SURFing the Point Clouds: Selective 3D Spatial Pyramids for Category-level Object Recognition

Carolina Redondo-Cabrera, Roberto J. López-Sastre,
Javier Acevedo-Rodríguez and Saturnino Maldonado-Bascón
GRAM, Dept. of Signal Theory and Communications, University of Alcalá, Alcalá de Henares, Spain
crc04057@alu.uah.es, robertoj.lopez@uah.es

Abstract

This paper proposes a novel approach to recognize object categories in point clouds. By quantizing 3D SURF local descriptors, computed on partial 3D shapes extracted from the point clouds, a vocabulary of 3D visual words is generated. Using this codebook, we build a Bag-of-Words representation in 3D, which is used in conjunction with a SVM classification machinery. We also introduce the 3D Spatial Pyramid Matching Kernel, which works by partitioning a working volume into fine sub-volumes, and computing a hierarchical weighted sum of histogram intersections at each level of the pyramid structure. With the aim of increasing both the classification accuracy and the computational efficiency of the kernel, we propose selective hierarchical volume decomposition strategies, based on representative and discriminative (sub-)volume selection processes, which drastically reduce the pyramid to consider. Results on the challenging large-scale RGB-D object dataset show that our kernels significantly outperform the state-of-the-art results by using a single 3D shape feature type extracted from individual depth images.

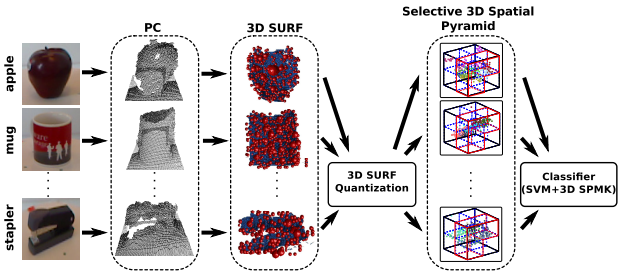


Figure 1. Proposed approach using Selective 3D Spatial Pyramid Matching Kernels for object recognition in point clouds. We quantize 3D SURF descriptors, extracted from partial 3D shapes computed from single depth images, into 3D visual words. This codebook is used to represent the objects in a BoW approach. The 3D SPMK repeatedly subdivides a cube inscribed in the 3D shape, 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.

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. Whatever the image depicts, we have the ability to interpret it. Furthermore, we do this with an astonishing ease.

Image categorization, including category-level object recognition and scene classification, remains to be a major challenge for the computer vision community. Given an image, the objective is to predict the presence/absence of at least one object of a particular class. In the last few years, this problem has been receiving a lot of attention. A popular strategy for representing images, within the context of category-level object recognition, is the Bag-of-Words (BoW) approach [5, 23]. The brilliant idea behind this type of representation is to characterize 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 that have obtained impressive results recently (*e.g.* [16, 24, 26, 27]), being the basic recipe for most of the methods submitted to the PASCAL VOC Challenge [6].

Various generative approaches have been proposed (*e.g.* [7]). However, nonlinear SVMs methods using Spatial Pyramid Matching Kernels (SPMKs) [9, 16] have been systematically obtaining the best results. The most recent improvements have been achieved by incorporating multiple local features such as SIFT [17], SURF [2] or color SIFT [26], into the BoW pipeline [3, 26].

So, we can say that the categorization problem in 2D images is a well established field of research. But, 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 object recognition and detection approaches [1, 14].

In this paper, we build a discriminative approach for recognizing object categories in point clouds, which *just* uses the information extracted from depth images. Inspired by the works of Lazebnik *et al*. [16] and Knopp *et al*. [13], we introduce a novel framework for object recognition, which uses 3D local shape features. The new methodology is depicted in Figure 1. We start extracting 3D SURF local descriptors [13] on a partial 3D shape obtained from a point cloud that a depth camera provides. Note that we do use a single depth image as input. These 3D SURF descriptors are then quantized, *e.g*. using K-means, so as to obtain the 3D visual words. We introduce a kernel-based category-level object recognition approach, which works adapting the SPMK [16] to work in 3D, *i.e*. the 3D SPMK. This novel mechanism involves repeatedly subdividing a cube inscribed in the 3D shape, 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 drastically reduce the volume to consider (see the red subvolumes selected in the cubes of Figure 1), while jointly increase the classification accuracy and the computational efficiency of the kernel. Results on the challenging RGB-D object dataset [14] show that our kernels significantly outperform the state-of-the-art results by just using a single 3D shape feature type extracted from individual depth images.

The rest of this paper is organized as follows. Section 2 describes related work. The 3D SPMK is detailed in Section 3. The experimental setup and results are presented in Sections 4 and 5 respectively. We conclude in Section 6.

2. Related Work

As there exists a large body of work on category-level object recognition (*e.g*. [5, 16, 24, 26, 27]), we briefly review in the following only the most relevant to this paper, *i.e*. on image categorization using point clouds, 3D shapes and/or depth images.

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 the shape moments [22] or the shape histograms [21], for example. Neither partial shapes, nor intra-class variations are successfully handled by global descriptions. Moreover, using depth cameras, we do not obtain perfect scans of the environment, and we capture all the neighboring clutter in addition to the relevant data coming from the object of interest. Therefore, 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 categorization methods using local features. For instance, Mian *et al*. introduce the use of local tensors [19], while scale-dependent and scale-invariant local 3D shape descriptors are proposed in [20]. Toldo *et al*. [25] 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*. [13] 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 3D SPMK for object categorization. Moreover, we extend their method to work with *partial* 3D shapes obtained from depth images and their corresponding point clouds.

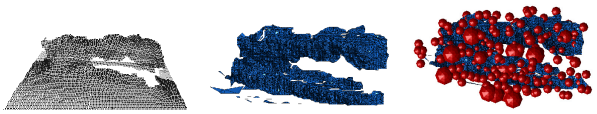
Recently, some approaches combine RGB and depth images so as to increase the performance in visual categorization (*e.g*. [1, 14, 15]). For instance, Lai *et al*. [14] benchmark the object categorization problem using a combination of RGB (SIFT [17]) and depth features (spin images [12]) in the very challenging large-scale hierarchical multi-view RGB-D object dataset, consisting of segmented RGB and depth images of 300 everyday objects distributed across 51 different categories. In [15], a sparse distance learning approach is designed for combining RGB and depth information. However, in this paper, we show how our selective 3D SPMK significantly outperforms the state-of-the-art results reported in the RGB-D object dataset, by just using a single feature type computed on partial 3D shapes obtained from depth images.

3. Categorizing Point Clouds

Our goal is to learn models for object categorization in point clouds. In this section, we detail our proposed representation for object categories, and introduce the 3D SPMK, which can be used in a SVM machinery for object recognition.

3.1. Category Representation

Our approach is shown in Figure 2. We start capturing a point cloud that contains the object of interest (Figure 2(a)), from a single depth image, for example captured with the Kinect. We then perform a point cloud triangulation so as to obtain a partial 3D shape. For this step, we use the greedy surface triangulation method proposed in [18]. For more



(a) (b) (c)

Figure 2. 3D SURF extraction pipeline from a depth image. (a) the input image is a point cloud obtained by the Kinect depth sensor. (b) we process the point cloud to obtain a partial 3D shape. (c) 3D SURF features are extracted and back-projected to the 3D shape.

details, see Section 4.2. Figure 2(b) shows the result of this triangulation step. Finally, 3D SURF features [13] are extracted on the partial 3D shape (Figure 2(c)). In contrast to a global representation, by using, for example, a dense or random coverage with spin images [12], the 3D SURF is equipped with an interest point detector, where the descriptors are computed. By following a traditional BoW approach, we quantize 3D SURF descriptors, into 3D visual words. Each image can be then characterized by a histogram of its 3D visual words.

3.2. 3D Spatial Pyramid Matching Kernel

Nonlinear SVMs methods using SPMKs [9, 16] have been offering the best performances in object categorization systems. The original formulation of pyramid matching was introduced in [9]. 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 feature sets. Grauman and Darrell [9] demonstrated that the pyramid match kernel satisfies the Mercer's condition, *i.e.* it guarantees an optimal solution to kernel-based algorithms based on convex optimization, such as SVMs.

Based on [9], Lazebnik *et al.* [16] 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 [16], we propose to extend the SPMK to the three-dimensional space, *i.e.* the 3D SPMK. As it was described in Section 3.1, we model a partial 3D shape 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 3D SPMK should be able to capture the spatial distribution of such codewords at different scales and locations in a working cube $\Omega^{(0)}$. Similar to [16], but in 3D, we define a pyramid structure by partitioning $\Omega^{(0)}$ 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 3). It 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.

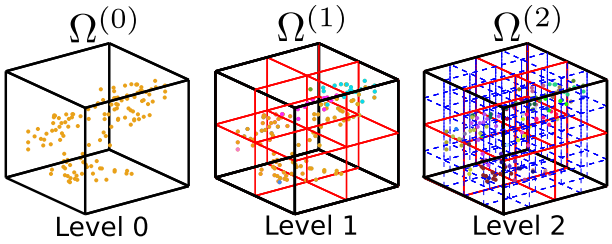


Figure 3. Example of a 3D spatial pyramid of three levels. The working volume $\Omega^{(0)}$ is recursively decomposed into eight sub-cubes.

Once the pyramid decomposition of L levels $P(L)$ is processed, we perform the pyramid matching in 3D. Let us define H_X^l and H_Y^l as the histograms for features X and Y in the level l of the pyramid. We also define $H_X^l(i)$ and $H_Y^l(i)$ as the histograms of features X and Y that fall into the ith sub-cube in the pyramid $P(L)$ for the level l, *i.e.* $\Omega_i^{(l)}$. Only features of the same type can be matched. So the number of matches at level l is given by the histogram intersection function as follows

$$\mathcal{I}(H_X^l, H_Y^l) = \sum_{i=1}^{D} \min(H_X^l(i), H_Y^l(i)) \,. \qquad (1)$$

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

$$K(X, Y) = \omega_0 \mathcal{I}(H_X^0, H_Y^0) + \sum_{l=1}^{L} \omega_l \mathcal{I}(H_X^l, H_Y^l) \,, \qquad (2)$$

where, w_l is set to $\frac{1}{2^{L-l}}$. By doing so, we penalize those matches found in larger volumes, because they involve increasingly dissimilar features.

3.3. Selective 3D SPMK

So far, our formulation can be seen as an extension of the original SPMK [16] to 3D. One clear disadvantage of the pyramid decomposition proposed is its high computational cost. For a pyramid of L levels and K features, we obtain a vector of dimensionality $K \sum_{l=0}^{L} 8^l$, *i.e.* 2^l times more bins in each level with respect to the SPMK [16]. With the aim reducing this dimensionality, but also increasing both the classification accuracy and the computational efficiency of the 3D SPMK, we introduce two *selective* volume decomposition schemes based on representative and discriminative (sub-)volume selection processes. Note that our approach significantly differs from [10], where neither discriminative feature-based, nor representativeness-based decomposition mechanisms are considered. Furthermore, in [10] the appearance information is transfered to the 3D

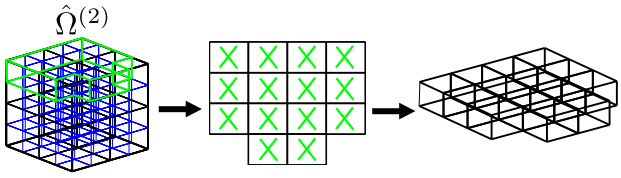


Figure 4. Toy example of the representativeness-based selective 3D SPMK. The process selects the green sub-volume.

points from 2D images where invariant descriptors are computed. However, we recover the invariant information directly from the sparse point clouds thanks to the 3D SURF descriptors.

3.3.1 Representativeness-based Selection

Unlike in the 2D case [16], where we can consider a uniform distribution of local features across the whole 2D pyramid (specially with a dense feature extraction), in our 3D formulation, the local features occupy sparse locations in the 3D working volume. Furthermore, the higher the level of the pyramid, the lower the size of each sub-cube (*e.g.* for $L = 2$, $\Omega_i^{(2)} = \Omega^{(0)}/64$), and the higher the number of empty sub-volumes.

Thus, 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 follow a selective approach that will incorporate into the pyramid, only those (sub-)cubes that are likely to represent images in our dataset. That is, our objective is to reduce the large number of 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. 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 (*e.g.*, in the experiments $N = 50$), for computing the representativeness-based selection. A toy example of this process is shown in Figure 4.

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 $\hat{H}_X^l(i)$ of the features that fall into the ith sub-cube $\hat{\Omega}_i^{(l)}$ at level l. These histograms will be used in Eq. (2).

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 when a particular feature is discriminative enough for a particular class. Assume we are given a set of images, and each image belongs to a class i, being N the total number of classes. As it has been described, we build a visual codebook of size K from 3D SURF 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 SURF descriptors, X_j, we assign each one to a feature $f_k \in \mathcal{F}$. For each class i, we define M_i, $i = 1, \ldots, N$, as the total number of descriptors extracted from the images of the class i. We also define $m_i^{(f_k)}$, as the number of descriptors for the class i that has been assigned to the 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 $s_{ik} = S(i,k)$ is computed as follows

$$s_{ik} = \Delta_k \frac{m_i^{(f_k)}}{M_i}\,, \tag{3}$$

where,

$$\Delta_k = \left(\sum_{i=1}^{N} \frac{m_i^{(f_k)}}{M_i} \right)^{-1}\,. \tag{4}$$

Each score s_{ik} can be seen as the ratio between the percentage of descriptors that belong to the feature k in the class i, 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

$$s'_{ik} = \begin{cases} 1 & \text{if } s_{ik} \geq \tau \\ 0 & \text{if } s_{ik} < \tau \end{cases} \tag{5}$$

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

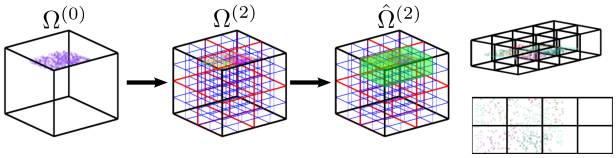


Figure 5. Toy example of the discriminative feature-based volume decomposition for the 3D SPMK. The discriminative features fall in the green sub-volume selected.

This time, we consider all the training images of all the classes to compute S', so we know which are the discriminative features. Given a pyramid of L levels $P(L)$, we inspect all its sub-volumes. For each sub-volume and each object class, we measure the proportion of images that contain at least one *discriminative* feature, 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. A toy example of this process is shown in Figure 5, where discriminative features fall in the green sub-volumes.

Note that we can run this procedure on top of either the original pyramid decomposition, or the pyramid decomposition selected by the representativeness-based criterion. Furthermore, both selective mechanisms can be run in parallel, and then define as the final decomposition, the intersection of the two solutions. In our experiments we have found that normally the representativeness-based is more restrictive than the other.

4. Experimental Setup

In this section, we briefly introduce the RGB-D Object dataset, and then we describe in detail the feature extraction process followed to compute the 3D SURF descriptors from point clouds.

4.1. RGB-D Object Dataset

In order to test our approach for the problem of object recognition in point clouds, we have used the challenging RGB-D Object dataset [14]. It is a large scale set 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 $\times$ 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 6 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.

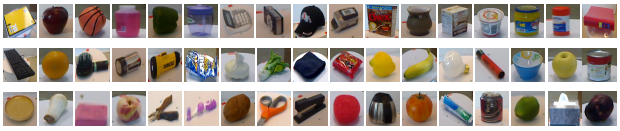


Figure 6. Object instances from RGB-D Object Dataset [14]. One example for each of the 51 object categories is shown.

We evaluate our object categorization approach on this dataset, following the same experimental setup described in [14]. For the experiments, we use all the 51 categories. We subsample the turntable data by taking every fifth video frame. For image categorization, we randomly leave one object out from each category for testing, and train the classifier using the 3D SPMK on all the views of the remaining objects. The final result is reported as the average per-class recognition rate. We also present confusion matrices for the 51 categories used.

4.2. Feature Extraction

3D SURF features [13] have been computed using the RGB-D Object dataset. For doing so, we start reading the point clouds provided in the dataset. We consider two different approaches for the feature extraction: with and without automatic object segmentation in the point cloud. We report results using both pipelines. When the object has to be automatically segmented, we use the known distance between the turntable and the camera, 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. Objects are placed on a turntable, so we can clean the turntable points by running a RANSAC [8] fit plane algorithm on the point cloud. Following this automatic procedure, we obtain clean point clouds for all the object classes in the dataset.

The 3D SURF descriptors have to be computed from a 3D shape. Therefore, the next step consist in obtaining the partial 3D shape defined by the point cloud. We perform a point cloud triangulation to each depth image. For doing so, we follow the greedy surface triangulation method proposed in [18][1]. 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. Above, Figure 2(b) shows an example of this automatic object segmentation process, and how the triangulation algorithm works.

Each partial 3D shape is uniformly scaled to fit a cube

[1]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°.

with a side of length 256. Then 3D SURF descriptors of 162 dimensions are computed using the original implementation provided in [13]. With the aim of covering the full 3D shape with 3D SURF descriptors, we have experimentally chosen the following parameters: the distance between triangle mesh and the border of the cube is 30, and the threshold is fixed to 10^{-8}. A result of this 3D SURF extraction step is shown in Figure 2(c).

5. Experiments

We have conducted two types of experiments. First, in Section 5.1, we integrate the automatic object segmentation algorithm in the feature extraction pipeline, and report classification results following this approach. Next, in Section 5.2, we conduct additional experiments where no object segmentation is performed, showing that our approach is able to deal with point clouds with the object of interest and the clutter coming from the rest of the scene.

For all the experiments, we use a visual vocabulary of size $K = 200$. The visual vocabulary is obtained performing K-means clustering on a subset of the local descriptors (taking the 3D SURF descriptors of 50 images per class). We represent each object 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. To process the 3D spatial pyramids for $L = 2$, we directly follow the selective algorithms described in Section 3.3, *i.e*. we do not report the performance of the 3D SPMK using the full volume because the amount of memory needed for representing all the images considerably exceeds our resources.

We use SVMs for classification. As kernel function, we use our 3D SPMK detailed in equation (2). The multi-class classification problem is solved training the SVM using the one-against-one strategy. We follow the approach in [11], 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 [11]: if one of the classifiers votes for the class 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. Specifically, we use libSVM [4] for training each binary classifier. A 10-fold cross-validation on the train set to tune SVM parameters is conducted.

5.1. Object Recognition with Object Segmentation

Table 1 shows the results obtained by our approaches, as well as a comparison with the state-of-the-art methods [14, 15].

First, let us analyze the performance of our 3D shape features, *i.e*. quantized 3D SURF descriptors. For a pyramid of level 0, our average classification rate for the 51 classes is 72.3%. If we compare with state-of-the-art results, when *only* shape features are used, we can see that our approach outperforms the best results reported in [15] (70.2%), where spin images are the shape features used. These results reveal the convenience of using the proposed codebooks of quantized 3D SURF local descriptors for the problem of object recognition in point clouds.

Table 1. Classification Accuracy of different approaches on the RGB-D Object dataset. kSVM, RF (Random Forest), IDL (Instance Distance Learning).

Classification Accuracy			
Method	Shape	Texture	All
[14] (kSVM)	64.7	74.5	83.8
[14] (RF)	68.8	74.7	79.6
[15] (IDL)	70.2	78.6	85.4
3D SPMK ($L = 0$)	72.3	n/a	n/a
3D SPMK ($L = 1$)	**93.1**	n/a	n/a
3D SPMK Representativeness ($L = 2$)	**94.8**	n/a	n/a
3D SPMK Discriminative Feature ($L = 2$)	**94.6**	n/a	n/a

Next, let us examine the performance of the 3D SPMK. Table 1 shows that the results improve dramatically as the pyramid level goes from $L = 0$ to $L = 2$. Moreover, it is important to note that our approach, which just uses a single feature type (extracted from a single depth image), significantly outperforms the state-of-the-art results [14, 15], which combine multiple features (SIFT and spin images). We consider this a remarkable point, which confirms, as it already did before in the 2D case [16], the capacity of strengthening image categorization strategies, via combining 3D spatial pyramid schemes into BoW approaches.

The best results have been obtained by pyramids with $L = 2$. We report results using the representativeness-based and the feature discriminative-based approaches, and it seems that both of them obtain similar results. For the former, with $p = 0.1$, the selective pyramid $\hat{\Omega}^{(2)}$ contains only 14 sub-cubes of the 64. For the latter, with τ and β fixed to 0.7 and 50% respectively, $\hat{\Omega}^{(2)}$ includes only 19 sub-volumes. These results reveal that, by following our selective decomposition strategies, the classification rate and the computational efficiency jointly increase. Furthermore, we did not fine tune the parameters β, τ and p, we experimentally observed that they are not critical for classification performance. We also found that the representativeness-based method is always more restrictive, *i.e*. its selected volume is included within the volume selected by the discriminative feature-based approach. So, the results of the representativeness-based would be equivalent to the results obtained by the intersection of the two selected volumes.

The RGB-D Object dataset is a large scale dataset, so it is also relevant to analyze how our methods perform for each particular class. Figure 7 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 84% of the classes. When

$L = 0$ only 13% of the classes attain an average accuracy higher than 90%. Figure 7 also shows that the higher the pyramid level, the higher the minimum classification rate which increases from 17% (class mushroom, $L = 0$), to 60% (class tomato, $L = 2$). Another conclusion is that, our approach is a shape-based approach, 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 *tomato*, *pear* and *ball*. In Figure 8, we also show confusion matrices for the 51 categories. In general, the higher the pyramid level, the lower the confusions.

5.2. Recognizing without 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.

For these experiments, we again follow the same experimental setup described, but we use the whole depth image, and not just the object segmented. Figure 9 compares the performances of 3D SPMK with and without automatic object segmentation. The first conclusion we draw is that for pyramids with $L = 0$, the classification rate dramatically increases (from 72.3% to 86.1%) when no automatic object segmentation is done. We think this increment is related with the impreciseness of the 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 $L = 1, 2$, the results of both approaches are comparable. The best results are obtained for $L = 1$ without segmentation (95.9%). For pyramids with $L = 2$ the approach with object segmentation performs slightly better. It is also interesting to note that the discriminative feature-based selection works better than the representativeness-based, which is not able to discard those features from the background that are not discriminative for an object class. Furthermore, the discriminative feature-based approach obtains very similar results with and without segmentation. We consider that this is related with the fact that the discriminative feature-based approach tends to select those sub-volumes that contain only discriminative features, discarding those sub-volumes with features that are common to all categories (like the background features), and we can consider this as an *object segmentation* approach but from the feature space.

6. Conclusion

We have presented a novel approach for recognizing object categories in point clouds. A BoW approach is proposed, using quantized 3D SURF local descriptors, which are computed on partial 3D shapes extracted from depth images. We have also introduced the 3D SPMK and two selective volume decomposition algorithms for increasing the classification performance while drastically reducing both the memory cost and runtime. Experiments on the RGB-D object dataset show that our kernels significantly outperform the state-of-the-art results by using a single 3D shape feature type.

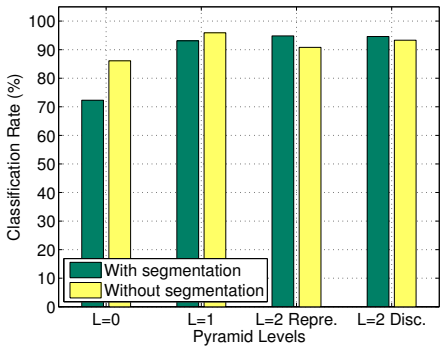


Figure 9. Category recognition performances of the 3D SPMK with and without automatic object segmentation.

We plan to integrate our methods in a multiple kernel learning approach, where we can combine the shape features with appearance features extracted from the RGB images too.

Acknowledgements

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

References

[1] A. Bar-Hillel, D. Hanukaev, and D. Levi. Fusing visual and range imaging for object class recognition. In *ICCV*, 2011.

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

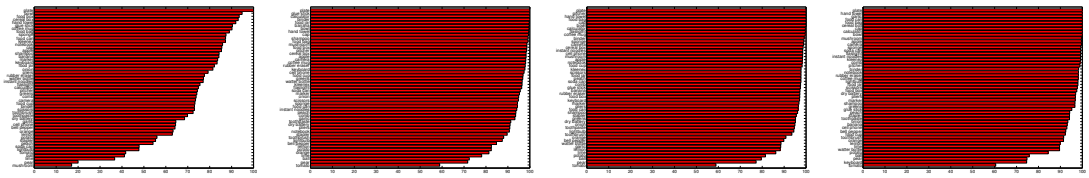
[3] A. Bosch, A. Zisserman, and X. Muñoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and Video retrieval*, pages 401–408, 2007.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

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

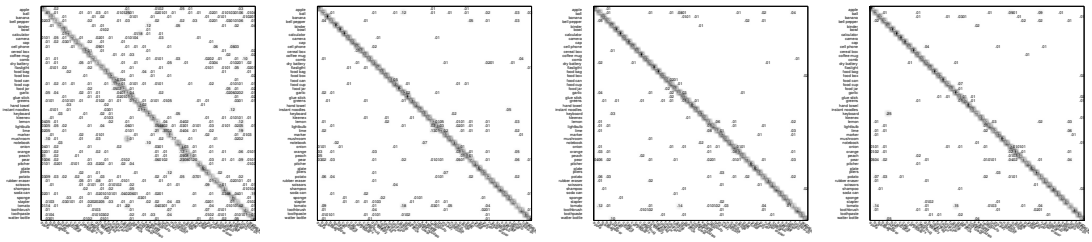
[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[7] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental



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

Figure 7. Classification accuracy for each category. We use a vocabulary size of $K = 200$, and the 3D SPMK with (a) $L = 0$, (b) $L = 1$, 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.



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

Figure 8. Confusion matrices for the 51 categories in RGB-D Object database. Average classification rates for individual categories are listed along the main diagonal. Results for the 3D SPMK with (a) $L = 0$, (b) $L = 1$, 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.

bayesian approach tested on 101 object categories. In *CVPR*, 2004.

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

[9] K. Grauman and T. Darrell. The pyramid match kernel:discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, 2005.

[10] P. Gupta, S. S. Arrabolu, M. Brown, and S. Savarese. Video scene categorization by 3d hierarchical histogram matching. In *ICCV*, 2009.

[11] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.

[12] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *PAMI*, 21, 1999.

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

[14] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *ICRA*, 2011.

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

[16] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178, 2006.

[17] D. Lowe. Object recognition from local scale-invariant features. In *ICCV*, September 1999.

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

[19] A. Mian, M. Bennamoun, and R. Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *PAMI*, 28, 2006.

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

[21] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobki. Shape distributions. *ACM Transactions on Graphics*, 21(4), 2002.

[22] D. Saupe and D. V. Vranic. 3D model retrieval with spherical harmonics and moments. In *DAGM-Symposium on Pattern Recognition*, 2001.

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

[24] Y. Su and F. Jurie. Visual word disambiguation by semantic contexts. In *ICCV*, 2011.

[25] R. Toldo, U. Castellani, and 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.

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

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