Shortened Color-Shape Descriptors for Point Cloud Classification from RGB-D Cameras

Viktor Seib
University of Koblenz-Landau
56070 Koblenz, Germany
vseib@uni-koblenz.de

Dietrich Paulus
University of Koblenz-Landau
56070 Koblenz, Germany
paulus@uni-koblenz.de

***Abstract*—Deep learning techniques have become the standard approach for computer vision tasks. However, traditional methods are still advantageous in application domains with limited computing or battery power such as mobile robots. In this work, we address the task of point cloud classification from RGB-D cameras with a traditional approach using local feature descriptors. We modify CSHOT, an established feature descriptor, with the goal of radically reducing its dimensionality. The resulting Short-CSHOT descriptor is evaluated on common point cloud data sets for keypoint matching and object classification. The obtained results suggest that Short-CSHOT is able match the original CSHOT descriptor, while at the same time having a much lower dimensionality. Further, the number of dimensions of Short-CSHOT can be varied, allowing to trade-off between the memory footprint and accuracy. All code is made publicly available on https://github.com/vseib/PointCloudDonkey.**

***Index Terms*—Classification, Point clouds, Descriptors, RGB-D Cameras, Mobile Robots**

I. Introduction

Deep neural networks have revolutionized the field of computer vision by achieving high recognition rates at short inference times. The price to pay for these powerful classifiers is a long training time and the need for a vast amount of annotated data. Mobile robots are often equipped with RGB-D cameras for perception. While large annotated data sets with images exist [1], such data sets for point clouds are rare. Recently, ModelNet [2] has become one of the mostly used data sets for classification of synthetic point cloud with neural networks. Less focus is put on data sets such as the Washington RGB-D [3], BigBird [4] and YCB [5] that provide colored point clouds from RGB-D cameras (Fig. 1).

The necessity to process point clouds directly is increasing due to LiDAR sensors and increased interest in autonomous cars. Still, many of the neural networks using ModelNet do not process point clouds directly, but transform them to depth or stereoscopic images or volumetric representations. [1]

Usually, mobile robots at competitions such as RoboCup are equipped with off-the-shelf notebook computers. For training neural networks a powerful computer equipped with at least one graphics processing unit (GPU) is needed. Running a trained network on the available notebook is often infeasible, thus another computing device is needed that is specialized for deep learning inference and is efficient in terms of power consumption. This constellation leads to a total of three devices which might exceed available budgets or battery power. On the other hand, training and inference of traditional approaches is viable on portable computers already available on mobile robot platforms. Trading off these advantages for a lower accuracy might be acceptable in many cases for robotic applications.

Fig. 1. Example object point cloud views from the Washington RGB-D [3] (top row), BigBird [4] (center row) and YCB [5] (bottom row) data sets. Individual objects are not to scale.

Even though research on traditional approaches has stagnated in the last years, we want to make some contributions in that area. Color SHOT (CSHOT) [6] is a well-known descriptor for point clouds combining color and shape cues. The surveys presented in [7] and [8] attribute the best balance between speed and accuracy to the CSHOT descriptor. Unfortunately, with 1344 dimensions CSHOT is a very high dimensional descriptor. In this work we several possibilities to modify the structure of this descriptor. This extends our previous work where we transformed the SHOT [9] descriptor to make it faster and more memory efficient [10].

Our contribution in this work is the *Short Color SHOT* (SCSH), a family of shortened descriptors that can be varied in size to trade-off between the memory footprint and accuracy. We present keypoint matching as well as object recognition experiments and show that SCSH is competitive with CSHOT in many cases. We evaluate the influence of SCSH dimensions on the classification accuracy in an ablation study. Further,

[1] Only 15 out of 63 methods reported on https://modelnet.cs.princeton.edu/ operate on point clouds directly.

we evaluate SCSH on common point cloud data sets from RGB-D cameras, including the YCB data set, that has been chosen as the standard object data set for the RoboCup@Home competition [11].

This work is structured as follows. Sections II and III present related work and review the CSHOT descriptor. We introduce the required prerequisites and the SCSH descriptor in Sections IV and V, followed by experiments in Section VI. Finally, the paper is concluded by a discussion in Section VII and a summary in Section VIII.

II. Related Work

Among the recent work exploiting traditional approaches is the people detection method of Hanten et al. [12]. They focus on a light-weight approach without GPUs and without deep neural networks for mobile and robotic applications. Their algorithm is capable of processing LiDAR and RGB-D point clouds. Lewandowski et al. [13] also target robotic applications with limited computing power for their proposed person detection approach. The authors focus on the FPFH descriptor [14], but also consider SHOT in their evaluation. Both works are also compared with deep learning methods, which perform better in the respective application domain, but require a significantly longer processing time.

The Point Cloud Library (PCL) [15] is a popular framework for point cloud processing with many feature descriptors. The authors of [8] systematically evaluate all PCL descriptors and conclude that the color version of FPFH and CSHOT perform best, while the latter is much faster to compute. Further, Guo et al. [16] and Yang et al. [17], while identifying some shortcomings of the SHOT descriptor, conclude that it is well suited for time-crucial applications and for classification of point clouds from RGB-D cameras. Based on these findings in literature we decided to focus on the CSHOT descriptor in this work.

Multivariate analysis methods such as t-SNE [18] can be used to reduce dimensionality. However, t-SNE and other nonlinear embedding methods are not suitable for our application because we aim at a fast computation on mobile platforms. Our goal is to obtain an efficient, short descriptor that can be plugged into existing object classification pipelines. An additional nonlinear embedding step would introduce an unwanted computational burden.

Transforming descriptors for point cloud data has been reported for example by Prakhya et al. [19]. They propose to transform the SHOT descriptor to create B-SHOT, a binary descriptor for 3-D data. In previous work, we also transformed the SHOT descriptor to obtain *Short SHOT* [10]. Both of these transformations had the goal of reducing the memory footprint and allowing for faster feature matching. However, same as SHOT, B-SHOT requires point normals to be computed in advance, which is not needed for Short SHOT. In this work we extend the Short SHOT formulation to additionally handle color cues of RGB-D point clouds.

III. Review of SHOT and CSHOT

The SHOT descriptor is a signature of histograms computed on a spherical 3-D grid. Each cell is a localization of points at a certain position relative to a local reference frame. Inside each cell each point's normal contributes to a histogram of values. More formally, the SHOT descriptor S_g is a concatenation of histograms H_i:

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

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 CSHOT descriptor S encodes geometric and color properties which are concatenated as

$$S = S_g \cup S_c\,. \tag{3}$$

The geometric part S_g is represented by the SHOT descriptor, while the color part S_c is again a concatenation of histograms according to (1) and (2). The difference is that the encoded value x (2)) is now the color distance in CIELAB color space to a reference point and the histogram size b is 31.

IV. Generalized Short SHOT Formulation

In [10] we proposed the Short SHOT (SSH) transformation to the SHOT descriptor by subsuming the 11 histogram bins to a single value:

$$\mathrm{SSH} = \bigcup_{i=1}^{n_g} \sum_{j=1}^{b} \sum_{k} I_j(x_k)\,. \tag{4}$$

Instead of creating a histogram of angles per cell, we propose to compute a sum of points 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 size. The points remain localized by the grid cell relative to the local reference frame. However, by omitting the histogram binning, SSH becomes independent of point normals. The evaluation in [10] showed a speed-up of 31% of the complete object recognition pipeline when the computation of normals is omitted.

Essentially for the Short CSHOT descriptor introduced in the next Section is to remove the constraints on the number of grid cells n. In the generalized SSH formulation, n_g will not be fixed to 32, but is instead determined by choosing suitable values for radial (r_g), elevation (e_g) and azimuth (a_g) subdivisions of the spherical grid (Fig. 2):

$$n_g = r_g \cdot e_g \cdot a_g\,. \tag{5}$$

The subscript g is added to emphasize the description of the geometric cues. For sake of notation, the SSH descriptor in (4) becomes SSH_{32} to indicate the chosen number of bins.

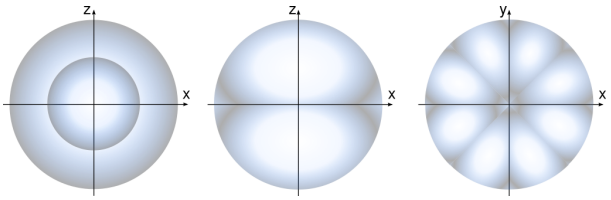

Fig. 2. Subdivisions of the spherical grid of SHOT, from left to right: 2 radial, 2 elevation and 8 azimuth subdivisions, resulting in a total of 32 grid cells.

V. Short Color Shot Descriptor

With 1344 dimensions, CSHOT leads to a high memory footprint for larger data sets and for traditional approaches applying a codebook of local features. We introduce a short version of the CSHOT descriptor, in the following dubbed *Short CSHOT* (SCSH), to reduce the number of dimensions, while maintaining a high accuracy for object recognition.

We find three possibilities to reduce the size of CSHOT. Instead of the original SHOT descriptor we will apply the SSH formulation to describe the geometric cues (Section IV). Further, we do not restrict the spherical grids to have the same number of subdivisions and will use a different number of cells to partition the point cloud for the geometric and the color part of the SCSH descriptor. Finally, the third property is a variable histogram size to encode the color distances per grid cell.

The reasoning behind the different spherical grids is that less subdivisions for the color part than for the geometric part are sufficient, since we still rely on histograms to describe color. Since RGB-D cameras often produce medium to low quality imagery, we see the opportunity to reduce the histogram size, since less bins should be sufficient to encode the color distribution. Also, more bins might introduce arbitrary partitioning of color values due to noise. The histograms are computed in the same way as for CSHOT.

The SCSH descriptor is a concatenation of a shortened geometric $\hat{S}_g$ and a shortened color descriptor $\hat{S}_c$:

$$\mathrm{SCSH} = \hat{S}_g \cup \hat{S}_c \,. \tag{6}$$

The shortened geometric part $\hat{S}_g$ is represented by the generalized SSH_* descriptor (4) with $*$ indicating different descriptor sizes. The shortened color part $\hat{S}_c$ resembles the definition of SHOT (1) as a concatenation of histograms H_i:

$$\hat{S}_c = \bigcup_{i=1}^{n_c} H_i = \bigcup_{i=1}^{n_c} \bigcup_{j=1}^{h} \sum_{k} I_j(x_k) \,, \tag{7}$$

where n_c is the number of color grid cells and h is the number of histogram bins. The total number of dimensions for the SCSH descriptor is computed similar to (5) as

$$n = n_g + n_c \cdot h = (r_g \cdot e_g \cdot a_g) + (r_c \cdot e_c \cdot a_c) \cdot h \,, \tag{8}$$

where r_c, e_c, a_c indicate the radial, elevation and azimuth subdivisions of the color grid (Fig. 2) and h is the number of bins in the color histogram. Note that we do not need to compute normals for the SCSH descriptor since we apply the SSH descriptor for the geometric part.

TABLE I
Two SCSH configurations used in the experiments, both have significantly less dimensions than CSHOT (1344).

	dimensions	relative size	n_g	n_c	h
Conf. 1	336	25%	96	16	15
Conf. 2	696	52%	96	24	25

VI. Experiments

The proposed SCSH descriptors will be evaluated in three experiments. In the first experiment we will determine their quality for keypoint matching. In the second experiment we will replace the original CSHOT descriptor by SCSH in a point cloud processing pipeline for object recognition [20] for a comparative evaluation. In the last experiment we will evaluate the impact of the number of dimensions of SCSH on the object recognition accuracy by varying the number of shape and color subdivisions as well as histogram sizes.

For the first two experiments we will use two distinct SCSH configurations. The first configuration has 336 dimensions, is slightly shorter than the geometric SHOT descriptor and has only about 25% of the dimensions of CSHOT. This configuration has only half of the CSHOT grid cells and bins per histogram. The second configuration has 696 dimensions, which is about 50% of the dimensions of CSHOT. While this configuration has more color grid cells and bins per histogram than the first configuration, it is still less than CSHOT. These two configurations will be denoted as SCSH-336 and SCSH-696 and are summarized in Tab. I.

A. Keypoint Matching

We perform a similar evaluation as proposed by Prakhya et al. [19] for their BSHOT descriptor for keypoint matching, using the Kinect data set provided by Tombari et al. [21]. The data set consists of several scenes comprising 2 to 4 objects each, with 49 object-scene pairs in total. For each object and scene, ground truth transformation data is given.

We uniformly extract keypoints on the scene and object and compute feature descriptors. This entails a high keypoint extraction ambiguity and promotes false correspondences with keypoints from the background of the scene. Then, reciprocal correspondences between the object and scene descriptors are established. Finally, we use RANSAC to filter outliers and estimate a 3-D transformation T_h of the object.

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

$$T_{\mathrm{diff}} = \sqrt{\sum_{i=0}^{m} \sum_{j=0}^{m} \left(T_{h_{ij}} - T_{g_{ij}}\right)^2} \,, \tag{9}$$

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

The comparative results are presented in Fig. 3 and Fig. 4. For the T_{diff} metric (Fig. 3) we observe an overall good

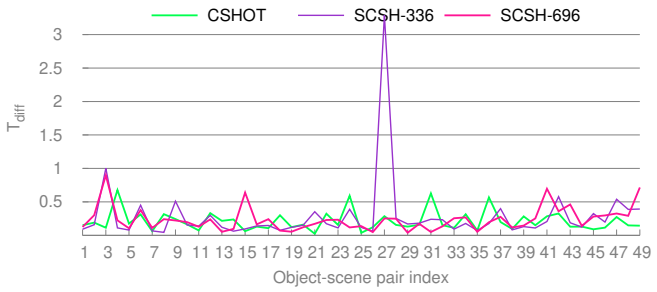

Fig. 3. Comparison of descriptors based on the T_{diff} metric (lower values are better). Despite the reduced size of SCSH these descriptors perform on a similar level. One exception is the high error of SCSH-336 on one of the object-scene pairs.

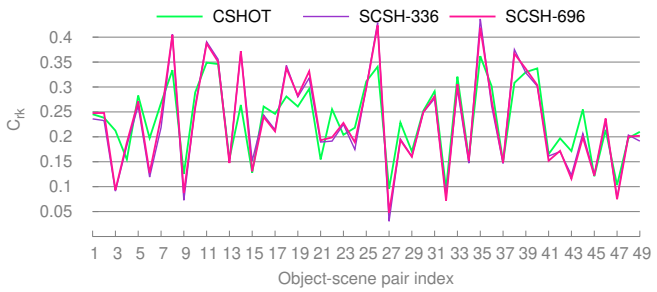

Fig. 4. Comparison of descriptors based on the C_{rk} metric (higher values are better). The number of correct keypoint correspondences is almost identical for both SCSH descriptors. It is also close to the number of correspondences obtained with CSHOT.

performance of both SCSH descriptors. In most cases the obtained transformation error is similar or close to the error obtained with the CSHOT descriptor. There is only one outlier for SCSH-336 with a high error (object-scene index 27), which comes from a very low number of matched keypoints (Fig. 4).

The C_{rk} metric shows that a similar number of keypoints is correctly matched for both SCSH configurations. Overall, there are several cases, where more keypoints are matched with CSHOT than SCSH and vice versa. Both metrics indicate that SCSH descriptors have a similar performance as CSHOT, which has much more dimensions.

B. Object Recognition

In this experiment we compare the recognition accuracy of SCSH descriptors with CSHOT, which serves as a baseline. For this purpose, we employ the object recognition pipeline presented in [20]. The evaluation is carried out using the Washington RGB-D data set (W-RGB-D) [3], the BigBird data set [4] and the YCB [5] data set that has been selected as the default data set for the RoboCup@Home competition [11]. Each of these data sets offers objects recorded from different views and elevation angles. Sample objects are shown in Fig. 1. Out of these data sets only the W-RGB-D defines a training and test split for evaluation.

We adapt the following evaluation strategy. We use a partial W-RGB-D data set, consisting of all classes, but only three

TABLE II
DATA SET OVERVIEW. CLASS LABELS FOR THE BIGBIRD AND YCB DATA SETS WERE ADDED BY US.

	# classes	# instances	# instances/class
W-RGB-D	51	153*	3
BigBird	66	125	mostly 1, max. 12
YCB	37	79	mostly 1, max. 13

*The full data set has 300 instances.

instances per class, to find suitable hyper parameters for the evaluation pipeline with CSHOT. With these hyper parameters we select the three most similar elevation angles corresponding to the W-RGB-D data set from BigBird and YCB. The views from the lowest and highest angles are used for training, while the medial view is used for testing. This evaluation strategy is called *leave-sequence-out* in the original W-RGB-D publication [3] which we adapt for BigBird and YCB as well.

Objects of the W-RGB-D data set are annotated with two labels: one indicating the class and another indicating the instance id. A class can be for example a *mug*, *ball*, or *tool* while instances ids simply enumerate objects belonging to that category. While BigBird and YCB do not provide class labels, many instances share a suffix such as *ball*, a brand name or simply *a*, *b* and *c*, offering candidate instances to be grouped into the same class. We add class labels for two reasons. First, we want to treat all data sets equally and exploit both labels. Second, in competitions such as the RoboCup@Home the robot may offer an alternative object if the actual object can not be found. In practice, this translates to offering a different instance from the same class. We therefore consider recognizing objects based on two labels, a class and instance label, as an important property of a recognition pipeline. A data set overview is given in Tab. II. Point clouds for YCB had to be reconstructed from the given RGB-D images. Some objects could not be reconstructed due to missing data (depth images or object masks), others were too flat or transparent and resulted in randomly scattered points. In total, we reconstructed 79 objects successfully. The grouping of instances to classes is detailed on the project website.

In the evaluation we report the metrics class accuracy (C.A.) and instance accuracy (I.A.). Both metrics indicate the percentage of object views from the test set that were assigned the correct class or instance label, respectively.

The recognition results on these data set are reported in Tab. III. Due to the significantly reduced dimensionality we expected a worse performance of the SCSH descriptors compared to CSHOT. However, the difference is quite small for the W-RGB-D data set. Compared to CSHOT, SCSH-696 has a 0.6 percentage points lower class accuracy and instance accuracy. The SCSH-336 loses 1.2 and 1.4 percentage points class and instance accuracy. We consider this loss in accuracy as small and as an acceptable trade-off.

The results on the BigBird data set are completely different. Compared to CSHOT, SCSH-336 has a 3.1 percentage points *higher* class accuracy and also a 4.3 percentage points *higher* instance accuracy. Results with the SCSH-696 descriptor im-

TABLE III
COMPARISON OF CLASSIFICATION ACCURACIES OF THE CSHOT AND SCSH DESCRIPTORS USING THE *leave-sequence-out* EVALUATION. ONLY WORK USING POINT CLOUDS IS REPORTED.

	W-RGB-D		BigBird		YCB	
	C.A.	I.A.	C.A.	I.A.	C.A.	I.A.
CSHOT (Ours)	91.6	83.4	84.0	71.2	87.9	73.2
SCSH-336 (Ours)	90.4	82.0	87.1	75.5	78.4	66.4
SCSH-696 (Ours)	91.0	82.8	91.0	81.7	81.4	68.6
CoSPAIR [22]	86.2*	74.0*	-	73.0	-	-
CSHOT [23]	-	94.1*	-	88.4	-	-
SURF + SHOT [24]	55.9	42.1	-	-	-	-
CSHOT [24]	51.0	36.4	-	-	-	-

*Reported on the full data set.

prove even further: it has a 7.0 and 10.5 percentage points *higher* class and instance accuracy. Unfortunately, we do not observe this positive effect on the YCB data set where the performance of SCSH descriptors can not match our baseline.

Tab. III also reports a comparison with other work. Please note that we only include work that uses point clouds on these data sets (rarely found in literature) in contrast to using RGB and depth images. Additionally, we are the first ones to report classification results on the YCB data set which is mostly used for pose estimation and grasping benchmarking. Although the W-RGB-D data set was primarily used to find suitable hyper parameters we report classification results from the same split and the whole data set for a fair comparison.

C. Varying the Dimensionality of SCSH

This section presents ablation experiments for the parameters n_g, n_c and h of the SCSH configuration (see Tab. I). We systematically tested different values for these parameters, however, only combinations that result in a total number of dimensions below 1344.

The number of geometric subdivisions n_g corresponds to the total number of descriptor dimensions of SSH. Keeping the color related parameters n_c and h constant and varying n_g we observed an increasing accuracy for classes and instances. However, the gain in accuracy of using 96 geometric subdivisions over 64 is already small. This indicates that shape cues are less significant on these data sets than color cues. Increasing n_g further has no positive impact on the accuracy, but adds more descriptor dimensions. Consequently, the following experiments were performed with a fixed $n_g = 96$.

Fig. 5 shows the effect of increasing the number of color subdivisions n_c (different graphs) and the number of different histogram sizes per subdivision h (axis of abscissae). While choosing an n_c of 24 (green graphs) over 16 (blue graphs) clearly increases the accuracies, a higher n_c of 32 (red graphs) or more has only a marginal effect while at the same time adding many additional dimensions to the descriptor.

Finally, the impact of a higher number of histogram bins h can be observed best on the BigBird and YCB data sets in Fig. 5. With an $h = 10$ we observe accuracy values that are a few percentage points below the baseline with CSHOT on the BigBird data set. Using an $h = 15$ or higher we obtain values that are close to the baseline on the W-RGB-D data set and values that are much higher than the baseline on the BigBird data set. However, the accuracy on the W-RGB-D data set stagnates for values higher than 15. Contrary, we observe a significant increase of the accuracy on the BigBird data set for a h of up to 30. Using a $h = 35$ also leads to marginal improvements, but adds an undesired number of dimensions to the descriptor. While a higher h also improved the accuracy on the YCB data set, it can not match the CSHOT baseline.

VII. DISCUSSION

The proposed SCSH descriptors perform well in the task of keypoint matching and object recognition. We have examined two configurations in detail, SCSH-336 and SCSH-696 which correspond to around 25% and 50% the size of CSHOT. With this substantial reduction both configurations have a similar key point matching performance as CSHOT, while having a much lower memory footprint. However, SCSH configurations with lower dimensions can miss some keypoints and lead to higher transformation errors (index 27 in Fig. 3).

The results of the object recognition experiments closely match the CSHOT baseline on the partial W-RGB-D data set. On the BigBird data set we observe a significant increase in recognition accuracy. One explanation are different data acquisition setups used to create these datasets. The color data of the W-RGB-D data set appear much brighter, indicating a better lighting during acquisition, while colors of BigBird appear darker. This leads to noisy color histograms in the CSHOT descriptor leading to lower classification accuracies. Further, the 32 histograms of CSHOT are locally more constrained than the 16 or 24 color histograms of SCSH spanning over a larger area. We assume that CSHOT can not exploit its whole potential on BigBird due to these factors. Contrary, CSHOT is better suited for the YCB data set than SCSH because of more distorted and incomplete object point clouds.

The geometric part of the CSHOT descriptor might introduce an additional drawback. Point clouds from RGB-D cameras are very noisy and the 352 dimensions representing the geometric cues as normal vector directions are too much disturbed by the noise in the data. On the W-RGB-D data set we see that setting the color parameters $n_c = 32$ and $h = 30$ to match CSHOT[2] and using $n_g = 96$ already matches the CSHOT performance. This indicates that for the purpose of object recognition from noisy point clouds the geometric part of CSHOT might have a negative effect.

VIII. SUMMARY

We have presented SCSH, a family of configurable color-shape descriptors based on CSHOT. SCSH possesses a similar keypoint matching quality as CSHOT, while having a significantly lower memory footprint, making it suitable for robotic applications. Using our proposed descriptors in a object recognition pipeline can match (W-RGB-D) or even outperform (BigBird) the CSHOT baseline. While SCSH could

[2]CSHOT has actually $h = 31$, the difference is neglected here.

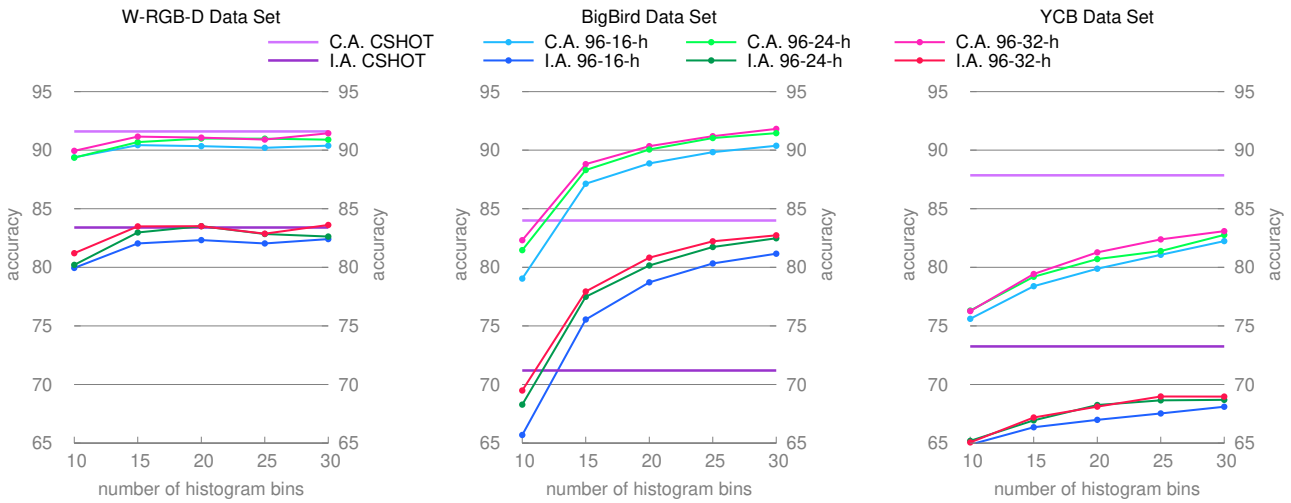

Fig. 5. Accuracies obtained with different configurations of the SCSH descriptor. All depicted configurations were evaluated with a fixed number of geometric subdivisions ($n_g = 96$). The different graphs show a varying number of color subdivisions n_c. The configurations are denoted as n_g-n_c-h, with h, the number of histogram bins per color subdivision, depicted on the axis of abscissae. Abbreviations: class accuracy (C.A.), instance accuracy (I.A.).

not match the CSHOT accuracy on the YCB data set, this work is the first to report a classification accuracy on the YCB data set, providing a baseline for the RoboCup@Home competition.

References

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[2] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3d shapenets: A deep representation for volumetric shapes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.

[3] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 1817–1824.

[4] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel, "Bigbird: A large-scale 3d database of object instances," in *2014 IEEE international conference on robotics and automation*. IEEE, 2014, pp. 509–516.

[5] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*. IEEE, 2015, pp. 510–517.

[6] F. Tombari, S. Salti, and L. Di Stefano, "A combined texture-shape descriptor for enhanced 3d feature matching," in *Proc. of the International Conference on Image Processing (ICIP)*. IEEE, 2011, pp. 809–812.

[7] L. 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 International Conference on Intelligent Robots and Systems (IROS), Vilamoura, Portugal*, vol. 1, no. 3, 2012, p. 7.

[8] N. Zrira, F. Z. Ouadiay, M. Hannat, E. H. Bouyakhf, and M. M. Himmi, "Evaluation of pcl's descriptors for 3d object recognition in cluttered scene," in *Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*, 2017, pp. 1–6.

[9] F. Tombari, S. Salti, and L. Di Stefano, "Unique signatures of histograms for local surface description," in *Proc. of the European Conf. on computer vision*, ser. ECCV'10. Springer-Verlag, 2010, pp. 356–369.

[10] V. Seib and D. Paulus, "A low-dimensional feature transform for keypoint matching and classification of point clouds without normal computation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2949–2953.

[11] M. Matamoros, A. Moriarty, J. Hart, and H. Okada, "Robocup@home 2020: Rules and regulations (draft)," https://athome.robocup.org/wp-content/uploads/2020_rulebook.pdf, 2020.

[12] R. Hanten, P. Kuhlmann, S. Otte, and A. Zell, "Robust real-time 3d person detection for indoor and outdoor applications," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2000–2006.

[13] B. Lewandowski, J. Liebner, T. Wengefeld, S. Müller, and H.-M. Gross, "Fast and robust 3d person detector and posture estimator for mobile robotic applications," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4869–4875.

[14] R. B. Rusu, N. Blodow, and M. 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.

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

[16] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok, "A comprehensive performance evaluation of 3d local feature descriptors," *International Journal of Computer Vision*, vol. 116, no. 1, pp. 66–89, 2016.

[17] J. Yang, Y. Xiao, and Z. Cao, "Toward the repeatability and robustness of the local reference frame for 3d shape matching: An evaluation," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3766–3781, 2018.

[18] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[19] S. M. Prakhya, B. Liu, and W. 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.

[20] V. Seib, N. Theisen, and D. Paulus, "Boosting 3d shape classification with global verification and redundancy-free codebooks." in *VISIGRAPP (5: VISAPP)*, 2019, pp. 257–264.

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

[22] K. B. Logoglu, S. Kalkan, and A. Temizel, "Cospair: colored histograms of spatial concentric surflet-pairs for 3d object recognition," *Robotics and Autonomous Systems*, vol. 75, pp. 558–570, 2016.

[23] C. Li, A. Reiter, and G. D. Hager, "Beyond spatial pooling: fine-grained representation learning in multiple domains," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4913–4922.

[24] V. Seib, R. Memmesheimer, and D. Paulus, "Ensemble classifier for joint object instance and category recognition on rgb-d data," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 143–147.