Learning Based Depth Map Estimation: Considering Noise and Scene Categories

A THESIS

submitted by

Seema Kumari

(PTS1501)

for the award of the degree

of

Master of Science

(by Research)

SCHOOL OF COMPUTING AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY MANDI

2019

"Difficulties in your life do not come to destroy you, but to help you realize your hidden potential and power, let difficulties know that you too are difficult" – Dr. A. P. J. Abdul Kalam

To My Parents

Smt. Santosh Devi *and* **Sri. Dinesh Kumar**

My Brothers

Dr. Rajeev Kumar*,* **Mr. Vinay Kumar** *and* **Late Mohit Kumar**

My Husband

Dr. Srimanta Mandal

DECLARATION

I hereby declare that the entire work embodied in this thesis is the result of the investigations carried out by me in the **School of Computing and Electrical Engineering, Indian Institute of Technology Mandi**, under the supervision of **Dr. Arnav Bhavsar**. This work has not been submitted elsewhere for any degree or diploma. In keeping with the general practice, due acknowledgments have been made wherever the work described is based on a finding of other investigators. In addition, I certify that no part of this work will, in future, be used for submission in my name, for the award of any other degree at any university.

Kamand, 175 005
Date:

Seema Kumari

THESIS CERTIFICATE

This is to certify that the thesis titled **Learning Based Depth Map Estimation: Considering Noise and Scene Categories**, submitted by **Seema Kumari**, to the Indian Institute of Technology, Mandi, for the award of the degree of **Master of Science** (by Research), is a bonafide record of the research work done by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other institute or university for the award of any degree or diploma.

Kamand, 175 005
Date:

Dr. Arnav Bhavsar
(Guide)

Acknowledgments

The success and final outcome of this research work required a lot of guidance and assistance from my supervisor and I am extremely privileged to give my deepest thanks to my guide Dr. Arnav Bhavsar for his expert advice and encouragement throughout my research work. He has always been there to motivate and encourage me in the right direction. Also, I would like to thank all the members of annual progress committee that consists of Dr. Bharat Singh Rajpurohit, Dr. Anil K. Sao, Dr. Aditya Nigam and Dr. Syed Abbas for their valuable inputs in my research work.

I really want to thank my job supervisors Dr. Bharat Singh Rajpurohit and Dr. Satyajitsinh A. Thakor for their constant support. I would like to also thank my colleagues Mr. Shivam Prajapti, Mr. Arun Kumar and Mr. Tarun Verma for all their support.

I would like to thank all MANAS Lab group members for all their support in all forms. Especially, I would like to extend my sincere gratitude to Dr. Renu M. Rameshan for her advice for research work as well as life. I also want to thank Sujeet Kumar, Krishan Sharma, Ranjeet R. Jha, Krati Gupta, Shikha Gupta and Arshdeep S. Boparai for their support in all forms.

I would like thank to my Pooja *bhabi*, Deepa *bhabi* and Puspa *bhabi* for taking care of me. I owe my gratitude to my Maa, Papa, Brother and other family members for their endless support, care, blessings in my entire journey. I also want to thank my brother Dr. Rajeev Kumar for his all support and care.

Importantly, I love to thank my dear husband Srimanta Mandal without his care, help, motivation, and support I could not have completed my research work smoothly. Last but not least, there are persons not mentioned here but, deserve my token of appreciation, I want to thank them all.

ABSTRACT

Keywords: *Depth map from single intensity image estimation, CNN, Residual connection, Encoder-decoder, Hourglass, Perceptual loss, AWGN, Image denoising, Sparsity, Non-local grouping, Dictionary, Edge preservation, Scene classification*

3D scene analysis can play a crucial role in different 3D vision-related applications, where depth information is pivotal. However, accurate dense depth sensing through active depth sensors (e.g. Laser depth scanner) is costly. An alternative is to employ low-cost depth sensors, which yields noisy depth information. Another common alternative which is still prevalent is that of depth estimation from intensity images using stereo. However, in this case, establishing correspondences among the multiple viewpoints is often not accurate due to various issues such as illumination, occlusion and so on. Thus, in recent years learning based depth estimation from single intensity images has been explored. However, intensity images can be noisy due to sensor characteristics.

On these lines, we propose an approach to estimate depth from a single intensity image using a learning-based strategy. Here, we have developed a novel convolutional neural network (CNN) encoder-decoder architecture, which learns the depth information using example pairs of color images and their corresponding depth maps. The proposed model is based on an integration of residual connections within pooling (down-sampling) and up-sampling layers, and hourglass module which operates on the encoded features, thus processing these at various scales. Furthermore, the model is optimized under the constraints of perceptual loss as well as the mean squared error loss. The perceptual loss considers the high-level features, thus operating at a different scale of abstraction, which is complementary to the mean squared error loss that considers a pixel-to-pixel error.

Considering that the training and testing dataset can be noisy, the estimated depth may not be accurate. Although our depth estimation framework can handle low-level noise in the intensity test image, a higher level of noise can distract the estimated depth map. For this scenario, we propose a denoising algorithm for both intensity images and depth maps that can address higher levels of noise. It has been shown that for denoising, non-local similar patches play an important role. Nevertheless, noise may create ambiguity in finding similar

patches, hence it may degrade the results. However, most of the non-local similarity-based approaches do not consider the issue of noisy patch grouping. Hence, we propose to denoise an image by mitigating the issue of grouping non-local similar patches in the presence of noise in the transform domain using sparsity and edge-preserving constraints. The effectiveness of the transform domain grouping of patches is utilized for learning dictionaries and is further extended for achieving an initial approximation of sparse coefficient vector for the clean image patches. The results are further improved by employing edge preserving constraints and processing at coarser scales. Our technique is useful to preserve the surface discontinuities and prominent details in depth and intensity images while suppressing noise, and we demonstrate clear benefits of denoising.

Another aspect that is considered in this work, is whether an apriori knowledge of scene type can benefit in depth estimation. We demonstrate the improvement in estimating the depth map by classifying different indoor scenes and building different depth estimation models for scene types. Such an approach may be useful in an application involving a small and fixed number of scenes. In order to build a classifier, we have used a smaller version of Residual Convolutional Neural Network (ResNet-18) that discriminates between different indoor scenes (e.g. bookstore, dining, bathroom, classroom, and kitchen, etc.) even in presence of noise in testing images. Here, our denoising method can help in accurate estimation of the depth map. Such an approach can not only serve as an initial step of depth estimation but it can also be useful in scene classification/retrieval application.

Contents

Abstract **ii**

List of Tables **vii**

List of Figures **vii**

Abbreviations **xii**

1 Introduction **1**

1.1 Depth Sensing/Estimation Techniques 2

1.1.1 Active Sensors 2

1.1.2 Passive Techniques from Multiple Images 3

1.1.3 Monocular Depth Estimation 4

1.2 Objective and Scope of the Work 5

1.2.1 Convolutional Neural Network for Depth Map Estimation from a Single Intensity Image 6

1.2.2 Image Denoising 7

1.2.3 Scene Classification 9

1.3 Contributions of the Thesis 10

1.4 Organization of the Thesis 11

2 Background **13**

2.1 Sparse Representation 13

2.1.1 Linear Systems of Equations 13

2.1.2 Regularization 15

2.1.3 Sparse Modelling 16

2.1.4 The Dictionary 18

2.2 Convolutional Neural Network (CNN) 18

2.2.1 Shallow vs Deep Convolutional Neural Network 19

2.2.2 Residual Connections 20

2.2.3 Basic Autoencoder Framework 21

2.2.4 Autoencoder with Skip Connections 23

2.3 Summary 24

3 Single Image Based Depth Map Estimation via Residual CNN Encoder-Decoder and Stacked Hourglass 25

3.1 Introduction 25

3.2 Related Work 26

3.3 Proposed Approach for Single Image Depth Estimation with CNN 29

3.3.1 Network Model 29

3.3.2 Implementation Details 31

3.3.3 Loss Function 32

3.4 Experimentation 33

3.4.1 Dataset 33

3.4.2 Assessment metrics 34

3.4.3 Results and Comparisons 35

3.4.4 Ablation Study 38

3.5 Summary 41

4 Sparse Representation Based Single Image Denoising with Transform Domain Patch Similarity 44

4.1 Introduction 44

4.2 Grouping Similar Patches: Noiseless vs Noisy 47

4.3 Patch Similarity in Transform Domain 49

4.3.1 Patch Grouping Using Transform Domain Similarity 49

4.3.1.1 Learning Dictionaries 50

4.3.1.2 Estimation of Sparse Coefficient Vector 51

4.4 Edge Preservation 51

4.4.1 Preserving edges at coarser scales 51

4.4.1.1 Multi-scale edge preservation 53
4.4.2 Preventing transitional smoothing 54
4.5 Experimentation 55
4.5.1 Comparison 55
4.5.2 Denoising in Depth Map Estimation 58
4.5.3 Comparisons and analysis among orthogonal transformations and filtering technique 59
4.6 Summary 61

5 Scene Classification for Depth Map Estimation 64

5.1 Introduction 64
5.2 Related Work 66
5.3 Proposed Framework 68
5.3.1 Residual Learning 68
5.3.2 Network Architecture 68
5.3.3 Implementation Detail 70
5.3.4 Depth Estimation with Scene Classification 70
5.4 Experimentation 70
5.4.1 Results for Classification 71
5.4.2 Effect of Scene Classification in Depth Map Estimation 73
5.5 Summary 76

6 Summary and Future Work 78

6.1 Summary 78
6.2 Future Work 80

References 81

List of Publications 90

List of Tables

3.1 Quantitative comparison of results on the Ikea chair dataset 35

3.2 Quantitative comparison of results on the NYU V2 Depth dataset with state-of-the-art approaches 38

4.1 Quantitative Comparison of Results for Depth Maps Produced by Denoising Approaches via PSNR (top) & SSIM (bottom) 56

4.2 Quantitative Comparison of Results for Intensity Images Produced by Denoising Approaches via PSNR (top) & SSIM (bottom) 56

5.1 Quantitative results of indoor scene classification by using ResNet-18 72

5.2 Quantitative comparison of results on the NYU V2 Depth dataset 74

5.3 Quantitative comparison of results in presence of noise in input intensity images on the NYU V2 Depth dataset 76

List of Figures

1.1 Applications of depth information in (a) robot navigation and (b) autonomous driving platform Annieway (Image Source: `http://www.cvlibs.net/datasets/kitti/index.php`). 1

1.2 Examples of active depth sensor (a) Microsoft Kinect and (b) Intel Realsense active depth sensors. 2

1.3 Principle of estimating depth using stereo technique. (Image Source: `http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/HENKEL/research/stereo/principle.html`) 3

1.4 Objective of depth estimation (a) Intensity image and (b) Depth map of left image [2]. 4

1.5 A basic framework for depth estimation from a single intensity image based on CNN auto-encoder. 7

1.6 Objective of denoising. 8

1.7 A basic framework of scene classification. 9

1.8 Different image modalities of indoor scene: (a) Intensity image of the scene, and (b) Depth map of the scene [2]. 10

2.1 Over-determined System. 14

2.2 Under-determined System. 14

2.3 Dense Solution. 16

2.4 From dense solution to sparse solution. 17

2.5 Basic block diagram of CNN based classification network [3]. 19

2.6 Plot of error convergence for 20 and 50 layers networks, in left first, training error and second, testing error [4]. 20

2.7 Block diagram of (a) Direct feed-forward network and (b) Neural network with residual connection. 21

2.8 Basic block diagram of autoencoder. 22

2.9 Basic block diagram of autoencoder with skip connections. 24

3.1 Block diagram of our network for depth estimation is consisted of multiple stacked layers with hourglass in encoder-decoder model. 30

3.2 Block diagram of an hourglass module. 31

3.3 Visual comparison on Ikea Chair dataset: First left column - intensity images, second column - results using proposed approach, third column - ground truth depth map. . 36

3.4 Visual comparison on NYU V2 Depth dataset: First left column - intensity images, second column - results of. [5], third column - results using proposed approach, fourth column - ground truth. 39

3.5 Visual comparison on NYU V2 Depth dataset: first row - intensity images, second row - ground truth depth map, third row - results of [5], fourth row - results of [6], fifth row - results using proposed approach. 40

3.6 First column represents the noisy intensity images for different amounts of noise across rows (row1: sigma=5; row2: sigma=10; row3: sigma=15). Second column - results of. [5] (PSNR = 16.71, 15.03, 14.46 in dB), third column - results using proposed approach (PSNR = 22.26, 21.55, 20.28 in dB), and fourth column - ground truth. 41

3.7 First column represents the noisy intensity images for different amounts of noise across rows (row1: sigma=5; row2: sigma=10; row3: sigma=15). Second column - results of. [5] (PSNR = 17.36, 16.01, 15.03 in dB), third column - results using proposed approach (PSNR = 21.23, 19.92, 18.58 in dB), and fourth column - ground truth. 42

3.8 Convergence plot with respect to epoch (a) without perceptual loss and (b) with perceptual loss. 42

3.9 Visual results on NYU V2 Depth dataset: first left - intensity image, second left - result without perceptual loss, third left - result with perceptual loss, fourth left - ground truth. 43

3.10 Visual results on NYU V2 Depth dataset: first left - intensity image, second left - result without using hourglass model, third left - result with hourglass, fourth column - ground truth. 43

4.1 First column represents the noisy intensity images for different amounts of noise across rows (row1: sigma=20; row2: sigma=50;). Second column contains the depth results of proposed approach (PSNR = 18.59, 13.98 in dB), and the third column shows the ground truth depth maps. 45

4.2 First column - represents the intensity images, Second column - estimated depth maps with noisy training depth maps at $\sigma = 50$ (Average PSNR = 13.89), third column-estimated depth maps in absence of noise, and in fourth column - ground truth. . . 46

4.3 Demonstration of patch grouping: (a) Ground-truth image, (b) noisy version (AWGN, $\sigma = 100$) of ground-truth image. Three groups of patches, achieved from K-means clustering of (c) the clean ground-truth image, (d) noisy image (AWGN, $\sigma = 100$), (e) noisy image based on transform domain similarity. 48

4.4 Demonstration of “Down-sampling has denoising effect” for cameraman image: (a) The edginess feature of the clean image, (b) the edginess feature of the down-sampled (by factor 3) cameraman image, (c) the edginess feature of the down-sampled (by factor 3) noisy cameraman image (noise level $\sigma = 50$), and (d) the edginess feature of the noisy cameraman image. 52

4.5 Block diagram representation of preserving edges at coarser scales. 53

4.6 Block diagram for preventing transitional smoothing. 54

4.7 Denoising results for noisy fingerprint image ($\sigma = 50$): (a) The noisy image (PSNR=14.14). Result of (b) K-SVD denoising [7] (PSNR=23.07), (c) NCSR [8] (PSNR=24.53),(d) BM3D [9] (PSNR=24.49), (e) proposed approach (PSNR=24.49), and (f) the ground truth image. 57

4.8 Denoising results for noisy Lena image ($\sigma = 100$): (a) The noisy image (PSNR=8.12). Result of (b) K-SVD denoising [7] (PSNR=24.49), (c) NCSR [8] (PSNR=25.66),(d) BM3D [9] (PSNR=25.91), (e) the proposed approach (PSNR=25.72), and (f) the ground truth image. 58

4.9 Denoising results for noisy Cones depth map ($\sigma = 50$): (a) The noisy image (PSNR=14.12). Result of (b) K-SVD denoising [7] (PSNR=31.89), (c) NCSR [8] (PSNR=33.98), (d) BM3D [9] (PSNR=33.96), (e) proposed approach (PSNR=34.17), and (f) the ground truth depth map. 59

4.10 Denoising results for noisy baby depth map (σ = 100): (a) The noisy image (PSNR=8.11). Result of (b) K-SVD denoising [7] (PSNR=28.09), (c) NCSR [8] (PSNR=32.11),(d) BM3D [9] (PSNR=31.25), (e) proposed approach (PSNR=32.43), and (f) the ground truth depth map. 60

4.11 First column represents the noisy intensity images for different amounts of noise across rows (row1: sigma=20; row2: sigma=50;). Second column contains depth maps for noisy intensity images (PSNR = 18.59, 13.98), in third column - the resultant depth maps for denoised intensity images (PSNR = 24.21, 20.11), and the fourth column shows the ground truth depth maps. 61

4.12 First column - represents the intensity images, Second column - estimated depth maps with noisy training depth maps at σ = 50 (Average PSNR = 13.89), third column - estimated depth maps with denoised training depth maps (Average PSNR = 20.76), fourth column - estimated depth maps in absence of noise, and in fifth column - ground truth. 62

4.13 Comparisons of results: DFT (Blue) vs. Wavelet (Cyan) vs. DCT (Yellow) vs. GIF (Red) [10] for different powers of noise (AWGN: $\sigma_n = 25, 50, 100$); (a) Average PSNR for intensity images; (b) Average SSIM for intensity images; (c) Average PSNR for depth maps and (d) Average SSIM for depth maps. 63

5.1 Visual comparison on NYU V2 Depth dataset: First left column - intensity images, second column - estimated depth map tested on single model for all classes, third column - estimated depth map tested on specific scene model for all classes, fourth column - ground truth depth map. 65

5.2 Block diagram of basic residual learning. 69

5.3 Block diagram of a residual convolutional neural network for scene classification. . 69

5.4 Block diagram of the combination of classification and depth map estimation framework. 71

5.5 Some example results for classification: in first row, first left - bookstore, second - bathroom, third - library image, in second row, first left - classroom, second - dinning, third - kitchen indoor images of NYU V2 Depth dataset. 72

5.6 Some example results for classification: in first row, first left - bookstore, second - classroom, third - bathroom, in second row, first left - kitchen, second - dinning, third - dinning indoor noisy images at σ = 20 of NYU V2 Depth dataset. 73

5.7 Some example results for classification: in first row, first left - kitchen, second - bathroom, third - library, in second row, first left - library, second - classroom, third - dinning indoor denoised images at σ = 20 of NYU V2 Depth dataset. 73

5.8 Some example results for classification: in first row, first left - kitchen, second - bathroom, third - classroom, in second row, first left - kitchen, second - bookstore, third - dinning indoor denoised images at σ = 50 of NYU V2 Depth dataset. 74

5.9 Visual comparison on NYU V2 Depth dataset: First left column - intensity images, second column - estimated depth map tested on single model for all classes, third column - estimated depth map tested on specific scene model for all classes, fourth column - ground truth depth map. 75

5.10 First column represents the noisy intensity images for different amounts of noise across rows (row1: sigma=5; row2: sigma=10; row3: sigma=15). Second column - results on single model for all classes (PSNR = 22.62, 21.89, 20.04 in dB), third column - results on specific scene model for all classes (PSNR = 22.79, 21.90, 20.05 in dB), and fourth column - ground truth. 76

Abbreviations

1D	- One Dimensional
2D	- Two Dimensional
3D	- Three Dimensional
AI	- Artificial Intelligence
IR	- Infrared
ToF	- Time-of-Flight
HF	- High Frequency
DCT	- Discrete Cosine Transform
DFT	- Discrete Fourier Transform
PCA	- Principal component analysis
AWGN	- Additive White Gaussian Noise
NLM	- Non Local Mean
RMSE	- Root Mean Square Error
PSNR	- Peak Signal-to-Noise Ratio
SSIM	- Structural SIMilarity index
AWGN	- Additive White Gaussian Noise
K-Means	- K times Mean computation
CNN	- Convolutional Neural Network
FCN	- Fully Connected Network
ResNet	- Residual Network
CRF	- Conditional Random Field
BM3D	- Block-Matching and 3D filtering
GIF	- Guided Image Filtering
NCSR	- Nonlocally Centralized Sparse Representation
K-SVD	- K times Singular Value Decomposition