Using Depth Information to Improve Object Recognition with Deep Learning

Awni Altabaa[1], Eduard Varshavsky[2], Jada Buchanan[3], Ngoc Bao Han “Mimi” Nguyen[4]

QMIND – Queen’s AI Hub
Queen’s University, Kingston, Ontario K7L 3N6, Canada.

1 e-mail: awni.altabaa@queensu.ca
2 e-mail: 18ev@queensu.ca
3 e-mail: 16jmb7@queensu.ca
4 e-mail: 19bhnn@queensu.ca

Abstract: *Over the past decade, deep learning has driven great progress in computer vision and 2D image understanding. On the other hand, 3-Dimensional image understanding is still comparatively immature. In recent years, RGB-D cameras combining visual and 3D shape information have become more accessible, enabling progress to be made in the field. In this paper, we test the hypothesis that RGB-D object recognition models can improve on state-of-the-art RGB models and propose a deep learning architecture that leverages the added depth information. Our proposed architecture encodes depth images into a geocentric embedding and makes use of two independent processing streams for the RGB and depth images. We train multiple models to control for and validate the effects of the added depth information. Our best model achieved an accuracy of 70.1% on a dataset of 51 classes.*

1. INTRODUCTION

1.1 Motivation

The extraction of a high-level understanding of three-dimensional (3D) images is a fundamental problem in the field of computer vision. 3D image understanding has applications in areas including remote sensing, mapping, monitoring, autonomous-driving, virtual/augmented reality, and robotics [1]. Thus, models which can autonomously extract high-level information (such as recognizing objects) from 3D images are in high demand.

In past decades, similar to 2D computer vision, research on 3D computer vision often employed classic machine learning methods like Support Vector Machines and Random Forests. [2]. However, with the increase of computational power and availability of data, deep learning has allowed for rapid development in both 2D and 3D computer vision [3].

Our focus is on 3D sensed data in the form of so called RGB-D images. The format consists of a pair of images; a standard RGB image and a depth image. Depth images provide additional information about the 3D structure of the scene, and unlike RGB images, are invariant to lighting and are particularly useful in background separation [4].

1.2 Related Works

One approach to the problem is to simply stack the RGB and depth images generating a 4-channel image and employ existing Convolutional Neural Network (CNN) architectures. However, Gupta et. al. [5] found that this approach does not make the most use of the geometric information encoded in the depth image.

In one of the earlier papers on the subject, Socher et. al. [6] proposed a CNN-RNN architecture in which CNNs extract low-level translation-invariant features for RGB and depth images independently, then RNNs generate high-level global features. Eitel et. al. [7] used a similar approach in which two different CNNs process the RGB and depth images independently, then fuse the output after passing it through two fully connected layers.

1.3 Problem Definition

In this paper, we aim to build an RGB-D deep learning object recognition model which makes good use of the depth information and outperforms RGB-only models. As input, the model takes an RGB-Depth image pair, and outputs a prediction of the class of the object present in the image. Through this process, we aim to develop an understanding of 3-dimensional images in the context of deep learning and gain insights that pave the way for further improvement in future research.

2. METHODOLOGY

2.1 Dataset

We used the Washington University RGB-D dataset containing images of 300 common household objects organized into 51 categories. The dataset was collected using a Kinect-style 3D camera. The dataset is 84 GB large and contains 207,920 RGB-D images [8].

2.2 HHA Geocentric Encoding of Depth Information

In 2014, Gupta et. al [5] proposed a geocentric embedding of depth images which transforms single-channel depth images into a 3-channel representation. This representation encodes horizontal disparity, height above ground, and angle with gravity for each pixel (referred to as the HHA embedding). In their paper, they demonstrated that extraction of features from HHA images using CNNs learned stronger representations and achieved higher performance than raw depth images.

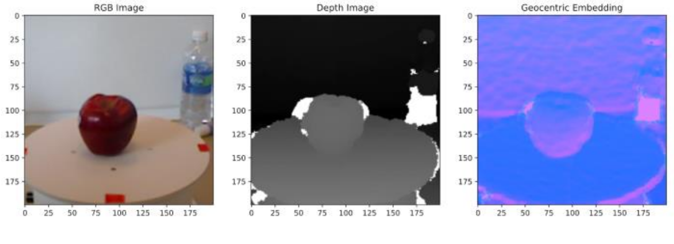

Figure 1: A sample image from the dataset

2.3 Proposed Architecture

We proposed and validated a deep learning architecture based on two independent CNN processing streams for the RGB and depth images, respectively. In the RGB processing stream, we make use of the proven ResNet50 2D CNN model to generate an RGB feature vector [9]. The depth image is transformed to the HHA geocentric encoding, then passed through a CNN feature extractor that we built from scratch. The RGB and depth feature vectors are then fused and passed through two fully connected layers before generating a class prediction.

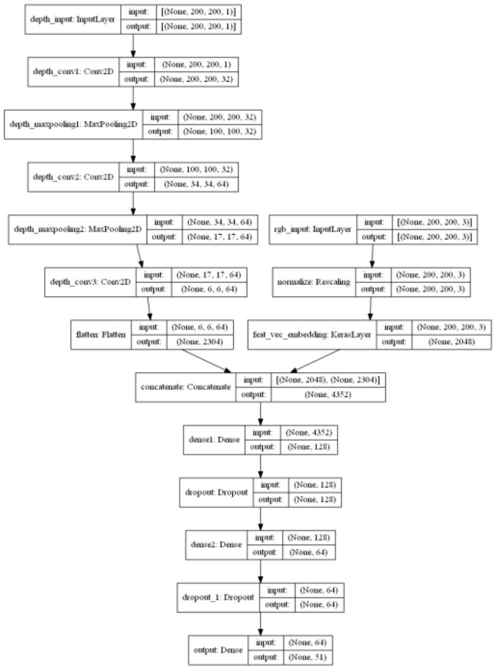
depth_input: InputLayer
input: [(None, 200, 200, 1)]
output: [(None, 200, 200, 1)]
depth_conv1: Conv2D
input: (None, 200, 200, 1)
output: (None, 200, 200, 32)
depth_maxpooling1: MaxPooling2D
input: (None, 200, 200, 32)
output: (None, 100, 100, 32)
depth_conv2: Conv2D
input: (None, 100, 100, 32)
output: (None, 34, 34, 64)
depth_maxpooling2: MaxPooling2D
input: (None, 34, 34, 64)
output: (None, 17, 17, 64)
rgb_input: InputLayer
input: [(None, 200, 200, 3)]
output: [(None, 200, 200, 3)]
depth_conv3: Conv2D
input: (None, 17, 17, 64)
output: (None, 6, 6, 64)
normalize: Rescaling
input: (None, 200, 200, 3)
output: (None, 200, 200, 3)
flatten: Flatten
input: (None, 6, 6, 64)
output: (None, 2304)
feat_vec_embedding: KerasLayer
input: (None, 200, 200, 3)
output: (None, 2048)
concatenate: Concatenate
input: [(None, 2048), (None, 2304)]
output: (None, 4352)
dense1: Dense
input: (None, 4352)
output: (None, 128)
dropout: Dropout
input: (None, 128)
output: (None, 128)
dense2: Dense
input: (None, 128)
output: (None, 64)
dropout_1: Dropout
input: (None, 64)
output: (None, 64)
output: Dense
input: (None, 64)
output: (None, 51)

Figure 2: The architecture of RGB/Raw depth model

2.4 Evaluation

We trained four different models using different combinations of RGB, raw depth, and HHA to isolate for the impact of each. Namely, we built and trained depth-only, HHA-only, RGB-Depth, and RGB-HHA models on our dataset.

We split the dataset into a training and test set by dedicating one object instance to the test set and leaving the rest for training. This was done to prevent data leakage since adjacent frames of the same object instance will look very similar.

3. RESULTS AND DISCUSSION

First, our results demonstrate that a deep learning model can achieve reasonable accuracy in recognizing objects in depth-only images. Our raw depth model achieved a test accuracy of 40.5%. Thus, depth images have utility beyond just providing auxiliary information to RGB images.

Table 1: Model performances

Model	Accuracy
Depth-Only	40.5%
HHA-only	48.0%
RGB w/ Raw Depth	54.7%
RGB w/ HHA	70.1%

Second, our results confirm that the HHA geocentric representation of depth images improves performance in deep learning object recognition models. The HHA-only model achieved notably higher accuracy than the raw depth model at 48.0%. The effect is even larger in the combined RGB and depth models. While RGB with raw depth achieved 54.7% accuracy, RGB with HHA achieved 70.1%. Thus, representing depth information in the HHA embedding improves performance for this model architecture.

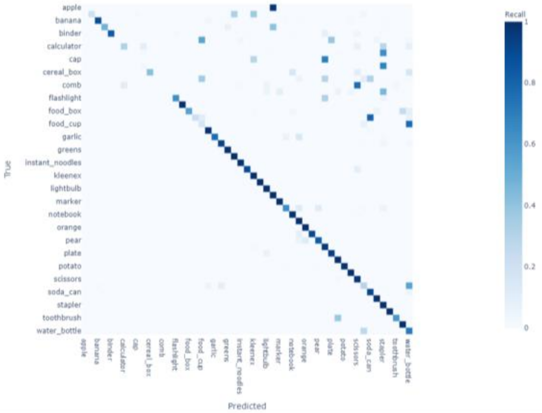

Figure 3: Confusion matrix of the RGB-HHA model

Finally, we examined the learned weights of the RGB-Depth models to further assess how useful the depth information was to the models. We did this by looking at the weights of the dense layer following the concatenated feature vector, and compared the weights associated with the RGB feature vector and the depth feature vector. We found the magnitudes of weights associated with depth information to be comparable to that of RGB information, especially when encoded in HHA.

Table 2: Statistics of learned weights

Statistic	*RGB w/ Raw Depth*		*RGB w/ HHA*	
	RGB	Raw Depth	RGB	HHA
Mean Abs. Value	0.0449	0.0243	0.0233	0.0234
Max	0.8412	0.1694	0.0467	0.0467
Min	-0.9376	-0.1694	-0.0467	-0.0467
Std. Dev.	0.0712	0.0286	0.0269	0.0270

4. CONCLUSIONS AND FUTURE WORK

With the utility of depth information in object recognition models demonstrated, future research may be directed at optimizing performance further and pursuing the maximum possible accuracy. Our architecture makes use of two independent processing streams for the RGB and depth images; perhaps an architecture where information from the opposite stream is allowed to gradually seep in before the final fusion would result in a model that is more aware of the "full picture." Another promising application of depth information in computer vision is object segmentation, where 3-dimensional structure is especially important. This would be an interesting avenue for further research.

5. REFERENCES

[1] S. Zia, B. Yuksel, D. Yuret and Y. Yemez, "RGB-D Object Recognition Using Deep Convolutional Neural Networks," *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[2] D. Griffiths and J. Boehm, "A Review on Deep Learning Techniques for 3D Sensed Data Classification," *Remote Sensing*, vol. 11, no. 12, p. 1499, 2019.

[3] Y. Gao, F. Sohel, M. Bennamoun, M. Lu and J. Wan, "Rotational Projection Statistics for 3D Local Surface Description and Object Recognition," *International Journal of Computer Vision*, vol. 105, no. 1, pp. 63-86, 2013.

[4] Y. LeCun, "Deep learning & convolutional networks," *IEEE Hot Chips 27 Symposium (HCS)*, 2015.

[5] S. Gupta, R. Girshick, P. Arbeláez and J. Malik, "Learning Rich Features from RGB-D Images for Object Detection and Segmentation," *Computer Vision – ECCV 2014*, pp. 345-360, 2014.

[6] R. Socher, B. Huval, B. Bhat, C. D. Manning and A. Y. Ng, "Convolutional-Recursive Deep Learning for 3D Object Classification," *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012.

[7] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[8] K. Lai, L. Bo, X. Ren and D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," *IEEE International Conference on Robotics and Automation*, 2011.

[9] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly and N. Houlsby, "Big Transfer (BiT): General Visual Representation Learning," 2020.

Our work can be found at:
https://github.com/Awni00/3d-object-classification