A 3D Keypoint Detector based on Biologically Motivated Bottom-Up Saliency Map

Sílvio Filipe
http://socia-lab.di.ubi.pt/~silvio

Luís A. Alexandre
http://di.ubi.pt/~lfbaa

IT - Instituto de Telecomunicações
Department of Computer Science
University of Beira Interior
6200-001 Covilhã, Portugal

Abstract

We present a new method for the detection of 3D keypoints on point clouds and we perform benchmarking between each pair of 3D keypoint detector and 3D descriptor to evaluate their performance on object and category recognition. Our keypoint detector is inspired by the behavior and neural architecture of the primate visual system. The 3D keypoints are extracted based on a bottom-up 3D saliency map, that is, a map that encodes the saliency of objects in the visual environment. The saliency map is determined by computing conspicuity maps of the orientation, intensity and color information in a bottom-up and in a purely stimulus-driven manner. Finally, the focus of attention (or “keypoint location") is sequentially directed to the most salient points in this map. The main conclusions are: with a similar average number of keypoints, our 3D keypoint detector outperforms the other 3D keypoint detectors evaluated in the category and object recognition experiments.

1 Introduction

The interest on using depth information on computer vision applications has been growing recently due to the decreasing prices of 3D cameras. Depth information improves object perception, as it allows for the determination of its shape or geometry.

This paper has two main focuses: the first is to present a new keypoint detector; the second an evaluation of our and others 3D keypoint detectors. Our keypoint detector is a saliency model based on spatial attention derived from the biologically plausible architecture. It uses three feature channels: color, intensity and orientation. The computational algorithm of this saliency model has been presented in [5] and the standard saliency benchmark in 2D images. We present the 3D version of this saliency detector and demonstrate how keypoints can be extracted from a saliency map.

In [2], the author focuses on the descriptors available in Point Cloud Library (PCL), explaining how they work and made a comparative evaluation on publicly available data. In our study, we will see that it is not enough, the results also depend on the keypoint location. The same author studies the accuracy of the distances both for objects and category recognition and finds that simple distances give competitive results [1].

The 3D keypoint detectors and descriptors that we will compare can be found in version 1.7 of the PCL [8]. With this, we will find what is the best pair of keypoint detector/descriptor for 3D point cloud objects. We propose to answer this question using a public large RGB-D Object Dataset [6], this is composed of 300 real objects and divided in 51 categories.

In [3], the invariance of 3D keypoint detectors according to rotations, scale changes and translations was evaluated. It also contains a more detailed description of the two keypoint detectors: 1) The Scale Invariant Feature Transform (SIFT) keypoint detector was proposed by [7]. In [4], the original algorithm for 3D data is presented, which uses a 3D version of the Hessian to select the interest points; 2) Intrinsic Shape Signatures 3D (ISS3D) [12] is a method relying on region-wise quality measurements. This method uses the magnitude of the smallest eigenvalue (to include only points with large variations along each principal direction) and the ratio between two successive eigenvalues (to exclude points having similar spread along principal directions). We compare our proposal against these ones.

One of our goals was to evaluate the four descriptors, two main descriptors and its variants based on color, in terms of category and object recognition: 1) Descriptors such as Point Feature Histograms (PFH) [9] can be categorized as geometry-based descriptors. This type of descriptor is represented by the surface normals, curvature estimates and distances, between point pairs. PFHRGB is an version of PFH in which is included information regarding the color of the object; 2) The Signature of Histograms of OrienTations (SHOT) descriptor [10] 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. For each spherical grid bin, a a one-dimensional histogram is obtained. In [11], they propose a color version (SHOTCOLOR), where use the CIELab color space as color information.

2 Proposed 3D Keypoint Detector

The Biologically Inspired 3D Keypoint based on Bottom-Up Saliency (BIK-BUS) is a keypoint detector that is based on the saliency maps, which are also known as visual attention. The saliency maps are determined by computing conspicuity maps of the features intensity and orientation in a bottom-up and data-driven manner. These conspicuity maps are fused into a saliency map and, finally, the focus of attention is sequentially directed to the most salient points in this map. Using this theory and following the work presented in [5], we will present our keypoint detector in six steps.

Step 1: Linear Filtering – The color channels (r, g, and b) of the input colored point cloud are normalized when $I = (r+g+b)/3$ is larger than $1/10$ of its maximum over the entire image, other locations yield zero. With these three normalized color channels, we create four broadly-tuned color channels: $R = r - (g+b)/2$, $G = g - (r+b)/2$, $B = b - (r+g)/2$ and $Y = (r+g)/2 - |r-g|/(2-b)$.

Five Gaussian pyramids $R(\sigma)$, $G(\sigma)$, $B(\sigma)$, $Y(\sigma)$ and $I(\sigma)$ are created from the color and intensity channels, where σ represents the standard deviation used in the Gaussian kernel. Each Gaussian pyramid is achieved by convolving the cloud with Gaussian kernels of increasing radius, resulting in a pyramid of clouds.

The orientation pyramids $O(\sigma, \theta)$ are obtained using the normals extracted from the intensity cloud I, where $\theta \in \{0^o, 45^o, 90^o, 135^o\}$ is the preferred orientation. In the primary visual cortex, the impulse response of orientation-selective neurons is approximated by Gabor filters.

Step 2: Center-Surround Differences – There are two types of center-surround structures in the retina: *on-center* and *off-center*. The *on-center* use a positive weighed center and negatively weighed neighbors and the *off-center* use exactly the opposite. The positive weighing is better known as excitatory and the negative as inhibitory.

Center-surround is implemented in the model as the difference between fine and coarse scales: the center is a pixel at scale $c \in \{2,3,4\}$, and the surround is the corresponding pixel at scale $s = c + \delta$, with $\delta \in \{3,4\}$. The across-scale difference between two maps, denoted by ‘$\ominus$’, is obtained by interpolation to the finer scale and point-by-point subtraction.

The first set of feature maps is concerned with intensity contrast. Here, both types of sensitivities are simultaneously computed in a set of six maps $I(c,s) = |I(c) \ominus I(s)|$. For the color channels, the process is similar, which, in the cortex, is called ‘color double-opponent’ system. The existence of a spatial and chromatic opponency between color channels in human primary visual cortex is described. With that, the maps $RG(c,s)$ and $BY(c,s)$ are created in the model to simultaneously take in account the red/green and green/red, and blue/yellow and yellow/blue double opponency, respectively, as: $RG(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|$ and $BY(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|$. Orientation feature maps, $O(c,s,\theta)$, encode local orientation contrast between the center and surround scales: $O(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)|$.

Step 3: Normalization – The salient objects appear only in a few maps, which can be masked by noise or by less-salient objects present in a larger number of maps. In the absence of top-down supervision, we use a map normalization operator $\mathcal{N}(.)$ and consists of: 1 – Normalizing the values in the map to a fixed range $[0..M]$, in order to eliminate large amplitude differences; 2 – Finding the location of the global maximum maps M and computing the average $\overline{m}$ of all its other local maxima; and 3

– Globally multiply the map by $(M-\overline{m})^2$.

Step 4: Across-Scale Combination – The conspicuity maps are the combination of the feature maps, for intensity, color, and orientation, at the scale s = 4 of the saliency map. They are obtained through the reduction of each map to scale four and point-by-point addition, called across-scale addition, '$\bigoplus$'. The conspicuity maps for the intensity, $\bar{I}$, and color channels, $\overline{C}$, are given by: $\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(I(c,s))$ and $\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(RG(c,s)) + \mathcal{N}(BY(c,s))]$.

For orientation, four intermediary maps are first created by combination of the six feature maps for a given θ and are then combined into a single orientation conspicuity map
$$\overline{O} = \sum_{\theta \in \{0^o, 45^o, 90^o, 135^o\}} \mathcal{N}\left[\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(O(c,s,\theta))\right].$$

Step 5: Linear Combination – The final saliency map is obtained by the normalization of each conspicuity map and calculating the mean between $\bar{I}$, $\overline{C}$ and $\overline{O}$.

Step 6: Inhibition-of-Return (IR) – The IR is part of the method that is responsible for the selection of keypoints. It detects the most salient location and directs attention towards it, considering that location a keypoint. After that, the IR mechanism transiently suppresses this location in the saliency map and its neighborhoods in a small radius, such that attention is autonomously directed to the next most salient image location. Computationally, the IR performs a similar process of selecting the global and local maximums.

3 Experimental Evaluation and Discussion

In order to perform this evaluation, we will use three measures, which are the Area Under the ROC Curve (AUC) and the decidability (DEC). The obtained AUC and DEC for category and object recognition are given in table 1. The decidability index represents 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*). Considering 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 is given by $DEC = \frac{|\mu_{intra} - \mu_{inter}|}{\sqrt{\frac{1}{2}(\sigma^2_{intra} + \sigma^2_{inter})}}$.

Analyzing the descriptors in a generic way, the best results were obtained with the PFHRGB. It is interesting to compare it to the PFH: improvement can only be attributed to the incorporation of color information. The same is true for the SHOTCOLOR versus the SHOT descriptor. The two best results in terms of category and object recognition are presented in the descriptors that use color information.

Considering only the accuracy, the best combination for the category recognition is BIK-BUS/PFHRGB, closely followed by BIK-BUS/ SHOTCOLOR, ISS3D/PFHRGB and ISS3D/SHOTCOLOR both in terms of AUC and DEC. The pairs BIK-BUS/PFHRGB and BIK-BUS/ SHOTCOLOR have exactly the same AUC, the difference is in the DEC where it is slightly higher in the case of PFHRGB. BIK-BUS turns out again the best performer among detectors: SHOT, SHOTCOLOR and PFHRGB.

In terms of object recognition, the best pair is BIK-BUS/PFHRGB, but only beats the second best combination, ISS3D/PFH. For SHOT and SHOTCOLOR descriptors, if we compare our keypoint detector with the ISS3D we obtain improvements for both of 1.5% in the case of category recognition, and 1.1% and 1.4% in object recognition, respectively.

4 Conclusions

In this paper we presented a novel 3D keypoint detector biologically motivated by the behavior and the neuronal architecture of the early primate visual system. The BIK-BUS is a keypoint detector on a computational technique to determine visual attention, which are also known as saliency maps. The saliency maps are determined by sets of features in a bottom-up and data-driven manner. The fusion of these sets produced the saliency map and the focus of attention is sequentially directed to the most salient points in this map, representing a keypoint location.

In the evaluation, we used some of the 3D keypoint detectors and the 3D descriptors available in the PCL library. The main conclusions of this paper are: 1) a descriptor that uses color information should be used instead of a similar one that uses only shape information; 2) in terms of keypoint detectors, to obtain an accurate recognition system, we recommend the use of the BIK-BUS, since its performance is by far the best among the keypoint detectors tested; 3) in terms of descriptors, if the focus is on accuracy we recommend the use of PFHRGB and for real-time a good choice is the SHOTCOLOR because it presents a good balance between recognition performance and time complexity.

Table 1: AUC and DEC values for the category and object recognition for each pair keypoint detector/descriptor. **BOLD** indicates the best (bigger) results in terms of AUC and DEC for each pair.

Descriptor	Keypoint	Category		Object	
		AUC	DEC	AUC	DEC
SHOT	BIK-BUS	**0.827**	**1.281**	**0.863**	**1.513**
	ISS3D	0.812	1.168	0.852	1.413
	SIFT3D	0.814	1.207	0.848	1.409
SHOTCOLOR	BIK-BUS	**0.867**	**1.571**	**0.916**	**2.012**
	ISS3D	0.852	1.465	0.902	1.873
	SIFT3D	0.839	1.394	0.896	1.792
PFH	BIK-BUS	**0.848**	1.488	0.893	1.832
	ISS3D	**0.848**	**1.489**	**0.895**	**1.855**
	SIFT3D	0.843	1.458	0.890	1.801
PFHRGB	BIK-BUS	**0.867**	**1.586**	**0.948**	**2.397**
	ISS3D	0.866	1.585	**0.948**	2.394
	SIFT3D	0.861	1.546	0.946	2.373

Acknowledgment

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*', subsidized by the European Social Fund and by Portuguese funds through 'MCTES'. We also acknowledge the support given by the *IT - Instituto de Telecomunicações* through 'PEst-OE/EEI/LA0008/2013'.

References

[1] L. A. Alexandre. Set Distance Functions for 3D Object Recognition. In *18th CIARP*, pages 57–64. Springer, 2013.

[2] Luís A. Alexandre. 3D descriptors for object and category recognition: a comparative evaluation. In *Workshop on Color-Depth Camera Fusion in Robotics at the IEEE/RSJ IROS*, Vilamoura, Portugal, October 2012.

[3] S. Filipe and L. A. Alexandre. A Comparative Evaluation of 3D Keypoint Detectors in a RGB-D Object Dataset. In *9th VISAPP*, Lisbon, Portugal, 5–8 January 2014.

[4] A. Flint, A. Dick, and A. Hengel. Thrift: Local 3D Structure Recognition. In *9th DICTA*, pages 182–188, December 2007.

[5] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, March 1998.

[6] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *ICRA*, pages 1817–1824, May 2011.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, November 2004.

[8] R. B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *ICRA*, Shanghai, China, May 9-13 2011.

[9] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz. Aligning point cloud views using persistent feature histograms. In *IEEE/RSJ IROS*, pages 3384–3391, Nice, France, September 2008.

[10] F. Tombari, S. Salti, and L. Di Stefano. Unique Signatures of Histograms for Local Surface Description. In *11th ECCV*, pages 356–369, Crete, Greece, 2010.

[11] F. Tombari, S. Salti, and L. Di Stefano. A combined texture-shape descriptor for enhanced 3D feature matching. In *18th IEEE ICIP*, pages 809–812, Brussels, September 2011.

[12] Y. Zhong. Intrinsic shape signatures: A shape descriptor for 3D object recognition. *ICCV*, pages 689–696, September 2009.