Feature Extraction from Depth Maps for Object Recognition

Caleb Jordan
Department of Computer Science
Stanford University
Stanford, California 94305
Email: grnstrnd@stanford.edu

***Abstract*—With widespread availability of depth sensors, recognition systems have an exciting new source of information to help bridge the gap between 2D and 3D. This work attempts to more fully explore possible 3D and multimodal features derived exclusively from depth data for the purpose of designing more descriptive images features. In object recognition systems, depth data provides otherwise absent geometric data, a potentially powerful tool for discrimination between objects. Depth images contains local information about object geometry but also provide information on global geometry, object position, and object shape. Traditional feature-extraction techniques, typically per-pixel or window operations, use local information but lose the global information available. This paper attempts to engineer a new, generalizable class of depth features based on shape distributions and uses a simple category recognition framework to evaluate their performance in contrast to more traditional approaches. The results suggest that the discriminative power of shape distributions in the context of depth maps is in fact quite limited; however, the alternative histogram of normals feature in fact out-performs standard HOG, demonstrating the importance of depth data.**

I. Introduction

Image classification is a task that is nearly omnipresent in modern computer science. It is applied in various forms in facial recognition, robotics, medical research, image search engines. It is one of the fundamental building blocks in the field of computer vision. The power of any image classification system ultimately derives from the discriminative power of its features–their ability to glean information from the images provided to them. Over several decades, we have developed strong features based on image data that allow us to find and distinguish objects and categories well in images. However, a new type of image is now widely available: depth images. This new channel has the potential to overcome some of the weaknesses of standard color images if we can harness its information.

Currently, depth data is used in a variety of different image recognition systems, often in parallel with RGB data and standard features. One common class of feature derived from depth images includes local information such as curvature and normals. At the other end of the spectrum lie techniques that compute a point cloud from a depth map and then match it with clouds in a database, a much more extensive process. Exploring the space in between these two options could combine the strengths of both: the speed and simplicity of simple feature vectors with the more global analysis that point cloud registration provides.

This paper does not propose a novel object recognition pipeline–admittedly, models that combine RGB and depth features will always have more potential than depth alone. Rather, this paper's purpose is to discuss the relative power of various depth-based features in a controlled environment and shed light on promising directions of study regarding depth-based features.

II. Related Work

There exists a variety of work in recognizing objects in 3D as well as depth maps. Common features generated from depth data include curvature estimates and object scale heuristics. Tangentially, point cloud registration has been explored for decades both to match images, commonly faces, as well as in environment mapping.

In 2002, Osada et al [3] explored shape distributions for object recognition. This method is typically applied to complete 3D models rather than depth maps, while this paper experiments with them on the partial point clouds derived from depth data. The concept of shape distributions is a much more general one; whereas this paper generates distributions over pairwise Euclidean distances, other formulations compute distributions of angles between triples, volumes of sampled tetrahedrons, and much more.

Lai, Bo, et al [1] presented the RGB-D dataset in 2011. They also proposed a few basic depth-based features and did their own recognition experiments using color data, depth, and a combination. This paper primarily makes use of the dataset. At the same time, they proposed the Hierarchical Kernel Descriptors method, suggesting kernels based on depth information, including local shape [2].

A wider variety of scale and rotation-invariant geometric features have been studied. Of particular inspiration to this paper is Drost's work with multimodal point pair features [4], which the feature variations in this work hope to expand upon. While traditionally, systems that use both RGB and depth modalities do so by concatenating independent feature vectors, Drost combines RGB and depth data in new ways, using RGB edges to inform the extraction of depth features. This is applied directly in the edge shape distribution in this paper; however,

I feel that this concept deserves much more exploration than can be done in this paper.

Finally, spin images are another object recognition method that matches instances in a scene to original object models [5]. This is another thoughtful method that makes use of more global geometric information; however, like many shape distribution applications, it presupposes the existence of a library of 3D object models, something that is not always feasible.

III. Data and Features

A. Dataset

The RGB-D dataset includes 51 different object categories with 300 household objects and video segments for each instance. Each frame includes and RGB image, a depth image (stored in meters), and a ground-truth segmentation of the pixels belonging to the object. See columns 1-3 in figure 1 as examples of images provided in the dataset. The features presented here are derived from the depth image, often projected into 3D points using camera parameters. One feature uses the RGB image to mask the point cloud similar to [4] but does not actually incorporate numerical color data.

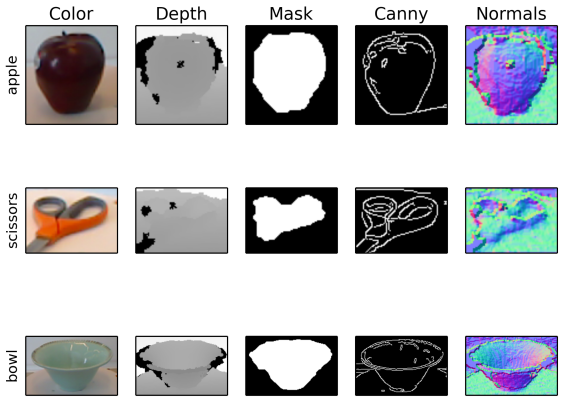

Fig. 1: Images provided in RGB-D dataset along with examples of Canny Edge segmentation and the a computed normals map.

B. Features

Two baseline features were used for comparison: standard HOG over the calculated RGB grayscale image itself and HOG using the depth map as a grayscale image. Both of these were intended as baselines, and HOG over the RGB grayscale is the only feature in this paper that directly incorporates numerical color data.

The first alternative feature is histogram of normals. It is intended to be similar to HOG and is a local geometric feature computed via sliding window. The normals themselves are estimated efficiently from the point cloud using a Sobel filter to calculate directional tangents and then crossing the tangent and bitangent. Column 5 in figure 1 shows an image representation of computed normal maps. The full feature vector is computed using a sliding window and binning the normals in each window. The histogram is a 2D histogram in spherical coordinates; all normals have length 1.0, so they can be binned according to elevation and azimuth.

The second class of features is based on shape distribution and pair point features. Each feature is derived by sampling pairs of points from the point cloud, then generating a histogram over different attributes of this sample set. The two attributes these features use are Euclidean distance and absolute distance in the z-direction (depth distance). I generated three different variations of this feature, each distinguished by the points it samples. The vanilla version collects pairs from all points for which depth data exists–this limitation is also enforced for the other two. The second version, masked shape distribution, collects pairs only from points belonging to the object, defined by some segmentation of the image. Admittedly, access to an accurate segmentation is impractical; a practical application of this method might be to actually score proposed image segmentations via distance from a learned ideal distribution. The third version first runs Canny edge detection on the RGB image, then samples from 3D points corresponding to edges. In particular, this idea came from [4], and it is an example of a hybrid feature that uses RGB to inform depth analysis. See figure 2 for examples of all three distributions.

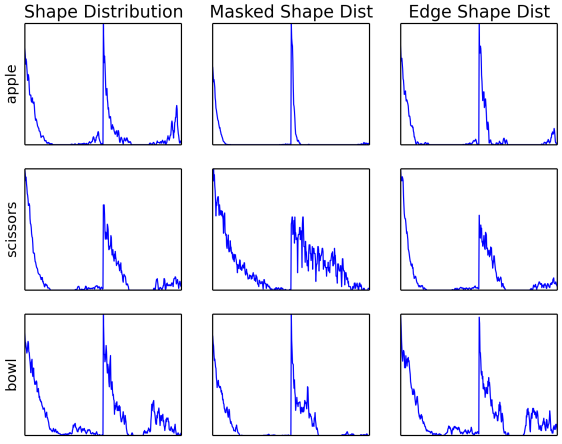

Fig. 2: Examples of all three shape distribution histograms.

IV. Experiments and Results

A. Technical Details

Each of the features has a number of different parameters to tune, and this tuning was done iteratively using scores on a 5-class subset of the data. As an attempt to normalize the expressive power of the features, parameters are chosen so that all feature vectors have on the order of 250-400 components.

Concatenated features are the exception to this, resulting in vectors around twice as long for the two concatenated features tested. For features that use sliding-windows, all images are resized to be of size 128x128 before feature extraction, and coarse sliding windows of 64x64 pixels are used.

HOG and depth image HOG are each 324 long. Longer feature vectors had only slightly more success, possibly due to limited training size. The normal histogram feature is the shortest with 256 components. Normals are binned into the 4x4 space over possible angle of orientation. As an side note from tuning, making either HOG feature finer–increasing the feature length by order of magnitude–makes it less competitive with histogram of normals, which performs as well even with fewer dimensions.

All of the shape distribution features are parametrized the same. Each takes 1000 samples (each a pair of points), calculates one histogram of 200 bins over Euclidean distance between the two points, another over absolute z-distance, and concatenates the two histograms. Each histogram is smoothed with a 1d, length 3 box filter to reduce sampling noise, reducing in a total length of 396. Of note is the effect of normalizing the histogram: probability distributions (sum is one) actually perform very poorly, whereas probability density histograms (integral is one) perform very well, possibly because they encode scale information very powerfully as they are normalized by bin width.

B. Experiment Setup

A simple category recognition task was set up to measure the descriptive power of the features relative to each other. I used OpenCV and Numpy in Python for most of the image processing, and used the OpenCV multi-class, linear SVM for the learning task.

The final training curves were computed using a full set of 5000 items, sampled from all 51 categories of the RGB-D dataset. A random test set of 1000 items was held out, and results were averaged over 10 independent runs.

C. Feature Performance

Of all features, only the histogram of normals performs as well as standard HOG for category recognition. This is significant in itself–it suggests that category recognition using only depth data could be competitive with color data. Depth image HOG, however, has about 15% more error than either, which is slightly unexpected in that the normals themselves are derived from the depth gradients.

None of the shape distribution features perform well. Of all, the basic variant performs the best. This is also surprising; I expected the two masked variants to outperform the standard as they only sample what are seemingly important points. The most probable cause is that both masks can be quite noisy–not all images are segmented correctly, and in either case, misclassifying even a couple of pixels can dramatically change the shape of the histogram. Experimentation with the histogram generation could shed light on or improve this result; using histograms of constant range presents different challenges but reduces the effect of outliers. Missing contextual information is not likely the cause for this discrepancy because the background for all images was the same.

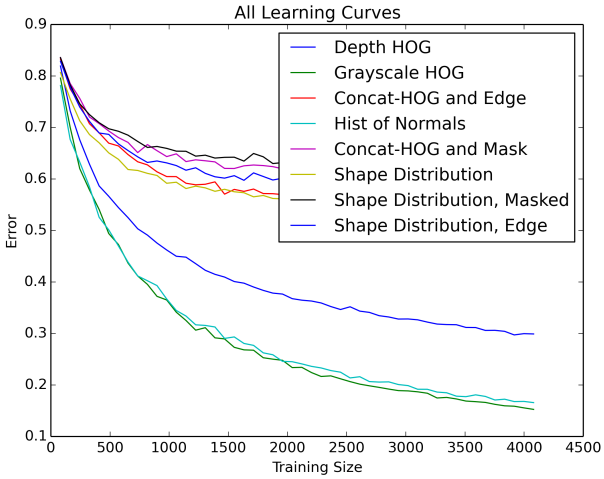

Fig. 3: Learning curves for all feature variants

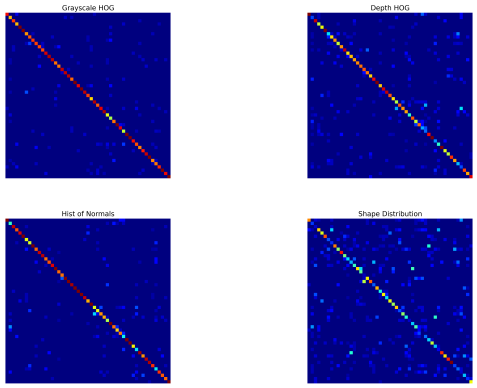

Fig. 4: Confusion matrices for both baselines, histogram of normals, and shape distribution feature over 51 categories

V. Conclusion

This work developed a framework for directly comparing the descriptive power of various features–a feature lab of sorts. Of the depth-based features explored in this framework, only histogram of normals matches the power of traditional HOG for category recognition. The concept of shape distribution, at least in its current formulation, clearly suffers when objects are occluded and therefore seems to be nearly powerless to distinguish objects accurately.

There remains much exploration to do in this area. While shape distribution itself fails in this particular task, the derived distributions have several strengths, including simplicity, generalizability, and natural scale and rotation invariance. Histogram of normals, on the other hand, is somewhat surprising in its descriptive power–in some circumstances, it outstrips traditional HOG, suggesting that a more thorough exploration

of its properties could reveal much for recognition.

An extension to this project would use these features in the context of a more elaborate model–certainly the SVM is a fairly simple model for this kind of task. It's possible that other models may better describe the relationship between these different features and make better use of shape distributions.

ACKNOWLEDGMENT

I would like to thank Roland Angst for his advice, encouragement, and time. Also, Silvio Saverese for helping to refine the project idea and for ideas on how to extend this work.

REFERENCES

[1] K. Lai, L. Bo, X. Ren, and D. Fox. "A Large-Scale Hierarchical Multi-View RGB-D Object Dataset". IEEE International Conference on on Robotics and Automation, 2011.

[2] L. Bo, K. Lai, Xiaofeng Ren, and D. Fox "Object Recognition with Hierarchical Kernel Descriptors".

[3] Osada, Robert, et al. "Matching 3D Models with Shape Distributions."

[4] B. Drost, S. Ilic. "3D Object Detection and Localization using Multimodal Point Pair Features"

[5] A. Johnson, M. Hebert. "Using Spin-Images for Efficient Object Recognition in Cluttered 3-D Scenes". IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998.

[6] Asari, M.A.; Supriyanto, E.; Sheikh, U.U. "The Evaluation of Shape Distribution for Object Recognition Based on Kinect-Like Depth Image", Computational Intelligence, Communication Systems and Networks (CICSyN), 2012 Fourth International Conference on, On page(s): 313 - 318

[7] G.G. Gordon, "Face Recognition Based on Depth and Curvature Features", Proc. CVPR'92, pp.108-110, 1992.

[8] Chenghua Xu; Yunhong Wang; Tieniu Tan; Long Quan "Automatic 3D face recognition combining global geometric features with local shape variation information", Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on, On page(s): 308 - 313.

[9] J. Cook , V. Chandran , S. Sridharan and C. Fookes "Face Recognition from 3D Data Using Iterative Closest Point Algorithm and Gaussian Mixture Models", Proc. IntÍ Symp. 3D Data Processing, Visualization, and Transmission, pp.502 -509 2004.