Recognizing in the Depth: Selective 3D Spatial Pyramid Matching Kernel for Object and Scene Categorization

Carolina Redondo-Cabrera[1], Roberto J. López-Sastre, Javier Acevedo-Rodríguez, Saturnino Maldonado-Bascón

GRAM, Department of Signal Theory and Communications. University of Alcalá. Spain

Abstract

This paper proposes a novel approach to recognize object and scene categories in depth images. We introduce a Bag of Words (BoW) representation in 3D, the Selective 3D Spatial Pyramid Matching Kernel (3DSPMK). It starts quantizing 3D local descriptors, computed from point clouds, to build a vocabulary of 3D visual words. This codebook is used to build the 3DSPMK, which starts partitioning a working volume into fine sub-volumes, and computing a hierarchical weighted sum of histogram intersections of visual words at each level of the 3D pyramid structure. With the aim of increasing both the classification accuracy and the computational efficiency of the kernel, we propose two selective hierarchical volume decomposition strategies, based on representative and discriminative sub-volume selection processes, which drastically reduce the pyramid to consider. Results on different RGBD datasets show that our approaches obtain state-of-the-art results for both object recognition and scene categorization.

Keywords: 3D Spatial Pyramid Matching Kernel, object recognition, scene classification, point clouds, depth images.

Figure 1: Samples of depth images included in the New York University Depth database [1].

1. Introduction

We humans look at a picture and are able not just to see a pattern of color and texture, but to comprehend it, categorizing all the objects and even the scene itself. Can we do the same with simply a depth image? For instance, look at the images in Figure 1. Can't we localize and recognize the objects? Can't we categorize the scenes?

Object recognition in RGB images has seen huge progress in recent years, much thanks to the popular Bag of Words (BoW) approach [2, 3]. The brilliant idea behind this type of representation consists in characterizing an image by an orderless set of quantized local features, *i.e.* the well-known visual words. This approach has inspired a lot of research efforts which have obtained impressive results recently (*e.g.* [4, 5, 6, 7, 8]). Furthermore, the BoW model is the basic recipe for most of the methods submitted to the PASCAL VOC Challenge [9]. Methods using Support Vector Machines (SVMs) with Spatial Pyramid Matching Kernels (SPMKs) [10, 4] have been systematically obtaining the best results. The most recent improvements have been achieved by incorporating multiple local features such as SIFT [11], SURF [12] or color SIFT [6], into the BoW pipeline [13, 6].

So, we can say that the categorization problem in RGB images is a well established field of research. However, nowadays, we are witnessing how a new generation of depth cameras, such as Kinect, are capable of offering quality synchronized images of both color and depth information. The introduction of these sensors represents an opportunity to explore how to increase the capabilities of scene categorization and object recognition approaches (*e.g.* [1, 14, 15]).

In this paper, we propose to go beyond the traditional BoW representation for RGBD images, and propose a characterization for the point clouds associated to the

[1]Corresponding author: e-mail (crc04057@alu.uah.es), phone (+34 918856732). Office S-201, Polytechnic School. University of Alcalá. 28805, Alcalá de Henares (SPAIN)

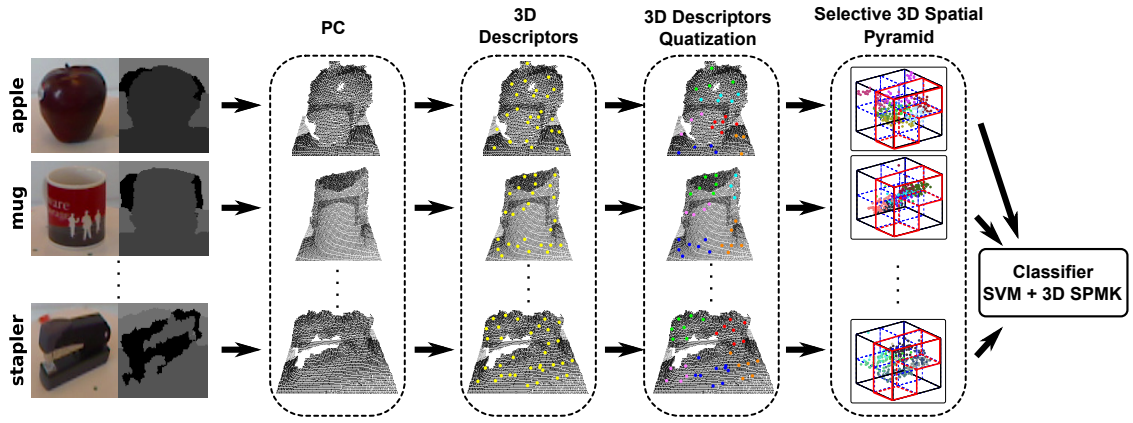


Figure 2: Proposed approach using the Selective 3DSPMK for object recognition in depth images. We quantize 3D descriptors, extracted from single depth images, into 3D visual words. This codebook is used to represent the objects in a BoW approach. The 3DSPMK repeatedly subdivides a cube inscribed in the 3D point cloud (PC), and computes a weighted sum of histogram intersections at increasingly fine sub-volumes. Selective volume decomposition strategies are proposed, based on representative and discriminative volume selection processes, which drastically reduce the volume to consider (see the red sub-volume selected), increasing both the classification accuracy and the computational efficiency of the kernel.

depth images.

We build a discriminative approach for recognizing object and scene categories in point clouds, which can simply use the information extracted from depth images. Inspired by the works of Lazebnik *et al.* [4] and Knopp *et al.* [16], we introduce a novel framework which uses 3D local features. This new methodology is depicted in Figure 2. We start extracting 3D local descriptors (such as 3D SURF [16] or NARF [17] descriptors) from a point cloud provided by a depth camera. Note that we do use a single depth image as input. These descriptors are then quantized, *e.g.* using K-means, so as to obtain a 3D visual vocabulary. We introduce a kernel-based image categorization approach, which works adapting the SPMK [4] to work in 3D, *i.e.* the 3D Spatial Pyramid Matching Kernel (3DSPMK). This novel strategy involves repeatedly subdividing a cube inscribed in the 3D point cloud, building histograms representations at increasingly fine sub- volumes, and computing a weighted sum of histogram intersections. We thoroughly explore how the 3D spatial binning and pyramids affect the performance, and propose selective hierarchical volume decomposition strategies, based on representative and discriminative sub-volume selection processes, which dramatically reduce the volume to consider, while jointly preserve the classification accuracy and increase the computational efficiency of the kernel.

A preliminary version of this paper appeared in [18], where we only addressed the problem of object recognition with 3D SURF descriptors in the RGB-D Object Dataset [15]. This paper contains a more detailed formulation of the Selective 3DSPMK approach. We also extend our approach to work with any type of 3D local descriptor (*e.g.* NARF [17]), and to solve novel problems, such as scene categorization. Additional experiments, using more RGBD datasets, have been incorporated as well. We also explore how to combine RGB and depth information into the same pipeline, in order to increase the global performance of the system.

The rest of this paper is organized as follows. Section 2 describes related work. Our novel approach, the 3DSPMK, is detailed in Section 3. Results are presented in Section 4. We finally conclude in Section 5.

2. Related Works

As there exists a large body of work on category-level object recognition and scene classification (*e.g.* [2, 4, 6, 5, 7]), we briefly review, in the following, only the most relevant to this paper, *i.e.* on image and scene categorization using point clouds, 3D shapes and depth images.

First, the problem of 3D shape class recognition has been extensively explored, and both local and global features have been proposed. A considerable variety of global descriptors have been detailed, such as shape moments [19] or shape histograms [20], for example. Neither partial shapes, nor intra-class variations are successfully handled by global descriptions. Moreover, using depth cameras, we do not get perfect scans of the environment, and we capture all the neighboring clutter in addition to the relevant data coming from the object

of interest. Hence global descriptors will be less successful at handling this type of data.

In the 2D case, it is well-known that the use of local features is beneficial for the object recognition problem. In the literature, there are also 3D shape and point cloud categorization methods using local features. For instance, Frome *et al.* develop the 3D shape context descriptors for object recognition in range data [21]. These descriptors are extracted in 3D from the point cloud which emerges from the depth image. The support region is discretized into bins, and a histogram is formed by counting the number of points falling within each bin. The obtained histogram is used directly as a descriptor. Mian *et al.* introduce the use of local tensors [22], while scale-dependent and scale-invariant local 3D shape descriptors are proposed in [23]. Toldo *et al.* [24] describe 3D shapes by splitting them into segments, which are then described on the basis of their curvature characteristics. These descriptors are then quantized into a visual vocabulary, using a SVM for classification.

Knopp *et al.* [16] introduce the 3D SURF descriptors in combination with a probabilistic Hough voting framework for the purpose of 3D shape class recognition. Our approaches also use their 3D SURF descriptors, but we propose to build a BoW based approach with them, in combination with the 3DSPMK for object categorization. Moreover, we extend their method to work on *partial* 3D shapes obtained from depth images and their corresponding point clouds.

Steder *et al.* [17] present a novel object detection and localization approach using 3D point cloud data, and introduce the Normal Aligned Radial Features (NARF), a novel interest point extraction method together with a feature descriptor for points in 3D range data. We also integrate their NARF descriptors into our 3DSPMK approach.

Recently, some approaches combine RGB and depth images so as to increase the performance in object categorization (*e.g.* [14, 15, 25]). For instance, in [15], Lai *et al.* benchmark the object categorization problem using a combination of RGB (SIFT [11]) and depth features (spin images [26]) in the very challenging large-scale RGB-D object dataset [15]. In [25], and using the same dataset, Lai *et al.* introduce a sparse distance learning approach for combining RGB and depth information for object categorization and instance recognition.

Bar-Hillel *et al.* [14] propose a system which fuses visual and range imaging for object category classification at multiple levels. For high-level fusion, existing techniques based on image-to-class nearest neighbors are used. For descriptor level fusion, they have abstracted existing descriptors into space and appearance histograms, trained by repetitive maximum likelihood optimization in growing image areas.

Gupta *et al.* [27] present a method for categorizing video sequences. Their hierarchical structure of histograms captures the typical spatial distribution of 3D points and codewords in the working volume and the scene is classified by SVMs equipped with a histogram matching kernel. Note that our approach significantly differs from this work. First, we do not need to attach to each 3D point an appearance based codeword obtained from RGB images: our approach is able to directly work with features extracted from the point clouds (or depth images). Second, our system does not need to apply any Structure From Motion algorithm to a video sequence to recover the geometry of the scene: we just need to integrate into our models the information provided by still depth images. Finally, our discriminative feature-based and representativeness-based decomposition mechanisms differ significantly from the occupancy-based decomposition strategy for the 3D reconstructions of video sequences.

As well as for the object recognition challenge, the scene classification problem has been extensively studied under many different settings in 2D images (*e.g.* [4, 6, 28, 29, 30]). Current scene recognition benchmarks [4, 30] define categories that are only relevant to the classification of single views of the scene. Only Xiao *et al.* [31] have introduced the problem of scene viewpoint recognition.

Within the context of scene recognition from RGBD images, it is very relevant the work of Silberman *et al.* [1], where they introduce the New York University (NYU) Depth dataset. The scene categorization problem is solved following the 2DSPMK approach in [4], using a standard BoW approach on SIFT descriptors extracted from RGB and depth images. For multi-class classification a SVM classifier is used. Additionally, for combining information from RGB and depth images, Silberman *et al.* [1] propose the following approach. SIFT descriptors are first extracted from both depth and RGB images. At that time, at each location, the 128 dimensional descriptors of both images are concatenated to form a single 256 dimensional descriptor. These descriptors are then used to build the 2D SPMK representation. In contrast to [1], our method builds a *3D representation* which lets us exploit the 3D information contained in the point clouds and depth images. That is, we do not use the 2D SPMK, but the novel 3D SPMK. Additionally, instead of extracting SIFT descriptors from depth images, we propose to use 3D fea-

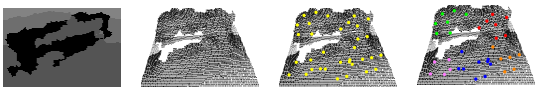
tures: 3D SURF [16] or NARF [17] descriptors, for instance. These feature extractors are able to extract relevant information directly from the point clouds. Moreover, for combining RGB and depth information, we propose to concatenate the RGB and depth pyramid structures instead of working at the descriptor level (as in [1]). In summary, we evaluate the performance of our novel 3DSPMK approach using the NYU benchmark, where we improve the state-of-the-art results previously reported on the problem of scene categorization.

3. Categorizing Point Clouds

Our goal is to learn models for object and scene categorization in point clouds. In this section, we detail our proposed category representation and introduce the 3DSPMK.

3.1. Category Representation

The feature extraction in our approach is shown in Figure 3. We start capturing a point cloud that contains the object (or the scene) of interest, from a single depth image, for example captured with the Kinect. Then, 3D local features are extracted (*e.g.* 3D SURF [16] or NARF [17] descriptors). For the 3D SURF descriptors we use the original implementation provided with [16]. In contrast to a global representation, by using, for example, a dense or random coverage with spin images [26], the 3D SURF [16] extractor is equipped with an interest point detector, where the descriptors are computed. The interest point detector picks out a repeatable and salient set of interest points in the shapes obtained from the point clouds. The local 3D SURF descriptors are computed in these points via uniformly sampling Haar-wavelet responses. NARF descriptors are also computed (using the original implementation [17]) calculating a normal aligned range value patch at each point, which is a small range image with the observer looking at the point along the normal. Then, a star pattern is overlaid onto the patch, where each beam corresponds to a value in the final descriptor that captures how much the pixels under the beam change. Finally, the algorithm extracts a unique orientation from the descriptor, and shifts the descriptor according to this value, in order to make it invariant to the rotation. The dimensionality of these descriptors used in our experiments is 36. Note that *any other* 3D local descriptor can be incorporated into our model. By following a traditional BoW approach, we quantize these descriptors, into 3D visual words. Each depth image can be then characterized by a histogram of its 3D visual words.



(a) (b) (c) (d)

Figure 3: 3D Descriptor extraction pipeline from a single depth image. (a) the depth image obtained by the Kinect sensor. (b) point cloud extracted from single depth image. (c) 3D descriptor features are extracted and back-projected to the partial 3D shape. (d) 3D descriptors are quantized into 3D visual words, therefore, following a traditional BoW approach, each depth image can be characterized by a histogram of its 3D visual words.

3.2. 3D Spatial Pyramid Matching Kernel

Nonlinear SVMs methods using SPMKs [10, 4] have been offering the best performances in object categorization systems. The original formulation of the pyramid matching strategy was introduced in [10]. The idea of pyramid matching consists in mapping a set of features to multi-resolution histograms. Then, a comparison between histograms is carried out using a histogram intersection function so as to approximate the similarity of the best partial matching between features sets. Grauman and Darrell [10] demonstrated that the histogram intersection and the pyramid match kernels satisfy the Mercer's condition, so they can be used in kernel-based algorithms based on convex optimization, such as SVMs.

Based on [10], Lazebnik *et al.* [4] introduced a different approach for image categorization: the SPMK. They propose to perform the pyramid matching in the two-dimensional image space, while using traditional quantization techniques in feature space.

Inspired by [4], we propose to extend the SPMK to the three-dimensional space, *i.e.* the 3DSPMK. As it was described in Section 3.1, we model a point cloud in 3D by an orderless set of 3D visual words. That is, if we define a visual codebook of size K, each 3D feature is associated to a codebook label $\{1, \ldots, K\}$. The 3DSPMK should be able to capture the spatial distribution of such codewords at different scales and locations in a working volume $\Omega^{(0)}$. Before building the pyramid structure, in order to achieve a spatial distribution of 3D local features that occupies the greatest possible proportion of volume in the working cube $\Omega^{(0)}$, we simply perform a centering and scaling process of the initial spatial distribution of the features. This process is detailed in Figure 4.

Once the centering and scaling processes have been performed, we define a pyramid structure by partitioning the working volume $\Omega^{(0)}$ into fine sub-cubes. For each level l, the volume of the previous level, *i.e.* $\Omega^{(l-1)}$, is decomposed into eight sub-cubes (see Figure 5). It

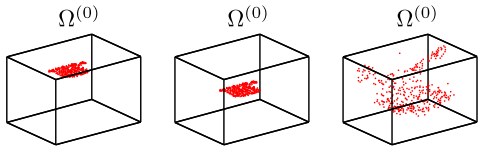


Figure 4: Example of centering and scaling process of a spatial distribution of 3D local features. In the first $\Omega^{(0)}$ cube, the initial spatial distribution of the features is represented. Second $\Omega^{(0)}$ cube shows a centered spatial distribution of features. This spatial distribution is then scaled to fit the whole $\Omega^{(0)}$ cube. The final result can be observed in the third $\Omega^{(0)}$ cube.

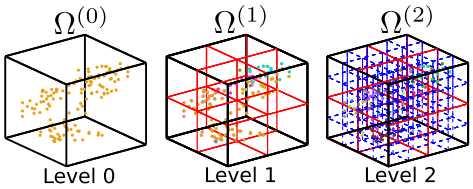


Figure 5: Example of a 3D spatial pyramid of three levels. The working cube $\Omega^{(0)}$ is recursively decomposed into eight sub-cubes.

is straightforward to see that, in our formulation, if we build a pyramid of L levels, $P(L)$, it will have $D = 8^L$ sub-cubes.

When the pyramid decomposition of L levels is processed, we perform the pyramid matching in 3D. Let define $H^l_{X_i}$ and $H^l_{Y_i}$ as the histograms of features for depth images X and Y that fall into the ith sub-cube in the pyramid $P(L)$ for the level l, *i.e.* $\Omega^{(l)}_i$. Only features of the same type can be matched. So, the number of matches at level l into the ith sub-cube is given by the histogram intersection function as follows

$$\mathcal{I}(H^l_{X_i}, H^l_{Y_i}) = \sum_{j=1}^{K} \min(H^l_{X_i}(j), H^l_{Y_i}(j)) \qquad (1)$$

where K is the number of components of histograms H_X and H_Y, and $H^l_{X_i}(j)$ represents the value of the $j-th$ bin of the histogram into the ith sub-cube at level l.

Accordingly, the number of matches at level l can be obtained by

$$\mathcal{I}(H^l_X, H^l_Y) = \sum_{i=1}^{S} \mathcal{I}(H^l_{X_i}, H^l_{Y_i}) \qquad (2)$$

where S represents the number of sub-cubes at level l.

The 3DSPMK is then defined as the following sum of weighted histogram intersections

$$K(X, Y) = \omega_0 \mathcal{I}(H^0_X, H^0_Y) + \sum_{l=1}^{L} \omega_l \mathcal{I}(H^l_X, H^l_Y)\,, \qquad (3)$$

where, w_l is set to $\frac{1}{2^{L-l}}$, *i.e.* a weight which is inversely proportional to the cell width at that level l. By doing so, we penalize those matches found in larger volumes, because they may involve increasingly dissimilar features.

We do normalize all histograms involved in our 3DSPMK representation. We use the number of descriptors extracted in each depth image to normalize the histograms, in effect forcing the number of features extracted in all images to be the same.

Note that one of the limitations of the 3DSPMK, as for the 2DSP [4], is its lack of invariance to rotation, that depends on the number of levels of the pyramid. The higher levels of the pyramid, *i.e.* $L > 0$, sacrifice the geometric invariance properties of BoW, *i.e.* $L = 0$, but these levels compensate this loss with increased discriminative power derived from the global spatial information. Essentially, when $L = 0$, our 3DSPMK is a standard (3D) BoW approach. Such an approach is invariant to rotation, only if the extracted local descriptors are also invariant under rotation and scale, which is the case for the 3D SURF or NARF descriptors used in this work. When the number of pyramid levels is increased, the spatial pyramid matching tends to "zero in"on these higher levels that contains the most discriminative spatial information. If a dataset happens to be so highly variable that the global position of features yields no useful cues at all, the matching scheme will simply "fall back"on level 0 ($L = 0$), which is equivalent to an orderless BoW.

Furthermore, our 3DSPMK representation is particularly interesting for the problem of scene recognition in depth images since it augments BoW representations with global spatial relations. For similar scenes, similar features tend to repeatedly appear in similar locations (floor, ceiling, ...), and our representation is able to capture this spatial distribution similarity. In the experiments, we analyze the performance of our method for this task with the NYU Depth [1] scenes dataset (see Section 4.2.1). In conclusion, our thorough experimental validation reveals that the proposed 3DSPMK is a discriminative and spatial representation based on aggregating statistics of local features over fixed subvolumes, which is able to compensate the lack of geometric invariance reporting state-of-the-art results for object and scene recognition problems.

3.3. Selective 3DSPMK

So far, our formulation can be seen as an extension of the original SPMK [4] to 3D. Although the computational complexity of histogram intersection operations is linear in the number of features, one clear disadvantage of the pyramid decomposition proposed is its high computational cost. For a pyramid of L levels and K features, we will obtain a vector of dimensionality $K\sum_{l=0}^{L} 8^l$, that is 2^l times more bins in each level with respect to to the SPMK [4]. With the aim of jointly increasing the classification accuracy and the computational efficiency of the 3DSPMK, we introduce two *selective* volume decomposition schemes based on representative and discriminative sub-volume selection processes.

3.3.1. Representativeness-based Selection

With the aim of increasing the computational efficiency of our approach, rather than simply decomposing the working volume as it was described in Section 3.2, we have designed the following selective approach. Our target consists in incorporating into the pyramid, *only* those sub-cubes that are likely to represent images in our dataset.

Unlike the 2D case, where we can consider a uniform distribution of local features across the whole 2D pyramid (specially if a dense feature extraction is carried out), in our 3D formulation, the local features occupy sparse locations in the 3D working volume. Furthermore, the higher the level of the pyramid, the higher the number of empty sub-cubes within it. So, our objective is to reduce the large number of these *uninformative* sub-cubes that yield unnecessary long histograms.

Let $\Omega^{(0)}$ be the working cube for level zero. We first perform the pyramid decomposition until level L, so we obtain $\Omega_i^{(L)}$ sub-volumes, where $i = 1, \ldots, 8^L$. We now redefine the working volume of level zero as $\hat{\Omega}^{(0)}$, where the decomposition only includes those sub-cubes $\hat{\Omega}_i^{(L)}$ in which a percentage p of the images are represented. With this selection criterion our approach ignores the uninformative sub-cubes, and thus those features that fall into them. We consider that an image I is represented if there is at least one feature of I falling in the sub-volume. The value of p can be determined empirically in the experiments. We perform this selective pyramid decomposition just once at the beginning of the training, and use a set of N randomly selected images per object category, for computing the representativeness-based selection. A toy example of this process is shown in Figure 6.

Once the new volume $\hat{\Omega}^{(0)}$ has been computed, we can define the associated pyramid $\hat{P}(L)$, where we can compute the histogram $\hat{H}^l_{X_i}$ of the features that fall into the ith sub-cube $\hat{\Omega}_i^{(l)}$ at level l. These histograms will be used in Equations 2 and 3.

3.3.2. Discriminative Feature-based Selection

The representativeness-based selective method drastically reduces the working volume. However, it does not exploit the fact that the volume selected may contain features that are not discriminative for the classes of interest. In this section, we propose the complementary discriminative feature-based decomposition, where the objective is to select those cubes that are likely to contain discriminative features. Our objective is two-fold: continue reducing the working volume, and improve the classification performance.

We start considering the following question: how can we measure that a particular feature is discriminative enough for a particular class?

We are given a set of images. Each image belongs to a class n, being N the total number of classes. As it has been described, we build a visual codebook of size K from 3D local descriptors extracted from the images of all the classes. Our notation is based on a set of features $\mathcal{F} = \{f_1, f_2, \ldots, f_K\}$, which form the visual vocabulary, and a set of measurements X_j extracted from the images. That is, for a set of 3D descriptors, X_j, we assign each one to a feature $f_k \in \mathcal{F}$. For each class n, we define M_n as the total number of descriptors extracted from the images of the class n. We also define $m_n^{(f_k)}$, as the number of descriptors for the class n assigned to feature f_k.

So, for a given visual codebook of size K, and a set of N different classes, we introduce a feature scoring technique which shall define the score matrix S, of size $N \times K$, where each score $s_{nk} = S(n,k)$ is computed as follows

$$s_{nk} = \Delta_k \frac{m_n^{(f_k)}}{M_n}, \tag{4}$$

where,

$$\Delta_k = \left(\sum_{n=1}^{N} \frac{m_n^{(f_k)}}{M_n} \right)^{-1}. \tag{5}$$

The score s_{nk} is the ratio between the percentage of descriptors that belong to the feature k in the class n, and the proportion of descriptors that belong to the feature k when all the categories are considered simultaneously.

Once the score matrix S has been computed, we define a threshold τ for considering whether a feature is discriminative for a class. We then obtain the binary matrix S', where,

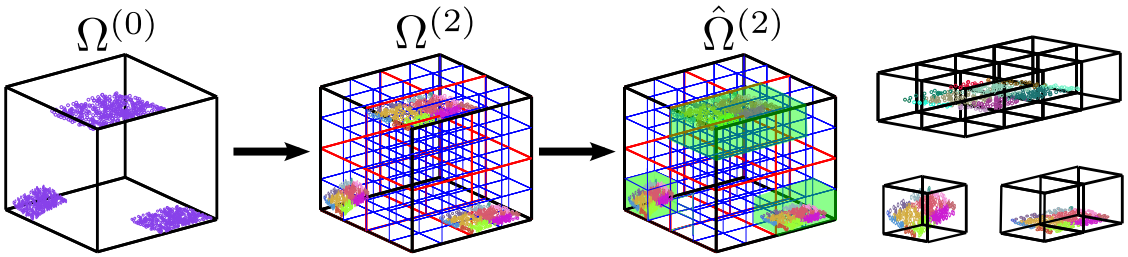


Figure 6: Toy example of the representativeness-based volume decomposition for the 3DSPMK. The representative features fall in the green sub-volume selected.

$$s'_{nk} = \begin{cases} 1 & \text{if } s_{nk} \geq \tau \\ 0 & \text{if } s_{nk} < \tau \end{cases} \quad (6)$$

Our next step consists in propagating this discriminative analysis from the feature-level to the pyramid-level. The question we want to address now is: how do we consider that a sub-cube $\Omega_i^{(l)}$ is discriminative?

Given a pyramid of L levels $P(L)$, we inspect all sub-volumes in level L, *i.e.* $\Omega_i^{(l)}$ for $i = 1, \ldots, 8^L$. For each object class and each sub-volume, we measure the proportion of images that contain at least one *discriminative* feature in each sub-volume, and we define this measure as $\mathcal{R}(\Omega_i^{(l)})$. If $\mathcal{R}(\Omega_i^{(l)}) > \beta$, where β is an empirically fixed threshold, then the sub-volume $\Omega_i^{(l)}$ is considered as discriminative for the analyzed object class. The final discriminative decomposition is obtained merging all the discriminative sub-volumes for each category.

Note that we can run this selection procedure on top of either the original pyramid decomposition, $\Omega^{(L)}$, or the pyramid decomposition selected by the representativeness-based criterion. Furthermore, both selective mechanism can be run in parallel, and then define as the definitive decomposition, the intersection of the two solutions. In general, our experiments show that the discriminative feature-based selection approach leads to more compact description, *i.e.* the number of sub-volumes selected by discriminative-based decomposition is normally lower than the number of sub-cubes selected by the representativeness-based approach.

4. Experiments

In this section, we present two applications of the proposed Selective 3DSPMK approach, namely object recognition (Section 4.1) and scene categorization (Section 4.2).

4.1. Object Recognition

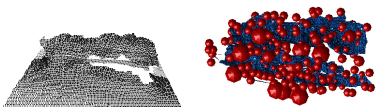
For the object recognition problem in depth images, we propose to use two challenging and novel datasets: the RGB-D Object dataset [15], and the dataset collected by Bar-Hillel *et al.* [14]. With the aim of really evaluating the performance of the 3DSPMK model, for these experiments, only depth images are going to be processed.

The 3D local features that we are going to use for the object recognition experiments are the 3D SURF descriptors introduced by Knopp *et al.* [16].

In order to characterize each depth image, we proceed as follows. We start computing the point cloud associated to each depth image (see Figure 7(a)). The 3D SURF descriptors were designed to be computed on 3D shapes, so, our next step consists in generating the partial 3D shape associated with each point cloud. For doing so, we follow the greedy surface triangulation method in [32][2] for each point cloud. The algorithm works by maintaining a list of points from which the mesh can be grown and extending it until all possible points are connected. Triangulation is performed locally, by projecting the local neighborhood of a point along the point's normal, and connecting unconnected points.

Once the partial shape has been obtained, the 3D SURF descriptors can be computed. First, each partial 3D shape is uniformly scaled to fit a cube with a side of length 256. Then, 3D SURF descriptors of 162 dimensions are computed using the original implementation provided in [16]. With the aim of covering the full partial 3D shape with 3D SURF descriptors, we have experimentally chosen the following parameters: the distance between the triangle mesh and the border of the cube is

[2] We have used the following parameters: number of neighborhood points = 100, maximum distance between neighborhood points = 2.5, minimum angle in each triangle = 10°, maximum angle in each triangle = 120°, maximum surface angle = 45°.



(a) (b)

Figure 7: 3D SURF extraction pipeline from a depth image. (a) A point cloud is obtained from the depth image. (b) We process the point cloud in order to obtain a partial 3D shape. 3D SURF features are extracted and back-projected to the partial 3D shape.

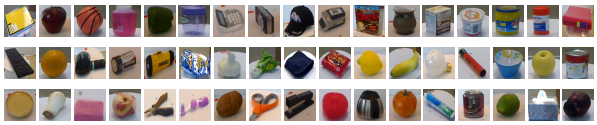


Figure 8: Object instances from RGB-D Dataset [15]. One example for each of the 51 object categories is shown.

fixed to 30, and the threshold is fixed to 10^{-8}. A result of this 3D SURF extraction step is shown in Figure 7(b).

4.1.1. RGB-D Object Dataset

Experimental Setup. The RGB-D Object Dataset [15] is a large scale collection of images, which contains 300 objects organized into 51 categories. The dataset provides between 3 to 12 instances in each category. The images were collected with a RGB-D sensor that simultaneously records both color images and depth data at 640×480 resolution. The dataset provides 250.000 RGB+Depth images in total, which were recorded from 3 different zenith directions and 250 azimuth angles. Figure 8 shows examples of objects of all the categories in the RGB-D Object database. As we can see, each image contains only a single object and it has little or no clutter.

We evaluate our object categorization approach on this dataset, following the same experimental setup described in [15]. For the experiments, we use all the 51 categories. We subsample the turntable data by taking every fifth video frame. During categorization, we randomly leave one object out from each category for testing, and train the classifier using the 3DSPMK on all the views of the remaining objects. The final result is reported as the average per-class recognition rate. Additionally, we present confusion matrices for the 51 categories used.

In the experiments, we use a visual vocabulary of different sizes ($K = 200$, $K = 800$, $K = 1000$). The visual codebook is obtained performing a K-means clustering on a subset of the 3D SURF descriptors (taking the 3D SURF descriptors of 50 images per class). We represent each image by a 3D spatial pyramid. Typical pyramid level values for our experiments are $L = 0, 1, 2$. Note that when $L = 0$, we just have a standard BoW, but in our case in 3D. We report the performance of the 3DSPMK using the full volume of the pyramid and also following the selective algorithms described in Section 3.3.

For classification we use SVMs. The kernel function used is our novel 3DSPMK, which was detailed in equation (3). The multi-class classification problem is solved training the SVM using the one-against-one strategy. We follow the approach in [33], and train $N(N-1)/2$ classifiers (being N the number of classes) where each one is trained on data from only two classes. For testing, we follow the *Max Wins* voting strategy [33]: if one of the classifiers votes for class with index i, then the vote for the i-th class is added by one. The class with the highest number of votes is selected for each image. In case that two classes have identical votes we select the one with smaller index i. Specifically, we use libSVM [34] for training each binary classifier. A 10-fold cross-validation on the training set to tune SVM parameters is conducted.

In this dataset, we have designed *two* types of experiments. First, and in order to *strictly* follow the experimental setup described in [15], we integrate the same automatic object segmentation proposed in [15] in our feature extraction pipeline. Second, we also run some experiments without using any object segmentation algorithm, *i.e.* we let our approach to work with depth images which contain not only the object of interest, but clutter coming from the rest of the scene. We consider this a harder problem.

Results with automatic object segmentation. Following [15], when the object has to be automatically segmented, we use the known distance between the turntable and the camera, so as to remove most of the background points by taking only the points within a 3D bounding box, *i.e.* the working volume, where we expect to find the turntable and the object. We clean the turntable points by running a RANSAC [35] fit plane algorithm on the point cloud. Following this automatic procedure, we obtain clean point clouds for all the object classes in the dataset, as in the experimental setup detailed in [15].

Table 1 shows the results obtained by our approaches, as well as a comparison with the state-of-the-art methods [15, 25].

First, let us analyze the performance of the 3D shape features used, *i.e.* quantized 3D SURF descriptors. For a pyramid of level 0 (3DSPMK ($L = 0$)), our average classification rate for the 51 classes is 52.8%. If we

Table 1: Classification Accuracy of different approaches on the RGB-D Object dataset. EBLocal (Exemplar-Based Local distance learning), LinSVM and kSVM (Linear and Gaussian kernel SVM), RF (Random Forest), IDL (Instance Distance Learning) and 3DSPMK with $K = 800$ and different pyramid levels ($L = 0$, $L = 1$ and $L = 2$).

Classification Accuracy	
Method	Shape
[25] (LinSVM)	53.1
[25] (EBLocal)	58.9
[15] (kSVM)	64.7
[15] (RF)	68.8
[25] (IDL)	**70.2**
3D SPMK with automatic object segmentation	
3D SPMK ($L = 0$)	52.8
3D SPMK ($L = 1$)	57.2
3D SPMK Full Volume ($L = 2$)	64.5
3D SPMK Representativeness ($L = 2$)	**67.8**
3D SPMK Discriminative Feature ($L = 2$)	60.1
3D SPMK without automatic object segmentation	
3D SPMK ($L = 0$)	59.8
3D SPMK ($L = 1$)	66.3
3D SPMK Full Volume ($L = 2$)	66
3D SPMK Representativeness ($L = 2$)	**69.4**
3D SPMK Discriminative Feature ($L = 2$)	67.3

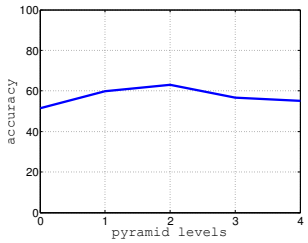


Figure 9: 3D SPMK with automatic object segmentation. Classification Accuracy for different pyramid levels and $K = 600$.

compare with state-of-the-art results, when *only* depth features are used, we can see that the spin image representations (RF, kSVM, LinSVM, EBLocal and IDL) [15, 25] (Table 1, rows 1-5) work slightly better than our codebooks of quantized 3D SURF local descriptors. If the performance of the 3DSPMK is examined, Table 1 shows that the results always improve as the pyramid level goes from $L = 0$ to $L = 2$. Our best result 67.8% is obtained with a pyramid of level $L = 2$, a codebook of size 800, and using the representativeness-based approach. Note that the 3DSPMK is able to report state-of-the-art results, even improving some of the previous results reported in the RGB-D Object dataset when just shape features were used ([25] (LinSVM, EBLocal), [15] (kSVM)). This confirms the discriminative power of the pyramidal decomposition proposed.

We have experimentally observed that the performance starts to decrease for pyramid with $L \geq 3$ (see Fig. 9). It is worth to mention that a similar behavior was observed in [4] but for the 2DSPMK. We explain this by the fact that for a pyramid with $L \geq 3$ the highest level is too finely subdivided in sub-volumes, which yield too few matches between features within them. To summarize, when using 3DSPMK with depth images a good choice is to use $L = 2$.

We follow analyzing the performance of our selective decomposition strategies and how their parameters (namely, p, τ and β) affect the final behavior of the system.

For this experiment, with the representativeness-based approach, using $p = 0.1$ and the same visual codebook used with the standard 3DSPMK, the selective pyramid $\hat{\Omega}^{(2)}$ is reduced to 61 sub-cubes from a total of 64, while the accuracy increases from 64.5% to 67.8%. These results reveal that, although the computational boost is not very prominent, our selective method is able to eliminate those three noisy sub-volumes while improving the 3DSPMK categorization performance.

When we use the discriminative-based approach, with τ and β fixed to 0.7 and 50%, respectively, $\hat{\Omega}^{(2)}$ includes only 32 sub-volumes. It is worth to mention that even with such a dramatic reduction in the volume to consider, the performance of the pyramid does not significantly decrease.

Experimentally, we have observed that the discriminative feature-based method is always more compact. This decomposition method includes less sub-cubes than the representativeness-based approach. Additionally, its selected sub-volume is normally included within the sub-volume selected by the representativeness-based algorithm. So, the results of the discriminative feature-based strategy are equivalent to the results obtained by an intersection of the two selected sub-volumes. For example, in an experiment with $p = 0.1$, and τ and β fixed to 0.7 and 50%, respectively, if we intersect the sub-volumes obtained by the two selective methods, the final working volume $\hat{\Omega}^{(2)}$ includes only 32 sub-cubes. Furthermore, the sub-cubes selected are the same cubes selected by the discriminative feature-based approach.

Our experiments also reveal that parameter β is not critical for the classification performance, but it is crucial in terms of computational efficiency. For the experiment in Figure 10(a), we fixed parameter τ to 0.7 and vary β from 30% to 70%. While the number of sub-cubes increases from 23 to 48, the classification accuracy does not significantly vary. This fact shows that not all features have the same relevance, and our selective algorithm is able to include those sub-cubes that contain relevant features for the classification performance, while the dimensionality of the final representation is drastically reduced.

Parameter τ evaluates whether a feature is discriminative enough for a particular class. As it is shown in Figure 10(b), the influence of τ in the classification rate is more relevant than that of β. If we choose a strong

discriminative threshold, *i.e.* $\tau > 0.7$, the final working volume is almost empty, and the classification accuracy tends to zero. However, if we select a less restrictive threshold, *i.e.* $\tau < 0.7$, the classification rate increases. For a good balance between classification performance, runtime and memory cost, we recommend to use the following values: $\tau < 0.7$ and $\beta < 0.5$.

With respect to the representativeness-based selection strategy, we have observed the following behavior. The representativeness threshold p is critical for the classification performance, see Figure 10(c). If threshold $p > 0.3$, the performance decreases, due to the high number of sub-cubes that are discarded. Our best result is obtained fixing $p = 0.1$. Observe that for $p = 0.3$, the performance does not decrease, compared with the classification accuracy reported when the whole working volume is used ($p = 0$). Again, as a recommendation, a good choice is to work with $p < 0.3$.

The RGB-D Object database is a large scale dataset, so it is also relevant to analyze how our methods perform for each particular class. Figure 11 shows the results reported per class for each of our approaches. First, it is important to note that, for pyramids with $L = 2$ we obtain a classification rate higher than 90% for $\approx$ 30% of the classes. For $L = 0$ only 10% of the classes attain an average accuracy higher than 90%. This fact shows, once more, the improvement obtained when the 3DSPMK approach is introduced in the pipeline. Figure 11 also shows that the higher the pyramid level, the higher the maximum classification rate for most of the classes. For example, we reach a 100% for the followings: *soda can*, *plate*, *cereal box* and *binder*.

In Figure 13, we also show confusion matrices for the 51 categories. In general, the higher the pyramid level, the lower the confusions. Our approach uses shape-based features, so it is straightforward to understand that the confusion between classes with a similar shape might be high. For instance, this is what happens for classes *apple* and *tomato*, or *ball* and *orange* (see Figure 13(d)). The poor performance of some classes, such as *mushroom*, is due to the low number of training/testing instances (3 or 4).

Results without automatic object segmentation. Our approaches are also able to perform visual categorization in the wild, *i.e.* to work with the whole point cloud, without using any automatic object segmentation approach which makes use of *a priori* knowledge of the scene (as [25]).

For these experiments, we follow exactly the same experimental setup described, but now considering the whole depth images provided, *i.e.* no automatic object segmentation is applied.

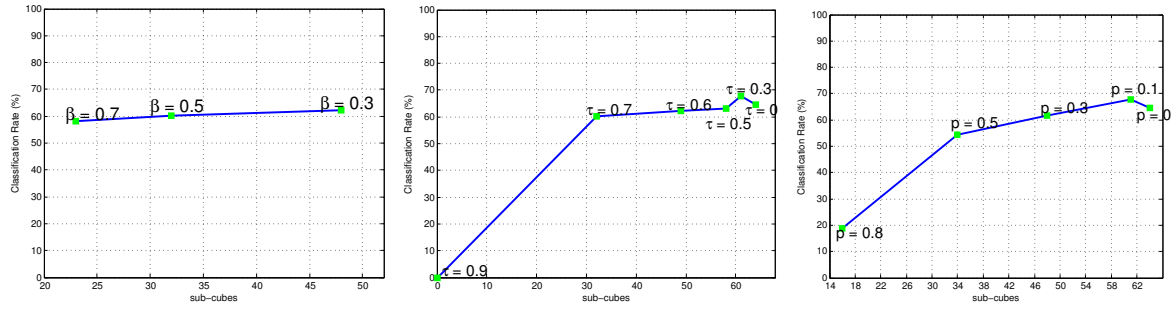
Table 2: Classification Performance on the RGB-D Object dataset using a 3D SPMK with $L = 0$ and different codebook sizes. Without object segmentation.

Classification Accuracy		
Codebook Size		
$K = 200$	$K = 800$	$K = 1000$
52.76	**59.76**	58.42

First, we inspect how the performance changes when different vocabulary sizes are used. For doing so, we fix the level of the pyramid to $L = 0$ and vary K. Table 2 shows that the best classification performance is obtained for a visual vocabulary size of $K = 800$.

Using the vocabulary of size $K = 800$, we now run our approach for different pyramid levels. Again, Table 1 shows that the results improve as the pyramid level goes from $L = 0$ to $L = 2$. The best result has been obtained with a pyramid with $L = 2$ and the representativeness-based approach. We report results using the representativeness-based and the feature discriminative-based approaches. For the former, with $p = 0.1$, the selective pyramid $\hat{\Omega}^{(2)}$ contains 56 sub-cubes of the 64. For the latter, with τ and β fixed to 0.7 and 50%, respectively, $\hat{\Omega}^{(2)}$ includes *only* 39 sub-volumes. We can say that, by following our selective decomposition strategies, the classification rate and the computational efficiency jointly increase. Note that the discriminative-based approach works only with 60% of sub-cubes, and it just loses a 2% of classification rate with respect to the representativeness-based approach.

With or without object segmentation? Figure 12 compares the performances of the 3DSPMK with and without the automatic object segmentation. The first conclusion we draw is that for pyramids with levels $L = 0, 1$, the classification rate increases (from 53% to 60% with $L = 0$, and from 57.2 to 66.3 with $L = 1$) when no automatic object segmentation is used. We think this increase is related to the impreciseness of the automatic segmentation process, where we lose local descriptors that can help in the recognition task, specially in the object boundaries. As soon as we increase the pyramid level, *e.g.* for $L = 2$, the results of both approaches are comparable, except for the discriminative-based approach. The best results are obtained for $L = 2$ using a representativeness-based approach when no automatic object segmentation is done (69.4%). In general, the experimental evaluation shows that we are able to report state-of-the-art results without introducing additional object segmentation algorithms in our pipeline, a fact that we consider an important contribution of our work.



(a) (b) (c)

Figure 10: Analysis on RGB-D Object dataset. Classification Accuracy vs number of sub-cubes when different Selective 3DSPMK approaches are used. (a) Classification rate and number of sub-cubes when the parameter τ is fixed to 0.7 and several discriminative thresholds, β, are used. (b) Classification rate and number of sub-cubes for different τ values, when β is fixed to 0.5. (c) Classification rate and number of sub-cubes when several representativeness thresholds, p, are used.

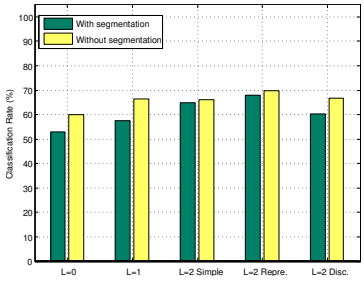


Figure 12: Classification performance of the 3DSPMK with and without automatic object segmentation.

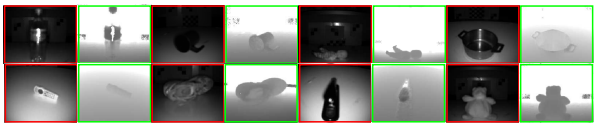


Figure 14: Some examples of intensity (red square) and depth (green square) images from Bar-Hillel *et al.* Dataset [14].

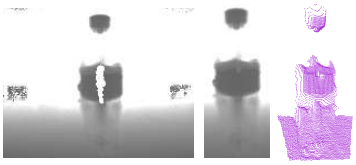


Figure 15: Preprocessing operations to the depth images. (Left) Captured depth image. (Middle) Preprocessed depth image. (Right) The cleaned 3D point cloud.

4.1.2. Bar-Hillel et al. Dataset [14]

Experimental Setup. The data in this dataset have been captured using a Baumer camera, which projects near-infrared light, and captures two 144 × 176 perfectly aligned images, containing range information and the light intensity as captured through the cameras near-infrared band filter. The dataset, as it is shown in Figure 14, consists of 8 classes of everyday objects (cup, bottle, doll, teddy bear, remote control, shoe, stapler, and pot), each with 10 objects per class. Each object was captured from 2 camera positions (side and upper side view), 3 object poses, and 2 illumination conditions, for a total of 12 images. Overall, the data offers 960 image pairs (depth and intensity).

We test our approach on this dataset following the experimental setup described in [14]. We randomly split the dataset into train and test sets with five different objects per class in each, and the results reported are averages over 5 such train-test splits. Due to the range camera is very sensitive to reflective properties of a surface, for transparent and specular objects (*e.g.* bottles, cups and pots), and on surfaces perpendicular to the camera, depth estimation is problematic. Following the steps described in [14], we apply several preprocessing operations to the depth images before obtaining the definitive 3D point clouds. These preprocessing operations are shown in Figure 15.

Results. Following [14], we do not perform any automatic object segmentation algorithm in the feature extraction pipeline. So, we let the 3DSPMK work directly with the full point clouds provided. We report results using different vocabulary sizes. In Table 3, we can see that the best classification performance, for a pyramid of level $L = 0$, is achieved by a codebook size of $K = 800$, so for the rest of experiments we used this codebook.

As the proposed discriminative-based selection criterion depends on empirically estimated parameters (τ and β), which are closely related to the spatial distri-

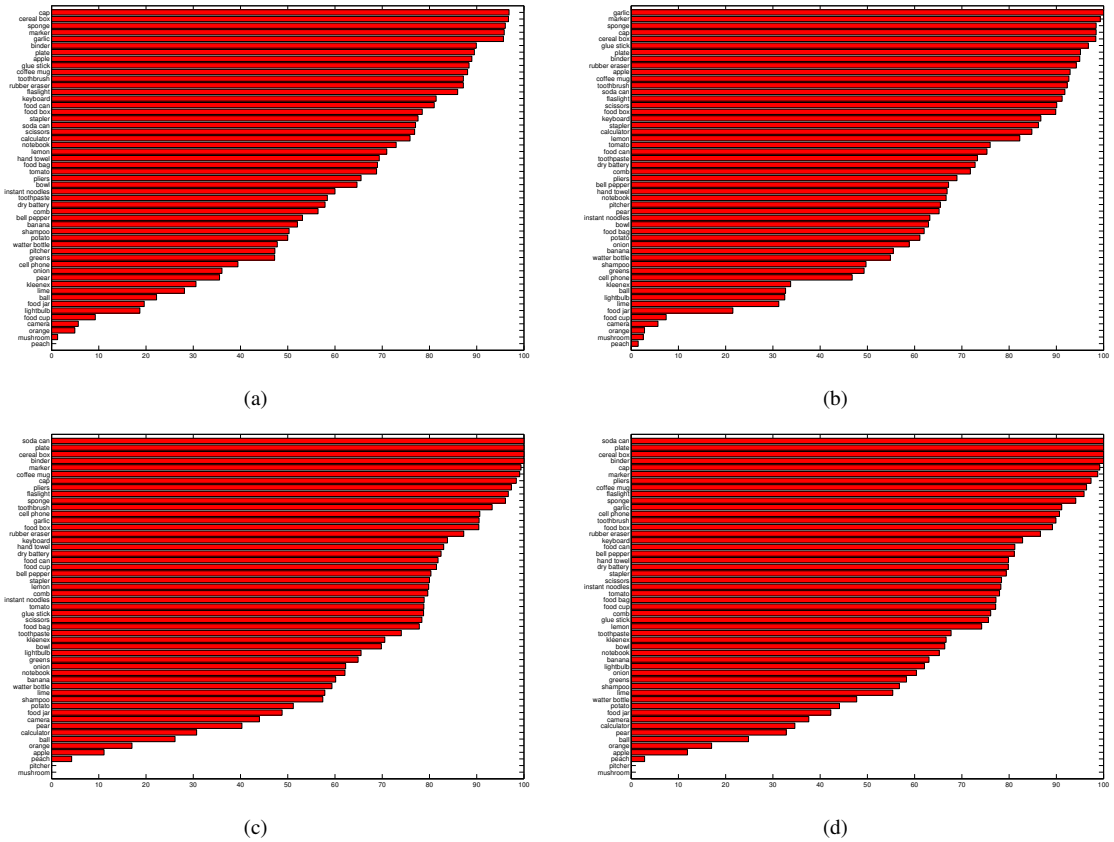


Figure 11: Classification accuracy for each category. We use a vocabulary size of $K = 800$, and the 3DSPMK with (a) $L = 0$, (b) $L = 2$, and $L = 2$ with the representativeness-based and the discriminative feature-based methods, in (c) and (d), respectively. This figure is best viewed with magnification.

Table 3: Classification Performance on Bar-Hillel *et al.* dataset using a 3DSPMK with $L = 0$ and different codebook sizes.

Classification Accuracy		
Codebook Size		
$K = 200$	$K = 800$	$K = 1000$
59.1	**66**	63.75

butions of feature points in the datasets, we repeat the same analysis than in Section 4.1.1 to inspect the influence of these parameters. Figure 16 shows the results obtained by different β and τ criteria. Note that again in this dataset most objects are well aligned, hence the histograms over dictionary features are quite similar in most representative regions, which makes β less important than τ. Once more, the best classification performances are obtained for the following values: $\tau < 0.7$ and $\beta < 0.5$.

Table 4 shows that our results improve as the pyramid level goes from $L = 0$ to $L = 2$. The best result has been obtained by pyramids with $L = 2$ using the full volume of the pyramid. Notice that for the

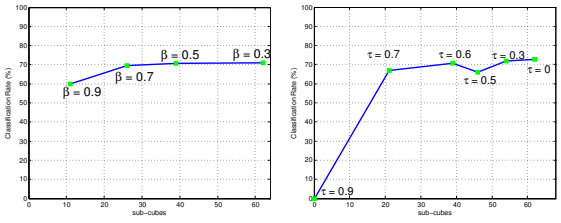
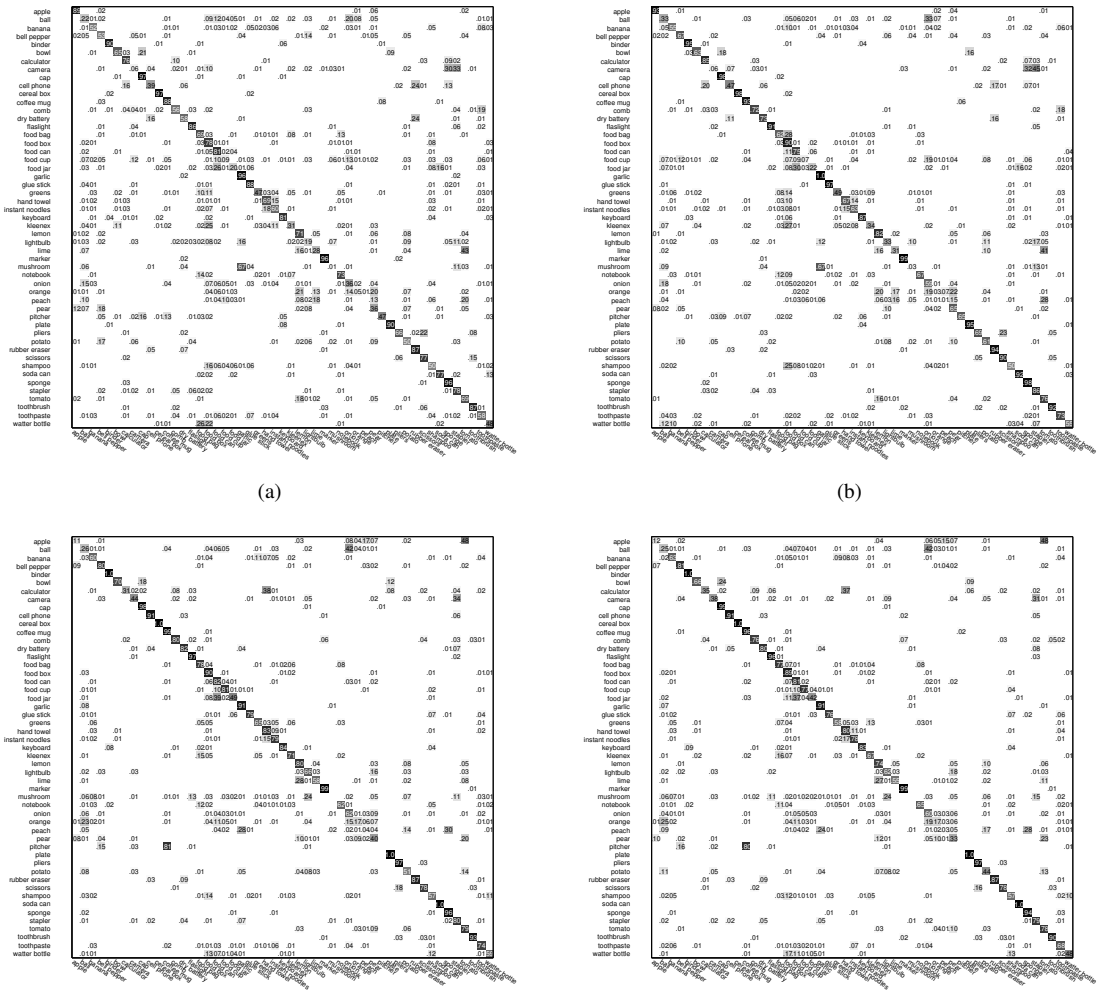


Figure 16: Analysis on Bar-Hillel *et al.* dataset. Classification Accuracy vs number of sub-cubes when Discriminative Selective 3DSPMK approach is used. (a) Classification rate and number of sub-cubes when the parameter τ is fixed to 0.7 and several discriminative thresholds, β, are used. (b) Classification rate and number of sub-cubes for different τ values, when β is fixed to 0.5.

representativeness-based approach, with $p = 0.1$, the selective pyramid $\hat{\Omega}^{(2)}$ includes only 49 sub-cubes (of 64), and for the feature discriminative-based approach, with τ and β fixed to 0.6 and 50%, respectively, $\hat{\Omega}^{(2)}$ includes only 39 sub-volumes. Therefore, these results



(a) (b) (c) (d)

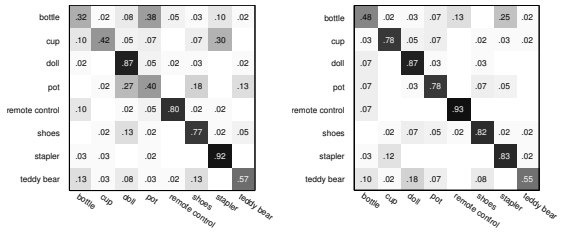
Figure 13: Confusion matrices for the 51 categories in the RGB-D Object database. Average classification rates for individual categories are listed along the main diagonal. Results for the 3DSPMK with $K = 800$ (a) $L = 0$, (b)$L = 2$, and $L = 2$ with the representativeness-based and the discriminative feature-based methods, in (c) and (d), respectively. This figure is best viewed with magnification.

show that, by following our selective decomposition strategies, the classification rate can remain constant while the computational efficiency increases. Again, the feature discriminative-based approach is the more compact, *i.e.* it selects lower number of sub-volumes than the representativeness-based approach, 39 vs 49, respectively. We found that unlike our Selective 3DSPMK approaches, the Bar-Hillel *et al.* features and classifiers obtain a high accuracy at classifying bottle and teddy bear classes. Our low performance for these two classes (see our confusion matrices in Figure 17), may help to explain the results reported in this dataset.

Evaluation per object class. In order to show the confusion between classes, we also report confusion

Table 4: Classification Performance of different approaches on the Bar-Hillel *et al.* Dataset. The results termed to single depth modalities, SIFT D-D and 3D-SC (more details in [14]). Selective 3DSPMK with a codebook of size $K = 800$ and different pyramid levels.

Classification Accuracy	
Single Depth Modalities	
3D-SC descriptors [14]	87.9
SIFT D-D descriptors [14]	80.3
3D SPMK	
3D SPMK ($L = 0$)	$65.6^{+}1.75$
3D SPMK ($L = 1$)	$67.2^{+}1.16$
3D SPMK Full Volume ($L = 2$)	$72^{+}1.5$
3D SPMK Representativeness ($L = 2$)	$71.7^{+}0.96$
3D SPMK Discriminative Feature ($L = 2$)	$70.1^{+}1.4$



(a) (b)

Figure 17: Confusion matrix for the 8 categories in Bar-Hillel *et al.* Dataset. Average classification rates for individual categories are listed along the main diagonal. Results for the 3D SPMK using the full volume of the pyramid with (a) $L = 0$ and (b) $L = 2$, with a vocabulary size of $K = 800$.

matrices for the 8 categories. Figure 17 shows the results achieved by the 3DSPMK approach using the full volume of the pyramid and a codebook size of $K = 800$. We show the results for pyramid levels $L = 0$ and $L = 2$, in Figures 17(a) and 17(b), respectively. It can be observed that the worst performance is obtained for the category *bottle*. We believe this is due to the quality of the depth information recovered by the sensor for transparent objects, which are really problematic. Although several preprocessing operations have been applied, the point cloud obtained does not characterize the object shape. Figure 15 actually shows the large portions of data that are actually missing for *bottle* class. Finally, except for *teddy bear* class, whose confusion with *doll* increases, the experiments confirm that the higher the pyramid level, the lower the confusions.

4.2. Scene Categorization

We also want to evaluate the performance of the detailed 3DSPMK for the particular problem of scene recognition using RGB-D images.

Here, we propose the following experiment to go further in our evaluation, and explore what the performance of the 3DSPMK is recognizing scenes in depth images. We use the New York University Depth (NYU Depth) video dataset [1]. Instead of 3D SURF descriptors, we incorporate to our approach a novel 3D feature, the NARF local descriptors [17]. Furthermore,we explore a "pre-classification" fusion strategy which consists in concatenating the image representations (for RGB and depth) prior to the application of any classifier, in order to increase the classification accuracy.

4.2.1. NYU Depth Database

Experimental Setup. The NYU Depth dataset [1] is a new and very challenging indoor video scene dataset, which is comprised of video sequences from a variety of

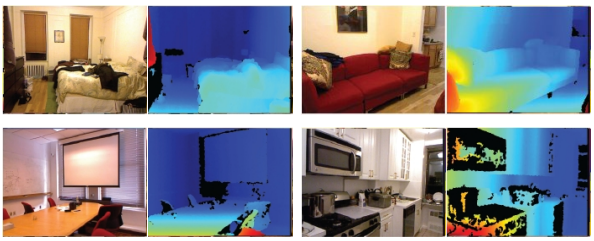


Figure 18: Samples of the RGB images and the raw depth images of the NYU Depth dataset.

Table 5: Statistics of captured sequences from NYU Depth Database.

Scenes Class	Scenes	Frames
Bathroom	6	5588
Bedroom	17	22764
Kitchen	10	12643
Living Room	13	19262
Office	14	19262

indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. For the scene classification benchmark, the dataset offers 20,000 images (10,000 for training and testing) distributed across 5 different scene-level classes: *bathroom*, *bedroom*, *kitchen*, *living room* and *office*. Figure 18 shows some samples of both the RGB images and the raw depth images provided. The 5 scene-level classes in the NYU Depth dataset are summarized in Table 5.

We follow the same experimental setup detailed in [1]. We use the splits for training and testing provided by the authors. In order to evaluate the performance of our methods, we use the mean confusion matrix diagonal, as in [1].

For RGB image representation, we use SIFT [11] descriptors of 16×16 pixel patches, computed over a grid with spacing of 8 pixels. For depth image representation, we now use NARF features [17]. We perform a dense extraction of NARF descriptors in the point clouds. We use the publicly available implementation of NARF descriptors[3].

We build visual vocabularies using K-means clustering on the SIFT and NARF descriptors independently (we fix the vocabulary size to $K = 200$). Finally, for the RGB images, we apply the SPMK scheme introduced in [4], and for the depth images, we use our own 3DSPMK. Again, SVMs are used for classification.

Results. We start comparing our 3DSPMK approach with the state-of-the-art results reported in the NYU Depth dataset. Table 6 shows that our integration of

[3] `http://pointclouds.org/documentation/tutorials/narf_feature_extraction.php`

Table 6: Mean Diagonal of Confusion Matrix on the NYU Depth.

Mean Diagonal of Confusion Matrix	
Method	Depth
[1] 2D SPMK ($K = 200$)	49
[1] 2D SPMK ($K = 800$)	48
3D SPMK ($L = 0$)	45
3D SPMK ($L = 1$)	48
3D SPMK Representativeness ($L = 1$)	48
3D SPMK Discriminative Feature ($L = 1$)	48
3D SPMK ($L = 2$)	50
3D SPMK Representativeness ($L = 2$)	**51**
3D SPMK Discriminative Feature ($L = 2$)	48

NARF features in the 3DSPMK pipeline outperforms the state-of-the-art results, when *only* depth images are considered. Furthermore, we can observe again that, as the pyramid level increases, from 0 to 2, the performance increases too. Our best result is obtained by the 3DSPMK with the representativeness-based approach (with $p = 0.1$). We also have experimentally observed that the performance of the discriminative features-based approach, with $\tau < 0.6$ and $\beta < 50\%$, plateaus at 48% – a behavior which coincides with the previous analysis reported in Figures 10 and 16.

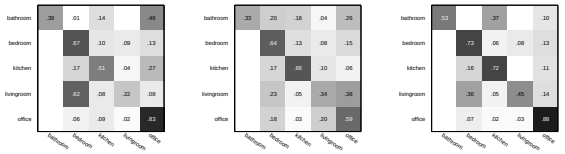
After the thorough experimental validation with the three different datasets proposed, we can conclude that, first, in general, the representativeness-based algorithm outperforms the standard 3DSPMK. The reason is that the former incorporates a smaller number of noisy sub-volumes to the pyramid structure. Second, with respect to the comparison with the discriminative-based technique, the performance of the representativeness-based algorithm seems to be always better. The reason is that the discriminative-based technique is too compact. That is, the sub-volume identified discards cubes that are selected by the representativeness-based method, and that help to better recognize the classes. However, the computational efficiency is higher for the discriminative-based approach than for the representativeness-based method. Overall, the discriminative-based approach manages the largest reduction on the dimensionality of the histograms, while its performance decreases lower than 3% in all databases, with respect to the performance reported by the representativeness-based algorithm.

Combining RGB and Depth. In order to study the relative contribution and optimal blend of depth and intensity information, we propose the following experiment. RGB and depth information can be fused into our representation using a "pre-classification" fusion approach. Basically, we propose to concatenate pyramid representations of both RGB (SPMK) and depth (3DSPMK) to form a single feature representation prior to being introduced into the SVM classifier.

For doing so, we take the following steps. First, in order to characterize the RGB images, we follow the SPMK approach in [4]. With a vocabulary of size $K = 200$, computed using SIFT descriptors, we represent each RGB image with an spatial pyramid as in [4]. Second, each depth image is represented following our 3DSPMK approach. Finally, each pair of RGB+depth images is represented by concatenating both histogram based representations.

Table 7: Mean Diagonal of Confusion Matrix on the NYU Depth. Combining RGB and Depth information.

Mean Diagonal of Confusion Matrix			
Methods	Depth	RGB	All
2D SPMK ($L = 2$) + 3D SPMK Representativeness ($L = 2$)	51	52.4	**66**
[1] 2D SPMK ($K = 200$)	49	55	55
[1] 2D SPMK ($K = 800$)	48	56	60



(a) (b) (c)

Figure 19: Confusion matrices for the 5 categories in the NYU Depth database. Average classification rates for individual categories are listed along the main diagonal. Results for the (a) SPMK with $K = 200$ and $L = 2$, (b) 3D SPMK Representativeness with $K = 200$ and $L = 2$, and (c) the combination of both.

Table 7 shows the results obtained by the combination of the two best approaches, according to our experimental evaluation: 1) the SPMK with $L = 2$ for RGB images; and 2) the 3DSPMK with $L = 2$ and the representativeness based approach for the depth images. Notice that our approach outperforms the state-of-the-art only using a vocabulary of size $K = 200$. We also show the corresponding confusion matrices for each of these approaches and their combination in Figure 19.

5. Conclusion

We have presented a novel approach for object and scene categorization using RGB-D images. We have introduced a 3D BoW based model, which uses the quantized local 3D descriptors extracted from point clouds. Our method incorporates the novel 3DSPMK, and two selective sub-volume decomposition strategies for jointly increasing the classification performance and the computational efficiency of the proposed approach.

We conclude that the 3DSPMK is a simple yet efficient strategy for categorization in point clouds, which can be easily integrated with the RGB information. In our experiments, we have evaluated our approach on three RGB-D datasets, and the results show that the

proposed kernels perform well compared to the state-of-the-art. All experiments are based on our publicly available source code [4].

Acknowledgment

This work was partially supported by projects CCG2013/EXP-047, IPT-2012-0808-370000, TIN2010-20845-C03-03, UAH2011/EXP-030 and IPT-2011-1366-390000.

References

[1] N. Silberman, R. Fergus, Indoor scene segmentation using a structured light sensor, in: Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition, 2011.

[2] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: ECCV International Workshop on Statistical Learning in Computer Vision, 2004.

[3] J. Sivic, A. Zisserman, Video google: A text retrieval approach to object matching in videos, in: ICCV, 2003.

[4] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories., in: CVPR, Vol. 2, 2006, pp. 2169–2178.

[5] Y. Su, F. Jurie, Visual word disambiguation by semantic contexts, in: ICCV, 2011.

[6] K. van de Sande, T. Gevers, C. Snoek, Evaluating color descriptors for object and scene recognition, PAMI 32 (2010) 1582–1596.

[7] J. Zhang, M. Marszalek, S. Lazebnik, C. Schimd, Local features and kernels for classification of texture and object categories: A comprehensive study, IJCV 73 (2) (2007) 213–238.

[8] S. A. Chatzichristofis, C. Iakovidou, Y. Boutalis, O. Marques, Co.vi.wo.: Color visual words based on non-predefined size codebooks, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on PP (99) (2012) 1 –14. doi:10.1109/TSMCB.2012.2203300.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes (VOC) challenge, IJCV 88 (2) (2010) 303–338.

[10] K. Grauman, T. Darrell, The pyramid match kernel:discriminative classification with sets of image features, in: ICCV, 2005, pp. 1458–1465.

[11] D. Lowe, Object recognition from local scale-invariant features, in: ICCV, 1999.

[12] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, SURF: Speeded up robust features, Computer Vision and Image Understanding 110 (3) (2008) 346–359.

[13] A. Bosch, A. Zisserman, X. Muñoz, Representing shape with a spatial pyramid kernel, in: Proceedings of the 6th ACM international conference on Image and Video retrieval, 2007, pp. 401–408.

[14] A. Bar-Hillel, D. Hanukaev, D. Levi, Fusing visual and range imaging for object class recognition, in: ICCV, 2011.

[15] K. Lai, L. Bo, X. Ren, D. Fox, A large-scale hierarchical multi-view RGB-D object dataset, in: ICRA, 2011.

[16] J. Knopp, M. Prasad, G. Willems, R. Timofte, L. Van Gool, Hough transform and 3D SURF for robust three dimensional classification, in: ECCV, 2010.

[17] B. Steder, R. B. Rusu, K. Konolige, W. Burgard, NARF: 3D range image features for object recognition, in: Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 2010.

[18] C. Redondo-Cabrera, R. J. Lopez-Sastre, J. Acevedo-Rodriguez, S. Maldonado-Bascon, SURFing the point clouds: Selective 3D spatial pyramids for category-level object recognition, in: IEEE CVPR, 2012.

[19] D. Saupe, D. V. Vranic, 3D model retrieval with spherical harmonics and moments., in: DAGM-Symposium on Pattern Recognition, 2001.

[20] R. Osada, T. Funkhouser, B. Chazelle, D. Dobki, Shape distributions, ACM Transactions on Graphics 21 (4).

[21] A. Frome, D. Huber, R. Kolluri, T. Blow, J. Malik, Recognizing objects in range data using regional point descriptors, in: ECCV (3)'04, 2004, pp. 224–237.

[22] A. Mian, M. Bennamoun, R. Owens, Three-dimensional model-based object recognition and segmentation in cluttered scenes., PAMI 28.

[23] J. Novatnack, K. Nishino, Scale-dependent/invariant local 3D shape descriptors for fully automatic registration of multiple sets of range images, in: ECCV, 2008.

[24] R. Toldo, U. Castellani, A. Fusiello, A bag of words approach for 3D object categorization, in: MIRAGE '09 Proceedings of the 4th International Conference on Computer Vision/Computer Graphics Collaboration Techniques, 2009.

[25] K. Lai, L. Bo, X. Ren, D. Fox, Sparse distance learning for object recognition combining RGB and depth information, in: ICRA, 2011.

[26] A. Johnson, M. Hebert, Using spin images for efficient object recognition in cluttered 3D scenes, PAMI 21.

[27] P. Gupta, S. S. Arrabolu, M. Brown, S. Savarese, Video scene categorization by 3d hierarchical histogram matching, in: ICCV, 2009.

[28] L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: CVPR, Vol. 2, 2005, pp. 524–531.

[29] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, IJCV 42 (3) (2001) 145–175.

[30] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba., Sun database: Large-scale scene recognition from abbey to zoo., in: CVPR, 2010.

[31] J. Xiao, K. Ehinger, A. Oliva, A. Torralba, Recognizing scene viewpoint using panoramic place representation, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2012, pp. 2695 –2702. doi:10.1109/CVPR.2012.6247991.

[32] Z. C. Marton, R. B. Rusu, M. Beetz, On fast surface reconstruction methods for large and noisy datasets, in: ICRA, 2009.

[33] C.-W. Hsu, C.-J. Lin, A comparison of methods for multi-class support vector machines, IEEE Transactions on Neural Networks 13 (2002) 415–425.

[34] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm` (2001).

[35] M. A. Fischler, R. C. Bolles, Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography, Comm. ACM 24 (6) (1981) 381–395.

[4] http://agamenon.tsc.uah.es/Personales/rlopez/data/3dspmk/