Good Appearance and Shape Descriptors for Object Category Recognition

Pedro F. Proença[1], Filipe Gaspar[1], Miguel Sales Dias[1, 2]

[1]ADETTI – IUL/ISCTE-Lisbon University Institute, Portugal
[2]Microsoft Language Development Center, Portugal

Abstract. In the problem of object category recognition, we have studied different families of descriptors exploiting RGB and 3D information. Furthermore, we have proven practically that 3D shape-based descriptors are more suitable for this type of recognition due to low shape intra-class variance, as opposed to image texture-based. In addition, we have also shown how an efficient Naive Bayes Nearest Neighbor (NBNN) classifier can scale to a large hierarchical RGB-D Object Dataset [2] and achieve, with a single descriptor type, an accuracy close to state-of-art learning based approaches using combined descriptors.

1 Introduction

Object category recognition is the task of classifying one object instance never seen before. Here instance stands for object physically unique and category consists of instances that share common features. The recent availability of RGB-D information provided by Microsoft Kinect Sensor encouraged researchers to use this combined information in computer vision problems. Progress was made in instance recognition [1][2][3][4], object categorization [2][3] and pose estimation [1]. In this context we have studied the performance of different families of feature descriptors, exploring this information on the task of generic object category recognition. For this purpose we rely on a publicly available large hierarchical RGB-D object dataset [2].

The bases of most computer vision applications are local features. Numerous feature descriptors have been proposed for intensity images [6][7][8] and for 3D point clouds [9][10][11] respectively: local image descriptors and local surface descriptors, and both share common principles. Since the last decade, SIFT [6] has been consistently the most accepted and used local image feature and Spin Image [9] is arguably the most popular local surface descriptor. However, in recent literature, some local image feature methods [7][8] faster than SIFT are reported and, some local surface descriptor methods [10][11], claim to be more noise resilient and discriminative than Spin Image. As our first contribution we have tested local surface descriptors and local image descriptors side by side.

In object recognition, state-of-art methods are usually based on a combination of bag of words (BoW) [12] with Support Vector Machine classifier. In image classification, Naive Bayes Nearest Neighbor (NBNN) classifier [4] was introduced as a competitive alternative to these learning based methods. This non-parametric

classifier does not require a quantization step, inherent of BOW, and thus features maintain their discriminative power. NBNN also generalizes well beyond the training data by exploiting Image–to-Class distance rather than Image-to-Image distance used in other NN approaches. Over the recent years several modified versions were proposed to deal with NBNN limitations.

In [14] a more powerful parametric version of NBNN than the original was introduced, supporting unbalanced datasets, where the number of features per class is strongly class-dependent. Thanks to the introduction of a learning phase, the bias towards more densely classes was corrected, resulting in 15-percentage points gain in several datasets. In [15] it was criticized the independence assumption of NBNN. The argument was that since each feature is treated separately, the information as a whole describing the image, is ignored. As a result, the accuracy of distinguishing classes that share similar local features is worse than in BOW, which encodes the feature distribution over the image. More recently the problem of scalability was addressed in [16]. It was shown that multi-way NBNN version using one merged search structure for all the training data instead of a separate search structure for each class, achieved a 100 times speed-up over original NBNN, with 256 classes. As a second contribution of our paper, we have extended NBNN to local surface descriptors in a dataset [2] dominated by learning based methods.

2 Classification Pipeline

An example of our 3D Object Classification Workflow approach is depicted in Figure 1. Our pipeline builds on the data provided by the RGB-D Object Dataset [2], which comprises RGB images and 3D point clouds, already segmented, from several views around the objects. At training time, for each class, we simply extract image and 3D descriptors from a set of views from all training objects belonging to a category and then, we build random kd-trees as an approximate search structure. Our simple NBNN training stage is class-independent and has no weight-learning phase hence it's much more suitable to online learning applications than learning-based approaches.

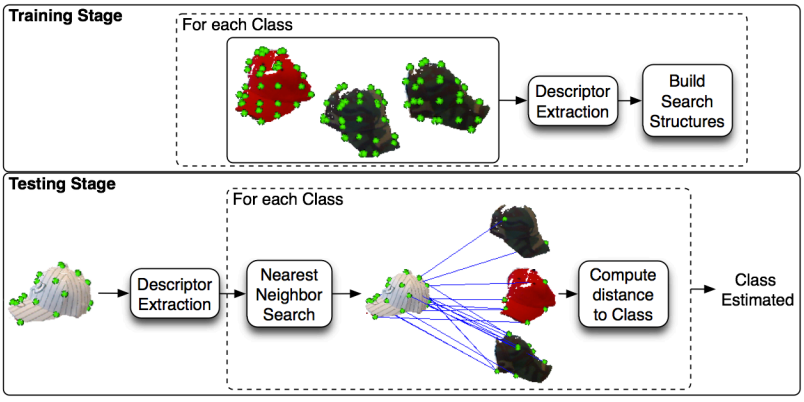

Fig. 1. Overview of the classification pipeline proposed in this work. We show an example of a cap class query for a cap point cloud using 3D surface descriptors

In testing time, given a query frame (i.e. image or point cloud) from a never-seen-before object, we extract descriptors from the test frame the same way we did in the training phase, although a sparser sampling in this phase causes a significantly speed-up, in exchange for a minor performance loss [15]. Our method is no different from NBNN algorithm described in Algorithm 1 below. To evaluate the likelihood of a query frame Q belonging to a class c, for each query descriptor d_i we search its approximated nearest neighbor in c: $NN_c(d_i)$. Then our distance to c is the sum of all the correspondence distances measured using the squared L_2 distance. This procedure is repeated to all training classes. Then the class estimated is the class with smaller distance.

1.	Compute descriptors d_1, ..., d_n from query image Q.
2.	$\forall d_i \ \forall_C$ compute the NN of d_i in c: $\mathrm{NN}_C(d_i)$
3.	$\hat{c} = \arg min_c \sum_{i=1}^{n} \|d_i - NN_c(d_i)\|^2$

Algorithm 1. NBNN [5]

3 Approximate Nearest Neighbor Search

The number of training descriptors in each class is very large: in a dense descriptor extraction, we can get 500 feature descriptors in one image of one object instance. Our regular number of training images for instance is around 120, if a class has at least 4 instances, the total number of descriptors in that class is 240,000. Hence a simple linear search is not a choice. A kd-tree has logarithmic time complexity for low dimensions though its efficiency tends to decrease with the feature dimensionality, which in our case is at least 128 dimensions.

Therefore approximate nearest neighbor search methods are required when using NBNN. Approximate search neighbor's search time-precision tradeoff is controlled by parameter c, the number of leaf nodes checked. We use FLANN [17] implementation and our chosen search structure is 4 random k-d-trees [18] with $c = 20$, as they are quite precise at a cost of relatively low memory footprint. Our search time complexity is then $O\left(cN_{DQ}N_C \log(N_I N_{DI})\right)$. Where c in number of checks, N_{DQ} is the number of query descriptors, N_C is the number of Classes, N_I is the number of instances per class and N_{DI} is the number of descriptors per instance. One query with 51 classes in our parallel x64 implementation takes between 30 to 160 ms, depending on N_{DQ} and building 4 random k-d trees per class takes between 0.56 to 4.4 seconds. Tests were performed in one 2.3 GHz core i5 with 4 GB of RAM.

4 Visual Appearance and Shape Descriptors

Our work focuses on the study of different kinds of descriptors, under the classification approach described in Section 2, and the selection of the one that we believe is more appropriate for object classification tasks. In Section 4.1 we describe

the descriptors that capture the visual appearance information present in RGB images. In section 4.2 we describe the surface descriptors that capture the 3D shape information from point clouds.

4.1. Visual Appearance Descriptors

We have selected SIFT [6] as representative of local image feature descriptors due to its superiority in precision, discriminability and popularity. We have used VLFeat [19] DSIFT implementation, to extract SIFT descriptors from regular dense grid points using one fixed scale patch (see parameter details in Section 5). We have discarded low contrast features as in [16] and unlike the original SIFT, the rotation invariance is disabled in these descriptors. We state that in this classification approach, the discriminative power of the descriptors is more relevant than the invariance propriety and that the feature orientation information, favors discriminability. Rotation invariance loses this information, leading consequently to lower discriminability.

In order to make the descriptors even more discriminative we have included the keypoint normalized coordinates in the descriptor histogram, keeping the aspect ratio. Hence the 128-dimensional SIFT descriptor becomes 130-D. This technique used initially in [5], represents spatial information in finer way than spatial pyramid [13]. In Section 5 we show the gain of including this information.

Local image descriptors explore the intensity-images and not color information, thus we include global Hue Histograms in our framework to capture this complementary information. One Hue histogram is extracted from each image, low saturation zones are unreliable therefore we discard them. We tested both a Hard and Soft Assignment version. In hard version a Hue sample is simply assigned to the closest histogram bin, whereas in a soft version, a Hue sample is count in the closest and in the second closest bin with weights proportional to proximity.

4.2. Shape Descriptors

As local surface feature descriptor we have selected Spin Image [9], already used by other authors in the RGB-D Object Dataset and the most recent SHOT descriptor [10]. Both are view-invariant. In Spin Image, the 3D surface around the point feature is represented by a 2D Histogram. Points falling into the neighborhood (support) of the point feature are count according to their cylindrical coordinates (r, z), without the azimuth angle, where the origin is the feature point location, the longitudinal axis (z the signed height) is the feature points normal, r the radius. Hence the descriptor is rotation invariant around the feature point normal.

On the other side, SHOT uses spherical coordinates and encodes the azimuth angle, the radial distance and the elevation angle, by estimating a unique reference frame instead of a single reference axis (i.e. feature point normal) and using local histograms like SIFT. Therefore we believe that SHOT yields more discriminative power than Spin Image.

As feature detection we have used a common method to both descriptors and analogous to DSIFT. Keypoints are detected using a uniform voxel grid over the point cloud. In each voxel, the centroid of points within the voxel generates a point feature 3D location. Our sampling parameter is the voxel size.

To enhance the descriptor descriptiveness we have included the feature point 3D location (relative to the bounding box enclosing the object) in the descriptor histogram. We normalize the 3D coordinates dividing them by the largest dimension found in the bounding box. In Section 5 we tune both descriptors support radius on the RGB-D Object dataset keeping the standard descriptor length: SHOT with 352 bins + 3 spatial bins and Spin Image with 153 dimensions + 3 spatial bins.

4.3. Comparing Local Descriptors

Comparing fairly image local descriptors with local surface descriptors is a difficult task, since each family of descriptors has its own particular parameterization namely, the sampling step. We believe that the simple and fairest way to compare them is to take into account the number of descriptors per frame.

We simply perform a coarse tuning of the sampling parameters to find close cardinality matches, in two different categories of objects: Apple and Cereal Box, the former a relatively smaller and low-textured object and the latter a larger object rich in textures. In Table 1 we show this cardinality in two types of sampling, a sparser and a denser, these parameters are further used in our official results. As can be observed there's a certain disparity that gives DSIFT some advantage. Our method is far from ideal considering that the density of local image features strongly depends on the textures, whereas the local surface descriptors depend rather on the size of the object. One could impose a limit of features per category as the minimum descriptors extracted between both descriptor types but instead we let the methods capitalize on the class features in order to perform a discriminative evaluation at the class level.

Table 1. Average number of local features, per frame, in each class in function of the sampling step. In DSIFT sampling is the grid resolution. In SHOT and Spin Image sampling is the grid voxel size in meters

Class	DSIFT		SHOT & SI	
	8x8	3x3	0.015	0.005
Apple	63	425	59	397
Cereal_box	555	3887	407	3005

5 Experiments and Results

5.1. Dataset for training and testing

Throughout our experiments we have used the RGB-D Object dataset [2]. This large dataset is comprised of sequences of 640x480 color and depth images of 300 instances of household objects grouped in 51 categories, it is unbalanced in the sense

that the number of instances per category ranges from 3 to 14, with the average being 6 instances. Each object was recorded from 3 elevation angles (30º, 45º, 60º) while it was rotating in a turntable rig. We subsampled the dataset as in [2] by taking every 5th frame of the video resulting in about 120 frames per Object instance. Our results reported in Section 5.3 follow the standard object category recognition evaluation method: We measure accuracy over 10 trials and in each trial one random instance from each category is left for testing and the remaining instances from all classes are used for training.

5.2. Tuning Descriptors

All results in this section were obtained for 20-train classes. Figure 2 shows the effect of adding spatial information. α is the weight assigned to scale the normalized spatial bins. In all descriptors there's a performance boost. However we realize that the weight for spatial information is descriptor-dependent. We emphasize the importance of finding optimum spatial information-descriptor tradeoff. For instances, Spin Image is better without spatial information than with $\alpha = 2$. Based on this results we set $\alpha = 0.25$ for Spin Image, $\alpha = 0.5$ for SHOT and $\alpha = 1.5$ to DSIFT. Still in Figure 2 we show the tuning of the descriptor length for the Hue Histogram. The soft assignment version reaches high accuracy with remarkably only 6 bins. The hard assignment only meets this value with 14 bins.

Experiments in Figure 3 and Figure 4 aims to discover the optimum descriptor size. As can be generally observed, accuracy has at least two distinctive zones: a growth and a saturation zone. In the former the size of the descriptor is not enough to fully exploit the object's features, along this zone discriminative power increases until the saturation zone where increasing the descriptor size doesn't add more discriminative information and ends on losing it (Figure 3).

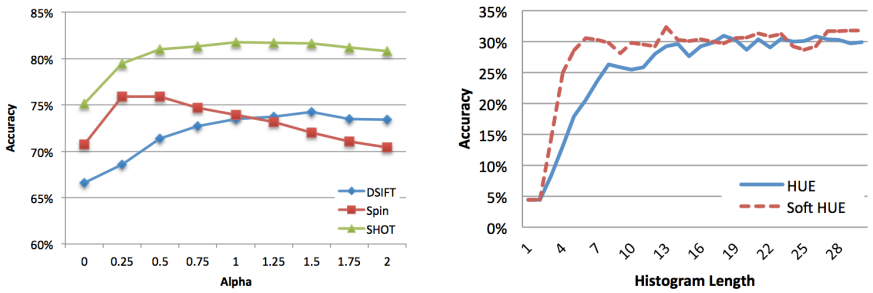

Fig. 2. (*Left*) Alpha Tuning. The effect of spatial information weight on the descriptors performance. (*Right*) Histogram length tuning in Hue Histograms. Soft Hue corresponds to our Hue descriptor implementation with soft assignment

For DSIFT in Figure 3, we used a sampling step of 8 pixels (which means descriptors always overlap). We also show the average feature extraction time for the two different classes considered in section 4.3. Based on these results we use a 24x24 patch instead of the traditional 16x16. For Spin Image and SHOT we used a voxel size of 3 cm due to efficiency. In SHOT we choose a support radius of 5 cm taking

into account the time complexity, which is exponential, this is not observed for the apple due to its size being smaller than the support at a certain level. Spin Image is faster to extract than SHOT however it requires a much larger support radius, as result Spin Image ends up being slower, we choose using a support radius of 30 cm to maximize accuracy. Here the cereal box time slope drops by the same reason as the apple in SHOT.

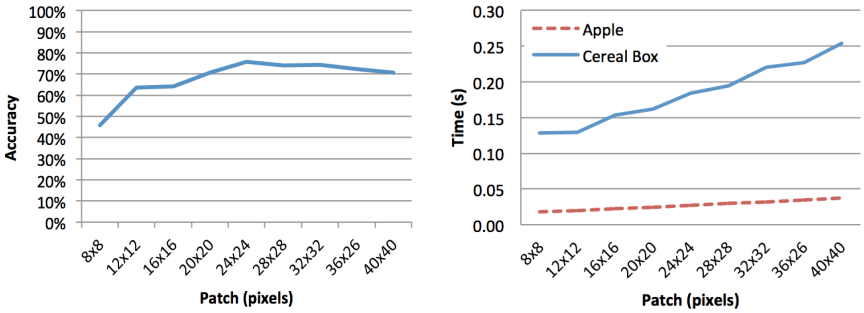

Fig. 3. Patch size tuning in DSIFT. *(Left)* Accuracy as function of the patch size. *(Right)* Feature extraction time per category

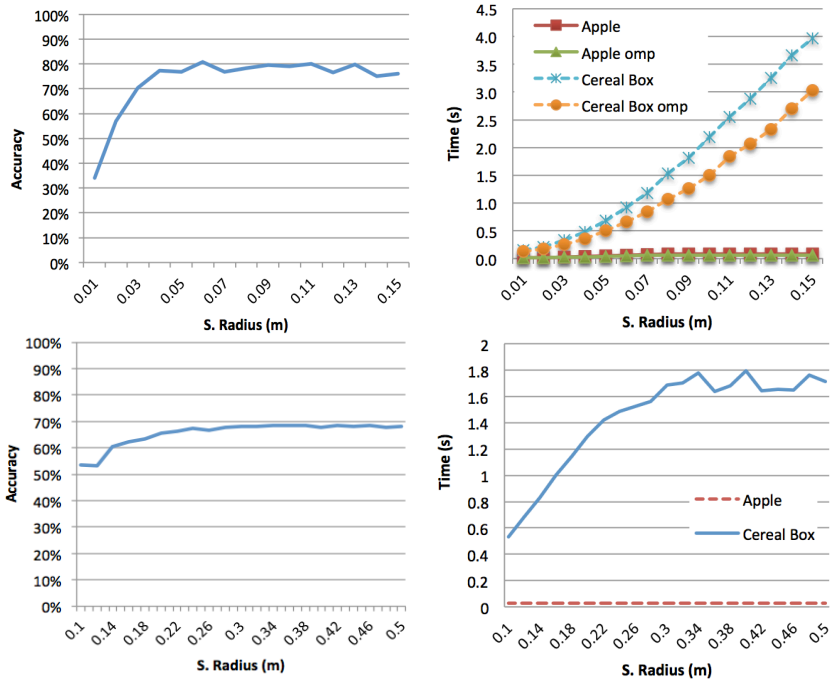

Fig. 4. Support size tuning in surface descriptors. SHOT in the top row, Spin Image in the bottom row. *(Left)* Accuracy as function of the support radius *(Right)* Feature extraction time for category, Omp stands for OpenMP, corresponds to the parallel implementation version available in PCL, running with two threads

5.3. Results

We have evaluated the RGB-D Object dataset and compared the different descriptors, using the methodology defined in section 5.1. Figure 6 depicts the respective confusion matrices. One could notice several similarities between them. We have observed that the 3 common worst classes are: Mushroom, Peach and Camera, that all have 3 instances each and suffer from high intra-class variance. We have also checked the classes with most bias: Food bag with 8 instances and Sponge with 12 instances. This is coherent with NBNN expected behavior in unbalanced datasets. The classes with most bias are visually observed by looking at the pronounced vertical lines (class predicted) that means low precision, despite the natural high recall. Hue confusion matrix (not shown) can be described as cluttered with a weak diagonal.

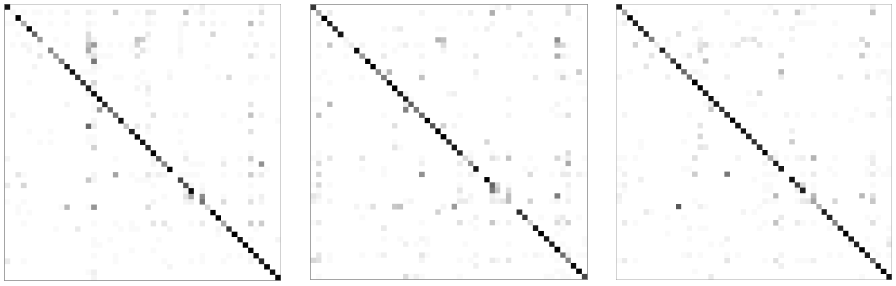

Fig. 5. Confusion matrices row normalized for all classes. Actual Classes along the ordinate and Predicted Classes along abscissa. *Left*: DSIFT, *Middle*: Spin Image, *Right*: SHOT

We have found some notable differences recorded in Table 2. Classes with low shape intra-variance and some texture variance (e.g. Cap, Cereal Box) are significantly better classified by local surface descriptors. On the other hand, we also show classes where appearance plays a big part, e.g. the light bulb specular reflection is discarded by local surface descriptors and the tomato has a common shape thus ambiguous for surface descriptors. Our color descriptor may not be as strong as the others for generic class classification, but it exceeds others performance in classes with unique and constrained colors, such as Greens.

Table 2. Class accuracy for Descriptor. Accuracy is measured as F1 Score. Recall and Precision are computed directly from the confusion matrices with false positives appearing in the class ordinate and false negatives in the class abscissa

Desc.	Ball	Bell pepper	Cap	Cereal box	Greens	Comb	Dry battery	Light bulb	Tomato
DSIFT	0.08	0.39	0.61	0.64	0.37	0.97	0.83	0.75	0.58
Spin	0.28	0.65	0.97	0.99	0.79	0.61	0.53	0.24	0.33
SHOT	0.42	0.81	0.92	0.86	0.90	0.70	0.61	0.46	0.42
Hue	0.19	0.23	0.00	0.01	0.78	0.48	0.06	0.19	0.63

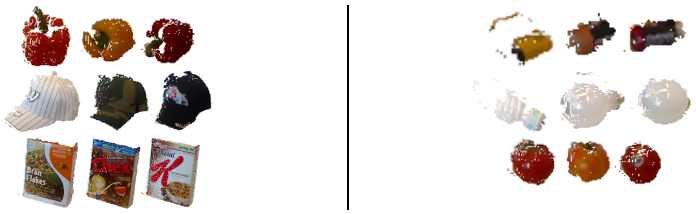

Fig. 6. RGB Point Clouds taken from the Categories in Table 2, except ball, greens and comb (same order). (*Left*) Classes in which surface descriptors performance is better. (*Right*) Classes in which image intensity descriptors performance is better

Although our classifier suffers from bias, we have achieved classification results that compare well with the literature [2][3]. The performance gap between SHOT and Spin Image accuracy proves that SHOT is more discriminative. The margin between SHOT and DSIFT and the density disparity mention in section 4.3 ensure us that local surface descriptors are more appropriate for object classification.

Table 3. Overall accuracy over 10 trials. Spin Image and SHOT sampling is the voxel size used, DSIFT sampling is the grid resolution. The second denser set improves accuracy at a cost of much higher memory use (e.g. SHOT reaches 17 Gb RAM and DSIFT 11 Gb)

Descriptor	Sampling	Accuracy
DSIFT	8x8	71.2 ± 2.3
Spin Image	0.015 m	68.4 ± 2.8
SHOT	0.015 m	**75.9 ± 2.1**
Hue	20 bins	19.8 ± 2.2
DSIFT	3x3	73.0 ± 2.1
SHOT	0.005 m	**77.4 ± 2.3**

6 Conclusion and Future work

In this paper we have extended NBNN to 3D descriptors and evaluated different families of visual and shape descriptors, on the task of object category classification. With this study we have concluded that SHOT outperforms the traditional Spin Image, most likely due to SHOT representation of azimuth and Spin image sensibility to the feature point's normal computation. We have also found that depth based descriptors are generally more reliable than intensity image (visual appearance) descriptors, in the task of object category recognition. This is due to fact that image texture intra-class variance is higher than shape intra-class variance. However we have found that some classes benefit from appearance descriptors due to low shape inter-class variance.

With only a single type descriptor and no learning phase, we have achieved accuracy close to a combined descriptor learn-based approaches. In the future, we plan to introduce a distance-learning phase such as in [3], in our method, in order to combine our selected visual and shape descriptors and correct the bias.

Acknowledgments: The work was partially supported by Project: QREN 7943 CNG

– Contents for Next Generation Networks, co-promotion, managed by Agência de Inovação (ADI), and also partially supported by project PEst-OE/EEI/UI0605/2011, managed by Fundação para a Ciência e Tecnologia (FCT) - Portugal, two R&D Projects funded by European Structural Funds for Portugal (FEDER) Through COMPETE.

References

1. Tang, J., Miller, S., Singh, A., and Abbeel, P.: A Textured Object Recognition Pipeline for Color and Depth Image Data. In: ICRA (2012)
2. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view RGB-D object dataset. In: ICRA (2011).
3. Lai, K., Bo, L., Ren, X., Fox, D.: Sparse Distance Learning for Object Recognition Combining RGB and Depth Information. In: ICRA (2011)
4. Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes. In: ICCV (2011)
5. Boiman, O., Shechtman, E., Irani, M.: In Defense of Nearest-Neighbor Based Image Classification. In: CVPR (2008)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60: 91–110 (2004)
7. Bastos, R., Dias, M. S.: FIRST - Fast Invariant to Rotation and Scale Transform. VDM Verlag Dr. Müller e.K. ISBN: 978-3-639-17489-2 (Junho 2009)
8. Calonder, M., Lepetit V., Strecha, C., Fua P.: Brief: Binary robust independent elementary features. Computer Vision– ECCV 2010, pages 778–792 (2010)
9. Johnson, A., Hebert, M.: Using spin images for efficient object recognition in cluttered 3d scenes. PAMI 21, 433–449 (1999)
10. Tombari, F., Salti, S., Stefano, L.D.: Unique signature of Histograms for local surface description. In: ECCV (2010)
11. Rusu, R. B., Marton, Z. C., Blodow, N., Beetz, M.: Learning Informative Point Classes for the Acquisition of Object Model Maps. In Proceedings of the 10th International Conference on Control, Automation, Robotics and Vision (2008)
12. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In Workshop on Statistical Learning in Computer Vision (2004)
13. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
14. Behmo, R., Marcombes, P., Dalalyan, A., Prinet, V.: Towards optimal naive bayes nearest neighbor. In ECCV (2010)
15. Tuytelaars, T., Fritz, M., Saenko, K., Darrell, T.: The NBNN kernel. In ICCV (2011)
16. McCann, S., Lowe, D.G.: Local Naive Bayes Nearest Neighbor for Image Classification. In *CVPR* (2012)
17. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: VISSAPP (2009)
18. Silpa-Anan, C., Hartley, R.: Optimised KD-trees for fast image descriptor matching. In: CVPR (2008)
19. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/ (2008)