Keypoint detection in RGBD images based on an efficient viewpoint-covariant multiscale representation

Maxim Karpushin, Giuseppe Valenzise, Frédéric Dufaux

KEYPOINT DETECTION IN RGBD IMAGES BASED ON AN EFFICIENT VIEWPOINT-COVARIANT MULTISCALE REPRESENTATION

Maxim Karpushin, Giuseppe Valenzise, Frédéric Dufaux

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay

ABSTRACT

Texture+depth (RGBD) images provide information about the geometry of a scene, which could help improve current image matching performance, e.g., in presence of large viewpoint changes. While depth has been mainly used for processing keypoint descriptors, in this paper we focus on the keypoint detection problem. In order to produce a computationally efficient viewpoint-covariant multiscale representation, we design an image smoothing procedure which locally smooths a texture image based on the corresponding depth. This yields an approximated scale space, where we can find keypoints using a multiscale detector approach. Our experiments on both synthetic and real-world images show substantial gains with respect to 2D and other RGBD feature extraction approaches.

Index Terms— RGBD, texture+depth, scale space, keypoint detection, visual odometry

1. INTRODUCTION

Keypoint detection, as a part of the classic image matching problem, is one of the basic task in computer vision. A number of application-level problems can be efficiently reduced to image matching, such as target tracking, visual odometry, and simultaneous localization and mapping (SLAM). Among the several signal representations on which to extract interest points, texture+depth (RGBD) content has been recently attracting a good deal of interest, thanks to the increasing availability of low-cost depth sensors such as Microsoft Kinect. In RGBD, the geometry of the scene is captured and represented through a depth map, in addition to the conventional texture map representation of 2D images. This additional geometrical information can be used to improve the matching performance by providing more robust, repeatable and distinctive features. In our previous works [1, 2] we show that a possible way to use depth maps in matching consists in improving feature robustness with respect to viewpoint position changes. However, this was obtained at the price of a significant computational effort, and the performance were highly afftected by depth sensor noise.

On the other hand, few approaches are currently available in the state of the art that are able to detect repeatable keypoints efficiently, and which can benefit from both texture *and* depth information. In this work, we address this problem by proposing an efficient keypoint detector for RGBD images. It discovers repeatable keypoints in the texture image, but uses the depth map as additional source of information in order to improve the repeatability, especially under viewpoint position changes. To achieve this, we exploit the idea of *viewpoint-covariant multiscale representation* by injecting the geometrical information into the keypoint detection modality. To this end, we design an approximated scale space representation that performs spatially adaptive smoothing, where the intensity of the filtering is determined by depth. Our results show significant gains with respect to simple 2D keypoint detectors and alternative RGBD approaches, both for a mid-level feature evaluation and for a visual odometry scenario.

The rest of the paper is organized as follows. In Section 2 we discuss related background and state-of-the-art techniques for computing multiscale image representations and using them in feature detection. Section 3 describes the proposed keypoint detection approach in detail. In Section 4 we illustrate the performance of the proposed detector in terms of keypoint repeatability and in a simple visual odometry scenario. Finally, Section 5 concludes the paper.

2. BACKGROUND AND RELATED WORK

Multiscale image representations play an important role in vision, especially in feature detection [3]. Among these representations, the *difference of Gaussians* (DoG) detector, used, e.g., in SIFT [4], has been a successful application of the Gaussian scale space [5] for the detection of interest points. Similar pyramidal image representations based on box filter have lower computational complexity and are used in other local feature detectors, such as SURF [6] or BRISK [7].

Gaussian scale space is constructed through a progressive, unstructured smoothing process. A function representing Gaussian scale space is the solution of a partial differential equation problem, set up for the standard (isotropic) heat diffusion equation and having the input image as the initial data [8]. Making the diffusion process *anisotropic* allows to establish a smoothing that respects the internal structure of the image. An example of this process is the Malik-Perona diffusion, that also produces a scale space [9]. Such anisotropic filtering processes were used for keypoint detection [10, 11, 12].

For texture+depth images, the keypoint detection task and, more widely, local feature extraction have been intensively studied in recent years. Since texture+depth image may be regarded as a mesh with associated photometric information, keypoint detection and mesh matching techniques could be used to produce feature representations, for example [13, 14, 15]. However, transforming a texture+depth image into such a mesh has some drawbacks. First of all, the occlusions problem raises, which does not exist in the mesh domain. Second, a texture+depth image is parametrized in the camera coordinates, contrarily to a typical mesh specified in its own coordinate system. Therefore, any viewpoint position change corresponds to a mesh resampling that affects keypoint repeatability and increases the sensibility to acquisition noise.

Thus, an image-level RGBD feature extraction is of interest. Viewpoint Invariant Patches (VIP) [16] extend SIFT detection principle to texture+depth images by detecting dominant planes in the scene and synthesizing their frontal views. A major limitation of this approach emerges when dealing with images having a complex geometry, where such dominant planes are little or not present at all. Some approaches focus only on keypoints in depth maps, such as 2.5D SIFT [17], NARF [18] or SIPF [19], just to name a few.

As the texture is not exploited, this allows to eliminate any feature instability caused by illumination changes. However, in some real scenes with rich texture maps, depth maps alone (without texture) are not representative enough to provide a rich and distinctive feature representation for an image matching-based application. A number of techniques for texture+depth image matching focus only on the descriptor side, e.g. BRAND [20], PIN [21], CSHOT [22] or our previous work [1]. In these works, conventional keypoint detectors are used, applied to the texture image only. A possible reason is that the design of a distinctive descriptor is perhaps an easier and less constrained task than constructing a repeatable detector. This exposes a lack of keypoint detectors that are able to benefit from the geometrical information in order to provide more repeatable keypoints, which is the main contribution of this paper.

3. THE PROPOSED DETECTOR DESIGN

A scale invariant keypoint detector is generally composed of two blocks: i) a signal representation where to look for keypoints, typically with a scale space-like structure; and ii) one or more keypoint selection criteria. In the following we describe in detail these two parts that form our proposed approach.

3.1. Scale space approximation

A scale space is typically engendered by progressively applying an image smoothing operator to the input image. Keeping in mind the goal of improved keypoint repeatability under viewpoint position changes, in this paper we exploit the idea of transferring smoothing from the image plane (as in the Gaussian scale space, for example) onto the scene surface given by the depth map. The resulting smoothing process becomes *viewpoint covariant*, and the keypoints will then be searched on the scene surface as if they were attached to it.

In our previous work, we formalized such an approach as a diffusion process on a manifold, whose internal metric is given by the depth map, and the texture represents the initial data [2]. Although that approach proves to satisfy scale space requirements, it has the drawback of requiring a computationally expensive iterative simulation of the diffusion process. In this work, we propose a simpler and faster approximation of [2], which adapts the intensity of smoothing locally by using depth information.

The proposed scale space approximation is based on the following observation. A texture image corresponds to the projection of objects in the scene onto the camera plane, followed by a sampling at the pixel granularity. As this sampling is uniform on the camera plane, the scene surfaces are sampled non uniformly. We leverage this simple observation to construct an approximated scale space, by varying locally the amount of smoothing, i.e., we vary the smoothing quantity from pixel to pixel as a function of the distance given by the depth map, so that *the further a given pixel is, the less it is smoothed*. More precisely, assume that we can smooth the input image up to a given smoothing quantity $\sigma(x, y)$ at each pixel location (x, y). Then, let us assume that $\sigma(x, y)$ depends on the depth map $D(x, y)$ in the following way:

$$\sigma(x, y) = \frac{\hat{\sigma}}{D(x, y)}. \quad (1)$$

$\hat{\sigma}$ is a constant value (in x, y image variables) representing a scale on the surface, or *spatial scale*, whereas σ is the corresponding scale in the image plane or *projected scale*. Using the pinhole camera model, it is straightforward to show that the projection on the camera plane of an object of characteristic spatial size $\hat{\sigma}$ is of size $\frac{\hat{\sigma}}{D(x,y)}$ pixels independently on the observer position. Thus, the smoothing quantity $\sigma(x, y)$ injected into the image becomes related to the surface and varies accordingly when the camera moves.

By progressively smoothing the original texture image I using a set $\{\hat{\sigma}_k\}_k$ of increasing $\hat{\sigma}$ values, it is possible to build a multiscale representation $I_k = I(\hat{\sigma}_k)$ that demonstrates the described approximated viewpoint-covariant behavior. The choice of $\hat{\sigma}_k$ and the structure of I_k are discussed in Section 3.2.

To complete the construction of scale space, we need an appropriate smoothing filter. Existing multiscale representations used for feature detection generally employ either a Gaussian or a box filter. Since $\sigma(x, y)$ varies across the image, filtering an N pixels image with a $M \times M$ pixels Gaussian kernel $K_G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$ requires $O(NM^2)$ operations, which can be computationally expensive for large filters (coarsest scales). On the other hand, a box filter consists in convolving the image with a constant square kernel, which can be done with $O(N)$ operations using integral images [23]. However, the box filter is not rotationally invariant.

In this paper, we adopt a smoothing filter we presented recently in [24], which offers a more accurate smoothing under image rotations, at the same computational complexity of the box filter. Applying this filter to the texture image with σ controlled by the depth map according to Eq. 1, we are able to synthesize $I(\hat{\sigma}_k)$. A visual example is given in Fig. 1 (a) – (c). The desired viewpoint-covariant behavior of the proposed approach might be observed on larger scales.

3.2. Detection criteria

Once a scale space is defined by a proper smoothing operator, the remaining part of keypoint detection consists in three steps: candidates selection; candidates filtering; and accurate localization. In this work, we use the *multiscale detector* approach proposed in [27] to select candidates. That is, we simply look for spatial local extrema in images obtained in a DoG-like fashion, by subtracting consecutive levels of our proposed multiscale representation. Specifically, candidates selection is performed as follows:

(i) The detection begins by setting $\hat{\sigma} = \hat{\sigma}_0$, a parameter tuned manually as a function of the depth measurement unit, which mainly depends on the used depth sensor (e.g., Kinect or LIDAR).

(ii) Next, we construct the pyramidal image representation analogous to [3], conventionally used in numerous scale-invariant detectors, including SIFT. It is based on a combination of smoothing and subsampling steps: we compute $I(\hat{\sigma}_k)$ and $I(2\hat{\sigma}_k)$ as described in the previous section and then downsample the input image by a factor of two horizontally and vertically. Here, k is an integer representing the octave index (counted from zero), and $\hat{\sigma}_k = 2^k\hat{\sigma}_0$. The smoothing filter we use allows to avoid explicit downsampling of the texture image, however, the depth map is properly filtered and resampled to avoid aliasing.

(iii) Finally, we compute the differences $J_k = I(2\hat{\sigma}_k) - I(\hat{\sigma}_k)$, which are analogous to DoG. It is known that local extrema of DoG reveal visual details of different scales and are repeatable under various transformations [28]. Based on this, our technique consists in taking the local extrema of J_k that should reveal visual details of a given spatial scale in octave k.

An illustration is given in Fig. 1 (d), (e). Distinctive red and blue blobs in the example images contain local maxima and minima of J_k that are taken as initial candidates.

For the candidate filtering and the accurate localization we reuse the methodology of SIFT, which filters out the keypoints situated on straight edges of J_k and performs an iterative accurate local extrema

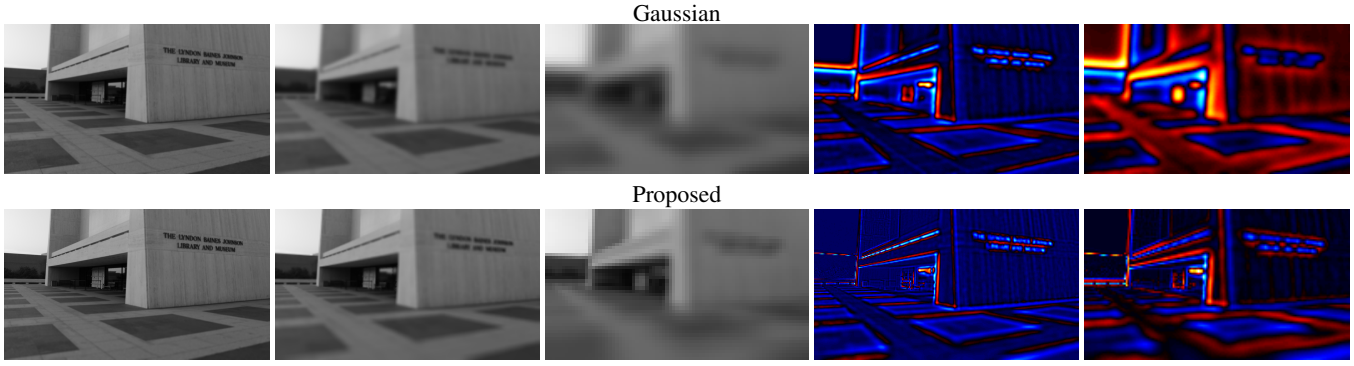

Fig. 1: Qualitative comparison of the Gaussian and the proposed multiscale representations for an RGBD image from the LIVE dataset [25, 26]. Top row: standard Gaussian scale space, where σ is constant within each image (no depth map used). Second row: proposed multiscale representation, where σ varies but $\hat{\sigma}$ remains constant. Images (a)–(c) in each row present different levels of smoothing. Images (d) and (e) obtained by subtracting adjacent smoothed images.

localization based on derivatives of J_k [4]. Keypoints detected on all octaves are put together and sent to the detector output. Each keypoint is thus characterized by its location on the image plane and its visual scale σ obtained according to Eq. 1 and interpolated properly after the accurate localization process.

4. EXPERIMENTS AND DISCUSSION

The proposed detector is evaluated in two scenarios: (1) a typical mid-level detector evaluation using repeatability score on synthetic RGBD sequences; and (2) an application-level evaluation on real RGBD images from KITTI dataset [29], where the depth maps are acquired with a LIDAR laser scanner.

4.1. Repeatability score

In this first scenario, keypoint detection is performed on a set of images (views) of a given scene. Keypoints extracted from the first (reference) view of the scene are then compared with keypoints obtained from other views. A keypoint from the reference view is *repeated* in another (test) view, if there is a keypoint detected in it that occupies approximatively the same area on the scene surface. The portion of repeated keypoints between two views is called *repeatability score* and is often used to evaluate the keypoint detection performance [30].

Following our previous works [1, 2], we control the keypoint area overlap by means of Jaccard index-based relative error η. Jaccard index is computed on spherical keypoint areas: its centers are obtained by projecting the keypoint positions onto the scene surface, and radii are approximated from the keypoint scale. For a given η value, the keypoint is counted as repeated if Jaccard index is greater than $1 - \eta$ (please refer to [1] for a detailed explanation). In a nutshell, the smaller the η, the more the keypoints that must be repeated and precisely localized in order to contribute to the repeatability score. We perform the experiment for two values of η: 0.5 and 0.25.

We use two artificial RGBD sequences both containing significant viewpoint position and scale changes: *Bricks* with 20 images and *House* with 25 images [31]. The proposed detector performance is compared to SIFT (VLFeat [32] implementation) and VIP (original implementation). The resulting repeatability scores as a function of the angle of view difference between test and reference views are shown in Fig. 2.

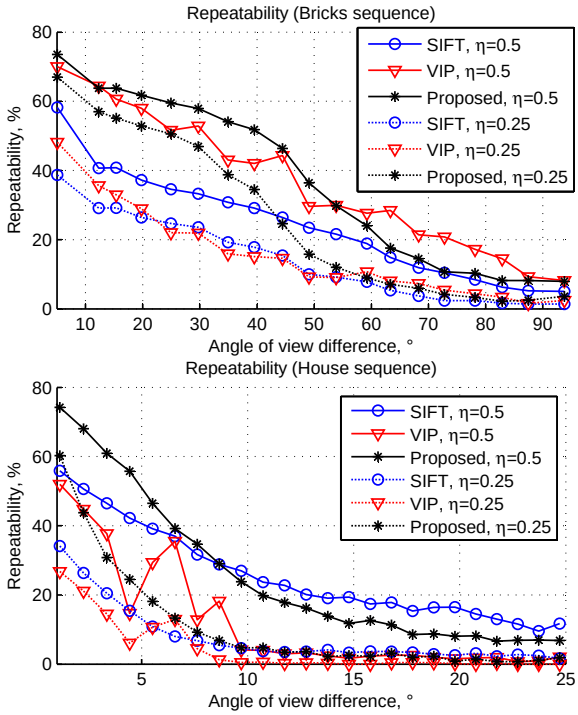

Fig. 2: Keypoint repeatability obtained with different detectors on two synthetic RGBD sequences.

The proposed detector demonstrates better overall repeatability except. A particularly higher accuracy is achieved on *Bricks*, as for tighter η our proposed detector demonstrates a significant gain up to 45° of rotation. At larger angles, the proposed scheme is outperformed by VIP on *Bricks* and SIFT on *House*, but the difference is at most 10 points. Furthermore, VIP detector fails on *House* sequence due to a more complex geometry, that may not be represented well by dominant planes. It is worth noticing that the number of detected keypoints by SIFT and the proposed detector are comparable and of order of 1000 to 2000, whereas VIP exhibits 2 or 3 times more keypoints. A visual example of repeated keypoints is given in Fig. 3.

4.2. Visual odometry

In the second part of our experiments, we evaluate the proposed approach in a visual odometry scenario. This is one of the typical applications of image matching, which consists in estimating the observer position and orientation solely based on data coming from visual sensors.

We match consecutive frames using different keypoints and descriptors, and then retrieve the pose change using Nister's calibrated fivepoint solver [33] (we used its implementation in OpenCV 3.0 beta). The obtained pose change is then cumulated within the current one, giving an estimation of the absolute pose with respect to the first frame. The result is then compared to the ground truth.

As the test data, we use ten sequences from KITTI dataset [29]. For each sequence we conduct the experiment on the first 300 frames, computing at each frame two types of error:

- translation error, simply equal to the absolute distance between estimated position and the real one from the ground truth;
- rotation error showing the inaccuracy in the estimated orientation; if $\mathbf{R}_{est}$ and $\mathbf{R}_{gt}$ are, respectively, the estimated and ground-truth 4×4 pose matrices in the coordinates of the first frame, the angular error is given by $\arccos\left(\frac{\mathrm{tr}(\mathbf{R}_{est}^{-1}\mathbf{R}_{gt})-1}{2}\right)$.

These errors are cumulative and strongly depend on the feature repeatability. Thus we present the error values reached on the last frame (final translation and rotation errors), as well as the error values averaged over all frames. In this test, we compare with the conventional SIFT descriptor and BRAND (Binary Robust Appearance and Normal Descriptor) [20], a recently proposed binary descriptor for RGBD images. It is originally used with CenSurE (Center Surround Extrema) keypoint detector [34], whose OpenCV implementation is referred to as STAR. For both descriptors we consider both the detectors suggested in the original papers, i.e., DoG for SIFT and STAR for BRAND, and the proposed one. On each image and with each detector we keep only 1000 keypoints with the highest detector response. We also conducted this test with VIP features, however, on all the test sequences VIP fails at the very first frames, revealing its inability to detect enough features to match two frames. Consequently, the relative pose can not be determined.

The results are shown in Table 1. For both descriptors, on almost all sequences the proposed keypoints reach smaller translation errors. The rotation errors are also small and comparable to the ones achieved by SIFT. Moreover, our method always outperforms STAR detector on all sequences and for all error types.

5. CONCLUSION

In this paper we present a simple and efficient keypoint detector for texture+depth images. Our proposed approach takes into account the depth map in order to provide an adapted multiscale representation of the texture map, that yields better keypoint repeatability. Our experiments show higher repeatability scores with respect to standard SIFT and VIP features, and improved tracking precision in a visual odometry application with SIFT and BRAND descriptors. Further improvement could be reached by using a more accurate filtering, that takes into account the edges of object in the scenes and adapts better to the local surfaces in the scenes. How to achieve this while keeping the computational complexity as limited as possible is currently one of our research directions.

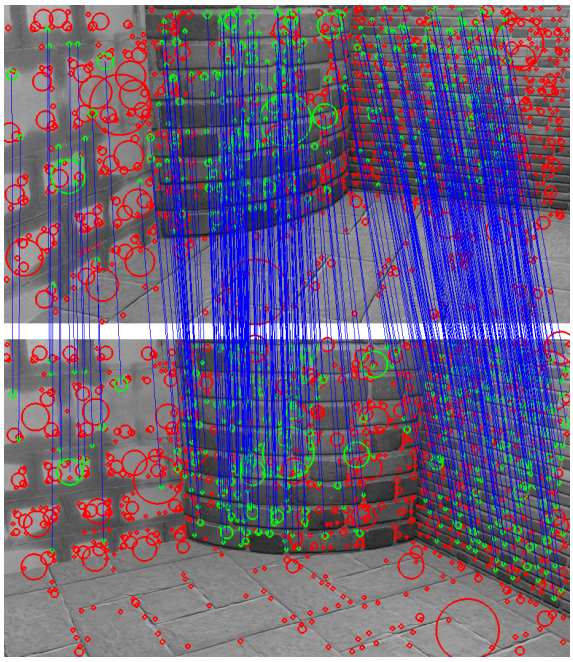

Fig. 3: Repeated keypoints of the proposed detector in two views of *Bricks* scene: 1257 / 1109 keypoints, 34.9% repeated for $\eta = 0.5$.

6. REFERENCES

[1] M. Karpushin, G. Valenzise, and F. Dufaux, "Improving distinctiveness of BRISK features using depth maps," in *Proceed. of IEEE Intern. Conf. Image Proc.*, Québec city, Canada, September 2015.

[2] ——, "A scale space for texture+depth images based on a discrete Laplacian operator," in *IEEE Intern. Conf. on Multimedia and Expo*, Torino, Italy, July 2015.

[3] T. Lindeberg and L. Bretzner, "Real-time scale selection in hybrid multi-scale representations," *Scale Space Methods in Computer Vision*, pp. 148–163, 2003.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Intern. J. of Comp. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] J. J. Koenderink, "The structure of images," *Biological cybernetics*, vol. 50, no. 5, pp. 363–370, 1984.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comp. Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[7] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proceed. of IEEE Intern. Conf. on Comp. Vision*, Barcelona, Spain, November 2011.

[8] J. Weickert, *Anisotropic diffusion in image processing*. Teubner Stuttgart, 1998, vol. 1.

[9] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, no. 7, pp. 629–639, 1990.

[10] M. Gobara and D. Suter, "Feature detection with an improved anisotropic filter," in *Computer Vision–ACCV*. Hyderabad, India: Springer, January 2006.

	SIFT descriptor								BRAND descriptor							
	Final translation error, m		Avg. translation error, m		Final rotation error, rad		Avg. rotation error, rad		Final translation error, m		Avg. translation error, m		Final rotation error, rad		Avg. rotation error, rad	
	DoG	Prop.	STAR	Prop.	DoG	Prop.	STAR	Prop.	DoG	Prop.	STAR	Prop.	DoG	Prop.	STAR	Prop.
01	139.2	**65.7**	57.0	**30.1**	0.26	**0.21**	0.22	**0.17**	511.3	**260.0**	166.5	**111.0**	2.85	**0.46**	1.70	**0.42**
02	61.5	**47.2**	32.8	**27.9**	**0.18**	**0.18**	0.19	**0.18**	303.9	**65.6**	100.1	**36.1**	2.84	**0.23**	1.52	**0.20**
03	11.5	**7.6**	6.5	**4.5**	0.19	**0.18**	**0.16**	**0.16**	17.3	**9.0**	10.9	**4.2**	0.70	**0.24**	0.27	**0.17**
04	78.9	**71.2**	35.3	**30.5**	**0.01**	0.03	**0.02**	**0.02**	153.2	**92.5**	64.7	**39.9**	1.19	**0.04**	0.19	**0.02**
05	**25.6**	29.4	17.4	**16.8**	**0.17**	0.18	**0.08**	**0.08**	57.6	**28.0**	31.1	**19.2**	0.58	**0.22**	0.33	**0.09**
06	75.5	**68.6**	42.4	**37.6**	**0.14**	0.17	0.06	**0.05**	183.7	**86.7**	92.8	**46.7**	0.84	**0.19**	0.48	**0.08**
07	**29.8**	**29.8**	**12.7**	13.0	**0.10**	0.11	**0.09**	**0.09**	57.0	**37.9**	16.2	**13.7**	0.56	**0.28**	0.42	**0.14**
08	**57.9**	75.0	**27.3**	36.4	**0.14**	0.17	**0.11**	0.14	**48.7**	69.9	40.3	**39.1**	0.85	**0.21**	0.49	**0.13**
09	**64.0**	**64.0**	**27.9**	28.6	0.14	**0.12**	0.09	**0.07**	136.1	**89.0**	47.5	**35.3**	1.36	**0.32**	0.45	**0.13**
10	**42.1**	47.8	**22.9**	24.5	0.22	**0.18**	0.28	**0.26**	47.2	**47.1**	33.9	**26.3**	0.92	**0.25**	0.65	**0.29**
Avg.	58.61	**50.62**	28.23	**24.98**	**0.15**	0.16	0.13	**0.12**	151.60	**78.57**	60.39	**37.15**	1.27	**0.25**	0.65	**0.17**

Table 1: Visual odometry results on KITTI dataset [29]. The lowest errors in each group are highlighted. The last line presents averaged results over the 10 sequences.

[11] S. Wang, H. You, and K. Fu, "BFSIFT: A novel method to find feature matches for SAR image registration," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 4, pp. 649–653, 2012.

[12] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," in *Computer Vision–ECCV 2012*. Florence, Italy: Springer, October 2012.

[13] A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud, "Surface feature detection and description with applications to mesh matching," in *Proceed. of IEEE Intern. Conf. on Comp. Vision and Pattern Rec.*, Miami, USA, June 2009.

[14] T. Darom and Y. Keller, "Scale-invariant features for 3-D mesh models," *IEEE Trans. Image Processing*, vol. 21, no. 5, pp. 2758–2769, 2012.

[15] I. Sipiran and B. 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, 2011.

[16] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys, "3D model matching with viewpoint-invariant patches (VIP)," in *Proceed. of IEEE Intern. Conf. on Comp. Vision and Pattern Rec.*, Anchorage, Alaska, USA, June 2008.

[17] T.-W. R. Lo and J. P. Siebert, "Local feature extraction and matching on range images: 2.5D SIFT," *Comp. Vision and Image Understanding*, vol. 113, no. 12, pp. 1235–1250, 2009.

[18] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "Point feature extraction on 3D range scans taking into accountobject boundaries," in *Proceed. of IEEE Intern. Conf. on Rob. and Autom.*, Shanghai, China, May 2011.

[19] B. Lin, F. Zhao, T. Tamaki, F. Wang, and L. Xiao, "SIPF: Scale invariant point feature for 3D point clouds," in *Proceed. of IEEE Intern. Conf. Image Proc.*, Qubec City, Canada, October 2015.

[20] E. R. do Nascimento, G. L. Oliveira, A. W. Vieira, and M. F. Campos, "On the development of a robust, fast and lightweight keypoint descriptor," *Neurocomputing*, vol. 120, pp. 141–155, 2013.

[21] K. Koser and R. Koch, "Perspectively invariant normal features," in *Proceed. of IEEE Intern. Conf. on Comp. Vision*, Rio de Janeiro, Brazil, October 2007.

[22] F. Tombari, S. Salti, and L. Di Stefano, "A combined texture-shape descriptor for enhanced 3D feature matching," in *Proceed. of IEEE Intern. Conf. Image Proc.*, Brussels, Belgium, September 2011.

[23] F. C. Crow, "Summed-area tables for texture mapping," *ACM SIGGRAPH computer graphics*, vol. 18, no. 3, pp. 207–212, 1984.

[24] M. Karpushin, G. Valenzise, and F. Dufaux, "An image smoothing operator for fast and accurate scale space approximation," in *Proceed. of IEEE Intern. Conf. Acoust., Speech and Sign. Proc.*, Shanghai, China, March 2016.

[25] C.-C. Su, L. K. Cormack, and A. C. Bovik, "Color and depth priors in natural images," *IEEE Trans. Image Processing*, vol. 22, no. 6, pp. 2259–2274, 2013.

[26] C.-C. Su, A. C. Bovik, and L. K. Cormack, "Natural scene statistics of color and range," in *Proceed. of IEEE Intern. Conf. Image Proc.*, Brussels, Belgium, 2011.

[27] K. Mikolajczyk and C. Schmid, "An affine invariant interest point detector," in *Computer Vision–ECCV 2002*. Springer, 2002, pp. 128–142.

[28] T. Lindeberg, "Feature detection with automatic scale selection," *Intern. J. of Comp. Vision*, vol. 30, no. 2, pp. 79–116, 1998.

[29] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proceed. of IEEE Intern. Conf. on Comp. Vision and Pattern Rec.*, Providence, Rhode Island, USA, June 2012.

[30] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *Intern. J. of Comp. Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[31] M. Karpushin, G. Valenzise, and F. Dufaux, "Local visual features extraction from texture+depth content based on depth image analysis," in *Proceed. of IEEE Intern. Conf. Image Proc.*, Paris, France, October 2014.

[32] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," in *Proceed. of Intern. Conf. on Multimedia*, ser. MM '10, New York, USA, 2010.

[33] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 6, pp. 756–770, 2004.

[34] M. Agrawal, K. Konolige, and M. R. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *Computer Vision–ECCV 2008*. Marseille, France: Springer, 2008.