Stereo image generation using Neural Networks

Siddhant Prakash
1211092724
Email: sprakas9@asu.edu

Anurag Solanki
1211084183
Email: asolank1@asu.edu

***Abstract*—Using stereo image to generate 3D views is a challenging problem to solve, especially due to increase in demand for 3D content with the advent of virtual reality systems like Microsoft Hololens and Oculus Rift. With big companies such as Google, Facebook, Amazon and a lot of start-ups investing heavily in virtual and augmented reality system, this demand is set to increase exponentially with time. In our project, we will try to train a neural network to come up with a model to estimate the stereo pair given a single RGB image for 3D scene reconstruction.**

1. Problem Statement

Finally, 3D images and videos are in the mainstream media after being ignored for so long. Since the advent of multi-view geometry, the potentials of stereo image pair in various computer vision problems, such as, depth estimation, object recognition, segmentation, simultaneous localization and mapping (SLAM) etc., has been exploited comprehensively. Stereo image pairs are essentially images of the same scene from two different view. The image pair differ from each other by a projective transformation. Stereo images has been studied extensively in 2-view geometry and geometrical constraints have been established between the pairs which can be exploited to efficiently generate depth map of the scene given the two views. The depth map along with the stereo pairs are all that is need to project a 3D scene on a display. The use of these stereo pairs for scene understanding has been the motivation behind such varied applications. Thus, the importance of having the stereo pair of an image increases manifold.

Given an image of a scene, can we generate its stereo pair image by training a neural networks end-to-end, is what we want to explore through our project. This can very well lead us to a way of estimating a better depth map than with previous methods. Thus the problem becomes that one of estimating a depth map by generating a stereo pair, rather than the traditional other way round as has been since so long.

2. Relevant Work

The problem of obtaining stereo pair from a a single image directly has not been much explored in the academia. Stereo image pairs have been used to deal with a number of computer vision problems. In depth estimation, number of algorithms [9] [10] [11] have been developed to optimally utilize the 3D scene information captured by the pairs. Stereo matching has been explored in many conventional computer vision algorithms [12] [13] problem which tries to estimate the cost of matching stereo image pairs. These algorithms act as the first stage of many stereo algorithms. Recently, neural network has been employed to compare these patches [14] in a fast as well as efficient manner, studying the trade offs between the two.

Although in past years, 3D view generation from single image gained momentum using learning based methods like Im2depth [19] and Make3D [20], which employed an MRF based algorithm to capture the 3D location and orientation of patches in an image. Eigen et. al. [21] was one of the first neural network based method to deal with depth estimation from single image which employed two deep network stacks for prediction. Recently published, competing directly with the work we are trying to do is Deep3D [15], which addresses the problem of generating stereo pairs using single image. Although, the accuracy achieved by the work is best, we would like to explore different network architecture mentioned in Section 3.2, to better the accuracy. Also, due to the availability of new and larger datasets like ScanNet, we are optimistic about our work.

3. Expected Experiments

3.1. Dataset

The first and foremost requirement to learn good models using deep neural networks are their hunger for large datasets. Until recently, not many large datasets were available for the problem of RGB-D scene understanding. But now with the advent of datasets like KITTI [1] and Middleburry [2] stereo datasets, we can delve into the domain of learning models for 3D scene understanding. While the KITTI dataset has 400 dynamic scenes, along with Middleburry dataset, they do not contribute to more than a few thousand frames. Another recent addition to this class of dataset is the ScanNet [3], which provides us with 2.5M views, which should mostly satisfy our requirement. Although, in [15] the dataset used were the 3D movies downloaded from internet, which provided about 5M views.

We will try to obtain the movie dataset although the ScanNet data should be good enough for our purpose.

For initial exploratory tasks, we also plan to simulate our own 3D dynamic scene using Matlab and capture the scene from two viewpoints corresponding to stereo pair of images.

3.2. Network model

We plan to exploit large datasets and use them to learn a model to generate stereo image pairs given an image. We can approach the problem in two ways. Given one image from the stereo pairs, we can use one (say left) two generate the other stereo pair (right). In another way, we may come up with a solution such as given a scene, we encode the scene with significant concepts and then decode it to produce two similar but different stereo pair of images. We will mostly formulate our network as the former model.

We will like to explore two network architectures for the problem. One architecture will be built using auto-encoders like, stacked auto-encoders which has proved very useful in pixel-level image manipulations or, transforming auto-encoders, because the problem can itself be formulated as the problem of forming the suitable transformation from one view to the other. Through the other architecture we will like to explore Generative Adversarial Networks, and try to estimate the stereo pair of image using an adversarial net framework.

Some of the experiments we look forward to are,

- We will experiment with both L_1 and L_2 loss, although for pixel level prediction, L_1 loss has been claimed to be a better loss function. [16]
- GANs are really tricky to train with generative layers favouring ReLU [17] and Sigmoid activation, while adversarial networks favoring, Maxout [18] activation. We will experiment with these to see what works out best.
- Using dropout with the discriminator model, and batch normalization with each convpool layer for autoencoders, may give us stable result in training the networks.

4. Analysis

3D scene understanding requires depth estimation from images. Most of the work done in this area is on depth estimation so there is no standard baseline to test our methods. Thus, we will resort to test our results using the baseline of [15] which uses pixel-wise reconstruction error using Mean Absolute Error (MAE) for methods used in [21].

We also plan to do a qualitative evaluation by creating anaglyphic images of our generated stereo pairs and viewing them through dual-colored 3D glasses. We can use them to gain a approval percentage of our 3D images through a human study.

5. Timeline

- Feb 17 - Mar 3 : Dataset obtaining and/or generation (Anurag). Hypothesizing network architectures to explore (Siddhant).
- Mar 4 - Mar 13 : Writing codes for the networks. Auto-encoders (Anurag) and GANs (Siddhant)
- Mar 13 - Mar 30 : Transforming / Stacked Auto-encoders analysis. (Anurag and Siddhant)
- Mar 24 - Apr 14 : Generative Adversarial Networks analysis. (Anurag and Siddhant)
- Apr 15 - Apr 28 : Results accumulation, analysis and verification (Siddhant). Report generation (Anurag).

References

[1] Geiger, Andreas, et al. "Vision meets robotics: The KITTI dataset." The International Journal of Robotics Research 32.11 (2013): 1231-1237.

[2] D. Scharstein, H. Hirschmller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In German Conference on Pattern Recognition (GCPR 2014), Mnster, Germany, September 2014.

[3] A. Dai, A. Chang, M. Savva, M. Halber, T. Funkhouser, M. Niebner. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. https://arxiv.org/pdf/1702.04405.pdf

[4] Hinton, Geoffrey E., Alex Krizhevsky, and Sida D. Wang. "Transforming auto-encoders." International Conference on Artificial Neural Networks. Springer Berlin Heidelberg, 2011.

[5] Dong, Chao, et al. "Image super-resolution using deep convolutional networks." IEEE transactions on pattern analysis and machine intelligence 38.2 (2016): 295-307.

[6] Vincent, Pascal, et al. "Extracting and composing robust features with denoising autoencoders." Proceedings of the 25th international conference on Machine learning. ACM, 2008.

[7] Mao, Xiaojiao, Chunhua Shen, and Yu-Bin Yang. "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections." Advances in Neural Information Processing Systems. 2016.

[8] Reed, Scott, et al. "Generative adversarial text to image synthesis." Proceedings of The 33rd International Conference on Machine Learning. Vol. 3. 2016.

[9] Zhang, Liang, and Wa James Tam. "Stereoscopic image generation based on depth images for 3D TV." IEEE Transactions on broadcasting 51.2 (2005): 191-199.

[10] Eigen, David, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network." Advances in neural information processing systems. 2014.

[11] Garg, Ravi, Gustavo Carneiro, and Ian Reid. "Unsupervised CNN for single view depth estimation: Geometry to the rescue." European Conference on Computer Vision. Springer International Publishing, 2016.

[12] Kolmogorov, Vladimir, Pascal Monasse, and Pauline Tan. "Kolmogorov and Zabihs graph cuts stereo matching algorithm." Image Processing On Line 4 (2014): 220-251.

[13] Luo, Wenjie, Alexander G. Schwing, and Raquel Urtasun. "Efficient deep learning for stereo matching." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[14] Zbontar, Jure, and Yann LeCun. "Stereo matching by training a convolutional neural network to compare image patches." Journal of Machine Learning Research 17.1-32 (2016): 2.

[15] Xie, Junyuan, Ross Girshick, and Ali Farhadi. "Deep3d: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks." European Conference on Computer Vision. Springer International Publishing, 2016.

[16] Mathieu, Michael, Camille Couprie, and Yann LeCun. "Deep multi-scale video prediction beyond mean square error." arXiv preprint arXiv:1511.05440 (2015).

[17] Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. "Deep Sparse Rectifier Neural Networks." Aistats. Vol. 15. No. 106. 2011.

[18] Goodfellow, Ian J., et al. "Maxout Networks." ICML (3) 28 (2013): 1319-1327.

[19] Baig, Mohammad Haris, et al. "Im2depth: Scalable exemplar based depth transfer." Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on. IEEE, 2014.

[20] Saxena, Ashutosh, Min Sun, and Andrew Y. Ng. "Make3d: Learning 3d scene structure from a single still image." IEEE transactions on pattern analysis and machine intelligence 31.5 (2009): 824-840.

[21] Eigen, David, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network." Advances in neural information processing systems. 2014.