Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

b-it Bonn-Aachen International Center for Information Technology

Master Thesis Proposal

Multi-view Stereo by Temporal Nonparametric Fusion

Manoj Kolpe Lingappa

Supervised by

Prof. Dr Nico Hochgeschwender

Second Supervisor

MSc. Deebul Sivarajan Nair

Month 2022

1 Introduction

Computer vision aims to understand the surrounding environment using various mathematical modelling techniques. First generation of depth estimation was based on pixel matching between multiple images taken from a calibrated cameras [24]. With the development of 3D reconstruction, depth sensors are becoming increasingly popular in areas such as self-driving cars. These sensors are used to obtain the information of the surrounding environment. However, the acquired depth maps from these sensors are sparse in nature due to low computational power resulting in information loss of the captured depth map. Another approach to reconstruct the 3D scene of an object is with the help of high quality images captured from the camera where the texture and lighting information are captured [38]. Reconstruction of three dimensional view from images is a classic problem in the computer vision domain. Multi view stereo algorithms can reconstruct the disparity maps or three dimensional view of an object from the images [2]. It is the process of reproducing the 3D scenes from the multiple images given the camera poses and internal camera matrix. Number of areas take advantage of the reconstruction such as 3D mapping, 3D printing, video games, online shopping in the consumer domain, visual effect industry, digital mapping [8], vehicle tracking, aircraft estimation and positioning [27], depth estimation [34]. Depth estimation is the process of extracting the depth of objects present in the images by capturing and processing multiple images of the object taken from different locations. Images can be obtained from a stereo camera or a monocular camera. This work is based on the monocular camera images. Estimation of depth from the unconstrained monocular camera images is a challenging task. Most of the state of the art depth estimation algorithms are based on deep learning and compute cost volume according to the hypothesized depths. 3D convolution is applied to this cost volume to regress and predict the depth map [11]. This work aims to reproduce the result and deploy depth disparity estimation algorithms in a mobile device Finally end the research work with feasibility study of depth estimation architecture to segmentation.

2 Problem Statement

Stereo vision is one of the field of computer vision that targets to construct the 3D model of a object using images. Over the past years a large number of algorithms and architectures have been proposed to find the 3D geometry of the object. However, a lack of dataset taken at varying environmental conditions made it difficult to compare the performance of the algorithms [30]. It takes a lot of time to process large images and bad reconstruction with the low textured images [17], [30], [31]. Most of the state of the art depth estimation algorithm are computationally heavy and cannot be deployed on the edge device. A light weight architecture with reasonable performance needs to be developed to deploy in a low computational power devices. Conventional approaches uses two view stereo rigs for reconstruction. However, estimation of depth from unconstrained monocular camera images is a challenging task. There are advantage of using the moving camera. Firstly with larger baseline the accuracy of the distant object can be improved. Secondly with multiple varying point images are able to fuse all the information for robust and stable depth estimation [13]. This work concentrate on the depth estimation from unconstrained monocular camera images and extension of the disparity map estimation architecture to the segmentation.

3 Related Work

Multi view stereo (MVS) is a general term given to the group of techniques that uses stereo correspondence to find the geometry of a object using the images captured from different viewpoint. 3D reconstruction of the target object can be done with classical [3], [5], [23], [22], [6], [10] or modern deep learning based approaches [35], [16], [36], [4]. Goal of the image based 3D reconstruction algorithm can be defined as "given a set of photographs of an object or a scene, estimate the most likely 3D shape that explains those photographs, under the assumptions of known materials, viewpoints, and lighting conditions" [8]. With a set of assumptions state of the art architecture can produce highly detailed reconstructions from large set of images. MVS is classified as follows (a) volumetric reconstruction method [20] (b) point cloud reconstruction [9] and (c) depth map based method [34]. In a pairwise

stereo method image rectification is performed to limit the correspondences finding in the horizontal epipolar lines. This problem is addressed by the volumetric representation of the view [21], [15], [7], [33], [25]. Due to high memory load the volumetric approach is not suitable to large scenes but it gives good performance for small objects. A light weight architecture is proposed by Wang et al. A matching cost volume is computed using the plane sweeping approach from the nearby images and then regard depth estimation as regression problem which is found using the deep neural network [32]. Plane sweep volume method does not require any rectified image. However, the approach require intrinsic and extrinsic camera parameter in advance or can be computed using structure from motion [15]. Learning based depth reconstruction can be described as finding a predictor f_θ that can find the depth maps $\hat{D}$ from the set of images I, which are close to the unknown depth map D. Mathematically, we are trying to find a function f_θ such that the loss function $L(I) = d(f_\theta(I), D)$ is minimized. Where θ is the learnable parameter, and $d(.)$ is the measure of distance between the real depth D and the predicted depth $\hat{D}$ [24]. There are two class of depth estimation methods. First class of method involves traditional stereo matching approaches to find the correspondences which in turn help to find the disparity map. Depth map can be found from these disparity map [29]. There are three stages for prediction of function f, first is the feature extraction, feature matching and cost aggregation, and finally depth estimation [24]. Second class of method involves end to end trainable network [37]. Training requires large amount of data and these approaches is similar to the traditional stereo matching algorithm by breaking the problem into small chunks and computing the result[24]. Early multi view stereo methods works on the finding the correspondences between multiple image patches[12]. Most of the state of the art disparity map estimation architecture requires high computational power thereby limiting the deployment in the low computational devices such as mobile phones, tablets. Work by [13] deploys the depth estimation architecture on a IOS device, current work aims to deploy the architecture in a android device also find the feasibility of the depth estimation architecture on to the segmentation task.

4 Project Plan

The following sections explains work packages, milestones and project schedule, and deliverable.

4.1 Work Packages

The bare minimum will include the following packages:

WP1 Literature Search

This section aims to extensively search for reference to papers that are related to multi view stereo.

T1.1 Literature review

In this task collection of literature related to multi view stereo is done and conceptual understanding of the disparity map estimation from images.

WP2 Data aggregation and preprocessing

This section explains the data collection and data preprocessing.

T2.1 Data collection

In this section data is collected from multiple sources and the nature of the data is examined and analyzed using visualization tools or statistical methods. Analysis is carried out to ensure data is diverse, unbiased and abundant in nature.

T2.2 Data preprocessing

Preprocessing of data is carried out based on the input requirement of the model. The preprocessing step converts the raw sourced data into a format that enables successful training of the model.

WP3 Model implementation

This section explains the development and implementation of the model.

T3.1 Evaluation of the model

In this task aims to reproducing of the Multi-view Stereo by Temporal Nonparametric Fusion architecture results.

T3.2 Extension of disparity map estimation architecture
In this section extension of multi view stereo disparity estimation architecture to segmentation task is carried out.

WP4 Evaluation
This package aims to evaluate the results based on the different metrics.

T4.1 Results reporting
In this task, the output of evaluation is reported.

WP5 Project Report
This work package involves writing the project report. It is done in parallel with all previous work packages.

T5.1 Documentation of literature reviewed
In this task, a detailed analysis of the state of the art is done and all the findings are documented in the project report.

T5.2 Documentation of reproduced results
In this task, the implementation result of Multi-view Stereo by Temporal Nonparametric Fusion is done.

T5.3 Documentation of extended architecture results of segmentation
In this task, the evaluation is conducted on researched segmentation architecture based on the disparity map estimation architecture.

4.2 Milestones

M1 Literature search

M2 Data collection and preprocessing

M3 Implementation of Multi-view Stereo by Temporal Nonparametric Fusion

M4 Experimental Analysis

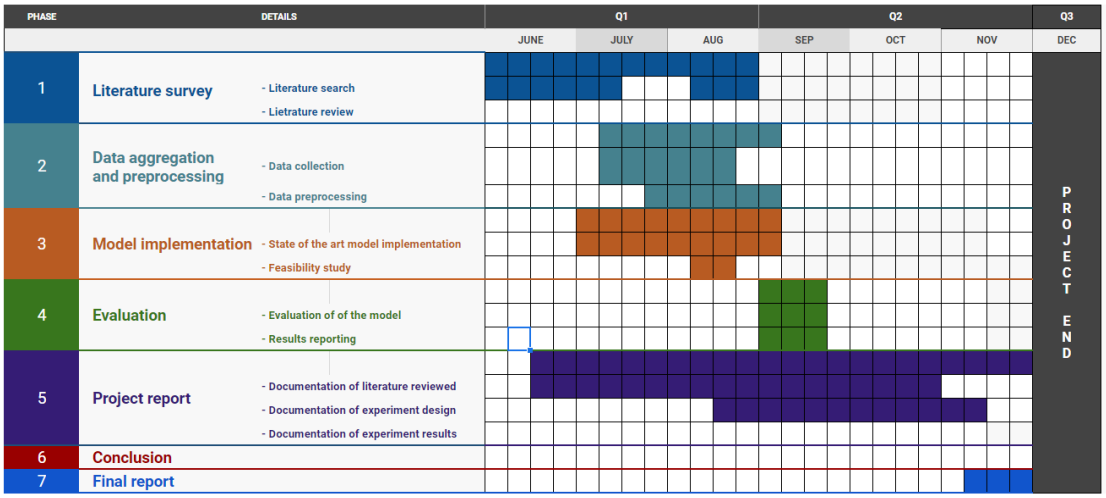
M5 Deployment of model in the android phone

M6 Researching on the feasibility of disparity map estimation architecture for segmentation

M7 Report submission

4.3 Project Schedule

Figure 1:



4.4 Deliverables

Minimum viable

- Survey
- Analysis of state of the art
- Reproducing of the existing paper results

Expected

- Understanding and executing Temporal non-parametric fusion
- Survey of method available in Temporal fusion.
- Cross transfer this methodology for other task like object classification.
- Simple simulated use case
- Experiment with different kernel functions and hyperparameter tuning

Desired

- Research on the feasibility study of disparity map estimation architecture to solve the segmentation task

References

[1] Gianfranco Bianco, Alessandro Gallo, Fabio Bruno, and Maurizio Muzzupappa. A comparative analysis between active and passive techniques for underwater 3d reconstruction of close-range objects. *Sensors*, 13(8):11007–11031, 2013.

[2] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.

[3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.

[4] Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2524–2534, 2020.

[5] Jeremy S De Bonet and Paul Viola. Poxels: Probabilistic voxelized volume reconstruction. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 418–425. Citeseer, 1999.

[6] Carlos Hernández Esteban and Francis Schmitt. Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3): 367–392, 2004.

[7] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5515–5524, 2016.

[8] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

[9] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[10] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[11] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo

matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

[12] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler. Learned multi-patch similarity. In *Proceedings of the IEEE international conference on computer vision*, pages 1586–1594, 2017.

[13] Yuxin Hou, Juho Kannala, and Arno Solin. Multi-view stereo by temporal nonparametric fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2019.

[14] Yuxin Hou, Arno Solin, and Juho Kannala. Unstructured multi-view depth estimation using mask-based multiplane representation. In *Scandinavian Conference on Image Analysis*, pages 54–66. Springer, 2019.

[15] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018.

[16] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019.

[17] Michal Jancosek and Tomás Pajdla. *Segmentation based multi-view stereo.* Citeseer, 2009.

[18] Ray A Jarvis. A perspective on range finding techniques for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):122–139, 1983.

[19] Ray A Jarvis. A perspective on range finding techniques for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2):122–139, 1983.

[20] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017.

[21] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. *Advances in neural information processing systems*, 30, 2017.

[22] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *European conference on computer vision*, pages 82–96. Springer, 2002.

[23] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 1, pages 307–314. IEEE, 1999.

[24] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[25] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 781–796, 2018.

[26] Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, et al. The digital michelangelo project: 3d scanning of large statues. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 131–144, 2000.

[27] Mendoza Guzmán Vıctor Manuel, Mejıa Munoz José Manuel, Moreno Márquez Nayeli Edith, Rodrıguez Azar Paula Ivone, and SRE Ramirez. Disparity map estimation with deep learning in stereo vision. *Proceedings of the Regional Consortium for Foundations, Research and Spread of Emerging Technologies in Computing Sciences (RCCS+ SPIDTEC2)*, pages 27–40, 2018.

[28] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

[29] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1):7–42, 2002.

[30] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, 1:519–528, 2006. ISSN 0302-2345.

[31] Christoph Strecha, Wolfgang Von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

[32] Kaixuan Wang and Shaojie Shen. Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International conference on 3d vision (3DV)*, pages 248–257. IEEE, 2018.

[33] Youze Xue, Jiansheng Chen, Weitao Wan, Yiqing Huang, Cheng Yu, Tianpeng Li, and Jiayu Bao. Mvscrf: Learning multi-view stereo with conditional random fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4312–4321, 2019.

[34] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[35] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018.

[36] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5525–5534, 2019.

[37] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang, and Feng Qian. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, 63(9):1612–1627, 2020.

[38] Qingtian Zhu, Chen Min, Zizhuang Wei, Yisong Chen, and Guoping Wang. Deep learning for multi-view stereo via plane sweep: A survey. *arXiv preprint arXiv:2106.15328*, 2021.