

2D/3D Object Recognition and Categorization Approaches for Robotic Grasping

Nabila Zrira, Mohamed Hannat, El Houssine Bouyakhf, Haris Ahmad Khan

▶ **To cite this version:**

Nabila Zrira, Mohamed Hannat, El Houssine Bouyakhf, Haris Ahmad Khan. 2D/3D Object Recognition and Categorization Approaches for Robotic Grasping. Advances in Soft Computing and Machine Learning in Image Processing, 730, pp.567-593, 2017, 978-3-319-63754-9. hal-01676387

HAL Id: hal-01676387
https://hal.science/hal-01676387

Submitted on 5 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

2D/3D Object Recognition and Categorization Approaches for Robotic Grasping

Nabila Zrira, Mohamed Hannat, El Houssine Bouyakhf and Haris Ahmad Khan

Abstract Object categorization and manipulation are critical tasks for a robot to operate in the household environment. In this paper, we propose new methods for visual recognition and categorization. We describe 2D object database and 3D point clouds with 2D/3D local descriptors which we quantify with the k-means clustering algorithm for obtaining the Bag of Words (BOW). Moreover, we develop a new global descriptor called VFH-Color that combines the original version of Viewpoint Feature Histogram (VFH) descriptor with the color quantization histogram, thus adding the appearance information that improves the recognition rate. The acquired 2D and 3D features are used for training Deep Belief Network (DBN) classifier. Results from our experiments for object recognition and categorization show an average of recognition rate between 91% and 99% which makes it very suitable for robot-assisted tasks.

Nabila Zrira
LIMIARF Laboratory, Mohammed V University Rabat, Faculty of Sciences Rabat, Morocco, e-mail: nabilazrira@gmail.com

Mohamed Hannat
LIMIARF Laboratory, Mohammed V University Rabat, Faculty of Sciences Rabat, Morocco, e-mail: mohamedhannat@gmail.com

El Houssine Bouyakhf
LIMIARF Laboratory, Mohammed V University Rabat, Faculty of Sciences Rabat, Morocco, e-mail: bouyakhf@mtds.com

Haris Ahmad Khan
NTNU, Norwegian University of Science and Technology, Gjøvik, Norway, e-mail: haris.a.khan@ntnu.no

1 Introduction

In recent years, robots are being deployed in many areas where automation and decision-making skills are required. Robots are not just mechanically advanced but are becoming intelligent as well, and the idea behind these intelligent machines is the creation of systems that imitate the human behavior to be able to perform tasks which are actually infeasible for humans. The type of tasks for which robots are well adapted includes those that are in unexplored environment such as outer-space [63, 66] and undersea [4, 16, 27]. However, the robot tasks are not limited just to complex and difficult problems, but they are covering some industrial [22], medical [18], and domestic applications [21] as well.

A human can search and find an object visually in a cluttered scene. It is a very simple task for human to pick an object and place it in the required place while avoiding obstacles along the path, and without damaging the fragile objects. These simple and trivial tasks for humans become challenging and complex for robots and can overcome their capabilities. The majority of pick-up and drop applications through robots are performed in fully known and structured environments. The key question that arises in this context is how robots can perform as well as humans in these tasks when the structure of the environment is varied?

Human vision is extremely robust and can easily classify objects among tens of thousands of possibilities [11] within a fraction of a second [45]. The human system is able to tolerate the tremendous changes in scale, illumination, noise, and viewing angles for object recognition. Contrary to the human vision, the object recognition is a very complex problem and still beyond the capabilities of artificial vision systems. This contrast between vision systems and the human brain for performing visual recognition and classification tasks gave rise to the development of several approaches to visual recognition.

The ability to recognize and manipulate a large variety of objects is critical for mobile robots. Indoor environment often contains several objects on which the robot should make different actions such as ”Pick-up the remote control TV!”, ”Drop it inside the box!”. So, how to represent and classify objects to be recognized by robots?

Several techniques have been explored in order to achieve this goal. Recently, appearance-based methods have been successfully applied to the problem of object recognition. These methods typically proceed with two phases. In the first phase, a model is constructed from a set of training images that includes the appearance of the object under different illuminants, scales, and multiple instances. Whereas, in the second phase, the methods try to extract parts from the input image through segmentation or by using the sliding windows over the whole image. The methods then compare extracted parts of the input image with the training set. A popular strategy of appearance-based methods is the Bag of Words (BoW). BoW is inspired from text-retrieval systems that count how many times a word appears in a document. It aims to represent an image as an orderless set of local regions. In general, local regions are discretized into a visual vocabulary. This method obtains excellent

results in image classification [5], image retrieval [68], object detection as well as scene classification [25].

With the advent of new 3D sensors like Microsoft Kinect, 3D perception became a fundamental vision research in mobile robotic applications. The Point Cloud Library (PCL) was developed by Rusu *et al.* [49] in 2010 and officially published in 2011. This open source library, licensed under Berkeley Software Distribution (BSD) terms, represents a collection of state-of-the-art algorithms and tools that operate with 3D point clouds. Several studies have been made based on PCL detectors and descriptors for 3D object recognition applications. PCL integrates several 3D detectors as well as 3D local and global descriptors. In 3D local descriptors, each point is described by its local geometry. They are developed for specific applications such as object recognition, and local surface categorization. This local category includes Signature of Histograms of OrienTation (SHOT) [59], Point Feature Histograms (PFH) [47], Fast Point Feature Histograms (FPFH) [46], and SHOT-COLOR [60]. On the other hand, the 3D global descriptors describe object geometry and they are not computed for individual points, but for a whole cluster instead. The global descriptors are high-dimensional representations of object geometry. They are usually calculated for subsets of the point clouds that are likely to be objects. The global category encodes only the shape information and includes Viewpoint Feature Histogram (VFH) [48], Clustered Viewpoint Feature Histogram (CVFH) [2], Oriented Unique and Repeatable CVFH (OUR-CVFH) [1], and Ensemble of Shape Functions (ESF) [64].

The ability to recognize objects is highly valuable for performing imperative tasks in mobile robotics. In several works, authors use classification methods such as Support Vector Machines (SVMs) [33] [67] [26], Nearest Neighbor (NN) [40], Artificial Neural Network (ANN) [7], or Hidden Markov Model (HMM) [61] in order to predict the object class. Recently, researchers got interested in deep learning algorithms because they can simultaneously and automatically discover both low-level and high-level features. Deep Belief Network (DBN) is a graphical model consisting of undirected networks at the top hidden layers and directed networks in the lower layers. The learning algorithm uses greedy layer-wise training by stacking restricted Boltzmann machines (RBMs) which contain hidden layer for modeling the probability distribution of visible variables [10].

In this paper, we present new 2D/3D object recognition and categorization approaches which are based on local and global descriptors. We describe 2D objects and 3D point clouds using 2D/3D local and global descriptors. Then, we train separately these features with Deep Belief Network. We summarize our contributions as follows:

1. We describe an object database with SURF feature points which are quantified with the k-means clustering algorithm to make the 2D Bag of Words;
2. We describe a point cloud with spin image features which we quantify with the k-means clustering algorithm to generate the 3D Bag of Words;
3. We propose VFH-Color descriptor that combines both the color information and geometric features extracted from the previous version of VFH descriptor. We extract the color information for point cloud data, and then we use the color

quantization technique to obtain the color histogram which is combined with VFH histogram.

We organize the rest of our paper in the following way. In Section 2, we provide a literature review on the relevant works of 2D/3D object recognition and categorization. We describe in details our proposed approaches in Section 3. The implementation details and the experimental results are presented in Section 4. Finally, we conclude the paper in Section 5.

2 State of The Art

2.1 2D recognition and categorization

Recently, the approaches that were based on Bag of Words (BoW), also known as Bag of features produced the promising results on several applications, such as object and scene recognition [13] [35], localization and mapping for mobile robots [20], video retrieval [54], text classification [6], and language modeling for image classification and retrieval [34] [39] [70].

Sivic *et al.* [53] used Latent Dirichlet Allocation (LDA) and probabilistic Latent Semantic Analysis (pLSA) in order to compute latent concepts in images from the cooccurrences of visual words. The authors aim to generate a consistent vocabulary of visual words that is insensitive to viewpoint changes and illumination. For this reason, they use vector quantized SIFT descriptors which are invariant to translation, rotations, and re-scaling of the image.

Csurka *et al.* [15] developed a generic visual categorization approach for identifying the object content of natural images. In the first step, their approach detects and describes image patches which are clustered with a vector quantization algorithm to generate a vocabulary. The second step constructs a bag of keypoints that counts the number of patches assigned to each cluster. Finally, they use Naive Bayes and SVM to determine image categories.

Fergus *et al.* [19] suggested an object class recognition method that learns and recognizes object class models from unlabeled and unsegmented cluttered scenes in a scale invariant manner. The approach exploits a probabilistic model that combines shape, appearance, occlusion and relative scale, as well as an entropy-based feature detector to select regions and their scale within an image.

Philbin *et al.* [44] proposed a large-scale object retrieval system with large vocabularies and fast spatial matching. They extract features from each image in a high-dimensional descriptor space which are quantized or clustered to map every feature to a ”visual word”. This visual word is used to index the images for the search engine.

Wu *et al.* [65] proposed a new scheme to utilize optimized bag of words models called Semantics Preserving Bag of Words (SPBoW) that aims to map semantically related features to the same visual words. SPBoW computes a distance between

identical features as a measurement of the semantic gap and tries to learn a codebook by minimizing this gap.

Larlus *et al.* [32] combined a bag of words recognition component with spatial regularization based on a random field and a Dirichlet process mixture for category-level object segmentation. The random field (RF) component assures short-range spatial contiguity of the segmentation while a Dirichlet process component assures mid-range spatial contiguity by modeling the image as a composition of blobs. Finally, the bag of words component allows strong intra-class imaging variations and appearance.

Vigo *et al.* [62] exploited color information in order to improve the bag of words technique. They select highly informative color-based regions for feature extraction. Then, feature description focuses on shape and can be improved with a color description of the local patches. The experiments show that color information should be used both in the feature detection as well as the feature extraction stages.

Khan *et al.* [29] suggested integration of spatial information in the bag of visual words. The approach model the global spatial distribution of visual words that consider the interaction among visual words regardless of their spatial distances. The first step consists of computing pair of identical visual words (PIW) that save all the pairs of visual words of the same type. The second step represents a spatial distribution of words as a histogram of orientations of the segments formed by PIW.

2.2 3D recognition and categorization

Most of the recent work on 3D object categorization focused on appearance, shapes, and Bag of Words (BoW) extracted from certain viewing point changes of the 3D objects.

Savarese and Fei-Fei [50] proposed a compact model for representing and learning 3D object categories. Their model aims to solve scale changes and arbitrary rotations problems using appearance and 3D geometric shape. Each object is considered as a linked set of parts that are composed of many local invariant features. Their approach can classify, localize and infer the scale as well as the pose estimation of objects in the given image.

Toldo *et al.* [58] introduced Bag of Words (BoW) approach for 3D object categorization. They used spectral clustering to select seed-regions then computed the geometric features of the object sub-parts. Vector quantization is applied to these features in order to obtain BoW histograms for each mesh. Finally, Support Vector Machine is used to classify different BoW histograms for 3D objects.

Nair and Hinton [42] presented a top-level model of Deep Belief Networks (DBNs) for 3D object recognition. This model is a third-order Boltzmann machine that is trained using a combination of both generative and discriminative gradients. The model performance is evaluated on NORB images where the dimensionality for each stereo-pair image is reduced by using a foveal image. The final represen-

tation consists of 8976-dimensional vectors that are learned with a top-level model for Deep Belief Nets (DBNs).

Zhong [69] introduced an approach for 3D point cloud recognition based on a new 3D shape descriptor called Intrinsic Shape Signature (ISS). ISS uses a view-dependent transform encoding for the viewing geometry to facilitate fast pose estimation, and a view-independent representation of the 3D shape in order to match shape patches from different views directly.

Bo *et al.* [12] introduced a set of kernel features for object recognition. The authors develop kernel descriptors on depth maps that model size, depth edges, and 3D shape. The main match kernel framework defines pixel attributes and designs match kernels in order to measure the similarities of image patches to determine low dimensional match kernels.

Lai *et al.* [30] built a new RGBD dataset and proposed methods for recognizing RGBD objects. They used SIFT descriptor to extract visual features and spin image descriptor to extract shape features that are used for computing efficient match kernel (EKM). Finally, linear support vector (LiSVM), gaussian kernel support vector machine (kSVM) and random forest (RF) are trained to recognize both the category and the instance of objects.

Mian *et al.* [41] suggested a 3D object retrieval approach from cluttered scenes based on the repeatability and quality of keypoints. The authors proposed a quality measure to select the best keypoints for extracting local features. They also introduced an automatic scale selection method for extracting scale and multi-scale invariant features in order to match objects at different unknown scales.

Madry *et al.* [38] proposed the Global Structure Histogram (GSH) to describe the point cloud information. Their approach encodes the structure of local feature response on a coarse global scale to retain low local variations and keep the advantage of global representativeness. GSH can be instantiated in partial object views and trained using complete or incomplete information about an object.

Tang *et al.* [57] proposed a Histogram of Oriented Normal Vectors (HONV) feature which is based on local geometric characteristics of an object captured from the depth sensor. They considered that the object category information is presented on its surface. This surface is described by the normal vector at each surface point and the local 3D geometry is presented as a local distribution of the normal vector orientation.

Socher *et al.* [56] introduced the first convolutional-recursive deep learning model for 3D object recognition. They computed a single CNN layer to extract low-level features from both color and depth images. These representations are provided as input to a set of RNNs with random weights that produce high-quality features. Finally, The concatenation of all the resulting vectors forms the final feature vector for a softmax classifier.

Schwarz *et al.* [51] developed a meaningful feature set that results from the pre-trained stage of Convolutional Neural Network (CNN). The depth and RGB images are processed independently by CNN and the resulting features are then concatenated to determine the category, instance, and pose of the object.

Eitel *et al.* [17] presented two separate CNN processing streams for RGBD object recognition. RGB and colorized depth images consist of five convolutional layers and two fully connected layers. Both streams are processed separately through several layers and converge into one fully connected layer and a softmax layer for the classification task.

Alex [3] proposed a new approach for RGBD object classification. Four independent Convolutional Neural Networks (CNNs) are trained, one for each depth data and three for RGB data and then trains these CNNs in a sequence. The decisions of each network are combined to obtain the final classification result.

Ouadiay *et al.* [43] proposed a new approach for real 3D object recognition and categorization using Deep Belief Networks. First, they extracted 3D keypoints from point clouds using 3D SIFT detector and then they computed SHOT/SHOTCOLOR descriptors. The performance of the approach is evaluated on two datasets: Washington RGBD object dataset and real 3D object dataset.

Madai *et al.* [37] reinvestigated Deep Convolutional Neural Networks (DCNNs) for RGBD object recognition. They proposed a new method for depth colorization based on surface normals, which colorized the surface normals for every pixel and computed the gradients in a horizontal direction (x-axis) and vertical direction (y-axis) using the Sobel operator. The authors defined two 3D vectors a and b in direction of the z-axis in order to calculate the surface normal n. As n has 3 dimensions, the authors map each of the three values of the surface normal to a corresponding RGB channels.

3 Our Recognition Pipelines

In this paper, we suggest new approaches for 2D/3D object recognition and categorization for mobile robotic applications. We introduce two different recognition pipelines, one relies on 2D/3D detectors and descriptors which are quantified with a k-means algorithm to obtain 2D/3D Bag of Words, while the other one uses our new 3D global descriptor called VFH-Color. Figure 1 summarizes the main steps of 2D/ 3D Bag of Words approaches.

1. **Training set:** represents a set of data (images or point clouds) used on our experiment. Training means, creating a dataset with all the objects which we want to recognize.
2. **Keypoint extraction:** is the first step of our approach where keypoints (interest points) are extracted from input data. It reduces the computational complexity by identifying particularly those regions of images which are important for descriptors, in term of high information density.
3. **Keypoint description:** once keypoints are extracted, descriptors are computed on the obtained keypoints and these form a description which is used to represent the data.
4. **Vocabulary:** after the extraction of descriptors, the approach uses the vector quantization technique to cluster descriptors in their feature space. Each clus-

ter is considered as "visual word vocabulary" that represents the specific local pattern shared by the keypoints in this cluster.

5. **Bag of Words:** is a vector containing the (weighted) count or occurrence of each visual word in the data which is used as the feature vector in the recognition and classification tasks.
6. **Classificiation:** all data in training set are represented by their Bag of Words vectors which represent the input of DBN classifier.

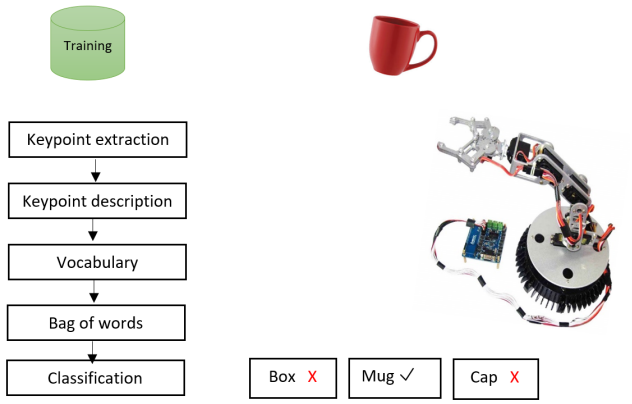


Fig. 1: Overview of 2D/3D Bag of Words approaches.

For the global pipeline, we present a new VFH-Color descriptor that combines both the color information and the geometric features extracted from the previous version of VFH descriptor. Figure 2 summarizes the main steps of the global approach.

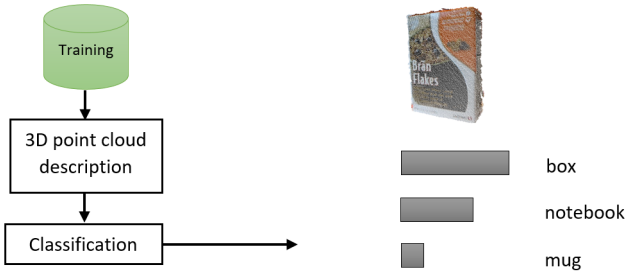


Fig. 2: Overview of 3D global approach.

1. **Training set:** represents a set of point clouds used on our experiment.

2. **3D point description:** extracts the color information for point cloud data, then uses the color quantization technique to obtain the color histogram which is combined with VFH histogram extracted from the previous version of VFH descriptor.
3. **Classificiation:** all point clouds in training set are represented by their VFH-Color features and are provided as the input to DBN classifier.

3.1 Object Representation

3.1.1 2D Bag of Words

2D Speeded-Up Robust Features (SURF) detector

Keypoints are important features that are becoming more and more widespread in image analysis. The Speeded-Up Robust Features (SURF) [8, 9] is based on the same steps and principles of SIFT detector [36], but it utilizes a different scheme and provides better results than those obtained with SIFT extractor. SURF is scale and rotation invariant keypoint detector that uses a very basic Hessian-matrix approximation because of its good performance in term of accuracy. Gaussian kernels are optimal for scale-space analysis. SURF divides the scale space into levels and octaves where each octave corresponds to a doubling of scale and is divided into uniformly spaced levels. The method builds a pyramid of response maps with various levels within octaves. The keypoints represent the points that are the extrema among 8 neighbors in the current level and its 2*9 neighbors in the above and below levels.

2D Speeded-Up Robust Features (SURF) descriptor

SURF descriptor provides a unique and robust description of a feature that can be generated on the area surrounding a keypoint. SURF descriptor is based on Haar Wavelet responses and can be calculated efficiently with integral images. SURF describes an interesting area with size 20s, then each interest area is divided into 4*4 sub-areas and is described by the values of a wavelet response in the x and y directions. The interest areas are weighted with a Gaussian centered at the keypoint for being robust in deformations and translations. For each sub-area, a vector v is calculated, based on 5*5 samples. The descriptor for keypoint consists of 16 vectors for the sub-areas being concatenated. Finally, the descriptor is normalized, to achieve invariance to contrast variations that will represent themselves as a linear scaling of the descriptor.

Visual vocabulary

Once the keypoint descriptors are obtained, the approach imposes a quantization on the feature space of these descriptors. The standard pipeline to obtain ”visual vocabulary” is also called ”codebook” which consists of (i) collecting a large sample of a local feature; (ii) quantizing the feature space according to their statistics. Most vector quantization or clustering algorithms are based on hierarchical or iterative square error partitioning methods. Hierarchical methods organize data on groups which can be displayed in the form of the tree. Whereas, square-error partitioning algorithms attempt to obtain which maximizes the between cluster scatter or minimizes the within-cluster scatter. In our work, we use a simple k-means clustering algorithm. The ”visual words” or ”codevector” represent the k cluster centers. A vector quantizer takes a feature vector and maps it to the index of the nearest codevector in the codebook using the Euclidean distance.

Bag of Words

Bag of Words is generated by computing the count or occurrence of each visual word in the image which is used as the feature vector in the recognition and classification tasks.

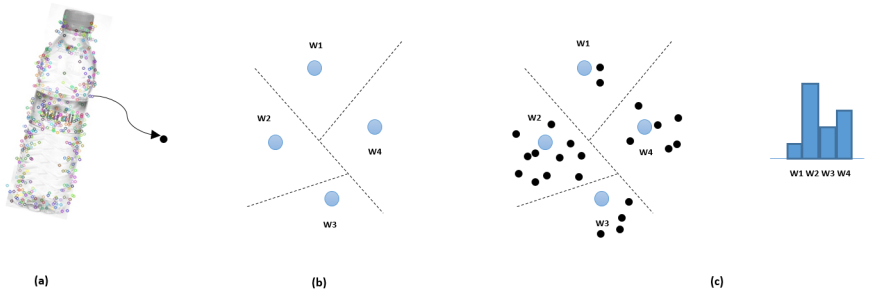


Fig. 3: The schematic illustrates visual vocabulary construction and word assignment. (a) the black dot represents SURF keypoint, the object contains in total 240 SURF keypoints. Next, the approach computes SURF descriptor on each keypoint. (b) Visual words (W1, W2, W3, and W4) denote cluster centers. (c) The sampled features are clustered in order to optimize the space into a discrete number of visual words. A bag of visual words histogram can be used to summarize the entire image. It counts the occurrence of each visual word in the image.

3.1.2 3D Bag of Words

3D Scale-invariant feature transform (SIFT) detector

Scale-invariant feature transform (SIFT) is an algorithm deployed in the field of computer vision to detect and describe regions in an image and identify similar elements between varying images. This process is called "matching". The algorithm consists of the detected feature points of an image which are used to characterize every point that needs to be recognized by comparing its characteristics with those of the points contained in other images. The general idea of SIFT is to find the keypoints that are invariant to several transformations/changes: rotation, scale, illumination and viewing angle. The 3D SIFT detector [52] use the Difference-of-Gaussian (DoG) function to extract the extrema points in both spatial and scale dimensions.

Spin image descriptor

The spin image was proposed to describe points of interest by [28]. This descriptor translates the local properties of the surface oriented in a coordinate system fixed and linked to the object. This system is independent of the viewing angle. The spin is defined at a point oriented and designated by its 3D position (p) as well as associated direction (n the normal to the local surface). A 2D local coordinate base is formed using the tangent plane P in the point p, oriented perpendicularly to the normal n, and the line L through p parallel to n. A cylindrical coordinate system (α, β) of the point p is then deduced. The radial coordinate defining the distance (non-negative) is perpendicular to L and the elevation coordinate of the defined distance is perpendicular to P (signed positive or negative). The resulting histogram is formed by counting the occurrences of different pairs of discretized distances.

Visual vocabulary

After describing each of the point clouds inside a class with the spin image, we need to make the visual categorization using the probabilistic approach. The method we use consists of applying a quantization operation with the k-means clustering and constructs visual words with the well-known method of the bag of features.

Bag of Words

Instead of considering each feature point a visual word, we consider thanks to the quantization that each of the clusters' center represent a word. The bag of words algorithm consists of computing the number of occurrences of each word in the

model database. It is like a probability of the number of words inside the class of objects.

3.1.3 Viewpoint Feature Histogram Color (VFH-Color)

The viewpoint feature histogram (VFH) [48] computes a global descriptor of the point cloud and consists of two components: a surface shape component and a viewpoint direction component. VFH aims to combine the viewpoint direction directly into the relative normal angle calculation in the FPFH descriptor [46]. The viewpoint-dependent component of the descriptor is a histogram of the angles between the vector $(p_c - p_v)$ and each point's normal. This component is binned into a 128-bin histogram. The other component is a simplified point feature histogram (SPFH) estimated for the centroid of the point cloud, and an additional histogram of distances of all points in the cloud to the cloud's centroid. The three angles (α, ϕ, θ) with the distance d between each point and the centroid are binned into a 45-bin histogram. The total length of VFH descriptor is the combination of these two histograms and is equal to 308 bins.

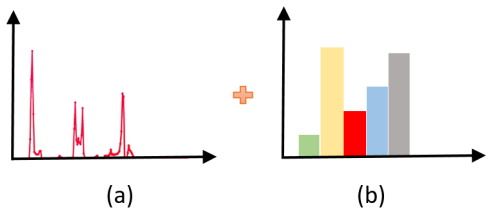


Fig. 4: VFH-Color. (a) VFH descriptor. (b) Color quantization.

Color quantization is a vector quantization that aims to select K vectors in N dimensional space in order to represent N vectors from that space ($K << N$). In general, color quantization is applied to reduce the number of colors in a given image while maintaining the visual appearance of the original image. Color quantization is applied in a 3-dimensional space RGB and follows the following steps:

1. Extract RGB features for each point from the point cloud data;
2. Obtain the matrix of RGB features (number of points$\times$3);
3. Compute k-means algorithm for the RGB matrix in order to generate the codebook (cluster centers);
4. Count the occurrence of each codebook in the point cloud.

The codebook size represents the bins of color quantization histogram. After a set of experiments, we fix the codebook size to 100 bins (see Figure 6). Therefore, VFH-Color histogram concatenates 308 values of original VFH descriptor and 100 values of color quantization histogram, thus giving the total size of 408 values.

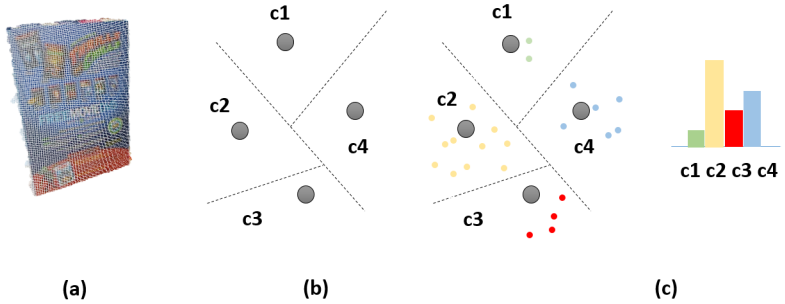


Fig. 5: Color quantization process. (a) point cloud data. (b) codebook (C1, C2, C3, and C4) denote cluster centers. (c) The RGB features are clustered in order to optimize the space. The histogram counts the occurrence of each codebook in the point cloud.

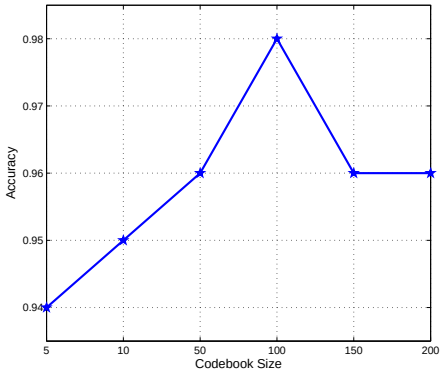


Fig. 6: The classification performance with respect to the codebook size.

3.2 Object Classification

3.2.1 Restricted Boltzmann Machines (RBMs)

Restricted Boltzmann Machines (RBMs) [55] are a specific category of energy based model which include hidden variables. RBMs are restricted in the sense so that no hidden-hidden or variable-variable connections exist. The architecture of a generative RBM is illustrated in Figure 7.

RBMs are a parameterized generative stochastic neural network which contain stochastic binary units on two layers: the visible layer and the hidden layer.

1. Visible units (the first layer): they contain visible units (x) that correspond to the components of an observation (i.e. 2D/3D features in this case of study);

2. Hidden units (the second layer): they contain hidden units (h) that model dependencies between the components of observations.

The stochastic nature of RBMs results from the fact that the visible and hidden units are stochastic. The units are binary, i.e. $x_i, h_j \in \{0,1\} \forall\ i$ and j, and the joint probability which characterize the RBM configuration is the Boltzmann distribution:

$$p(x,h) = \frac{1}{Z} e^{-E(x,h)} \tag{1}$$

The normalization constant is $Z = \sum_{x,h} e^{-E(x,h)}$ and the energy function of an RBM is defined as:

$$E(x,h) = -b^{'} x - c^{'} h - h^{'} W x \tag{2}$$

where:

- W represents the symmetric interaction term between visible units (x) and hidden units (h);
- b and c are vectors that store the visible (input) and hidden biases (respectively).

RBMs are proposed as building blocks of multi-layer learning deep architectures called deep belief networks. The idea behind is that the hidden neurons extract pertinent features from the visible neurons. These features can work as the input to another RBM. By stacking RBMs in this way, the model can learn features for a high-level representation.

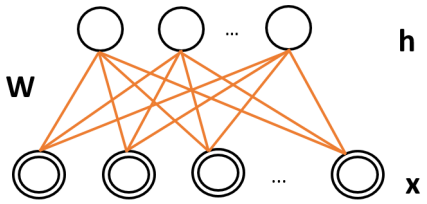


Fig. 7: RBM model. The visible units x and hidden units h are connected through undirected and symmetric connections. There are no intra-layer connections.

3.2.2 Deep Belief Network (DBN)

Deep Belief Network (DBN) is the probabilistic generative model with many layers of stochastic and hidden variables. Hinton *et al.* [24] introduced the motivation for using a deep network versus a single hidden layer (i.e. a DBN vs. an RBM). The power of deep networks is achieved by having more hidden layers. However, one of the major problems for training deep network is how to initialize the weights W between the units of two consecutive layers ($j-1$ and j), and the bias b of layer j.

Random initializations of these parameters can cause poor local minima of the error function resulting in low generalization. For this reason, Hinton *et al.* introduced a DBN architecture based on training sequence of RBMs. DBN train sequentially as many RBMs as the number of hidden layers that constitute its architecture, i.e for a DBN architecture with l hidden layers, the model has to train l RBMs. For the first RBM, the inputs consist of the DBN's input layer (visible units) and the first hidden layer. For the second RBM, the inputs consist of the hidden unit activations of the previous RBM and the second hidden layer. The same holds for the remaining RBMs to browse through the l layers. After the model performs this layer-wise algorithm, a good initialization of the biases and the hidden weights of the DBN is obtained. At this stage, the model should determine the weights from the last hidden layer for the outputs. To obtain a successfully supervised learning, the model "fine-tunes" the resulting weights of all layers together. Figure 8 illustrates a DBN architecture with one visible layer and three hidden layers.

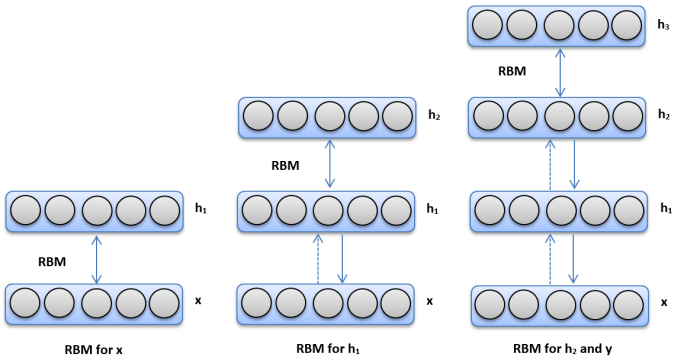


Fig. 8: DBN architecture with one visible layer x and three hidden layers $h1$, $h2$, and $h3$.

4 Experimental Results and Discussion

4.1 Datasets

4.1.1 ALOI dataset

Amsterdam Library of Object Images (ALOI) [23] dataset is an image collection of 1000 small objects recorded for recognition task. 111,250 images are captured by Sony DXC390P 3CCD cameras varying viewing angle, illumination angle and illu-

mination color for each object, and additionally images are captured wide-baseline stereo images.



Fig. 9: The sample images extracted from Amsterdam Library of Object Images (ALOI) dataset.

4.1.2 Washington RGBD Dataset

Washington RGBD dataset is a large dataset built for 3D object recognition and categorization applications. It is a collection of 300 common household objects which are organized into 51 categories. Each object is placed on a turntable and is captured for one whole rotation in order to obtain all object views using Kinect camera that records synchronized and aligned 640x480 RGB and depth images at 30 Hz [31] (see Figure 10).



Fig. 10: The sample point clouds extracted from Washington RGBD Dataset.

4.2 2D/3D object classification

DBN aims to allow each RBM model in the sequence to receive a different representation of the data. In other words, after RBM has been learned, the activity values of its hidden units are used as the training data for learning a higher-level RBM. The input layer has a number N of units, equal to the size of sample data x (size of 2D/3D features). The number of units for hidden layers, currently, are pre-defined according to the experiment. We fixed DBN with three hidden layers $h1$, $h2$ and $h3$. The general DBN characteristics are shown in Table 1.

Table 1: DBN characteristics that are used in our experiments.

Characteristic	**Value**
Hidden layers	3
Hidden layer units	600
Learn rates	0.01
Learn rate decays	0.9
Epochs	200
Input layer units	size of descriptor

4.2.1 2D Bag of Words

Images contain local points or keypoints defined as salient region patches which represent rich local information of the image. We used SURF to automatically detect and describe keypoints from images. Then, we used the vector quantization method in order to cluster the keypoint descriptors in their feature space into a large number of clusters using the k-means clustering algorithm. We test in a set of experiments the impact of the number of clusters on classifier accuracy and we select k=1500 as the size of the codebook (number of visual words) that represents the best accuracy value. We conduct the experiments on ALOI dataset on which we select ten categories: teddy, jam, ball, mug, food_box, towel, shoes, pen, can, and bottle. Figure 9 shows some examples from ALOI dataset which are used in our experiments.

As shown in the confusion matrix (Figure 12), the classes which are consistently misclassified are the teddy, ball, shoes, can, mug, and bottle which are very similar in appearance (Figure 11). The results show also that 2D Bag of Words approach which is based on SURF features works perfectly with the accuracy rate of 91%. BoW representation encodes only the occurrence of the appearance of the local patches and ignores the object geometry. The lack of geometric features can provide some misclassification especially when the objects are similar in appearance. In Table 5, we report accuracy values for 2D Bag of Words with both SVM and DBN classifiers. The first row reports the accuracy value of SVM whereas the second row shows the accuracy value of DBN. We notice that the combination of 2D Bag of Words and

DBN outperforms the 2D Bag of Words with SVM and rises steadily from 88.86% to 90.83%. This result shows the power of deep learning architectures that learn multiple levels of representation depending on the depth of the architecture.



Fig. 11: The objects which are missclassified using 2D Bag of Words classification.

Classes	Metrics			
	wrong class	f1-score	recall	precision
(1)	(3,4,7,9,10)	86%	86%	87%
(2)	(1,7,10)	96%	98%	95%
(3)	(1,4,7)	95%	92%	98%
(4)	(1,3,5,8,9)	95%	96%	93%
(5)	(4,8)	96%	96%	96%
(6)	(1)	100%	99%	100%
(7)	(1,2,8,9,10)	78%	79%	77%
(8)	(2,4,5,7,9,10)	82%	81%	84%
(9)	(5,8,10)	92%	93%	92%
(10)	(1,2,4,5,7,9)	86%	86%	85%
Average	–	91%	91%	91%

Table 2: The performance of 2D Bag of Words.

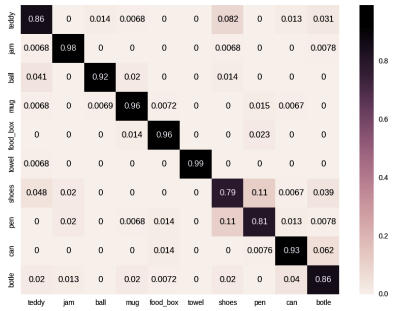


Fig. 12: Confusion Matrix of 2D Bag of Words model.

4.2.2 3D Bag of Words

After extracting the spin image for the set of point clouds, we constructed a shape dictionary whose size is fixed at k=250, by clustering all spin image acquired from the whole training set with k-means method. For each bin, a representative local 3D feature description is required. These descriptions are taken from the centroids of each cluster (visual words) determined by k-means clustering on precomputed spin image descriptors.

Figure 13 represents the confusion matrix across all 10 classes. Most model's results are reasonable showing that 3D Bag of Words can provide high-quality features. The classes that are consistently misclassified are ball-mug-can, bowl-mug-notebook-plate, and food_box-cereal_box-notebook which are very similar in shape. Table 3 illustrates the performance metrics of 3D Bag of Words that encodes only the surface shape of 3D point clouds thanks to the use of spin images descriptor.

Classes	Metrics			
	wrong class	f1-score	recall	precision
(1)	(5,7,10)	94%	95%	94%
(2)	(5,8,9,10)	97%	98%	96%
(3)	(5,6,7,10)	95%	94%	96%
(4)	(6,8)	84%	80%	89%
(5)	(2,3,7,10)	91%	91%	91%
(6)	(4,5,8,10)	87%	90%	84%
(7)	(1,3,5,10)	89%	90%	88%
(8)	(4,6,9,10)	98%	98%	98%
(9)	(2,8)	99%	99%	100%
(10)	(1,2,3,5,6,7,8)	87%	85%	88%
Average	–	92%	92%	92%

Table 3: The performance of 3D Bag of Words.

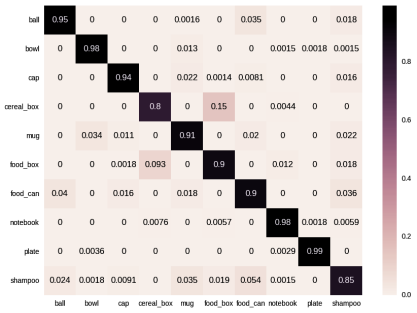


Fig. 13: Confusion Matrix of 3D Bag of Words.

4.2.3 Global Pipeline

VFH-Color descriptor combines both the color information and the geometric features extracted from the previous version of VFH descriptor. We extract the color information for point cloud data, then we use the color quantization technique to obtain the color histogram which is combined with VFH histogram. For each point cloud, we extract two types of features: 1) geometric features extracted from Viewpoint Feature Histogram (VFH) (308 bins), and 2) color features extracted from color quantization (100 bins). These features are then combined into a single vector, being 308+100=408 dimensional. Figure 15 represents the confusion matrix across all 10 classes. Most model's results are very reasonable showing that VFH-Color can provide meaningful features. The classes that are consistently misclassified are mug-cap, cereal_box-food_box, and shampoo-cap-mug-food_can which are very similar in appearance and shape.

Classes	Metrics			
	wrong class	f1-score	recall	precision
(1)	(6)	100%	100%	100%
(2)	(–)	100%	100%	100%
(3)	(4,5,6)	99%	99%	99%
(4)	(6,10)	86%	95%	79%
(5)	(6)	100%	100%	100%
(6)	(3,4,7,10)	88%	83%	94%
(7)	(1,6)	97%	98%	96%
(8)	(–)	100%	100%	100%
(9)	(–)	100%	100%	100%
(10)	(3,4,6,7,8)	95%	92%	97%
Average	–	95%	96%	97%

Table 4: The performance of global pipeline using VFH descriptor.

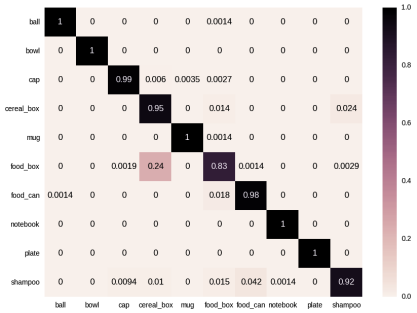


Fig. 14: Confusion Matrix of global pipeline using VFH descriptor.

Moreover, we evaluate the performance of VFH-Color against the previous version of VFH and SHOTCOLOR. The accuracy using VFH-Color performs 3% better than VFH that models only the geometric features. This result shows the effectiveness of the approach after adding the color information. We also notice that SHOTCOLOR presents a good accuracy (Table 7), although this descriptor encounters a problem when it is not able to compute the local reference frame for some point clouds. In this set of experiments, 15% of point clouds from the dataset are not computed with SHOTCOLOR. This problem becomes significant when 3D object recognition is in real time. Indeed, VFH-Color descriptor can be used in the real-time applications thanks to its estimation for every point cloud as well as its good recognition rate.

Table 6 shows also that our global pipeline works perfectly with the accuracy rate of 99.63% with DBN architecture that performs the use of SVM classifier. In general, the use of DBN instead of SVM in our approaches increases the accuracy rate thanks to the performance of deep learning algorithms which outperformed the shallow architectures (SVM).

Table 5: Accuracy of different proposed approaches using DBN and SVM classifiers.

Classifier	**BOW2D**	**BOW3D**	**VFH**	**VFH-Color**	**SHOTCOLOR**
SVM	88.86%	86.68%	95.01 %	98.34%	97.21 %
DBN	90.83%	92.03%	96.41 %	**99.63%**	98.63 %

Classes	**Metrics**			
	wrong class	**f1-score**	**recall**	**precision**
(1)	(–)	100%	100%	100%
(2)	(–)	100%	100%	100%
(3)	(–)	100%	100%	99%
(4)	(6)	99%	99%	99%
(5)	(3)	100%	100%	100%
(6)	(4)	99%	99%	99%
(7)	(–)	100%	100%	99%
(8)	(–)	100%	100%	100%
(9)	(–)	100%	100%	100%
(10)	(3,4,5,6,7)	99%	98%	100%
Average	–	99%	99%	99%

Table 6: The performance of global pipeline using VFH-Color descriptor.

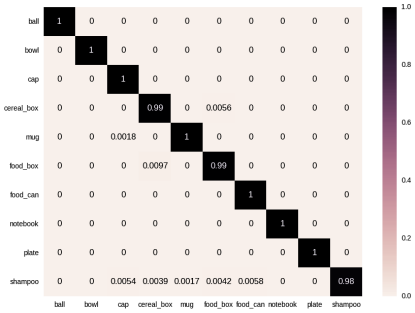


Fig. 15: Confusion Matrix of global pipeline using our VFH-Color descriptor.

Classes	Metrics			
	wrong class	f1-score	recall	precision
(1)	(7)	100%	99%	100%
(2)	(–)	100 %	100%	100%
(3)	(–)	100%	100%	100%
(4)	(6,8)	93%	92%	95%
(5)	(–)	100%	100%	100%
(6)	(4,7,10)	95%	96%	94%
(7)	(6)	99%	100%	99%
(8)	(–)	100%	100%	99%
(9)	(–)	100%	100%	100%
(10)	(7)	99%	100%	99%
Average	–	99 %	99%	99%

Table 7: The performance of SHOTCOLOR descriptor.

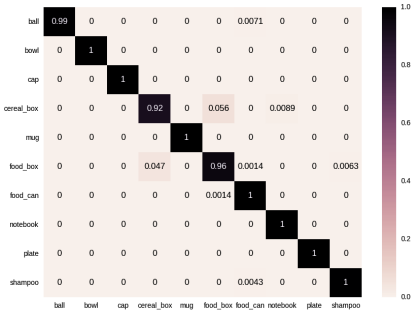


Fig. 16: Confusion Matrix of SHOTCOLOR descriptor.

4.3 Comparison to Other Methods

In this subsection, we compare our contributions to related state-of-the-art approaches. Table 8 shows the main accuracy values and compares our recognition pipelines to the published results [12, 30, 51] and [17, 37]. Lai *et al.* [30] extract a set of features that captures the shape of the object view using a spin image and another set which captures the visual appearance using SIFT descriptors. These features are extracted separately from both depth and RGB images. A recent work by Schwarz *et al.* [51] uses both colorizing depth and RGB images that are processed independently by a convolutional neural network. CNN features are then learned using SVM classifier in order to successively determine the category, instance, and pose. The previous approaches [17, 37] used the color-coding depth maps and RGB images for training separately CNN architecture.

Table 8: The comparison of 3D object recognition accuracies and PCL descriptors on the Washington RGBD dataset.

Approaches	Accuracy rates
Lai *et al.* [30]	90.6%
Bo *et al.* [12]	84.5%
Eitel *et al.* [17]	91%
Madai *et al.* [37]	94%
Schwarz *et al.* [51]	94.1%
VFH and DBN	96.41%
3D BoW and DBN	92%
VFH-Color and DBN	**99.63%**
SHOTCOLOR and DBN	98.63%

In our work, we learn our 3D features using DBN with three hidden layers that model a deep network architecture. The results show also that our global pipeline

works perfectly with the accuracy rate of 99.63% thanks to the efficiency of our VFH-Color descriptor and outperforms all methods that are mentioned in the state-of-the-art.

5 Conclusion and Future Work

In this paper, we proposed new approaches for object categorization and recognition in real-world environment. We used the Bag of Words (BoW) that aims to represent images and point clouds as an orderless of local regions that are discretized into a visual vocabulary. Also, we proposed the VFH-Color descriptor which combined geometric features extracted from Viewpoint Feature Histogram (VFH) descriptor and color information extracted from color quantization method. Then, we learned the 2D and 3D features with Deep Belief Network (DBN) classifier.

The experimental results on ALOI dataset and Washington RGBD dataset clearly ascertain that the proposed algorithms are able of categorizing objects and 3D point clouds. These results are encouraging, especially that our new VFH-Color descriptor performed the state-of-the-art methods in recognizing 3D objects under different views. Also, our approach improved the recognition rates thanks to the use of color information.

In a future work, we will attempt to embed our algorithms in a mobile robot in order for it to recognize and manipulate the real-world objects. We will also develop a new approach using 3D sensors and other deep learning methods.

References

1. ALDOMA, A., TOMBARI, F., RUSU, R., AND VINCZE, M. *OUR-CVFH–Oriented, Unique and Repeatable Clustered Viewpoint Feature Histogram for Object Recognition and 6DOF Pose Estimation*. Springer, 2012.
2. ALDOMA, A., VINCZE, M., BLODOW, N., GOSSOW, D., GEDIKLI, S., RUSU, R., AND BRADSKI, G. Cad-model recognition and 6dof pose estimation using 3d cues. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on* (2011), IEEE, pp. 585–592.
3. ALEXANDRE, L. A. 3d object recognition using convolutional neural networks with transfer learning between input channels. In *Intelligent Autonomous Systems 13*. Springer, 2016, pp. 889–898.
4. ANTONELLI, G., FOSSEN, T. I., AND YOERGER, D. R. Underwater robotics. In *Springer handbook of robotics*. Springer, 2008, pp. 987–1008.
5. AVILA, S., THOME, N., CORD, M., VALLE, E., AND ARAÚJO, A. D. A. Bossa: Extended bow formalism for image classification. In *2011 18th IEEE International Conference on Image Processing* (2011), IEEE, pp. 2909–2912.
6. BAI, J., NIE, J.-Y., AND PARADIS, F. Using language models for text classification. In *Proceedings of the Asia Information Retrieval Symposium, Beijing, China* (2004).
7. BASU, J. K., BHATTACHARYYA, D., AND KIM, T.-H. Use of artificial neural network in pattern recognition. *International Journal of Software Engineering and Its Applications 4*, 2 (2010).

8. BAY, H., ESS, A., TUYTELAARS, T., AND VAN GOOL, L. Speeded-up robust features (surf). *Computer vision and image understanding 110*, 3 (2008), 346–359.
9. BAY, H., TUYTELAARS, T., AND VAN GOOL, L. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*. Springer, 2006, pp. 404–417.
10. BENGIO, Y. Learning deep architectures for ai. *Foundations and trends® in Machine Learning 2*, 1 (2009), 1–127.
11. BIEDERMAN, I. Recognition-by-components: a theory of human image understanding. *Psychological review 94*, 2 (1987), 115.
12. BO, L., REN, X., AND FOX, D. Depth kernel descriptors for object recognition. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on* (2011), IEEE, pp. 821–826.
13. BOLOVINOU, A., PRATIKAKIS, I., AND PERANTONIS, S. Bag of spatio-visual words for context inference in scene classification. *Pattern Recognition 46*, 3 (2013), 1039–1053.
14. BOSER, B. E., GUYON, I. M., AND VAPNIK, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (1992), ACM, pp. 144–152.
15. CSURKA, G., DANCE, C., FAN, L., WILLAMOWSKI, J., AND BRAY, C. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (2004), vol. 1, Prague, pp. 1–2.
16. DUNBABIN, M., CORKE, P., VASILESCU, I., AND RUS, D. Data muling over underwater wireless sensor networks using an autonomous underwater vehicle. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.* (2006), IEEE, pp. 2091–2098.
17. EITEL, A., SPRINGENBERG, J. T., SPINELLO, L., RIEDMILLER, M., AND BURGARD, W. Multimodal deep learning for robust rgb-d object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on* (2015), IEEE, pp. 681–687.
18. FEI, B., NG, W. S., CHAUHAN, S., AND KWOH, C. K. The safety issues of medical robotics. *Reliability Engineering & System Safety 73*, 2 (2001), 183–192.
19. FERGUS, R., PERONA, P., AND ZISSERMAN, A. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on* (2003), vol. 2, IEEE, pp. II–264.
20. FILLIAT, D. A visual bag of words method for interactive qualitative localization and mapping. In *Robotics and Automation, 2007 IEEE International Conference on* (2007), IEEE, pp. 3921–3926.
21. FORLIZZI, J., AND DISALVO, C. Service robots in the domestic environment: a study of the roomba vacuum in the home. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* (2006), ACM, pp. 258–265.
22. FREUND, E. Fast nonlinear control with arbitrary pole-placement for industrial robots and manipulators. *The International Journal of Robotics Research 1*, 1 (1982), 65–78.
23. GEUSEBROEK, J.-M., BURGHOUTS, G. J., AND SMEULDERS, A. W. The amsterdam library of object images. *International Journal of Computer Vision 61*, 1 (2005), 103–112.
24. HINTON, G. E., OSINDERO, S., AND TEH, Y.-W. A fast learning algorithm for deep belief nets. *Neural computation 18*, 7 (2006), 1527–1554.
25. HU, F., XIA, G.-S., WANG, Z., HUANG, X., ZHANG, L., AND SUN, H. Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8*, 5 (2015).
26. JANOCH, A., KARAYEV, S., JIA, Y., BARRON, J. T., FRITZ, M., SAENKO, K., AND DARRELL, T. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*. Springer, 2013, pp. 141–165.
27. JAULIN, L. Robust set-membership state estimation; application to underwater robotics. *Automatica 45*, 1 (2009), 202–206.
28. JOHNSON, A., AND HEBERT, M. Using spin images for efficient object recognition in cluttered 3d scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 21*, 5 (1999), 433–449.

29. KHAN, R., BARAT, C., MUSELET, D., AND DUCOTTET, C. Spatial orientations of visual word pairs to improve bag-of-visual-words model. In *Proceedings of the British Machine Vision Conference* (2012), BMVA Press, pp. 89–1.
30. LAI, K., BO, L., REN, X., AND FOX, D. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on* (2011), IEEE, pp. 1817–1824.
31. LAI, K., BO, L., REN, X., AND FOX, D. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on* (2011), IEEE, pp. 1817–1824.
32. LARLUS, D., VERBEEK, J., AND JURIE, F. Category level object segmentation by combining bag-of-words models with dirichlet processes and random fields. *International Journal of Computer Vision 88*, 2 (2010), 238–253.
33. LECUN, Y., HUANG, F. J., AND BOTTOU, L. Learning methods for generic object recognition with invariance to pose and lighting. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on* (2004), vol. 2, IEEE, pp. II–97.
34. LI, M., MA, W.-Y., LI, Z., AND WU, L. Visual language modeling for image classification, Feb. 28 2012. US Patent 8,126,274.
35. LI, T., MEI, T., KWEON, I.-S., AND HUA, X.-S. Contextual bag-of-words for visual categorization. *Circuits and Systems for Video Technology, IEEE Transactions on 21*, 4 (2011), 381–392.
36. LOWE, D. G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on* (1999), vol. 2, Ieee, pp. 1150–1157.
37. MADAI-TAHY, L., OTTE, S., HANTEN, R., AND ZELL, A. Revisiting deep convolutional neural networks for rgb-d based object recognition. In *International Conference on Artificial Neural Networks* (2016), Springer, pp. 29–37.
38. MADRY, M., EK, C. H., DETRY, R., HANG, K., AND KRAGIC, D. Improving generalization for 3d object categorization with global structure histograms. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on* (2012), IEEE, pp. 1379–1386.
39. MC DONALD, K. R. *Discrete language models for video retrieval*. PhD thesis, Dublin City University, 2005.
40. MCCANN, S., AND LOWE, D. G. Local naive bayes nearest neighbor for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (2012), IEEE, pp. 3650–3656.
41. MIAN, A., BENNAMOUN, M., AND OWENS, R. On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *International Journal of Computer Vision 89*, 2-3 (2010), 348–361.
42. NAIR, V., AND HINTON, G. E. 3d object recognition with deep belief nets. In *Advances in Neural Information Processing Systems* (2009), pp. 1339–1347.
43. OUADIAY, F. Z., ZRIRA, N., BOUYAKHF, E. H., AND HIMMI, M. M. 3d object categorization and recognition based on deep belief networks and point clouds. In *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics* (2016), pp. 311–318.
44. PHILBIN, J., CHUM, O., ISARD, M., SIVIC, J., AND ZISSERMAN, A. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (2007), IEEE, pp. 1–8.
45. POTTER, M. C. Short-term conceptual memory for pictures. *Journal of experimental psychology: human learning and memory 2*, 5 (1976), 509.
46. RUSU, R., BLODOW, N., AND BEETZ, M. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on* (2009), IEEE, pp. 3212–3217.
47. RUSU, R., BLODOW, N., MARTON, Z., AND BEETZ, M. Aligning point cloud views using persistent feature histograms. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pp. 3384–3391.

48. RUSU, R., BRADSKI, G., THIBAUX, R., AND HSU, J. Fast 3d recognition and pose using the viewpoint feature histogram. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on* (2010), IEEE, pp. 2155–2162.
49. RUSU, R., AND COUSINS, S. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)* (Shanghai, China, May 9-13 2011).
50. SAVARESE, S., AND FEI-FEI, L. 3d generic object categorization, localization and pose estimation. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (2007), IEEE, pp. 1–8.
51. SCHWARZ, M., SCHULZ, H., AND BEHNKE, S. Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on* (2015), IEEE, pp. 1329–1335.
52. SCOVANNER, P., ALI, S., AND SHAH, M. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia* (2007), ACM, pp. 357–360.
53. SIVIC, J., RUSSELL, B. C., EFROS, A. A., ZISSERMAN, A., AND FREEMAN, W. T. Discovering object categories in image collections.
54. SIVIC, J., AND ZISSERMAN, A. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (2003), IEEE, pp. 1470–1477.
55. SMOLENSKY, P. Information processing in dynamical systems: Foundations of harmony theory.
56. SOCHER, R., HUVAL, B., BATH, B., MANNING, C. D., AND NG, A. Y. Convolutional-recursive deep learning for 3d object classification. In *Advances in Neural Information Processing Systems* (2012), pp. 665–673.
57. TANG, S., WANG, X., LV, X., HAN, T. X., KELLER, J., HE, Z., SKUBIC, M., AND LAO, S. Histogram of oriented normal vectors for object recognition with a depth sensor. In *Asian conference on computer vision* (2012), Springer, pp. 525–538.
58. TOLDO, R., CASTELLANI, U., AND FUSIELLO, A. A bag of words approach for 3d object categorization. In *Computer Vision/Computer Graphics CollaborationTechniques*. Springer, 2009, pp. 116–127.
59. TOMBARI, F., SALTI, S., AND D. STEFANO, L. Unique signatures of histograms for local surface description. In *Computer Vision–ECCV 2010*. Springer, 2010, pp. 356–369.
60. TOMBARI, F., SALTI, S., AND STEFANO, L. A combined texture-shape descriptor for enhanced 3d feature matching. In *Image Processing (ICIP), 2011 18th IEEE International Conference on* (2011), IEEE, pp. 809–812.
61. TORRALBA, A., MURPHY, K. P., FREEMAN, W. T., AND RUBIN, M. A. Context-based vision system for place and object recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (2003), IEEE, pp. 273–280.
62. VIGO, D. A. R., KHAN, F. S., VAN DE WEIJER, J., AND GEVERS, T. The impact of color on bag-of-words based object recognition. In *Pattern Recognition (ICPR), 2010 20th International Conference on* (2010), IEEE, pp. 1549–1553.
63. VISENTIN, G., VAN WINNENDAEL, M., AND PUTZ, P. Advanced mechatronics in esa's space robotics developments. In *Advanced Intelligent Mechatronics, 2001. Proceedings. 2001 IEEE/ASME International Conference on* (2001), vol. 2, IEEE, pp. 1261–1266.
64. WOHLKINGER, W., AND VINCZE, M. Ensemble of shape functions for 3d object classification. In *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on* (2011), IEEE, pp. 2987–2992.
65. WU, L., HOI, S. C., AND YU, N. Semantics-preserving bag-of-words models and applications. *Image Processing, IEEE Transactions on 19*, 7 (2010), 1908–1920.
66. YOSHIDA, K. Achievements in space robotics. *IEEE Robot. Automat. Mag. 16*, 4 (2009), 20–28.
67. ZHANG, H., BERG, A. C., MAIRE, M., AND MALIK, J. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (2006), vol. 2, IEEE, pp. 2126–2136.

68. ZHENG, L., WANG, S., LIU, Z., AND TIAN, Q. Packing and padding: Coupled multi-index for accurate image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 1939–1946.
69. ZHONG, Y. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on* (2009), IEEE, pp. 689–696.
70. ZHU, L., RAO, A. B., AND ZHANG, A. Theory of keyblock-based image retrieval. *ACM Transactions on Information Systems (TOIS) 20*, 2 (2002), 224–257.