due to the type of camera used.

2.1 2D Keypoint Detectors

The Scale Invariant Feature Transform (SIFT) keypoint detector was proposed by [3]. This method shares similar properties with neurons in inferior temporal cortex that are used for object recognition in primate vision. The image I is convolved with a number of Gaussian filters whose standard deviations differ by a fixed scale factor. That is, $\sigma_{j+1} = k\sigma_j$ where k is a constant scalar that should be set to $\sqrt{2}$. The convolutions yield smoothed images, denoted by $G(x, y, \sigma_j), i = 1, \ldots, n$. The adjacent smoothed images are then subtracted by $D(x, y, \sigma_j) = G(x, y, \sigma_{j+1}) - G(x, y, \sigma_j)$. These two steps are repeated, yielding a number of Difference-of-Gaussians (DoGs) over the scale space. Once DoGs have been obtained, keypoints are identified as local minima/maxima of the DoGs across scales. This is done by comparing each point in the DoGs to its eight neighbors at the same scale and nine corresponding neighborhood points in each of the neighborhood scales. The dominant orientations are assigned to localized keypoints.

Speeded-Up Robust Features (SURF) [4] is partly inspired by the SIFT descriptor. SURF is based on sums of 2D Haar wavelet responses and makes an efficient use of integral images. It uses an integer approximation to the determinant of Hessian blob detector, which can be computed extremely quickly with an integral image. For features, it uses the sum of the Haar wavelet response around the point of interest.

Biologically Inspired keyPoints (BIMP) [1] is a cortical keypoint detector for extracting meaningful points from images, solving the computational problem of [15]. The keypoints are extracted by a series of filtering operations: simple cells, complex cells, end-stopped cells and inhibition cells. Simple cells are modeled using complex Gabor filters with phases in quadrature are given by:

$$g_{\lambda,\sigma,\theta,\phi}(x, y) = \exp\left(-\frac{\tilde{x}^2 + \gamma\tilde{y}^2}{2\sigma^2}\right)\exp\left(i\frac{2\pi\tilde{x}}{\lambda}\right), \tag{1}$$

where $\tilde{x} = x\cos(\theta) + y\sin(\theta)$, $\tilde{y} = y\cos(\theta) - x\sin(\theta)$, with σ the receptive field size, θ the filter orientation, λ is the wavelength and $\gamma = 0.5$. Simple cell responses are obtained by convolving the image with the complex Gabor filter: $R_{\lambda,\theta} = I * g_{\lambda,\theta}$. Complex cells are the modulus of simple cell responses $C_{\lambda,\theta} = |R_{\lambda,\theta}|$. Remaining kernels are sums of Dirac functions (δ). If $ds = 0.6\lambda\sin(\theta)$ and $dc = 0.6\lambda\cos(\theta)$, double-stopped cell kernels are defined by

$$k^D_{\lambda,\theta} = \delta(x, y) - \frac{\delta(x-2ds, y+2dc) + \delta(x+2ds, y-2dc)}{2} \tag{2}$$

and the final keypoints is given by

$$K^D_\lambda = \sum_{\theta=0}^{\pi} |C_{\lambda,\theta}k^D_{\lambda,\theta}|^+ - \sum_{\theta=0}^{2\pi} |C_{\lambda,\theta}k^{TI}_{\lambda,\theta} + C_{\lambda,\theta^\perp}k^{RI}_{\lambda,\theta} - C_{\lambda,\theta}|^+, \tag{3}$$

where $\theta^\perp$ is orthogonal to θ, $|.|^+$ represents the suppression of negatives values. $k^{TI}_{\lambda,\theta}$ is the tangential inhibition kernel and $k^{RI}_{\lambda,\theta}$ the radial.

$$k^{TI}_{\lambda,\theta} = -2\delta(x, y) + \delta(x+dc, y+ds) + \delta(x-dc, y-ds) \tag{4}$$

$$k^{RI}_{\lambda,\theta} = \delta(x + dc/2, y + ds/2) + \delta(x - dc/2, y - ds/2). \tag{5}$$

2.2 3D Descriptors

Point Feature Histograms (PFH) [16] can be categorized as geometry-based descriptor [17]. This descriptor is represented by the surface normals, curvature estimates and distances between point pair which are generated by a point and its local neighborhood. They are represented by three angles and they are stored into an histogram that encodes the correlations between the normal angles of point pairs in a neighborhood. PFHRGB variant includes three more histograms, one for the ratio between each color channel.

The Signature of Histograms of OrienTations (SHOT) descriptor [18] is based on a signature histograms representing topological features, that make it invariant to translation and rotation. For a given keypoint, it computes a repeatable local reference frame using the eigenvalue decomposition around it, in order to incorporate geometric information of point locations in a spherical grid. A histogram is constructed by summing point counts of the angle between the normal of the keypoint and the normal of each point belonging to the spherical grid. In [19], a color version (SHOTCOLOR) is proposed, where the CIELab color space is used as color information.

3 OUR Method

Our Biological Motivated Multi-Scale Keypoint Detector (BMMSKD) is a color information extension of BIMP. The way we add the color information is based on an neural architecture of the primate visual system [20,21]. Figure 1 presents the block diagram of our keypoint detector.

For a given color image, we create three images from the RGB channels, which are: RG, BY and grayscale image I (shown in the left column of the figure 2). The r, g, and b channels are normalized by I in order to decouple hue from intensity. However, because hue variations are not perceivable at very low luminance (and hence are not salient), normalization is only applied at the locations where I is larger than 1/10 of its maximum over the entire image (other locations yield zero r, g, and b). Four broadly-tuned color channels are created: R for red channel, G for green, B for blue and Y for yellow: $R = r - (g + b)/2$, $G = g-(r+b)/2$, $B = b-(r+g)/2$ and $Y = (r+g)/2-|r-g|/(2-b)$. Accordingly, maps $RG = R-G$ are created in the model to represent the red/green opponency and $BY = B - Y$ for blue/yellow opponency (negative values are set to zero).

For each color channels RG, BY and I, we apply the BIMP keypoint detector and fuse the keypoint locations. Given the application of BIMP method on each channel, we obtain three sets of keypoints k_{RG}, k_{BY} and k_I, respectively (shown in the right column of the second to fouth rows of figure 2). With these three sets, we create a keypoint map K_m given by $K_m = k_{RG} \cup k_{BY} \cup k_I$. A location is considered a keypoint, if there exists another color channel, in its neighborhood, which indicates that there exists one keypoint in the region. This is: $k_l \in K_m : \#K_m^r(k_l) > 1$, where k_l is a keypoint location, r the neighborhood radius and $K_m^r(k_l)$ is a sub-set of K_m centered in the point k_l and

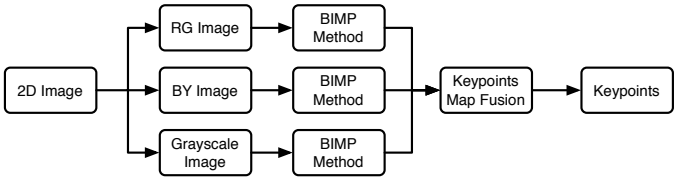

Fig. 1. Block diagram of our method. Our method receives an image directly from the camera and generates the three new images (RG, BY and I). In each of these images we apply the BIMP keypoint detector and fuse the result of the three detections. See the text for details.

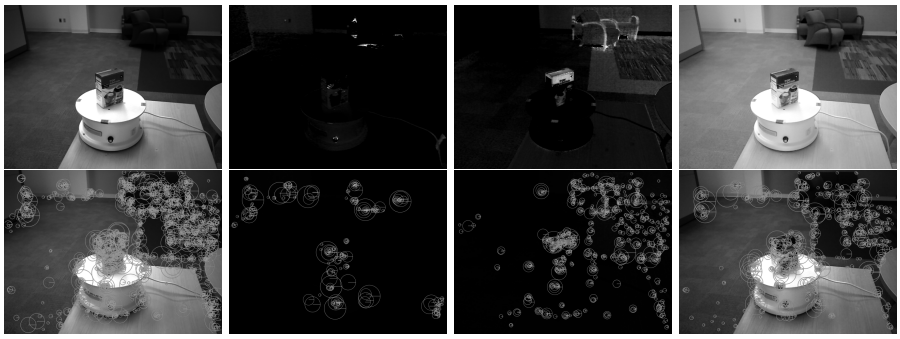

Fig. 2. Our keypoint detection method. In the first column, we have the original image on the top and the keypoint fusion on the bottom. The second, third and fourth columns contain the RG, BY and gray color channels (top) and the respective keypoint detection on the bottom.

with radius r. An example of the fusion result is presented in the bottom of the first column in figure 2.

4 Object Recognition Pipeline

In this section, we present the pipeline used in this work, shown in figure 3. Each block will be explained in the following subsections.

4.1 Segmented Objects

The input camera and segmentation process is simulated by the large RGB-D Object Dataset [1] [10]. This dataset was collected using an RGB-D camera and contains a total of 207621 segmented objects and these were saved through both 3D point clouds and 2D images. The dataset contains 300 physically distinct objects taken on a turntable and the objects are organized into 51 categories. The chosen objects are commonly found in home and office environments, where personal robots are expected to operate. In this work, we use 5 images/point clouds of each physically distinct object, using a total of 1500 from each of them.

[1] The dataset is publicly available at http://www.cs.washington.edu/rgbd-dataset.

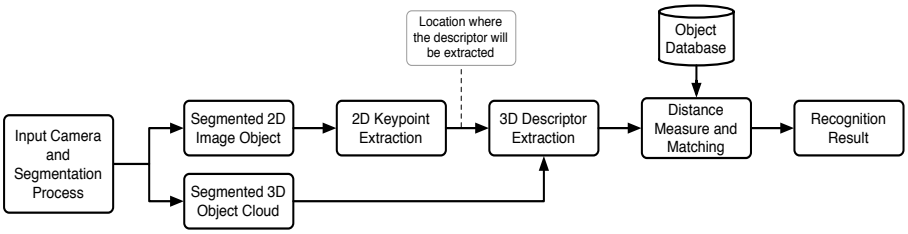

Fig. 3. Block diagram of the 3D recognition pipeline. See the text in section 4 for more details.

Table 1. Keypoints statistics. The number of points, time in seconds (s) and size in kilobytes (KB) presented are related to each cloud in the processing of the test set.

Keypoints	N. of Points	Time (s)	Size (KB)
BMMSKD	142.03±141.00	10.65±1.61	6.55±6.22
BIMP [1]	56.05±53.07	3.93±0.90	2.69±2.35
SIFT [3]	46.83±63.02	0.26±0.07	2.27±2.78
SURF [4]	47.77±60.06	0.28±0.07	2.32±2.67
Average	73.17±85.15	3.78±3.62	3.45±3.77
Original	5740.06±6851.42		316.86±375.73

4.2 2D Keypoint Detectors

The 2D image of the segmented object, present in the database, will feed the keypoint extraction process, which is used to reduce the computational cost of the recognition system. The keypoints implementation used in this work is done in Open Source Computer Vision (OpenCV) library [22]. In table 1, we present the average number of keypoints, mean computation time (in seconds) spent by each method to extract the keypoints and the file size (in KiloBytes). These times were obtained on a computer with *Intel®Core^{TM}i7-980X Extreme Edition 3.33GHz* with 24 *GB* of RAM memory.

4.3 3D Descriptors

The descriptors are extracted at the locations given by the keypoint detector obtained from the 2D images, but the processing of descriptors is done in point clouds. The point clouds have the 3D information of the segmented object, which is composed by: color (in the RGB color space) and depth information. The 3D descriptors that present better results in [11] are: PFH, SHOT and their color based versions PFHRGB and SHOTCOLOR. In table 2, we present some statistics about the extracted descriptors using the presented keypoint detectors (like in table 1).

Table 2. Descriptors statistics (for more details see caption of table 1).

Descriptor	Time (s)	Size (KB)
PFH	1.01±1.76	24.00±31.99
PFHRGB	1.97±4.01	53.35±77.49
SHOT	0.05±0.06	90.67±121.62
SHOTCOLOR	0.05±0.06	277.54±372.96
Average	0.79±2.36	111.39±223.57

4.4 Object Database

Using the 1500 point clouds selected, the observations is given by the Leave-One-Out Cross-Validation (LOOCV) method [23]. As the name suggests, LOOCV involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. This is the same as a K-fold cross-validation with K being equal to the number of observations in the original sampling. With 1500 point clouds and LOOCV method, we perform more than 1200000 comparisons for each pair keypoint detector/descriptor and we have a total of 16 pairs (4 keypoint detectors $\times$ 4 descriptors).

4.5 Distance Measure and Matching

One of the stages in 3D object recognition is the correspondence between a input cloud and a known object cloud (stored in the database). The correspondence is typically done using a distance function between the sets of descriptors. In [13], multiple distance functions are studied. In this work, we will use the distance

$$D_6 = L_1(c_A, c_B) + L_1(std_A, std_B) \tag{6}$$

that presents good results, in terms of recognition and run time, where c_A and c_B are the centroids of the sets A and B, respectively, and

$$std_A(i) = \sqrt{\frac{1}{|A|-1}\sum_{j=1}^{|A|}(a_j(i) - c_A(i))^2}, i = 1, \ldots, n, \tag{7}$$

$a_j(i)$ refers to the coordinate i of the descriptor j, and likewise for std_B. The L_1 distance is between descriptor (not sets) $x, y \in X$ and is given by

$$L_1(x, y) = \sum_{i=1}^{n} |x(i) - y(i)|. \tag{8}$$

5 Results and Discussion

In order to perform the evaluation, we will use three measures, which are the Receiver Operator Characteristic (ROC) Curve, the Area Under the ROC Curve

Table 3. AUC and DEC values for the category recognition for each pair keypoints/descriptor. The underline value is the best result for this descriptor and the best pair is the bold one.

Keypoints	Category Recognition							
	PFH		PFHRGB		SHOT		SHOTCOLOR	
	AUC	DEC	AUC	DEC	AUC	DEC	AUC	DEC
BMMSKD	0.743	0.850	**0.772**	**1.001**	0.630	0.369	0.701	0.615
BIMP [1]	0.716	0.743	0.758	0.929	0.615	0.360	0.673	0.562
SIFT [3]	0.739	0.870	0.764	0.980	0.610	0.326	0.683	0.589
SURF [4]	0.737	0.854	0.764	0.978	0.606	0.304	0.678	0.571

Table 4. AUC and DEC values for the object recognition for each pair keypoints/descriptor. The underline value is the best result for this descriptor and the best pair is the bold one.

Keypoints	Object Recognition							
	PFH		PFHRGB		SHOT		SHOTCOLOR	
	AUC	DEC	AUC	DEC	AUC	DEC	AUC	DEC
BMMSKD	0.796	1.076	**0.920**	**1.926**	0.666	0.473	0.791	0.931
BIMP [1]	0.766	0.941	0.904	1.781	0.627	0.393	0.738	0.790
SIFT [3]	0.789	1.093	0.892	1.724	0.620	0.371	0.729	0.760
SURF [4]	0.799	1.120	0.894	1.749	0.615	0.337	0.721	0.723

(AUC) and the decidability (DEC). The decidability index

$$DEC = |\mu_{intra} - \mu_{inter}| / \sqrt{\frac{1}{2}(\sigma^2_{intra} + \sigma^2_{inter})} \tag{9}$$

is the distance between the distributions obtained for the two classical types of comparisons: between descriptors extracted from the same (*intra-class*) and different objects (*inter-class*). The μ_{intra} and μ_{inter} denote the means of the intra- and inter-class comparisons, σ^2_{intra} and σ^2_{inter} the respective standard deviations and the decidability can vary between $[0, \infty[$. The obtained AUC and DEC for category and object recognition are given in tables 3 and 4.

As shown in table 3 and 4, our method increases the recognition results in both category and object recognition. Comparing our method with the original approach, we can see that color information has introduced a significant improvement in both category and object recognition. Only in the PFH descriptor is where the SURF detector presents slightly better results, but for the AUC and DEC measures this improvement is highlighted only in object recognition. For this descriptor in terms of category recognition, SIFT reached better results than DEC. The fact that we are using color information and the color information fusion justifies these improvements. For grayscale images, the results will be the same as the BIMP, since our method generalizes it to color.

6 Conclusions

In this paper we focused on keypoint detectors and we present a novel keypoint detector biologically motivated by the behavior and the neuronal architecture of the early primate visual system. We made a recognition evaluation of the proposed approach on public available data with real 3D objects. For this evaluation, we used the keypoint detectors in the OpenCV library and we projected the keypoint locations to the 3D space to use available 3D descriptors on the PCL library.

The main conclusions of this paper are: 1) the keypoint locations can help or degrade the recognition process; 2) a descriptor that uses color information should be used instead of a similar one that uses only shape information; 3) since there are big differences in terms of recognition performance, size and time requirements, the descriptor should be matched to the desired task; 4) to recognize the category of an object or a real-time system, we recommend the use the SHOTCOLOR method because it presents a recognition rate of 7% below of the PFHRGB but with a much lower computational cost; and 5) to do the object recognition, we recommend PFHRGB because it presents a recognition rate 12.9% higher than SHOTCOLOR.

Acknowledgments

This work is supported by '*FCT - Fundação para a Ciência e Tecnologia*' (Portugal) through the research grant 'SFRH/BD/72575/2010', and the funding from 'FEDER - QREN - Type 4.1 - *Formação Avançada*', co-founded by the European Social Fund and by national funds through Portuguese 'MEC - *Ministério da Educação e Ciência*'. It is also supported by the *IT - Instituto de Telecomunicações* through 'PEst-OE/EEI/LA0008/2013'.

References

1. Terzić, K., Rodrigues, J., du Buf, J.: Fast Cortical Keypoints for Real-Time Object Recognition. In: IEEE International Conference on Image Processing, Melbourne (2013) 3372–3376
2. Shi, J., Tomasi, C.: Good features to track. In: IEEE Conference on Computer Vision and Pattern Recognition. (1994) 593–600
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
4. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding **110** (2008) 346–359
5. Forstner, W., Dickscheid, T., Schindler, F.: Detecting interpretable and accurate scale-invariant keypoints. In: IEEE 12th International Conference on Computer Vision. (2009) 2256–2263
6. Pinto, N., Cox, D.D., DiCarlo, J.J.: Why is real-world visual object recognition hard? PLoS Computational Biology **4** (2008) 151–156

7. Boiman, O., Shechtman, E., Irani, M.: In defense of Nearest-Neighbor based image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK (2008) 1–8
8. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Robust object recognition with cortex-like mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 411–426
9. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision **73** (2007) 213–238
10. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view RGB-D object dataset. In: International Conference on Robotics and Automation. (2011) 1817–1824
11. Alexandre, L.A.: 3D descriptors for object and category recognition: a comparative evaluation. In: Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vilamoura, Portugal (2012)
12. Rusu, R.B., Cousins, S.: 3D is here: Point Cloud Library (PCL). In: International Conference on Robotics and Automation, Shanghai, China (2011)
13. Alexandre, L.A.: Set Distance Functions for 3D Object Recognition. In: 18th Iberoamerican Congress on Pattern Recognition, Springer (2013) 57–64
14. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: A survey. Found. Trends. Comput. Graph. Vis. **3** (2008) 177–280
15. Rodrigues, J., du Buf, J.: Multi-scale keypoints in V1 and beyond: object segregation, scale selection, saliency maps and face detection. Biosystems **86** (2006) 75–90
16. Rusu, R.B., Blodow, N., Marton, Z.C., Beetz, M.: Aligning point cloud views using persistent feature histograms. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, Nice, France (2008) 3384–3391
17. Aldoma, A., Marton, Z., Tombari, F., Wohlkinger, W., Potthast, C., Zeisl, B., Rusu, R., Gedikli, S., Vincze, M.: Tutorial: Point Cloud Library: Three-Dimensional Object Recognition and 6 DOF Pose Estimation. IEEE Robotics & Automation Magazine **19** (2012) 80–91
18. Tombari, F., Salti, S., Di Stefano, L.: Unique Signatures of Histograms for Local Surface Description. In: 11th European Conference on Computer Vision, Crete, Greece (2010) 356–369
19. Tombari, F., Salti, S., Di Stefano, L.: A combined texture-shape descriptor for enhanced 3D feature matching. In: 18th IEEE International Conference on Image Processing, Brussels (2011) 809–812
20. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1254–1259
21. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research **40** (2000) 1489–1506
22. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
23. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: 14th International Joint Conference on Artificial Intelligence. Volume 2., San Francisco, CA, USA (1995) 1137–1143