A LOW-DIMENSIONAL FEATURE TRANSFORM FOR KEYPOINT MATCHING AND CLASSIFICATION OF POINT CLOUDS WITHOUT NORMAL COMPUTATION

Viktor Seib and Dietrich Paulus

Active Vision Group (AGAS), Institute for Computational Visualistics,
University of Koblenz-Landau, 56070 Koblenz, Germany
{vseib, paulus}@uni-koblenz.de

ABSTRACT

Most feature descriptors need point normal information to be computed prior to computing the descriptor itself. We present a descriptor transform and apply it to the SHOT descriptor that allows to entirely omit the computation of normals. Further, our transform reduces the number of descriptor dimensions of SHOT by more than 90%, decreasing the computational requirements for feature matching and the memory footprint for feature storage. Despite the heavy reduction in the number of dimensions, our approach retains a high descriptiveness and the computational efficiency of SHOT. We evaluate the proposed transform on datasets for keypoint matching and 3D shape classification and show that it is competitive with state of the art descriptors.

Index Terms— Descriptor Transform, Keypoint Matching, Point Cloud Registration, Shape Classification

1. INTRODUCTION

Reliably matching keypoints has become a wide-spread task with many application domains such as object classification, shape retrieval, registration and simultaneous localization and mapping. Nowadays, affordable RGBD-cameras are widely available and bring the task of keypoint matching into the domain of 3D data. For matching, keypoints need to have a unique representation that is usually achieved by computing a feature descriptor around the keypoint location.

A recent survey on the performance of feature descriptors [1] concluded that the best performing feature descriptors are RoPS [2], FPFH [3] and SHOT [4]. According to that survey, for time-critical application FPFH is best suited for data with a low number of points, while SHOT can efficiently handle a large number of points. On the other hand RoPS is less efficient to compute, but produces good results on point clouds with different noise levels and point densities. These handcrafted feature descriptors do not require prior training and can be computed directly on the data. However, a recent trend is to learn descriptors with deep neural networks [5], [6].

In this work we adhere to SHOT, one of the classic descriptors. SHOT achieves a good performance in descriptiveness and computational efficiency. Out of the three best performing feature descriptors reported in [1] it has the highest number of dimensions: 352. For comparison, FPFH has 33 and RoPS has 135 dimensions. Further, all of these descriptors require point normals to be computed prior to the descriptor computation. Computing normals for point clouds, especially if the point clouds are not organized, takes a significant amount of time. Real-time feature matching application therefore benefit from features that do not rely on point normals. Specific features for real-time classification were designed for object recognition [7] or traffic scene analysis [8].

In this paper we tackle these two shortcomings of SHOT. We present a transform called *Short SHOT* that efficiently reduces the number of dimensions of the SHOT descriptor from 352 to 32, which is 9% of its original size. Moreover, as we will show in Sec. 3 this transform omits the necessity to compute point normals prior to computing the descriptor. Thus, an entire pipeline step in descriptor computation is left out. The decrease of dimensions reduces computational requirements for feature matching and makes Short SHOT lightweight for feature storage. At the same time, Short SHOT retains a high descriptiveness and the computational efficiency of SHOT.

Transforming the SHOT descriptor was already proposed in [9] with the goal of creating B-SHOT, a binary descriptor for 3D data. The goals were similar to ours, reducing the memory footprint and allowing for faster matching. However, same as SHOT, B-SHOT also requires point normals to be computed in advance. Further, B-SHOT acquires its efficiency from the binary nature of the descriptor, whereas we reduce the number of dimensions. We therefore regard our contribution in this work as complementary to B-SHOT and will use a similar evaluation as [9] for keypoint matching and transformation matrix estimation based on the matched keypoints. Additionally, we evaluate the performance of Short SHOT in a probabilistic Hough-voting pipeline [10] for 3D shape classification.

To make this paper self-contained we briefly present the SHOT descriptor in Sec. 2. Our proposed feature transform is introduced in Sec. 3 and extensively evaluated in Sec 4. Finally, Sec. 5 summarizes and concludes the paper.

2. SHOT DESCRIPTOR

When presenting the Signature of Histograms of Orientations (SHOT) [4], Tombari et al. start by categorizing existing 3D feature descriptors into the categories *signatures* and *histograms*. By proposing the SHOT descriptor they aim at combining the advantages of signatures and histograms in one descriptor. The authors of SHOT further emphasize the importance of a unique and unambiguous local reference frame to compute the descriptors. Given their reference frame, a spherical 3D grid with 32 cells is superimposed on the local point neighborhood. A local histogram based on the normal angles is computed for each of the cells. The final descriptor has 352 dimensions and is the concatenation of these histograms, including some interpolation to avoid boundary effects.

3. SHORT SHOT TRANSFORM

The SHOT descriptor is a signature of histograms computed on a spherical 3D grid. It is crucial to emphasize the differences between a signature and a histogram at this point [11]. A signature is a *localization* of a specific (geometric) property, while a histogram is an *accumulation* of such a property. In the SHOT descriptor, the localization happens by dividing the spherical 3D grid into cells, while the accumulation occurs within the cells.

Each cell is a localization of points and their respective normals at a certain position relative to the local reference frame. Inside the cell and independent of its position each point's normal contributes to a histogram of values. More formally, the SHOT descriptor S is a concatenation of histograms H_i:

$$S = \bigcup_{i=1}^{n} H_i = \bigcup_{i=1}^{n} \bigcup_{j=1}^{b} \sum_{k} I_j(x_k)\,, \tag{1}$$

where $n = 32$ is the number of grid cells, $b = 11$ is the number of histogram bins and $I_j(\cdot)$ an indicator function

$$I_j(x) = \begin{cases} 1 & x \in j, \\ 0 & \text{otherwise}\,, \end{cases} \tag{2}$$

determining whether the value x falls into bin j. In case of SHOT, the value x is the angle between the z-axis of the reference frame in the query point and the normal of a neighboring point. The histogram of each cell encodes all values x_k of the neighboring points (within the support radius) that fall into that cell.

While the resulting descriptor is highly descriptive, it has the disadvantage of requiring the computation of point normals. Although point normals can be computed efficiently, especially on organized point clouds, this might not be sufficient enough for some real-time scenarios or unorganized point clouds [7], [8]. Further, when numerous descriptors need to be stored, a low-dimensional descriptor has the advantage of a lower memory footprint. The proposed transform targets application domains where time and memory load is more critical than accuracy.

We propose to transform the SHOT descriptor by subsuming the 11 bins to a single value to obtain the Short SHOT descriptor S_{32}:

$$S_{32} = \bigcup_{i=1}^{n} \sum_{j=1}^{b} \sum_{k} I_j(x_k)\,. \tag{3}$$

Essentially, instead of creating a histogram per cell, we propose to compute a sum per cell on the spherical grid. The dimensionality of the descriptor is thereby reduced from 352 to 32 which is only 9% of the original length. The points remain localized by the grid cell relative to the local reference frame. However, by omitting the histogram binning each point's contribution is the same regardless of the angle formed by the z-axis of the reference frame in a point and its normal. This step makes the proposed descriptor transform independent of point normals. As a consequence, the normal computation can be entirely omitted.

Each of the SHOT histograms of angles becomes a localized value representing the number of points that fall into each of the grid cells. These single values from all grid cells represent a per cell accumulated quantity, i.e. a histogram. In this case, however, the histogram does not accumulate over angles, but over spatial point positions, the grid cells. The result of the transform, the Short SHOT descriptor, can thus be regarded as a histogram of signatures - although the signatures are degenerated to single values. After applying Eq. 3 the descriptor is normalized as the original SHOT descriptor to make it robust against point density variations.

As the original SHOT descriptor, we expect Short SHOT to be highly descriptive and robust. Due to the significant reduction of dimensionality we expect some minor loss in descriptiveness. On the other hand, this descriptor does not require point normal information to be computed in advance.

The proposed transform was tested and implemented using the SHOT descriptor. Nonetheless, it is generalizable to other descriptors employing histograms. This includes other descriptors for 3D data, but also descriptors for 2D images. In this case SHOT was chosen because of its high descriptiveness and efficient computation. Further, by applying the proposed transform we can entirely omit a pipeline step, namely the point normals computation (not present in 2D data).

4. EVALUATION

We evaluate our descriptor transform in two distinct experiments. The first experiment determines the quality of keypoint matching and estimation of a transformation matrix. The second experiment applies Short SHOT in a probabilistic Hough-voting scheme [10] for shape classification. A recent

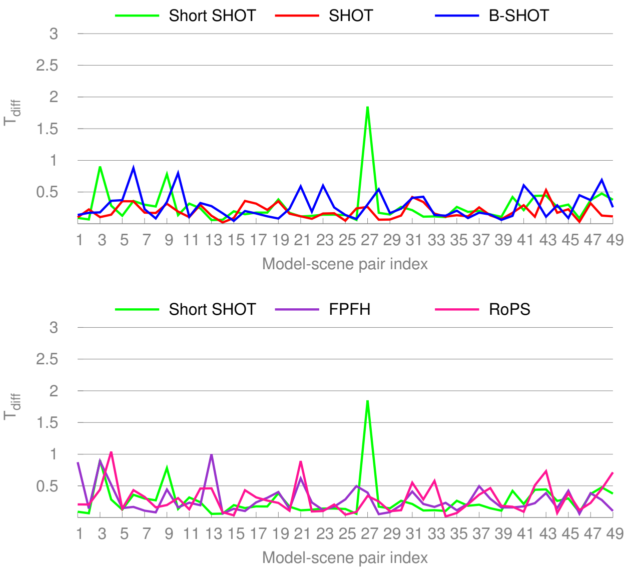

Fig. 1. Comparison of descriptors based on the T_{diff} metric. Despite the significantly reduced length of Short SHOT this descriptor allows to estimate transformation matrices with a low error, matching the low errors obtained with other descriptors in many cases.

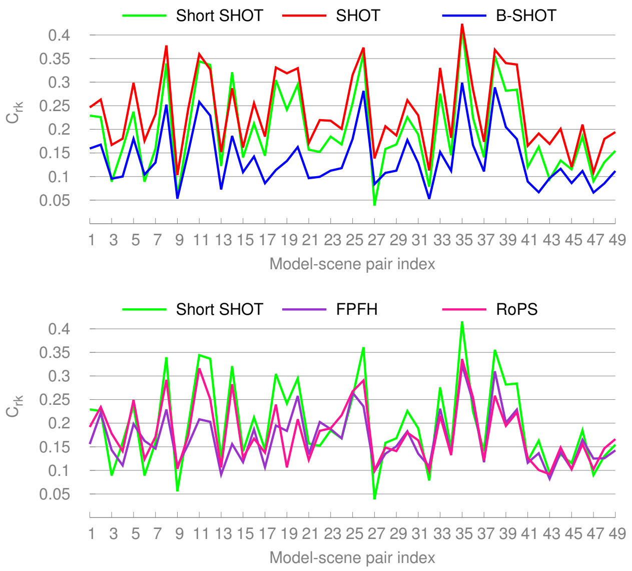

Fig. 2. Comparison of descriptors based on the C_{rk} metric. The number of correct keypoint correspondences with Short SHOT is close to the number of correspondences obtained with the original SHOT descriptor. In some cases Short SHOT even produces more correct matches than FPFH and RoPS.

survey [1] identified SHOT [4], FPFH [3] and RoPS [2] as the best performing descriptors. We therefore compare Short SHOT with these descriptors and additionally with a more recent binary descriptor, B-SHOT [9], in both experiments. Further, we include a comparison with CGF [5], a deep learned descriptor, in the second experiment.

4.1. Keypoint Matching

We perform a similar evaluation as Prakhya et al. [9] for their B-SHOT descriptor for keypoint matching. We employ an extended version of the publicly available Kinect dataset[1] of Tombari et al. [12]. The dataset consists of several scenes comprising 2 to 4 models each. For each model and scene, ground truth transformation data is available. In total there are 49 model-scene pairs available.

First, we uniformly extract keypoints on a regular grid on the scene and the model and compute feature descriptors. On one hand, this entails a high keypoint extraction ambiguity and on the other hand this promotes false correspondences with keypoints from the background of the scene. Then, reciprocal correspondences between the model and scene descriptors are established. Finally, we use RANSAC to filter outliers and estimate a 3D transformation T_h of the model.

The quality of the estimated transformation hypothesis T_h is assessed by computing the difference T_{diff} between the ground truth transformation T_g and T_h with the Euclidean metric

$$T_{diff} = \sqrt{\sum_{i=0}^{n} \sum_{j=0}^{n} \left(T_{h_{ij}} - T_{g_{ij}}\right)^2}\,, \qquad (4)$$

with $n = 3$ for homogeneous 3D transformation matrices. Further, the quality of the descriptor matching is assessed with the correspondence ratio $C_{rk} = \frac{c_r}{k_m}$ as the ratio between the RANSAC correspondences c_r and the number of keypoints extracted from the model k_m.

The comparative results are presented in Fig. 1 and Fig. 2. Keypoints were extracted on a grid of $0.01m$ and the descriptor support radius was set to $0.05m$ in all experiments. For the T_{diff} metric (Fig. 1) we observe an overall good performance of Short SHOT. Although being the descriptor with the smallest dimensionality, it is sufficient to compute a transformation matrix with small errors. The errors are very similar to SHOT in many cases and mostly below the errors achieved with FPFH and RoPS. There is only one outlier (index 27) stemming from very few keypoints that could be matched in that model-scene pair. In the other cases (Fig. 2) the number of correspondences that were matched correctly is again similar to the SHOT descriptor that has significantly more dimensions and above B-SHOT, FPFH and RoPS. This result highlights the ability of Short SHOT to reliably match keypoints that serve for a correct pose estimation with a small error and without computing point normal information.

[1] Kinect dataset: `http://vision.deis.unibo.it/keypoints3d/?page_id=2`

Table 1. Classification accuracy in percent achieved with the evaluated descriptors on the 3D shape datasets. Short SHOT achieves same performance as RoPS on ASW, beats FPFH on SH12 and beats all descriptors on the MCG dataset.

	ASW	PSB	SH12	MCG
Short SHOT	86.0	60.5	62.7	80.7
SHOT	88.5	64.7	69.2	80.3
B-SHOT	87.0	62.0	64.3	78.9
FPFH	88.0	61.4	60.3	77.6
RoPS	86.0	62.1	66.5	78.5
CGF	80.5	58.7	58.3	73.1

With the same keypoint grid, but a higher support radius $(0.1m)$, the number of matched keypoints with Short SHOT is even closer to the number of SHOT matches. When matched with a sparser grid of $0.03m$ and a radius of $0.05m$ the value T_{diff} increases for all descriptors. Still, Short SHOT stays on a similar level as SHOT and B-SHOT.

4.2. Shape Classification

Since a descriptor that is reduced to 9% of its original length is expected to be less descriptive we evaluate Short SHOT in a probabilistic Hough-voting pipeline for shape classification [10]. We employ typical 3D shape datasets used to benchmark 3D classification and shape retrieval algorithms:

ASW [13]: a dataset with 400 rigid objects in 20 classes, half for training, half for classification

PSB [14]: 1814 rigid objects in 7 classes, half for training, half for classification

SH12 [15]: a dataset with 1200 rigid objects in 60 classes, half for training, half for classification

MCG [16]: 457 articulated objects in 19 classes, 234 object for training and 224 for classification

We briefly introduce the pipeline and kindly refer the reader to [10] for details. The Hough-voting pipeline in this evaluation is an adaptation of the Implicit Shape Model approach of Leibe et al. [17] to 3D data. Keypoints are extracted on a dense grid and descriptors computed at the corresponding positions. Then a codebook is built by finding k-nearest neighbors of each descriptor and storing voting vectors for each of them. Descriptors obtained on the input model are matched with the codebook and the voting vectors cast votes into a continuous Hough-space for classification. Maxima in the Hough-space indicate an object's identity and position.

Tab. 1 reports the classification results for all evaluated descriptors on the four different 3D shape datasets. Short SHOT was matched with the Chi-Squared distance and the Hamming distance was used for B-SHOT, due to its binary nature. All other descriptors were matched with the Euclidean distance as the default in the literature. CGF [5] is available in different variants. We report the variant with the best results $(40$ dimensions learned from the laser dataset).

Table 2. Speed-up achieved when using Short SHOT for 3D object classification in the framework from [10]. Tests were run on a notebook with a Core i7-2760QM CPU @ 2.40GHz. Reported times are per object on the ASW dataset averaged over 10 runs.

	Normals	Feature	Complete Time
Short SHOT	-	168 ms	423 ms
SHOT	130 ms	179 ms	616 ms

As expected, compared to SHOT our proposed transform looses some descriptiveness due the much shorter descriptor. However, for MCG, the dataset with articulated objects, we observe a small increase in classification accuracy and an overall best performance of Short SHOT. We attribute this to the broad range of shape variants per class. The classic descriptors slightly overfit on this dataset, while Short SHOT captures the more general appearance of the shapes. On the other three datasets Short SHOT beats FPFH additionally on the SH12 dataset and achieves same performance as RoPS on the ASW dataset, while being considerably faster to compute and match. B-SHOT on the other hand performs slightly better than Short SHOT on these datasets. Surprisingly, the deep learned CGF performs worst. We attribute this to the different type of dataset that was used for learning in [5] which is not completely transferable to our use case. For the classification task on these datasets the RoPS and especially the FPHF descriptor have a very long computation time, while B-SHOT is in the same order of magnitude as Short SHOT. To better estimate the speed-up gained with Short SHOT we compare it to SHOT in Tab. 2. The most speed-up is gained by omitting the normal computation. Further the feature matching is faster with Short SHOT (not separately shown). The complete classification time per object reduces to 69% when compared to the SHOT descriptor, while the memory for storing the descriptors is reduced to 9%.

5. CONCLUSION

We present a feature transform called Short SHOT, which reduces the SHOT descriptor length by over 90%. Short SHOT is competitive with state of the art descriptors like SHOT, FPFH and RoPS for the task of feature matching. Further, Short SHOT achieves better results on some datasets for 3D shape classification than the well established descriptors, while being faster to compute and match without the need to compute point normals. Since the dimensionality reduction comes with a small loss in descriptiveness, Short SHOT is suited for applications where computation time and the memory footprint are more crucial than classification accuracy.

6. REFERENCES

[1] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, Jianwei Wan, and Ngai Ming Kwok, "A comprehensive performance evaluation of 3d local feature descriptors," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 66–89, 2016.

[2] Yulan Guo, Ferdous A Sohel, Mohammed Bennamoun, Jianwei Wan, and Min Lu, "Rops: A local feature descriptor for 3d rigid objects based on rotational projection statistics," in *Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on*. IEEE, 2013, pp. 1–6.

[3] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 3212–3217.

[4] Federico Tombari, Samuele Salti, and Luigi Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. of the European Conf. on computer vision (ECCV)*. 2010, ECCV'10, pp. 356–369, Springer-Verlag.

[5] Marc Khoury, Qian-Yi Zhou, and Vladlen Koltun, "Learning compact geometric features," in *International Conference on Computer Vision (ICCV)*, 2017.

[6] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 199–208.

[7] Guan Pang and Ulrich Neumann, "Training-based object recognition in cluttered 3d point clouds," in *3D Vision-3DV 2013, 2013 International Conference on*. IEEE, 2013, pp. 87–94.

[8] Michael Kusenbach, Michael Himmelsbach, and Hans-Joachim Wuensche, "A new geometric 3d lidar feature for model creation and classification of moving objects," in *Intelligent Vehicles Symposium (IV), 2016 IEEE*. IEEE, 2016, pp. 272–278.

[9] Sai Manoj Prakhya, Bingbing Liu, and Weisi Lin, "B-shot: A binary feature descriptor for fast and efficient keypoint matching on 3d point clouds," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 1929–1934.

[10] Viktor Seib, Norman Link, and Dietrich Paulus, "Pose estimation and shape retrieval with hough voting in a continuous voting space," in *German Conference on Pattern Recognition*. Springer, 2015, pp. 458–469.

[11] Samuele Salti, Federico Tombari, and Luigi Di Stefano, "Shot: Unique signatures of histograms for surface and texture description," *Computer Vision and Image Understanding*, vol. 125, pp. 251–264, 2014.

[12] Federico Tombari, Samuele Salti, and Luigi Di Stefano, "Performance evaluation of 3d keypoint detectors," *International Journal of Computer Vision*, vol. 102, no. 1-3, pp. 198–220, 2013.

[13] Daniela Giorgi, Silvia Biasotti, and Laura Paraboschi, "Shape retrieval contest 2007: Watertight models track," *SHREC competition*, vol. 8, no. 7, 2007.

[14] Philip Shilane, Patrick Min, Michael Kazhdan, and Thomas Funkhouser, "The princeton shape benchmark," in *Shape modeling applications, 2004. Proceedings*. IEEE, 2004, pp. 167–178.

[15] Bo Li, Afzal Godil, Masaki Aono, X Bai, Takahiko Furuya, L Li, Roberto Javier López-Sastre, Henry Johan, Ryutarou Ohbuchi, Carolina Redondo-Cabrera, et al., "Shrec'12 track: Generic 3d shape retrieval.," *3DOR*, vol. 6, 2012.

[16] Kaleem Siddiqi, Juan Zhang, Diego Macrini, Ali Shokoufandeh, Sylvain Bouix, and Sven Dickinson, "Retrieving articulated 3-d models using medial surfaces," *Machine vision and applications*, vol. 19, no. 4, pp. 261–275, 2008.

[17] Bastian Leibe, Ales Leonardis, and Bernt Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Workshop on statistical learning in computer vision, ECCV*, 2004, vol. 2, p. 7.