Eurographics Workshop on 3D Object Retrieval (2015)
I. Pratikakis, M. Spagnuolo, T. Theoharis, L. Van Gool, and R. Veltkamp (Editors)

SHREC'15 Track: Retrieval of Objects Captured with Kinect One Camera

Pedro B. Pascoal [†1,2], Pedro Proença[†3], Filipe Gaspar[†2,3], Miguel Sales Dias[†2,3], Filipe Teixeira[†1], Alfredo Ferreira[†1],

Viktor Seib[4], Norman Link[4], Dietrich Paulus[4], Atsushi Tatsuma[5], Masaki Aono[5]

[1]INESC-ID / IST / Universidade de Lisboa
[2]Microsoft Language and Development Center, Lisbon, Portugal
[3]ISCTE - Instituto Universitário de Lisboa/ISTAR-IUL, Lisbon, Portugal
[4]Active Vision Group (AGAS), University of Koblenz-Landau, Koblenz, Germany
[5]Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

Abstract

Low-cost RGB-D sensing technology, such as the Microsoft Kinect, is gaining acceptance in the scientific community and even entering into our homes. This technology enables ordinary users to capture everyday object into digital 3D representations. Considering the image retrieval context, whereas the ability to digitalize photos led to a rapid increase of large collections of images, which in turn raised the need of efficient search and retrieval techniques. We believe the same is happening now for the 3D domain. Therefore, it is essential to identify which 3D shape descriptors, provide better matching and retrieval of such digitalized objects. In this paper, we start by presenting a collection of 3D objects acquired using the latest version of Microsoft Kinect, namely, Kinect One. This dataset, comprising 175 common household objects classified into 18 different classes, was then used for the SHape REtrieval Contest (SHREC). Two groups have submitted their 3D matching techniques, providing the rank list with top 10 results, using the complete set of 175 objects as queries.

Categories and Subject Descriptors (according to ACM CCS): H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Relevance feedback. I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Geometric algorithms, languages, and systems.

1. Introduction

Due to the growing popularity of low-cost scanners, several RGB-D object datasets have been emerging in the research community. While designed for different purposes, such devices have proven to be able to digitize 3D objects in real-time with sufficient quality [NIH*11], at least considering the myriad of contexts where before the presence of such 3D capturing devices, was virtually non-existent before.

In this context, we have built a dataset, which will provide the community with a benchmark for the study of computer vision, object category classification or object retrieval algorithms.

This work goes along in the lines of the work that was done by Machado et al. [MFP*13]. In this work, the authors have created a dataset using the previous generation Kinect sensor. The authors concluded that the 3D information captured by those datasets, brought challenges for tasks such as object retrieval. When capturing more complex shapes, these datasets fail to provide accurate representations of the objects. With our approach, using the latest version of Microsoft Kinect One, will help us capturing more details of smaller and more complex objects.

In the scope of the SHape REtrieval Contest track, we have created a semi-automated process for point-cloud capture and registration, which collects multiple point-clouds

† Organizers of the SHREC Track. Both the dataset and corresponding human classification are available at: http://1drv.ms/19tu1RY

DOI: 10.2312/3dor.20151068

corresponding to multiple viewpoints, whereas human-feedback is only required for ground-truth classification. A description of our data capture and 3D reconstruction pipeline, is presented in the next section. Then, in the last two sections, we present the submitted object classification techniques for the SHREC context, the evaluation process and its results and, finally, we extract some conclusions and lines of further research.

2. The Dataset

This work, follows on Machado et al. [MFP*13], but now using the new Kinect One sensor, which provides more detail. The fast spread of such cameras, has increased the need of better and more detailed datasets. Considering the characteristics of this context, we captured 175 common household objects. These range from cups and dishes, to staplers, ash trays and so on.

Our dataset provides up to 90 frame pairs of RGB and Depth images for each captured object, its corresponding registered and segmented point cloud and a polygon mesh model of the object, carefully processed, instead of just collections of local views of the same. Each model in the dataset is assigned to a category, of a hierarchical class structure, whereas the root categories represent the class of the object, e.g.: "vehicle", "animal", etc. The final leaf category of the assigned object will be, for example, the name by which the object is known, e.g.: "cars", "airplane". With this classification, we will consider similar objects that share the same leaf category.

Similarly to the work done by Machado et al. [MFP*13], we organized a study with 27 users which evaluated, for a specific query (one of the 175 dataset objects), which objects were more similar. The complete set of the 175 physical objects captured have been used in this study.

2.1. Data Collection Setup

Our data collection approach presents an easy to build solution that can be easily replicated by the scientific community, or even by common users. The capture setup and reconstruction method, collects point-clouds in multiple viewpoints, by means of a turntable rig and a fixed Kinect One sensor. For each capture session, one object is rotated 360° on a regular turntable while the Kinect One sensor, mounted on a tripod, records the rotation sequence from a fixed elevation angle, as illustrated in Figure 1. Three capture sessions are performed for each object to cover different elevation angles, and we collect frames pairs of RGB and Depth images for each object.

To allow object pose estimation per frame, physical markers are included in the capture, in a grid configuration, placed on the turntable. Once the visible markers are detected in the RGB image, we use a Robust Planar Pose algorithm to estimate the common markerboard pose, relatively to the sensor reference frame.

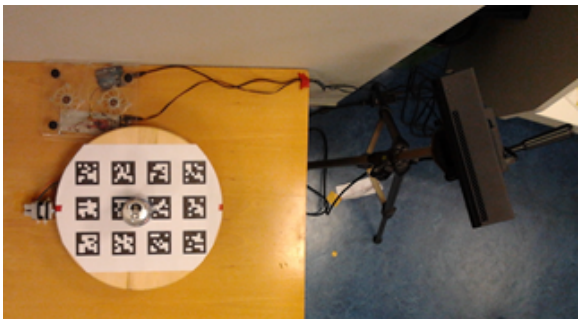

Figure 1: *Lab setup used for the object capture.*

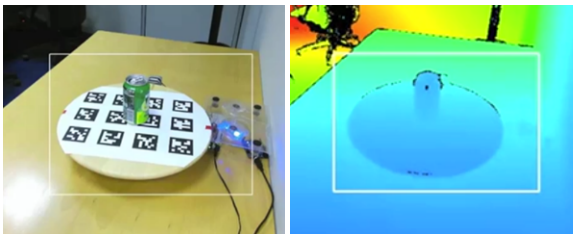

Figure 2: *Example of a pair of RGB and depth images cropped.*

The pairs of RGB and depth images are cropped semi-manually in our framework by letting the user create one bounding box for each session using only one representation view of the session, since the camera pose relatively to the scene, is the same acrosse the capture session, as depicted in Figure 2. Then, the RGB segmentation mask can be generated, by simply mapping the remainging points from the depth image to the RGB image. This process yielded good results for most of the objects in our collection.

To generate a global point cloud, all segmented local point clouds from the object, corresponding to the segmented frames, are registered in the common reference frame of the markerboard. Finally, we apply a filter to smooth the surface of the global point cloud.

2.2. Dataset Construction

As for the mesh construction we used an off-the-shelf triangulation algorithm proposed by Kazhdan et al. [KBH06] called Poisson Surface Reconstruction. This method considers all the points at once and is therefore highly resilient to data noise, creating smooth polygonal meshes, Figure 3. Unfortunatly, this method forces small holes and crevaces to be closed, which modifies some of the objects topology and shape.

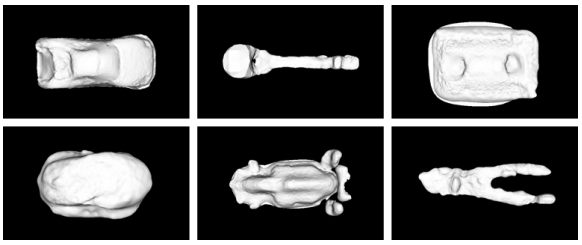

Figure 3: *Examples using Poisson Surface Reconstruction.*

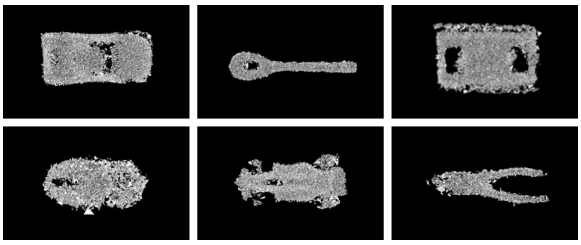

Figure 4: *Examples using the basic triangulation approach.*

Considering the limitations of the Poisson Surface Reconstruction, we additionally used a brute force approach for the triangulation of the point clouds. This method is performed, by projecting the local neighborhood of a point along the point's normal, and connecting unconnected points, until all possible points are connected. This method solves the problem of connected holes, but on the other hand creates noisy meshes, as depicted in Figure 4.

2.3. Human Relevance Evaluation

Usually, the tracks of the SHape REtrieval Contest use 3D model collections that are constructed around a finite set of classes, and the results can easily be evaluated according to whether or not a retrieval object fall into one of such class set. As simple as this evaluation might seem, classifications may be biased by the semantics of the selected classes. Also, a single object can only exist in one class at a time, and again, shape similarities may dictate that an object's shape may be confused for another from some other class. Therefore, the class-based results may not be what is expected by the users.

Our purpose was to explore a human-based relevance evaluation based on shape similarity. By presenting the list of results for a specific query, to a number of human judges to evaluate as true positive and false positive, we extract the information required for the evaluation of algorithm. With this aim, we developed an online survey, where a query and its corresponding list of results are presented to a human judge, when using one of the shape descriptors as the object retrieval technique. For each set of query-results, the judge would select which were true positives, and which were false positives, as illustrated in Figure 5. Since it was not viable to present the whole ranking of the collection for each query, we only present the top 10 results. Also, no information of which shape descriptor was used was presented to the judge, so that the evaluation was performed in a blind-test setting.

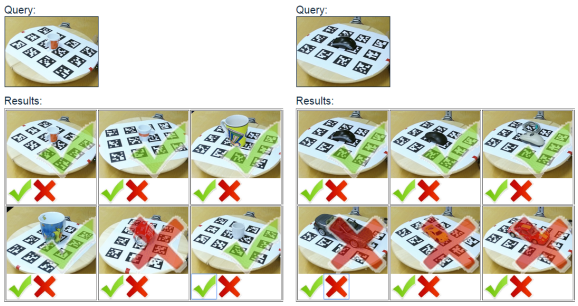

Figure 5: *Example of two user evaluations*

3. Submissions

For this contest, two distinct groups participated with their respective methods:

- V. Seib, N. Link and D. Paulus from the University of Koblenz-Landau have participated with a shape descriptor that uses a Hough-Voting in a continuous voting space. They submitted 4 sets of ranked lists using different sets of parameters. For the first strategy all models were scaled to an unit circle (radius 1m) and a descriptor radius of 0.3m with different parameter values for k was used. For the second strategy the models were scaled down, while their relative size was maintained. In this case a descriptor radius of 0.4m was used. In the comparison table and plots, this approach is denoted as CHV with the corresponding descriptor radius r and parameter k.
- A. Tatsuna and M. Aono from Toyohashi University of Technology have participated with a shape feature called Local Feature Correlation Descriptor (LCoD), producing just one set of results.

3.1. Continuous Hough-Voting (CHV)

The Continuous Hough-Voting (CHV), is related to the Implicit Shape Model formulation by Leibe et al. [LLS04]. Recently, adaptations of this method to 3D data were proposed [KPW*10, STDS10, WZS13]. In contrast to the original formulation, the adaptations to 3D data all use a discrete Hough-space for voting. Here, a continuous voting space is used and the vector quantization of features is omitted in order not to lose the feature's descriptiveness. To be able to generalize from learned shapes, each extracted feature is matched with the k best matches in the learned dictionary. Since the Continuous Hough-Voting (CHV) works

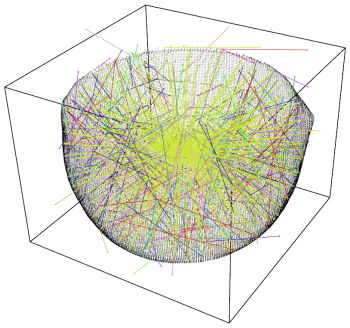

Figure 6: *Recognized object and its estimated bounding box. The green lines indicate votes of the highest ranked hypothesis, forming a maximum in the object's center. Other lines are votes for other hypotheses that are scattered and do not form a maximum.*

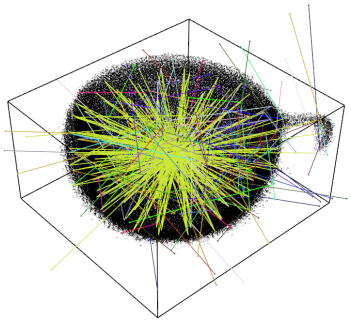

Figure 7: *Recognition of the same object as in Figure 6, however with noise.*

with point cloud data, it was required a first step to convert the provided meshes to point clouds.

For training, key points are extracted from full 3D models using a uniform voxel grid and a SHOT descriptor [TSDS10] is computed for each key point. In the next step, spatial relations between detected features on the training model are computed. For each feature, a vector pointing from the feature to the object's centroid is obtained, in the following referred to as center vector. The final data pool after training contains all features that were computed on all training models. Along with each feature, a center vector and the class of the corresponding object is stored.

To classify objects, features are detected on the input data in the same manner as in the training stage. Matching detected features with the previously trained data pool yields a list of feature correspondences. Correspondences are established at locations where the input data is assumed to match the trained object models. The distance between learned feature descriptor f_l and detected feature descriptor f_d is determined by the distance function $d(f_l, f_d) = \|f_l - f_d\|_2$. Since we can not expect to encounter the same objects during classification as were used in training, each detected feature is associated with the k best matching features from the learned data pool.

The center vectors of the created correspondences are used to create hypotheses on object center locations in a continuous voting space. A separate voting space for each class is used. Since weighted votes do not provide much benefit as suggested by the experiments of Salti et al. [STDS10], we do not use vote weighting.

Each voting space can be seen as a sparse representation of a probability density function. Maxima in the probability density function are detected using the Mean-shift algorithm. To create seed points for the Mean-shift algorithm a regular grid is superimposed on the data. Each cell containing at least a minimum number of data points creates a seed point. In a final step, the found maxima positions from all voting spaces of individual classes are merged.

The presented algorithm returns a list of results ranked by the common weight of the contributing votes. Since this challenge requires to report shape similarities, we apply this simple transformation from weights to similarities for each object i: $s = \frac{\omega_i}{\omega_{max}}$ (where ω_{max} is the weight of most likely object hypothesis).

An example image of the recognition is presented in Figure 6. The recognition of the same object with noise is shown in Figure 7.

In general, the second scaling strategy (maintaining the relative object scale) leads to better results compared to downscaled point clouds where a uniform scale is used. Further, it is better to use higher values for k.

3.2. Local Feature Correlation Descriptor (LCoD)

Previously, A. Tasuma and M. Aono proposed the Local Feature Correlation Descriptor (LCoD) as a view-based 3D shape descriptor in [TA13, MFP*13]. The LCoD comprises the correlations of local features extracted from depth buffer images.

Figure 8 illustrates the generation of LCoD feature. As pre-processing, it first performs a pose normalization using Point SVD [TA09] to determine the scale, position, and rotation of a 3D model.

Next, it encloses the 3D model within a unit geodesic sphere. From each vertex of the unit geodesic sphere, its rendered the depth and color buffer images with 256×256 resolution, and a total of 38 viewpoints are defined.

Finally, its calculated the correlation matrix of local features for each depth buffer image. Let $\mathbf{v}_i^{(j)} \in R^d (i = 1, \ldots, n)$ be d-dimensional local features extracted from a depth buffer image $L_j (j = 1, \ldots, 38)$. In this implementation, its extracted a SURF descriptor [BETVG08] from 48×48

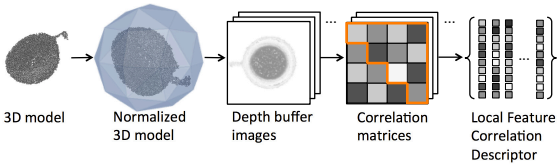

Figure 8: *Overview of the Local Feature Correlation Descriptor (LCoD)*

pixel patches arranged every 4 pixels as a local feature. The correlation matrix $R^{(j)}$ is obtained as follows: $\mathrm{R}^{(j)} = \frac{1}{n}\sum_{s,t=1}^{n} \mathbf{v}_s^{(j)} \mathbf{v}_t^{(j)\top}$. The vector $\mathbf{r}^{(j)}$ consists of concatenating the elements in the upper triangular part of the correlation matrix $R^{(j)}$.

$$\mathbf{r}^{(j)} = [R^{(j)}_{1,1}, \ldots, R^{(j)}_{1,d}, R^{(j)}_{2,2}, \ldots, R^{(j)}_{2,d}, \ldots, R^{(j)}_{d,d}]. \quad (1)$$

The vector $\mathbf{r}^{(j)}$ is normalized using the power and ℓ_2 normalization [PSM10] to diminish its sparseness.

The LCoD feature of a 3D model is defined as a set of the vectors $\{\mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(38)}\}$. To compare two LCoD features, we apply the Hungarian method [Kuh55] to all pair Euclidean distances between their vectors.

3.3. Further evaluation

In order to have some grounds for comparison and further build this study, we chose to evaluate some existing algorithms, from different classification branches and proven efficiency [TV08], as we could compare them with the track results to evaluate the gain achieved with the new contributions. For that purpose, we selected the following shape descriptors:

- **Coord and Angle Histograms (CAH)** [PR99]
  A cord is defined as a ray segment which joins the barycenter of the mesh with a triangle center. Since only global features are used to characterize the overall shape of objects this method is not very discriminative about object details, but the implementation is straightforward. These methods can be used as an active filter or in combination with other methods to improve results.
- **Spherical Harmonics (SH)** [KFR03]
  This approach, a kind of spacial based similarity feature descriptor, was originally proposed by Funkhouser et al. [KFR03] and outperforms many other approaches.
- **Lightfield Descriptors (LFD)** [CTSO03]
  The Light Field Descriptor is a view-based geometry feature retriever. Its authors claim, for studies driven in different databases, that its retrieval rate is distinctively higher than other view-based and feature-based methods, and that it should be tested with other benchmarks [SMKF04].

4. Evaluation Results

All SHREC participants submitted for the requested query set, at least one rank listing (one for each run). Each rank list has the top 10 results. We considered this information sufficient, since for our human relevance evaluation to be effective, it would not be viable to present too many result to the judges. Furthermore, using the top 10 results, would already enable us to extract enough information from precision-recall curves, to identify which techniques performed the best.

We employed the following evaluation measures on the results: Nearest Neighbor (NN), First-Tier (FT) and Discounted Cumulative Gain (DCG) [SMKF04]. These measures are based on the Precision and Recall evaluations of the queries and were chosen to give a general overview of the proposed methods.

Based on the results presented on Table 1, both view-based approaches, LCoD and LFD, provided the best results of the group. These methods are proven to be more robust to topology errors, surface deformations and noise, which are frequently present in the dataset models, thanks to the low degree of accuracy of the used camera. View-based methods generally work by extracting features from a range of different views taken from separate angles of the models, much like the way our testers evaluated and compared the physical models presented to them. We could surmise that some of the techniques used by view based methods are the ones that best mimic the same method employed by the evaluation judges.

Notwithstanding, the numbers are still considerably low when compared to the typical evaluation results with more accurate databases [SMKF04]. The major raionale to this fact are the state of the art on low-cost depth cameras hardware and respective capture software, which are still unable to provide a degree of accuracy to be realistically used in real-time scenarios.

Additionally, we can clearly notice that there is not much difference between the precision-recall curves of the collection using Poisson Surface Reconstruction and the basic triangulation approach. With the exception of the CAH, all method provided better results when using the basic triangulation approach. This was mostly due to the Poisson Surface Reconstruction connecting small holes and crevaces,

Participant	Method	NN	FT	NDCG (p=10)
A. Tatsuma & M. Aono	LCoD	0.623	0.435	0.533
V. Seib, N. Link and D. Paulus	CHV (r=0.30, k=3)	0.531	0.357	0.459
	CHV (r=0.30, k=5)	0.514	0.366	0.467
	CHV (r=0.40, k=5)	0.589	0.381	0.488
	CHV (r=0.40, k=7)	0.589	0.389	0.498
	LFD	**0.640**	**0.463**	0.556
	SH	0.537	0.435	0.519
	CAH	0.451	0.339	0.434

Table 1: *Retrieval performances of the algorithms.*

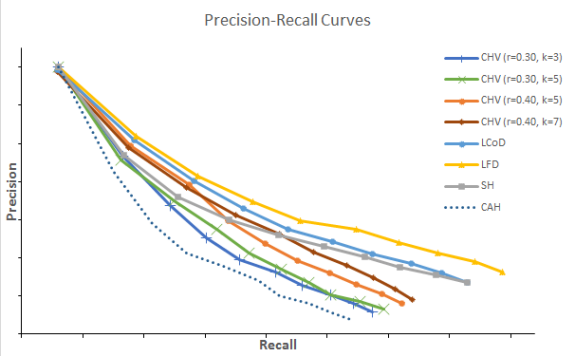

Figure 9: *Precision-Recall curves for the meshes built using Poisson Surface Reconstruction.*

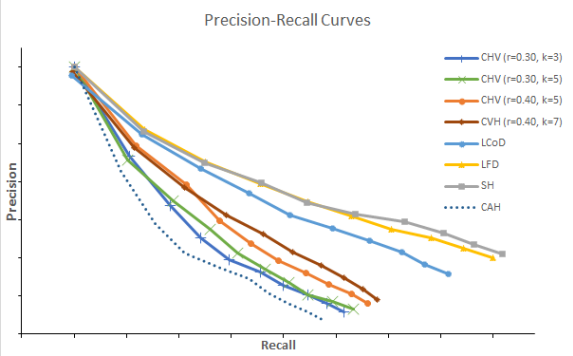

Figure 10: *Precision-Recall curves for the meshes built using the basic triangulation approach.*

whereas most approaches are robust to surface deformations and noise. This also explains the great improvement in the SH results, which similarly to the view based methods, is robust to noise and topology errors.

5. Conclusions

In this paper, we have described and compared object retrieval and classification algorithms from each of the two different research groups that participated in the SHREC'15 Track: Retrieval of Objects Captured with Kinect One Camera. Each participant was presented with a collection of 175 3D polygon mesh models, to use as query set, and asked to submit a top 10 ranked list of results for each of their respective matching algorithms and possible variants.

The ranked list of results were evaluated by human judges that had no prior experience in 3D computer graphics or computer vision, and compared these numbers against some of the state-of-the-art 3D retrieval descriptors available. Analyzing the statistics, view-based methods provided the best retrieval results. View-based methods are considerably better suited for recognizing similar objects with very low levels of detail, including surface noise and topological errors. These methods are in fact proven to be robust, and able to ignore defects across different objects, which makes them ideal to be used in the context of databases featuring 3D models captured using low-cost depth-sensing cameras.

6. Acknowledgements

The work described in this paper was supported by the following institutions:

- This work was supported by the ACP-EU Support Programme to ACP Cultural Sectors, with the Financial contribution of the European Union (European Development Fund) and the assistance of the ACP Group of States, with reference ACPStreetLibraries/FED/2013/328445.
- Was also supported by national funds through "Fundação para a Ciência e a Tecnologia" (FCT) with reference UID/CEC/50021/2013, and partially supported by the "Agência Nacional de Inovação" (ANI) through the project CLUTCH, reference ADI/QREN/22984.

References

[BETVG08] BAY H., ESS A., TUYTELAARS T., VAN GOOL L.: Speeded-up robust features (SURF). *Computer Vision Image Understanding 110*, 3 (June 2008), 346–359. 4

[CTSO03] CHEN D.-Y., TIAN X.-P., SHEN Y.-T., OUHYOUNG M.: On Visual Similarity Based 3D Model Retrieval. vol. 22 of *EUROGRAPHICS 2003 Proceedings*, pp. 223–232. 5

[KBH06] KAZHDAN M., BOLITHO M., HOPPE H.: Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing* (2006), vol. 7. 2

[KFR03] KAZHDAN M., FUNKHOUSER T., RUSINKIEWICZ S.: Rotation invariant spherical harmonic representation of 3D shape descriptors. In *Symposium on Geometry Processing* (June 2003). 5

[KPW*10] KNOPP J., PRASAD M., WILLEMS G., TIMOFTE R., VAN GOOL L.: Hough transform and 3d surf for robust three dimensional classification. In *ECCV (6)* (2010), pp. 589–602. 3

[Kuh55] KUHN H. W.: The hungarian method for the assignment problem. *Naval Research Logistics Quarterly 2* (1955), 83–97. 5

[LLS04] LEIBE B., LEONARDIS A., SCHIELE B.: Combined object categorization and segmentation with an implicit shape model. In *ECCV' 04 Workshop on Statistical Learning in Computer Vision* (2004), pp. 17–32. 3

[MFP*13] MACHADO J., FERREIRA A., PASCOAL P. B., ABDELRAHMAN M., AONO M., EL-MELEGY M., FARAG A., JOHAN H., LI B., LU Y., TATSUMA A.: Shrec'13 track: Retrieval of objects captured with low-cost depth-sensing cameras. In *Proceedings of the Sixth Eurographics Workshop on 3D Object Retrieval* (Aire-la-Ville, Switzerland, Switzerland, 2013), 3DOR '13, Eurographics Association, pp. 65–71. URL: http://dx.doi.org/10.2312/3DOR/3DOR13/065-071, doi:10.2312/3DOR/3DOR13/065-071. 1, 2, 4

[NIH*11] NEWCOMBE R. A., IZADI S., HILLIGES O., MOLYNEAUX D., KIM D., DAVISON A. J., KOHLI P., SHOTTON J., HODGES S., FITZGIBBON A.: Kinectfusion: Real-time

dense surface mapping and tracking. In *Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality* (Washington, DC, USA, 2011), ISMAR '11, IEEE Computer Society, pp. 127–136. URL: http://dx.doi.org/10.1109/ISMAR.2011.6092378, doi:10.1109/ISMAR.2011.6092378. 1

[PR99] PAQUET E., RIOUX M.: Nefertiti: a query by content software for three-dimensional models databases management. In *Proceedings of the International Conference on Recent Advances in 3-D Digital Imaging and Modeling* (Washington, DC, USA, 1999), NRC '99, IEEE Computer Society, pp. 345—-. URL: http://dl.acm.org/citation.cfm?id=523428.825366. 5

[PSM10] PERRONNIN F., SÁNCHEZ J., MENSINK T.: Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision: Part IV* (Berlin, Heidelberg, 2010), ECCV '10, Springer-Verlag, pp. 143–156. 5

[SMKF04] SHILANE P., MIN P., KAZHDAN M., FUNKHOUSER T.: The Princeton Shape Benchmark. *Shape Modeling and Applications, International Conference on 0* (2004), 167–178. doi:http://doi.ieeecomputersociety.org/10.1109/SMI.2004.1314504. 5

[STDS10] SALTI S., TOMBARI F., DI STEFANO L.: On the use of implicit shape models for recognition of object categories in 3d data. In *ACCV (3)* (2010), Lecture Notes in Computer Science, pp. 653–666. 3, 4

[TA09] TATSUMA A., AONO M.: Multi-fourier spectra descriptor and augmentation with spectral clustering for 3D shape retrieval. *The Visual Computer 25*, 8 (2009), 785–804. 4

[TA13] TATSUMA A., AONO M.: 3D object retrieval based on correlation of multi-view image local feature. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics (in Japanese) 25*, 1 (2013), 556–567. doi:10.3156/jsoft.25.556. 4

[TSDS10] TOMBARI F., SALTI S., DI STEFANO L.: Unique signatures of histograms for local surface description. In *Proc. of the European conference on computer vision (ECCV)* (Berlin, Heidelberg, 2010), ECCV'10, Springer-Verlag, pp. 356–369. 4

[TV08] TANGELDER J. W., VELTKAMP R. C.: A survey of content based 3D shape retrieval methods. *Multimedia Tools Appl. 39*, 3 (2008), 441–471. URL: http://portal.acm.org/citation.cfm?id=1395016.1395041, doi:10.1007/s11042-007-0181-0. 5

[WZS13] WITTROWSKI J., ZIEGLER L., SWADZBA A.: 3d implicit shape models using ray based hough voting for furniture recognition. In *3DTV-Conference, 2013 International Conference on* (2013), IEEE, pp. 366–373. 3