An Evaluation of Local Shape Descriptors in Probabilistic Volumetric Scenes

Maria I. Restrepo
http://vision.lems.brown.edu/students/restrepo
Joseph L. Mundy
http://vision.lems.brown.edu/faculty/mundy

LEMS Laboratory
School of Engineering
Brown University
Providence, RI. USA

Abstract

This paper presents the first performance evaluation of local shape descriptors in probabilistic volumetric models (PVM) that are learned from multi-view aerial imagery of large scale urban scenes. The PVM offers a dense solution to the multi-view stereo problem, handling in a probabilistic manner the ambiguities caused by highly reflective surfaces, varying illumination conditions, registration errors, and sensor noise. A GPU-based octree implementation guarantees scalability of the PVM to large urban models, encouraging its application to higher level 3-d computer vision tasks. Furthermore, local descriptors form the basis of many shape-based applications and their performance has been studied extensively for images-based applications. Local descriptors have also been popular for 3-d shape understanding, but most 3-d descriptors have been used within the context of point clouds data collected with range sensors or polygonal meshes of CAD models. This work presents an investigation of the performance of several local shape descriptors in the PVM, which is learned from image data and where surfaces ambiguities are common and explicitly modeled. Descriptors are evaluated on accuracy for object classification using *Bag-of-Words* models. This evaluation forms a step towards the characterization of a new 3-d probabilistic representation for 3-d scene understanding.

1 Introduction

Understanding the three-dimensional world from images is a long standing goal in computer vision, with numerous applications in the areas of robotics, autonomous navigation, city mapping and surveillance. Multi-view stereo can be thought of as the initial step towards the goal of 3-d scene understanding, and it has been widely studied by the scientific community. Surface estimation for isolated and non-occluded objects has received much attention, and the quality of the reconstruction can be competitive with the accuracy of laser scanners [35]. Few of the available methods can perform image-based 3-d modeling of outdoor, crowded, large scale scenes [5, 10], where accurate modeling is difficult due to severe occlusion, varying illumination conditions, the presence of highly reflective surfaces and sensor errors. In this realm, probabilistic volumetric methods offer a dense representation for the solution of the multi-view stereo problem, modeling explicitly the scene's uncertainty. This new representation of surface geometry [5, 26] has been used in the areas of 2-d change detection [24, 26], and more recently it has shown promising results in the areas of 3-d object classification [29].

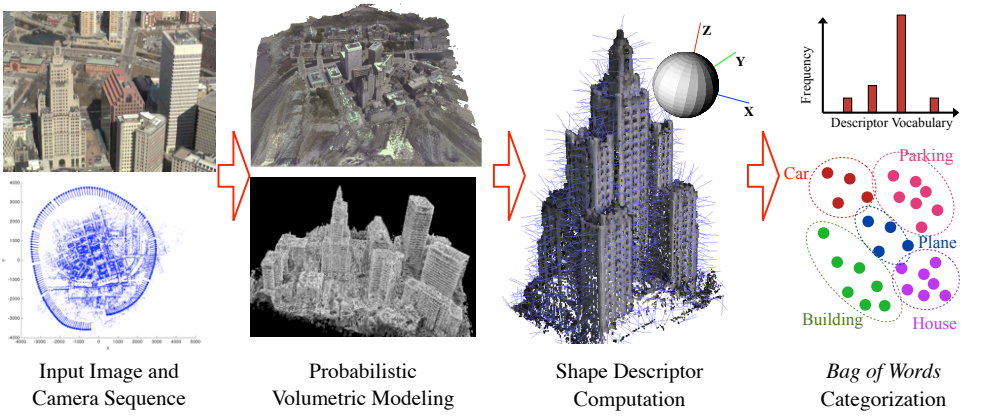


Figure 1: Framework overview. First, an input image and camera sequence are used to learn the PVM. Then, surface normals and shape descriptors are computed at highly likely surface locations. Finally, a *Bag of Words* model is used to learn and classify 5 object categories.

Inspired by the growing number of applications of the probabilistic volumetric model (PVM) to 3-d scene modeling and understanding, this work aims to provide the first evaluation of the performance of several local shape descriptors extracted from the PVM in terms of accuracy for object classification. Descriptors based on local histograms are of particular interest as they are the most popular and most successful for many image indexing applications. While the performance of many 3-d shape descriptors has been studied in point cloud data (from range sensors, or computer generated meshes) it is unclear that their descriptiveness and robustness to noise successfully extends to the diffuse surface probability distributions of the PVM. Surfaces can be poorly localized due to the presence of highly reflective materials, large regions of constant intensity and challenging illumination conditions in the input image data. This work takes a step towards the characterization of the PVM as a new representation for 3-d scene understanding that provides a dense continuous representation of scene geometry despite of all these sources of ambiguity.

This papers evaluates Spin Images (SI) [14], 3-d Shape Contexts (SC) [9, 17] , the SHOT [41] and the FPFH [32] descriptors. For evaluation, this paper proposes a framework that starts by learning the probabilistic models for 17 urban scenes from publicly available aerial imagery [29]. The probabilistic models are learned using an octree-GPU-based implementation [21]. After learning the PVM, surface normals are computed by convolving three-dimensional Gaussian derivative kernels with the volumetric surface probabilities. Surface normals and their locations are used to compute the local 3-d descriptors at highly-probable surface elements. Finally, the descriptors are used to learn *Bag of Words* models for five object categories: planes, cars, houses, buildings and parking lots. During learning and categorization, the objects are manually segmented and labeled using the bounding boxes provided in [29]. The *Bag of Words* framework follows common practices, where a common vocabulary is learned for all categories using k-means clustering. Naive Bayes, SVM and Nearest Neighbors classifiers are compared during the classification stage. To ensure robustness of the evaluation, the results are reported for various vocabulary sizes, descriptor parameters and classifiers across multiple splits of the train/test data. It is emphasized that the key contribution of this work is on the evaluation of various 3-d features and not on novel approaches to classification. A special effort is made to report all parameters used through out this work, such that this evaluation can provide precise guidance to future works using the PVM. An overview of the proposed pipeline is presented in Fig. 1.

2 Prior Work

2.1 3-d Object Description

There exist many different approaches to 3-d shape analysis. For a review, the reader is referred to [40]. In the rigid shape retrieval community, global descriptors have been studied extensively. Global features characterize the overall shape of a 3-d model; examples are: features based on volume and area [46], reflective symmetry descriptors [15], and 3-d Zernike invariants [23] among others. In contrast, local feature-based methods take into account the information in the neighborhood around points on the surface. There are a large number of local descriptors for 3-d shapes, including shape contexts [9, 17], local patches [22], spin images [14], tensors [20], heat kernels [27], 3-d SURF [16], 3-d SIFT [3, 8, 33] and the SHOT descriptor [41]. Although early works in the area generally describe the 3-d surface around a point by encoding in a local coordinate system the geometric properties computed from the local support [4, 38], more recent works have favored histogram based descriptors [9, 14, 41] inspired by the success of descriptors such as HOG [7] and SIFT [19] for 2-d image-based applications. This evaluation focuses on distribution based descriptors as they have shown robustness to clutter, noise, and missing data [9, 11, 16]; characteristics that frequently occur in image-based reconstructions of outdoor urban scenes.

Several evaluations of local shape descriptors have been performed. For example, Heider *et al*. [12] survey and evaluate several descriptors for shape matching tasks under changing mesh quality, noise, and smoothing. The evaluated descriptors belong to two classes: descriptors that locally sample a geometric property, *e.g*., curvature, normals; and descriptors that locally fit a model, *e.g*., polynomials. Shilane and Funkhouser [36] evaluate the distinctive characteristics of local descriptors for similarity retrieval applications. Evaluations of descriptors using *Bag of Words* models have been performed by Bronstein *et al*. [2], where the Heat Kernel Signatures are used as the local shape descriptors. Similarly, Tang and Godil use *Bag of Words* models to evaluate the best performing descriptors reported in [17].

Most of the works mentioned have focused on objects that are acquired in controlled settings using 3-d scanners or CAD models. Fewer works have focused on image-based reconstructions from real life images. Recently, Knopp *et al*. [16] have shown promising results for combined segmentation and recognition. A related study on the PVM is the work in [29], where the authors propose local descriptors based on the PCA analysis and Taylor series expansion of the appearance of local neighborhoods in the PVM. These descriptors are used in *Bag of Words Models* for object categorization tasks and achieve adequate classification rates when using appearance information. However, they significantly under-perform when using occupancy information alone. This paper uses the data-set and manually segmented bounding boxes as provided in [29], to investigate whether robust, distribution-based descriptors can successfully learn the invariant shape information in the PVM without regard to the associated appearance information.

2.2 Probabilistic Volumetric Modeling

Reconstructing 3-d surfaces from 2-d image projections is ill-posed due to the existence of multiple photo-consistent solutions (areas of constant appearance), the presence of sensor noise, camera calibration errors, violations of a surface reflectance model, and occlusions. Pollard and Mundy [26] propose a volumetric Bayesian framework that is robust to large variations in appearance from a video sequence. However, the model is implemented in a regular grid and it does not scale up to large scenes due to its cubic space complexity. Crispell *et al*. [5] address these limitations with a continuous formulation of Pollard and

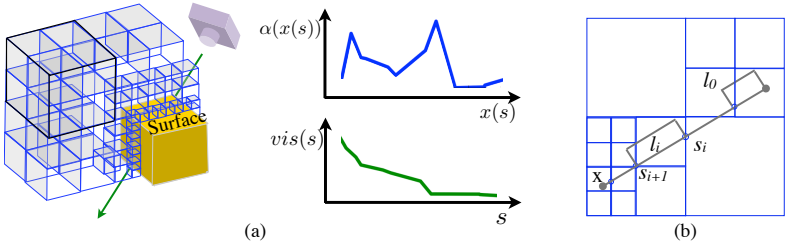


Figure 2: (a) Plots of occlusion density α and *vis* as a ray travels through volume. α peaks (and visibility drops) as the ray pierces the two outside surfaces of the (almost) hollow yellow cube. (b) A parametrization of the ray into intervals according to octree cell intersections.

Mundy's method implemented in an octree. Moreover, a recent GPU implementation [21] is capable of learning large, high resolution volumetric models efficiently where one pass of the online update can be processed at approximately one frame per second for HD resolution.

In Crispell's work, surface probabilities are closely related to a scalar function termed the *occlusion density* $\alpha(x)$ that measures the likelihood that a point occludes points behind it along any line of sight, assuming that the point itself is not occluded. Points along a ray are parametrized by a distance s from q (*e.g.*, camera center) as $x(s) = q + sr \quad s \geq 0$. The visibility probability of a point $x(s)$ is related to the integration of occlusion density as given by the following equation (see [5] for derivation):

$$vis(s) = e^{-\int_0^s \alpha(t)dt}. \tag{1}$$

Intuitively, the visibility along a ray drops significantly when it hits a point of high occlusion, *i.e.*, a surface. An example of this relationship is shown in Fig. 2a.

Learning the occlusion density from images follows an online Bayesian learning algorithm similar to that proposed by Pollard and Mundy [26], which is given by:

$$P(X \in S|I_{N+1}) = P^N(X \in S)\frac{P^N(I_{N+1}|X \in S)}{P^N(I_{N+1})}, \tag{2}$$

where the probability of a voxel being part of a surface, $P(X \in S)$, is updated with the intensity, I, observed in the pixel associated with a corresponding projection ray. $P(X \in S)$ increases if the appearance model at the voxel explains I better that any other voxel along the ray. The appearance at a voxel is modeled with a Gaussian mixture distribution.

The definitions proposed by Crispell *et al.* allow for a generalization of (2), where instead of reasoning about discrete voxels, the update equations account for the exact geometry of ray/voxel intersection, *i.e.*, the length of intersection segment. Fig. 2b illustrates the parametrization of a ray into segment lengths for the octree discretization.

3 Evaluation Framework

3.1 Shape Descriptors

Spin Images: Spin-images [14] are one of the best known 3-d shape descriptors. The neighborhood of a point is described by fitting a tangent plane to the surface at the point and accumulating 2-d histograms of points falling within the support region. The dimension of the descriptor is determined by the number of bins in each side of the spin image. The support radius is determined by the number of bins and the bin size, where bin size affects the

descriptiveness of the spin image. Finally, since spin images describe the relative position of points, they are invariant to rigid transformations.

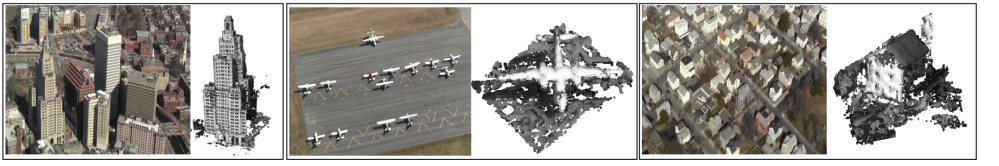
3-d Shape Contexts: Frome *et al.* [9] proposed a 3-d extension of the 2-d shape context [10]. The neighborhood of an oriented point is described by centering a sphere at that point, with the north pole aligned with the surface normal, and dividing the support region into bins with boundaries along the azimuth, elevation and radial dimensions. The subdivisions are logarithmically spaced along the radial dimension and equally along the other two. The dimension of the descriptor is determined by the number of subdivisions. Each bin accumulates a weighted count of points. Finally, there is a degree of freedom in the azimuth direction for 3-d shape contexts, therefore the the local reference frame is not unique.

SHOT: Signatures of Histograms of OrienTations (SHOT) [41] is a recently proposed descriptor that defines an invariant local reference frame to describe the local surface characteristics around a point and uses a spherical grid around a point to accumulate local topological measurements. In the SHOT descriptor, the spherical grid is coarsely divided along the radial, azimuth and elevation dimensions. Each bin contains a local histogram, formed by accumulating point counts into bins according to the cosine of the relative angle between the reference normal and the normals at points within the support of the local bin. The dimension of the descriptor is determined by the number of bins in the local histograms and the number of spatial subdivisions in the spherical grid.

Fast Point Feature Histogram (FPFH): With real-time applications in robotics in mind, the FPFH descriptor [32] is proposed to significantly reduce computation time of the original Point Feature Histogram [31]. In the PFH, a sphere with a given support radius is centered around each point p. For every pair of points (p_i, p_j) and their corresponding surface normals (n_i, n_j), with $(i \neq j)$, within the support sphere of p, a Darboux frame coordinate system at p_i is defined $\{uvw : u = ni, v = (pj - pi) \times u, w = u \times v\}$ and three angular measurements are accumulated: $\beta = v \cdot n_j$, $\phi = (u \cdot v))/||v||$ and $\theta = \arctan(w \cdot n_j, u \cdot n_j)$. During FPFH computation not all neighbors are interconnected, instead only the query point is connected to its neighbors. Finally, the bins of the histogram are decorrelated and concatenated into a low dimension feature vector.

3.2 Bag of Words Model

Bag of Words (BoW) models have their origins in texture recognition [42] and text categorization [13]. Their application to categorization of visual data has been studied extensively [6, 37]. The independence assumptions inherent to bag-of-features representation make learning models for few object categories a simple task, hence its popularity. The method has also been popular for 3-d representations [7, 18, 39]. Bag of words models are simple, efficient, general and intrinsically invariant, and through its use, this evaluation seeks to provide a baseline for future methods to compare against. The implementation of BoW used in this paper follows common practices. Once all features have been extracted (to be explained in subsequent section) and computed, a common vocabulary for all object categories is learned using k-means clustering. After a common vocabulary has been learned, objects are represented by a feature vector corresponding to the vector quantization of its 3-d shape descriptors. Note that only one type of descriptor is used per experiment and they are not intermixed. After the quantization step, the categorization problem is reduced to that of multi-class supervised learning. This paper reports results for three common classifiers, namely nearest neighbor (NN) classifier, multinomial naive Bayes classifier and support Support Vector Machines (SVM).



(a) (b) (c)

Figure 3: Example of the evaluation data-set. An example of a frame and a representative object for three of the categories: (a) building, (b) plane and (c) house. To facilitate visualization, the top 50% of the octree cell locations are shown with the corresponding mean intensity.

4 Implementation Details

Data: For this evaluation 17 probabilistic scenes were used, where a scene corresponds to a site for which a video was collected from a helicopter following a ring-shaped trajectory. All aerial videos were collected near Providence, RI, USA. The data represents a variety of urban and semi-urban sites, including residential neighborhoods, industrial sites, downtown, apartment, hotel and hospital buildings, parking lots, university campus buildings and fields among others. The helicopter flew at an average height of between 300 and 450 meters for all sites, therefore sampling rate is similar for all objects. An approximate resolution of 30 cm/pixel is obtained in the imagery and matched by having the highest resolution voxels span 30 cm on a side in the 3-d models. All camera matrices were computed using the VisualSFM [43, 44, 45] software. To carryout multi-class object learning/classification, this paper uses the bounding boxes and corresponding class labels provided in [29]. A total of 301 objects were labeled corresponding to: 32 planes, 83 cars, 106 residential houses (1-3 story-houses), 44 buildings (large city buildings) and 36 parking lot lanes. Fig. 3 shows an example of an input frame and a representative object for the building, plane and house categories. All data is publicly available, for more details see [28, 29].

Normal Estimation: Surface normals are computed using the gradient of the occlusion density $\nabla\alpha$, which is computed by convolving first order derivatives of the 3-d Gaussian kernel with the volume. Three oriented kernels, *i.e.*, G_x, G_y, G_z, were used and their responses interpolated into an estimate of $\nabla\alpha$. To estimate the direction of the surface normals from the gradient information it is necessary to resolve the ambiguity of inward-pointing vs. outward pointing normal vectors. Surface normals are oriented towards the outside of objects by determining the hemisphere that yields the maximum visibility. A visibility measure $vis\_score(x)$ is computed as the sum of the visibilities of that cell along 12 sample directions from a unit sphere (dodecahedron vertices).

Location Sampling: Cells lying on object surfaces should correspond to areas of high occlusion density $\alpha(x)$ and high visibility $vis\_score(x)$. Therefore, for each object, the shape descriptors are computed for the top 10% of the cells that maximize the following surface measure: $s(x) = \alpha(x) \times vis\_score(x) \times \|\nabla\alpha(x)\|$, where $\|\nabla\alpha(x)\|$ increases robustness to outliers. This filtering criteria avoids computation of the descriptors at locations external or internal to object surfaces.

Parameters of Shape Descriptors: This work uses the open source implementations of the descriptors available at the Point Cloud Library [30]. The most important parameters correspond to the number of bins and the bin sizes. The number of bins was chosen as to best fit the recommended values in the original formulations and the PCL implementation. The number of bins determines the size of the descriptor and has an effect on storage requirements and dimensionality of the classification problem. All parameter values are reported in Table

1. For a fixed number of bins, the support radius of the descriptor determines the bin size and affects descriptiveness of the feature vector. This evaluation reports results for various support radii: 10, 20, 30 and 45 measured in the resolution of the finest cell of the octree.

Table 1: Input parameter for the different shape descriptors.

Descriptor Type	Dimensions	Number of bins	Other
Spin Image	153	width=8	min support cosine = 0.5 min num of neighbors =2
3-d Shape Context	1980	azimuth=12, elevation = 11 radial=15	point density = radius/5 minimal radius =radius/10
SHOT	320	azimuth=8, elevation =2 radial=2, cosθ_i=10	- - - -
FPFH	33	nbins$_\beta = 11$, nbins$_\phi = 11$ nbins$_\theta = 11$	- - - -

K-Means Clustering: The algorithm proposed by Sculley [34] was used as it runs several orders of magnitude faster than the traditional batch algorithm for a large number of samples. The algorithm requires as input parameters the number of mini batches, set to $100 \cdot K$, where $K \in \{20, 50, 100, 200, 500\}$ is the number of clusters. The maximum number of iterations was set to 500, the initialization of the means was random, and the algorithm was run 10 times with a different initialization of the clusters. The solution with the smallest sum of distances between the samples and the centers was chosen as the solution.

Classifiers: For all classifiers, an open source implementation [25] was used. For NN-classifier the weighting scheme of the neighbors was uniform. For naive Bayes all object categories were considered equally likely and to avoid probabilities of zero, Laplace smoothing was used. The SVM was non-linear with the radial basis function (RBF) used as the kernel function. The error weighting constant and RBF parameter were set to 100 and 10 respectively. The parameters were chosen heuristically. Feature vectors were normalized using the L1 norm in the case of the NN and SVM classifiers. This is not necessary in the case of the Naive Bayes classifier, as feature likelihoods correspond to relative frequencies. To achieve multi-class classification with SVM, the one-vs-one approach was used.

Evaluation: For robustness all experiments were conducted over 5 random splits of the data. Classification performance is measured using the *total recall* and *average confusion matrix* across all validation sets.

5 Experiments and Results

Effects of varying the support radius (r_{supp}) and the number of clusters (K): Figure 4 presents the classification performance for the different descriptors as a function of K. For each descriptor and support radius the total recall across all validation sets is reported. All descriptors are classified using SVM. Few observations can be made from Fig. 4. In general, classification performance increases as K is increases. However, not a significant increase was observed after $K = 100$. Regarding the size of the support radius, the effects are different for each descriptor. In the case of the FPFH and the SI descriptors, the performance was almost identical for all tested values. In the case of the SHOT descriptor performance improved from $r_{supp} = 10$ to $r_{supp} = 30$, for $r_{supp} = 45$ the total recall was similar to that of $r = 30$. In the case of SC, the performance increased as the support radius increased, with $r_{supp} = 45$ achieving the best performance. The error bars in the figure refer to the best and the worst performance across the validation trials. Little variability is observed for FPFH

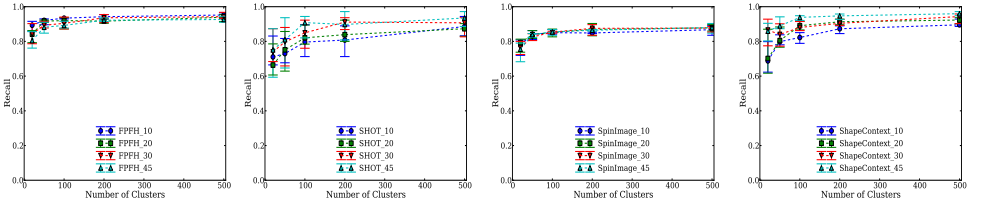


Figure 4: Total recall across all object categories as a function of the number of clusters in the vocabulary. Values 10, 20, 30, and 45 for the support radii are presented. From left to right the figures correspond to the following descriptors: FPFH, SHOT, SI, SC.

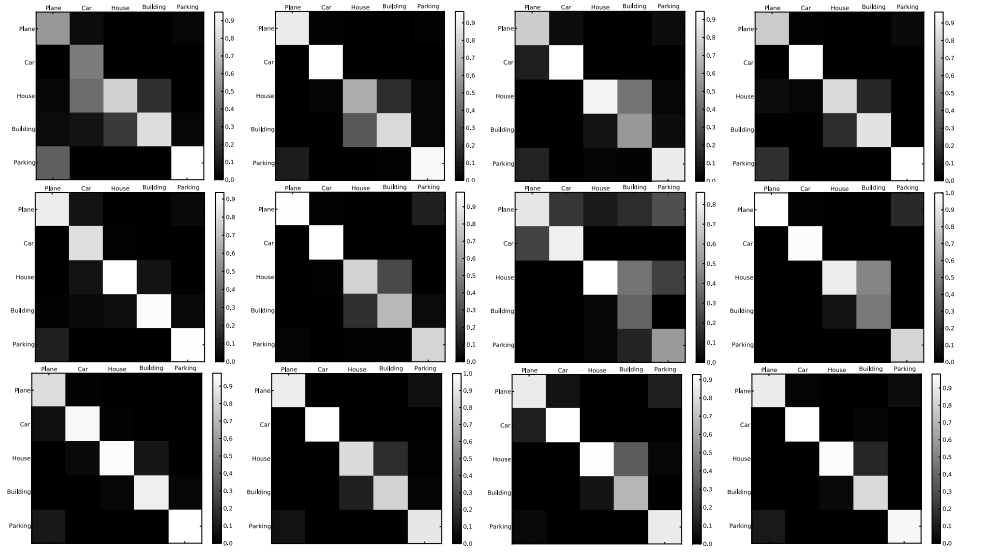


Figure 5: Average confusion matrices (across trials) . Columns (left to right): FPFH, SHOT, SI, SC. Rows (top to bottom): Nearest Neighbor, Bayes, SVM. In all cases $r_{supp} = 30$ and $K = 500$

and SI. The SHOT descriptor reveals the largest variation. In Fig. 4 the FPFH descriptor demonstrates the best performance across different values of K and r_{supp}.

Effects of varying the classifier, evaluating across different categories: Figure 5 presents the confusion matrices for all descriptors (FPFP, SHOT, SI, SC - column wise) and all classifiers (NN, Bayes, SVM - row wise). The number of means is fixed at $K = 500$ and the support radius at $r_{supp} = 30$. The SVM classifier exhibits better performance across all different features. In the case of FPFH, naive Bayes and SVM achieve similar classification accuracy However, NN-classifier exhibits significantly lower performance. The performance of the SHOT descriptor is similar for all classifiers. For Spin Images, naive Bayes has the lowest accuracy, particularly at recognizing buildings. In the case of Shape Contexts, all three classifiers exhibit acceptable performance, but SVM does a better job at differentiating houses from buildings. From the confusion matrices it is possible to notice that in general, houses and buildings are often confused. The SVM classifier appears to better differentiate these two categories.

Figure 6 presents the total recall across validation sets, as well as, the maximum and minimum recall achieved in a trial. In these plots $K = 500$ and $r_{supp} = 30$. It is apparent that the building category is not recognized well when using Spin Images. Naive Bayes does not

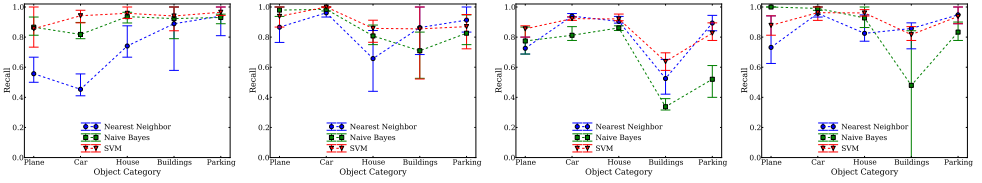
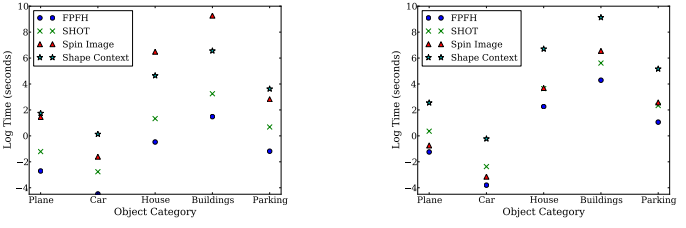


Figure 6: Total recall for all object categories using different classifiers. In all cases, $r_{supp} = 30$ and $K = 500$. From left to right: FPFH, SHOT, SI, SC. Markers correspond to the total recall in all validation trials. Error bars span the maximum and minimum recall attained by a trial.



(a) (b)

Figure 7: Average running time per object for different descriptors. (a) $r_{supp} = 10$, (b) $r_{supp} = 45$.

model the building category very well. In the case of Shape Contexts, naive Bayes classifies all buildings as houses for one of the validation trials. It can also be observed that the Nearest Neighbor classifier performs poorly with the FPFH descriptor. The variability across trials depends on the type of descriptor, the category in question, and the classifiers used.

Storage and time complexity: The probabilistic models used in this work contain high resolution occupancy information. Although only the top 10% of the available locations are sampled, for large objects such as buildings several hundred thousand descriptors are computed. During the experiments, the vast majority of the time was spent computing the shape descriptors. For all objects, the FPFH descriptor was several orders of magnitude faster to compute than the Spin Images and the Shape Contexts. Fig. 7 presents the average time spent computing the descriptors per object. In Fig. 7(a) $r_{supp} = 10$ and in Fig. 7(b) for $r_{supp} = 45$. Time is presented in logarithmic scale as the differences were significant between descriptors. These results are reported for a computer using two 2.93 GHz, Quad-Core Intel Xeon processors. For the FPFH and the SHOT descriptors multi-threaded implementations were available and 8 cores were used. Note that running times of parallel tasks are affected by the availability of cores, and the reported times include these waiting times.

The storage requirement varied significantly across descriptors. Table 1 reports the length of each descriptor. During traditional k-means clustering algorithm, all training descriptors are loaded into memory. For these experiments, approximately 2 million descriptors are used during training. The large dimensionality of 3-d Shape Contexts posed storage challenges.

6 Conclusion and Further Work

This paper presented the first evaluation of local shape descriptors in probabilistic volumetric scenes. The evaluation focused on histogram-based descriptors. FPFH [32], SHOT [41], Spin Images [14] and 3-d Shape Contexts [9] were evaluated for use in object categorization using *Bag of Words* models. Under the different scenarios, the FPFH obtained high recall while having the advantage of being compact and fast to compute. Spin Images underper-

formed when recognizing buildings. The SVM classifier was the more effective classifier. The results indicate that distribution-based descriptors effectively extract salient characteristics of the shape information in the PVM for object categorization. This work provides guidance on the selection of descriptors and parameters for characterization of the PVM, making a fundamental step on the understanding of the shape information in the PVM. However, it is recognized that the presented evaluation is not exhaustive and it would be interesting to include more descriptors and more object categories. It would also be informative to evaluate the effects of dimensionality compression and sampling on classification performance. Finally, this evaluation can be extended to other vision tasks in the PVM, such as registration.

References

[1] S Belongie, J Malik, and J Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 2002.

[2] Alexander M Bronstein, Michael M Broinstein, Leonidas J Guibas, and Maks Ovsjanikov. Shape Google: Geometric Words and Expressions for Invariant Shape Retrieval. *ACM Transactions on Graphics*, 2011.

[3] W Cheung and G Hamarneh. n-SIFT: n-Dimensional Scale Invariant Feature Transform. *IEEE Transactions on Image Processing*, 2009.

[4] Chin Seng Chua and Ray Jarvis. Point Signatures: A New Representation for 3D Object Recognition. *Int. J. Comput. Vision*, 1997.

[5] D Crispell, J Mundy, and G Taubin. A Variable-Resolution Probabilistic Three-Dimensional Model for Change Detection. *Geoscience and Remote Sensing, IEEE Transactions on*, 2011.

[6] Gabriella Csurka, Christopher R Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[7] N Dalal and B Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] G Flitton, TP Breckon, and N Megherbi. Object Recognition using 3D SIFT in Complex CT Volumes. In *BMVC*, 2010.

[9] Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik. Recognizing Objects in Range Data Using Regional Point Descriptors. In *ECCV*, 2004.

[10] Y Furukawa and J Ponce. Accurate, Dense, and Robust Multiview Stereopsis. *PAMI*, 2010.

[11] A Golovinskiy, V.G Kim, and T Funkhouser. Shape-based recognition of 3D point clouds in urban environments. In *ICCV*, 2009.

[12] Paul Heider, Alain Pierre-Pierre, Ruosi Li, and Cindy Grimm. Local shape descriptors, a survey and evaluation. *Eurographics Workshop on 3D Object Retrieval*, 2011.

[13] Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *ICML*, 1997.

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

[15] M Kazhdan, B Chazelle, D Dobkin, and T Funkhouser. A Reflective Symmetry Descriptor for 3D Models. *Algorithmica*, 2003.

[16] 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.

[17] M Körtgen, GJ Park, and M Novotni. 3D Shape Matching with 3D Shape Contexts. In *7th Central European Seminar on Computer Graphics*, 2003.

[18] X. Li and A Godil. Exploring the Bag-of-Words method for 3D shape retrieval. *International Conference on Image Processing*, 2009.

[19] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 2004.

[20] A S Mian, M Bennamoun, and R Owens. Three-Dimensional Model-Based Object Recognition and Segmentation in Cluttered Scenes. *PAMI*, 2006.

[21] Andrew Miller, Vishal Jain, and Joseph Mundy. Real-time Rendering and Dynamic Updating

of 3-d Volumetric Data. In *Workshop on General Purpose Processing on Graphics Processing Units*, 2011.

[22] Niloy J. Mitra, Leonidas Guibas, Joachim Giesen, and Mark Pauly. Probabilistic fingerprints for shapes. In *Eurographics Symposium on Geometry Processing*, 2006.

[23] Marcin Novotni and Reinhard Klein. 3D Zernike Descriptors for Content Based Shape Retrieval. In *ACM symposium on Solid modeling and applications*, 2003.

[24] O.C Özcanli and J.L Mundy. Vehicle Recognition as Changes in Satellite Imagery. In *International Conference on Pattern Recognition*, 2010.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011.

[26] T Pollard and J.L Mundy. Change Detection in a 3-d World. In *CVPR*, 2007.

[27] Dan Raviv, Michael M Bronstein, Alexander M Bronstein, and Ron Kimmel. Volumetric heat kernel signatures. In *ACM workshop on 3D object retrieval*, 2010.

[28] Maria I. Restrepo. Providence Aerial Multi-view, 2011. URL http://vision.lems.brown.edu/datasets/aerial-multiview.

[29] Restrepo, M.I, Mayer, B.A, Ulusoy, A.O. and Mundy, J.L. Characterization of 3-d Volumetric Scenes for Object Recognition. *IEEE Journal of Selected Topics in Signal Processing.*, 2012.

[30] R.B Rusu and S Cousins. 3D is here: Point Cloud Library (PCL). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011.

[31] R.B Rusu, Z.C Marton, N Blodow, and M Dolha. Towards 3D Point cloud based object maps for household environments. *Robotics and Autonomous Systems Journal (Special Issue on Semantic Knowledge)*, 2008.

[32] R.B Rusu, N Blodow, and M Beetz. Fast Point Feature Histograms (FPFH) for 3D Registration. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[33] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-Dimensional SIFT Descriptor and its Application to Action Recognition . In *15th International Conference on Multimedia*, 2007.

[34] D Sculley. Web-Scale K-Means Clustering. In *WWW '10: Proceedings of the 19th international conference on World wide web*, 2010.

[35] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In *CVPR*, 2006.

[36] P Shilane and T Funkhouser. Selecting Distinctive 3D Shape Descriptors for Similarity Retrieval. In *IEEE International Conference on Shape Modeling and Applications*, 2006.

[37] J Sivic, B.C Russell, A.A Efros, A Zisserman, and W.T Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[38] Y. Sun and M.A. Abidi. Surface matching by 3D point's fingerprint. In *ICCV*, 2001.

[39] Sarah Tang and Afzal Godil. An evaluation of local shape descriptors for 3D shape retrieval. In *Proceedings of SPIE*, 2012.

[40] J.W.H Tangelder and R.C Veltkamp. A survey of content based 3D shape retrieval methods. In *Shape Modeling Applications*, 2004.

[41] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *ECCV*, 2010.

[42] M Varma and A Zisserman. A Statistical Approach to Material Classification Using Image Patch Exemplars. *PAMI*, 2009.

[43] Changchang Wu. SiftGPU: A GPU implementation of Scale Invariant Feature Transform (SIFT), 2007. URL http://cs.unc.edu/~ccwu/siftgpu.

[44] Changchang Wu. VisualSFM: a Visual Structure From Motion System, 2011. URL http://www.cs.washington.edu/homes/ccwu/vsfm/.

[45] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M Seitz. Multicore bundle adjustment. In *CVPR*, 2011.

[46] Cha Zhang and Tsuhan Chen. Efficient feature extraction for 2D/3D objects in mesh representation. In *International Conference on Image Processing*, 2001.