Proceedings of the 2011 IEEE
International Conference on Robotics and Biomimetics
December 7-11, 2011, Phuket, Thailand

Ensemble of Shape Functions for 3D Object Classification

Walter Wohlkinger and Markus Vincze

***Abstract*— This work addresses the problem of real-time 3D shape based object class recognition, its scaling to many categories and the reliable perception of categories. A novel shape descriptor for partial point clouds based on shape functions is presented, capable of training on synthetic data and classifying objects from a depth sensor in a single partial view in a fast and robust manner. The classification task is stated as a 3D retrieval task finding the nearest neighbors from synthetically generated views of CAD-models to the sensed point cloud with a Kinect-style depth sensor. The presented shape descriptor shows that the combination of angle, point-distance and area shape functions gives a significant boost in recognition rate against the baseline descriptor and outperforms the state-of-the-art descriptors in our experimental evaluation on a publicly available dataset of real-world objects in table scene contexts with up to 200 categories.**

I. Introduction

With recent steps towards affordable 3D sensing devices delivering real-time high quality 3D data, the opportunities and possibilities are now at hand to put these sensors to work. One specific area of interest we are focusing on is domestic robotics where personal robots should be able to find, grasp and manipulate objects found in human living space. These are clearly not specific object instances, but a large variety of object classes. Thus, the goal is to come up with methods to easily learn new categories and recognize them in real-time in the home environment.

Especially the domestic setting with its plethora of categories and their intraclass variance demands strong generalization skills from a vision system. These categories are mostly characterized by their shape, ranging from low intraclass variance as in the case of fruits and simple objects like bottles up to high intraclass variance of classes such as liquid containers, furniture, and especially toys. In robotic manipulation where object recognition and object classification have to work from all possible viewpoints of an object, data collection for training becomes a bottleneck. We therefore use a large database of synthetic 3D CAD models collected throughout the Internet to ease the training stage. The database is freely available at 3D-Net (3d-net.org).

Our contribution is the introduction of a novel shape descriptor dubbed ESF (Ensemble of Shape Functions) which is a global shape descriptor based on three distinct shape functions describing distance, angle and area distributions on the surface of the partial point cloud. The descriptor can be calculated efficiently on the raw point clouds data, scales to hundreds of classes, is able to cope with the differences in sensor characteristics and thus enables robust real-time classification.

This work was conducted within the EU Cognitive Systems project GRASP (FP7-215821) funded by the European Commission.

W. Wohlkinger and M. Vincze are with Vision4Robotics Group, Automation and Control Institute, Vienna University of Technology, Vienna, Austria [ww,vm]@acin.tuwien.ac.at

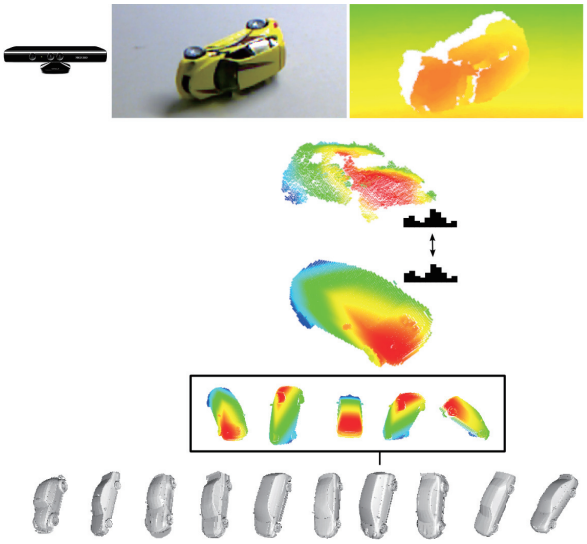


Fig. 1. System overview: For classification, the RGB-D image is segmented to obtain a point cloud cluster to calculate a descriptor on it. The descriptor is matched to synthetically rendered views of a database of CAD-models downloaded from the web, thus giving the most similar class, model and view.

After reviewing related work we briefly present the 3D model database and the synthetic view generation in Section III followed by a detailed explanation of the ESF descriptor in Section IV and present benchmarks and evaluation results against state-of-the-art descriptors in Section V.

II. Related Work

Shape Distributions were introduced as a pure 3D descriptor for content based 3D model retrieval by Osada [6] who found that the distribution of distances between two randomly chosen points on the surface of 3D meshes results in a robust shape descriptor. Osada proposed several other shape functions based on angles (A3), length (D1 and D2), areas (D3) and volumes (D4), but stated that distances between two randomly chosen points (D2) gives the best results and that a combination of the shape functions results in marginal improvements of only 2-3% classification rate. Although D2 shape distributions give a good overall shape description, using only the distance as descriptor is not descriptive enough to differentiate between classes once the number increases. In addition to shape distributions, coarse features are used in [3] to boost the classification performance and spherical harmonics [5] are used together

978-1-4577-2138-0/11/$26.00 © 2011 IEEE

with shape distributions in a voting scheme in [10] to get the needed performance. To increase the descriptiveness of the shape distribution itself, Ip [4] proposes an extension in the context of mechanical parts by classifying each line created by two randomly sampled points on the surface to be either inside the 3D model, outside or both (mixed). This extension boosts the descriptiveness, but is clearly only applicable to full 3D models.

To successfully use shape functions with the extension of line classification in the context of depth images delivered by RGB-D sensors like the Microsoft Kinect, the methodology has to be adapted to the new domain of partial point clouds by classifying lines as on/off the voxelized surface of the object. Furthermore, the combination of selected shape functions in this new context results in improved classification rates with more than 13% increase over using pure D2 shape functions.

III. 3D-Net Database

The database is created semi-automatically by downloading 3D models mainly from Google's 3D Warehouse and storing the models in a hierarchy according to WordNet [2] for further semantic linking. The database consists of 3876 CAD-models which cover more than 200 object categories found in home environments reaching from fruits and vegetables to furniture, tools, toys (such as vehicles, airplanes) and objects usually found in office and kitchen environments.

The training on these CAD models is done by rendering and sampling the z-buffer from views around each model and storing the generated partial views as point clouds. Descriptors are computed on these partial views. The number of views can be chosen from as few as 12 to several hundreds, depending on the descriptor and application in mind. The number of views used for the experiments in this paper is empirically set to 80 which provides sufficient views even for complex objects.

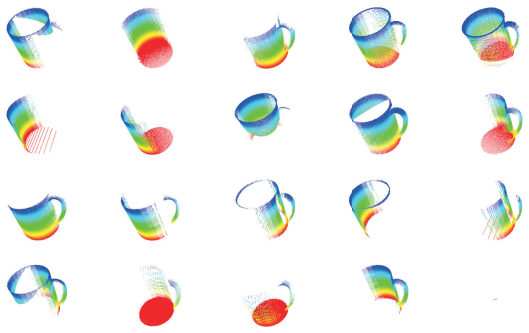


Fig. 2. Generation of the synthetic partial views point clouds of a mug CAD-model by sampling the depth-buffer while rendering views around the object.

The database and code for training on this database and the proposed descriptor are available at 3D-Net (3d-net.org) in our open-source PCL-based framework which is targeted at real-time classification and object instance recognition with pose estimation for robotics.

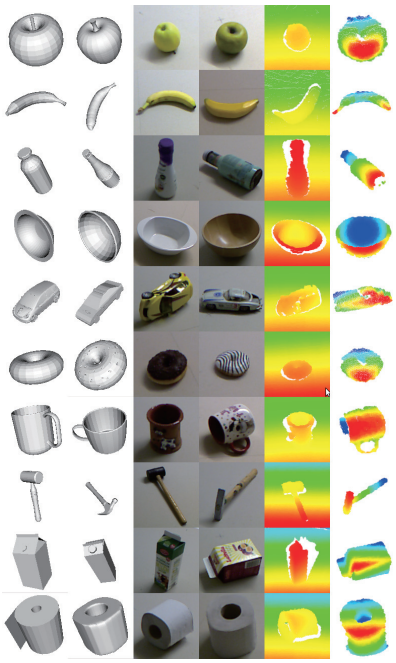


Fig. 3. Examples from the 10 test categories: First two colums show sample models from the CAD-database, middle colums show example images of the test database with example 3D depth data in the rightmost columns.

The 3D-Net database provides four databases with 10, 50, 100 and 200 categories together with evaluation databases. Figure 3 depicts CAD-models from 10 categories together with the evaluation scenes. As the CAD-models do not exactly match the objects in the evaluation scenes, the shape descriptor has to robustly cope with the variations between the synthetic data and the real-world data.

IV. Ensemble of Shape Functions

The ESF descriptor is an ensemble of ten 64-bin sized histograms of shape functions describing characteristic properties of the point cloud cluster. A sample ESF histogram of a banana is depicted in Figure 4 representing the combined histograms of three angle, three area, three distance and one distance ratio shape functions.

The D2 shape function, as introduced by Osada [6], is created by sampling point-pairs from the point cloud and forming a histogram of the distances between these two points. An illustration of the D2 shape function on a view of a mug is given in Figure 5.

D2 shape distributions can differentiate between rough shapes but once the number of classes increases or if, as in our case, only a partial view of the object is available, the descriptive power of this feature is not sufficient any more. In the context of 3D mechanical parts Ip [4] extended the D2 shape distributions by classifying each line created by two randomly sampled points to be either inside the 3D model, outside or a mixture of both. This approach is adapted to be applicable to partial point clouds. Lines between two randomly sampled points are classified to be on/off the surface of the point cloud. This is accomplished by tracing the line with the help of the 3D Bresenham algorithm

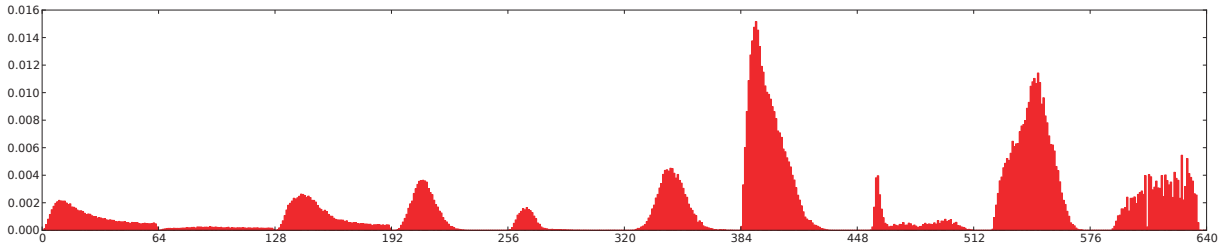


Fig. 4. The ESF descriptor calculated on a view of a banana with its ten 64-bin sub-histograms of shape functions. The shape functions from left to right: A3:Angle (in,out,mixed), D3:Area (in,out,mixed), D2:Distance (in,out,mixed) and ratio of line distances.

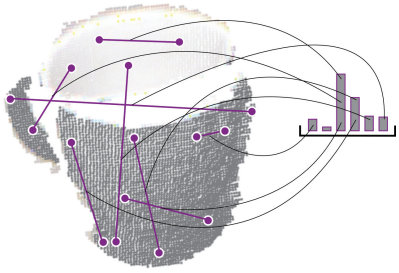


Fig. 5. Distances between randomly sampled points from the partial object view are put – with proper normalization – into a histogram which forms the D2 shape distribution descriptor.

in a voxel grid with side length of 64 as shown in Figure 6. This coarse voxel grid serves as an approximation of the real surface but can be created in a computationally efficient manner. After the tracing, the classified lines are put into three distinct histograms representing ON, OFF and MIXED distances as depicted in Figure 7.

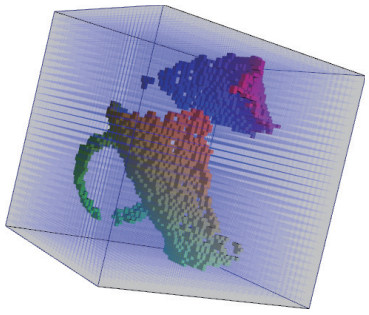


Fig. 6. A 64x64x64 voxel grid is created from the point cloud and used as a surface approximation to efficiently trace the line distances within.

To further exploit the information captured by the lines classified as MIXED, the ratios of these lines are used to construct another distinct shape histogram as depicted in Figure 8. The histograms of the ON, OFF, MIXED classified lines and the ratio of the lines form the rightmost four sub-histograms of the ESF descriptor as shown in Figure 4.

The A3 shape function encodes the angle enclosed by two lines created by randomly sampling three points from the point cloud. Figure 9 shows the construction of the A3 shape function. This shape function encapsulates a different aspect of the underlying object and therefore increases the descriptiveness of the overall descriptor. As with the D2

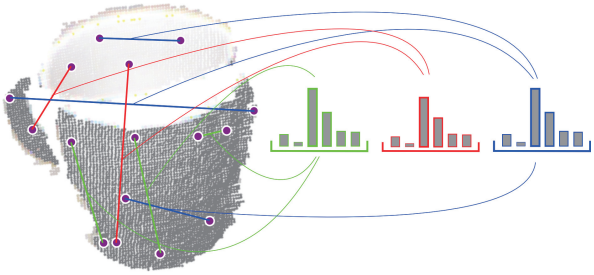


Fig. 7. The classified distances. Green represents the connecting lines lying ON the surface of the object, red represents the lines where only the endpoints are on the surface and the connecting line is OFF surface and the blue colored line distances are classified as MIXED as they are partly ON and OFF the surface.

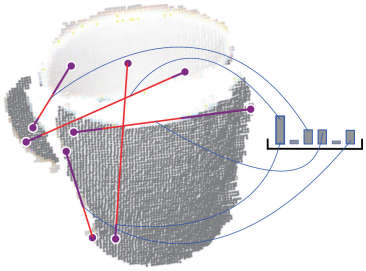


Fig. 8. The ratio of the lines lying in free space (red) and on the surface of the object (purple) are put into a separate shape histogram.

shape function, the descriptiveness is also increased by separating the angle-histograms into three distinct histograms representing the ON, OFF and MIXED angles. The classification is done by using the line opposing the angle to find the appropriate histogram for the angle, as depicted by the red dotted line in Figure 9. These three histograms can be found as the leftmost histograms in Figure 4.

The last three histograms are composed of the D3 shape function, which is the D2 shape function with one dimension increased, from distance of two points to area formed by three points. The same methodology of classifying the area as ON, OFF and MIXED is applied by using the three enclosing lines to calculate the appropriate histogram. The creation of the histogram with the D3 shape function is depicted in Figure 10. These three histograms form the fourth to sixth sub-histogram in the ESF descriptor shown in Figure 4.

The ESF descriptor can be efficiently calculated directly from the point cloud with no preprocessing necessary such as smoothing, hole filling or surface normal calculation

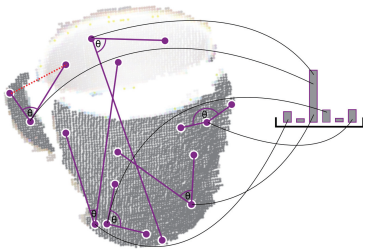


Fig. 9. The A3 shape function is calculated by taking the enclosed angle between two lines to create the shape histogram.

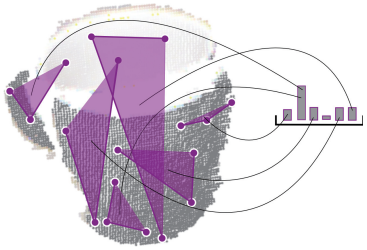


Fig. 10. The D3 shape function is build up by constructing a histogram of areas spanned by three randomly sampled points.

and handles data errors such as outliers, holes, noise and coarse object boundaries gracefully. As CAD-models are used and therefore no size information is available, the point cloud clusters are scaled into the unit sphere and the voxel cube is constructed from that. This methodology allows for an efficient implementation of the ESF descriptor leading to a calculation time of approximately 40 ms with a non-optimized single threaded C++ implementation. In our implementation the number of point-pair and point-triplet samples is set to 40000, resulting in a fixed calculation time of the descriptor, independent of the number of points in the point cloud.

Matching of these histograms is done with L1-distance.The distinct varying shapes of the ESF descriptor histograms for differing models can be seen in Figure 11 and for differing views and models from the same class in Figure 12.

A. *Weight Learning on Synthetic Views*

The influence of the ten sub-histograms of the ESF descriptor in the comparison is equally distributed in the initial shape descriptor. Given synthetic views of hundreds of models, setting the weights for the sub-histograms and thus optimizing the descriptor is possible. Parameters and descriptor weights can be learned on the synthetic views without having to see a single real scene. The improvement of the descriptor performance is showcased in Figure 13 where an overall improvement of 2 % and specific improvements of up to 13 % for certain classes are achieved by learning the descriptor sub-histogram weights on a small sample of the synthetic views. The advantage here is that this can be done offline, without having a test database of real objects to split into training and evaluation parts and thus can be done for new object classes before having real test data for them.

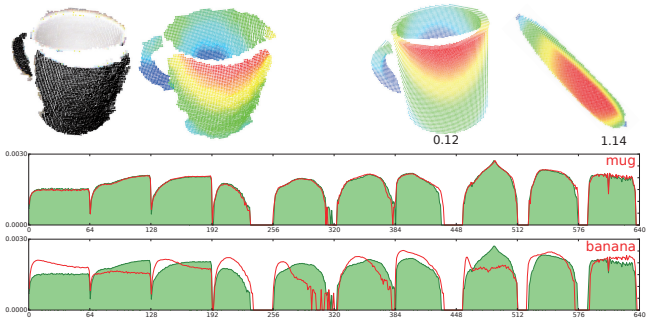
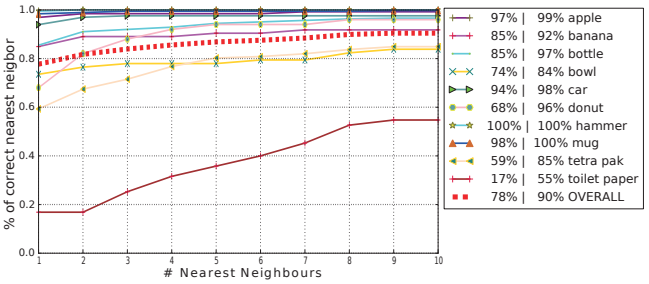
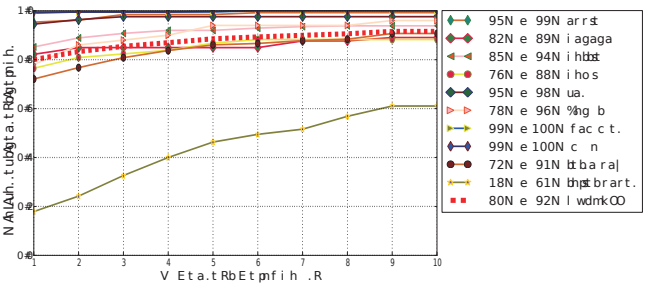


Fig. 11. The query object – the black mug, depicted as image on the left and as point cloud next to it – is compared to two sample database objects, a mug and a banana. The green histogram represents the query object, the overlayed red curves represent the two database objects. The similarity of the query object to the mug (0.12) is clearly visible as their histograms tightly overlap to a high percentage. The banana-view generates a completely different shape distribution histogram which results in a high dissimilarity score. For better viewing of the histogram shape differences, the histograms are plotted in log-scale.

The overall performance of the ESF descriptor can be seen in Figure 13 and in Figure 14. The performance on specific classes varies depending on the presence of similarly shaped classes, i.e. mug and toilet paper share some common shape. Similar partial views are also causing confusion, i.e. an apple seen from the top is similar to a donut, round, curved and a hole in the middle. Despite these challenging problems, the overall performance of the proposed ESF descriptor is superior to other 3D descriptors as presented in the next section.



(a) ESF descriptor with equal weights. Nearest neighbor classification with 78% overall performance and 90 % for ten nearest neigbor.



(b) ESF descriptor with learned weights from synthetic views shows an overall 2 % performance improvement, and up to 13 % for specific classes.

Fig. 13. Learning weights on synthetic views for increased classification. The ESF descriptor performs

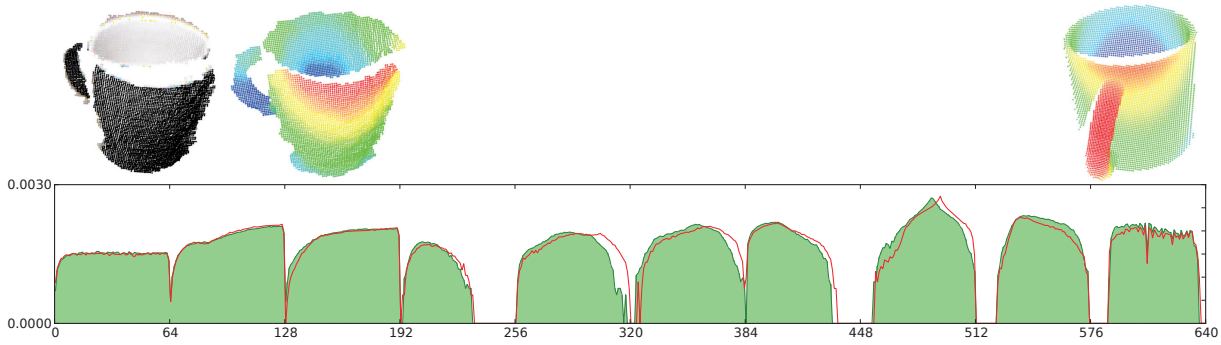


Fig. 12. The query object as in Figure 12 is compared to a different mug with a slightly different pose. The red curve resembles the green histogram only in a coarse manner as the classes are similar but the pose is different.

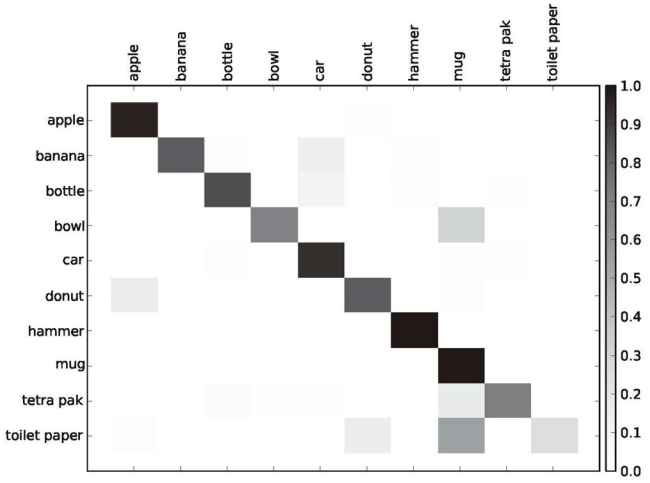


Fig. 14. The confusion matrix for the ESF descriptor shows the overall good separability of the classes, except for the toilet paper class, which is confused with the class mug due to its similar shape in some viewpoints.

V. BENCHMARK & EVALUATION

The benchmark consists of 1600 test scenes captured with a Kinect-style depth sensor representing multiple objects in various poses from 10 object categories. The target databases for nearest neighbor matching are our synthetic databases with 10 and 200 categories having 360 and 3800 models respectively with 80 views of each model.

As competing features the following global descriptors are used:

- SDVS: The Shape Distribution on Voxel Surfaces descriptor uses purely the D2 shape function and was introduced in [9]. This descriptor serves as our baseline descriptor and takes an average of 25 ms for calculation and matching, see Figure 15.
- VFH: The Viewpoint Feature Histogram is a descriptor based on normal vectors and was introduced in [7]. The average time for calculation and matching is approximately 70 ms, see Figure 16.
- CVFH: The Clustered Viewpoint Feature Histogram [1] is a semi-global view based descriptor based on VFH. Because of its semi-global nature, only certain parts of the objects are used to build the reference systems on which the computation is based but uses the whole available view information to build the angular normal distribution histograms. The average time for computation and search is approx. 208 ms, see Figure 17..
- SHOT: The SHOT descriptor introduced in [8] is aimed at surface matching with local descriptors, but is used here as a global descriptor for the whole object. The feature calculation and matching takes from 130 ms to 2 sec on our test database, see Figure 18.

The ESF descriptor achieves to retrieve 80% matching views in the first nearest neighbor and finds a correct view with 92% in the ten nearest neighbors as shown in Figure 13. Classifying the 1600 test scenes against 200 classes with ESF shows an overall good performance as presented in I and showcases the good scalability of the descriptor as the most confusing class still has a high similarity in shape, e.g. banana with pistol, car with SUV, hammer with axe. A supplemental video of classifying objects with the ESF descriptor is available on 3DNet (3d-net.org/video).

TABLE I
ESF NEAREST NEIGHBOR CLASSIFICATION RESULTS AGAINST 200 CLASSES WITH MOST CONFUSING CLASS ON THE RIGHT

class name	1-NN	10-NN	confusing class
per scenes OVERALL	58.22 %	78.23 %	
per class OVERALL	49.10 %	71.39 %	
apple	81.40 %	98.45 %	pumpkin
banana	54.79 %	69.86 %	pistol
bottle	48.77 %	79.01 %	suv
bowl	50.00 %	76.47 %	hat
car	11.52 %	43.64 %	suv
donut	20.00 %	62.00 %	cap
hammer	83.41 %	96.10 %	axe
mug	91.96 %	99.46 %	watch
tetra pak	47.09 %	72.09 %	mug
toilet paper	2.11 %	16.84 %	armchair

VI. CONCLUSIONS

A novel shape descriptor is introduced combining three shape functions into a high performing global shape descriptor aimed at real-time classification of objects sensed with a Kinect-style depth sensor while learned from synthetic CAD-models. The ensemble of angle, area and distance shape functions clearly results in a more discriminative descriptor able

to robustly retrieve similar shapes in a database containing more than 300000 synthetic partial views. Despite the good performance of the ESF descriptor, the full potential of shape functions in shape based retrieval is not yet fully exploited as additional specialized information such as normals, curvature and object contours have not been used with shape functions on depth data. This leaves a lot of scope for future improvements especially regarding occlusions and speed and we hope our freely available source code will help accelerating the development of even faster and better methods.

REFERENCES

[1] A. Aldoma, N. Blodow, D. Gossow, S. Gedikli, R.B. Rusu, M. Vincze, and G. Bradski. Cad-model recognition and 6dof pose estimation using 3d cues. *3rd IEEE Workshop on 3D Representation and Recognition*, 2011.

[2] Christiane Fellbaum. Wordnet: An electronic lexical database. *Cambridge, MA: MIT Press*, 1998.

[3] D. Gonzalez-Aguirre, J. Hoch, S. Roehl, T. Asfour, E. Bayro-Corrochano, and R. Dillmann. Towards shape-based visual object categorization for humanoid robots. In *International Conference on Robotics and Automation (ICRA)*, 2011.

[4] Cheuk Yiu Ip, Daniel Lapadat, Leonard Sieger, and William C. Regli. Using shape distributions to compare solid models. In *Proceedings of the seventh ACM symposium on Solid modeling and applications*, SMA '02, pages 273–280, New York, NY, USA, 2002. ACM.

[5] Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. *SGP*, pages 156–164, 2003.

[6] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Matching 3d models with shape distributions. In *Shape Modeling and Applications, SMI 2001 International Conference on.*, pages 154 –166, May 2001.

[7] Radu Bogdan Rusu, Gary Bradski, Romain Thibaux, and John Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 2155 –2162, 2010.

[8] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. *11th European Conference on Computer Vision*, 2010.

[9] W. Wohlkinger and M. Vincze. Shape distributions on voxel surfaces for 3d object classification from depth images. *IEEE International Conference on Signal and Image Processing Applications*, 2011.

[10] Walter Wohlkinger and Markus Vincze. Shape-based depth image to 3d model matching and classification with inter-view similarity. *International Conference on Intelligent Robots and Systems*, 2011.

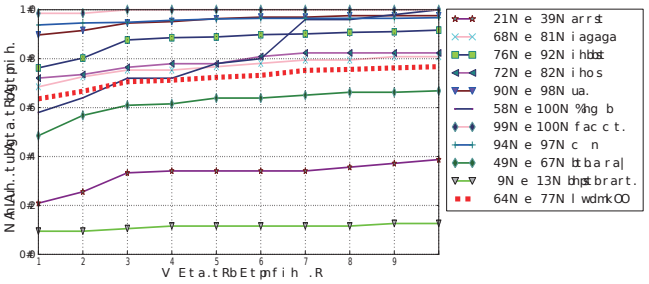


Fig. 15. SDVS rank plot on the 10 classes test database against 10 Classes.

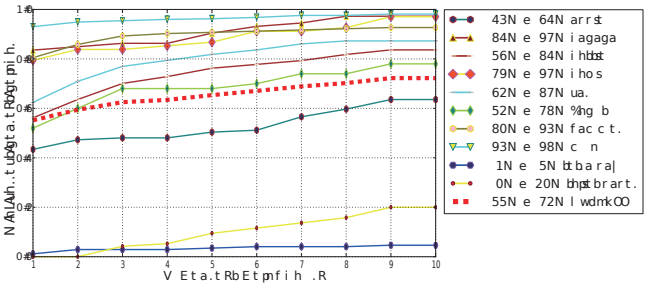


Fig. 16. VFH rank plot on the 10 classes test database against 10 Classes.

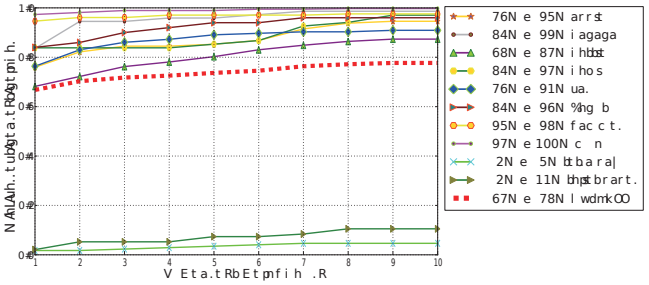


Fig. 17. CVFH rank plot shows improvement over VFH on 10 classes, but also has problems with two classes.

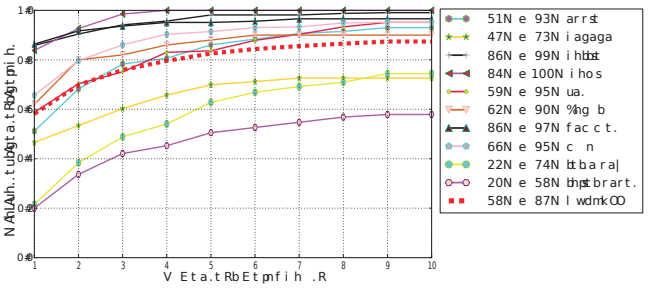


Fig. 18. SHOT rank plot on 10 classes provides good results with no class less then 20%, but is also the slowest descriptor in this benchmark.