Applicability of Deep Learned vs Traditional Features for Depth Based Classification

Fabio Bracci, Mo Li, Ingo Kossyk and Zoltan-Csaba Marton

Institute of Robotics and Mechatronics, German Aerospace Center (DLR), Weßling, Germany
{fabio.bracci**, mo.li, zoltan.marton}@dlr.de, inkoss74@gmail.com

Abstract. In robotic applications often highly specific objects need to be recognized, e.g. industrial parts, for which methods can't rely on the online availability of large labeled training data sets or pre-trained models. This is especially valid for depth data, thus making it challenging for deep learning (DL) approaches. Therefore, this work analyzes the performance of various traditional (global or part-based) and DL features on a restricted depth data set, depending on the tasks complexity. While the sample size is small, we can conclude that pre-trained DL descriptors are the most descriptive but not by a statistically significant margin and therefore part-based descriptors are still a viable option for small but difficult 3D data sets.

Keywords: 3D shape descriptor, point cloud descriptor, deep learning features, object recognition, scene analysis

1 Introduction - Research Question

Most robotics applications involve some degree of perception of the environment and with the availability of inexpensive imaging sensors delivering depth data, robotic vision takes a major role in such applications. This in turn mandates processing techniques for 3D data. One of the required capabilities is 3D object classification. Given some image measurement taken by the robotic unit, the question arises, what kind of objects and items can be detected in the measured 3D data.

This is often done through the concept of point clouds, a collection of points representing a surface, in this case of an object. Objects can be differentiated by their geometric properties, which are encoded by means of several descriptors. In the past two decades a number of handcrafted descriptors have been studied in the literature. Recently from the field of Neural Networks the so called Deep Learning techniques gained momentum, some of which are able to learn descriptors suitable for the same purpose.

We want to investigate the descriptiveness of traditional point-based features as well as of pre-trained or trained (on a specific data set) deep learned features

** Alternative Email: fabio.bracci@freenet.de

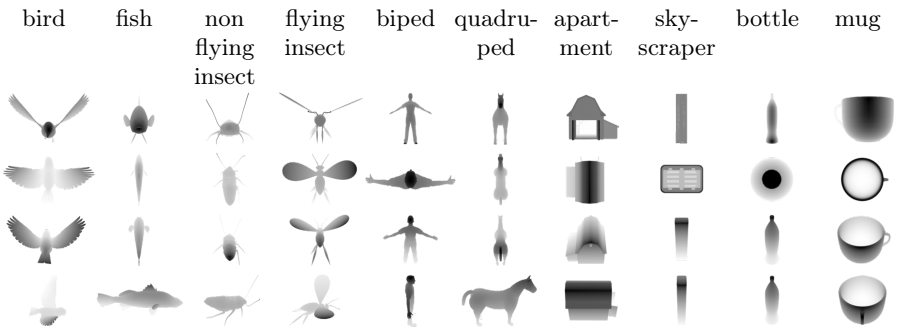


Fig. 1. Example Object Views. The rows in the table are from top to bottom: front view, top view, rear inclined view, right inclined view. The objects are taken form the SHREC 2010 object database.

for 3D object recognition, depending on the shape complexity. Considering the growing popularity of Deep Neural Networks (DNN), we want to evaluate if the feature sets extracted from a DNN are always the best choice for describing 3D object shape. We also want to investigate how the considered features perform when the shape differences are at the coarse level and when those are at the fine grain level.

A general problem for DNNs is the huge quantity of data needed for the training phase. This impedes the deployment of DNNs in robotics as large 3D data sets are scarce (there are few public data sets available with a medium-small number of objects) and within robotic setups to collect large samplings is usually not possible. Therefore, we want to perform this analysis in a context of limited data, as often it is the case in robotics applications.

2 Method

In our approach we consider view-based object classification. That is, given a 2.5D surface representation of an object view, we want to infer its class by finding the most similar surface in a collection of known ones. Those surfaces can come from real world observations as well as synthetic sources. The shape of the objects is encoded through traditional descriptors (VFH [14], CVFH [2] and OUR-VFH [1]) and deep learned ones (CaffeNet pre-trained features [5], VAE learned features [6] and DLR-VAE learned features [7]). The considered features are resumed in table 1.

Traditional 3D Descriptors

The Viewpoint Feature Histogram (VFH, [14]) is an evolution of the Fast Point Feature Histogram (FPFH, [13]) which keeps scale invariance and adds viewpoint variance. The VFH encodes a surface patch by means of a histogram of pan, tilt,

Table 1. Feature Types Overview

	handcrafted features	part-based	pre-trained	fully trained	supervised training
VFH	✔				
CVFH	✔	✔			
OUR-CVFH	✔	✔			
CaffeNet			✔		
VAE				✔	
DLR-VAE				✔	✔

yaw and surface normal angle relative to a given viewpoint vector. The VFH describes the whole object with a normalized histogram of 263 (45,45,45,128) bins (45 bins for pan, tilt and yaw angles, 128 bins for the surface normals angles with the viewpoint vector) and qualifies itself as a global descriptor.

The Clustered Viewpoint Feature Histogram (CVFH, [2]) is based on object parts, the so called stable regions. Those regions are obtained by applying on the point cloud a region growing algorithm with a maximum point distance and normal angle difference, as well as a minimum point number to accept the region. The stable regions are meant to represent the object and are robust against occlusions and missing parts and an example of such regions is shown in figure 2. The CVFH describes each region by extending the VFH vector with an unnormalized histogram of 45 bins, the Shape Distribution Component (SDC), computed by accumulating the quadratic point distances to the region centroid. For hypothesis verification, the Camera Roll Histogram (CRH) is added, a 90-bins histogram of the relative angles between the region normals and the camera view-up vector. This last histogram is useful for pose estimation problems. For our purposes, we regard the CVFH descriptor as a set of extended VFH descriptors, one for each part with 308 (45,45,45,45,128) bins and with the total descriptor size depending on the number of parts.

The Oriented, Unique and Repeatable Clustered Viewpoint Feature Histogram descriptor (OUR-VFH, [1]) is an evolution of the CVFH descriptor, where the CRH component is removed and the normal angles histogram is halved as the normals always point to the hemisphere where the viewing point lies. The authors propose Semi-Global Unique Reference Frames (SGURF), a method to compute a reference frame for every patch. With this reference frame they subdivide the region's points in octants, and for each they compute a 13-bins histogram of the distances to the centroid which gives a total histogram size of 104 bins The final descriptor consists of 303 (45*3,13*8,64) bins.

Both CVFH and OUR-VFH rely on the stable regions which represent parts of the analyzed objects and this qualifies them as part-based descriptors. All these methods were developed for fully 3D point clouds.

Deep Learned Descriptors

DNNs enjoy a growing popularity but require large amounts of training data and depth information requires more complex approaches [9,19], or involves transforming it into RGB images [19,17]. For example, in [17] transfer learning was attempted, where an RGB-based Places-CNN was fine-tuned with depth data using a proposed HHA embedding. On depth data this approach was found less effective than training from scratch.

We consider three kinds of DNNs, a pre-trained one and two fully trained ones, which synthesize global descriptors. The use of a pre-trained Convolutional Neural Network (CNN) to extract features is well known. For example, in [11] such features are benchmarked and deep learning models are publicly accessible since the high performances shown in the ImageNet Large Scale Visual Recognition Challenge [12], for instance, the networks studied in [5] and in [16].

The first DL descriptors are synthesized with the well known CaffeNet pre-trained network presented in [5]. There fully-connected Neural Networks are trained on RGB images subdivided in 1000 classes coming from ImageNet [8]. To one side CNN's synthesize representations which are increasingly abstract with the layer depth, to the other side the deeper layers tune the abstraction to the training problem. The challenge is to find the layer with an abstract representation which generalizes to a different problem, like the depth images which are mono-dimensional and untextured. We extract the features produced at the fc6 layer, since it was shown that it is the the best level of abstraction on depth data in the CaffeNet pre-trained network as shown in [18]. Because of the limited amount of training data, no fine-tuning is possible for this network.

The second kind of DL descriptors are fully trained and come from the Variational Auto Encoder (VAE) [6]. We trained this net on our unlabeled images data set and we extracted the features from the learned latent representation.

The last kind of DL descriptors is fully trained as well and come from a new flavor of VAE, the DLR-VAE [7]. Here the training is similar to the VAE's one, where the DLR-VAE considers also class labels and is therefore a supervised learner.

The last two DNNs need a smaller amount of data compared to CaffeNet, and the DLR-VAE is semi-supervised and better able than the original VAE to learn from small data sets, as shown in [7]. In our study, the training data is very limited and therefore challenging for both of the VAE methods. These limitations might lead to an advantage for the hand-crafted descriptors.

Descriptors and Metrics

This work focusses on a simple nearest neighbor classifier (1NN) because we want to focus on the descriptiveness of the features instead of on the tuning of various classification techniques from the vast machine learning field. The nearest neighbor classifier compares the feature vectors based on a given metric. 1NN prefers data which is cleanly clustered into classes in the feature space and is sensitive to class overlap, therefore it reveals feature weaknesses.

Fig. 2. Example of Stable Regions in CVFH and OUR-CVFH. Each color encodes a different stable region.

The VFH features in [14] are compared with a simple Euclidean distance. In [2] the CVFH are compared with a histogram metric defined as the sum of bin intersections divided by sum of bin unions. With this metric they retrieve the nearest ten candidates and find geometrically the best match by Iterative Closest Points (ICP, [3]) alignment and inliers count. This last stage is known as hypothesis verification.

The OUR-CVFH features are compared in a similar way to the CVFH: the candidates number is raised to fourteen and the hypothesis verification is done with inliers and outliers count. Both approaches are not suitable to our 1NN classification scheme because of the hypothesis verification stage; for each surface patch both approaches represent the geometry with a set of multidimensional features (VFH histograms) instead of a single multidimensional feature (a single VFH histogram).

Generally the Euclidean metric is applied on DNN features, while different metrics are recommended for the traditional descriptors. In order to compare properly all the features we consider a set of metrics, namely the L1, L2, Hellinger and the χ^2 distance. However, CVFH and OUR-CVFH are based on the segmentation in stable regions which are mapped to a set of descriptors and cannot be compared trivially, hence a scoring scheme is needed. The average vector, minimum distance and weighted confidence voting were considered; we report only the best performing, which in our setup is the weighted voting with a Gaussian distance scoring. For the DL descriptors, only the L1 and L2 distances were considered, Since the Hellinger and the χ^2 distances are meant for histograms and distributions we considered instead the cosine distance, which is the distance used by the DLR-VAE and the VAE learners.

3 Experimental Setup

The descriptiveness of the studied features is evaluated by measuring the retrieval accuracy on a subset of objects taken from the SHREC 2010 database - the Shape Retrieval Contest of Range Scans [10]. The subset consists of 10 objects taken from each of 11 object classes and for every object we take four selected views

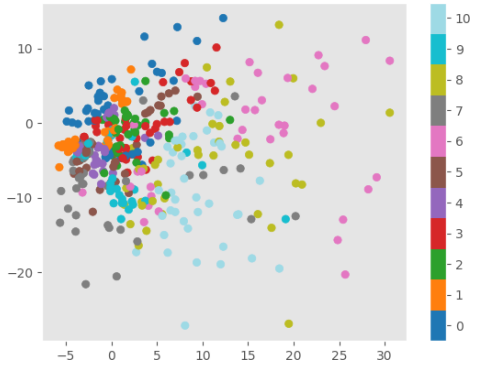


Fig. 3. DLR-VAE 2D Latent Space visualization. The eleven classes are encoded by different colors.

(front, top, top-rear and top-left) for a total of 440 labeled depth images. This setup comes from [4]; some examples of surfaces are shown in figure 1.

For simplicity we evaluate the considered features with a simple 1NN classifier while more advanced classifiers than 1NN are better able to reach high accuracies. This would happen at the cost of shifting the focus from feature quality to classifier quality. We choose 1NN in order to highlight the shape descriptiveness of the considered features as explained in subsection 2.

Each pair of surfaces is compared with the metrics explained in sec. 2. We perform a leave-one-out cross-validation and we collect the classification confusion matrices as well as the average accuracies. This procedure is applied to the the whole set of 440 surfaces from 11 classes as well as to two subsets: distinct objects {Biped, Bird, Quadruped, Fish, Mug} where objects are distinguishable at a coarse level through few details and similar objects {Apartment House, Flying Insect, Non Flying Insect, Single House, Skyscraper} where objects are distinguishable at a fine grained level with larger amount of details. In this second subset we have semantically related classes. The goal here is to highlight differences in the feature descriptiveness for such cases. Each of those subsets is made of 5 classes for which ten instances with the before mentioned views are considered. The 11 classes set is the union of the distinct objects set with the similar objects set and the class Bottle.

The methods are implemented with the commonly available Point Cloud Library [15]. For both the VAE and DLR-VAE learners we used the whole 440 surfaces. This is a small training data set for learning methods in general and for DNNs in particular, as such methods may overfit or fail to reach their maximal learning potential. Because of the data scarcity for robotic applications this is a relevant setup.

4 Results

We performed the 1NN classification for all the considered features and we collected the outcomes. Figure 5 shows the confusion matrices for the 1NN classification of the considered features for all classes with the metric giving the best accuracy. Figure 6 shows the same for the distinct classes and Figure 7 for the similar classes. An overview of the best classification accuracies is given in figure 4 and the results for the classification of the three objects sets are listed in table 2, 3 and 4. The 1NN classification was performed based on the 100 dimensional latent space learned by the DLR-VAE and the learned 2D latent space representation is shown in figure 3. Finally, a 2D projection of the CaffeNet features and the VFH features of the 440 patches is shown in figure 8.

On average, the classification of the similar objects shows a lower accuracy for all the used feature-metric pair, than the classification of the distinct objects. The classification accuracies of all the objects are in between the accuracies of the classification of the similar objects and distinct objects groups, except in the case of VAE features. The spread among the accuracies also varies with the classification problem: it is the smallest on the distinct classes, the largest on all classes and almost as large as on all classes on the similar classes. The CaffeNet features with L2 metric show the best 1NN classification accuracy, closely followed by CVFH features with L1 metric and OUR-CVFH features with the χ^2 metric. The VFH features with the L1 metric, the VAE and DLR-VAE features with the cosine metric follow with an accuracy gap on the all classes problem and are almost indistinguishable from each other.

The inspection of the confusion matrices reveals some specific features. In the similar objects case, all the global methods confuse Apartment House for Single House, Flying Insect for Non Flying Insect and vice versa, with the degree of confusion varying with the total accuracy. The VAE based methods tend to confuse a single Non Flying Insect and a Flying Insect for Apartment House and Skyscraper. For the part-based features Flying Insect is mistaken for Apartment House and Single House for Non Flying Insect. The VAE features confuse Skyscraper and Single House for Non Flying Insect.

In the distinct objects case the overall accuracies are higher and the scatter in the confusion matrices is lower. With global descriptors except the CaffeNet ones Biped objects are confused for Quadruped objects and, including the CaffeNet features, the Quadruped objects for Biped objects and Bird objects and Fish objects for Quadruped objects. For both the VAE features, Biped objects are confused for Fish objects.

In the classification problem for all classes, we see that the confusions observed in the two subsets translate to the total classification, with the addition of some cross set confusion, where objects from the distinct classes are confused for objects of the similar classes, and vice versa. We observe that both the part based features confuse Bird and Biped objects for Non Flying Insects objects as well as Bottle objects for Mug, Quadruped, Apartment House, Single House and Skyscraper objects. The CaffeNet features confuse Bottle objects for Single House and Skyscraper objects, Skyscraper objects for Bottle objects, as well as Flying

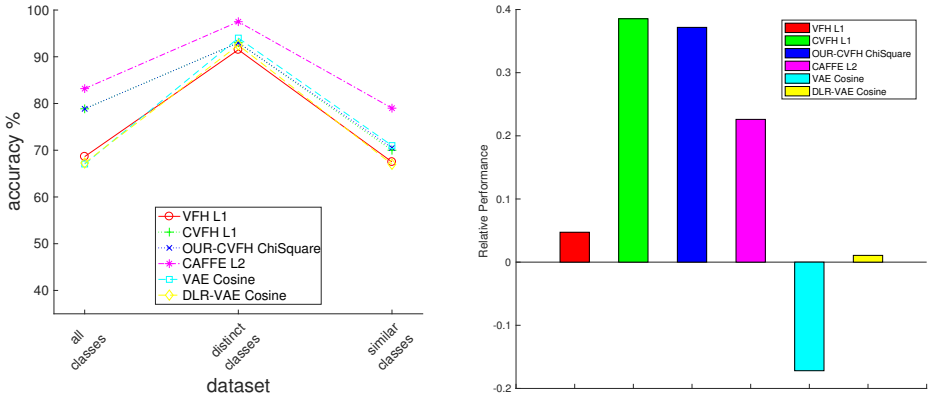


Fig. 4. Classification results. Left: Classification results. the vertical axis shows the classification accuracy, the horizontal axis shows the classification problem. Right: Relative accuracy plot. The horizontal axis represents the different combination of feature and metric, the vertical one represents the accuracy on the complete data relative to the two data sets. See the main text for details.

and Non Flying Insect objects for Biped, Bird and Fish objects. The VAE features show confusion for Bottle objects mistaken for Mug, Quadruped, Apartment House, Single House and Skyscraper objects, Skyscraper objects for Bottle objects, and both Flying and Non Flying Insect objects mistaken for objects of almost all the other classes.

We also estimate the stability of these performances by relating the accuracy for all objects with the ones for distinct objects and for the similar objects. We estimate the following score: $(acc_{all} - acc_{similar}) \,/\, (acc_{distinct} - acc_{similar})$. The obtained ratios are shown in in figure 4.

Finally we analyzed the Jeffreys intervals for the best accuracies reported in table 2 according to [7]. The 95% credible interval for the best performing CaffeNet features with L2 metric is [79.5% 86.5%], for the next best performing CVFH features with L1 metric and the OUR-CVFH features with χ^2 metric is [74.8% 82.5%], while for the following VFH features with L1 metric is [64.2% 0.72.8%]. The first two intervals are not disjoint while the last two are disjoint, therefore the accuracy difference between the CaffeNet features and the CVFH and OUR-CVFH features is statistically not significant while the accuracy difference between the second best performing features and the third best VFH features is statistically significant.

Table 2. Classification accuracy percentages on all classes. Features' best performance is marked in bold, second best in italics.

	VFH	CVFH	OUR-CVFH	CaffeNet	VAE	DLR-VAE
L1	**68.64%**	**78.86%**	77.73%	*82.95%*	52.05%	58.64%
L2	59.82%	74.09%	*78.18%*	**83.18%**	*57.73%*	*60.00%*
χ^2	66.82%	*78.41%*	**78.86%**	-	-	-
Hellinger	*67.05%*	76.36%	**78.86%**	-	-	-
Cosine	-	-	-	-	**67.04%**	**67.27%**

Table 3. Classification accuracy percentages on similar classes. Features' best performance is marked in bold, second best in italics.

	VFH	CVFH	OUR-CVFH	CaffeNet	VAE	DLR-VAE
L1	**67.5%**	*70.0%*	69.0%	**79.0%**	60.5%	52.0%
L2	58.5%	*70.0%*	69.0%	**79.0%**	*62.5%*	*59.5%*
χ^2	*67.0%*	**71.5%**	**70.5%**	-	-	-
Hellinger	*67.0%*	69.0%	*69.5%*	-	-	-
Cosine	-	-	-	-	**71.0%**	**67.0%**

Table 4. Classification accuracy percentages on distinct classes. Features' best performance is marked in bold, second best in italics.

	VFH	CVFH	OUR-CVFH	CaffeNet	VAE	DLR-VAE
L1	**91.5%**	**93.0%**	**93.0%**	**98.5%**	89.5%	82.5%
L2	86.0%	91.0%	90.0%	*97.5%*	*90.5%*	*90.0%*
χ^2	*90.5%*	*92.5%*	**93.0%**	-	-	-
Hellinger	90.0%	92.0%	*92.5%*	-	-	-
Cosine	-	-	-	-	**94.0%**	**92.5%**

5 Conclusions

This study focuses on a 3D object classification scenario with limited data, which is a plausible scenarion in robotics applications. The classification of 3D data, whether it being point clouds or other sparse and volumetric data, is a subject of current research and yet to be fully solved.

First of all, we see that the performance of the CaffeNet features is best and the VAE features is worst. The 2D projection of the feature data with the t-SNE method as shown in figure 8 qualitatively indicates that the CaffeNet features group into separated clusters, while the lower quality VFH ones result in a more scattered distribution with many overlapping regions between different classes.

Second, we notice that the similar objects are more difficult to discriminate than distinct objects, as expected. Also we note that semantically similar objects here imply a certain degree of geometrical similarity, which in turn implies sim-

Fig. 5. Confusion Matrices for all classes

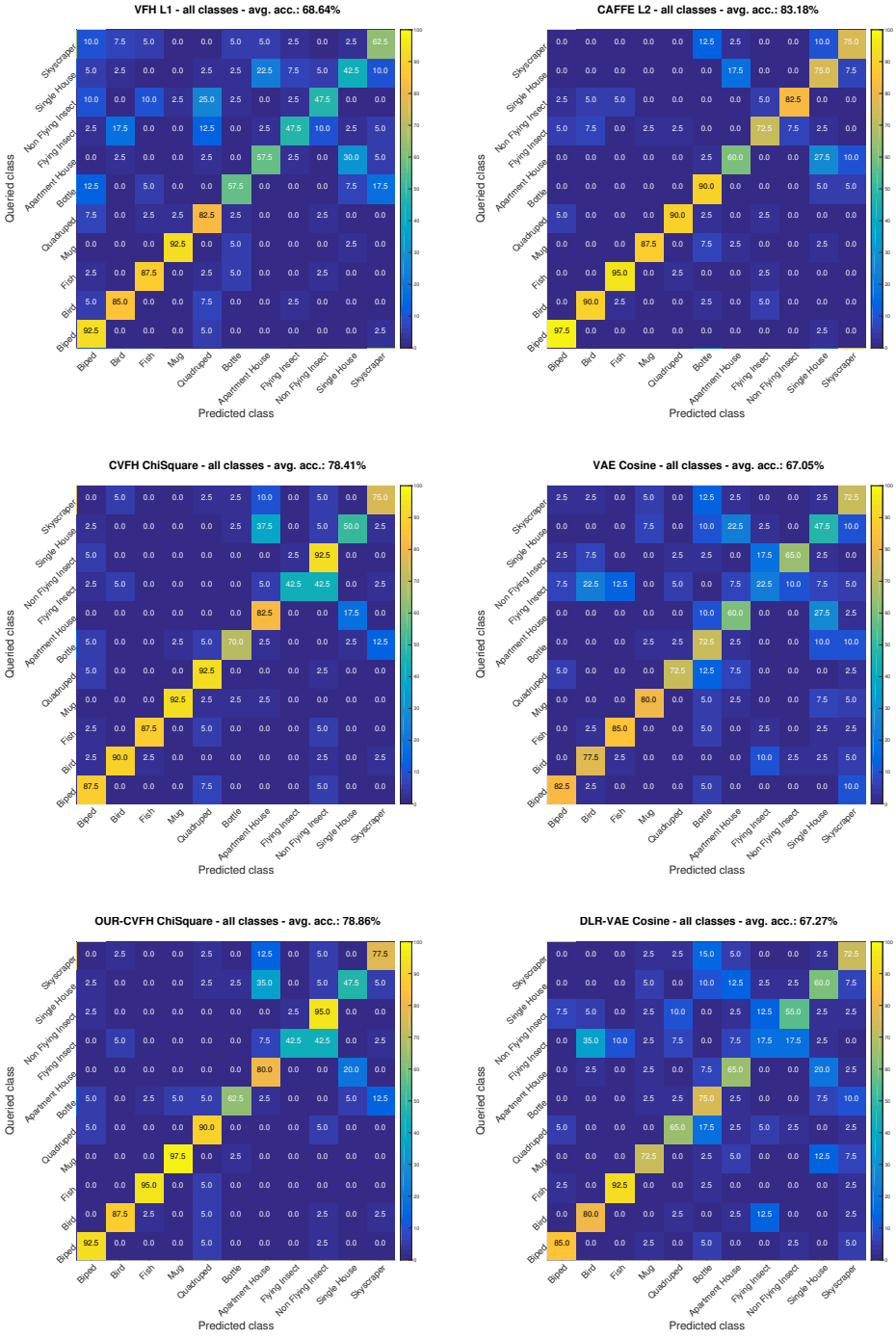
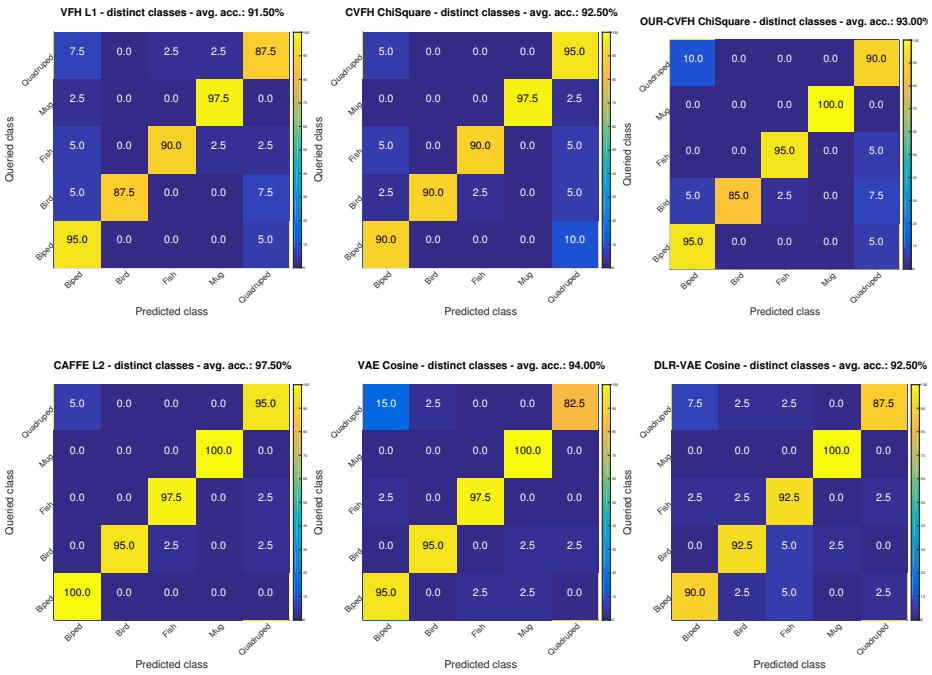


Fig. 6. Confusion Matrices for the distinct classes



ilarity between the final features. This way, semantic ambiguity among classes translates into low feature discriminability.

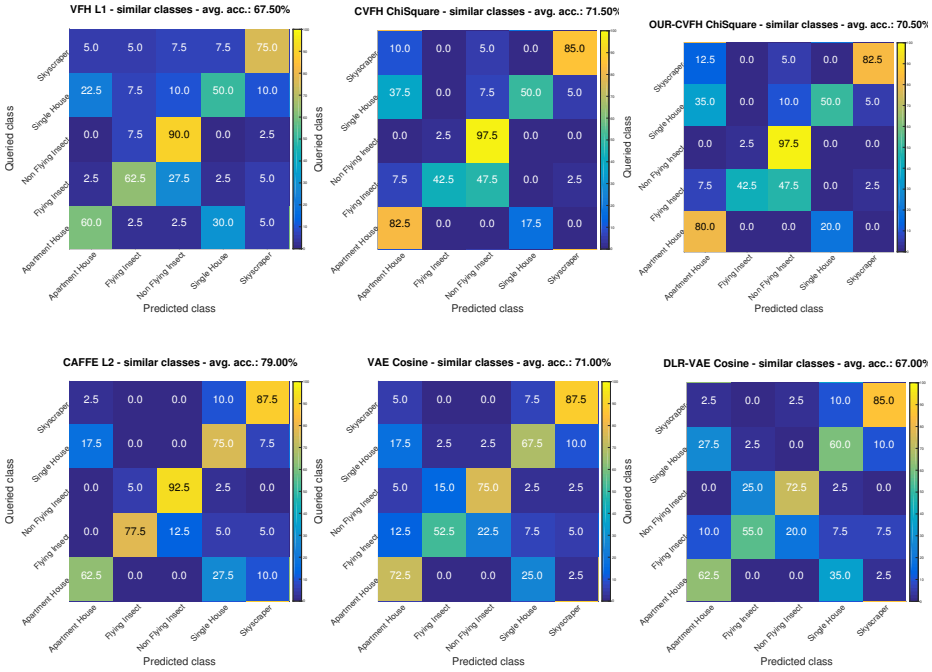
Third, as expected the part based methods performances are very correlated, as CVFH and OUR-CVFH are both based on the same concept of stable regions.

Further, in this study the CaffeNet features perform best for discriminating the considered objects, and the second best features by a small accuracy difference are the part-based methods. We also showed how the accuracy difference on all objects classification between the CaffeNet features and the part-based features is not statistically significant. This comparable performance obtained using part-based descriptors and a large (pre-trained) CNN holds for depth data, where comparatively less information is encoded than in RGB data. Depth images are usually smoother and have less high-frequency information.

Moreover, the VAE descriptors perform comparably to the DLR-VAE descriptors. We believe that unsupervised pre-training and data augmentation might help building a more robust embedding, where the DLR-VAE's supervised training might provide an improvement for classification tasks.

On one hand, the CaffeNet features are likely to benefit from the extensive pre-training and work better than the VAE and DLR-VAE methods trained on this limited surfaces set. This might be improved by a larger (unlabeled) pre-training data set, as the minimal training requirements for both VAE methods

Fig. 7. Confusion Matrices for the similar classes



need to be assessed. On the other hand, despite the clear disadvantage constituted by such a limited training data set, both the VAE methods perform similarly to each other and to the VFH descriptor.

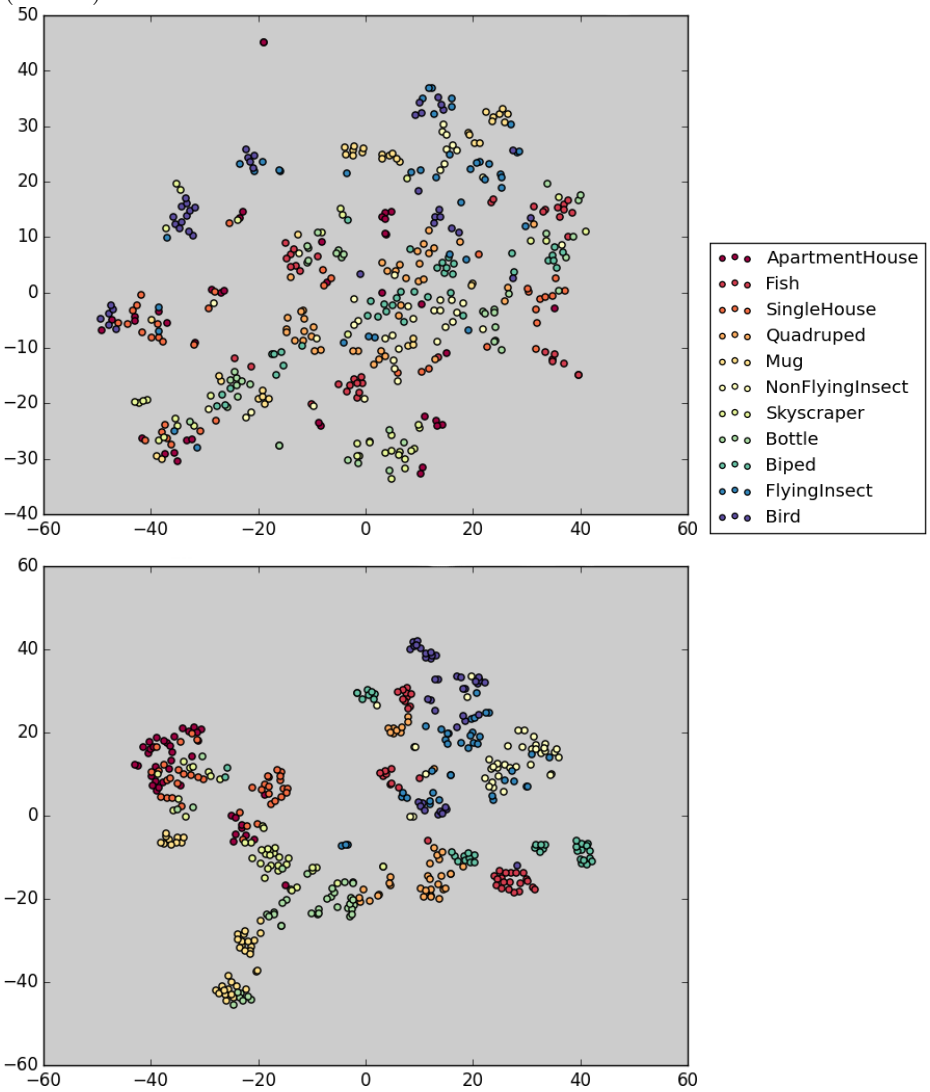
Finally, the bar plot in figure 4 shows that when more similar objects are added to the classification problem, in terms of discriminative power the degradation for the part based method is the least, the CaffeNet features are second best and the global methods are the worst, with the performance dropping near or below the accuracy of the similar objects classification. All in all for limited data scenarios the use of part-based descriptors remains an option to consider.

In the future we plan to investigate more complex part based approaches where part relationships are encoded, as well as the use of DNN features within the part-based methods. Also other classifiers like k-Nearest-Neighbors (kNN) could be investigated, where for this specific classifier we don't intuitively expect to see different trends in the accuracies, rather only higher values.

References

1. Aldoma, A., Tombari, F., Rusu, R., Vincze, M.: OUR-CVFH - Oriented, Unique and Repeatable Clustered Viewpoint Feature Histogram for Object Recognition and 6DOF Pose Estimation. In: Pinz, A., Pock, T., Bischof, H., Leberl, F.

Fig. 8. 2D projection using t-SNE of the VFH features (top) and the CaffeNet features (bottom).



(eds.) Pattern Recognition, Lecture Notes in Computer Science, vol. 7476, pp. 113–122. Springer Berlin Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-32717-9_12

2. Aldoma, A., Vincze, M., Blodow, N., Gossow, D., Gedikli, S., Rusu, R.B., Bradski, G.: CAD-model recognition and 6DOF pose estimation using 3D cues. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). pp. 585–592. IEEE (Nov 2011), http://dx.doi.org/10.1109/iccvw.2011.6130296

3. Besl, P.J., McKay, N.D.: A method for registration of 3-D shapes. Pattern Analysis and Machine Intelligence, IEEE Transactions on 14(2), 239–256 (Feb 1992), http://dx.doi.org/10.1109/34.121791
4. Bracci, F., Hillenbrand, U., Marton, Z.C., Wilkinson, M.: On the Use of the Tree Structure of Depth Levels for Comparing 3D Object Views. In: Felsberg, M., Heyden, A., Krüger, N. (eds.) Computer Analysis of Images and Patterns, Lecture Notes in Computer Science, vol. 10424, pp. 251–263. Springer International Publishing (2017), http://dx.doi.org/10.1007/978-3-319-64689-3_21
5. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia. pp. 675–678. MM '14, ACM, New York, NY, USA (2014), http://doi.acm.org/10.1145/2647868.2654889
6. Kingma, D.P.: Variational inference & deep learning: A new synthesis. Intelligent Sensory Information Systems (IVI, FNWI) (2017), http://dare.uva.nl/personal/pure/en/publications/variational-inference–deep-learning(8e55e07f-e4be-458f-a929-2f9bc2d169e8).html
7. Kossyk, I., Marton, Z.S.: Discriminative regularization of the latent manifold of variational auto-encoders for semi-supervised recognition. Online (2017), https://tinyurl.com/y8p3tjle
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. vol. 25 (2012), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.299.205
9. Masci, J., Rodolà, E., Boscaini, D., Bronstein, M.M., Li, H.: Geometric Deep Learning. In: SIGGRAPH ASIA 2016 Courses. SA '16, ACM, New York, NY, USA (2016), http://dx.doi.org/10.1145/2988458.2988485
10. Pratikakis, I., Spagnuolo, M., Theoharis, T., Editors, R.V., Dutagaci, H., Godil, A., Cheung, C.P., Furuya, T., Hillenbr, U., Ohbuchi, R.: SHREC 2010 - Shape Retrieval Contest of Range Scans http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.361.8068
11. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN Features off-the-shelf: an Astounding Baseline for Recognition (May 2014), http://arxiv.org/abs/1403.6382
12. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge 115(3), 211–252 (2015), http://dx.doi.org/10.1007/s11263-015-0816-y
13. Rusu, R.B., Blodow, N., Beetz, M.: Fast Point Feature Histograms (FPFH) for 3D Registration. In: The IEEE International Conference on Robotics and Automation (ICRA). Kobe, Japan (2009), http://files.rbrusu.com/publications/Rusu09ICRA.pdf
14. Rusu, R.B., Bradski, G., Thibaux, R., Hsu, J.: Fast 3D recognition and pose using the Viewpoint Feature Histogram. In: Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on. pp. 2155–2162. IEEE (Oct 2010), http://dx.doi.org/10.1109/iros.2010.5651280
15. Rusu, R.B., Cousins, S.: 3D is here: Point Cloud Library (PCL). In: 2011 IEEE International Conference on Robotics and Automation. pp. 1–4. IEEE (May 2011), http://dx.doi.org/10.1109/icra.2011.5980567
16. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (Apr 2015), http://arxiv.org/abs/1409.1556v5.pdf

17. Song, X., Herranz, L., Jiang, S.: Depth CNNs for RGB-D scene recognition: learning from scratch better than transferring from RGB-CNNs. ArXiv e-prints (Jan 2018), http://arxiv.org/abs/1801.06797
18. Ullrich, M., Ali, H., Durner, M., Marton, Z.C., Triebel, R.: Selecting CNN features for online learning of 3D objects. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 5086–5091. IEEE (Sep 2017), http://dx.doi.org/10.1109/iros.2017.8206393
19. Zaki, H.F.M., Shafait, F., Mian, A.: Convolutional hypercube pyramid for accurate RGB-D object category and instance recognition. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). pp. 1685–1692. IEEE (May 2016), http://dx.doi.org/10.1109/icra.2016.7487310