DOCTORAL DISSERTATION

Robustness of Feature Based Calibration in New Age 3D Applications

DEEPAK DWARAKANATH

July 2017

Submitted to the Faculty of Mathematics and Natural Sciences at the University of Oslo in partial fulfilment of the requirements for the degree of Philosophiae Doctor

Abstract

There has been an increasing demand for multimedia systems in various areas. An innovative vision for multimedia systems has paved way for the advent of several interesting and useful applications, especially in the 3D arena. Such current and new age 3D applications based on single or multiple camera images can, for example, be seen in the field of vision based inspection, mixed reality art performance, sports analytics, augmented reality and image metrology. For high quality performance in these applications, the underlying focus is on the quality of image based 3D reconstruction.

To achieve high quality performance in image based 3D reconstruction, an accurate *camera calibration* is necessary. Camera calibration provides a priori knowledge about the camera's intrinsic parameters (such as focal length, principal axes, skewness and lens distortion) and extrinsic parameters (such as spatial position and orientation). In certain application scenarios, traditional checkerboard calibration process (requires a checkerboard target) and marker-based calibration process (requires an identifiable marker) are impossible or inconvenient. In such cases, 3D systems have to rely on an alternate solution, i.e., *Feature Based Calibration (FBC)*, where interesting feature points in the camera images are extracted and used for the calibration process. Therefore, the accuracy of FBC is an important factor defining the quality of single or multiple camera 3D systems.

Although, the FBC can be integrated in 3D systems, there are several practical issues involved, e.g., (1) misalignment, arrangement and changes in the properties of one or more cameras; (2) misalignment of captured object scene; and (3) noisy feature points extracted from the images. Therefore, in this thesis, the aim is to explore the challenges in designing FBC to achieve high accuracy and robustness in 3D systems.

In order to explore the influence of practical issues on FBC, relevant evaluation procedures that relates to specific application scenarios was setup. Extensive tests were carried out using both real and virtual datasets and simulations. The effects of camera misalignment, adoption of FBC, characterization of the state-of-the-art feature extractors and camera pose estimator were studied for obtaining an accurate and robust 3D reconstruction. The evaluation of results are discussed by assessing the accuracy and robustness of FBC against practical issues. Consequently, tolerances for camera misalignment, operational limits of state-of-the-art feature extractors and an estimation of camera density to capture the scene are presented. Finally, recommendations are given for researchers and system developers to design better 3D systems considering practical issues for their applications scenarios.

Acknowledgements

The completion of this thesis was possible with the support of many people. I would like to take this opportunity to thank all of them for this journey.

With lots of gratitude, I would like to thank my advisors, Carsten Griwodz, Pål Halvorsen and Alexander Eichhorn for providing constant inspiration, guidance and supervision during my research work. All of them have been co-authors of some of the papers presented in the thesis. My sincere thanks to Jacob Lildballe - Image House PantoInspect A/S, Denmark, who was also a co-author for some of the papers presented in the thesis.

I thank Simula Research Laboratory, University of Oslo and Norwegian Research Council for have given me a platform to start my research career as a PhD candidate. It was great to work with Sabita Maharjan, Rajwinder Panesar-Walawege, Shaukat Ali, Ahmed Elmokashfi and Thomas Kupka. I greatly appreciate the collaborative work with the master students, Kjetil Endal and Steffan Gullichsen.

At PantoInspect A/S, Denmark, I wish to thank Lars Baunegaard With, Morten Langschwager and Claus Hoelgaard Olsen for their valuable discussions and encouragement. The PantoInspect system was used as an application scenario to conduct research and publish the results.

I thank Jan Friis Jorgensen, Ander Kühle and Ole Brydensholt - Image Metrology A/S, Denmark, and Christophe Mignot - Digital Surf, France, for their support and encouragement in the process of my research. The product from Digital Surf was used to validate a certain part of the results of the research.

I have had opportunities to work in the related area of vision systems through small projects and hence, I would like to thank, A-Star Research and Development, Singapore and Qtehcnology A/S, Denmark.

Specially, I thank my parents Sudharani A. and Dwarakanath G.R. for all their love and support in my life. I greatly thank Sphoorthi S.P., who took care of all the family affairs and provided invaluable support, in order to help me finish my research. I have received sincere support from the family in India, thanks to Chetak, Rashmi, Saanvi, Puttaswamaiah, Shashikala, Sujatha, Savitha, Shyla and many more. I would also like to thank the family in Denmark, Sri Sai Das, Kalpana Das, Raj Ponnambalam, Shilpa Kanakraj and Jayanth, for their constant support and encouragement.

Finally, I once again thank Sri Sai Das for thorough proofreading of this thesis.

Contents

I Overview **xvii**

1 Introduction **1**
- 1.1 Application Scenarios . . . 2
- 1.2 Practical Challenges . . . 3
- 1.3 Goal and Scope . . . 5
- 1.4 Problem Statement . . . 5
- 1.5 Research Methods . . . 9
- 1.6 Main Contributions . . . 9
  - 1.6.1 Publications . . . 11
  - 1.6.2 Software Development . . . 12
- 1.7 Limitations . . . 12
- 1.8 Thesis Outline . . . 13

2 Preliminary Concepts **15**
- 2.1 Image Based 3D Systems . . . 15
  - 2.1.1 Camera Calibration . . . 16
  - 2.1.2 3D Reconstruction . . . 22
- 2.2 Deeper Look Into Application Scenarios . . . 23
  - 2.2.1 Virtually Enhanced Real-life synchronizeD Interaction - ON the Edge (VERDIONE) . . . 23
  - 2.2.2 An Integrated System for Soccer Analysis (BAGADUS) . . . 28
  - 2.2.3 PantoInspect Train Monitoring System (PTMS) . . . 31
  - 2.2.4 Previz for On-set Production - Adaptive Real-time Tracking (POPART) . . . 35
  - 2.2.5 Scanning Electron Microscopes Reconstruction (SEMRECON) . . . 38
- 2.3 Conclusions for Preliminary Concepts . . . 39

3 Feature Based Calibration (FBC) **41**
- 3.1 Misalignment in Single Camera System . . . 42
  - 3.1.1 Evaluation . . . 43
  - 3.1.2 Error Analysis . . . 46
  - 3.1.3 Error Modeling . . . 49
  - 3.1.4 Discussions . . . 54
- 3.2 Adoption of Feature Based Calibration . . . 55
  - 3.2.1 Proposed Re-calibration Methodology . . . 57

3.2.2 Evaluation . . . 60
3.2.3 State-of-the-art FBC algorithms . . . 62
3.2.4 Accuracy of Measurements . . . 62
3.2.5 Error Distribution . . . 64
3.2.6 Resilience . . . 64
3.2.7 Discussions . . . 67
3.3 Misalignment in Stereo Camera System . . . 68
3.3.1 Evaluation . . . 68
3.3.2 Pure Translation Misalignment . . . 71
3.3.3 Pure Rotation Misalignment . . . 71
3.3.4 Combined Misalignment . . . 74
3.3.5 Variable Object Size . . . 74
3.3.6 Discussions . . . 76
3.4 Conclusions for FBC . . . 77

4 Feature Extraction **79**
4.1 State-of-the-art Feature Extractors . . . 80
4.2 Robustness against Camera Intrinsic . . . 83
4.2.1 Evaluation . . . 84
4.2.2 Accuracy Vs Speed . . . 89
4.2.3 Image Blur . . . 90
4.2.4 Lens Distortion . . . 90
4.2.5 Sensor Noise . . . 90
4.2.6 Discussions . . . 97
4.3 Robustness against Camera Extrinsic . . . 97
4.3.1 Evaluation . . . 101
4.3.2 2D Pixel Error . . . 105
4.3.3 Camera Pose Error . . . 106
4.3.4 Penality . . . 106
4.3.5 3D Estimation Error . . . 111
4.3.6 Comparative Performance . . . 111
4.3.7 Discussions . . . 115
4.4 FBC using SIFT for Wide Baseline . . . 119
4.4.1 Proposed Algorithm . . . 120
4.4.2 Evaluation . . . 123
4.4.3 Performance of Proposed Algorithm . . . 124
4.4.4 Discussions . . . 127
4.5 Conclusions for Feature Extraction . . . 129

5 Pose Estimation **131**
5.1 Sensitivity of Pose estimation . . . 132
5.1.1 Performance Metrics . . . 133
5.1.2 Evaluation . . . 134
5.1.3 Number of Feature Correspondence . . . 135

5.1.4 Noise in Feature Correspondence . . . 136
5.1.5 Sparsity of Feature Correspondence . . . 139
5.1.6 3D Reconstruction Metric Evaluation . . . 142
5.1.7 Discussions . . . 142
5.2 Conclusions for Pose Estimation . . . 145

6 Conclusions **147**
6.1 Main Contributions . . . 147
6.2 Practical Implications . . . 149
6.3 Practical Insight . . . 149
6.4 Future work . . . 150

II Research Papers **159**

7 Paper I: Faster and More Accurate Feature-Based Calibration for Widely Spaced Camera Pairs **161**

8 Paper II: Evaluating Performance of Feature Extraction Methods for Practical 3D Imaging Systems **169**

9 Paper III: Study the Effects of Camera Misalignment on 3D Measurements for Efficient Design of Vision-Based Inspection Systems **177**

10 Paper IV: Online Re-calibration for Robust 3D Measurement Using Single Camera-PantoInspect Train Monitoring System **193**

11 Paper V: Robustness of 3D Point Positions to Camera Baselines in Markerless AR Systems **209**

12 Poster I: 3-D Video Processing for Mixed Reality Art Performances **223**

13 Poster II: 3D Multi-view Acquisition and Rendering System **225**

14 Poster III: Multiple Camera Arrays for Real-time 3D Rendering Systems **227**

List of Figures

2.1 Typical 3D system illustrating two different workflows for 3D applications, using the knowledge of camera calibration. . . . 15
2.2 Pin-hole camera model . . . 16
2.3 Radial lens distortion. Undistorted image (left), barrel distortion (center), pincushion distortion (right) . . . 18
2.4 Tangential lens distortion . . . 19
2.5 Epipolar geometry stereo camera setup . . . 20
2.6 Feature point correspondences in stereo images. . . . 21
2.7 Illustration of image rectification. . . . 23
2.8 Illustration of depth estimation. . . . 23
2.9 World opera distributed stage performance. . . . 24
2.10 Illustration of VERDIONE capture and render subsystems. . . . 24
2.11 Multiple camera acquisition subsystem for VERDIONE. . . . 25
2.12 Overall BAGADUS architecture. . . . 28
2.13 Camera setup in Alfheim soccer stadium. . . . 29
2.14 PTMS: inspects defects on the pantographs mounted on electric trains. . . . 31
2.15 PTMS defects illustrated and shown on a pantograph. . . . 32
2.16 User Interface of PTMS inspection system. . . . 33
2.17 Pantograph image analysis. . . . 34
2.18 POPART System. . . . 36
2.19 The 3D Point cloud of the real filming set. . . . 36
2.20 The 3D reconstruction and surface analysis - waviness and roughness surfaces and ISO 25178 height parameters. . . . 38

3.1 PTMS inspection scenario: world coordinates (X_w, Y_w, Z_w) and camera coordinates (X_c, Y_c, Z_c). . . . 44
3.2 Simulation procedure. . . . 45
3.3 Variation of error in 3D width measurements of the defects, due to changes in camera position and orientation about its camera center. . . . 46
3.4 Variation of error in 3D depth measurements of the defects, due to changes in camera position and orientation about its camera center. . . . 47
3.5 Projective geometric effects of camera tilt angle in PTMS. . . . 48
3.6 Linear model fitting and residual plots for variation of width error with camera translations. . . . 50

3.7 Linear model fitting and residual plots for variation of depth error with camera translations. . . . 51
3.8 Curvilinear model fitting and residual plots for variation of width error with camera rotations. . . . 52
3.9 Curvilinear model fitting and residual plots for variation of depth error with camera rotations. . . . 53
3.10 Proposed feature based calibration for the PTMS. . . . 57
3.11 Profile image with a representation of camera and world coordinates. . . . 58
3.12 Defect identification and measurement. . . . 59
3.13 Evaluation of feature based calibration for PTMS. . . . 60
3.14 Mean difference of width and depth measurements for two schemes. . . . 63
3.15 Cumulative density function (CDf) for scheme 1 and 2. . . . 64
3.16 Resilience over pixel noise. . . . 65
3.17 Resilience over pantograph vertical displacement (uplift). . . . 66
3.18 Resilience over pantograph angular displacement (yaw). . . . 66
3.19 Resilience over pantograph angular displacement (roll). . . . 66
3.20 Resilience over pantograph angular displacement (pitch). . . . 67
3.21 Evaluation procedure for stereo camera misalignment. . . . 69
3.22 Camera axes and rotations around X-Tilt, Y-Pan, Z-Roll. . . . 69
3.23 Object and their projected stereo images. . . . 70
3.24 Variation of 3D error versus camera misalignment in terms of pure translations 72
3.25 Variation of 3D error versus camera misalignment in terms of pure rotations . . 73
3.26 Variation of Total 3D error versus camera misalignment in terms of translations and rotations . . . 74
3.27 Variation of Total 3D error averaged over range of misalignment versus object sizes. . . . 75

4.1 Illustraing feature extraction process - detection, description and matching, between stereo pairs. . . . 80
4.2 Evaluation Pipeline . . . 85
4.3 Stereo images from various datasets low resolution 320x240. . . . 87
4.4 Illustration of epipolar geometry. . . . 88
4.5 Accuracy Vs Computational time. The post-fixes refers to the size of the images: L-low resolution (320x240), M- medium resolution (640x480), H- high resolution (1280x960) . . . 89
4.6 Feature extraction on blurred (radius level 5) stereo images from Tromsø dataset with wide lens of L-low resolution 320x240. . . . 91
4.7 Performance of feature extractors for simulation of blur levels over various resolutions . . . 92
4.8 Feature extraction on barrel distortion (level 40%) stereo images from Microsoft dataset of L-low resolution 320x240. . . . 93
4.9 Performance of feature extractors for simulation of distortion levels over various resolutions . . . 94

4.10 Feature extraction on noisy (15dB) stereo images from Tromsø dataset with narrow lens and L-low resolution 320x240. 95
4.11 Performance of feature extractors for simulation of noise levels over various resolutions 96
4.12 Scatterplots of matched feature points and 2D pixel error with 3D accuracy. . . 100
4.13 Experimental setup. 102
4.14 Cameras arranged in a circular configuration around the 3D model. 103
4.15 The 3D models used for the experiment. From each model, 50 stereo image pairs are generated, corresponding to various baselines. 103
4.16 The 2D error (Squared Sampson) based on epipolar constraint over varied baselines 105
4.17 Rotation error of stereo camera over varied camera baselines. 107
4.18 Translation error of stereo camera over varied camera baselines. 108
4.19 Penalty values for all feature extractors. 109
4.20 Mean 3D estimation error, categorized based on feature descriptors over varied camera baselines 110
4.21 Standard deviation of 3D estimation over varied baselines. 111
4.22 Mean 3D estimation error, categorized based on feature detectors (SIFT, SURF and BRISK) over varied camera baselines 112
4.23 Mean 3D estimation error, categorized based on feature detectors (ORB, KAZE, AKAZE) over varied camera baselines 113
4.24 System overview. 121
4.25 Process of outlier detection: outliers (solid), inliers (dotted) 121
4.26 Illustration of setup used by Microsoft to produce the multiview dataset. 124
4.27 Epipolar error (E_p) computed for three different methods 125
4.28 Re-projection error (R_p) computed for different algorithms 126
4.29 Execution time of various algorithms relative to $FullSIFT - RANSAC$. . . 127
4.30 Deduction of relationship between object distance (D) and the baseline distance between the cameras (B). 128

5.1 The extended 3D performance metric explained. 133
5.2 Experimental setup 134
5.3 Mean 3D error for different number of total feature correspondences. 135
5.4 Measure of camera rotation and translation error over various noise levels with different number feature points for 3 different camera baselines. 137
5.5 Measure of 3D rotation accuracy and 3D position accuracy over various noise levels with different number feature points for 3 different camera baselines. . . 138
5.6 Measure of 3D orthogonality over various noise levels with different number feature points for 3 different camera baselines. 139
5.7 Measure of 3D rotation accuracy and 3D position accuracy over various noise levels with different sparsity (dispersion of points in 2D space) and various camera baselines. 140
5.8 Measure of 3D orthogonality over various noise levels with different sparsity (dispersion of points in 2D space) and various camera baselines. 141

5.9 Comparing 2D error and 3D error variation with noise for given N=75 pts and Sparsity = 100% and three baselines. 143
5.10 SEM reconstruction with Mountains software. 145

List of Tables

1.1 Outlining research questions and hypotheses that reflects the problem statement 8

3.1 Model parameters estimated for translation 54
3.2 Model parameters estimated for rotations 55
3.3 Tolerances for camera misalignment, given the system inaccuracy limit as 0.5mm. 55
3.4 Reference measurements of defects of two pantograph types. 61
3.5 Single camera pose estimation algorithms and their description. 61
3.6 Absolute angular difference in degrees between CBC and FBC - scheme 1. . . 62
3.7 Kullback-Leibler Divergence values for total (width + depth) error. 65

4.1 Overview of the state-of-the-art feature extractors. 83
4.2 Quality - accuracy, reliability and execution time of 24 feature extractors, which provides practical recommendation for 3D applications(section 4.3.7). Here "Rotation" is the mean 3D rotational change (expressed in degrees) and "Position" is the mean 3D positional shift (expressed in model units) of all the estimation 3D unit vectors that represent a model in 3D space. 117
4.3 Comparing known and estimated camera rotational parameters. 127

Part I

Overview

Chapter 1

Introduction

The perception of depth is the natural ability for human eyes. As early as in 1838, Charles Wheatstone first explained the "Physiology of Vision" (Wheatstone, 1838): ... *the mind perceives an object of three dimensions by means of the two dissimilar pictures projected by it on the two retinæ*.... Since then, there have been many attempts to imitate such depth perceptive ability in the digital visual media. Creating depth illusions from photographic images have been studied since a long time by research communities such as photogrammetry and computer vision. This has led to interesting applications in various fields.

The 3D multimedia systems have grown big in terms of applicability, deployment and maintenance. The quality of such systems are measured by accuracy, resilience and speed efficiency. Commonly, all these systems capture the scene of interest using one or more cameras from different viewpoints. Then, the camera feeds from all the cameras, along with other relevant information (e.g., metadata, calibration information, etc.) is compressed and transmitted. At the receiving end, the information is decompressed, processed and rendered to a suitable display. Some of the interesting 3D applications are found in the areas of immersive visual communication systems, structure from motion systems, visual inspection systems, etc. Each of them have different requirements for quality of service (QoS), in terms of accuracy, robustness and execution time.

Immersive systems aim at providing a realistic experience through applications such as freeviewpoint rendering, telepresence etc. Freeviewpoint rendering (Min, Kim, Yun, and Sohn, 2009; Tanimoto, 2010) enables us to view a 3D scene freely from any viewpoint, by synthesizing new viewpoints from a limited number of views captured by multiple cameras. Telepresence enables interactivity between human-human or real-virtual scenes. This includes distributed orchestra[1], cisco telepresence[2], immersive virtual-reality environments for entertainment (Muhanna, 2015), system to aid medical surgery (Chen, Xu, Wang, Wang, Wang, Zeng, Wang, and Egger, 2015), etc. Structure-from-motion systems aim at reconstructing 3D structures from a large number of images taken by one or more cameras at various viewpoints. These systems exploit the structural geometric information hidden in multiple images. Interesting examples of building 3D structures from shared photo collection are 'Rome in a day' (Agarwal, Furukawa, Snavely, Simon, Curless, Seitz, and Szeliski, 2011) and 'Photo Tour'

[1]Open Orchestra - `http://openorchestra.cim.mcgill.ca`

[2]Cisco Telepresence - `http://www.cisco.com/c/en/us/solutions/telepresence/telepresence_video_teleconference.html`

(Goesele, Snavely, Curless, Hoppe, and Seitz, 2007; Snavely, Seitz, and Szeliski, 2006). Visual inspection systems aim at real 3D measurements using 2D images. Currently, there are several such systems in the market, which use 3D vision principles to measure the 3D structure, for example, Scorpion 3D stringer[3], industrial vision systems[4], etc.

1.1 Application Scenarios

Currently, new-age 3D multimedia systems that utilizes advanced 3D imaging capabilities are emerging. In this thesis, the focus is on few application scenarios that have variety of applications, various working distances and differs in the requirement of QoS.

VERDIONE is a research project that aims at creating a platform for mixed reality art performances. These performances are highly interactive, where the actors at remote location are projected on a real stage. Multiple cameras capture the actor in a remote location and a particular view is rendered in the performance location. One such experiment for the world opera performance was carried out in Tromsø in 2012[5]. More details about VERDIONE and its challenges are discussed in section 2.2.1.

BAGADUS is a research project that aims at developing an integrated system for soccer analysis. In this project, multiple cameras capture the soccer field and creates a panorama video. The players are tracked, and their analytic information is annotated in the panorama video. In this way, the manual analysis of the players is replaced by this advanced technology. A prototype of this system is currently deployed in Alfheim soccer stadium, Tromsø. More details about BAGADUS and its challenges are discussed in section 2.2.2.

PTMS The PantoInspect Train Monitoring System (PTMS) is a product developed by the company, ImageHouse - PantoInspect, Denmark. The PTMS inspects, detects and reports defects occurring on the pantographs of electric trains. These systems are installed on the bridges or at locations where the PTMS can scan the pantographs. When the train passes under the PTMS, a camera captures the laser lines projected on the top of the pantographs. The image is then analyzed for measuring the defects. More details about the PTMS and its challenges are discussed in section 2.2.3.

POPART is a research project that aims at on-set visualization and post-production of films. Nowadays, the production of films focus on visual effects by integrating virtual scenes and real actors. In such productions, the director wishes to see the output right away. POPART allows the flexibility for the director or any other technician to view the integration of virtual and real scenes, while the film shoot is in progress. POPART also provides on-set post-production capabilities, which save a lot of time for filmmakers. More details about POPART and its challenges are discussed in section 2.2.4.

SEMRECON represents a module of a software product (Spip - Mountains) co-developed by the companies, Image Metrology, Denmark and Digital Surf, France. This is an image

[3]Scorpion 3D Stringer - `https://scorpion3dstinger.com`
[4]IVS - `http://www.industrialvision.co.uk/applications/3d-vision-systems`
[5]World Opera Performance in Tromsø - `http://www.geistweidt.com/publications.html`

based analysis tool for metrology purposes. The SEMRECON aims at reconstructing the 3D surface of a sample using series of images produced by scanning electron microscope. More details about SEMRECON and its challenges are discussed in section 2.2.5.

In most 3D multimedia applications where 3D reconstruction from 2D images is involved, one of the important factors that decide the quality of the system is camera calibration, i.e., the knowledge of the camera geometry and motion (Pedersini, Sarti, and Tubaro, 1998). Camera geometry is characterized by intrinsic parameters (such as focal length, principle point, pixel aspect ratio and skewness), and the camera motion is characterized by its extrinsic parameters (such as spatial position and orientation, also known as camera pose). Estimating both intrinsic and extrinsic parameters of one or more cameras is essentially the primary step in the 3D reconstruction process. Traditionally, a checkerboard pattern of a known measurement is used mostly for intrinsic camera calibration and sometimes for extrinsic camera calibration. This is termed as Checkerboard Based Calibration (CBC).

1.2 Practical Challenges

Although the 3D imaging technology has evolved in various application areas, practical challenges in their deployment and maintenance still exist, especially in new-age advanced 3D imaging systems.

In **large space scenarios**, e.g., VERDIONE (section 2.2.1) or BAGADUS (section 2.2.2), it is impractical to carry out the traditional camera calibration process because an extremely large checkerboard is required to calibrate the cameras. To obtain proper calibration, images must capture a large part of the checkerboard pattern in various angles. The cameras in these scenarios are located at large distances from the center of the space (in this case, a performance stage or a soccer stadium). Therefore, for calibration, the checkerboard pattern to be placed in the center of the space needs to be of a considerably large size in order to cover the image view of the camera. In such spaces, the application such as freeview rendering, greatly depends on the accurate estimation of camera calibration parameters. In the cases where cameras are misaligned after calibration, a re-calibration process is required, and this is again a cumbersome process while a stage performance or a soccer event is taking place. This leads to using in-accurate camera parameters for freeview rendering and the related application, and consequently, a distorted view is created, which may cause viewing discomfort for the end-users. A solution to overcome such misalignment problems in image based 3D systems could probably be to use a structured light based approach. Kinect is a product based on structured light, used for 3D reconstruction. In outdoor situation such as the BAGADUS scenario, the infrared rays in the sunlight interferes with structured light and degrades the Kinect system. In indoor situation, Kinect has an operating range of maximum 10 meters, but the opera stage in VERDIONE scenario or the soccer stadium in BAGADUS scenario is too big an area for the Kinect system to handle. In order to cover a larger area, more than one Kinect systems can be used but then the interference between the structured light will become a challenge.

In **medium space scenarios**, such as PTMS (section 2.2.3) or POPART (section 2.2.4), traditional CBC process is possible. However, the challenge here is to handle the camera misalignment. There is a serious effect of camera misalignment in PTMS. The critical damage of carbon

strips on the pantograph needs to be accurately measured. The camera misalignment caused by vibrations or wind or physical force during deployment leads to wrong measurements, which might then fail to detect the critical damage. This compromises safety in railways. There are several incidents where the electric lines have been damaged due to the failure of identifying damages on the pantographs and replacing them. In order to avoid such a catastrophe and to obtain high accuracy measurements with camera misalignment, a re-calibration process is required. However, it is difficult or inconvenient. After the system is deployed, a checkerboard cannot be used for calibration, unless the train traffic on the tracks are suspended, causing huge delays in the public transport system. In POPART, the camera misalignment can have an impact on wrong estimation of position cues to place virtual objects in the scene. The effect of misaligned virtual objects is not an expected outcome of the movie production. Once a scene is shot with wrong camera parameters, it is difficult to correct them during post-production. Of course, the cameras could be re-calibrated, but with a cost of time to re-shoot the scene.

In **small space scenario** such as in the SEMRECON scenario (section 2.2.5), which reconstructs a surface from a series of micro/nano-scopic images, a direct way for traditional calibration is not available. Placing a checkerboard is not possible because its not a camera-based imaging system but rather a electron-based imaging system.

From all scenarios discussed, the main practical problem is either that the traditional CBC is not possible at all, or if it is possible, then the re-calibration process becomes cumbersome due to misalignment of cameras and has a cost on safety, usability, reliability and time. In situations where CBC becomes impractical or inconvenient to use, various types of markers could be placed in the scene at known positions. The cameras in the scene register these markers' positions, and in turn, estimate each camera's relative position and orientation with respect to other cameras. This type of calibration is termed as Marker Based Calibration (MBC). Sometimes, it is not convenient to use markers within the scene, as it might disturb the scene setting, e.g., identifiable markers place on a opera stage for VERDIONE scenario, or placing at the target of microscope for SEMRECON scenario.

In order to overcome these problems, it is required to focus on an alternative way to calibrate the cameras and that is Feature Based Calibration (FBC). Unlike other techniques, feature based calibration estimates the camera calibration parameters based on point features extracted in the images instead of known positions of checkerboard corners (CBC) or any identifiable pattern in the scene (MBC).

The FBC is a process to estimate the camera parameters based on matches of interesting points, i.e., feature correspondences in two or more images. The FBC is comprised of the following steps: *Feature Extraction* and *Pose Estimation*. In the feature extraction step, the interesting points from two or more images are detected and matched against each other to obtain feature correspondences. In the pose estimation step, the feature correspondences are used to estimate the relative position and orientation of a camera with respect to another. The estimated camera parameters are used to define the scene geometry, and its accuracy has an effect on 3D reconstruction. For example, Scale Invariant Feature Transform (SIFT) (Lowe, 2004) is one of the state-of-the-art feature detectors known for obtaining feature correspondences between stereo camera pairs. The SIFT detector has been used as a feature extractor in the context of FBC (Liu, Zhang, Liu, Xia, and Hu, 2009).

1.3 Goal and Scope

In order to achieve high quality applications in new-age 3D systems and overcome the practical challenges, highly accurate and robust FBC is required. Hence, the goal of this thesis is to explore the adoption of FBC in 3D systems and assess the accuracy and robustness against practical challenges.

In this thesis, the focus is specifically on application scenarios such as mixed reality art performance (VERDIONE), soccer player tracking (BAGADUS), pantograph train monitoring system (PTMS), movie production on-set real-time tracking (POPART) and scanning electron microscopic image reconstruction (SEMRECON). A detailed view of these application scenarios is explained in section 2.2.

Although each of the application scenarios have different use cases, this thesis focuses on the common underlying concept, i.e., 3D reconstruction. Therefore, the entire thesis refers to a *3D multimedia system*, which is capable of reconstructing 3D data based on single or stereo images. Typically, such systems comprise both hardware and software units for scene acquisition, image processing, object reconstruction and 3D display and are referred to as *Image Based 3D Systems* (details in section 2.1).

1.4 Problem Statement

In accordance with the goal of the thesis, the accuracy and robustness of FBC on 3D multimedia systems were evaluated for several variations in scene properties and camera properties. Thus, the main question worth exploring was:

What are the challenges in designing FBC to achieve high accuracy and robustness against practical issues in 3D multimedia systems?

As a part of the scientific method, the approach of research is driven by one or more testable hypothesis. For a research question, a hypothesis is a tentative statement that might include a possible explanation or prediction of an answer, in the form of relation between the variables that change in the experiment and observations in the experiment. A null hypothesis[6] (H_0) represents that a change in the variables has an no effect on the observations of the experiment.

The main question above, was further categorized into the following research scopes: 3D systems, feature extraction and pose estimation. Within each research scope, relevant questions were posed and null hypotheses (H_0) were formulated to help guide the research.

1. *3D multimedia systems*:

   In 3D systems, the cameras are prone to misalignment due to several factors, i.e., the system deployment might not be sturdy or the system can be physically disturbed by human intervention or natural causes (e.g., by wind). Camera misalignment refers to the change in the rotation and translation of the camera with respect to its axis (single camera

[6]Null hypotheses can be proven/disproven wrong by an example, which allows us to make general statements, whereas we are unable to prove the opposite, that certain errors always have negative effects. In fact, we are certain that the opposite statements cannot be made so broadly.

system) or another camera (stereo camera system). Due to the camera misalignment, the 3D system that uses the initially calibrated camera parameters will then yield wrong results. Hence, the quality of such 3D systems is affected by the camera misalignment. The significance of the effect of the camera misalignment is determined by an acceptable reconstruction error in 3D systems, which is entirely specific to the application.

From the complete 3D system point of view, the following questions were posed:

- What is the effect of single/stereo camera misalignment on the quality of 3D system?

  It is important to study the effects in order to quantify the relationship between 3D error and the camera misalignment and thereby find the limits/tolerances of camera misalignment for an acceptable error for a specific application. This helps in designing better 3D systems that provides a good user experience for 3D reconstruction application and ensures safety for PTMS application scenario.

  Hypothesis I:

  H_0 : The 3D reconstruction accuracy has insignificant effect when the camera is misaligned.

- Can FBC ensure a good online re-calibration capability in comparison to CBC, in a single camera 3D system?

  In certain applications, the system deployment is such that the online re-calibration process becomes impossible with traditional CBC techniques, especially PTMS. In this case, it becomes important to understand the impact of using FBC on a single camera system in terms of quality comparable to CBC techniques. This helps in designing a good quality 3D system that is robust to any physical disturbance that causes camera misalignment.

  Hypothesis II:

  H_0 : The accuracy and robustness of 3D reconstruction has an insignificant effect, when the 3D system replaces CBC by FBC techniques.

2. *Feature extraction*:

In 3D stereo systems that adopts FBC, the feature extraction is one of the processes that comprises feature detection, description and matching process between stereo images. Therefore, the accuracy of feature matching has an effect on the the quality of 3D reconstruction. This builds up the curiosity to find out what are really the "good" features for FBC. Hence, characterizing the state-of-the-art feature extractors for changes in camera properties or scene properties is necessary. The significance of the performance of feature extractors are determined by an acceptable 3D reconstruction error that is specific to applications.

With a view to characterize the feature extractors, the following questions were posed:

- How does the quality and robustness of feature extractors vary with intrinsic and extrinsic camera properties?

Here, the feature extractors are characterized for changes in camera properties, i.e., internal (blur, lens distortion, resolution and noise) and external (camera baseline - relative camera displacement in stereo systems) camera properties. The change in both internal and external camera properties has an impact on the quality of feature correspondence matching in stereo systems for the discussed scenarios. This helps system builders to make a better choice of feature extractor suitable for their applications. On the other hand, it is also possible to design the camera density based on the choice of feature extractor.

Hypothesis III:

H_0 : The performances of the state-of-the-art feature extractors have insignificant differences to the change in intrinsic and extrinsic camera parameters.

- How to characterize SIFT features used for FBC in 3D systems, for wide baseline setup?

  SIFT, being a popular feature extractor, has a limitation of rotational invariance between stereo pairs for upto 30 degrees. However, a wider baseline is suitable for certain applications, especially in large space scenarios. So, an exploration of SIFT feature extractor for a wide baseline setup is necessary. This helps in achieving the same quality with reduced number of cameras and thereby also saves the cost of the system, in terms of storage, transmission and processing of multiple images.

  Hypothesis IV:

  H_0 : Accuracy of SIFT features for wide baseline FBC is maintained at an acceptable level, only up to 30 degrees angular separation between stereo cameras.

3. *Pose estimation*:

In stereo systems, the pose estimation determines the camera rotation and translation relative to another camera. In FBC, the pose estimation quality is affected by the feature correspondence matching quality. Hence, for a given set of feature correspondences, it becomes necessary to explore the characteristics of pose estimation based on attributes of the matched feature points. The attributes of the matched features are noise in the feature matching and the distribution of matched features in 2D space, which affects the pose estimation.

The significance of the error in pose estimation is determined by an acceptable 3D reconstruction error that is specific to the applications.

With this point of view, the following question was posed:

- How does the attributes of the matched feature points, i.e., noise and sparsity in a 2D space, affect the quality and robustness of pose estimation?

  This study is relevant to understand the influence of noise on the quality of pose estimation, which helps in underlining the noise limits of the feature matches for

good pose estimation. This study also explores if the selection of feature matches based on their sparsity affects the pose estimation. This helps in better selection of feature matches for robust 3D systems.

Hypothesis V:

H_0 : The pose estimation accuracy has an insignificant change with the increase in noise and sparsity of matched feature points.

Table 1.1 provides a summary and an outline of the research questions posed, problem areas identified and the corresponding hypotheses framed. The research on each of the scopes: 3D systems, feature extraction and pose estimation are presented as separate chapters in this thesis.

MAIN RESEARCH QUESTION:			
What are the challenges in designing FBC to achieve high accuracy and robustness in 3D systems?			
No.	Category	Resesarch Question	Hypothesis
1	3D Systems	What is the effect of single/stereo camera misalignment on the quality of 3D system?	H_0 : The 3D reconstruction accuracy has insignificant effect when the camera is misaligned.
2	3D Systems	Can FBC ensure a good online recalibration capability in comparison to CBC in a single camera 3D system?	H_0 : The accuracy and robustness of 3D reconstruction has an insignificant effect, when the 3D system replaces CBC by FBC techniques.
3	Feature Extraction	How does the quality and robustness of feature extractors vary with intrinsic and extrinsic camera properties?	H_0 : The performances of the state-of-the-art feature extractors have insignificant differences to the change in intrinsic and extrinsic camera parameters.
4	Feature Extraction	How to characterize SIFT features used for FBC in 3D systems, for wide baseline setup?	H_0 : Accuracy of SIFT features for wide baseline FBC is maintained at an acceptable level, only upto 30 degrees angular separation between stereo cameras.
5	Pose Estimation	How does attributes of the matched feature points, i.e., noise and sparsity in 2D space affect the quality and robustness of pose estimation?	H_0 : The pose estimation accuracy has an insignificant change with the increase in noise and sparsity of matched feature points.

Table 1.1: Outlining research questions and hypotheses that reflects the problem statement

1.5 Research Methods

A new intellectual framework for the discipline of computing was presented in the final report of ACM Task Force on the Core of Computer Science (Comer, Gries, Mulder, Tucker, Turner, and Young, 1989). This report provides a detailed description on how research can be organized based on three major paradigms: Theory, Abstraction and Design. Each of them has roots in different areas of science, although all can be applied to computing. It is stated that all paradigms are so intricately intertwined that it is irrational to say that any one is fundamental in the discipline.

The three paradigms to computer science are defined as follows:

- The *Theory paradigm* is rooted in mathematics. First, it specifies the objects of study and hypothesizes relationships between the objects. Then, the hypothesis is proven logically.

- The *Abstraction paradigm* is rooted in experimental scientific method. A scientist forms a hypothesis, constructs a model, makes a prediction before designing an experiment. Finally, data is collected and analyzed.

- The *Design paradigm* is rooted in engineering. A scientist states requirements and specifications, followed by design and implementation of the said system. Finally, the system is tested to see if the stated requirements and specifications were met.

This thesis followed the *Abstraction paradigm* and was initiated by formulating a research question, in order to explore and scientifically assess the accuracy and robustness of FBC against practical challenges in the current and new-age 3D applications. Accordingly, relevant hypotheses was framed in all research scopes, i.e., 3D systems, feature extraction and pose estimation. To test the hypotheses, relevant experimental evaluations were conducted focusing on real scenarios which relates to real systems in use and potential new age application scenarios that is under research.

A currently deployed system such as PTMS was considered to determine the effects of camera misalignment and adoption of feature based calibration. The VERDIONE, BAGADUS and POPART scenarios motivated the characterization of feature extractors and pose estimation for all practical challenges related to change in the camera's internal and external properties and scene properties. The SEMRECON scenario was used for validating the methodology of 3D reconstruction. Finally, recommendations for operational ranges for feature extraction and pose estimation were provided, and their practical implications to improve the robustness of image based 3D systems were discussed.

1.6 Main Contributions

In order to test the stated hypotheses, the practical problems of the application scenarios (mentioned in section 1.2) were explored by assessment of accuracy and robustness of FBC through experiments. Here, "system users" refers to system developers, builders, designers or researchers.

The main contributions of this thesis are as follows:

1. A statistical tool was developed for single camera 3D systems to determine the mechanical tolerances of the camera rigs that minimize the camera misalignment error in the PTMS. This helps to improve the robustness to practical error such as camera misalignment.

2. Feature based calibration was adopted in the PTMS by replacing the traditional checkerboard calibration, to improve the flexibility and maintainability of the PTMS without manual intervention. This also helps to improve robustness to practical error, such as pantograph misalignment and image analysis error of the PTMS.

3. The adverse effects of camera misalignment in stereo 3D applications were exhibited. This helps system users to build stable camera rigs to improve the accuracy of the 3D system by restricted erroneous camera misalignment in application scenarios such as VERDIONE, BAGADUS and POPART.

4. The state-of-the-art feature extractors (SIFT, SURF and ORB) were characterized and their operating limits were determined in the presence of image defocus, lens distortion and sensor noise, at various resolutions in VERDIONE and BAGADUS like application scenarios. This helps the system users to choose a feature extractor based on the requirement for accuracy, execution time and robustness.

5. The state-of-the-art feature extractors (SIFT, SURF, ORB, KAZE, AKAZE, MSER, BRISK, FAST, STAR, BRIEF, FREAK) were characterized and design considerations were recommended for using state-of-the-art feature extractors at different camera baselines (angular displacement between the stereo pair) using virtual dataset that mimics POPART, VERDIONE or BAGADUS like application scenarios. The design considerations were based on the 3D accuracy, deformation of 3D object and execution time. This helps system users to choose a feature extractor based on design parameters. This also helps system users to determine the camera density required to capture the scene.

6. A new algorithm - $NewSIFTcalib$ was proposed, which modified the existing SIFT to yield better accuracy and computation time, especially in wide baseline camera setups. This helps to improve the usability and scalability of 3D multiview capture systems. This also helps to reduce the camera density for capturing the scene and thereby is cost effective (in terms of storage, transmission and processing of multiple images) for VERDIONE and BAGADUS like application scenario.

7. The state-of-the-art pose estimation algorithm was characterized and the camera baselines and feature selection criteria were recommended to minimize noise in the feature correspondences of a stereo pair and thereby maximize the 3D accuracy. The experiments were carried out using virtual dataset that mimics VERDIONE or POPART application scenarios. The effect feature selection based on the sparsity of feature correspondences on the 3D accuracy was validated using SEMRECON scenario. This study helps system users to make a choice of camera baseline and a subset of feature correspondences and improve the robustness of pose estimation.

1.6.1 Publications

Here, is an outline of all the papers and posters published. This thesis is composed of all these publications rearranged and with the extended work that is yet to be published.

Refereed Proceedings

MMSys 2016 *Robustness of 3D Point Positions to Camera Baselines in Markerless AR Systems*. Deepak Dwarakanath, Carsten Griwodz and Pål Halvorsen. In Proceedings of the 7th International Conference on Multimedia Systems (MMSys), 2016 (more details in chapter 11).

ICVS 2015 *Online Re-calibration for Robust 3D Measurement Using Single Camera-PantoInspect Train Monitoring System*. Deepak Dwarakanath, Carsten Griwodz, Pål Halvorsen and Jacob Lildballe. In Proceedings of the International Conference on Computer Vision Systems (ICVS), 2015 (more details in chapter 10).

SETN 2014 *Study the Effects of Camera Misalignment on 3D Measurements for Efficient Design of Vision-Based Inspection Systems*. Deepak Dwarakanath, Carsten Griwodz, Pål Halvorsen and Jacob Lildballe. In Proceedings of the 8th Hellenic Conference on Artificial Intelligence (SETN), 2014 (more details in chapter 9).

IVCNZ 2012 *Evaluating Performance of Feature Extraction Methods for Practical 3D Imaging Systems*. Deepak Dwarakanath, Alexander Eichhorn, Pål Halvorsen and Carsten Griwodz. In Proceedings of the 27th International Conference Image and Vision Computing New Zealand (IVCNZ), 2012 (more details in chapter 8).

DICTAP 2012 *Faster and More Accurate Feature-Based Calibration for Widely Spaced Camera Pairs*. Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz and Pål Halvorsen. In Proceedings of the Digital Information and Communication Technology and it's Applications (DICTAP), 2012 (more details in chapter 7).

Poster Presentations

VERDIKT 2012 *Multiple Camera Arrays for Real-time 3D Rendering Systems*. Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz and Pål Halvorsen. In VERDIKT 2012, Norwegian Research Council, Oslo, Norway (BEST POSTER AWARD 2012) (more details in chapter 14).

VERDIKT 2010 *3D Multi-view Acquisition and Rendering System*. Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz and Pål Halvorsen. In VERDIKT 2010, Norwegian Research Council, Oslo, Norway (BEST POSTER AWARD 2010) (more details in chapter 13).

VERDIKT 2009 *3-D Video Processing for Mixed Reality Art Performances*. Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz and Pål Halvorsen. In VERDIKT 2009, Norwegian Research Council, Oslo, Norway (BEST POSTER AWARD 2009). (more details in chapter 12).

1.6.2 Software Development

The source code for all the implementation and testing are as follows:

- Libfacs: Library for feature based auto-calibration for multiple camera array systems. The functionality involves feature detection, feature description, feature matching, intrinsic camera calibration, pose estimation and other mathematical utilities.[7].
- 3DMars- 3D multiview acquisition and rendering system, includes evaluation test implementation[8].
- Matlab Projects- contains evaluation test implementation[9].

All programming implementation was carried out using the following tools:

- C++, Object oriented programming language[10].
- MATLAB, Matrix laboratory toolbox[11].
- OpenCV, Open source library for computer vision library[12].
- OpenFrameworks, Open source C++ toolkit for creative coding[13].
- OpenMVG, Open source multiview geometry library[14].
- NorthLight, VERDIONE project related multimedia library[15].
- OpenGL, Open source graphics library[16].
- VLFeat, Open source library for computer vision[17].
- Magick++ Image processing library[18].

1.7 Limitations

The accuracy and robustness of FBC for 3D reconstruction were tested on real datasets, i.e., PTMS and to some extent SEMRECON. However, the experimentation using real datasets was not achieved for VERDIONE, BAGADUS and POPART scenarios, i.e., instead virtual datasets were used. It was in a way better to use virtual datasets, because of ground-truth values that was

[7]Libfacs - `https://bitbucket.org/deepakdnath/facs`
[8]3DMars - `https://bitbucket.org/mpg_code/3dmars`
[9]Matlab Implementation - `https://bitbucket.org/mpg_code/matlab-projects`
[10]C++ - `http://www.cplusplus.com`
[11]Mathworks - `https://www.mathworks.com`
[12]OpenCV - `http://opencv.org`
[13]OpenFrameworks - `http://openframeworks.cc`
[14]OpenMVG - `https://github.com/openMVG`
[15]Verdione - `http://verdione.wiki.ifi.uio.no/Main_Page`
[16]openGL - `https://www.opengl.org`
[17]VLFeat - `http://www.vlfeat.org`
[18]Magick++ - `http://www.imagemagick.org/Magick++`

useful for testing. The virtual datasets were generated to mimic the application scenarios with a specific focus on the foreground objects, but for scenarios such as VERDIONE, BAGADUS and POPART, a scene capture involves background with a large depth of field.

A real dataset with background scene and ground-truth values from large space applications (VERDIONE and BAGADUS) was not available and although POPART had developed a system that considered the background, the current algorithms used for 3D reconstruction were not capable of handling the large depth reconstruction.

Therefore, the experimentation was limited to virtual datasets that does not consider the textured background of the scene. This could potentially be the further scope for exploration.

1.8 Thesis Outline

The thesis is outlined chapter-wise as follows:

Chapter 1: Introduces practical challenges in new age 3D applications and discusses the problem statement and the main contributions of this thesis.

Chapter 2: Discusses preliminary concepts so that it provides sufficient background for further discussions in the thesis. This includes brief introduction to conceptual and mathematical understanding of image based 3D systems and application scenarios that motivates this thesis.

Chapter 3: Describes the camera misalignment effects on single or stereo camera system and adoption of FBC in the PTMS system.

Chapter 4: Describes the robustness of state-of-the-art feature extractors and explains a proposed feature extractor for wide baselines.

Chapter 5: Explores the robustness of pose estimator against noise, number of features and selection of features in stereo pair feature matches

Chapter 6: Summarizes and concludes the thesis work and presents new ideas and concepts for further work.

Chapters 7 - 11: Contain each of the included publications.

Chapters 12 - 14: Contain each of the included posters.

This thesis is organized by topics, but the time of the experiments is not necessarily in the same order. So, in some cases, new features were learned and included in the later experiments.

Chapter 2

Preliminary Concepts

This chapter is an introduction to the preliminary concepts about image based 3D systems that provides necessary mathematical and conceptual background to understand the technical aspects of this thesis. Image based 3D systems are explained in terms of basic concepts, procedures and operations related to multiview geometry. Next, the system architecture and practical challenges in the application scenarios, used in this thesis, are explained in detail.

2.1 Image Based 3D Systems

An illustration of a typical image based 3D system as shown in figure 2.1 provides a platform to discuss the details of this research work. A single or multiple cameras are extensively used as an integral part of current or new age 3D systems in the field of vision based inspection, mixed reality art performance, sports analytics, augmented reality and image metrology. All these systems operate on camera images and focus on rendering 3D data, especially the focus is on 3D reconstruction in the form surface reconstruction, full volume 3D reconstruction, free rendering, augmenting virtual objects etc. Hence, they are usually referred as *image based 3D systems*.

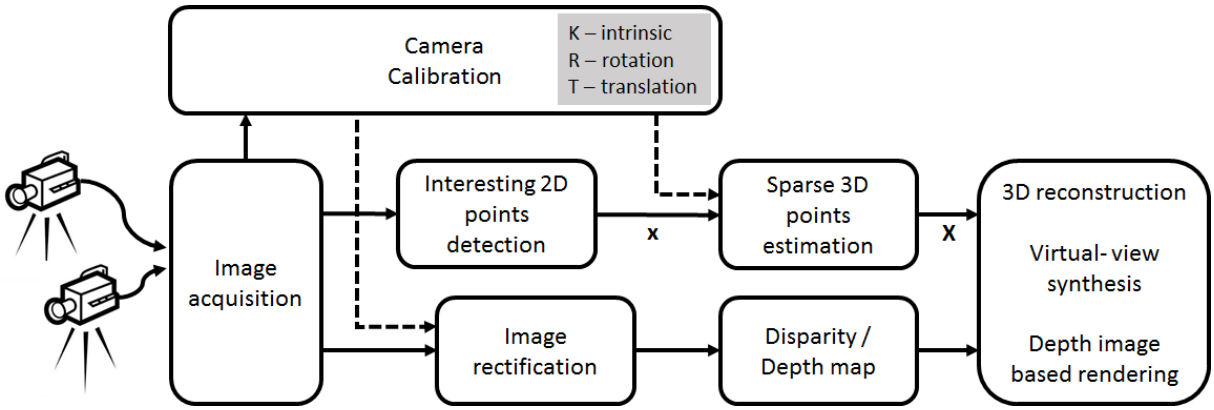


Figure 2.1: Typical 3D system illustrating two different workflows for 3D applications, using the knowledge of camera calibration.

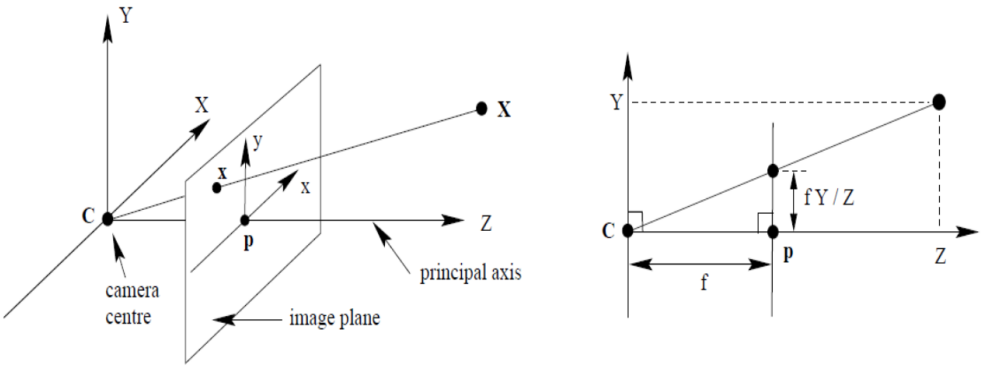


Figure 2.2: Pin-hole camera model. Courtesy: Multiview Geometry (Hartley and Zisserman, 2004).

2.1.1 Camera Calibration

Camera Model

One of the most simplest model that one has come across is the pin-hole model for the camera, which is as shown in the figure 2.2. The camera center with center at C, captures an image of 3D object on the image plane. The line originating from C, which is perpendicular to image plane is the principal axis. Point p is the principal point, which is the intersection of principal axis and the image plane. According to the pinhole camera model, the 3D point X in space is projected on the image at point x, where the line joining point X to the center of projection C meets the image plane.

In this ideal pinhole camera, the size of an image is related to the real three-dimensional object by the focal length, which is the distance between C and p. Figure 2.2 shows the focal length f, the height of an object X in space and the height of the same object x projected on the image plane. Using similar triangles, the mapping between Euclidean 3D space and Euclidean 2D space is represented as in equation 2.1.

$$x_i = f\frac{X_w}{Z_w}, y_i = f\frac{Y_w}{Z_w} \tag{2.1}$$

The 3D point is expressed in world coordinates (X_w, Y_w, Z_w) with origin in space. The same point can also be expressed in camera coordinates (X_c, Y_c, Z_c) with origin at the camera center C. The corresponding image coordinates are expressed in image coordinates (x_i, y_i) with origin at top-left corner of the image.

Extrinsic Parameters

The extrinsic parameter constitute rigid geometrical transformation of an object in 3D space. The mapping between world coordinate and camera coordinate can be represented by a rigid transformation that involves 3-by-3 rotation (R) and 3-dimension translation (T) of the coordinate system. If the mapping is represented in homogeneous space (Hartley and Zisserman,

2004), then the mapping can be expressed as a linear relationship as in equation 2.2. Here, $[R|T]$ constitute the *camera extrinsic parameters*.

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \underbrace{\left[\begin{array}{ccc|c} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{array}\right]}_{[R|T]} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{2.2}$$

Intrinsic Parameters

The internal camera parameters constitute geometrical relationship between the camera center, the image plane and physical properties of the pixels on the image sensor .

Focal Length is the distance from the center of the lens to the image plane along the optical axis. This defines the size of the object project onto the image plane.

Principal Point is intuitively thought to be located at the center of the image. This is rarely the case in real cameras, since that would require the image sensor to be perfectly positioned in the manufacturing process of the camera. In addition, it is common to place the origin of the image coordinate system at the top left corner in image/video processing systems. Equation 2.1 is rewritten with an offset p_x and p_y along the x-axis and the y-axis respectively, as in equation 2.3

$$x_i = f\frac{X_w}{Z_w} + p_x, y_i = f\frac{Y_w}{Z_w} + p_y \tag{2.3}$$

Pixel Aspect Ratio is the ratio of the width and the height of a physical pixel on an image sensor. Most image sensors have square pixels, although some system use rectangular pixels. To account for this, the focal length is scaled with the width (s_x) and the height (s_y) of the pixel. Equation 2.3 is then rewritten as equation 2.4.

$$x_i = s_x.f\frac{X_w}{Z_w} + p_x, y_i = s_y.f\frac{Y_w}{Z_w} + p_y \tag{2.4}$$

Skewness is similar to pixel aspect ratio which compensates for non-square pixels. The skewness of the pixel width and height is compensated based on the skewness factor (s).

The mapping between camera coordinates to image coordinates involves focal length (f) and principle point offset (p_x, p_y), pixel aspect ratio (α_x, α_y) and skewness (s). From pinhole equation 2.1, this mapping between the camera coordinates and image coordinates is deduced, as in equation 2.5. Here, K is the *camera intrinsic matrix*.

$$\begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \underbrace{\begin{bmatrix} \alpha_x f & s & p_x \\ 0 & f\alpha_y & p_y \\ 0 & 0 & 1 \end{bmatrix}}_{K} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \frac{1}{Z_c} \tag{2.5}$$

Calibration

By combining the equations 2.2 and 2.5, the linear relationship between 3D object points and 2D image points can be expressed as in equation 2.6.

$$\begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = K[R|T] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \frac{1}{Z_c} \tag{2.6}$$

The process of estimating these camera intrinsic and extrinsic parameters is known as *Camera Calibration*. Typically, the cameras are calibrated using a checkerboard target (Bouguet, 2008; Tsai, 1992; Zhang, 2000). Here, it is assumed that both 3D and 2D coordinates are known. Hence using a checkerboard is convenient and the corners are considered as 3D reference points.

Lens Distortion

The use of a lens adds distortions to the image, in particular radial lens distortion and tangential lens distortions. Complex systems of lenses are sometimes used (in more expensive cameras) to minimize the distortion, but any real camera system has distortions from the use of lenses. The lens distortion needs to be compensated as well, for high quality 3D reconstruction.

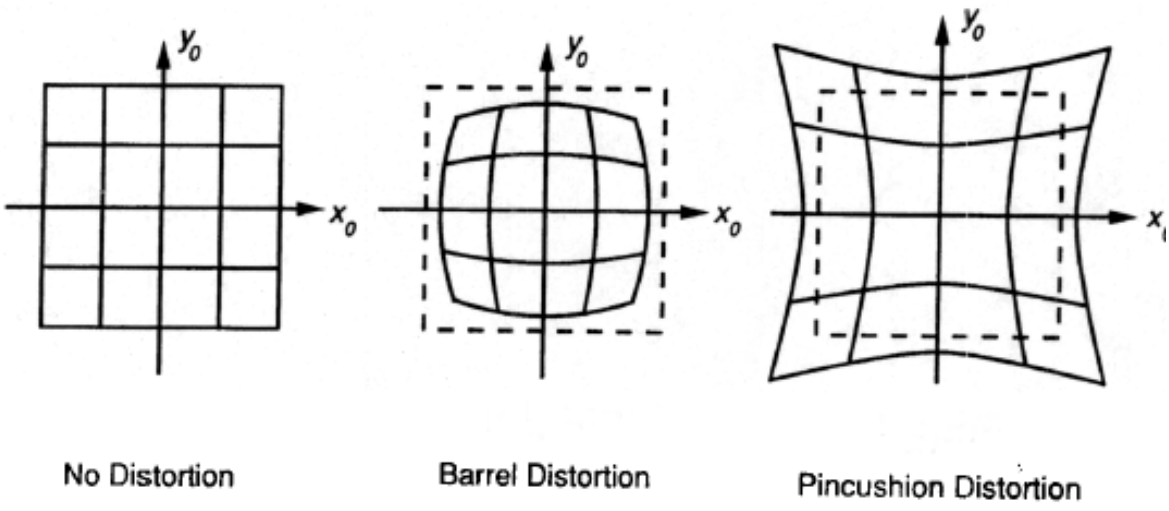


Figure 2.3: Radial lens distortion. Undistorted image (left), barrel distortion (center), pincushion distortion (right). Courtesy: University of Cologne[2].

Radial lens distortion comes from the spherical shape of the lens which causes the light to bend. The distortions extends outwards from the center of the lens (hence the name radial). In practice, the lens distortion affects the straight lines in the real world to be mapped as curved

[2]Courtesy: `http://www.uni-koeln.de/~al001/radcor_files/hs100.htm`
[3]Courtesy: `http://www.flickr.com/photos/riseriyo/4558440101/`

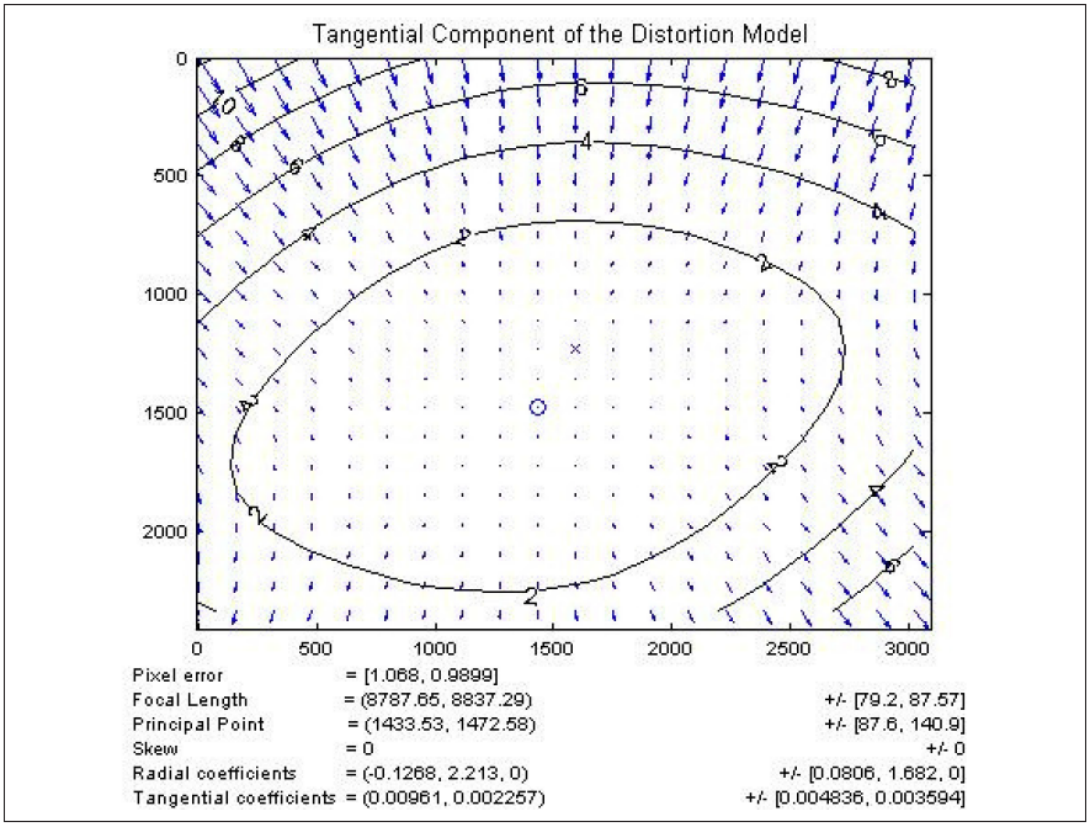


Figure 2.4: Tangential lens distortion. Courtesy: Public photo collection[3].

lines in the image (see figure 2.3). However, this distortion can be corrected so that the pinhole camera model still remains valid. The corrected coordinates of each pixel is then given by equations 2.7 and 2.8 where, x_u and y_u are the undistorted pixel positions, x_d and y_d are the distorted pixel positions in the x and y direction respectively, k_1 and k_2 are the radial distortion coefficients and r^2 is the normalized radial distance from the principal point.

$$x_u = x_d(1 + k_1 r^2 + k_2 r^4) \tag{2.7}$$

$$y_u = y_d(1 + k_1 r^2 + k_2 r^4) \tag{2.8}$$

Tangential lens distortion appears when the lens is not parallel to the image sensor (see figure 2.4). The corrected coordinates for tangential distortion is given by equations 2.9 and 2.10 where, p_1 and p_2 are the tangential distortion coefficients.

$$x_u = x_d + 2p_1 y_d + p_2(r^2 + 2x_d^2) \tag{2.9}$$

$$y_u = y_d + p_1(r^2 + 2y_d^2) + 2p_2 x_u \tag{2.10}$$

Epipolar Geometry

Epipolar geometry is the geometry between two views, which encapsulates both the intrinsic and extrinsic parameters of a stereo camera pair. Consider a point P in 3D space projected through the centers of projection O_L and O_R onto the image planes at position p_L and p_R as

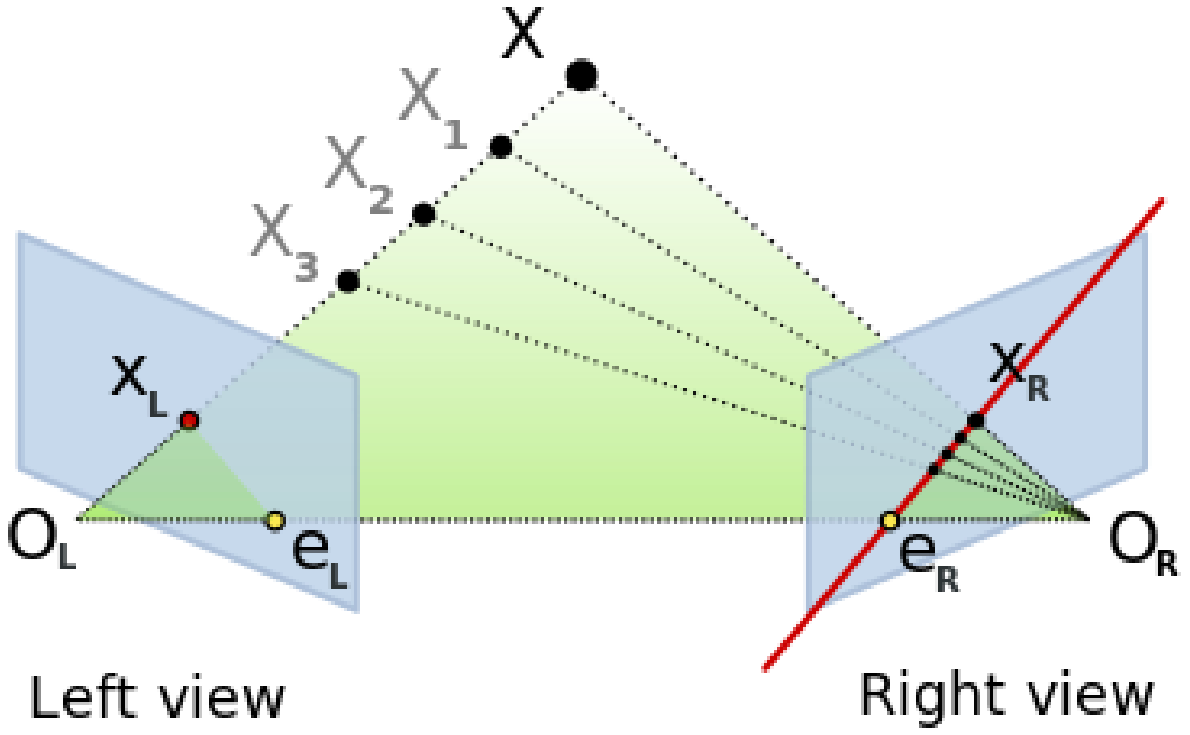


Figure 2.5: Epipolar geometry for stereo camera setup. Courtesy: Wikipedia[4]

shown in figure 2.5. Various terms defining the epipolar geometry are as follows:

Epipolar plane The plane, which is determined by a three-dimensional point X and the two camera centers O_L and O_R.

Epipolar line The line where the epipolar plane intersects the image plane, e.g., the line joining e_L and X_L, or e_R and X_R.

Baseline The line between the camera centers O_L and O_R.

Epipole The point where the baseline intersects the image plane, e_L and e_R. This is also the same as the projection of one camera center onto the other image plane.

Feature correspondences The projection of 3D point X on two cameras with non-coincident camera centers, i.e., X_L and X_R.

Given the point X_L, the corresponding point X_R must lie on the epipolar line, this is known as the *epipolar constraint*.

[4]Courtesy: Wikipedia, Epipolar Geometry - `http://en.wikipedia.org/wiki/Epipolar_geometry`.

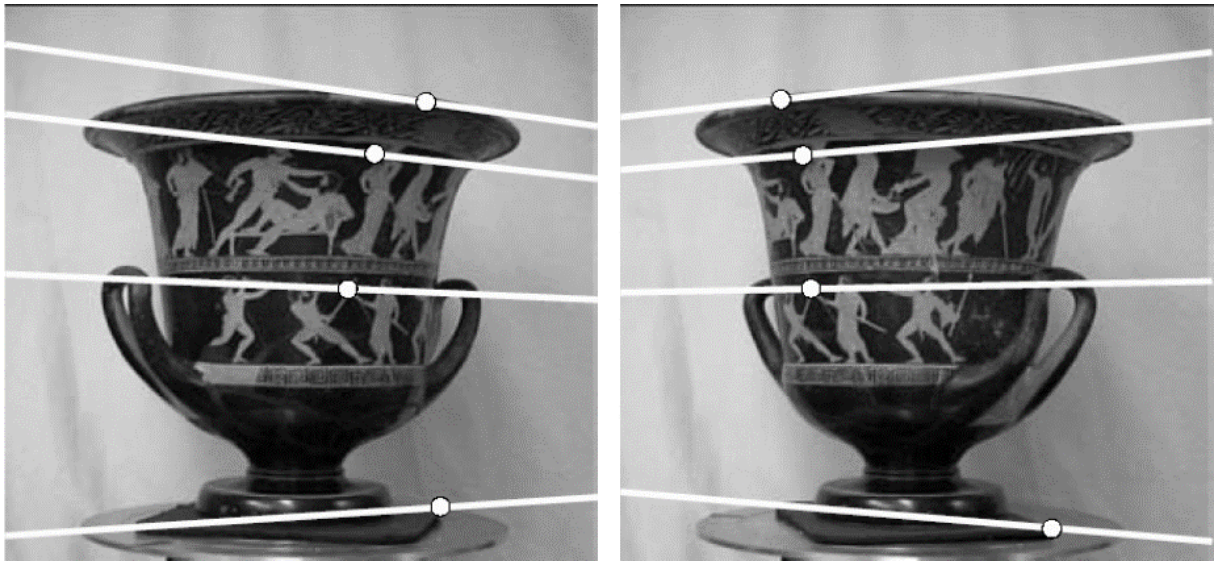


Figure 2.6: Feature point correspondences in stereo images. Courtesy: Multiview Geometry (Hartley and Zisserman, 2004).

Fundamental Matrix

The epipolar geometry of two views is algebraically represented by the fundamental matrix. The fundamental matrix maps a point in one image to it epipolar line in another image. This matrix encapsulates the camera internal parameters and general motion.

Mathematically, the fundamental matrix is represented as a homogeneous 3x3 matrix of rank 2 with 7 degrees of freedom. Given a set of feature correspondences between two camera views, i.e. x_i and x'_i, the fundamental matrix F satisfies the condition as in equation 2.11.

$$x_i F x'_i = 0, \forall i. \tag{2.11}$$

Every feature correspondence generates a linear equation with unknown entry in F. Therefore, F can be estimated using known values of feature correspondences. One such method to compute F is 8-point algorithm (Hartley and Zisserman, 2004), where atleast 8 feature correspondences are used to solve the linear equations based on least squares.

An example of stereo camera system that is dictated by epipolar geometric principles are as shown in the figure 2.6. Feature points are detected in both image and matched with each other. Every point in one image corresponds to a point in another image. Using these feature correspondences, F matrix is estimated. Normally for calibrating a stereo camera without any calibration target, the feature correspondences are used to estimate the F matrix. Further, there are different procedures in either calibrated (known camera intrinsic) and uncalibrated (unknown camera intrinsic) case.

For calibrated case (camera intrinsic K is known), similar to F, an Essential matrix E is estimated as $E = K'FK$. Form E matrix camera pose, i.e., rotation and translation is estimated. Once the camera pose is known, for given image feature correspondences, the corresponding 3D points can be obtained using triangulation.

For uncalibrated case (camera intrinsic K is unknown), a rectification homography is estimated using F and epipole points (details explained in (Hartley and Zisserman, 2004)). This

homography encapsulates the camera parameters and therefore, 3D reconstruction is carried out based on disparity and depth computation.

2.1.2 3D Reconstruction

A typical Image based 3D system is illustrated as shown in figure 2.1. These systems are installed with one or more cameras fixed onto a rig or a solid fixture to capture the scene. The cameras must be hardware synchronized so that they capture the images at the same time. This is very important for any further processing on these images.

Sparse 3D Workflow

In a single camera system, interesting points that are required to be reconstructed in 3D are detected and back-projected using the camera calibration parameters to obtain corresponding 3D points in world space.

In stereo or multiple camera setup every stereo pair is used for detecting feature correspondences in stereo pair of images. Each feature corresponding point is back-projected, using their respective camera calibration parameters (Hartley and Zisserman, 2004)), to obtain 3D sparse points. This process is called *Triangulation*. In this case, it is required to know the camera intrinsic parameters.

Dense 3D Workflow

Based on the feature correspondences, a fundamental matrix (F) is estimated, which gives a geometrical relationship between the stereo pair. The F matrix encapsulates the camera intrinsic matrix.

The images can be transformed onto a common synthetic image plane (i.e., the images are coplanar), so that the epipolar lines are horizontal and parallel. This transformation is referred as *image rectification*. Using F matrix, the images are rectified so that their image centers are aligned. In rectified images, every point in one image has a corresponding point on a horizontal line in the other image. This is illustrated in figure 2.7.

Normally the disparity of every pixel is estimated and using the similar triangle concept as in figure 2.2, depth of that pixel is estimated. Hence, a *depth estimation* from rectified stereo images as illustrated in the figure 2.8. For dense 3D reconstruction, it is normally assumed that the camera intrinsic are unknown.

[5]Courtesy: Wikipedia, Image Rectification - `https://en.wikipedia.org/wiki/Image_rectification`.

[6]Courtesy: Middlebury Stereo Vision Dataset - `http://vision.middlebury.edu/stereo`

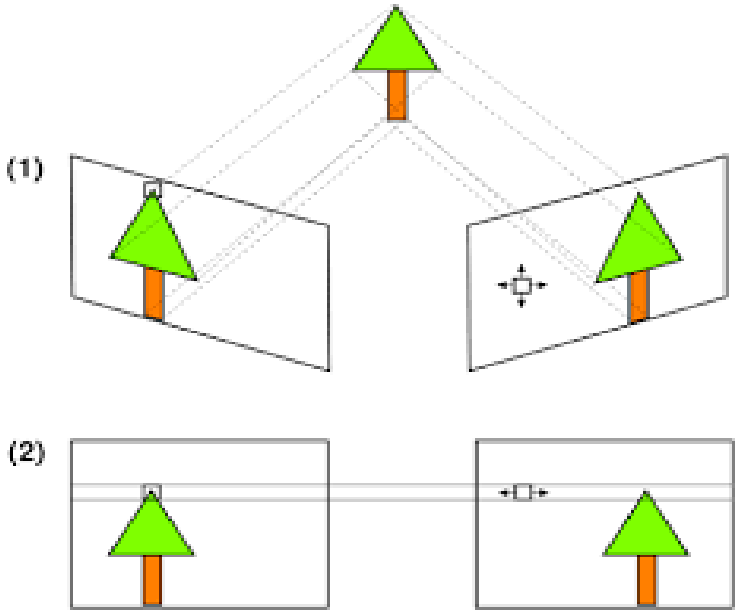


Figure 2.7: Illustration of image rectification. Courtesy: Wikipedia[5]

Figure 2.8: Illustration of depth estimation. Courtesy: Middlebury Stereo Vision Dataset[6]

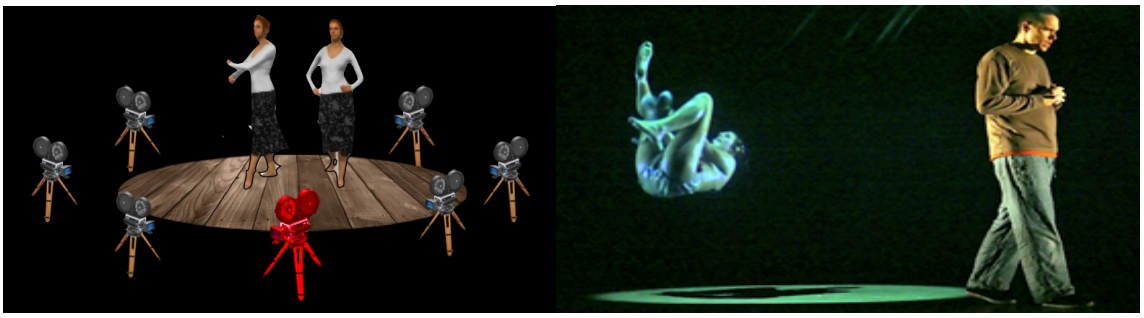
2.2 Deeper Look Into Application Scenarios

2.2.1 Virtually Enhanced Real-life synchronizeD Interaction - ON the Edge (VERDIONE)

The World Opera Consortium[7] envisioned the creation of a mixed-reality distributed stage for opera performances as shown in figure 2.9, where artists - real and virtual (artists who are physically at remote location) are performing together, immersively on one stage. Artists include singers, musicians or dancers, who would interact with their peers who are virtually projected on the stage. Such interactions pose a hard requirement on the quality of service of the system that projects the virtual video images captured in a remote location. For artists and audience to have a realistic experience of the whole performance, requirement of high quality video and a robust 3D video in real-time is necessary.

[7]The World Opera,
`http://www.geistweidt.com/pdf/WorldOperaWhitePaper.pdf`
`http://www.geistweidt.com/pdf/WorldOperaIntroduction.pdf`

Figure 2.9: World opera distributed stage performance. Courtesy: VERDIONE.



(a) Volumetric Capturing (b) Immersive Rendering

Figure 2.10: Illustration of VERDIONE capture and render subsystems. Courtesy: VERDIONE.

This idea of distributed stage performances has motivated the beginning of VERDIONE - Virtually Enhanced Real-life synchronizeD Interaction - ON the Edge project[8] to build a platform for such interactions between people around the world and to allow them to feel that they are co-located. VERDIONE is an image-based system consisting of volumetric capturing and immersive rendering as depicted in figure 2.10. To accomplish volumetric capturing, a multiple camera setup is necessary to synchronously capture the scene. For immersive environment volumetric displays are required. However, such a system aims at a high quality, robust 3D reconstruction and real-time rendering.

To meet the challenge of the world opera consortium, VERDIONE addresses problems in transmitting information from remote places to a virtual stage in terms of audio/visual acquisi-

[8]VERDIONE, Technology For Mixed Reality Stages, `http://verdione.wiki.ifi.uio.no/Main_Page`. Sponsored by Norwegian Research Council (Project No. 187828.)

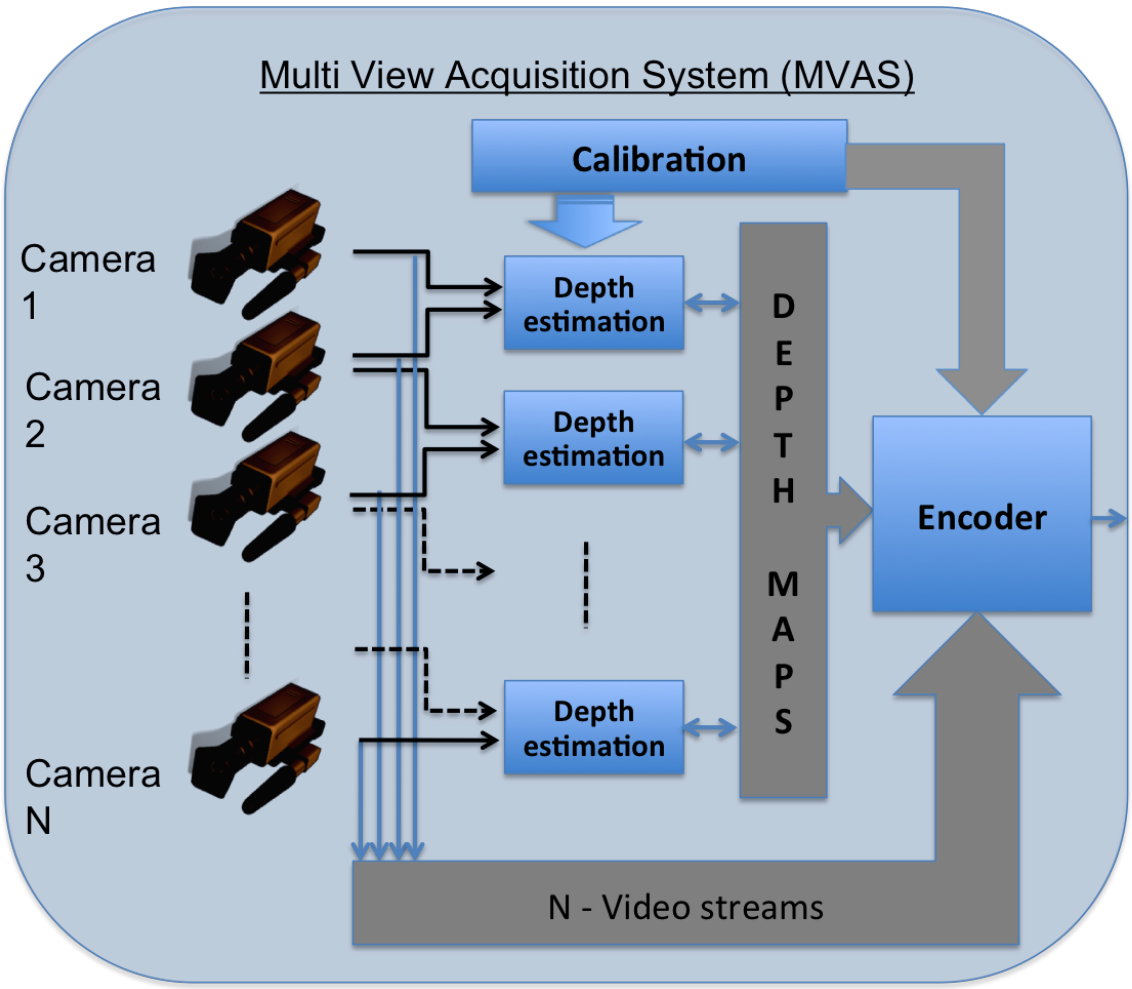


Figure 2.11: Multiple camera acquisition subsystem for VERDIONE.

tion and rendering. As envisioned multiple camera acquisition setup using depth based approach is as illustrated in figure 2.11. The disparity map and the video streams are sent to the network and on the receiving end, the information is used to depth image based rendering process.

Challenges

The image-based systems working in a larger volume, like an opera stage, is a challenge that has never been dealt before. Systems capturing the scene by markerless-based approach, is yet another challenge under the opera scenario. 3D reconstruction aims at photorealistic rendering at high resolution, but rendering images for both artists on the stage and the audience who are at different distances from the display is still a challenge.

Let us consider every aspect of this system and the challenges lying around them for the application of mixed reality art performances.

1. **Camera Density**
   In a multiple camera system, the first question that arises is how many cameras are required to efficiently obtain 3D data of the scene? The answer is based on empirical

observations, which depends on the object of interest, dimension of the 3D space and placement of cameras.

2. **Camera placement**
   Depending on the objects or the scene that are required to be captured and displayed remotely, the cameras need to be optimally placed around the stage during performance. Optimal placement depends on the approach used for 3D reconstruction; this is because the approach used for 3D reconstruction poses certain constraints on scene geometry of the images captured from multiple cameras. In the stereo algorithm based reconstruction requires that point correspondences should exist within 2 or 3 images, which in turn constraints that 2 or 3 cameras must look at the same object such that sufficient point correspondences are obtained.

3. **Camera Array**
   In a homogeneous camera array, all cameras have the same focal lengths. However, there exists variation in image resolution as the object distance varies proportionally. In such cases, it is worthwhile to measure and study the lower limits of the resolution for a better reconstruction. Also, a heterogeneous camera array, where the focal lengths of certain cameras in the array are different from each other, is also interesting to study.

4. **Camera Calibration**
   Obtaining 3D space metrics and camera characteristics is related to the camera calibration problem in computer vision. This is the first and foremost step, and the purpose is to obtain the intrinsic parameters of the camera such as focal length, radial distortion, etc., and extrinsic parameters such as rotation and translation orientation of the cameras in 3D space. Such geometrical calibration of cameras is required for accurate 3D reconstruction. The challenge here is to obtain an accurate calibration of cameras, in turn to obtain an efficient reconstruction of 3D data. The process of calibrating the cameras plays an important role: by using of 3D objects to calibrate or by adopting feature based calibration approach, where calibration method must be able to collect information directly from the scene. Photometric calibration is necessary to yield more realistic appearance of the object or region of interest. Color checkerboards are sometimes used for this purpose. The challenge is to accurately color calibrate the cameras for uniform color composition while rendering at different viewpoints.

5. **Feature Extraction** For images from two or more cameras, feature correspondences are required for feature based calibration and sparse depth map. For good quality camera calibration high quality feature correspondences are required. So the challenge is to find a robust feature extractor that provides high quality correspondence.

6. **Synchronization**
   The camera system must be synchronized so that multiple cameras capturing the same scene should do so, at the same instant of time. This must be done in order to compute the calibration parameters with high accuracy; otherwise determination of point correspondence during the calibration process will be hampered. In terms of reconstruction, there would exist inconsistencies due to unsynchronized data.

7. **Depth estimation problem**
   One of the main challenges in this section is to obtain accurate depth-maps of multiple images. Depth estimations in a multiview system depends on two things (a) matching problem - determining the point correspondences in the images and (b) disparity - distance between the matching points. Variations in depth estimation result in structural artifacts in the borders of the reconstructed image, which are sensitive to human eyes. Hence, the accuracy in depth estimation determines the robustness of the reconstruction algorithm for rendering.

8. **3D Representation for Transmission**
   Several multiview algorithms use one or many of the following representations: meshes, polygons, images+depth maps, point clouds, level sets or voxels. Choice of representation depends upon the underlying rendering scheme. A new representation intended to yield an efficient coding scheme is also interesting to explore.

9. **3D Reconstruction and Rendering**
   Image based rendering methods often suffer from holes in the warped images due to difference in re-sampling and occlusions. The main challenge for reconstruction and rendering schemes is high quality view synthesis in terms of speed, robustness against artifacts and resolution.

   Virtual viewpoint images for view-dependent and view-independent (free viewpoint) rendering use techniques like interpolation, warping and re-sampling, and render the images of the closest captured camera. Each of these techniques will in turn account for the quality of the view synthesis. Hence, an appropriate technique adoption is a challenge.

10. **Physical dimension of the stages**
    A stage dimension occupies a large volume of 3D space. It seems that no system has been developed to work with this huge volume as an application of multiview capture systems. Hence, certain challenges exist in terms of resolution and robustness of feature extraction, pose estimation and rendering quality images.

11. **End to End delay**
    For VERDIONE, this is one of the most important design parameters for real-time performances. End-to-end delay can be roughly estimated as the time expired from the image acquisition stage to the image displayed on the remote screen. This is a summation of delays contributed by several steps: capturing, processing, distribution, transmission and rendering.

12. **Backdrop**
    If the backdrop were of a contrast color (unnatural background for opera performances like a blue or green screen), it would be easier to obtain a bounding box around the scene object, by using background subtraction techniques. It is necessary to consider other parameters like the uniformity of luminance on the backdrop and its shading variation. If the scene has similar appearance of the backdrop then the task is more challenging. One of the ways of dealing with this problem could be to have prior knowledge about the scene backdrop.

13. **Illumination and Shadow**
    The illumination on the stage would affect the appearance of the scene object, which needs to be detected and tracked. Also a shadow on the foreground or the background affects the system performance. It seems that ambient light would be a solution, but it is difficult to achieve ambient light. Hence, studying the effects of lighting becomes necessary to find an optimal solution .

14. **Parallax and occlusion effects**
    It is important to study the effects of motion parallax on depth perception, reconstruction errors and other factors that the parallax depends on. Occlusion problem is not simple to solve in stereo algorithms, and hence, it is worth investigating.

2.2.2 An Integrated System for Soccer Analysis (BAGADUS)

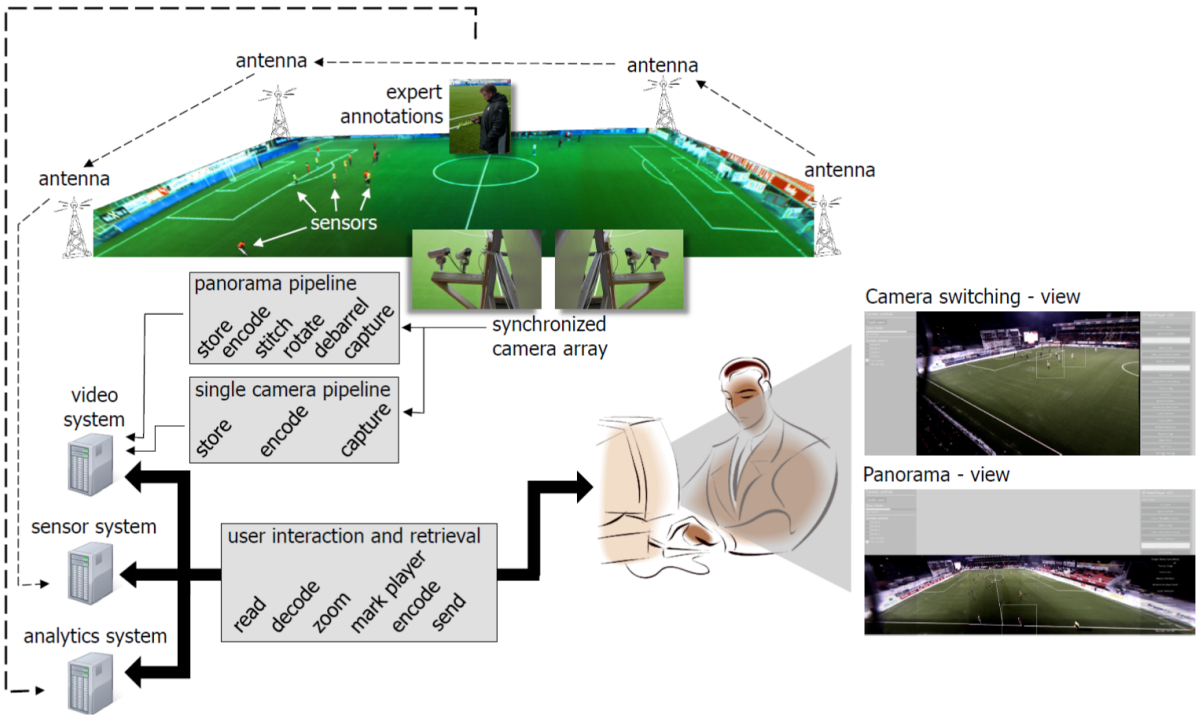


Figure 2.12: Overall BAGADUS architecture. Courtesy: BAGADUS.

Sports analysis, especially in soccer has been very important in order to keep up the competitiveness in the game. The manual analysis that are carried out by coaches or other analysts consume a lot of time. Today's technology provides a solution to conduct such sports analysis in real-time. BAGADUS[9] (Halvorsen, Sægrov, Mortensen, Kristensen, Eichhorn, Stenhaug, Dahl, Stensland, Gaddam, Griwodz, and Johansen, 2013) is a project that aims to fully integrate existing systems and enable real-time presentation of sport events. This system is built in cooperation with the Tromsø IL soccer club and the ZXY sport tracking company for soccer analysis. A brief overview of the architecture and interaction of the different components is given in figure 2.12.

[9]BAGADUS - `http://site.uit.no/iad/sports/bagadus`

The BAGADUS system is divided into three different subsystems, which are integrated in the soccer analysis application. The subsystems are, *Video*, *Tracking* and *Analytics*.

In *video subsystem*, multiple small shutter and synchronized cameras capture high resolution video of the soccer field. This subsystem is also responsible for playback of the videos in two different modes. The first allows the viewer to watch separate camera feeds and the second mode plays a panorama video stitched from multiple camera feeds.

The *tracking subsystem*, identifies and tracks players through the camera arrays. BAGADUS uses a sensor based solution to get the positions of the players and tracks the players in single camera view or stitched view.

The *analytics subsystem* equips the members of the team with a tablet or even a mobile phone, where they can register predefined events quickly with the press of a button or provide textual annotations. In BAGADUS, the registered events are stored in an analytics database, and can later be extracted automatically and shown along with a video of the event.

BAGADUS focuses mainly on combination and integration of components enabling automatic presentation of video events based on the sensor and analytics data that are synchronized with the video system. Thus BAGADUS will, for example, be able to automatically present a video clip of all the situations where a given player runs faster than 10 meters per second or when all the defenders were located in the opponent's 18-yard box (penalty box). Furthermore, BAGADUS system can follow single players and groups of players in the video, and retrieve and playout the events annotated by expert users. Thus, where people earlier used a huge amount of time for analyzing the game manually, BAGADUS provides an integrated system whereby the required operations and the synchronization with video is automatically managed.

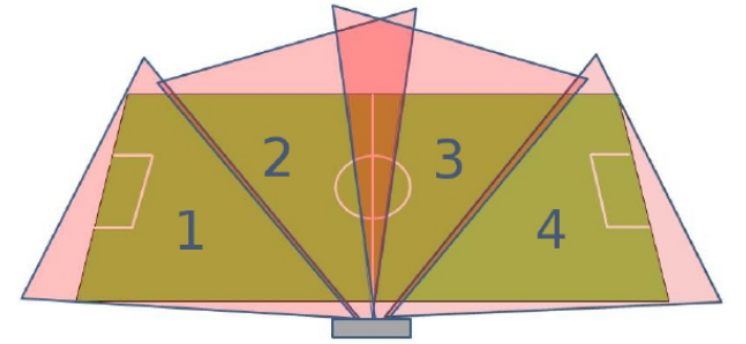


Figure 2.13: Camera setup in Alfheim soccer stadium. Courtesy: BAGADUS.

To record high resolution video of the entire soccer field, a camera array consisting of 4 Basler industry cameras with a 1/3-inch image sensor supporting 30 fps and a resolution of 1280x960 is used (see figure 2.13). The cameras are synchronized by an external trigger signal in order to enable a video stitching process that produces a panorama video picture. For a minimal installation, the cameras are mounted close to the middle line under the roof covering the spectator area, i.e., approximately 10 meters from the side line and 10 meters above the ground. With a 3.5 mm wide-angle lens, each camera covers a field-of-view of about 68 degrees, i.e., all four cover the full field with sufficient overlap to identify common features necessary for camera calibration and stitching.

The pipeline of creating a panorama video involves image capture, remove barrel lens distortion and stitching. Four cameras are installed to view the entire soccer field with certain overlapping regions. Cameras are calibrated offline and their calibration parameters are used for removal of lens distortion. Stitching is desired to have a complete view of the soccer field. Using the overlapping regions of the camera views, a homography between the images are computed. Feature correspondence between two images are detected in the overlapping areas and then used to estimate the homography transform based on multiview geometric principles. Finally, the images are warped based on the homographies obtained, and thus stitched image is obtained.

Challenges

Some of the main challenges pertaining to image acquisition are as follows:

1. **Feature extraction**
   It is very likely to have bad weather conditions, i.e., snow, fog and rain. Such situations make it very difficult to obtain a decent image for feature extraction that is required for homography estimation. Moreover feature correspondences obtained on the grass field are not very reliable.

2. **Lens distortion**
   Barrel distortion errors are minimized using the camera calibration parameters. However, it is hard to achieve a perfect calibration. This makes finding the homography between planes difficult and error-prone, and thereby affects the stitching quality.

3. **Latency**
   The execution time for the whole process of video subsystem from capture to storing the stitched images can be challenging. Most time is spent on stitching process. If the annotations of players information needs to be shown while the game is on, then a real-time processing is a critical aspect of quality of service.

4. **Parallax**
   The problem of parallax has been identified in this project. In certain areas of the field, parallax have been found to be more prominent. This problem is worth investigating.

5. **Camera Intrinsic**
   It has been a problem to maintain the same exposure in all the images, which causes an uncomfortable viewing experience when images are stitched.

6. **Camera placement**
   The camera placement is such that the center of projection is not well aligned. Due to this parallax effects have been observed. Therefore new camera arrangement needs to be found, which minimizes the parallax effects while obtaining the sufficient view of the soccer field. Based on the arrangement in Alfheim stadium, overlap seems to be very less with a large angular separation.

2.2.3 PantoInspect Train Monitoring System (PTMS)

PantoInspect Train Monitoring System (PTMS)[10] is a fault inspection system, which inspects pantographs and ensures automatic quality control of the entire network of pantographs, while the train is in motion. This system, which has been installed in Banedanmark (Denmark), Sydney Trains (Australia), Transnet (South Africa) and Austrian Federal Railways(Austria), is currently in operation.

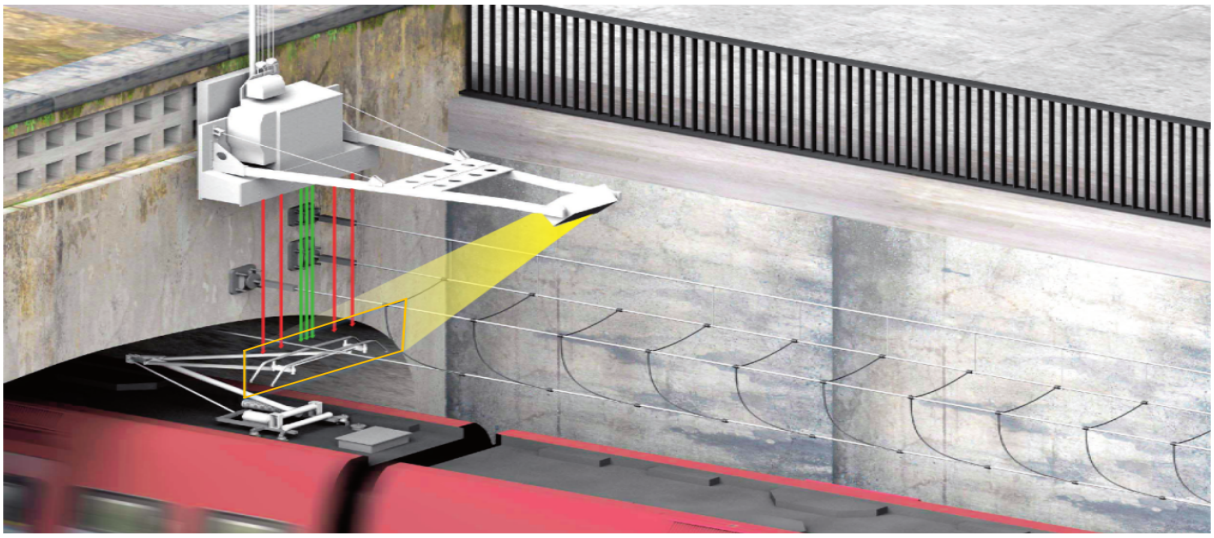


Figure 2.14: PTMS: inspects defects on the pantographs mounted on electric trains. Courtesy: PantoInspect A/S.

PTMS is normally installed, as shown in figure 2.14, over railway tracks to inspect trains running with electric locomotives that are equipped with pantographs. Pantographs are mechanical components placed on one or more wagons of the train, which can be raised in height so that they touch the contact wire for electricity. PTMS measures the dimensions of the defects in their pantograph's carbon strips, detects misalignment in the pantograph's position and estimates the train speed.

Pantographs have one or more carbon strips that are actually in contact with the wire. Over time, due to constant contact, the carbon strips may wear out. This can eventually result in tear down of contact wires, if necessary action is not taken to replace the pantographs in time. The uplift force of the pantograph controls the pressure applied by the pantograph on the contact wires. The variation in uplift force can also be the reason for tear down of the contact wires. While the train is in motion, the pantographs may move sidewards based on the speed of the train and the pressure on the contact wire, eventually tearing them down.

Thus, there are several factors resulting in the contact wire tear down, which leads to a serious consequence. These factors are related to various defects occurring on the carbon strip. There can be several types of defects, which are *(1) thickness of carbon wear, (2) vertical carbon cracks, (3) carbon edge chips, (4) missing carbon and (5) abnormal carbon wear*, and can be seen in figure 2.15.

[10]Pantoinspect A/S - `http://www.pantoinspect.com`

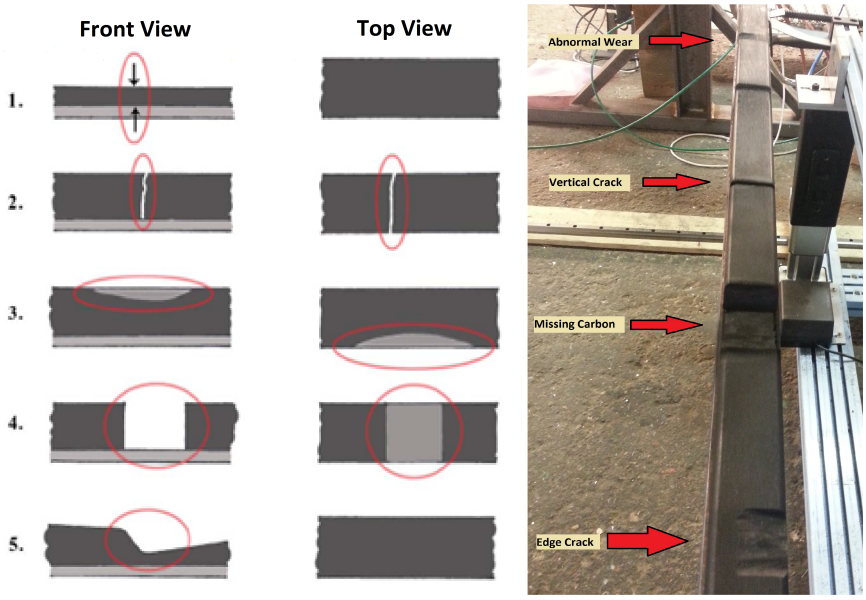


Figure 2.15: PTMS defects illustrated and shown on a pantograph. Courtesy: PantoInspect A/S.

The PTMS is a complex system involving several hardware and software units. It also interfaces to the real-time train monitoring system for receiving notifications about the trains and transmit the defect inspection report. Soon after the system detects and analyzes the defect to a be critical one, it alarms the authorities for taking necessary action. A snapshot of the user interface of PTMS is shown in figure 2.16.

The working principle of PTMS can be classified into the following stages: data acquisition, profile image analysis, offline calibration and defect measurement.

Data Acquisition

The PTMS is mounted right above the train tracks, usually on bridges or other fixtures as shown in figure 2.14. The PantoInspect system receives a notification when the train is approaching and prepares itself. When the train passes right below the system, the range finders (depicted as red lasers in figure 2.14) detect the pantograph's carbon strip, when it is right below the system. At this instance, three line lasers are projected onto the carbon strips (depicted as green line in the figure 2.14), and the camera located about 2 - 3 meters away from from the pantograph captures the near infrared image of the laser, termed as the profile image, which can be seen in figure 2.17(a).

Profile Image Analysis

The profile image is pre-processed by noise reduction and image enhancement techniques to obtain a good image of laser lines for further analysis. When defects are present, the line deforms instead of remaining a straight line in the image. In this way, the laser line defines the geometry of the defect, which helps to precisely measure the defect in physical units. Figure 2.17(b) also shows how a carbon missing defect is identified as the deformed laser lines obtained by processing the profile image.

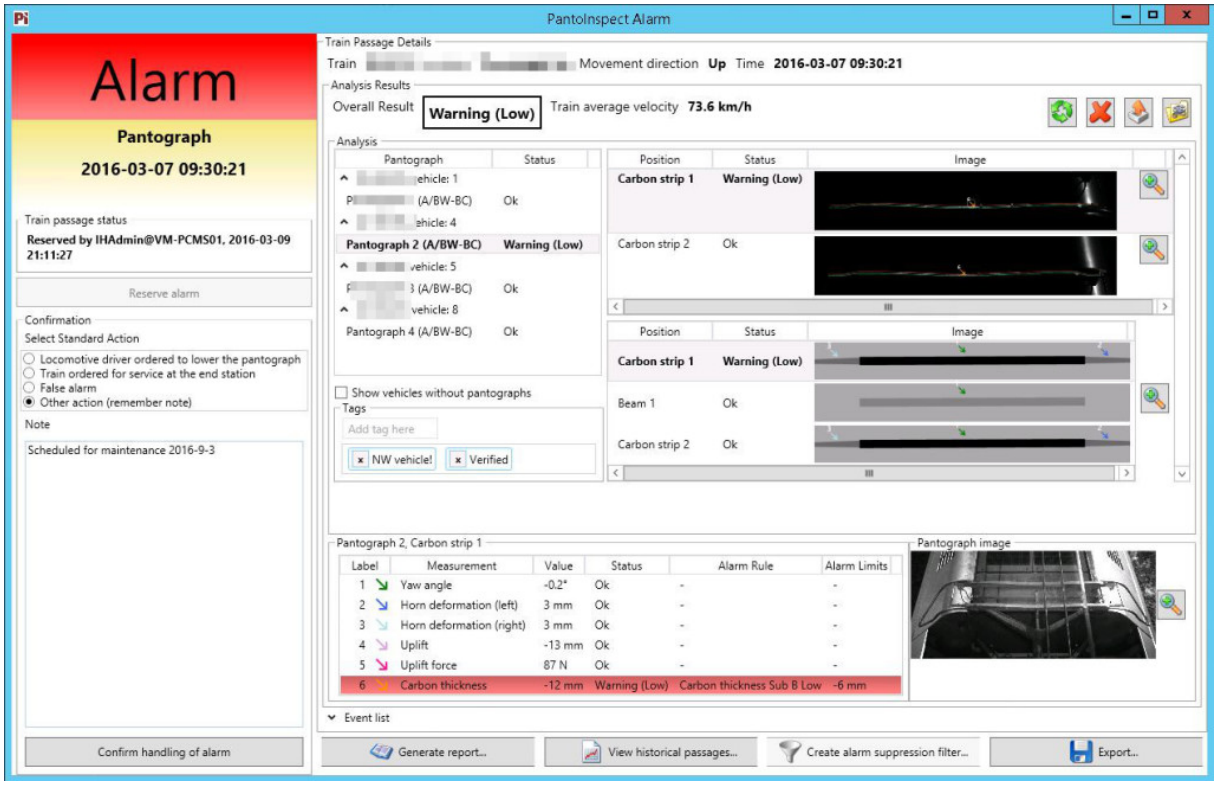


Figure 2.16: User Interface of PTMS inspection system. Courtesy: PantoInspect A/S.

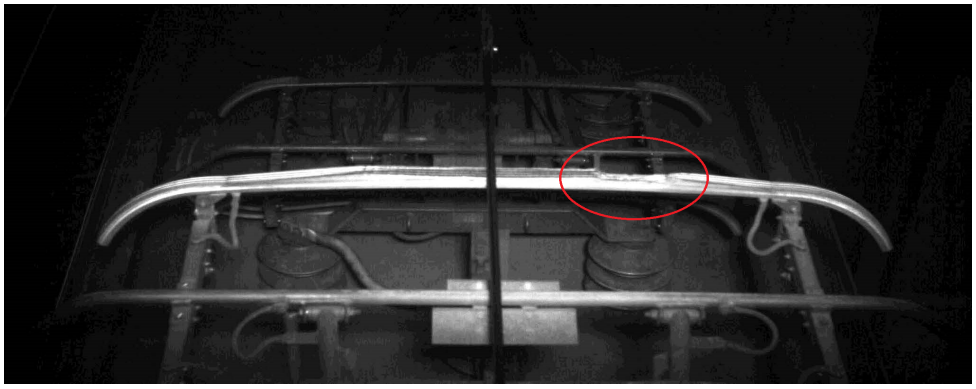
Once the defect is identified from the profile image, intricate profile matching techniques are used on all three laser lines to obtain the actual geometry of the defect. Figure 2.17(c) also shows how the laser line geometry is detected for the defect in the profile image. In the figure, the 'System' indicates the width measurement of the defect estimated by the system through image analysis and 'Caliper' indicates the actual measurement of the width measured using an industrial caliper for obtaining ground-truth measurements.

Camera Calibration:

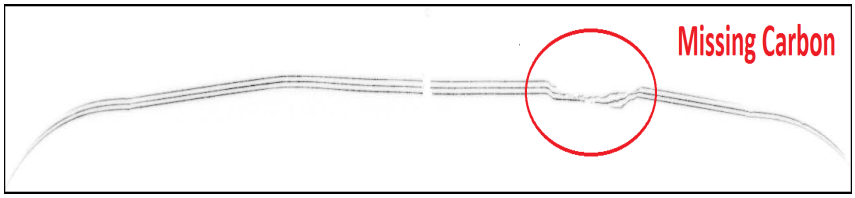
Camera calibration is an important step in obtaining such 3D measurements from 2D image points. For PTMS, this is carried out in the factory before deploying the system, using Bouguet's method (Bouguet, 2008). A number of checkerboard images are used to estimate the intrinsic parameter K of the camera that constitutes focal length and principle axes of the camera. Next, a single image of the checkerboard that is placed exactly on the laser plane, is used to estimate the extrinsic parameter of the camera - position T and orientation R, with respect to the checkerboard coordinates. For more mathematical details about camera calibration, please refer section 2.1.1.

Defect measurement

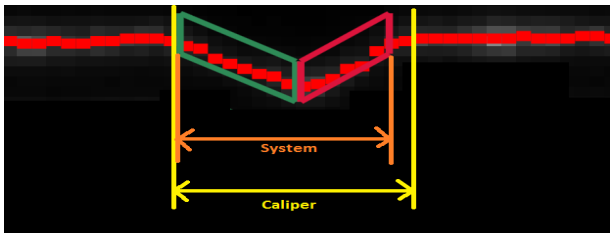
At this stage, the PTMS has acquired 2D image points that define the defect. The next step is to estimate the corresponding 3D points, and thereby, estimate the defect measurements in real world metrics. This is carried out with the help of camera calibration that is carried out



(a) Profile image captured.



(b) Laser line image.



(c) Defect detection in a laser line image.

Figure 2.17: Pantograph image analysis. Courtesy: PantoInspect A/S.

offline as described above. It is considered that the 3D defect points lie on an imaginary laser, and therefore, estimating 3D points is merely a ray-plane intersection (Hartley and Zisserman, 2004), where the ray is drawn from the camera center through 2D points.

This is how the PTMS captures laser profile images, analyzes them, measures the dimension of the defects and notifies the user, if the measurement is above certain threshold values.

Challenges

Some of the main challenges that the PTMS faces are as follows:

1. **Camera misalignment**
   The camera used in the PTMS is fixed at its position and orientation during the offline calibration process that takes place in the factory before deployment of the system. The camera position is not allowed to change after the calibration process to acquire accurate 3D measurements. Before deployment the camera might be misaligned due to transportation. After deployment, the camera might be misaligned due to manual intervention or natural causes (e.g., wind). This misalignment causes inaccuracy in the 3D measurements.

2. **Pantograph misalignment**
   Due to the tension between the pantograph carbon strip and the contact wires, and the speed of the train, the pantograph can disorient itself from its original position. This leads to misalignment of laser lines projected on the carbon strip, giving raise to partial capture of defect which results in inaccurate measurement of the defects. Even if a good image of the laser lines is obtained, due to the pantograph misalignment, the calibration parameters will no longer be valid because the camera was calibrated based on an imaginary plane that lies on the pantograph. This again results in inaccuracy of the system.

3. **Noisy Profile Image**
   The camera used in the PantoInspect system captures infrared images of the laser lines. When PTMS is installed in sunlit areas such as in South Africa, Australia etc., then the infrared component in the sun rays interfere with the laser lines and a noisy image is captured by the cameras. Another source of noisy images is overexposure of the infrared images, which makes it difficult to detect any features in the images. So, the noisy image is a challenge to extract the line features for further analysis and measurement of defects.

2.2.4 Previz for On-set Production - Adaptive Real-time Tracking (POPART)

In modern film production, animated effects are used extensively. The EU-project POPART[11] - Previz for On-set Production - Adaptive Real-time Tracking aimed at developing product where it is possible to preview digital effects on set, and POPART further aims to heighten the efficiency of post-production work considerably. POPART contributes to film making from the planning stage to filming with actors on set, where the final composition of mixed reality scenes can be reviewed during and right after the shoot. POPART provides an adaptive and integrated solution to conduct live on-set production in order to assess the quality of filming and greatly help the post-production process.

The POPART system is illustrated in figure 2.18, which comprises 4 steps:

- Shooting Preparation
- Filming with real-time on-set visualisation
- Deferred on-set post-production
- Post-production

During the shooting preparation step, a series of high-resolution images are captured from distinct viewpoints. These images are sent on-the-fly to a computer that incrementally computes the 3D points cloud and the camera poses. At the end of preparation, a final 3D reconstruction is generated by simultaneously taking into account all collected images, which ensures a greater accuracy and quality. Such global reconstruction of point cloud of the set with camera poses is shown in figure 2.19. Using POPART's photo-modeling application, the matte-painter creates the 3D model of the set based upon the 3D points cloud. The input snapshots are projected to create the textures. Eventually, a 3D visual database is created during the 3D reconstruction.

[11]EU POPART Project - `http://www.popartproject.eu`

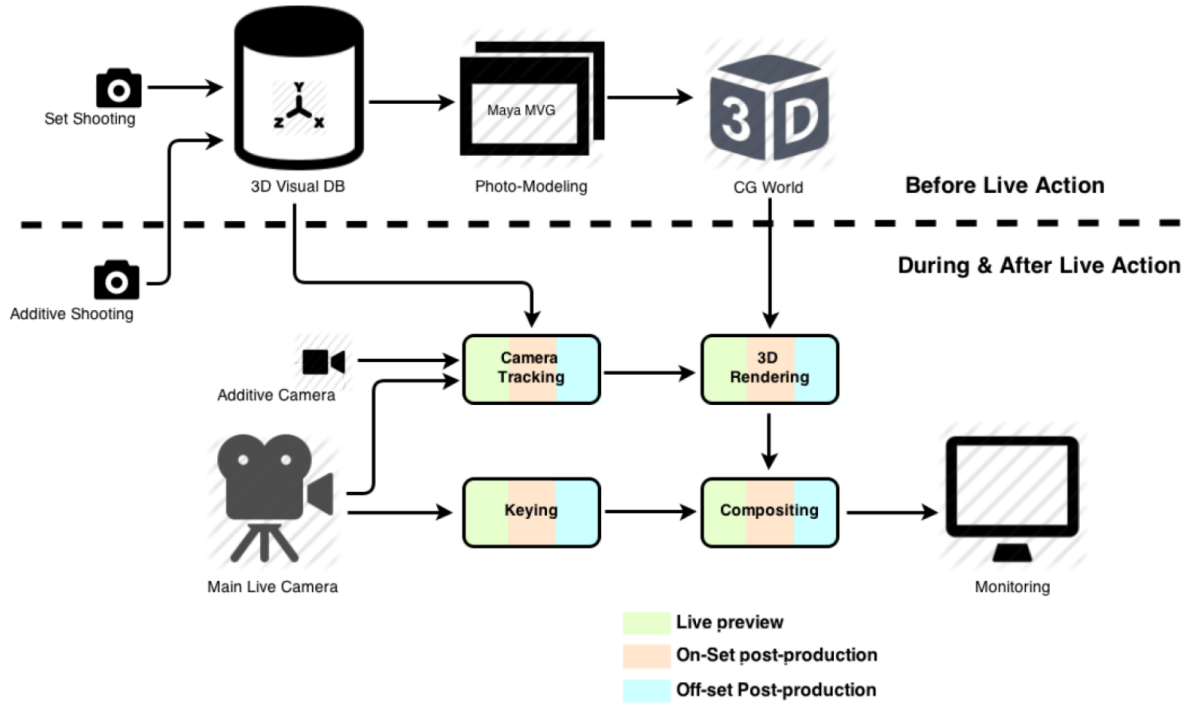


Figure 2.18: POPART System. Courtesy: EU POPART Project.

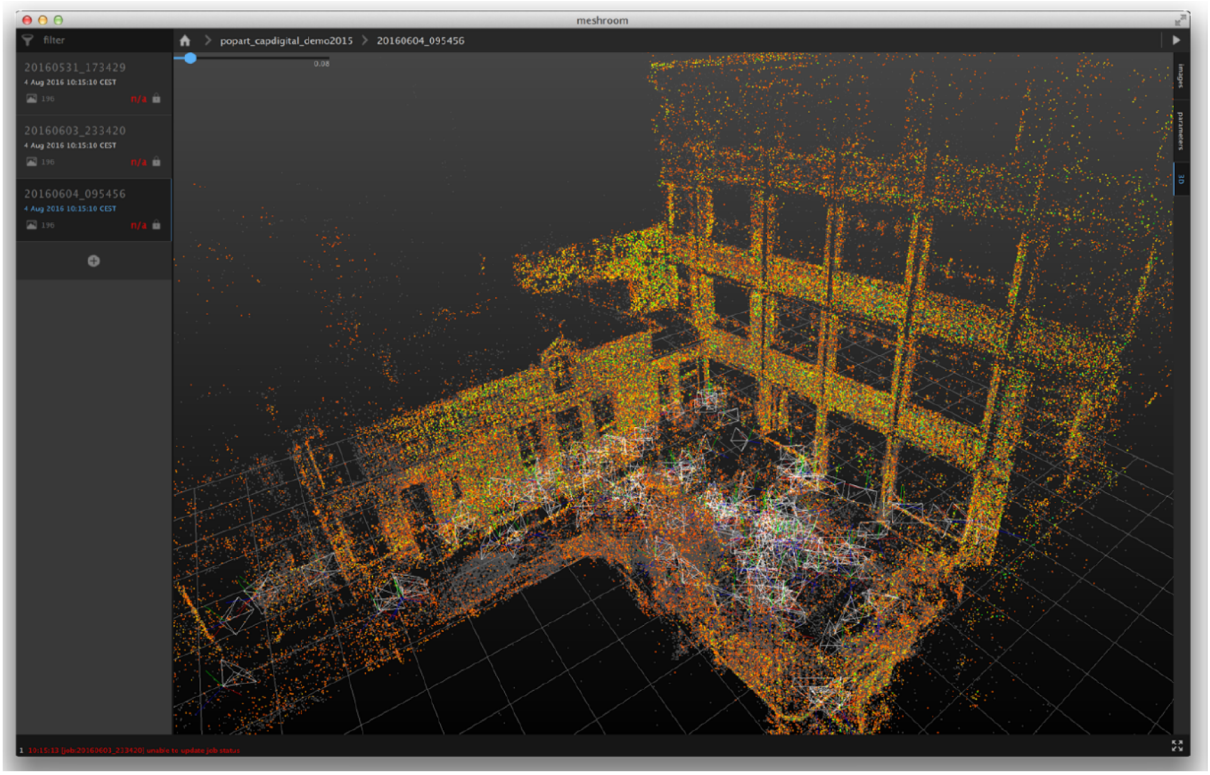


Figure 2.19: The 3D Point cloud of the real filming set. Courtesy: EU POPART Project.

During the filming stage, a multiple camera system is setup. POPART rig is equipped with additional cameras rigidly fixed to the main action camera. Before starting to film, the multiple camera system is calibrated using the 3D visual database. During filming, the camera tracker estimates the pose of the multiple camera system relying on the 3D visual database. With all this information, the main camera feed is seamlessly integrated into the digital world. The final composited image including color-keying adjustment, color matching and grading are streamed for real-time on-set visualization of the mixed scene.

It is also possible to provide a "on-set post-production" with deferred processing that yields higher-quality results somewhat slower than the real-time visualization provided in the previous step. In this step a high quality 3D rendering with effects can be previewed.

All the raw data collected during the shooting preparation and the on-set visualization are a precious source of information, i.e., 3D database, camera tracking, 3D rendering and compositing. All this can be exploited in the post-production step. In this way, POPART framework will be fully compatible with the traditional VFX software for high-end post-production.

Challenges

1. **3D Visual Database**
   An accurate and robust 3D reconstruction is the primary focus to achieve high quality 3D point cloud of the real set. Obtaining such high accuracy and robustness is absolutely necessary, because it is used for a continuous camera pose estimation during the shooting stage. The challenge here is to incorporate all factors (i.e. change in scene and camera properties) that affect the accuracy in the whole pipeline of 3D reconstruction.

2. **Feature Extraction**
   Feature extraction is the fist step in analysis of images for either matching the main camera feed to the images in the database or for estimating camera pose. The detection of features becomes extremely challenging due to poor informative images or dynamic and cluttered scenes: typical cases include shooting in poorly textured environments (like unified green backgrounds), natural factors like smoke, mist etc. and dynamic scenes where motion blur occurs.

3. **Low Latency**
   For real-time live preview, every frame must be blended well into digital 3D world. This requires an accurate camera tracking with low latency. Therefore, a real-time camera tracking based on search in visual database is a challenge.

4. **Deferred Preview Processing Time**
   On-set pre-production which includes 3D rendering and high quality compositing with extra effects, i.e., motion blur, defocus, noise processing are currently too computer intensive. So challenges here are to obtain a small delay in the intermediate step to provide an on-set pre-production capability.

2.2.5 Scanning Electron Microscopes Reconstruction (SEMRECON)

The Scanning Electron Microscopes (SEM) are instruments that produce images of a sample by focusing beams of electrons. Today, softwares are available such as SPIP[12] and Mountains[13] to analyze SEM images for image enhancement or to characterize the surface of the sample by reconstructing 3D surfaces and estimating their roughness.

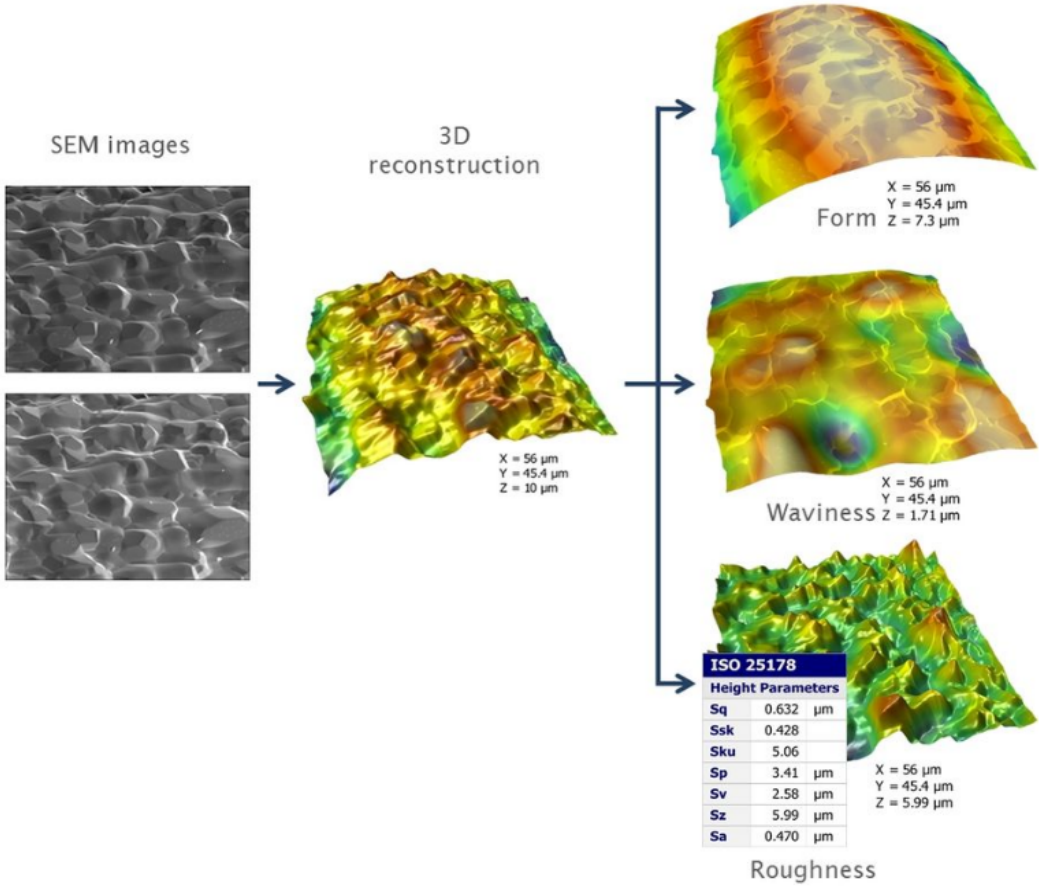


Figure 2.20: The 3D reconstruction and surface analysis - waviness and roughness surfaces and ISO 25178 height parameters. Courtesy: MountainsMap SEM

An example of SEM image analysis is as shown in the figure 2.20. Here, the aim is to analyze the surface texture, and therefore, 3D reconstruction of the SEM image is necessary. After the surface is reconstructed, roughness measurements based on ISO 25178[14] is carried out for industrial purposes.

Challenges

Roughness measurement has been very important using SEM images. However, the new trend is to measure roughness for complex structures, especially on the sides of the structure, which is impossible to see in a single image. Therefore, the sample is rotating on a turntable and several

[12]SPIP - `http://www.imagemet.com/`
[13]MountainsMap SEM - `http://www.digitalsurf.com/en/mntsem.html`
[14]ISO 25178 - Surface texture: `https://en.wikipedia.org/wiki/ISO_25178`

images are produced using SEM instruments. The aim here is to reconstruct a full 3D structure of the sample based on multiple SEM images and analyze the surface. The challenge here is to model the SEM as a camera and use multiview geometric principles to achieve high quality 3D reconstruction and carry out advanced roughness analysis on SEM samples.

2.3 Conclusions for Preliminary Concepts

In this chapter, important concepts and terminologies of image based 3D multimedia systems were briefly described. New age application scenarios that this thesis focuses on, were also explained. The challenges faced by the application scenarios were outlined, and some of them were used as a motivation for the research question.

In all above scenarios, the common aspect is that the 3D reconstruction requires the knowledge of camera calibration parameters. The inconvenience to use the calibration target in the above scenarios poses one of the main challenges to the camera calibration process and thereby 3D reconstruction. Therefore, the next chapter goes deeply into the exploration of camera calibration in 3D multimedia system, especially using distinctive features in the images and without the use of any type of calibration target.

Chapter 3

Feature Based Calibration (FBC)

A brief overview of image based 3D multimedia system was outlined in the previous chapter. The estimation and application of camera calibration in 3D systems was also described for both single and stereo camera. Such calibration process can be carried out using traditional Checkerboard Based Calibration (CBC) (Bouguet, 2008; Tsai, 1992; Zhang, 2000) or Marker Based Calibration (MBC) (Kurillo, Li, and Bajcsy, 2008). There are also solutions existing for calibrating distributed cameras capturing the scene (Kassebaum, Bulusu, and Peng, 2010). As an alternate solution, in this chapter, Feature Based Calibration (FBC) is discussed.

In application scenarios such as VERDIONE (section 2.2.1) or BAGADUS (section 2.2.2), carrying out CBC is almost impossible, as a really big checkerboard is required to be placed in the center of a stadium or soccer field to calibrate the cameras installed at a large distance. Alternatively, MBC techniques might be inconvenient in terms of capturing undesirable markers in the scene. In the PTMS scenario (section 2.2.3), placing a checkerboard on the rail tracks is very inconvenient and is at the cost of interrupting the railway traffic. Hence, the 3D applications must rely on FBC. In this chapter, adopting FBC in 3D systems and studying the effects of camera misalignment of both single and stereo 3D systems are discussed in detail.

In 3D systems with either single or stereo cameras, a major practical problem is camera misalignment, i.e., change in position or orientation of the camera fixed to the rig. Before installation of the system, the camera misalignment might be caused due to transportation or deployment errors and after installation, any manual intervention or natural causes (e.g., wind) might be the reason for camera misalignment. When cameras are misaligned, the quality of reconstruction and hence the quality of 3D systems is hampered. Quantifying the quality of 3D systems depends on the acceptable reconstruction error which in-turn depends solely on what the application demands. To study such effects of camera misalignment in more detail, a relevant research question was put forth, and correspondingly, *Hypothesis I* was stated as in section 1.4.

When camera misalignment occurs during the operation of PTMS and the performance degrades, only a recalibration process can be the solution to maintain the quality of 3D systems without shutting down the system for servicing. But in most cases, the recalibration process based on traditional methods is practically impossible to achieve, i.e., by trying to hold a checkerboard on the rail site, and hence, adoption of FBC is necessary. To study the implication of using FBC instead of CBC and comparatively analyze the quality of 3D systems, *Hypothesis II* was stated in section 1.4.

In this chapter, the aim is to test *Hypothesis I* and *Hypothesis II*. In order to test these hypotheses, the effects of camera misalignment and adopted FBC in 3D multimedia systems were studied. In order to investigate the performance degradation of 3D systems due to camera misalignment, the effects of camera misalignment on both single and stereo camera systems were studied.

3.1 Misalignment in Single Camera System

3D vision systems using single camera technology finds its applications in various industries. In games and entertainment industries, popular products like Kinect use a single camera to capture a projected structured pattern on the object and thereby estimate the depth information. Depth estimation from structured light has been studied quite extensively (Rocchini, Cignoni, Montani, Pingi, and Scopigno, 2001; Turk, Kim, Yi, and Park, 2005). In robotics, companies like ABB Robotics[1] and DTI[2] adhere to single camera technology solution by using a camera attached to the robotic arm to assist picking and sorting objects in industries. Nowadays, one of the most popular industries where 3D vision systems are used is automation and inspection. Single cameras have been installed to inspect the potatoes, eggs, chicken etc for grading purposes by companies like IHFood[3], QTechnology[4] and Newtec[5]. Similarly automatic inspection systems are used to inspect various types of faults or defects for sorting and quality assurance in food industry (Brosnan and Sun, 2004; G and S, 2010), inspection of cracks in roads (Cord and Chambon, 2012), crack detection of mechanical units in manufacturing industries (Mar, Fookes, and Yarlagadda, 2009; Zhao and Li, 2005) and so on. Vision based inspection systems are increasingly growing with the advance in computer vision techniques and algorithms.

The quality of vision based inspection systems, where 3D reconstruction is involved, is strongly dependent on the quality of camera calibration. Usually camera calibration process is carried out offline and the corresponding calibration parameters are used to recover 3D measurements from single 2D image. The use of calibration in 3D recovery can be seen in many applications (Araki, Sato, Konishi, and Ishigaki, 2009; Heimonen, Hannuksela, Heikkila, Leinonen, and Manninen, 2001; Le Flohic, Parpoil, Bouissou, Poncelet, and Leclerc, 2014).

One such interesting application that uses single camera vision system, i.e., PantoInspect Train monitoring System (PTMS), is considered for studying the effects of camera misalignment in a single camera system. The PTMS is chosen because it is a very good example, where a catastrophic consequence might occur due to error in the 3D system in real scenario. As explained in section 2.2.3, PTMS detects and measures defects in pantographs in real-time and alarms the train monitoring system for further action. Small magnitude of error in defect detection might cause false alarms, for example, one which wrongly inspects a wornout pantograph. This can indicate a wrong signal to the train authorities that the replacement of pantograph is not required. This in fact, compromises railway safety. The camera unit of the PTMS is respon-

[1]ABB - Leading supplier of Robot software, equipment and a complete application solution - `http://new.abb.com/products/robotics`

[2]Danish Technological Institute - `http://www.dti.dk`

[3]IHFood A/S - Vision technology for inspection, grading and sorting - `http://www.ihfood.dk`

[4]Q Technology A/S - All-in-one camera with vision system - `http://www.qtec.com`

[5]NEWTEC - weighing, packing, sorting machinery - `http://www.newtec.com`

sible to acquire image data of the cracks. This unit is prone to movement during transportation, deployment, due to wind or manual intervention during servicing. Due to this misalignment, quality of defect detection and measurement is prone to errors.

The effects on digital imaging systems from electronic, mechanical and optical influences have been analyzed analytically and discussed (Godding, 2000). However, the study motivates the need for system calibration as a result of the errors occurred due to the external influences. This paper explains more about the need and methodology of camera calibration. In (Hayman and Murray, 2003), effects on estimation of focus is measured due to translation misalignment of self calibrating camera. In another article (Hijazi, Friedl, and Kähler, 2011), the effect of non perpendicularity of camera's optical axis with object plane, on mechanical strain was studied. Here, the misalignment was represented only as camera tilt and shift, i.e., camera orientation and translation in one direction.

The study in this section of the thesis extends the camera orientation and translation in all three directions, which is important to address for image-based 3D measurement applications. Moreover, in this case, the calibration parameters are assumed to be highly accurate and that they stay the same even after the camera is misaligned. Hence, 3D measurements estimated using original calibration parameters results in errors. This might seem obvious that the quality of 3D measurement is affected by camera misalignment, but interesting part of this study is to explore more in the aspect of how significant are the effects of camera misalignment on single camera system and how to use the knowledge to improve the system design.

3.1.1 Evaluation

In our paper titled "Study the Effects of Camera Misalignment on 3D measurements for Efficient Design of Vision-based Inspection Systems" [details in chapter 9], a statistical tool in the form of a methodology was proposed. This involved the following:

- Studying the significance of the effects of 3D measurement errors due to camera misalignment.
- Modeling the error data using regression models.
- Deducing expressions to determine tolerances of camera misalignment for an acceptable inaccuracy of the system.

As the PTMS application motivated this study, the accuracy of 3D measurements of the PTMS was observed in order to evaluate the effects of camera misalignment. Although the PTMS detects and measures various types of defects as seen in figure 2.15, the common attributes in these measurements were width and depth of the defects. Therefore, in the rest of the evaluation, width and depth were considered the measurements obtained from the PTMS.

The PTMS inspection scenario is illustrated in figure 3.1. The 3D points (X_w, Y_w, Z_w) representing the defects on pantographs lies at the intersection of the pantograph surface and an imaginary laser plane. These 3D points are projected on an image by the camera as 2D points (p, q). The world coordinate and camera coordinate system are also illustrated in the figure. The mapping between 2D-3D points, as in equation 3.1, is represented by a planar homography transformation, because 3D points lie on a plane, and therefore, their 'z' component is zero.

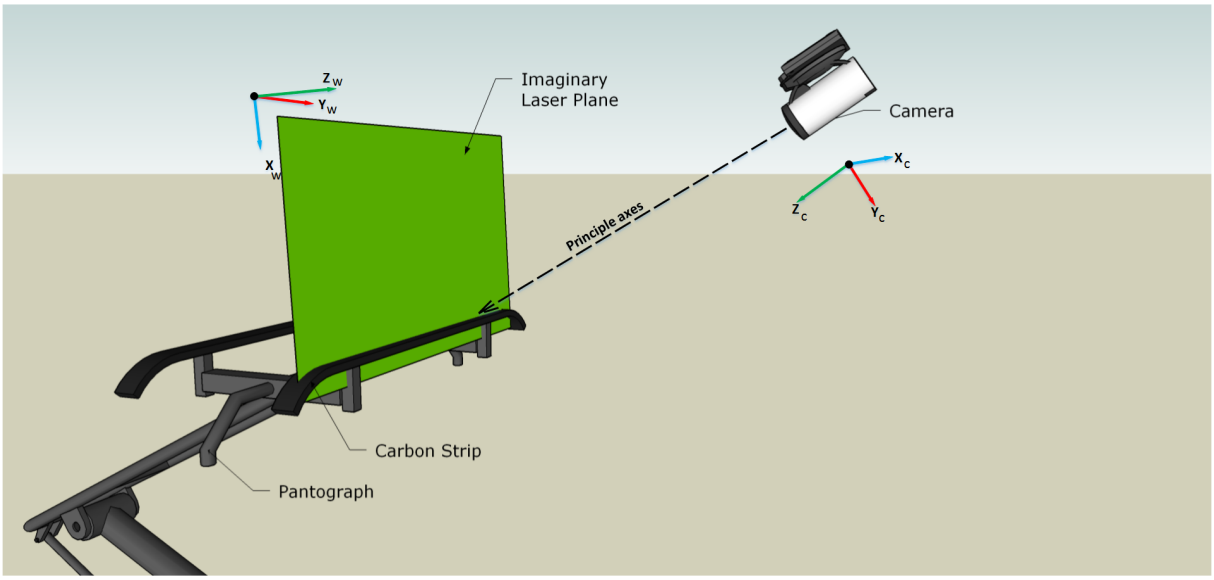


Figure 3.1: PTMS inspection scenario: world coordinates (X_w, Y_w, Z_w) and camera coordinates (X_c, Y_c, Z_c).

Such planar homography is given by equation 3.2, where K, R and T are the camera intrinsic, rotation and translation matrices, respectively.

$$\begin{bmatrix} p \\ q \\ 1 \end{bmatrix} = K[R|T] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \frac{1}{Z_c} \tag{3.1}$$

$$\begin{bmatrix} p \\ q \\ 1 \end{bmatrix} = K * \left[\begin{array}{ccc|c} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{array}\right] * \begin{bmatrix} X_w \\ Y_w \\ 0 \\ 1 \end{bmatrix} \frac{1}{Z_c} = \underbrace{K * \begin{bmatrix} r_{11} & r_{12} & t_1 \\ r_{21} & r_{22} & t_2 \\ r_{31} & r_{32} & t_3 \end{bmatrix}}_{H} * \begin{bmatrix} X_w \\ Y_w \\ 1 \end{bmatrix} \frac{1}{Z_c} \tag{3.2}$$

In this study, the simulation of PTMS was carried out under the conditions of camera misalignment, in terms of camera orientation (r_x, r_y, r_z) and translation (t_x, t_y, t_z) in all three directions. Figure 3.2 shows the simulation procedure used for evaluating the effects of camera misalignment.

The simulation database was generated based on modeling a pantograph with carbon strips of length 1.2 meters in length and about 30-50 mm in width and 30 mm in thickness. Five various defects (as illustrated in figure 2.15) were considered to be randomly present on a pantograph when the train passes under the PTMS. Each defect represents errors measured in terms of width (max. 50mm) and depth (max. 30mm). Assuming 200 such train passages under the PTMS, 1000 measurements in the database were obtained for testing, which were considered as known 3D measurements or points representing width and depth of defects (W^{known} and D^{known}). Then, with the help of known K, R, T values, known 2D points on the image were determined.

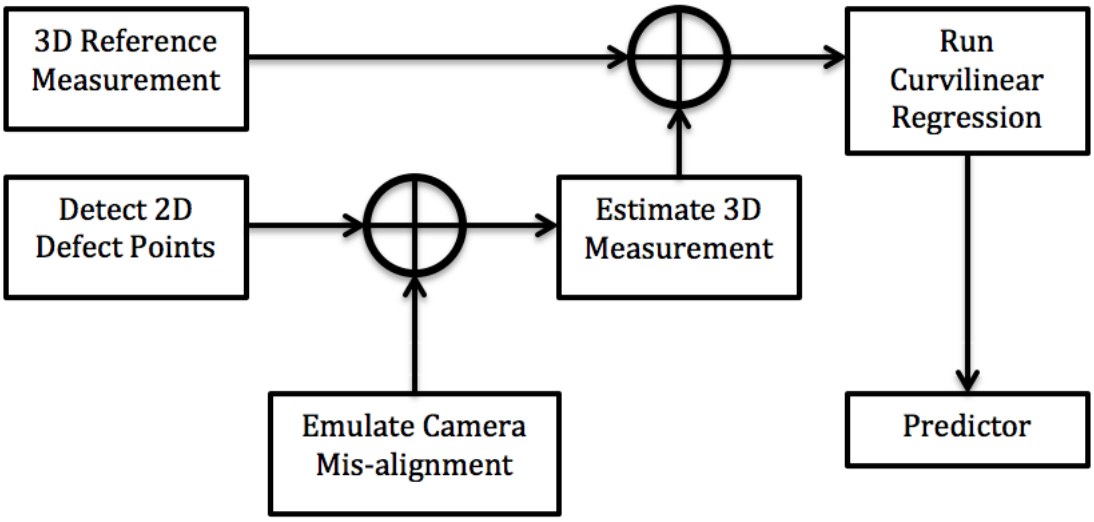


Figure 3.2: Simulation procedure.

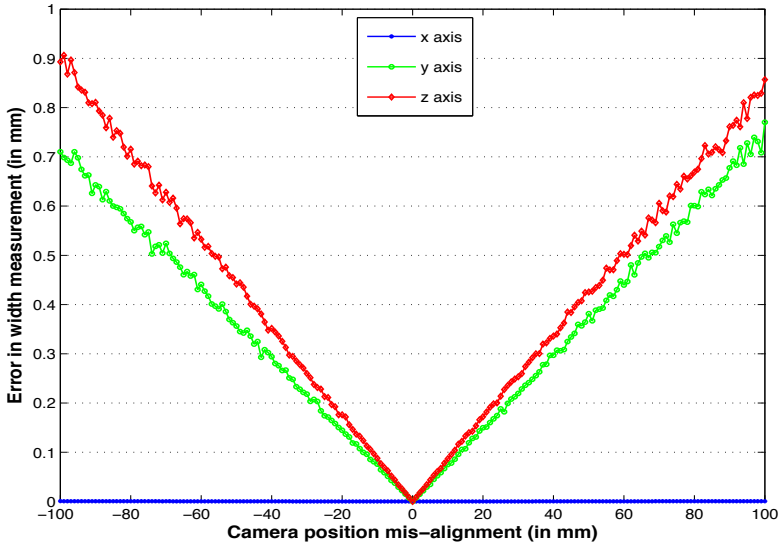
The camera misalignment was emulated in terms of orientations (ranging between -40 to +40 degrees) and position shifts (ranging between -100 to +100 millimeter). As a result of emulation, misaligned 2D points representing noisy defect points were obtained. As explained in the working principle of the PTMS in section 2.2.3, the 3D measurements (W^{est} and D^{est}) were estimated based on back-projection of 2D misaligned points from the image plane onto the laser plane using the known planar homography (Hartley and Zisserman, 2004), which was initially estimated by offline calibration process. Here, the camera intrinsic and extrinsic properties were assumed to be known. This means that the cameras are calibrated offline using Matlab tool (Bouguet, 2008).

In this way, the mean squared error was computed as in equations 3.3 and 3.4, which represents the accuracy of PTMS system when camera was misaligned by a specific degree of orientation or position shift in a specific direction.

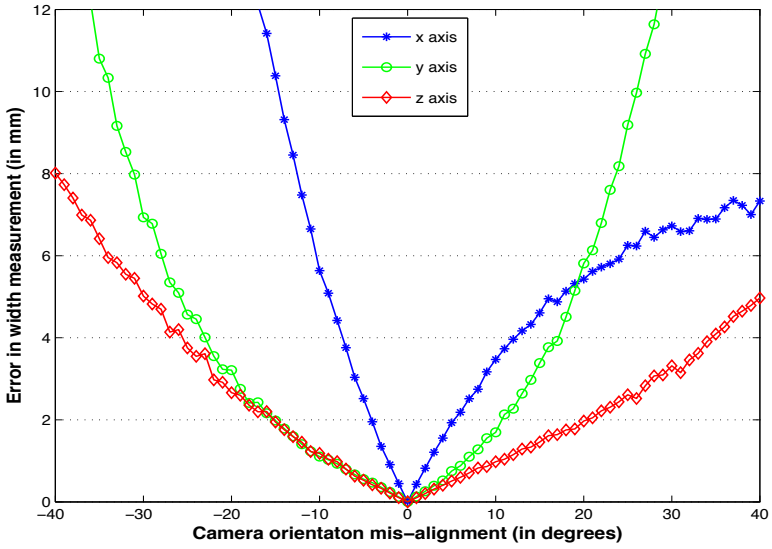
$$Error_{width} = ||W^{known} - W^{est}||_2 \tag{3.3}$$

$$Error_{depth} = ||D^{known} - D^{est}||_2 \tag{3.4}$$

By observing the $Error_{width}$ and $Error_{depth}$ for various camera misalignment, the effects of camera misalignment were studied. Further, the error was fed into a regression process. Considering each camera misalignment component as a variable and estimated error as a response, the error was modeled using appropriate regression models. When the error data was modeled, the parameters of that model was further utilized to predict tolerances of misalignment that the camera unit can withstand without affecting the system accuracy upto the acceptable levels.



(a) $Error_{Width}$ Vs t_x, t_y, t_z

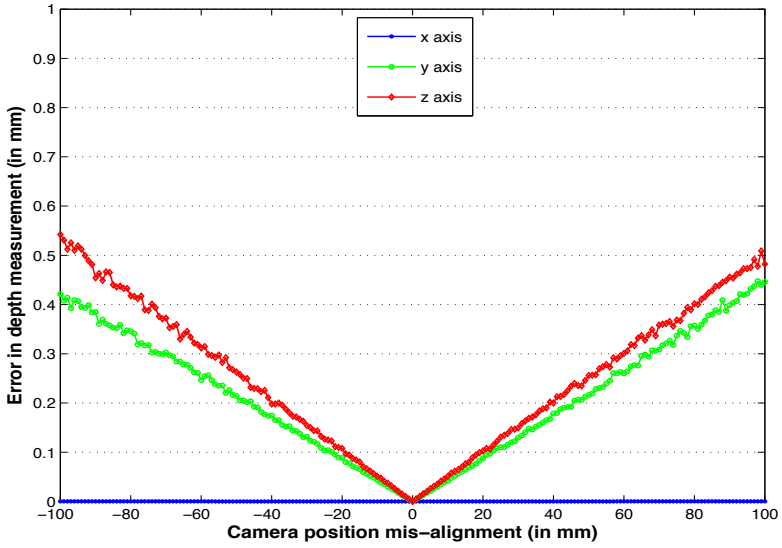


(b) $Error_{Width}$ Vs r_x, r_y, r_z

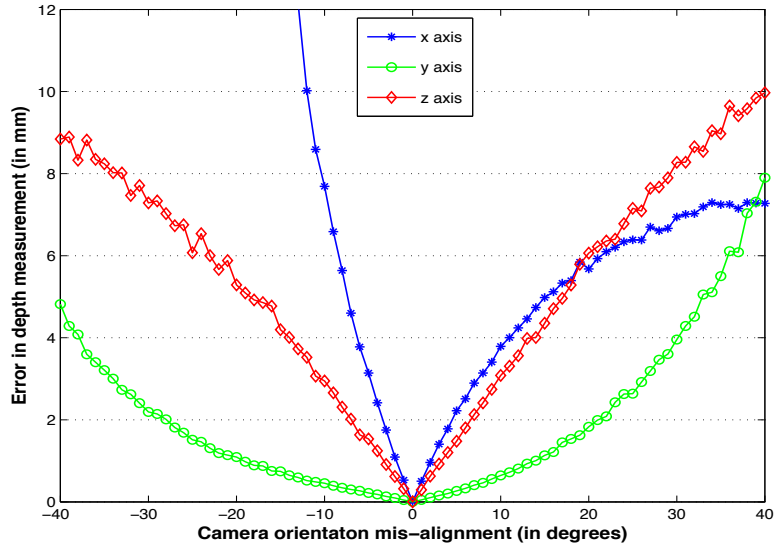
Figure 3.3: Variation of error in 3D width measurements of the defects, due to changes in camera position and orientation about its camera center.

3.1.2 Error Analysis

The result of error variation due to the camera misalignment in three directions can be seen in figures 3.3 and 3.4. Here, the error was measured in millimeters, camera orientation (r_x, r_y, r_z) in degrees and camera position (t_x, t_y, t_z) in millimeters. Please refer to figure 3.1, for insight into camera coordinate system for the direction of camera's rotation and translation. Looking at figures 3.3 and 3.4, it is obvious that the camera misalignment has an effect on the system's accuracy. Another work (Dosovitskiy, Springenberg, Riedmiller, and Brox, 2014), also showed



(a) $Error_{Depth}$ Vs t_x, t_y, t_z



(b) $Error_{Depth}$ Vs r_x, r_y, r_z

Figure 3.4: Variation of error in 3D depth measurements of the defects, due to changes in camera position and orientation about its camera center.

similar results of the effect of camera rotation.

The effect of camera translation was not very significant. The accuracy of the system was least affected by the translation in 'x' direction (t_x). Camera translation t_y and t_z did not affect the accuracy to a large extent but acquired an error of less than 1mm. This is the case for both width error in figure 3.3(a) and depth error in figure 3.4(a), however, the depth error was less affected by camera translation than the width error.

The effect of camera rotation was more interesting and significant, because of the camera arrangement, i.e., non-perpendicularity to the object plane. In figures 3.3(b) and 3.4(b), a

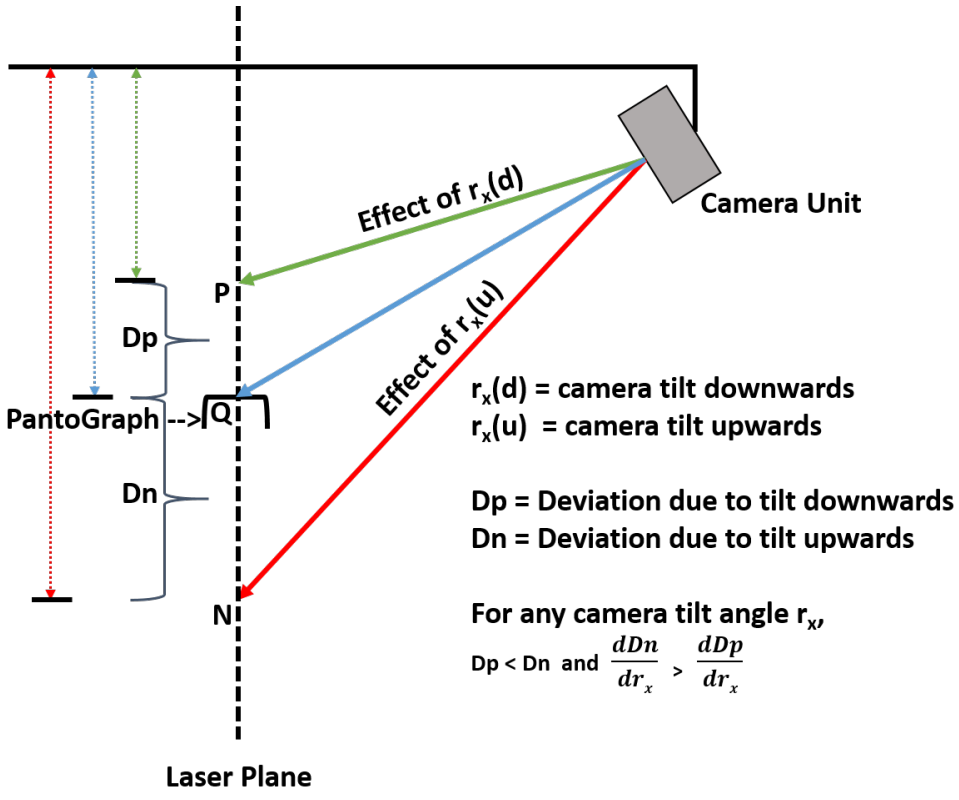


Figure 3.5: Projective geometric effects of camera tilt angle in PTMS.

symmetrical variation in both width and depth error due to camera rotations r_y and r_z, and an unsymmetrical variation due to camera rotation r_x were observed.

Camera pan (rotation r_y) affected the width error more than the depth error. As the camera rotates around the y axis, the horizontal x component of the 2D defect points is displaced more than the vertical y component. Since the width is a function of the x component, the width error is greater than the depth error. Exactly the opposite was observed when the camera roll (rotation r_z) affected the depth error more than width error. This is because, when the camera rotates around z axis, the y component of the 2D defect points are displaced more and depth is a function of the y component. Even few degrees of rotation resulted in a significant change in the error of more than 1mm.

However, special cases are the errors due to camera rotation r_x. The camera tilt (rotation r_x) seemed to be very critical and affected the accuracy of the system more than any other camera motions. This is due to the non-perpendicularity and the camera placement at an angle with respect to the laser plane, which results in an unsymmetrical variation of both depth and width errors. Moreover, it was also observed that the error increases more rapidly when the camera tilts upwards than when the camera tilts downwards. This behavior can be explained using projective geometric properties.

Consider the illustration in figure 3.5, where the camera is placed at the same position and orientation as in PTMS. Let us consider that the camera captures the image of the defect point that lies on the laser plane. By ray-to-plane intersection the 3D defect point is estimated to be at the location Q. Now, when the camera is tilted upwards with (rotation $r_x(u)$), the defect point

is seen slightly lowered in the image and its corresponding 3D estimated point is at N and the deviation of the reconstructed defect point is represented as D_n. Similarly, when the camera is tilted downwards (rotation $r_x(d)$), the estimated defect point is at P and the deviation is D_p. It is important to note that for the same angle of rotation (r_x), the deviation in the reconstructed defect point is different due to the alignment of the camera with respect to the laser plane, i.e., $D_p < D_n$. If the camera is further tilted, i.e., as the angle of rotation increases, the rate of increase in D_n is higher than the rate of increase in D_p. This is the reason why in figures 3.3(b) and 3.4(b), the errors rapidly increase (corresponding to D_n) with camera rotation $-r_x$ compared to the camera rotation $+r_x$.

3.1.3 Error Modeling

After visually inspecting the effects of camera misalignment on the width and depth error, a linear model was used to fit the error data variation due to camera translation, and a curvilinear model was used to fit the error data variation due to camera rotation.

The linear and curvilinear models for translation and rotation, respectively, are mathematically shown in equations 3.5 and 3.6.

$$error = p0 + p1 * (component) \tag{3.5}$$

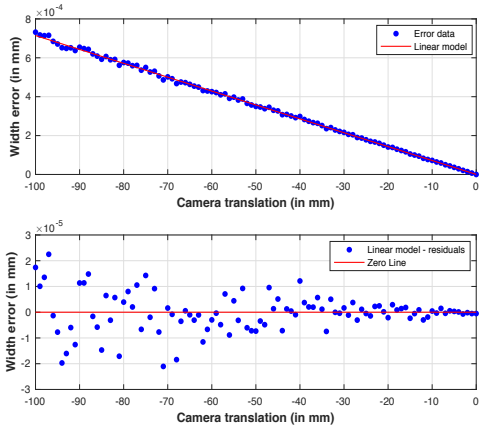
$$error = p0 + p1 * (component) + p2 * (component)^2 \tag{3.6}$$

where, $p0, p1, p2$ are the model parameters and $component$ represents the misalignment component, e.g., camera rotation around x axis, in positive direction r_x^+. The line/curve fitting was carried out separately for all camera misalignment components, i.e., camera rotations (r_x^+, r_x^-, r_y^+, r_y^-, r_z^+, r_z^-) and camera translations (t_x^+, t_x^-, t_y^+, t_y^-, t_z^+, t_z^-), where " $+$ " and " $-$ " represents the direction of camera rotation and translation.

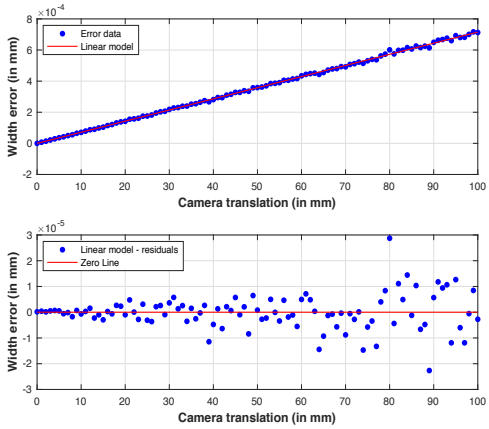
The model fitting of width error with camera translations and rotations are as shown in figures 3.6 and 3.8 respectively. Model fitting for depth error are shown in figures 3.7 and 3.9 respectively. The figures also show the residual plot in which one can see how good was the data fit. In figures 3.8(a) and 3.9(a), there are fewer error points compared to the other error data. This is because after ≈25 degrees of camera tilt upwards, the error is so high that they are treated as outliers in the statistical sense.

The resulting model parameters for both width and depth error over all camera misalignment components were computed and are shown in tables 3.1 and 3.2. Here, $p0, p1, p2$ are the model parameters and the model fitting quality is represented by the Root Mean Squared Error (RMSE) values. The RMSE values for all the data are less than unity and that signifies a good model fit (estimation of model parameters) with a confidence level of 95%.

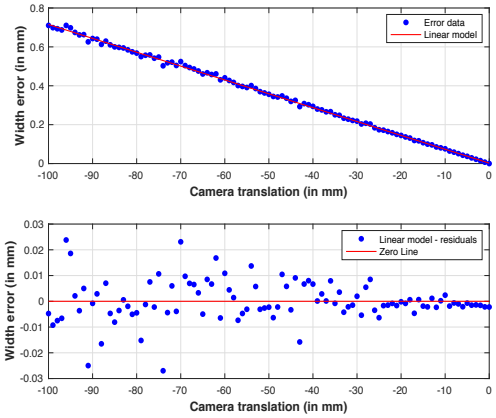
The model parameters were further used as a predictor as mentioned in (Jain, 1991), after which the camera misalignment could be computed for a given error value. This was used to determine the tolerances of camera misalignment for an acceptable error in the system.



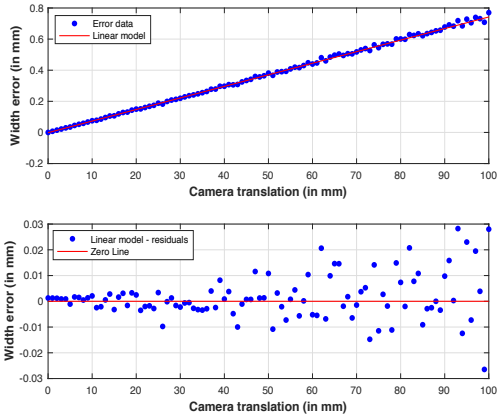
(a) Width versus t_x^-.



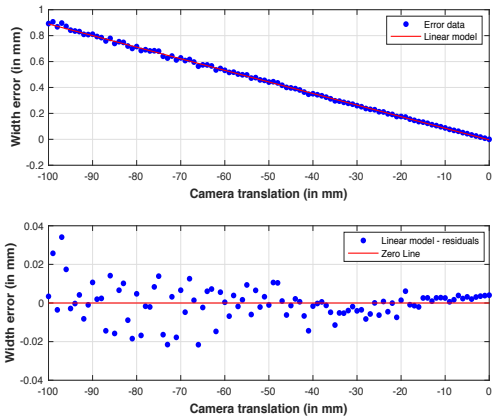
(b) Width versus t_x^+.



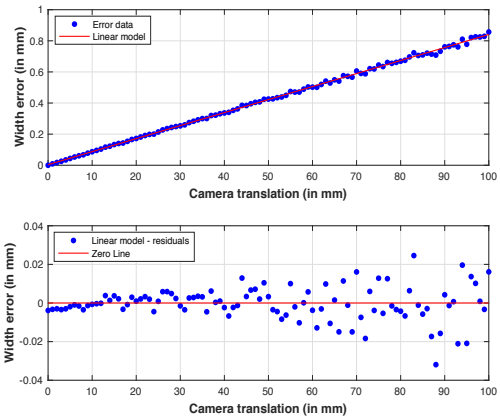
(c) Width versus t_y^-.



(d) Width versus t_y^+.

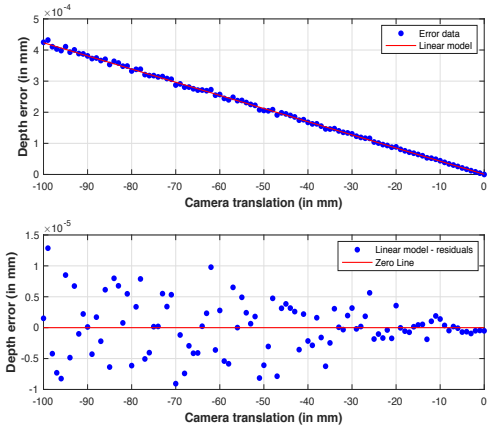


(e) Width versus t_z^-.

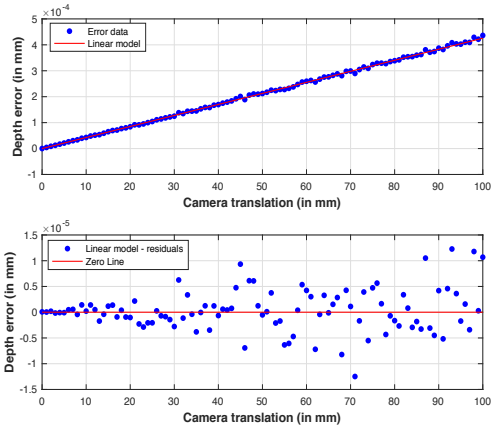


(f) Width versus t_z^+.

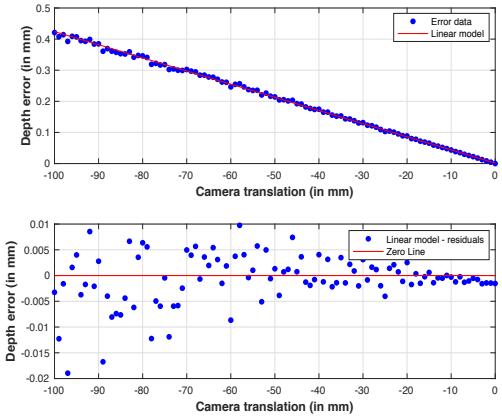
Figure 3.6: Linear model fitting and residual plots for variation of width error with camera translations.



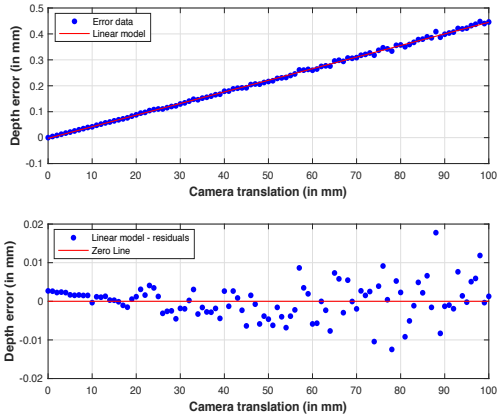
(a) Width versus t_x^-.



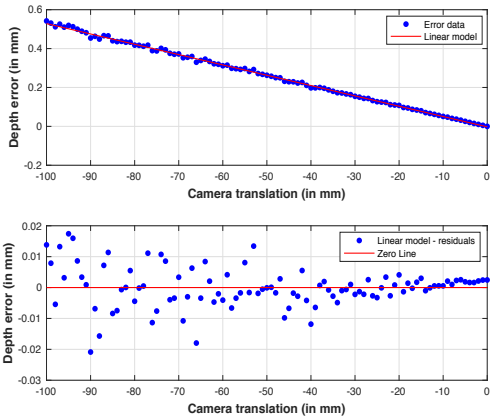
(b) Depth versus t_x^+.



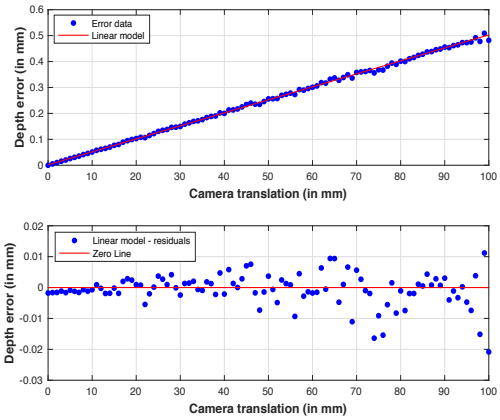
(c) Depth versus t_y^-.



(d) Depth versus t_y^+.

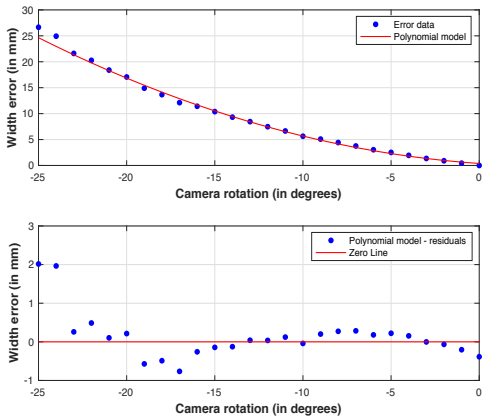


(e) Depth versus t_z^-.

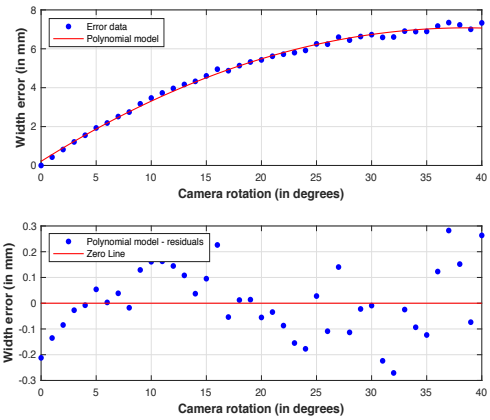


(f) Depth versus t_z^+.

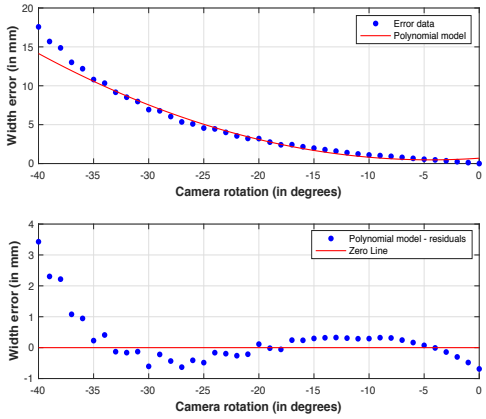
Figure 3.7: Linear model fitting and residual plots for variation of depth error with camera translations.



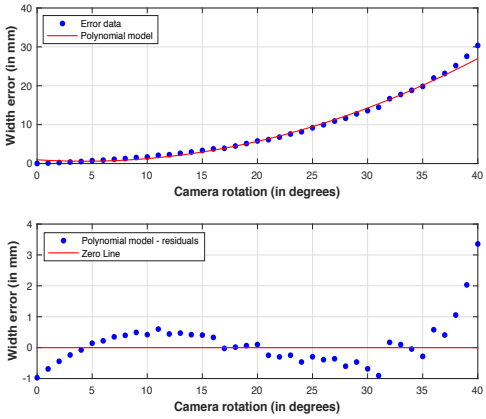
(a) Width versus t_x^-.



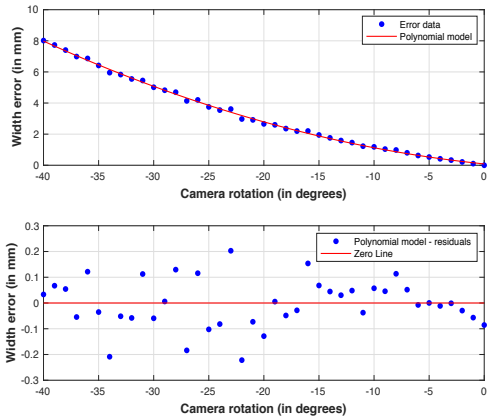
(b) Width versus t_x^+.



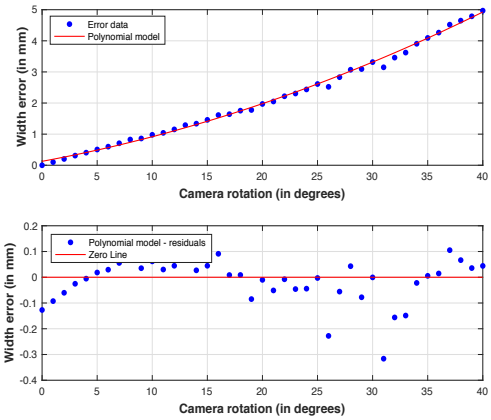
(c) Width versus t_y^-.



(d) Width versus t_y^+.

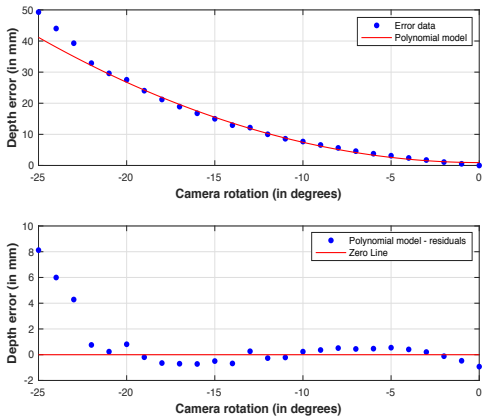


(e) Width versus t_z^-.

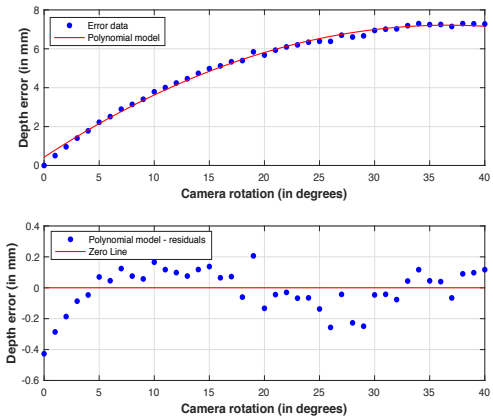


(f) Width versus t_z^+.

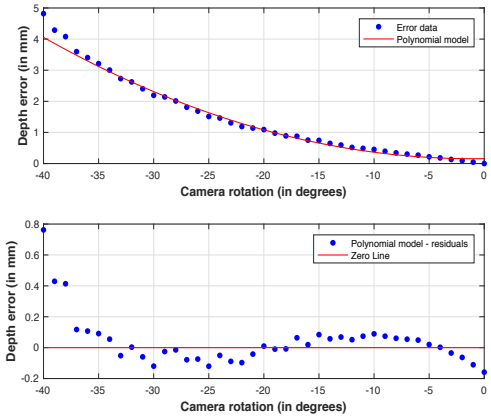
Figure 3.8: Curvilinear model fitting and residual plots for variation of width error with camera rotations.



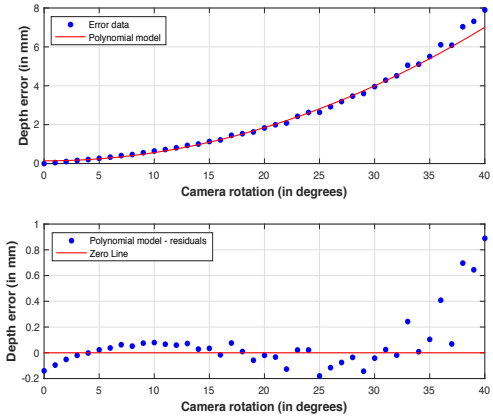
(a) Width versus t_x^-.



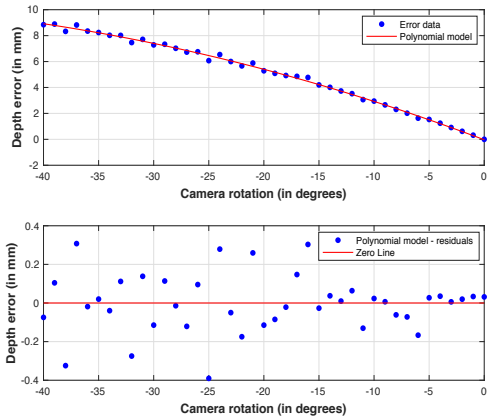
(b) Depth versus t_x^+.



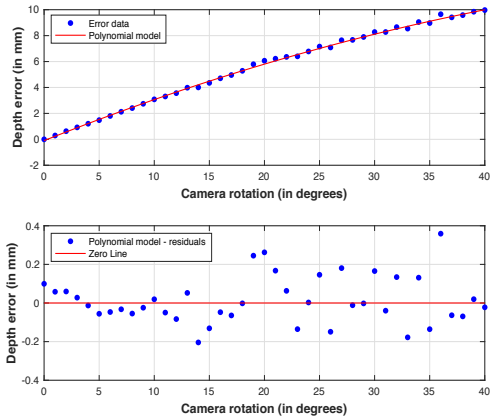
(c) Depth versus t_y^-.



(d) Depth versus t_y^+.



(e) Depth versus t_z^-.



(f) Depth versus t_z^+.

Figure 3.9: Curvilinear model fitting and residual plots for variation of depth error with camera rotations.

Data	Linear		
f(x),x	p_0	p_1	RMSE
width, t_x^-	5.13e-07	-7.14e-06	6.36e-06
width, t_x^+	-1.73e-07	7.15e-06	5.47e-06
width, t_y^-	2.22e-03	-7.13e-03	6.47e-03
width, t_y^+	-1.28e-03	7.43e-03	6.41e-03
width, t_z^-	-4.04e-03	-8.93e-03	7.34e-03
width, t_z^+	3.84e-03	8.37e-03	7.20e-03
depth, t_x^-	5.09e-07	-4.23e-03	4.33e-06
depth, t_x^+	-6.48e-08	4.26e-06	3.51e-06
depth, t_y^-	1.54e-03	-4.23e-03	4.11e-03
depth, t_y^+	-2.7e-03	4.47e-03	4.118e-03
depth, t_z^-	-2.45e-03	-5.30e-03	5.62e-03
depth, t_z^+	1.77e-03	5.01e-03	3.83e-03

Table 3.1: Model parameters estimated for translation

3.1.4 Discussions

It became clear that there is an effect of camera misalignment on the accuracy of PTMS. However, the effect of whether the camera misalignment is significant or not, depends on the application, i.e., the PTMS. Before deploying the PTMS, the system was calibrated using the Matlab toolbox (Bouguet, 2008) and tested for accuracy in the factory to obtain acceptance for deployment. This factory acceptance test adheres to customer requirements. And BaneDanmark (Danish Railway Network) has specified the requirement for an acceptable width or depth error upto a maximum of 1mm. So, for the PTMS, 1mm is the critical error beyond which the system fails to adhere the accuracy specified and is not reliable for the inspection of defects on the pantographs.

So, considering the critical error of 1mm, the effect of camera misalignment was very significant especially in terms of rotation but was not very significant in terms of translations misalignment for the PTMS. The PTMS system showed high sensitivity in accuracy for camera tilt compared to any other camera motion.

Additionally, an acceptable error of 0.5mm was considered and the predictor in equations 3.5 and 3.6 was used to determine the tolerances of camera misalignment in terms of rotation and translations in all three directions. These computed tolerances are given in table 3.3. These tolerances defined the maximum misalignment the camera can have without compromising the accuracy and reliability of the PTMS system. Tolerances defined in the table are related to camera misalignment only in one dimension. The tolerances over a combination of the camera misalignment in all three dimensions can be easily estimated using the proposed simulation procedure (shown in figure 3.2), where emulation of camera misalignment takes place over three dimensions. However, for practical purposes, tolerances over one dimension is more useful for testing the sturdiness of the systems (e.g. camera rigs) in those specific dimensions.

The significance of the effect and tolerances of camera misalignment, both were dependent on the PTMS system. Although this result is not universal, the methodology used (explained

Data	Polynomial			
f(x),x	p_0	p_1	p_2	RMSE
width, r_x^-	0.386	-0.235	0.029	0.394
width, r_x^+	-0.212	0.356	-0.005	0.142
width, r_y^-	0.688	0.095	0.011	0.468
width, r_y^+	0.974	-0.177	0.020	0.563
width, r_z^-	0.085	-0.072	0.003	0.102
width, r_z^+	0.127	0.065	0.001	0.076
depth, r_x^-	0.926	-0.014	0.064	0.912
depth, r_x^+	0.426	0.370	-0.005	0.141
depth, r_y^-	0.158	0.005	0.002	0.096
depth, r_y^+	0.140	-0.001	0.004	0.106
depth, r_z^-	-0.032	-0.319	-0.002	0.150
depth, r_z^+	-0.099	0.338	-0.002	0.130

Table 3.2: Model parameters estimated for rotations

Tolerances	X axis (deg/mm)	Y axis (deg/mm)	Z axis (deg/mm)
Rotation (width)	-0.46 to 0.82	-2.96 to 4.27	-4.73 to 5.12
Rotation (depth)	-0.11 to 0.19	-12.57 to 9.21	-1.68 to 1.79
Translation (width)	-6.97e04 to 6.98e04	-69.83 to 67.42	-56.41 to 59.20
Translation (depth)	-11.82e05 to 11.75e04	-117.93 to 112.35	-94.67 to 99.44

Table 3.3: Tolerances for camera misalignment, given the system inaccuracy limit as 0.5mm.

in section 3.1.1) is universal, which is independent of any single camera 3D inspection system. The methodology can be repeated for any single camera 3D measurement system to obtain the tolerances of the camera misalignment given a known 3D reference, its corresponding 2D points. Thus, obtained tolerance values help in a better mechanical design of the camera fixtures and rigs to minimize defect measurement errors caused by camera misalignment.

All the above tests and discussions regarding the significance and tolerances of effect of camera misalignment, are evident enough to reject a part of *Null Hypothesis I* (stated in section 1.4). From this, it can be concluded that the 3D reconstruction accuracy significantly decreases when the camera is misaligned in a single camera system, where the significance is determined by the application scenario.

3.2 Adoption of Feature Based Calibration

The concept of camera misalignment and its effects on single camera system, i.e., the PTMS was discussed in section 3.1. In the PTMS, the camera is prone to misalignment during transportation and after deployment due to natural causes (e.g., wind) or manual intervention that can occur during servicing. Consequently, this has an effect on the accuracy of the 3D reconstructed data in PTMS. However, pantograph misalignment can also adversely affect the accuracy of the system. The pantographs are allowed to move upwards and downwards to have sufficient up-

thrust to the contact wires in the operation of PTMS. This is termed as Uplift. The motion and speed of the train in combination with the uplift of the pantographs, causes pantograph misalignment from its original position (prominently in the vertically direction) and orientations (roll, pitch and yaw directions). In such cases of camera and/or pantograph misalignment, the original camera calibration data is not useful anymore and therefore leads to inaccurate 3D measurements. The solution to obtain a robust system is only through re-calibrating the camera, i.e., to re-estimate the position and orientation of camera with respect to pantograph and use the new calibration parameters to recover 3D measurements.

In the PTMS, the camera is calibrated offline using traditional checkerboard calibration (CBC) technique (Bouguet, 2008). The same calibration technique cannot be used for online re-calibration process because, it is impractical to hold a checkerboard at the position of the pantograph on top of the rails, when the PTMS is deployed and operational. Moreover, unmounting the PTMS for offline calibration will require manual intervention and this leads to a large maintenance and service time. This surely affects the train network infrastructure.

In geographical areas where sunlight is quite prominent, the PTMS operation is affected invariably. Sunlight contains infrared rays, and hence, masks the laser lines, when the camera captures the infrared images. Therefore, the images captured by the camera can be very noisy. Rain or dirt on the camera box, will not affect the quality of image, because the cameras capture infrared image of the laser lines, which does not capture raindrops or dirt that are seen in the normal cameras. Flash under or over exposure is also a cause for noisy images similar to that of sunlight. Poorly visible or occluded profile image or laser misalignment can cause noise in the profile image that makes it difficult for the PTMS to analyze the images for profile detection. Motion blur can also be caused due to vibration in the mount. This could indeed mis-detect laser line positions and thereby deteriorate the quality of operation of the PTMS.

In this case, a robust PTMS is required to maintain the quality of the 3D measurements in spite of camera/pantograph misalignment or noisy images, and save servicing and maintenance costs at the same time. Therefore a feature based calibration (FBC) process is proposed for the PTMS in order to overcome the inaccuracies in the system and save practical costs. For the PTMS, the FBC refers to the estimation of camera pose with respect to the pantograph by using interesting points detected in 2D image and their corresponding known 3D points.

An online re-calibration framework (Li and Lu, 2010) aimed at automatic calibration using SIFT (Scale-Invariant Feature Transform) features cannot be used directly on the PTMS because the framework was proposed for stereo systems, which focuses on obtaining feature matches between the image pairs. However, in the PTMS the infrared image contains captured laser lines and hence SIFT would not be suitable to figure out interesting feature points on the lines. There exists other FBC solutions (Basso, Levorato, and Menegatti, 2014; Mavrinac, Chen, and Tepe, 2008), which focus more on stereo vision and hence not adoptable for the PTMS.

For single camera systems, there exists markerless re-calibration schemes which do not rely on the points (like in checkerboard), but on the structure of the scene (Carr, Sheikh, and Matthews, 2012) or the structure of objects (Drummond and Cipolla, 1999). A camera laser online calibration (Levinson and Thrun, 2013) is applied for autonomous robots/vehicles for augmenting dense color information from the camera image to the sparse depth measurements obtained by a laser. A feature based single camera vision system (Cesetti, Frontoni, Mancini, Zingaretti, and Longhi, 2009) was proposed for the safe landing of an unmanned aerial vehicle.

Most of the above discussed solutions cannot be used for the PTMS, because their methodologies vary in the way the features are detected for calibration. In PTMS, the target structure is not complex but a well defined structure, which makes it simpler to model for calibration. However, the challenge is to detect feature points in the noisy image and calibrate using very few feature points based on a strategic scheme defined by the frequency of online re-calibration. Thus, obtained calibration must also be resilient to PTMS specific camera or pantograph misalignment.

3.2.1 Proposed Re-calibration Methodology

In our paper titled "Online Re-calibration for Robust 3D measurement using Single Camera - PantoInspect Train Monitoring System" [details in chapter 10], an integrated solution of feature based calibration methodology was proposed to increase the usability of the PTMS. This involved the following:

- Adoption of feature based calibration in the PTMS instead of CBC over various practical schemes of implementation.
- Evaluation of the performance of four state-of-art pose estimation algorithm using fewer points.
- Evaluation of FBC in comparison to CBC in terms of performance and robustness against pixel noise and camera or pantograph misalignment effects.

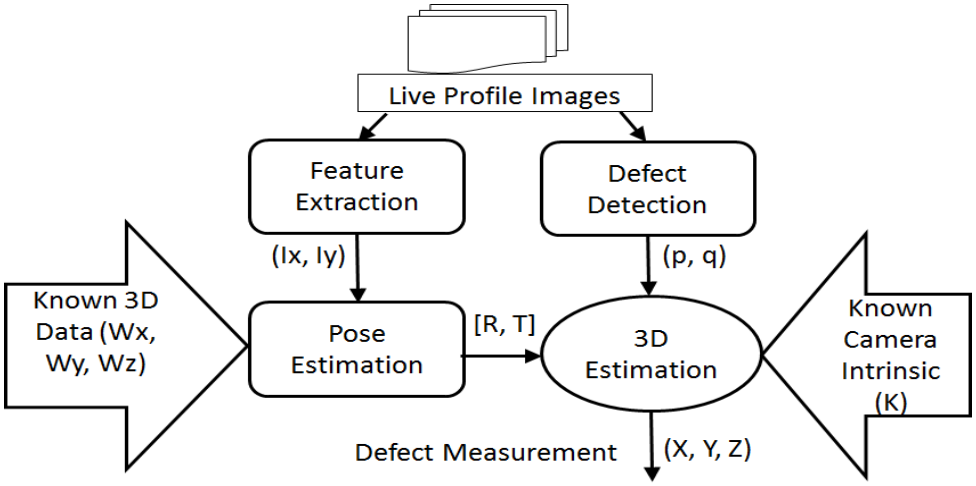


Figure 3.10: Proposed feature based calibration for the PTMS.

The proposed feature based calibration for the PTMS is shown in figure 3.10. The near infrared images of the laser lines that define the surface of the carbon strip on the pantograph was termed as *Profile Image*. These profile images were used to detect defects and estimate the 3D measurements of it. The proposed FBC used the same profile image to re-calibrate the system on-the-fly. Typically, FBC methodology consists of a 2-step process, (a) Feature extraction and (b) Pose estimation. A number of feature points were extracted from an image

and were used to estimate the camera pose, i.e., translation and rotation, and subsequently 3D estimation of the defects.

The proposed FBC method relied on known 3D points and their corresponding 2D point features extracted in the profile images. The camera is calibrated offline, meaning that the camera intrinsic parameters (focal point, principle axes) were known prior to FBC operation.

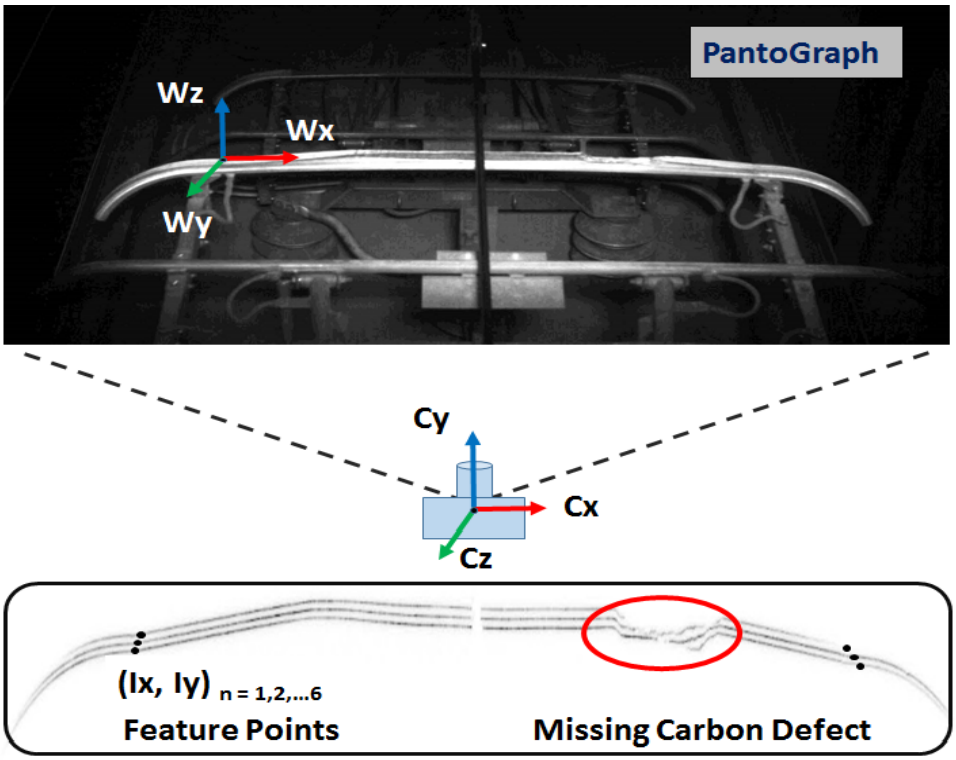


Figure 3.11: Profile image with a representation of camera and world coordinates.

Feature Extraction

In the profile images, the laser lines traverse along the carbon strips of the pantographs. Consider figure 3.11, where a missing carbon defect was captured by the infrared camera. Figure 3.11 also shows the shape of the 3 laser lines that traverse the shape of the pantograph and bends on both the ends, where carbon strip ends. Each laser line has two end points where the line bends and totally 6 such points in 3 laser lines represents the most distinctive feature that can be used for pose estimation. Therefore, these 2D feature points (I_x, I_y), in principle, can be extracted from the profile image using edge detection and a priori knowledge of the physical model of the pantograph. However, feature points thus obtained will not be perfectly noiseless. So, to obtain a noiseless feature point for the sake of evaluation of this study, manually annotated feature points are considered. The world coordinate origin lies on the pantograph as shown in the figure 3.11. Since each pantograph had standard dimensions, the 3D points corresponding to the six 2D feature points were known and were expressed in world coordinates.

Pose Estimation

Using both known 3D points in world coordinate system (W_x, W_y, W_z) and 2D feature points in image coordinate system (I_x, I_y), the camera pose with respect to the world coordinate system

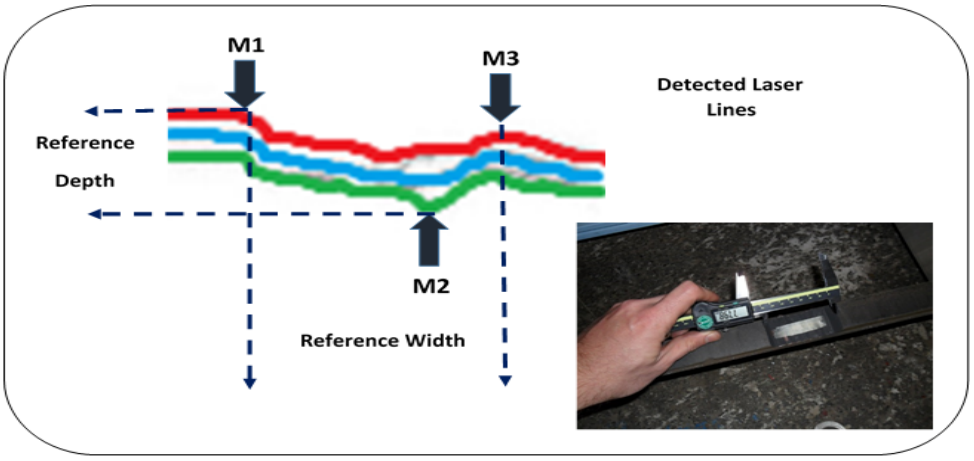


Figure 3.12: Defect identification and measurement.

was estimated. Since the 3D points are lying on the imaginary plane spanning all points lying on the pantograph, the relationship between 3D points and the 2D feature points merely becomes a projective homography mapping, which is represented as a projective matrix. This projective homography can be further decomposed into the rotation matrix and translation vector that refers to as the camera pose parameters ($[R, T]$).

3D Estimation

The 3D reference points is assumed to lie on a plane. So, conversion from 2D to 3D is merely a ray-plane intersection. From each of the 2D points (say (p, q)) that defines the defect, their 3D counterparts (say (X, Y, Z)) were estimated as a back-projection of 2D point using the projective homography (explained in section 2.1.1). Mathematically, the homography relationship between 2D and 3D points are expressed as in equations 3.1 and 3.2.

Defect Detection

The profile images were analyzed to detect the defects. Whenever the lines in the profile images are not straight, then that is potentially a defect, e.g., carbon missing defect, as highlighted in figure 3.11.

Each defect was characterized by width and depth measurements. For each defect, potentially three control points were identified that can be used to calculate width and depth of the defect. The missing carbon defect detected in 3.11 is processed to identify the control points M_1, M_2, M_3 as shown in figure 3.12. These control points are back-projected to their 3D positions as $\widehat{M_1}$, $\widehat{M_2}$, $\widehat{M_3}$ using equation 3.2.

$$Width = \widehat{M_3} - \widehat{M_1} \qquad and \qquad Depth = AbsMax(H1, H2) \tag{3.7}$$

$$where, \qquad H1 = \widehat{M_1} - \widehat{M_2} \qquad H2 = \widehat{M_3} - \widehat{M_2}$$

The width and depth of the defects were computed using equation 3.7.

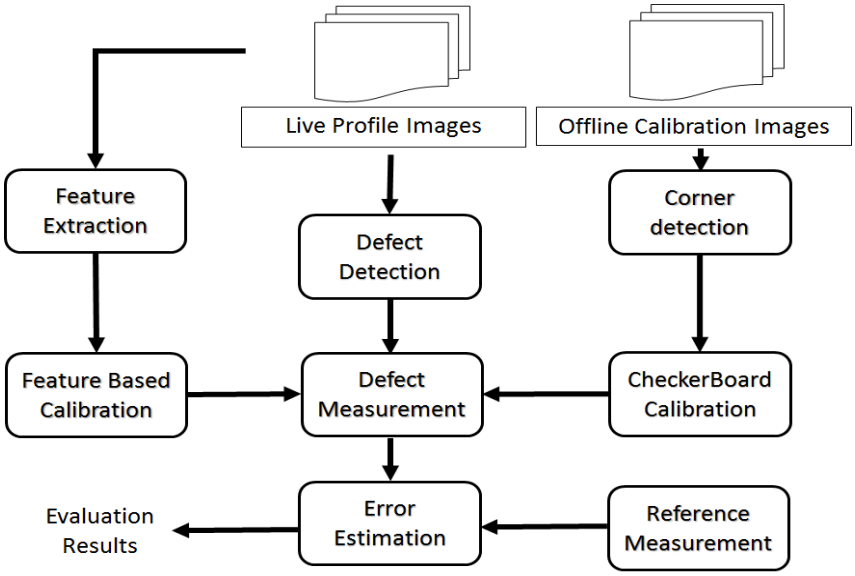


Figure 3.13: Evaluation of feature based calibration for PTMS.

3.2.2 Evaluation

The evaluation procedure was carried out to measure the performance traits of the FBC adopted PTMS in comparison to traditional CBC adopted PTMS. Figure 3.13 depicts the methodology used for evaluating the FBC adopted PTMS. The live profile images of the PTMS were analyzed to detect the defects. The camera of PTMS was calibrated offline using 20 images of checkerboard pattern. Live profile images of PTMS was used to calibrate the camera using FBC with only 6 points to potentially achieve the calibration quality as for CBC. To test the accuracy of both CBC and FBC, the calibration parameters estimated by both of the techniques were used in conjunction with back-projection of 2D defects point, to measure the defect dimensions (width and depth) in real metrics. A reference defect dimensions were measured using industrial caliper as shown in figure 3.12. The defect measurements obtained from both the techniques, was compared with the reference measurements to obtain an error metric for the evaluation.

Reference measurements and evaluation tests were carried out on real pantographs (BR and EG types) of trains in Denmark, BaneDanmark (Rail Net Denmark). Each of the pantographs had 4 types of defects and 5 profile images from each pantograph comprising of 40 samples were used for the test. Each of the defects was manually annotated for obtaining a proper reference dataset. Table 3.4 shows the reference measurements of pantographs used for evaluation tests.

The FBC could possibly be carried out under various operational modes. Every profile image can be used for FBC, but noisy profile image will worsen the re-calibration quality. Alternatively, FBC can be carried out at regular intervals, however, a stable re-calibration could be carried out only during servicing or maintenance periods. In this accord, two different operational schemes that was tested are as follows:

Scheme 1: FBC was carried out on every profile image and the defect is measured on those images using its respective calibration parameter.

Scheme 2: FBC is carried out on a random profile image and the measurement is carried out on all the remaining profile images with the same calibration parameters.

Measurement	Pantograph-type			
(in millimeters)	BR-type		EG-type	
Defects	Width	Depth	Width	Depth
Vertical Crack	2.38	17	5.88	20
Missing Carbon	77.98	17	39.04	20
Edge Crack	24.21	6	21.18	5
Abnormal Wear	19.36	6	14.78	5

Table 3.4: Reference measurements of defects of two pantograph types.

There are several pose estimation algorithms used for single camera system (Bouguet, 2008; Faugeras, 1993; Hartley and Zisserman, 2004; Lepetit, Moreno-Noguer, and Fua, 2009; Lu, Hager, and Mjolsness, 2000; Tsai, 1992; Zhang, 2000), which assumes that the camera intrinsic and reference 3D point coordinates are known. Out of many, four extensively used single camera pose algorithms, i.e., FBC-boug, FBC-zhang, FBC-gold and FBC-epnp, were chosen for evaluating adoption of FBC against CBC for PTMS. A brief description of these algorithms are shown in table 3.5. All these algorithms operate with $n \geq 4$ points.

Algorithm	Description
FBC-epfl (Lepetit, Moreno-Noguer, and Fua, 2009)	Unlike other methods, this is a non-iterative approach to PnP problem. Under PnP problem, 3D points are expressed in camera coordinate system and then, the Euclidean motion that aligns both world and camera references is used to retrieve [R,T]. This method adopts the idea of expressing n 3D points as weighted sum of four virtual control points, which reduces complexity and noise sensitivity.
FBC-boug (Bouguet, 2008)	This method initially estimates planar homography using the Quasi-Linear algorithm and recovers [R,T] parameters, which are further optimized to minimize re-projection error through Gradient Descent.
FBC-gold (Hartley and Zisserman, 2004)	This method estimates a projective geometric transformation using Gold Standard algorithm before recovering [R,T].
FBC-zhang (Zhang, 2000)	This method estimates planar homography using the Direct Linear Transformation followed by a non-linear optimization (Levenberg Marquardt) based on Maximum Likelihood criterion. Then, [R,T] are recovered using orthogonal enforcement.

Table 3.5: Single camera pose estimation algorithms and their description.

There are several challenges that has practical implications on the accuracy of PTMS. Poorly visible profile image, laser misalignment, flash under/over exposure, motion blur or sunlight will affect in detection of very few feature points and can introduce noise in the detected feature point locations. The pantograph can be linearly displaced in the vertical direction (uplift) and

FBC-type	R_x (Tilt)	R_y (Roll)	R_z (Pan)
epfl	7.17	3.50	9.46
boug	9.99	8.21	11.93
gold	11.51	5.96	10.99
zhang	10.86	6.7	11.30

Table 3.6: Absolute angular difference in degrees between CBC and FBC - scheme 1.

can be angular displaced in all three rotations (yaw, pitch and roll angles) during upthrust of pantograph to the catenary wire and/or the motion of the train.

3.2.3 State-of-the-art FBC algorithms

FBC was carried out using four algorithms, i.e., FBC-boug, FBC-zhang, FBC-gold and FBC-epnp on the data samples and obtained calibration parameters. The rotational parametric difference between FBC algorithms and CBC is noted as in table 3.6. This table shows the absolute angular difference (in degrees) of rotational parameters estimated using FBC (scheme 1) and CBC. Although the numbers gave an indication that one of the FBC method was better than the others, it is difficult to conclude without evaluating the accuracy and robustness of algorithms when adopted in the PTMS.

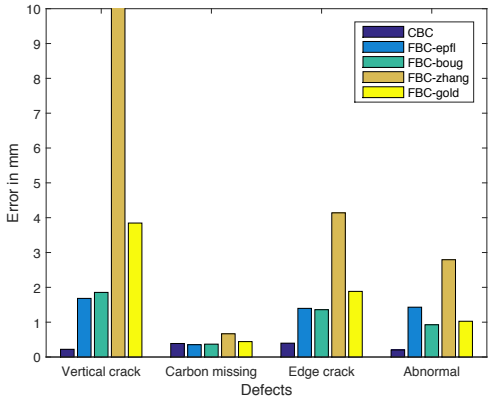
3.2.4 Accuracy of Measurements

Based on the evaluation procedure described in figure 3.13, the accuracy of width and depth measurements in millimeters was measured in terms of mean error of the measurements over the dataset samples and defects for two different operational schemes of PTMS.

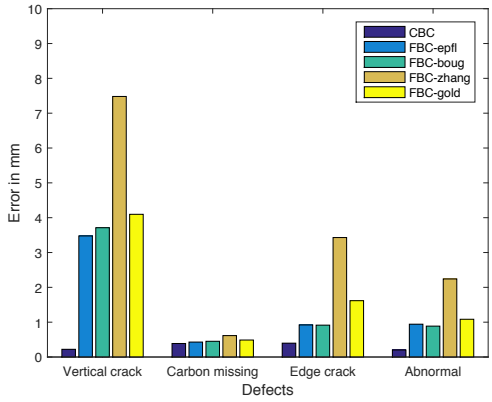
The results of width error measurements are as shown in figures 3.14(a) and 3.14(b). The observations of width error measurements are as follows, however more details are given in the publication specified in chapter 10:

- For large width defect (*missing carbon* defect), all FBC types performed equally and close to CBC.
- For lowest width defect (*vertical crack*), the width error in FBC-zhang increased drastically, however other types resulted 2-4 mm of error.
- For *edge crack* and *abnormal wear*, FBC-epfl, FBC-boug and FBC-gold introduced only 1-2 mm mean error compared to CBC.
- All FBC types are more sensitive to narrow widths ($< 5mm$) in scheme 2 than scheme 1.
- Overall, FBC-epfl and FBC-boug were observed to have performed close to CBC, with a maximum increase in mean error of about 1mm in *scheme 1* and 1.5mm in *scheme 2*.

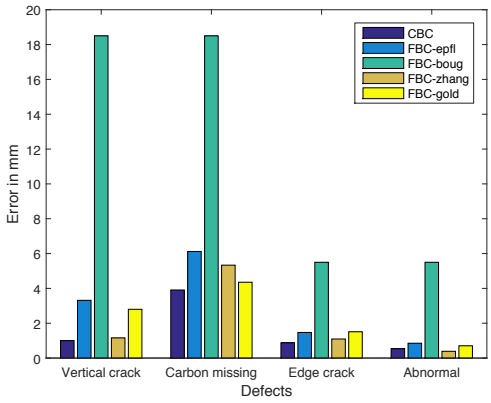
The results of depth error measurements are as shown in figures 3.14(c) and 3.14(d). The observations of depth error measurements are as follows, however more details are given in the publication specified in chapter 10:



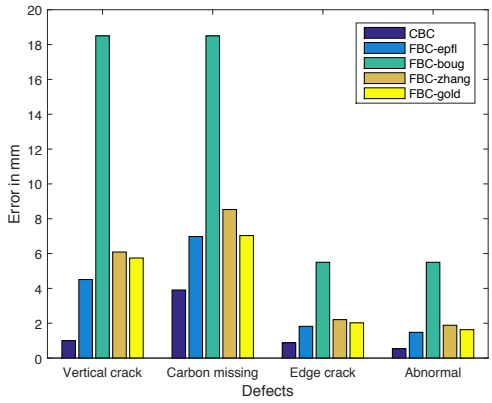
(a) Width error for scheme 1.



(b) Width error for scheme 2.



(c) Depth error for scheme 1.



(d) Depth error for scheme 2.

Figure 3.14: Mean difference of width and depth measurements for two schemes.

- All FBC types except FBC-boug performed very close to CBC with mean errors between 1-3mm in both schemes.

- For *edge crack* and *abnormal wear*, FBC types performed with mean error difference of about 1mm compared to CBC.

- For *vertical crack* and *missing carbon*, all FBC types introduced about 2mm error in *scheme 1* and 3mm error in *scheme 2*.

- Overall, FBC-epfl, FBC-zhang and FBC-gold performed the best compared to CBC, with a maximum increase in mean error of about 1.5mm for *scheme 1* and 3mm for *scheme 2*.

In both the cases, mean width and depth error for scheme is observe dto be less than scheme 2. However, sometimes it is observed that the width errors in scheme 1 are higher than in scheme 2, for example, FBC-epfl for abnormal & edge crack and FBC-zhang for vertical crack. These are due to the randomness in the occurrence of feature detection errors or pantograph misalignment in the real dataset used for evaluation.

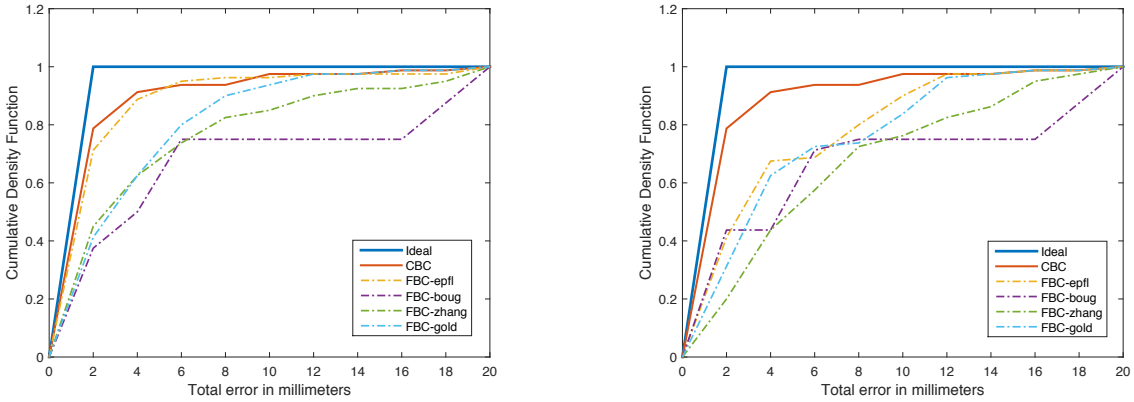


Figure 3.15: Cumulative density function (CDf) for scheme 1 and 2.

3.2.5 Error Distribution

The width and depth errors were measured, but to consider randomness in the error, the error distribution was observed using computed Cumulative Density Function (CDF) for all (width and depth) errors for both *scheme 1* and *scheme 2*. An ideal CDF was used as a baseline to compare the error distributions of FBC/CBC techniques. Figure 3.15(a) and 3.15(b) show the tendency of the divergence of FBC/CBC from the ideal baseline case. To quantify the divergence of FBC/CBC techniques from the ideal case, Kullback-Leibler Distance (KLD) (Kullback and Leibler, 1951) was used. For discrete CDFs P and Q, the Kullback-Leibler divergence of Q from P is computed as in equation 3.8. The lower value of KLD metric signifies higher accuracy as a result of small divergence from the ideal.

$$D_{KL}(P||Q) = \sum_{i} P(i) ln \frac{P(i)}{Q(i)} \tag{3.8}$$

The KLD for CBC and all FBC techniques were computed and listed in the table 3.7 for width, depth and total error for both the schemes. From the table, FBC-boug showed the best performance for width and FBC-epfl showed the best performance for depth. This confirmed the observations made in figure 3.14. According to the error distribution analysis, FBC-epfl performed better than CBC and FBC-boug performed close to FBC-epfl. FBC-zhang and FBC-gold performed alternatively better than each other in various configurations. Overall by observing the KLD values for all algorithms in both the schemes, FBC techniques performed with better accuracy in *scheme 1* configuration compared to *scheme 2*.

3.2.6 Resilience

Here, the robustness of FBC was tested against the perturbations such as feature detection error (pixel noise) and pantograph misalignment (uplift, yaw angle, roll angle, pitch angle). To emulate pixel noise, Gaussian noise with variance between +10 and -10 was added to the signal. Pantograph uplift was emulated by varying the vertical axes of 3D points from -0.5mm to +0.5mm. All pantograph rotations (yaw angle, roll angle, pitch angle) were ranging between -10

Measurment	Ideal	CBC	FBC-epfl	FBC-boug	FBC-zhang	FBC-gold
Scheme-1						
width	0	1.29	1.39	**0.92**	1.39	1.20
depth	0	0.51	**0.39**	0.69	0.52	0.80
total	0	0.24	**0.34**	0.98	0.80	0.88
Scheme-2						
width	0	1.29	1.29	**0.80**	1.61	1.39
depth	0	0.52	0.92	**0.70**	1.39	1.12
total	0	0.24	0.88	**0.83**	1.61	1.16

Table 3.7: Kullback-Leibler Divergence values for total (width + depth) error.

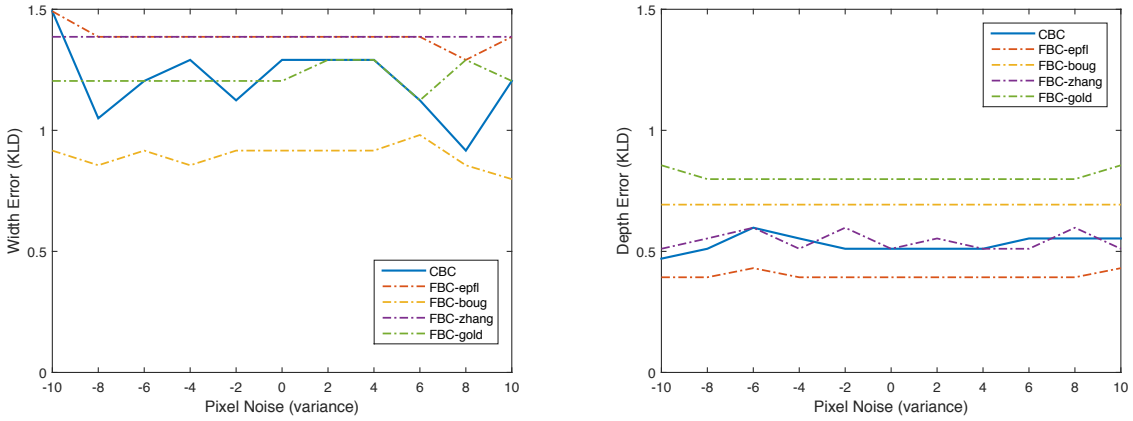


Figure 3.16: Resilience over pixel noise.

to +10 degrees. For all these perturbations, the FBC was carried out for *scheme 1* configuration and KLD was computed.

Figures 3.16-3.20 shows the results of both width (all subfigures in the first column) and depth (all subfigures in the second column) measurements measured in KLD metric. This showed the divergence of each FBC technique and CBC from the baseline error distribution. The lower value of KLD signified better accuracy of the techniques.

The observations can be summarized as follows, however, more details are given in the publication specified in chapter 10:

- For pantograph misalignment errors, i.e., uplift (figure 3.17), and rotation errors (figures 3.18-3.20), several FBC types were more robust than CBC. This was because the reference world axis was fixed in space for CBC and any misalignment in pantograph would affect the measurement from reference, whereas for FBC the reference axes was located on the pantograph itself.

- For feature detection errors (figures 3.16), FBC types were obviously more sensitive than CBC, because FBC relied on noise-free feature points for calibration. However, FBC-boug (for width) and FBC-epfl (for depth) showed better resilience compared to CBC in

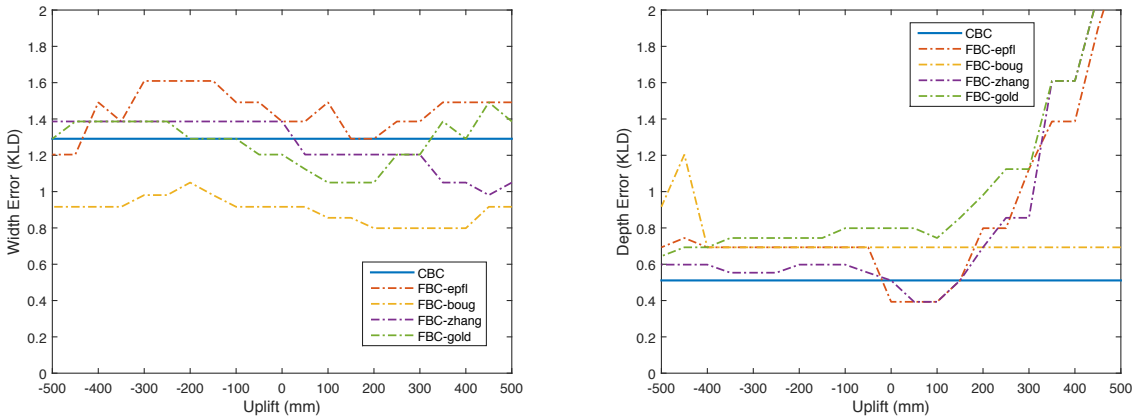


Figure 3.17: Resilience over pantograph vertical displacement (uplift).

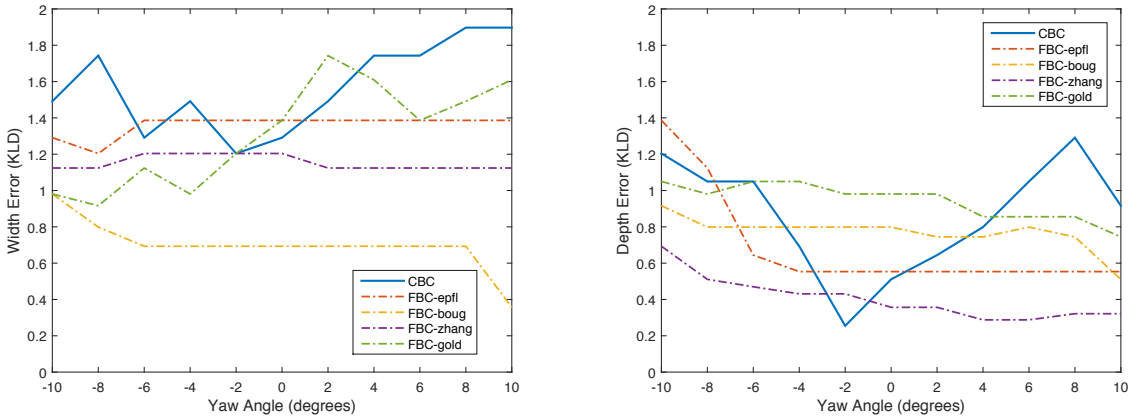


Figure 3.18: Resilience over pantograph angular displacement (yaw).

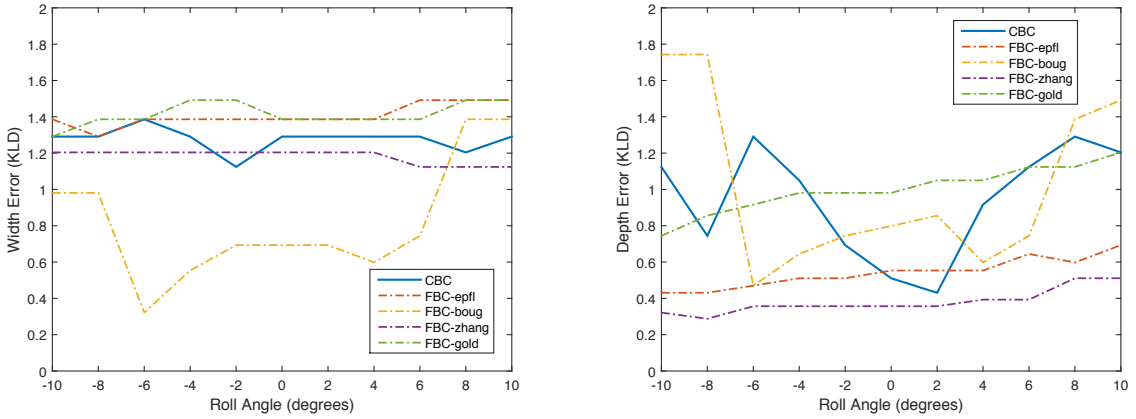


Figure 3.19: Resilience over pantograph angular displacement (roll).

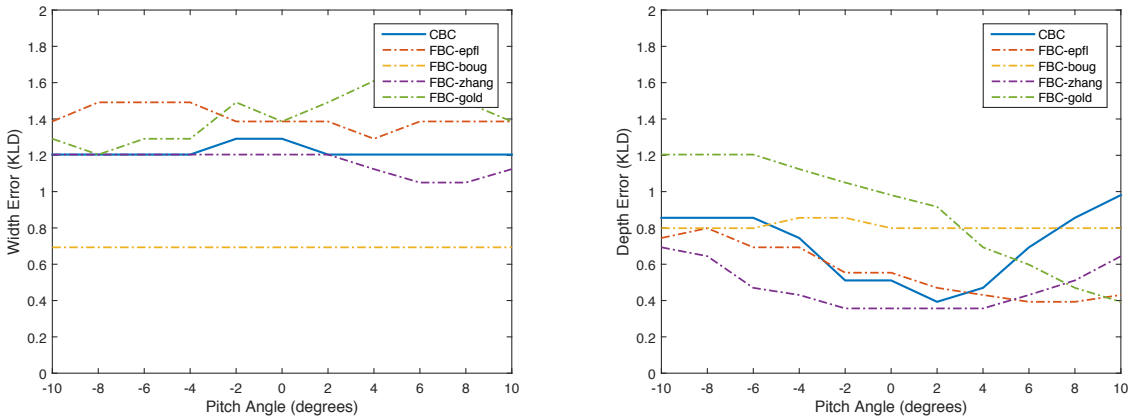


Figure 3.20: Resilience over pantograph angular displacement (pitch).

handling noisy feature points. This is because the optimization routine of FBC yields better results with the localization of world coordinate system on the pantograph. Therefore, FBC is preferred over CBC when the profile images are noisy, however in this case, detection of six control points for FBC can also be affected, but not to a large extent because a priori knowledge of the physical model of the pantograph is available.

- In all width errors (figures 3.16(a), 3.17(a), 3.18(a), 3.19(a), 3.20(a)), FBC-boug consistently showed best resilience compared to CBC. FBC-boug uses only 6 points to calibrate compared to CBC using more than 200 points for bundle optimization.

- In all depth errors (figures 3.16(b), 3.17(b), 3.18(b), 3.19(b), 3.20(b)), especially for rotational perturbations, CBC was very sensitive, where as FBC types, i.e., FBC-zhang and FBC-epfl showed a flatter response.

3.2.7 Discussions

Four of the popular state-of-art pose estimation algorithms, i.e., FBC-boug, FBC-epfl, FBC-zhang and FBC-gold, were evaluated in order to show that FBC can replace CBC in 3D systems. Keeping the PTMS as the main application focus, CBC was replaced with four FBC types in two configurations (*scheme 1 & 2*). It was seen that *scheme 1*, which used every profile image to carry out FBC and use the calibration parameters to estimate the 3D measurements provided more accurate results. In *scheme 2*, after FBC is carried out using an image of laser lines defining the pantograph, there are high chances that the pantograph is misaligned and the new profile image is used with outdated FBC parameters for 3D measurements. This problem is overcome in *scheme 1*.

Overall, it was found that FBC techniques outperformed CBC technique in both width and depth measurements in various configuration. It was also found that FBC techniques were more robust to pixel error or pantograph misalignment. By adopting FBC in the PTMS instead of CBC, it was evident that FBC technique used only 6 points to carry out calibration; it could handle noisy feature points; provide higher accuracy and robustness.

All the above evaluation tests were sufficient to reject the *Null Hypothesis II* (stated in section 1.4). Hence, it can be concluded that online re-calibration for error-sensitive 3D measurement systems (PTMS) is possible using FBC methods that give effectively better performance and robustness than CBC. This tremendously increases the usability of 3D vision inspections systems with greater flexibility of using online re-calibration without any manual intervention.

3.3 Misalignment in Stereo Camera System

In the previous section 3.1, the effects of single camera misalignment with application to vision based inspection systems, i.e., Pantoinspect (details in section 2.2.3) was discussed. The essence of this study cannot be used directly on other 3D applications scenarios such as VERDIONE (details 2.2.1), BAGADUS (details 2.2.2), POPART (details 2.2.4), where stereo cameras are used for image acquisition.

In 3D stereo capture systems, the cameras are susceptible to misalignment especially, due to manual intervention after deployment. This can happen when the cameras are not stable and fixed onto the rigs. If the system builders knew, which of the camera motion (represented by position and orientation in x,y,z, direction) affects the 3D reconstruction quality, then they can take care either by manufacturing stable rigs or by restricting the camera to move in a certain direction while using the camera. Hence, studying the misalignment of stereo camera is important as much as for single camera.

Previously, there were many investigations on the misalignment of stereo camera and how to correct them (Santoro, AlRegib, and Altunbasak, 2012). The influence of camera misalignment error on stereoscopic reconstruction was studied (Bolecek and Ricny, 2015), but the evaluation was limited to accuracy of feature correspondence. Another article (Zhao and Nandhakumar, 1996) evaluated individual spatial misalignment of camera but was limited to only camera rotations. However, in this study the focus was on measuring the performance due to camera misalignment in 3D space, which is more relevant for 3D applications and all camera misalignment, i.e., spatial, positional and the combinations, in all directions, i.e., both positive and negative x,y,z axis were considered.

The error for every component (x,y,z directions) of 3D reconstruction was measured. Error occurring in 3D components can represent deformation of 3D reconstruction. This helps us in understanding the nature of deformation of 3D reconstruction due to the camera misalignment and this knowledge helps system designers build stable systems.

3.3.1 Evaluation

The camera misalignment can be either a pure translation, pure rotation or in most practical cases, a combination of both translation and rotation. This is very important because the practical reason for misalignment is manual intervention that has no control over the camera misalignment. Camera misalignment has 6 degrees of freedom for both translation and rotation. The effects of all combinations of the camera misalignment on the 3D reconstruction accuracy were studied. Also how the effects of camera misalignment influenced the variation of object size was studied. The evaluation study is illustrated as in figure 3.21, and was simulation based.

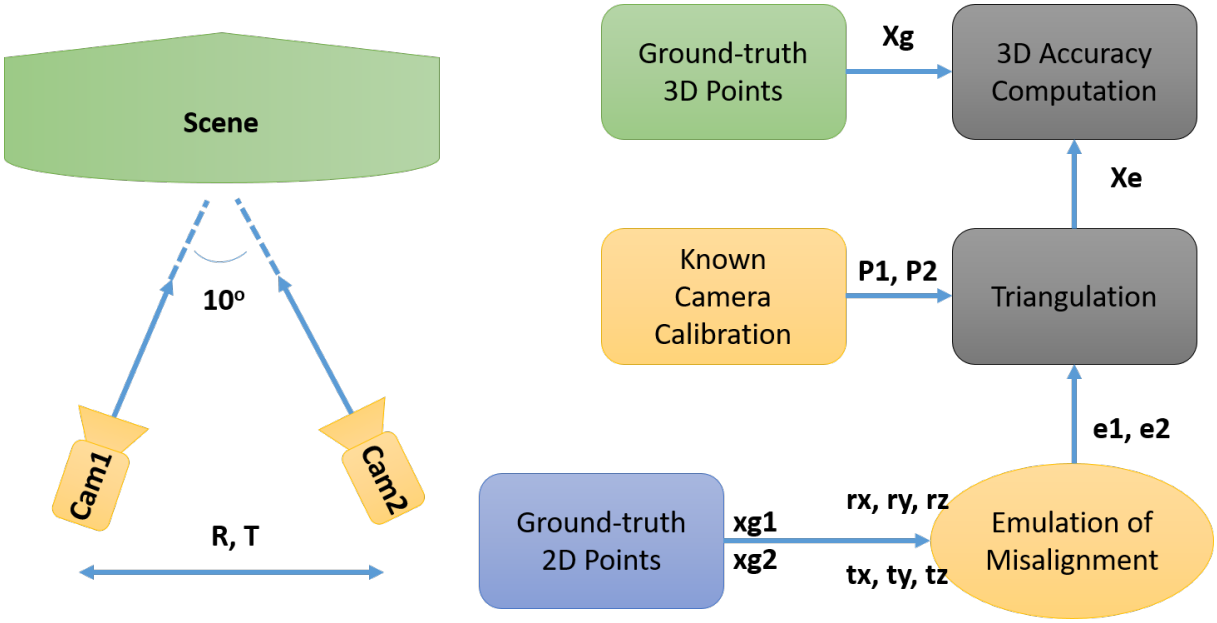


Figure 3.21: Evaluation procedure for stereo camera misalignment.

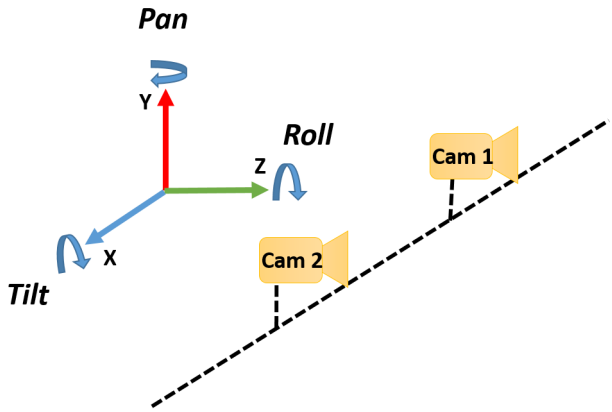


Figure 3.22: Camera axes and rotations around X-Tilt, Y-Pan, Z-Roll.

Procedure

The second camera on the right is used for emulating camera misalignment in terms of change in rotation and translation with respect to the left camera. The rotational misalignment are represented as r_x, r_y, r_z and translation misalignment as t_x, t_y, t_z. Camera axes and rotations (pan, tilt, roll) are as shown in the figure 3.22. The ground-truth 2D points (x_g1, x_g2) were filtered using the misalignment emulator, to obtain the misaligned image coordinates $(e1, e2)$.

The misaligned points were triangulated using the camera projection matrices $(P1, P2)$, which were computed as in equation 3.9. The camera intrinsic (K) was assumed to be the same for both the cameras. The left camera center is treated as the reference and hence, the rotation (R) and translation (T) relates to right camera.

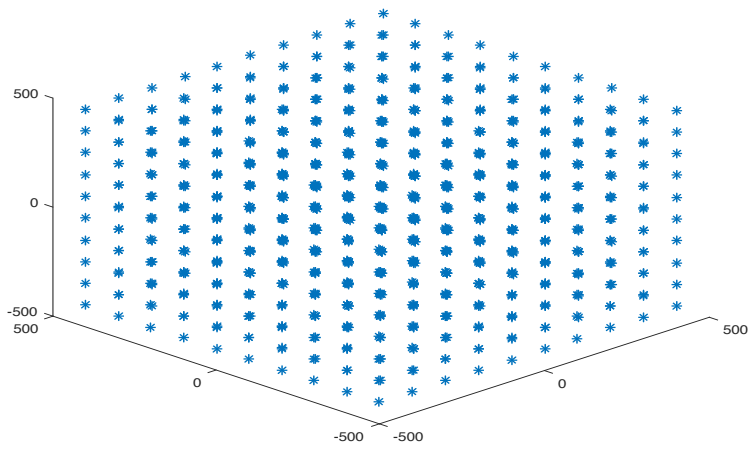
$$P1 = K * [I, 0], P2 = K * [R, T] \tag{3.9}$$

where,

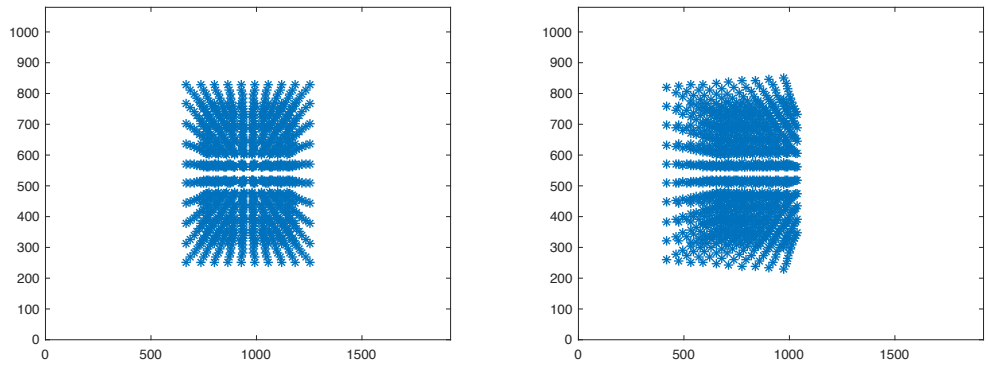
$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, K = \begin{bmatrix} f & 0 & px \\ 0 & f & py \\ 0 & 0 & 1 \end{bmatrix}$$

$$3Derror = ||X_g - X_e||_2 \tag{3.10}$$

As a result of triangulation 3D points were estimated. The performance metric used in this case is the 3D accuracy, which is computed as mean squared difference between the estimated 3D points (X_e) and the ground-truth 3D points (X_g) (equation 3.10). Also x, y, z components of 3D error were measured, to find out in which direction the deformation of 3D reconstruction was affected more, comparatively.



(a) Object point cloud for the experiment.



(b) Image view from stereo cameras - left camera facing the cube and right camera is rotated and translated with horizontal angular displacement of 10 degrees.

Figure 3.23: Object and their projected stereo images.

Dataset

The focus was on large space scenario such as VERDIONE (details in section 2.2.1). So, the test database was generated to mimic the VERDIONE stage performance with object in the scene at a distance of about 2 meters. The object is assumed to be uniformly spaced point cloud representing the vertices of a volumetric cube (figure 3.23(a)). A stereo camera setup was assumed to capture the object in the scene, with a focal length of 55mm at full resolution. Two cameras, without lens distortion, are positioned at an angular deviation of 10°. The perspective image from two cameras are shown in figure 3.23(b). Hence, ground-truth values for 3D and 2D points were obtained, for known camera calibration parameters, i.e., K-intrinsic matrix, R-relative rotation and T-relative translation of right camera with respect to the left one.

3.3.2 Pure Translation Misalignment

The results of effects on 3D accuracy for camera misalignment in terms of pure translation (all combinations of their components) is shown in figure 3.24. The camera translation was varied from -10mm to 10mm (sign represents the direction). The 3D error was measured in mm as well.

As the translation misalignment increases, the error is observed to increase linearly. In figure 3.24(a), 3.24(d), 3.24(f) and 3.24(g), the 3D-Z increased rapidly compared to the other component, for variation in t_x. The 3D - X, 3D-Y components increased upto around 5mm when the camera was translated to 10mm. On the other hand, in figures 3.24(b), 3.24(c) and 3.24(e), all 3D error components shown accuracy around 5mm. Therefore, a higher error rate occurred in 3D-Z component especially, when the camera motion involves translates in X direction. So, this means that when the right camera in a stereo setup is misaligned on its horizontal axis, the reconstruction is deformed along the z axis, i.e., towards or away from the camera.

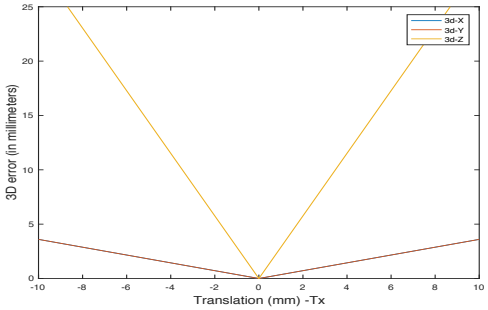
Overall, the 3D error is most sensitive to translation along the horizontal axis that is collinear to the stereo camera centers.

3.3.3 Pure Rotation Misalignment

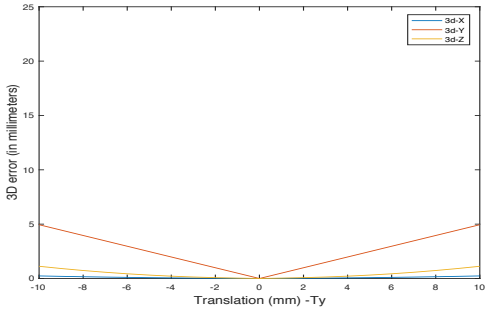
The results of effects on 3D accuracy for camera misalignment in terms of pure rotation (all combinations of their components) is shown in figure 3.25. The camera rotation was varied from -10° to $+10^{\circ}$ (sign represents the direction). The 3D error was measured in mm as well.

As the camera is misaligned by either pan, tilt or roll direction, the 3D error increase as a curvilinear function. However in most cases, 3D-Z component error rate is the highest, and then is the 3D-Y component that is affected due to camera rotations in all directions. Compared to camera tilt (figure 3.25(a)), pan (figure 3.25(b)) and roll (figure 3.25(c)), the camera roll type of misalignment affects the 3D accuracy more.

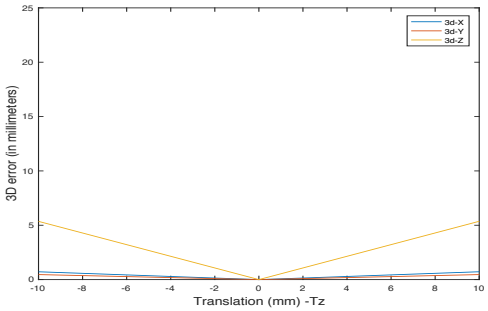
The camera pan affects the error in an unsymmetrical fashion (figures 3.25(b), 3.25(d), 3.25(d) and 3.25(g)). When the camera is panned in the positive direction, i.e., rotated towards the left camera (figure 3.22), then triangulation yields 3D points closer to its true position. When the camera is panned in the negative direction, i.e., rotated away from the left camera, then the triangulation yields 3D points that are much further away from their true position.



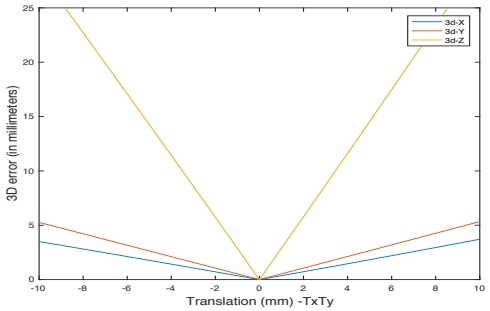
(a) Translation - X axis t_x



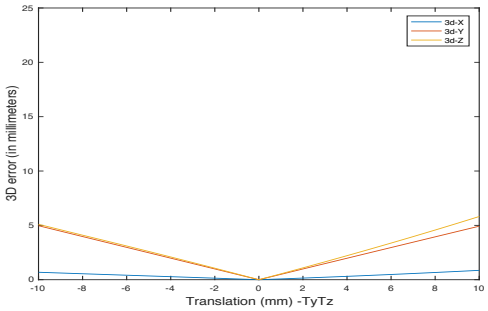
(b) Translation - Y axis Vs t_y



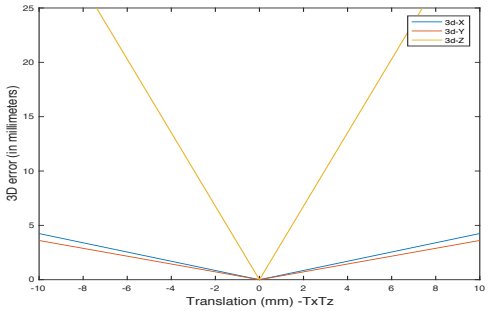
(c) Translation - Z axis t_z



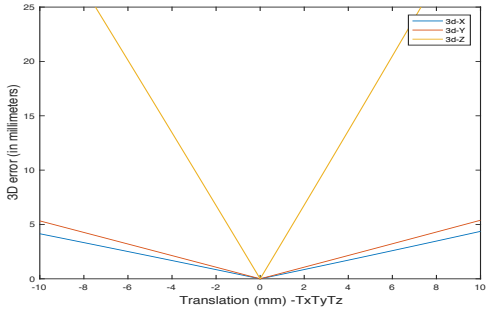
(d) Translation - X,Y axis t_x, t_y



(e) Translation - Y,Z axis t_y, t_z



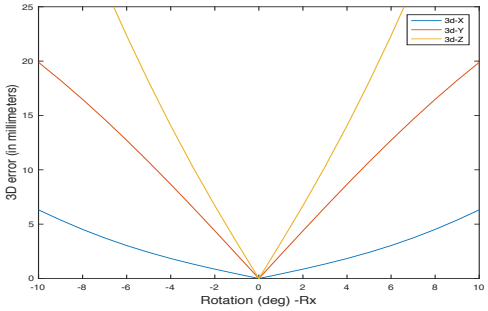
(f) Translation - X,Z axis t_x, t_z



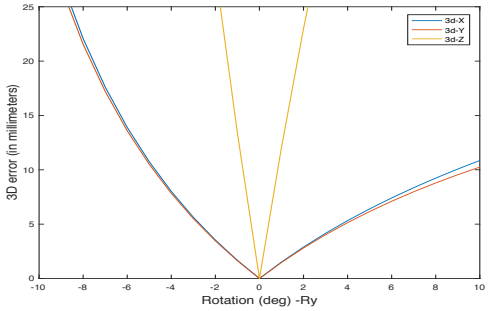
(g) Translation - X,Y,Z axis t_x, t_y, t_z

Figure 3.24: Variation of 3D error versus camera misalignment in terms of pure translations

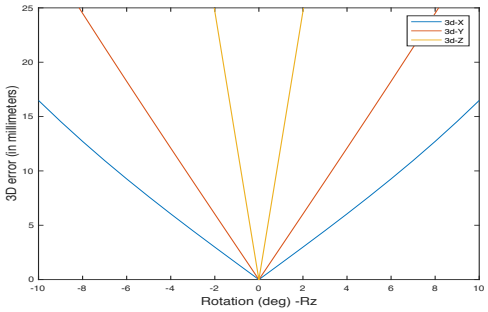
Overall, 3D accuracy is more sensitive to rotation more than translation misalignment. The camera roll affects the accuracy the most, and camera pan affects the symmetric error pattern.



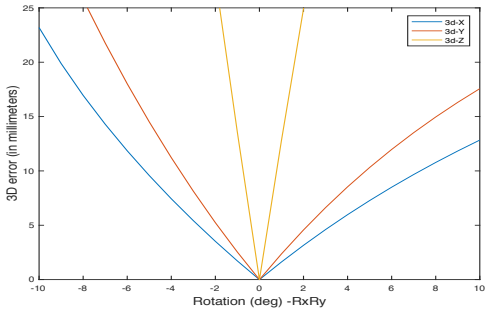
(a) Rotation - X axis r_x



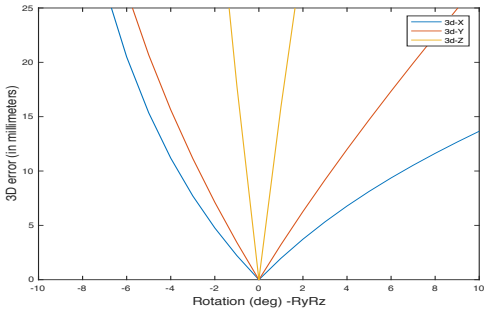
(b) Rotation - Y axis r_y



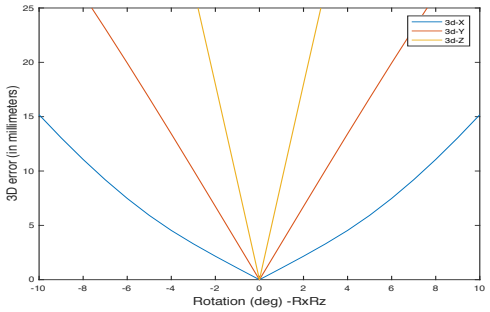
(c) Rotation - Z axis r_z



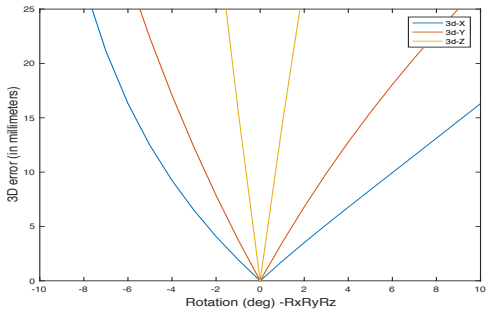
(d) Rotation - X,Y axis r_x, r_y



(e) Rotation - Y,Z axis r_y, r_z



(f) Rotation - X,Z axis r_x, r_z



(g) Rotation - X,Y,Z axis r_x, r_y, r_z

Figure 3.25: Variation of 3D error versus camera misalignment in terms of pure rotations

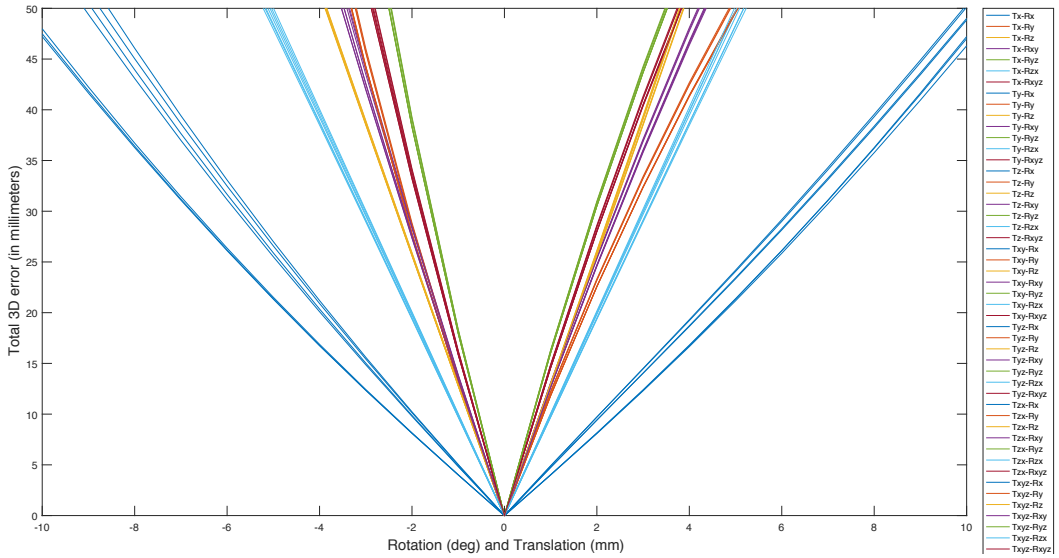


Figure 3.26: Variation of Total 3D error versus camera misalignment in terms of translations and rotations

3.3.4 Combined Misalignment

Combined misalignment refers to the combination of both translations and rotations. The results of effects on 3D accuracy for complex camera misalignment is shown in figure 3.26. The camera translation was varied from -10mm to 10mm and the camera rotation was varied from $-10°$ to $+10°$ (sign represents the direction). The 3D error was measured in mm as well.

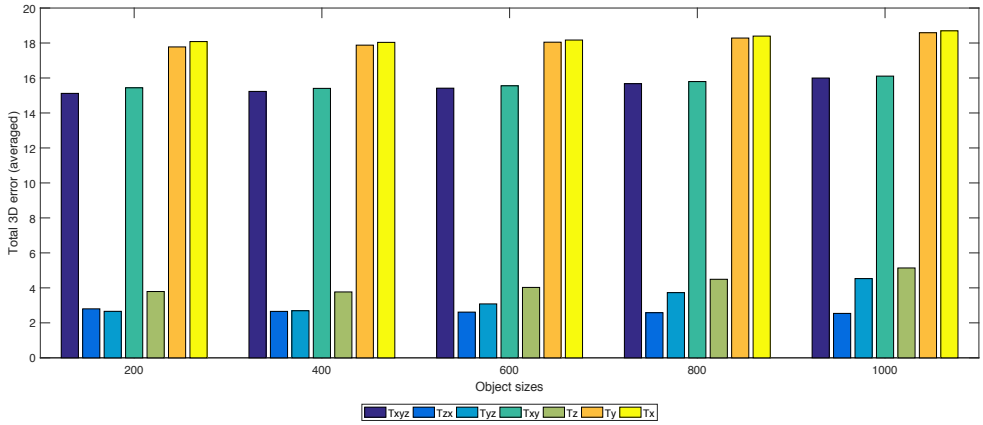
The results for a combined effect of camera translation and rotation is as shown in figure 3.26. There is an overall increase in the rate of 3D error with increase in the camera misalignment. This type of combined misalignment is more likely to practical scenario, where manual intervention has no control over misalignment direction. So, in practical scenarios the effect of camera misalignment on 3D errors seemed severe. For example, if any application requires an accuracy of about 5mm in 3D reconstruction, the tolerance of camera rotation and translation should be less than 1 degree and 1 mm, respectively.

3.3.5 Variable Object Size

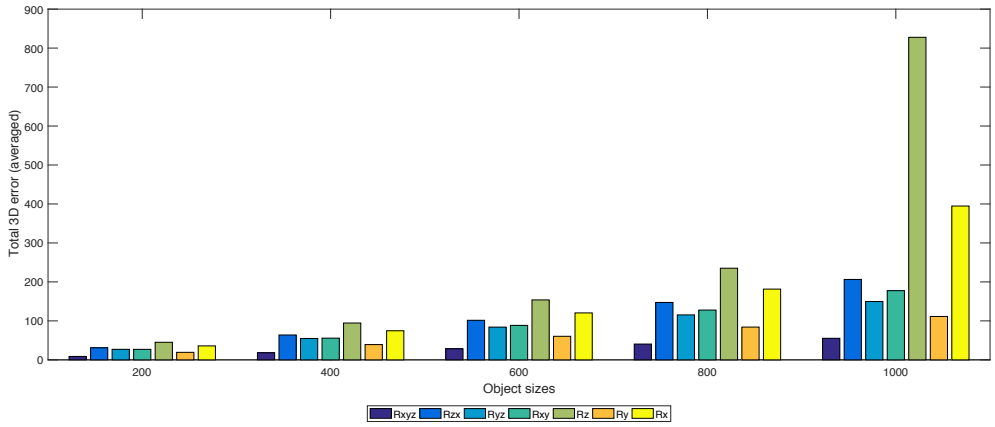
Here, the influence of the size of the object on the effect of camera misalignment was studied. The object sizes were varied between 200 and 1000 vertices of a cube, which is the test data.

The results are shown in figure 3.27, where average 3D error for various object sizes, over different translations, rotations and combination of rotations $\&$ translations are shown in figures 3.27(a), 3.27(b) and 3.27(c), respectively. In every sub figure, a similar pattern for each individual object sizes shows that the results on translation and rotations were consistent over various object sizes.

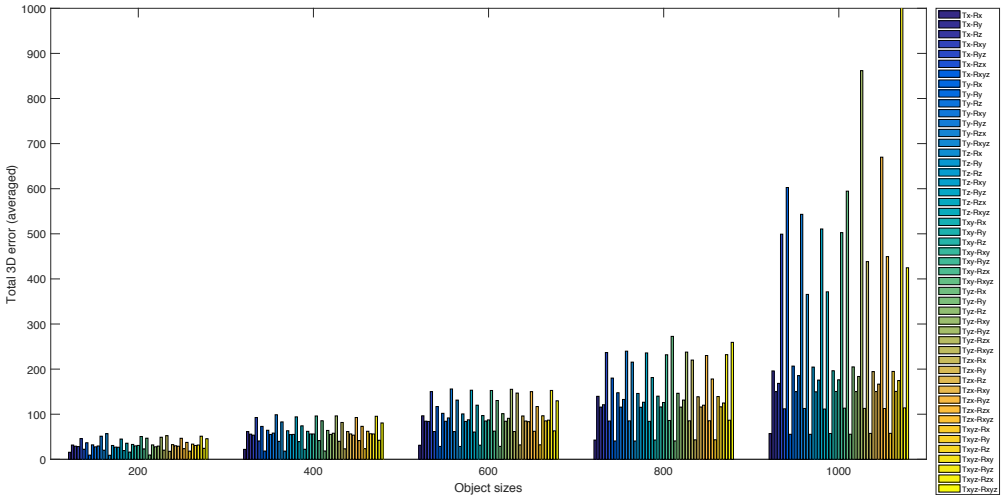
Even though the effect of translation has been the same for all object sizes, the effect of camera rotation increases with increase in object size. This shows a direct relation between object size and error. Both increase by the same factor, because both error and object points are represented in 3D space. This knowledge can be used to balance the error due to misalignment



(a) Total 3D error (pure translations) for various object sizes.



(b) Total 3D error (pure rotations) for various object sizes.



(c) Total 3D error (translations and rotations) for various object sizes.

Figure 3.27: Variation of Total 3D error averaged over range of misalignment versus object sizes.

from camera rotation. If the camera is susceptible to unavoidable misalignment in rotation, then the camera can be setup to capture the objects in lesser resolution (equivalent to reducing the object size) and achieve a reduction in 3D reconstruction error.

3.3.6 Discussions

Overall, the experiments were conducted to study the effects of camera misalignment in terms of translation, rotation and their combination. The following is the summary of the results obtained:

- 3D error is sensitive to misalignment in terms of translation along the line that is collinear to the camera centers.
- 3D error is sensitive to camera roll misalignment more than to the tilt or pan.
- In both of the above 3D error measurement, z component which represents the depth error is more that any other component. This signified a low tolerance to depth related applications compared to metrology measurement where only x or y component could be of interest.

This study can be used by the system builders to understand the sensitivity of camera misalignment and build the rigs accordingly, i.e., to restrict camera roll and camera translation along the line of camera centers. However, the extent to which restriction are held, i.e., tolerance is determined by application scenario.

The tolerances on each individual aspect of camera misalignment, i.e., translation or rotation in x,y,z directions can be determined, by modeling the error and using a predictor function, similar to section 3.1.3. However, to predict the tolerances, an acceptable error that depends on the application scenario needs to be known. As the evaluation was carried out only on a virtual simulated dataset, assuming an acceptable error for an application scenario is not fair.

Therefore, the study in this section was limited only to show the effects of camera misalignment on a stereo camera system and leave the significance of the effect to be application specific. This study can further be used as a methodology to model the error and predict the tolerances based on the application's acceptable accuracy.

All the above tests and discussions regarding the significance and tolerances of effect of camera misalignment are evident enough to reject a part of *Null Hypothesis I* (stated in section 1.4). From this, it can be concluded that the 3D reconstruction accuracy significantly decreases when the camera is misaligned in stereo camera setup, where the significance is determined by the application scenario.

The experiments were carried out on a simulation dataset, which tried to mimic the application scenario such as VERDIONE and BAGADUS, in order to obtain ground-truth values for determining the 3D accuracy. However, the application scenario testing was limited to only a foreground model object. The situation where the captured scene involves background with a large depth of field, is more likely to occur in the application scenario such as VERDIONE, BAGADUS and POPART. Such a situation was not tested in this study.

3.4 Conclusions for FBC

In this chapter, the research *Hypotheses I* and *II*, as stated in section 1.4, were tested. In the process of testing the research hypotheses, for each of the topics discussed, this chapter stated the proposed ideas, explained the experiments setup and discussed the results.

Accordingly, the effects of camera misalignment on single camera systems (for the PTMS scenario) and stereo camera systems (for the VERDIONE, BAGADUS and POPART scenarios) were discussed, and hence, it was shown that the 3D reconstruction error significantly increases when the camera is misaligned. Adoption of feature based calibration over traditional checkerboard calibration (for the PTMS scenario) was also discussed, and hence, it was shown that the accuracy and robustness of 3D reconstruction significantly improves when the 3D system replaces CBC by FBC techniques.

Overall, in a practical perspective, the contributions from the study in this chapter, is summarized as follows:

1. A statistical tool or methodology that is easily implementable and reproducible was developed. This tool can be used by any single camera system to determine the mechanical tolerances of camera rigs to minimize errors caused by camera misalignment. This helps in deployment of robust 3D systems, especially PTMS.

2. A feature based calibration methodology was adopted for 3D measurements system (PTMS) by replacing the traditional checkerboard calibration. This provides an extended flexibility in using the deployed 3D system without manual intervention, when practical problems occur.

3. It was shown how the 3D system gets affected by stereo camera misalignment. This helps the system designers to build stable camera rigs to improve the accuracy of the 3D system by restricted erroneous camera misalignment. The evaluation was carried out using simulation dataset that mimicked VERDIONE or BAGADUS scenarios, however, application scenario testing was limited.

In this chapter, the robustness of FBC as a complete system was explored. FBC technique relies on the features extracted from the images. Hence, the quality of FBC and thereby 3D reconstruction, depends on the quality of feature extraction. Therefore, the exploration of the robustness of feature extraction is described in the next chapter.

Chapter 4

Feature Extraction

The need for feature based calibration (FBC) and their adoption in 3D system was discussed in the previous chapter in the context of achieving high quality calibration of cameras in terms of accuracy and robustness. Although the performance of 3D systems was studied as a complete system, there are still few important questions to investigate pertaining to the building blocks of the feature calibration subsystem. As explained in figure 2.1, the main building blocks of FBC for any 3D system are feature extraction and pose estimation.

Feature extraction, in the context of 3D applications, is the first step in feature based calibration. As explained in section 2.1, in 3D stereo systems the feature extraction process first detects feature points in both the images and then matches them to obtain feature correspondences in stereo image pair. The unknown 3D points are typically estimated as sparse 3D points using triangulation process or dense 3D points using rectification followed by depth estimation process. In all cases of 3D reconstruction, the calibration parameters estimated are an important factor that determines the accuracy of reconstruction (Pedersini, Sarti, and Tubaro, 1998). Since the calibration parameters are estimated using the feature correspondences in FBC, the quality of feature extraction also plays an important role in determining the 3D reconstruction quality. Therefore, the important question is, what are good features for high quality 3D reconstruction systems.

For 3D reconstruction application in application scenarios such as VERDIONE (details in section 2.2.1), BAGADUS (details in section 2.2.2) and POPART (details in section 2.2.4), multiple stereo pair of cameras were considered to capture the scene. In these scenarios, the cameras are prone to change in their properties. The internal properties of the camera constitute focal length, resolution, lens distortion, noise etc. The external properties of the camera refers to relative position and orientation between the neighboring cameras. Any change in the camera properties, will have repercussions on the accuracy of 3D reconstruction. With all these perturbations, a good feature extraction that can be used for FBC to achieve high quality 3D reconstruction is not guaranteed. Hence, it is very important to evaluate the robustness of state-of-the-art feature extractors when used in 3D systems for high quality 3D reconstruction. Accordingly, *Hypothesis III* is stated in section 1.4, to investigate the impact of change in both internal and external camera properties on the quality of feature extractors in stereo systems.

One of the popular point feature extractors today is SIFT - Scale Invariant Feature Transform (Lowe, 2004). SIFT is known for scale and rotational invariance between a stereo pair. However, it is observed that SIFT has a limitation in terms of maintaining a good accuracy for

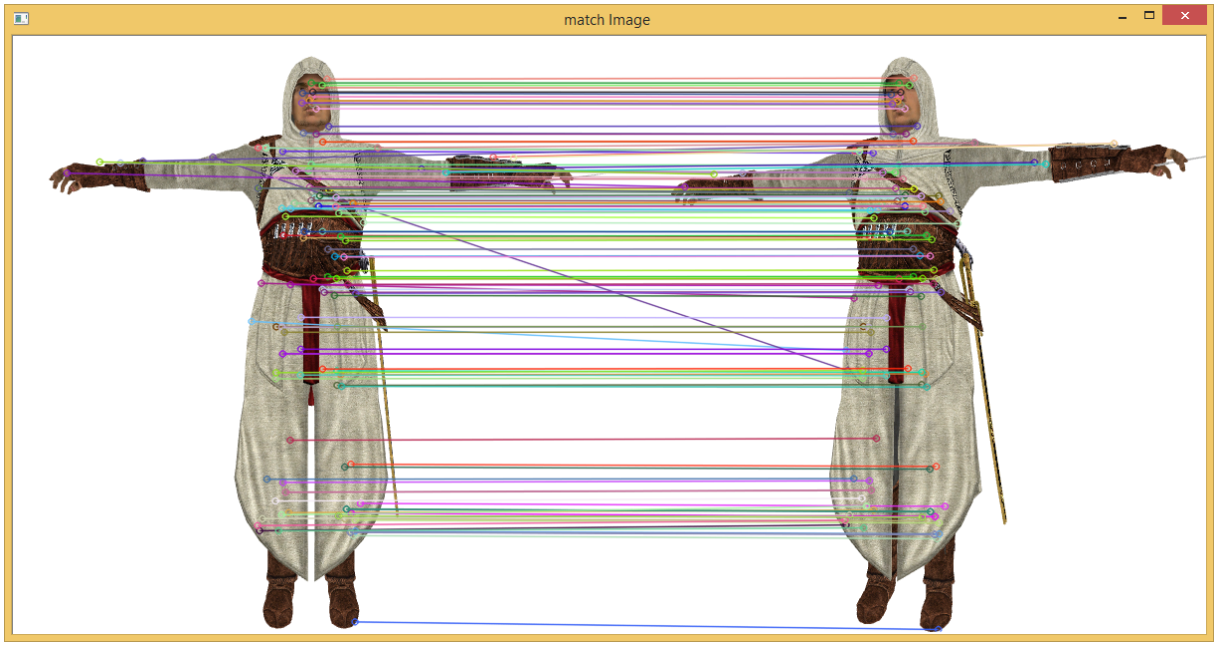


Figure 4.1: Illustraing feature extraction process - detection, description and matching, between stereo pairs.

an angular separation greater than 30 degrees between the stereo camera pair. For large space scenarios, a wide baseline camera setup is very suitable in order to use a smaller number of cameras to cover the large area of the scene, and maintain the same quality of 3D reconstruction. Therefore, to explore the possibility of using SIFT for wide baseline FBC, *Hypothesis IV* is stated in section 1.4.

In this chapter, the aim is to test *Hypothesis III* and *Hypothesis IV*. In order to test these hypotheses, the robustness of several state-of-art feature extractors over variation in camera intrinsic and camera extrinsic were evaluated. Also, SIFT feature extractor was explored for its usage in FBC for wide baseline camera setup.

4.1 State-of-the-art Feature Extractors

Feature extractors, in this context, refers to the combination of feature detector and descriptor. Feature detectors identify the interesting keypoints in an image and feature descriptors tag each of these points with a unique representation. Point features, obtained by feature extractors are then matched between the stereo pair. Figure 4.1 shows the point features extracted in a stereo camera pair, where an outlier is also visible. Normally, the outliers are removed using optimization algorithms (e.g., RANSAC (Fischler and Bolles, 1981)) to obtain feature correspondences between stereo images.

In this chapter, several of the state-of-the-art feature detectors and descriptors were selected as candidates for the evaluation. The selection was based on their extensive use in different areas of application. These state-of-the-art feature extractors are briefly explained with their properties in the table 4.1.

Feature Extractor	**Properties**	**Detection**	**Description**
SIFT (Lowe, 2004)	Scale and rotation invariant. Robust to change in illumination, 3D viewpoint, noise.	Interesting points are identified using Difference of Gaussian (DoG) over several linear scales of images. Then, the location and scale of key-points are accurately computed using neighbor pixels.	The descriptor is represented by histograms of image gradients that are computed at every image point around the key-points detected.
SURF (Bay, Ess, Tuytelaars, and Van Gool, 2008)	Scale and rotation invariant. Speeded-Up Robust features which are distinctive, robust to noise, geometric & photometric deformations and specially can be computed faster.	Using integral images makes the image convolution faster. The detector is based on Hessian matrix-based approximation of blob like interesting points using Gaussian scale space.	Descriptor is based on distribution of interesting points in its neighborhood. This is similar to SIFT but instead of using gradients, distribution of first order Haar Wavelets responses is considered.
ORB (Rublee, Rabaud, Konolige, and Bradski, 2011)	Designed to perform two magnitude faster than SIFT.	This is a FAST detector with addition of an accurate orientation component using intensity centroid.	Rotation-Aware binary descriptor based on BRIEF descriptor. This is computed by introducing a learning method for decorrelating the BRIEF features under rotational invariance.
BRISK (Leutenegger, Chli, and Siegwart, 2011)	Adaptive and high quality feature detector and descriptor designed to lower computational complexity compared to SURF.	It is a combination of FAST detector in scale space and identifying key-points by fitting a quadratic function.	The descriptor is a bit-string assembly from intensity comparisons retrieved by dedicated sampling of each key-point neighborhood.

KAZE (Alcantarilla, Bartoli, and Davison, 2012)	Scale and rotation invariant. Attains high accuracy in object boundaries and robust to noise.	Similar to SIFT, except that the key-points are detected in nonlinear scale space using Additive Operator Splitting techniques and variable conductance diffusion.	This uses a modified SURF descriptor, which is SURF descriptor with an intelligent two-stage Gaussian weighting scheme.
AKAZE (Pablo Alcantarilla and Bartoli, 2013)	Accelerated KAZE - motivated to compute faster with similar scale & rotational invariance and low storage requirement properties.	Instead of using non-linear scale space as in KAZE, a numerical scheme called Fast Explicit Diffusion in a pyramid framework.	A Modified-Local Difference Binary descriptor, which exploits gradient and intensity information from nonlinear scale space.
MSER (Matas, Chum, Urban, and Pajdla, 2002)	Affine-invariant feature extractor suitable for wide baselines in stereo. Robust to change in scale, illumination, out-of-plane rotation, occlusion and viewpoints.	Distinguished regions are detected and affine invariant procedure is carried out to estimate the stable invariant regions, from which the key-points are measured.	n/a
STAR (Agrawal, Konolige, and Blas, 2008)	A suite of scale invariant center-surround detectors focused on visual odometry applications. Stable and repeatable in viewpoint changes. (CenSurE)	The CenSurE features are computed at the extrema over multiple scales using full image resolution using center-surround filters. There is na approximation to scale space based of Laplacian of Gaussian.	n/a
FAST (Rosten and Drummond, 2006)	High Speed corner detector extensively used in machine learning methods, suitable for real-time applications.	Considers a circle comprising of 16 pixels in an image. Then every pixel is compared with only 4 neighbors to classify if its a corner or not.	n/a

BRIEF (Calonder, Lepetit, Strecha, and Fua, 2010)	A highly discriminative binary descriptor designed to compute faster. Invariant to large in-plane rotation.	n/a	Binary sting descriptor relies on the image patches-pairwise intensity comparisons. A classifier is trained with image patches form various viewpoints.
FREAK (Alahi, Ortiz, and Vandergheynst, 2012)	Inspired by the human visual system - retina, this descriptor is a cascade of binary strings aimed at faster computation.	n/a	Computed by efficiently comparing image intensities over a retinal sampling pattern containing Gaussian kernel information.

Table 4.1: Overview of the state-of-the-art feature extractors.

4.2 Robustness against Camera Intrinsic

The intrinsic and extrinsic camera parameters are vital information for 3D multimedia applications which rely on data acquired by stereo camera pairs, such as free-view rendering (Min, Kim, Yun, and Sohn, 2009), motion tracking (Moeslund and Granum, 2001), structure from motion (Agarwal, Furukawa, Snavely, Simon, Curless, Seitz, and Szeliski, 2011) or 3D scene reconstruction (Matusik, Buehler, Raskar, Gortler, and McMillan, 2000). In large space scenarios such as Mixed reality art performance, e.g., VERDIONE (details in section 2.2.1) or BAGADUS (details in section 2.2.2), it is very difficult or sometimes impossible to obtain camera calibration parameters using traditional Checkerboard Based Calibration (CBC) techniques (Bouguet, 2008; Tsai, 1992; Zhang, 2000). Sometimes, it is inconvenient to use markers in the scene, and hence, Marker Based Calibration (MBC) techniques (Kurillo, Li, and Bajcsy, 2008) are ruled out. So, such scenarios have to rely on Feature Based Calibration (FBC) techniques (Li and Lu, 2010; Liu, Zhang, Liu, Xia, and Hu, 2009). In FBC, camera calibration parameters are estimated using feature correspondences in a stereo/multiple camera images.

In practical 3D imaging systems, image sensors suffer from practical perturbations such as defocus, image blur, lens distortion, thermal noise, offsets in exposure time and whit balance. Camera defocus is a known problem which causes blurring of the image. Image blur also occurs when either the camera or the object moves faster than the camera's shutter speed. If the camera is set for high ISO during low light conditions, image noise is inevitable. Thermal noise might exist in camera sensors which depends on the operating temperature. The use of variety of lenses introduces lens distortion. Resolution of the captured image or video might be another variable based on the needs of the application. All these perturbations in the camera degrades the performance of feature extraction, and thereby, eventually impacts the accuracy of the geometric 3D reconstruction. The feature extractors such as SIFT (Lowe, 2004), SURF (Bay, Ess, Tuytelaars, and Van Gool, 2008) and ORB (Rublee, Rabaud, Konolige, and Bradski,

2011) are widely used for FBC due to their performance traits or ease of availability. However, the robustness of these feature extractors against the perturbations of internal camera parameters is still not known to a large extent.

It was shown that the accuracy of the camera calibration is sensitive to the quality of feature points in terms of data quantity, noise and distortions (Sun and Cooperstock, 2006). This work focused on evaluating the CBC techniques, especially from Tsai (Tsai, 1992), Hekkilä (Heikkilä, 2000) and Zhang (Zhang, 2000) using checkerboard corner points. Although it gave an idea of how the perturbations affect the accuracy, the robustness of feature extractors was not explored. In another article (Moreels and Perona, 2007), feature extractors were evaluated over the variation in the viewpoint, light and scale change, but not over the variation of camera intrinsic. SIFT, variants of SIFT, SURF and FAST features were evaluated for change in scale, illumination and blur (El-gayar, Soliman, and meky, 2013). Here, evaluation is based on a number of feature matches, which does not give any information about accuracy of 3D reconstruction whatsoever. Another article (Feng, Feng, Wyatt, and Liu, 2016), shows how reduced resolution or frame quality can negatively impact feature detection and tracking, using SIFT features. Other articles (Şahin IşÄśk and Özkan, 2014; El-Mashad and Shoukry, 2014) evaluated another set of feature extractors, but use evaluation metric as repeatability, recall or precision. These metrics are very important and gives insight to accuracy of detection and detection rate of the feature extractor. These metrics do not, however, give an insight into accuracy of 3D reconstruction, which is interesting for the application scenarios discussed in this thesis.

Therefore, in this study, the feature extractors were evaluated using the epipolar constraint as defined in the book (Hartley and Zisserman, 2004), which describes the geometrical constraint to obtain an accurate 3D point representation of point correspondences in stereo images. Moreover, the evaluation was carried out to provide practical insights about the robustness of the feature extractors, especially the prominent ones, i.e., SIFT, SURF and ORB, against variation in internal camera perturbations (defocus / image blur, lens distortion, thermal noise, resolution change). This helps in understanding the operating ranges of the feature extractors for various perturbations.

4.2.1 Evaluation

In our paper titled "Evaluating Performance of Feature Extraction Methods for Practical 3D Imaging Systems" [details in chapter 8], the behaviors of SIFT, SURF, and ORB were investigated. This involved the following:

- Evaluation of robustness of feature extractors against change in scale (resolution), motion blur/focus, lens distortion and thermal noise.
- Evaluation on real video dataset, which is degraded by simulating the practical perturbations.
- Performance metrics were accuracy, detectability and computation time which represents the quality and cost of 3D representation.
- Identification of operating ranges of feature extractors that aids researchers and developers for design decisions of multiview 3D applications.

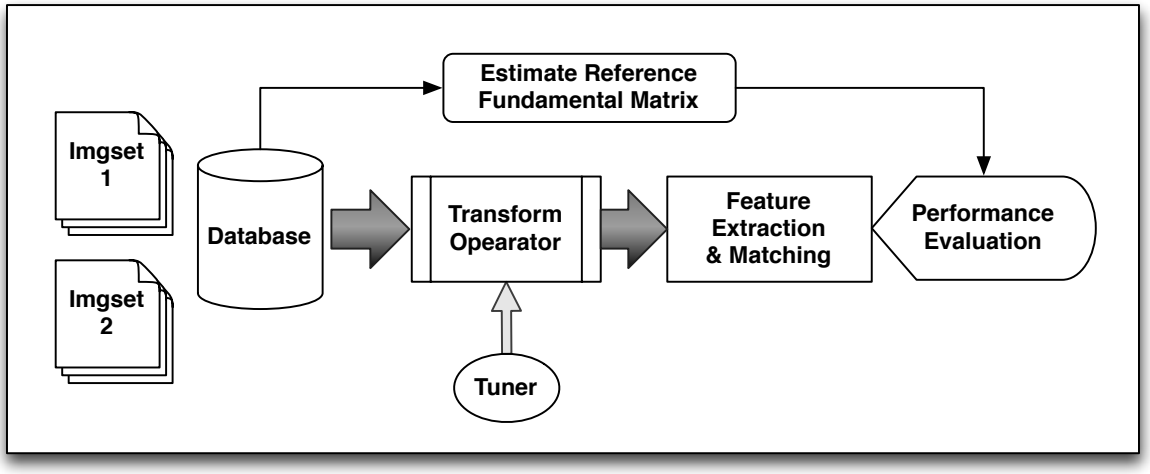


Figure 4.2: Evaluation Pipeline

The evaluation setup as illustrated in figure 4.2, is comprised of test dataset, image degradation module, feature extraction and matching module and finally a performance evaluation module.

Datasets

The first dataset is comprised of 30 stereo images of an opera performance conducted as an experimental mixed reality art performance that took place in Tromsø. Here, there were 8 cameras (2 camera arrays, each consisting of 4 cameras of narrow and wide lenses) to capture the scene. The second dataset comprised of 35 stereo images of a popular breakdance video sequence of Microsoft (Zitnick, Kang, Uyttendaele, Winder, and Szeliski, 2004).

Image Transformation

To study the performance of the feature extractors under practical scenarios, change of resolution, defocus, lens distortion and noise were simulated using the mathematical models. Using the simulated perturbations, the test stereo images from the dataset were transformed.

Image blur is the loss of image sharpness caused due to defocus, shallow depth of field and motion of the camera or the scene objects and quantization process. In this study, the focus was on image blur due to defocus only, because this study assumed multiview capture using only stationary cameras and hence motion blur was of lesser significance. Defocus $I_b(u, v)$ was accomplished by smoothing an image $I(u, v)$ with a linear 2D Gaussian filter $G(u, v)$, as in equations 4.1 and 4.2. Various defocus levels was controlled by the variance σ_b of the Gaussian kernel, which represented blur radius.

$$I_b(u, v) = I(u, v) * G(u, v) \tag{4.1}$$

$$G(u,v) = \frac{1}{2\pi\sigma_b^2} e^{-\frac{u^2+v^2}{2\sigma_b^2}} \tag{4.2}$$

Radial lens distortion is an optical aberration caused by spherical surfaces of the camera lenses. It produces aberrations symmetrically and radially from the image center. Barrel and pincushion (as explained in section 2.1) are the two types of radial distortions where the image aberration increases and decreases, as the radial distance from image center increases. Lens distortion was modeled as a 3rd order polynomial, as given by equation 4.3, where R_u and R_d are undistorted and distorted pixel radius, respectively. The distortion co-efficient k_1 was changed to obtain various levels of distortion.

$$R_d = \frac{R_u}{1 + k_1 R_u^2} \tag{4.3}$$

Image thermal noise appears as random speckles in an image which is random variation in the luminosity or color information of the pixels. This is caused by the camera sensor and its circuitry. Thermal noise was modeled as Gaussian distribution. A noisy image $I_n(u,v)$ was obtained by adding Gaussian random noise $N(u,v)$ with zero mean and variance σ_n to an image $I(u,v)$, as in equation 4.4. To obtain various noise levels N_l, measured in decibels, the variance σ_n was controlled as, $\sigma_n = 10^{N_l/10}$.

$$I_n(u,v) = I(u,v) + N(u,v) \tag{4.4}$$

Scale/Resolution is characterized as image size in pixels. The choice of setting the resolution depends on the application needs. However, evaluating feature extractors for various resolutions would help to make the right choice. All images in the dataset were of HD resolution. These images were re-scaled into three categories H-high (1280x960), M-medium (640x480) and L-low (320x240) resolutions for the evaluation.

Feature Extraction and Matching

This module detects interesting features in the stereo images, computes descriptors for them and then matches the descriptors to obtain feature correspondences between the stereo pair of images. At the time of this experiment, prominent extractors such as SIFT, SURF and ORB detectors-cum-descriptors were considered. For matching features, brute-force method was used. The working principle of each of the feature extractors are briefly described in table 4.1. Feature correspondences were extracted on both, the original image dataset and the pre-transformed (through simulated perturbations) dataset for comparing the performance. An example of feature extraction in stereo images for various datasets are shown in figure 4.3.

Performance Measure

The performance of feature extractors are measured using the following metrics:

Accuracy was measured in terms of *Epipolar Error*, which is based on epipolar geometry of stereo images. Researchers (Faugeras, 1993; Hartley and Zisserman, 2004; Ma, Soatto,

(a) Opera dataset - Wide lens with SIFT detected points

(b) Microsoft dataset with ORB detected points

(c) Opera dataset - Narrow lens with SURF detected points

Figure 4.3: Stereo images from various datasets low resolution 320x240.

Figure 4.4: Illustration of epipolar geometry. Courtesy Multiview Geometry (Hartley and Zisserman, 2004)

Kosecka, and Sastry, 2003) have shown that in 3D imaging systems, the geometrical relationship between the point correspondences between stereo images is important and is characterized by a mapping matrix called *Fundamental Matrix* (F). More details on the properties and computation of fundamental matrix are found in section 2.1.1.

The epipolar geometry is illustrated in figure 4.4. Ideally, for every point in one of the stereo images (say $\hat{x}$), a corresponding point on the other stereo image ($\hat{x'}$) should lie on a line, called *epipolar line* (l'), which was computed using the matrix F (Hartley and Zisserman, 2004). To obtain a reference baseline measurement, F matrix was estimated using the original stereo images without any added perturbations.

Feature extractors operated on the perturbed image dataset and the resulting feature correspondences (x and x') lied outside the line and thus produced a deviance (d'). Such a deviance averaged over all feature points was termed as *Epipolar Error* (E_p), which is expressed as the squared *Sampson error*.

Sampson error is the first-order approximation to the geometric error (Hartley and Zisserman, 2004). The E_p between feature correspondences (x, x') in a stereo pair is computed as in equation 4.5, where F is the fundamental matrix computed using N_p feature correspondences. This aided in measuring the accuracy of feature extractors, in pixels. Typically, the sub-pixel errors, i.e., $Ep < 1$ pixel, is considered an acceptable value for good performance in most of the relevant applications.

$$E_p = \sum_{i=1}^{N_p} \frac{x_i' F x_i}{(Fx_i)_1^2 + (Fx_i)_2^2 + (F^T x\prime_i)_1^2 + (F^T x\prime_i)_2^2} \tag{4.5}$$

Detectability measures the ability to obtain sufficient feature point correspondences in stereo images required to estimate a sufficiently good quality estimation of fundamental matrix (atleast 7 feature corresponding points for 7-point algorithm (Hartley and Zisserman, 2004)). Therefore, the percentage of trials resulting in at least 7 feature correspondences over all tested dataset represented the detectability of a feature extractor. This measure is similar to the repeatability measure of the feature detectors.

Computation time measures the computational speed of the feature extractors. It is computed as the total time required for detection, description and matching of features.

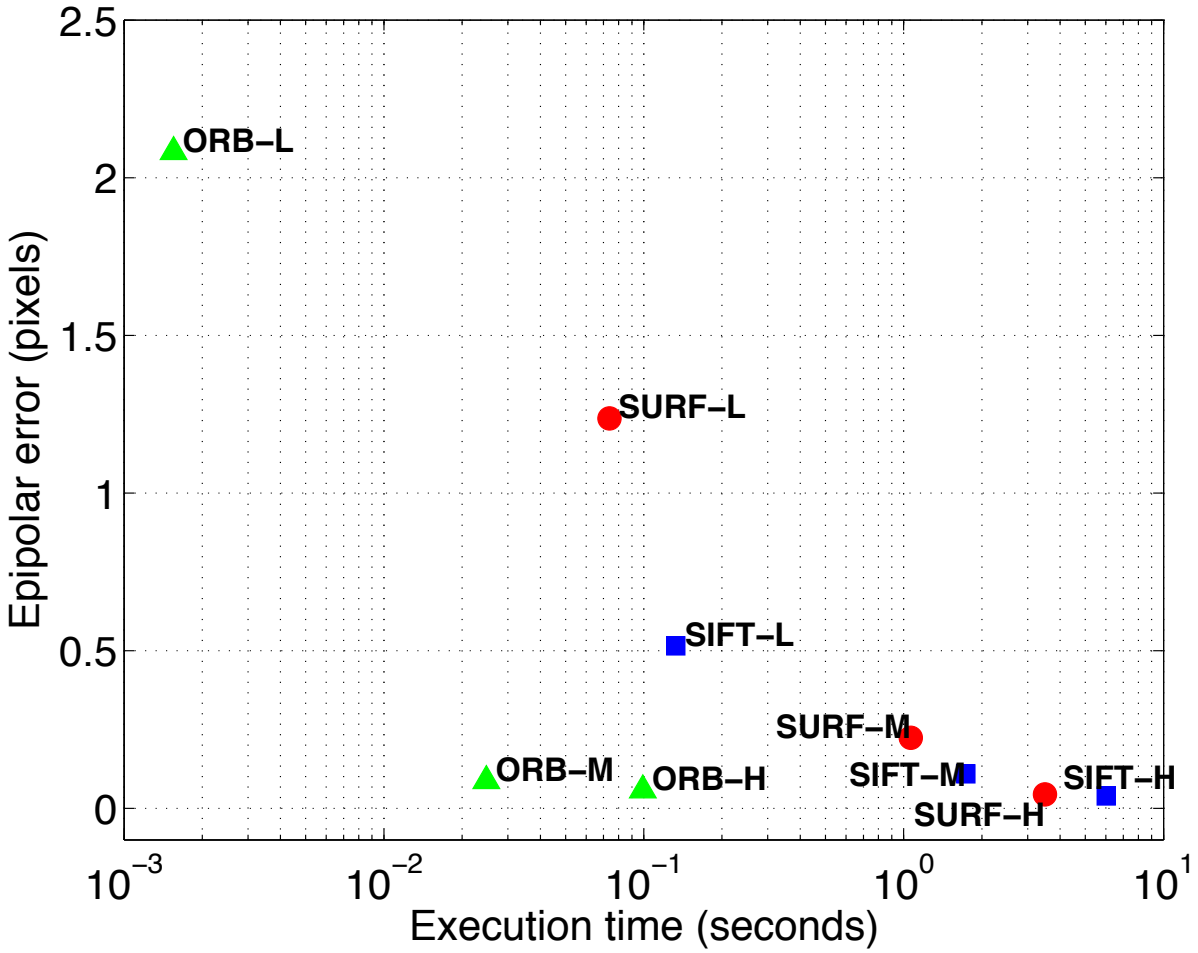


Figure 4.5: Accuracy Vs Computational time. The post-fixes refers to the size of the images: L-low resolution (320x240), M- medium resolution (640x480), H- high resolution (1280x960)

4.2.2 Accuracy Vs Speed

Ideally, any 3D system would require high accuracy and high speed computation. Accordingly, experiments to measure accuracy of the feature extractors at different resolutions and their computational speeds were conducted. Figure 4.5 shows the results of the test (note the execution time is plotted in logarithmic scale). Overall, ORB was computationally efficient compared to SIFT and SURF at all resolutions. A relative difference in execution time between SIFT and SURF was significant; SURF reduced the computational cost by 48% at all resolutions.

For individual feature extractors, accuracy increased with the increase in resolution. Comparatively, SIFT, SURF and ORB resulted in acceptable (sub-pixel) accuracy, except for SURF-L and ORB-L. This shows that SIFT is more robust to change in scale. At lower resolutions, very few features are detected, and even if there is a substantial number of features, they are erroneous due to pixel resolution.

It is shown how the perturbations (blur, distortion and noise) affect the performance at specific resolutions and eventually leads to identification of operational limits of feature extractors under various conditions modeled by simulated parameters. For every perturbation, *Epipolar Error* and *Detectability* was computed at low, medium and high resolutions.

4.2.3 Image Blur

Feature extraction was evaluated on blur levels ranging from 1.5 - 6.0. These values are the variances of the Gaussian filter used for blurring the images. An example of feature extraction on blurry images for level 5.0 is as shown in the figure 4.6. The performance of feature extractors for blurred images is shown in figure 4.7.

SIFT outperformed over all resolutions. SIFT seems to be more robust to blur levels probably because of its way of finding key-points, which uses scale space representation with various blur levels. At lower resolutions, blur-ness has a greater effect and hence SIFT showed an acceptable accuracy up to blur level 4.5 (figure 4.7(c)). SURF performed marginally at acceptable accuracy ($E_p <= 1$, figure 4.7(c)) up to blur level 4.5, at low resolutions, for the same reasons mentioned for SIFT. However, the difference in accuracy between SURF and SIFT is due to the nature of descriptor construction. SURF integrates the gradient information and loses distinctiveness when blur-ness increases, while SIFT uses individual gradient to create the descriptor and sustains the performance to a larger extent of blur. ORB performed to an acceptable accuracy at medium and high resolutions (figures 4.7(b) and 4.7(a)) up to blur level 3.5.

The detectability measure (figure 4.7(d)) for both SIFT and SURF reduced drastically with increase in blur level at low resolution, which makes them unsuitable to use when low resolution images are blurred, especially at levels > 4.5. ORB performed the least in terms of detectability. After level 4 of blur, ORB features were not found at all (see also figure 4.6(c)). The use of huge box filters in ORB to obtain descriptors seems to limit its performance on blurry images. Additional blur worsens the efficiency of the descriptor. Hence, ORB failed at low resolutions.

4.2.4 Lens Distortion

Feature extraction was evaluated on barrel and pincushion distortion levels from 10% to 50%. An example of feature extraction on barrel distorted image level 40%, is shown in the figure 4.8. Performance of feature extractors for lens distorted images is shown in figure 4.9.

All the feature extractors performed well and similar at high and medium resolution. At low resolutions SIFT outperformed SURF, which in turn outperformed ORB; however, all of them exhibited a constant detectability. Overall, the performance of SIFT, SURF and ORB at high and medium resolutions seems to be unaffected by lens distortion. It should be noted that this result was for a homogeneous stereo pair where the distortions are assumed to be of same degree in both the cameras.

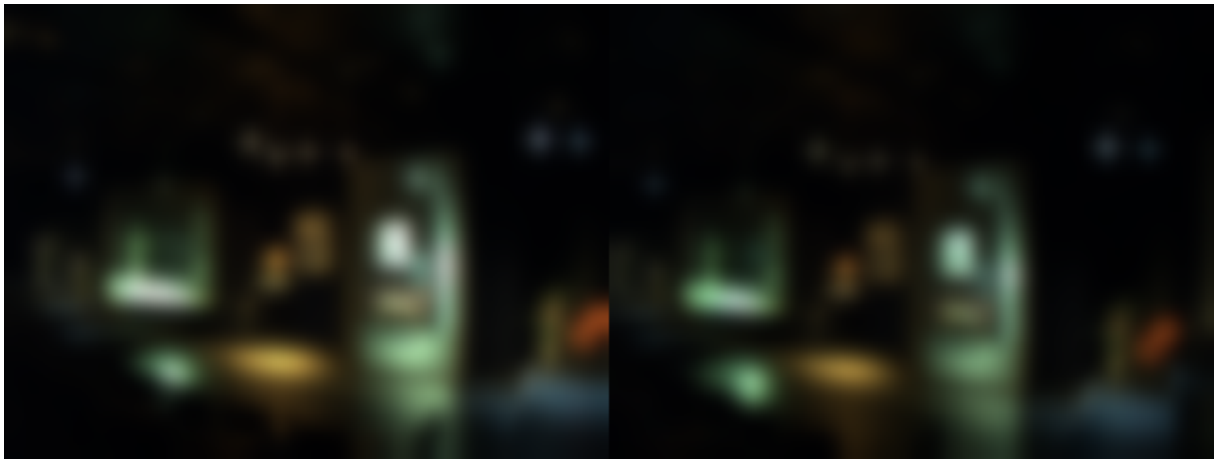
4.2.5 Sensor Noise

Feature extraction was evaluated on noisy images, where noise level ranged from 5dB - 50dB. An example of feature extraction on noisy images for level 15dB is as shown in the figure 4.10. The performance of feature extractors for noisy images is shown in figure 4.11.

SIFT outperformed SURF and ORB, at all resolutions and exhibited resilience to thermal noise, but became sensitive to noise at around 15dB for low resolution images. This effect can also be seen in figure 4.10(a). SURF and ORB showed resilience to noise up to 20dB and 15dB, respectively, at both high and medium resolutions. Importantly, a high and constant detectability rate (figure 4.11(d)) was observed for SURF and ORB, suggesting that the performance of

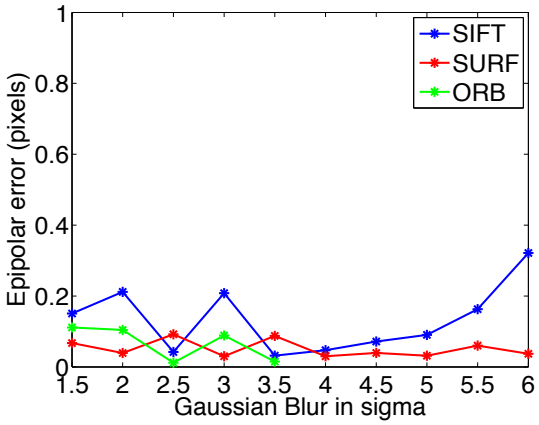
(a) SIFT on blurred images

(b) SURF on blurred images

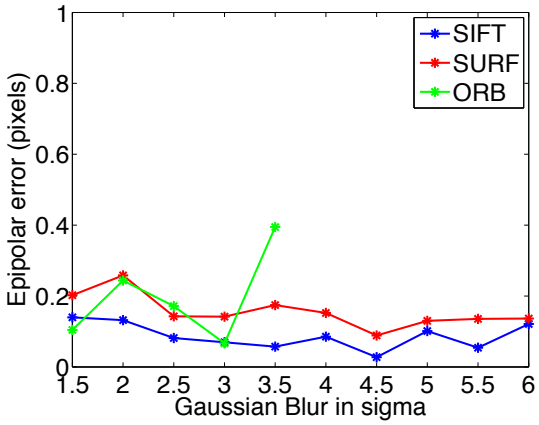


(c) ORB on blurred images

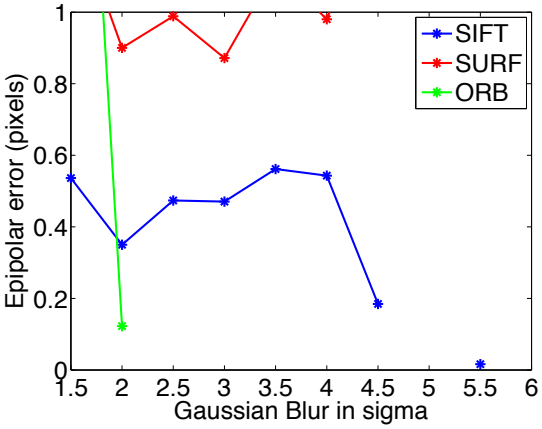
Figure 4.6: Feature extraction on blurred (radius level 5) stereo images from Tromsø dataset with wide lens of L-low resolution 320x240.



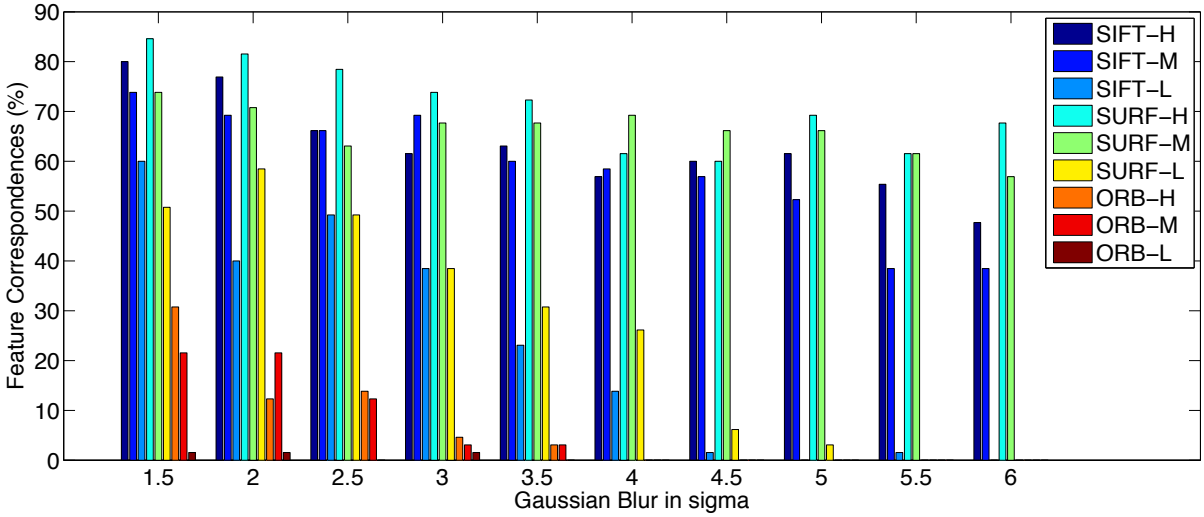
(a) Blur at High Resolution



(b) Blur at Medium Resolution

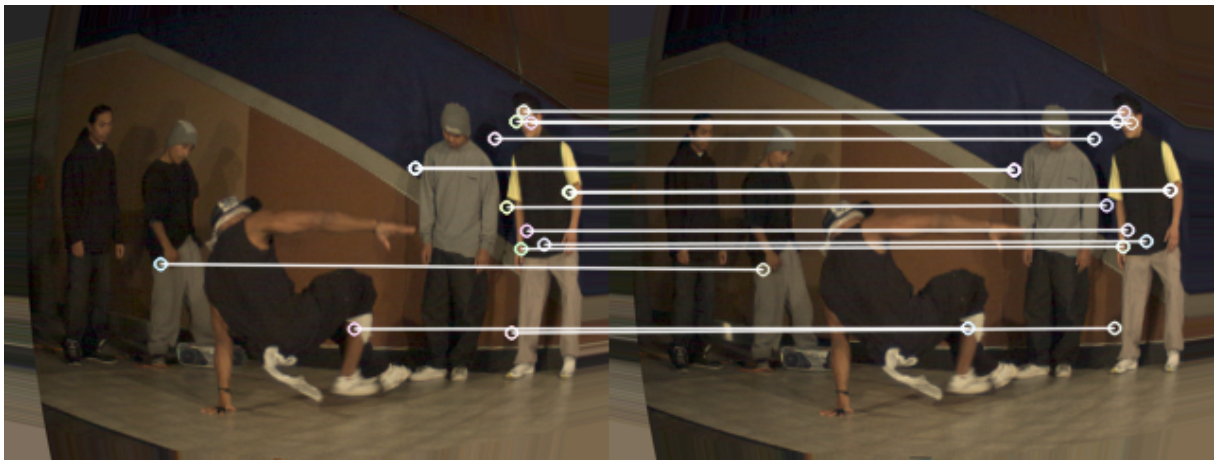


(c) Blur at Low Resolution

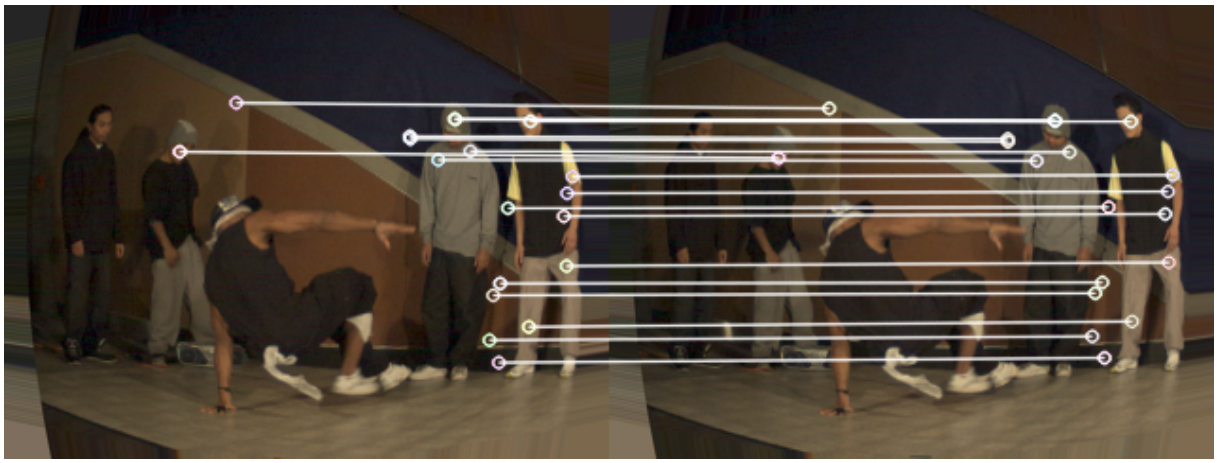


(d) Detectability for blurred images

Figure 4.7: Performance of feature extractors for simulation of blur levels over various resolutions



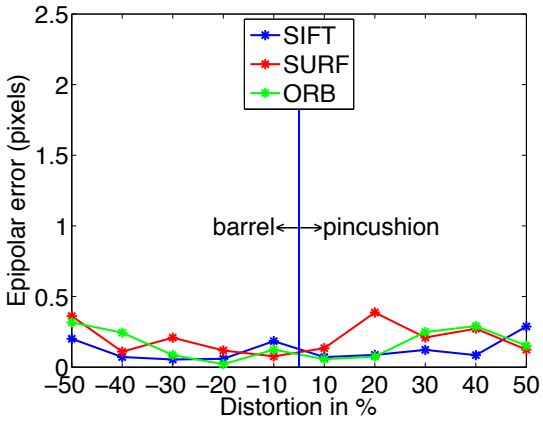
(a) SIFT on distorted images



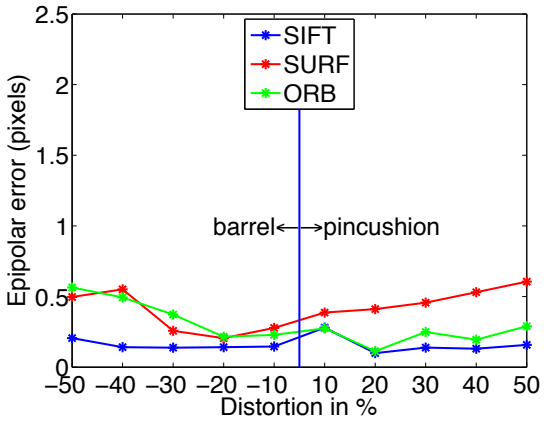
(b) SURF on distorted images

(c) ORB on distorted images

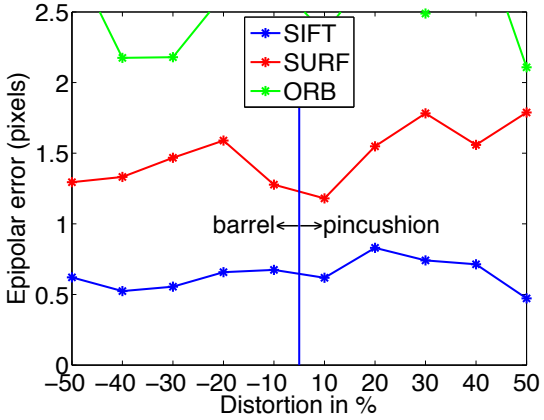
Figure 4.8: Feature extraction on barrel distortion (level 40%) stereo images from Microsoft dataset of L-low resolution 320x240.



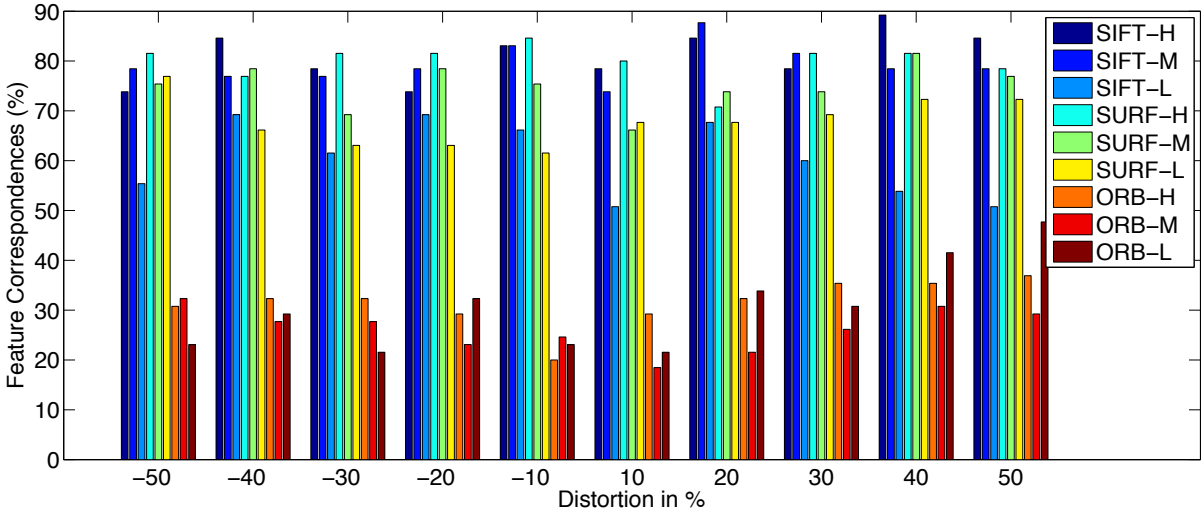
(a) Distortion at High Resolution



(b) Distortion at Medium Resolution



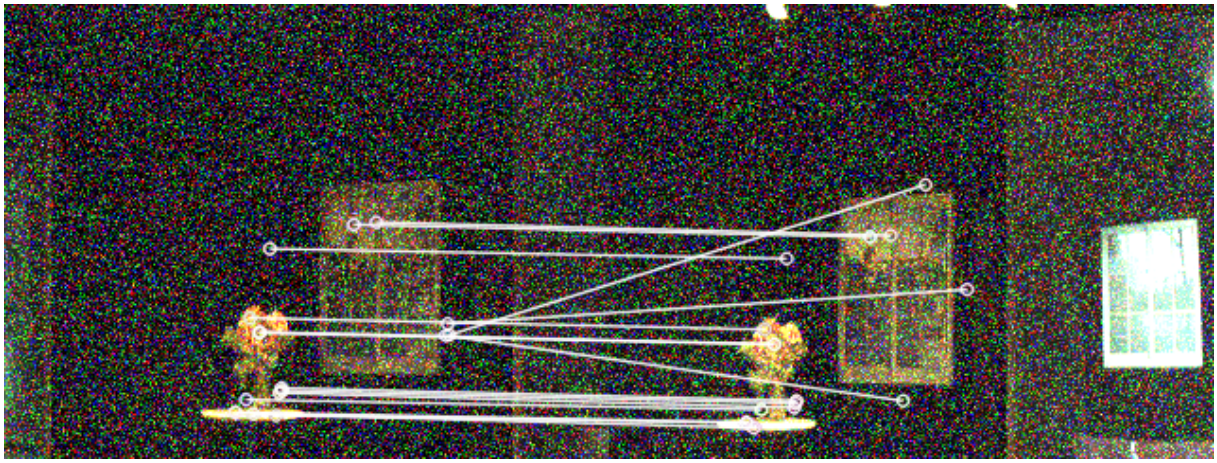
(c) Distortion at Low Resolution



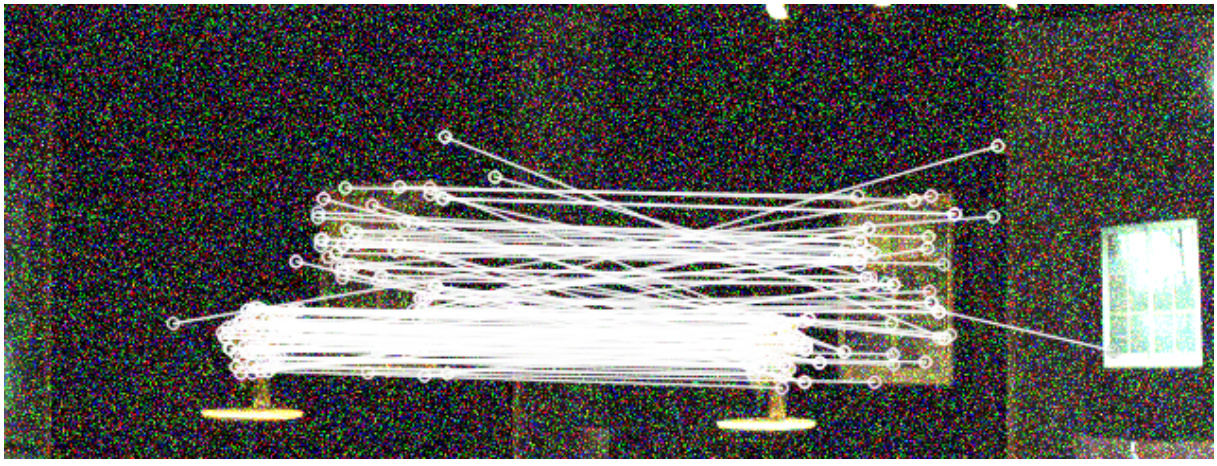
(d) Detectability for Distorted images

Figure 4.9: Performance of feature extractors for simulation of distortion levels over various resolutions

(a) SIFT on noisy images

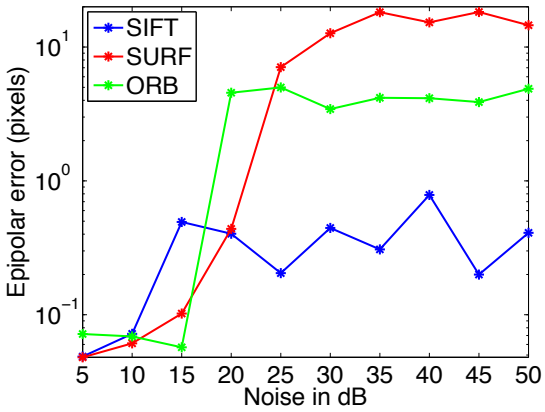


(b) SURF on noisy images

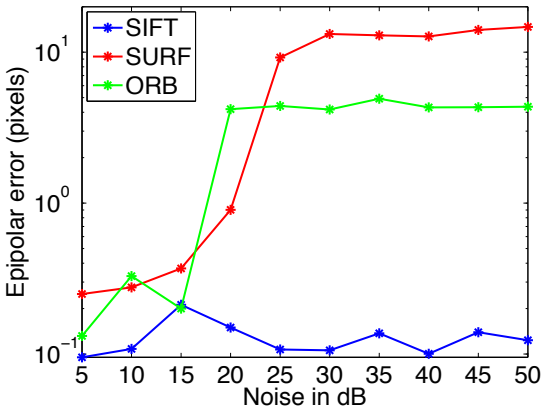


(c) ORB on noisy images

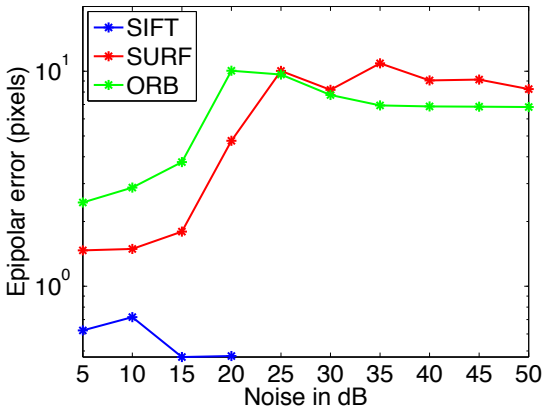
Figure 4.10: Feature extraction on noisy (15dB) stereo images from Tromsø dataset with narrow lens and L-low resolution 320x240.



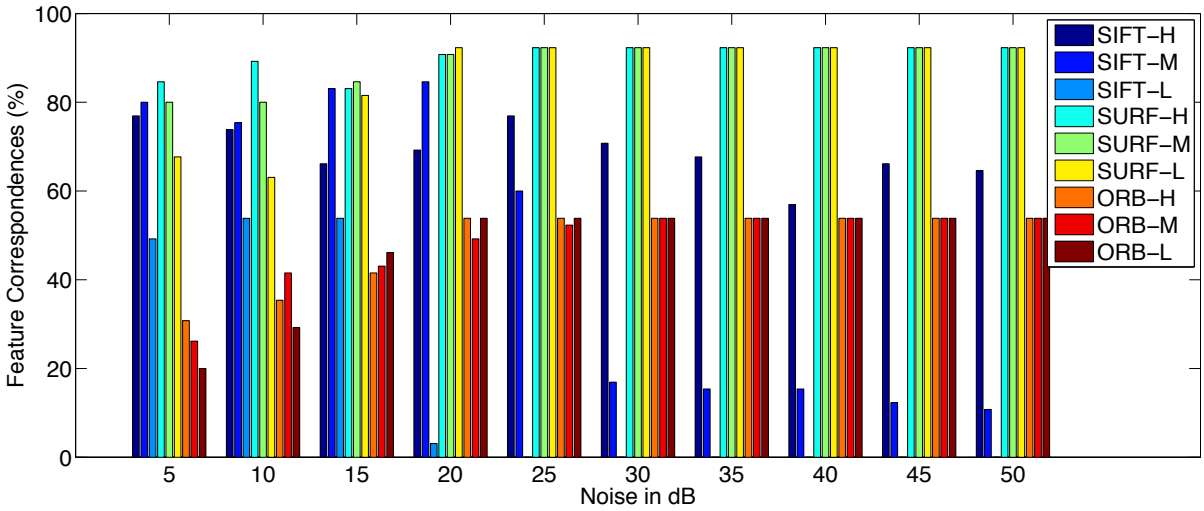
(a) Noise at High Resolution



(b) Noise at Medium Resolution



(c) Noise at Low Resolution



(d) Detectability for Noisy images

Figure 4.11: Performance of feature extractors for simulation of noise levels over various resolutions

SURF and ORB are not affected by noise, but the accuracy is too low ($E_p > 10$ pixels). This behavior was because of wrong detection of features in noisy images. Hence, under noisy conditions, above 15dB none of the feature extractors perform within the acceptable accuracy.

4.2.6 Discussions

The effects of blur, distortion and noise on SIFT, SURF and ORB feature extractors were investigated and the variation of accuracy, detectability and computational time was noted. From all these, the observations were as follows:

At resolutions $>$ 320x240:

- SIFT and SURF were the best choices. However, choosing SURF would save execution time of 48%, on an average, with a cost of around 0.10 pixels in accuracy.

At resolution 640x480:

- For blurry images, SIFT is the best choice. However, using SURF would save 48%, on an average with a cost of 0.22 pixels in accuracy.
- For lens distorted images, SIFT, SURF and ORB all are good choices. By using ORB, the execution time reduces by 98.12% and 95.27% with a cost of 0.69 pixels and 0.33 pixels in accuracy compared to SIFT and SURF, respectively.
- For noisy images, SIFT and SURF are good choices. SURF saves 32% time with a cost of 0.67 pixels in accuracy.

Unlike other feature evaluations, the *Epipolar Error* was used to measure the accuracy of the feature correspondences between the stereo pair. The *Epipolar Error* represents the geometrical error, which is close to evaluating the 3D reconstruction compared to other metric, such as re-projection error. Therefore, *Epipolar Error* was used to aid the selection of feature extractors for FBC and other 3D applications. A choice of feature extractor can be made considering the above conclusions on robustness and the resolution requirements.

All the above evaluation tests are sufficient to reject part of *Null Hypothesis III* (as stated in section 1.4), and hence, it can be concluded that the performance of state-of-the-art feature extractors have significant differences to the change in internal parameters represented by scale, defocus, lens distortion and noise. The state-of-the-art feature extractors, especially SIFT, SURF and ORB were evaluated in terms of accuracy and computational time. This study has identified the operational limits of the feature extractors that aids researchers and developers of multiview applications.

4.3 Robustness against Camera Extrinsic

In the previous section, the effects of changes in camera internal parameters on the accuracy of feature extractors were discussed. In this chapter, the effects of changes in camera external properties such as relative rotation and translation of one camera with respect to another in a stereo camera setup is discussed.

In several multimedia 3D applications such as head mount virtual reality systems (Ribo, Pinz, and Fuhrmann, 2001; Yuan, 2006), augmented reality (AR) mobile applications (Bres and Tellez, 2009; Fragoso, Gauglitz, Zamora, Kleban, and Turk, 2011; Ventura and Höllerer, 2012), interactive augmented reality systems (Lima, Simões, Figueiredo, and Kelner, 2010; Suenaga, Tran, Liao, Masamune, Dohi, Hoshi, and Takato, 2015), free-viewpoint rendering (Min, Kim, Yun, and Sohn, 2009), etc, two or more cameras are used to perform tasks such as augmenting 3D models in video sequences, depth estimation, virtual view synthesis, etc. The underlying principle of such multiple camera systems is the estimation of camera pose, i.e., relative camera position and orientation with respect to other cameras. In AR applications, high quality relates to an accurate augmentation of virtual objects in the real scene. For this, it is required to know the accurate position of the observer. This is regarded as solving image-based location problem by an accurate camera pose (relative position and orientation).

Let us consider one such AR application scenario in movie production - POPART (more details in section 2.2.4), which aims at providing an augmented preview of the scene shot during the movie production. In this setup, a primary camera shoots the movie and two auxiliary stereo pair estimates the camera pose with respect to the scene, so that virtual objects can be placed in the scene accurately for the preview. In this case, the accuracy of placing the virtual objects is highly dependent on the accuracy of the camera pose estimation, i.e., the camera extrinsic calibration. These calibration pose parameters are used to integrate the animated 3D models into the view of real-life actors on the set. This helps the directors or cinematographers to preview the augmented scene and analysis way ahead of the post production time.

The camera extrinsic calibration or pose estimation is carried out based on the detection of sparse feature correspondence that are extracted in the scene using the two auxiliary stereo camera pair located as described in the POPART setup. This process refers to feature based calibration (FBC).

The accuracy of FBC can be measured in 3D space using the metric *Normalized Correlation Coefficients* (η) as stated in the equation 4.6. η provides a similarity measure of estimated 3D points (X_e) with the ground-truth 3D points (X_g), which is represented as a normalized accuracy value [0-low and 1-high].

$$\eta^{\dagger} = \frac{\sum(X_e^{\dagger} - mean(X_e^{\dagger})) * (X_g^{\dagger} - mean(X_g^{\dagger}))}{\sqrt{\sum(X_e^{\dagger} - mean(X_e^{\dagger}))^2 * \sum(X_g^{\dagger} - mean(X_g^{\dagger}))^2}} \tag{4.6}$$

$$\eta = \sum_{\dagger=x,y,z} \frac{\eta^{\dagger}}{3}$$

where † represents 3D axes components x, y and z.

One important factor that decides the accuracy of the FBC and thereby the system itself, is the change in camera baseline (the angular displacement between the two camera positions). In multiple camera systems, the following statements are commonly accepted:

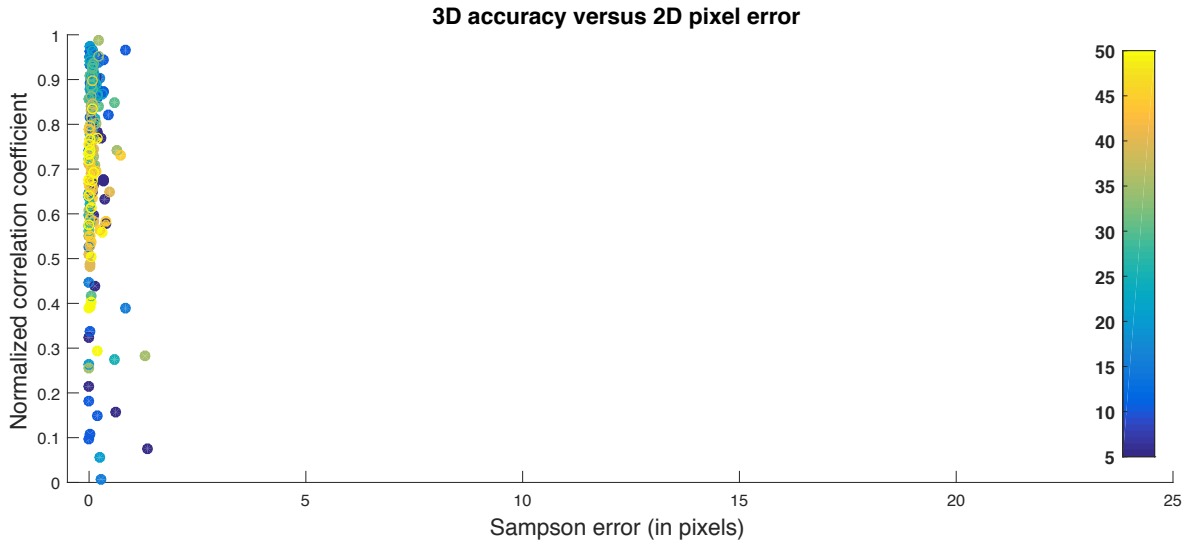
- A high number of matched feature points in a stereo pair results in a better camera pose estimation.
- Minimizing 2D pixel error calculated between matched pairs results in higher accuracy of 3D estimation, based on epipolar geometry (Hartley and Zisserman, 2004).

The first point holds good for iteration-based estimation algorithms (e.g., RANSAC (Fischler and Bolles, 1981)). The second point, however, is not always true. This is illustrated in figure 4.12, which represents a scatter plot of 3D accuracy (measured in Normalized Correlation Co-efficient as in equation 4.6) versus 2D pixel error (measured in Sampson error as in equation 4.5) and number of matched feature points extracted from images of stereo pair at various baselines (relative displacement between the stereo cameras). In figure 4.12, the colors represent different camera baselines and there exists many data points of each color that represent various feature extractors. Figure 4.12(a) showed that low pixel error does not guarantee high 3D accuracy and, similarly, figure 4.12(b) showed that high 3D accuracy is not always obtained by a larger number of feature matches. Therefore, it becomes very important to explore one of the important factors determining the accuracy of FBC, i.e., change in the camera baseline, which breaks the common assumptions made above.

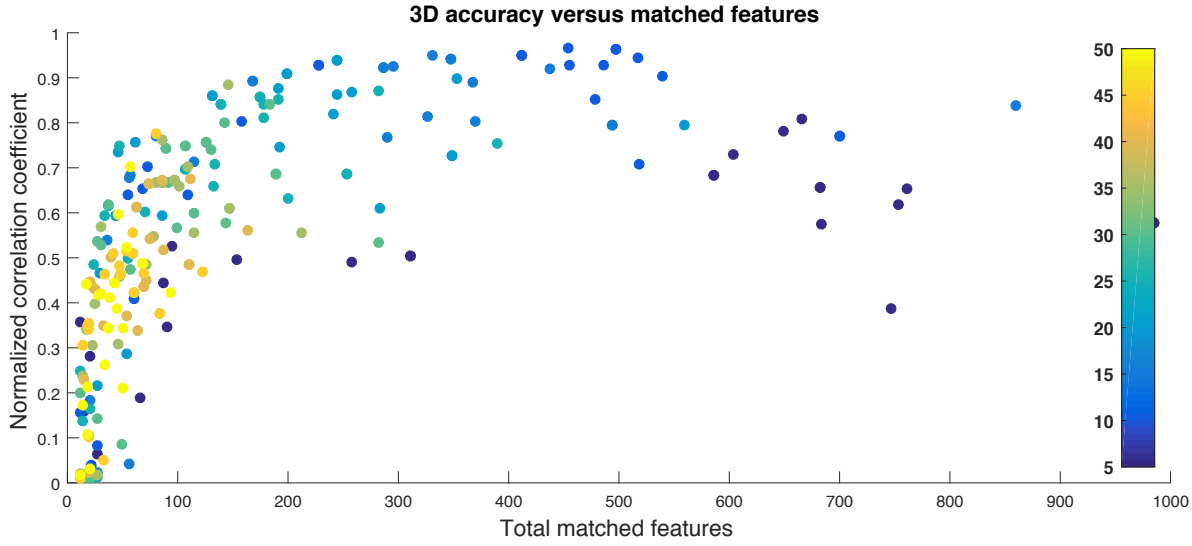
The combination of detectors & descriptors (Agrawal, Konolige, and Blas, 2008; Alahi, Ortiz, and Vandergheynst, 2012; Alcantarilla, Bartoli, and Davison, 2012; Bay, Ess, Tuytelaars, and Van Gool, 2008; Calonder, Lepetit, Strecha, and Fua, 2010; Leutenegger, Chli, and Siegwart, 2011; Lowe, 2004; Matas, Chum, Urban, and Pajdla, 2002; Pablo Alcantarilla and Bartoli, 2013; Rosten and Drummond, 2006; Rublee, Rabaud, Konolige, and Bradski, 2011) are used for FBC or camera pose estimation today. Each of these feature extractors has its own behavioral traits. Some of them claim invariance to change in camera baseline, but the extent of their tolerance is uncertain.

Previously, the evaluation of most of the state-of-the-art feature extractors, i.e., detectors or descriptors, have used various evaluation criteria. The feature detector KAZE (Alcantarilla, Bartoli, and Davison, 2012), feature descriptors FREAK (Alahi, Ortiz, and Vandergheynst, 2012) and BRIEF (Calonder, Lepetit, Strecha, and Fua, 2010) evaluated themselves with other known feature detectors using recall and precision metrics, which relates to a total number of correct feature matches found. Along with recall and precision, BRISK (Leutenegger, Chli, and Siegwart, 2011), STAR (Agrawal, Konolige, and Blas, 2008), FAST (Rosten and Drummond, 2006) and AKAZE (Pablo Alcantarilla and Bartoli, 2013), evaluated themselves in comparison to others, using the metric called repeatability, which measures the extent of overlap between the detected regions in an image pair. In both SIFT (Lowe, 2004) and SURF (Bay, Ess, Tuytelaars, and Van Gool, 2008), the evaluation was carried out on various viewpoints, but not in comparison with other features. However, the performance criteria was still repeatability. Sometimes, the distance between the descriptors was considered to be an evaluation metric, as in ORB (Rublee, Rabaud, Konolige, and Bradski, 2011). In all the above cases, the evaluation criteria focused only on the correctness of the feature matches and this may not be enough to evaluate the feature extractors for accuracy in 3D applications and robustness to camera baseline changes.

Point feature matching algorithms for stereo were evaluated (Juhász, Tanács, and Kato, 2013), but only for a particular baseline based on the re-projection error metric. In this study, a range of baselines were evaluated and their effects were studied. Interest point detectors and descriptors were evaluated for tracking applications (Gauglitz, Höllerer, and Turk, 2011), where detectors were tested on various conditions such as scale, rotation, baseline, light, etc., using repeatability metric. Further, feature detectors were compared based on tracking success rate, which was computed based on the re-projection error. However, KAZE, AKAZE, BRISK,



(a)



(b)

Figure 4.12: Scatterplots of matched feature points and 2D pixel error with 3D accuracy.

BRIEF and FREAK were not included in their study. Moreover, instead of measuring the re-projection error in 2D, the accuracy was measured in 3D space directly, relying on a dataset consisting of known 3D models. Intuitively, 3D space metrics seems more suitable for AR related applications.

SIFT feature extractor was evaluated (Michael Ying Yang, 2011) for viewpoint invariance, by comparing the descriptor properties over various baselines. Their evaluation basically outlined the quality of obtaining correct matches, but that does not guarantee high 3D accuracy. Feature tracking for pose estimation in underwater environment was evaluated (Shkurti, Rekleitis, and Dudek, 2011). However, their evaluation was limited to very few feature detectors and descriptors with a very specific testing condition. Feature extractors for 3D object recognition applications over various viewpoints and lighting conditions was evaluated (Moreels and

Perona, 2007), but with a limited number of candidates for evaluation.

Therefore, in this study, a wide range of feature extractor combinations were evaluated with a brute-force matcher to determine their robustness to change in the camera baseline. This study is meant to provide system builders with a better understanding of the operational limits of the state-of-the-art feature detectors and descriptors for various baselines. It helps system builders to make better choices while designing 3D multimedia applications using multiple camera systems. Besides the choice of algorithm, the study is also helpful in estimating the number and position of cameras that are required for reconstructing rigid structures in a well-known space, with a desired accuracy.

4.3.1 Evaluation

A virtual dataset was used for evaluation. So, complete control was obtained on the dataset. From the virtual dataset, known camera calibration parameters were extracted, which mapped ground-truth 2D and 3D points. First, these values were tested on the evaluation pipeline, before conducting the experiments. Later, these ground-truth values were used for computation or as a reference for comparing the experimental values.

In our paper titled "Robustness of 3D Positions to Camera Baselines in Markerless Augmented Reality (AR) Systems" [details in chapter 11], state-of-the-art feature extractors were characterized for its robustness against the camera baseline changes. This study involved the following:

- Evaluation of many state-of-the-art feature extractors over change in camera baseline.
- Evaluation based on complete test pipeline, i.e., feature based calibration and 3D reconstruction from stereo images.
- Design recommendations given the accuracy, execution time and reliability factor for each feature extractor combination.
- Performance metric used was Normalized Correlation Co-efficient (η), based on ground-truth data obtained from virtual dataset.

The evaluation setup is depicted as in figure 4.13. Every stereo pair from the dataset was used to extract feature correspondences between the stereo pairs, followed by camera pose estimation. The estimated and ground-truth camera poses were used for triangulating the ground-truth 2D test data points. The resulting 3D points from both estimated and ground-truth camera poses were compared to obtain the error in the system.

Dataset Generation

Ground-truth data was generated based on the assumed camera arrangement as in figure 4.14. The camera was assumed to be placed on a circular configuration and the displacement between the stereo camera centers was considered the *camera baseline*. The circular configuration was chosen because it nullifies the scaling factor of the object or the object distance from the camera and hence the focus lies on the baseline variation. Based on this configuration, stereo images of

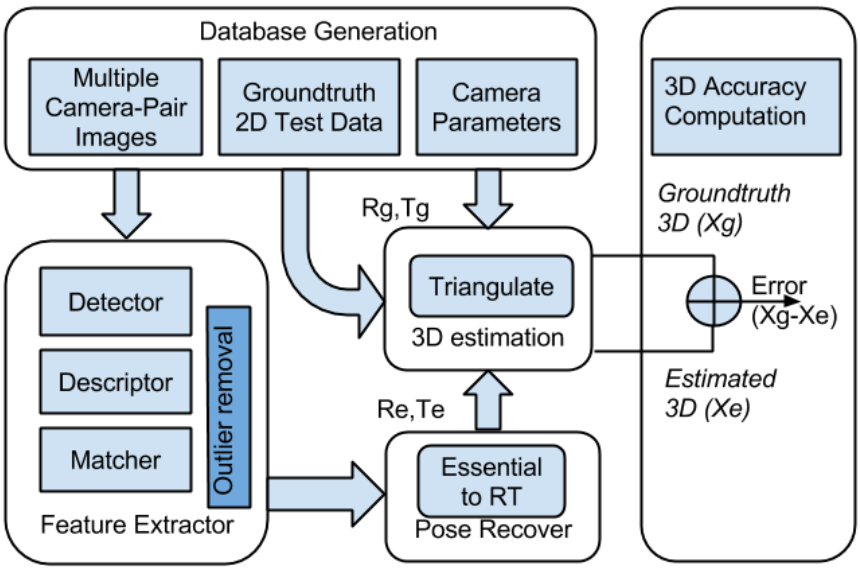


Figure 4.13: Experimental setup.

3D models were captures at various baselines from a distance of about 300 model units. Captured 3D models are as shown in figure 4.15 (obtained from CG Trader[1]), with stereo camera of known camera calibration parameters. The camera intrinsic parameters $[K]$ comprises camera's focal lengths (f_x,f_y) and principal axes (p_x,p_y). The camera extrinsic parameters represents relative rotation and translation of stereo pair ($[R_g$,$T_g]$).

$$K = \begin{bmatrix} f_x & 0 & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{bmatrix}, R_{g_{3X3}} = \begin{bmatrix} r_{x1} & r_{y1} & r_{z1} \\ r_{x2} & r_{y2} & r_{z2} \\ r_{x1} & r_{y3} & r_{z3} \end{bmatrix}, T_{g_{3X1}} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}$$

Four points that represents an origin and three unit vectors in three orthonormal directions was considered as the ground-truth 3D points $[X_g]$. These 3D points are sufficient to represent a model measured in world coordinate system, with a geometric center of the model as the origin. This was well suited ground-truth data, which when compared with estimated 3D data, resulted in changes in both position and rotation of the model in 3D space. By projecting the ground-truth 3D points onto the image plane using calibration parameters, ground-truth 2D points $[x_g^1$ and $x_g^2]$ were generated.

With cameras of similar focal length of 520 pixels, principal axes of 300 and image resolution of 600x600 with 24 bit depth, about 450 stereo pairs were generated with the camera baseline variation from 1 to 50 degrees with a step of 1 degree angular displacement.

[1] http://www.cgtrader.com

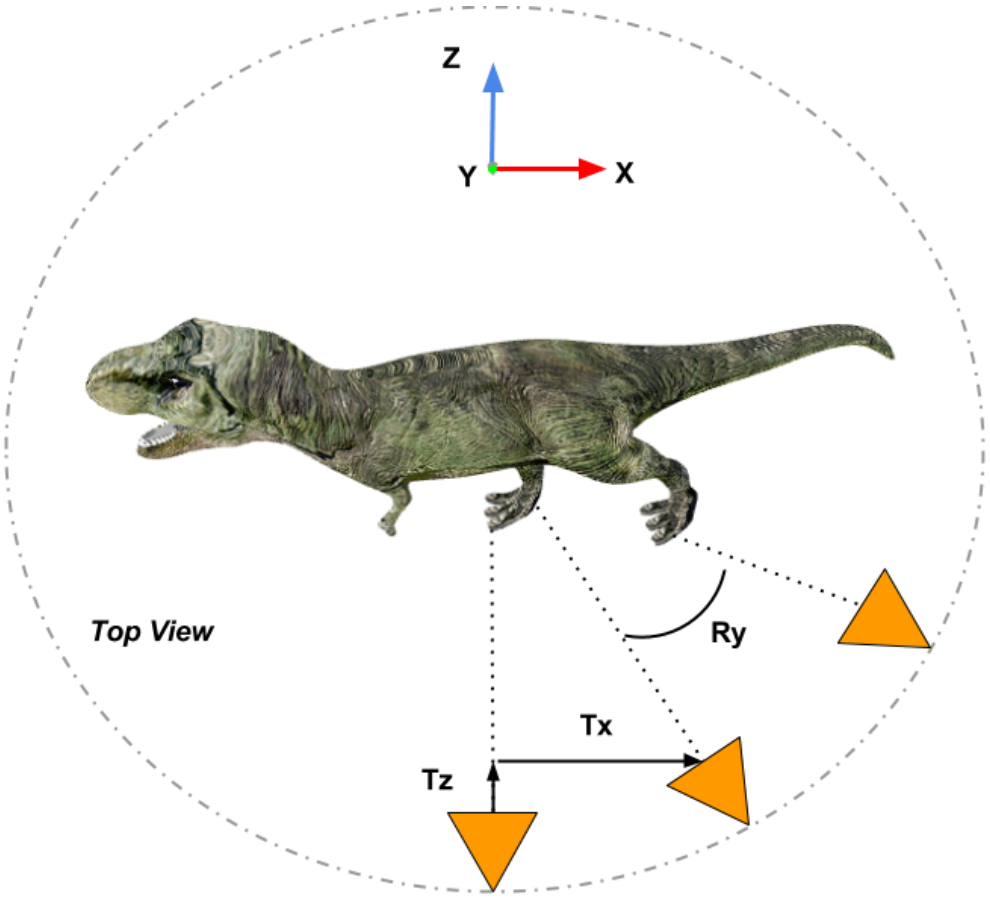


Figure 4.14: Cameras arranged in a circular configuration around the 3D model.

(a) (b) (c) (d) (e)

(f) (g) (h) (i)

Figure 4.15: The 3D models used for the experiment. From each model, 50 stereo image pairs are generated, corresponding to various baselines. Courtesy: CG Trader.

Feature Extraction

Feature extractors obtained by combining the detectors SIFT, SURF, BRISK, KAZE, AKAZE, ORB, MSER, STAR and FAST, with their own descriptors, and in combination with BRIEF, FREAK and other descriptors (All detectors and descriptors are briefly explained in section 4.1) were tested. In total, 26 feature extractor combinations were evaluated. To compute feature correspondences in a stereo pair, a brute-force matcher on the descriptors was applied combined with Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981) for removal of outliers. Each feature extractor was applied to every camera pair configuration to extract feature correspondences [x_e^1, x_e^2] between the stereo images.

Pose Recovery

Pose recovery estimates the pose (camera position and orientation [R_e,T_e]) of the right camera with respect to the left camera in a stereo pair. The Essential matrix (E) was estimated using feature correspondences [x_e^1, x_e^2] and camera intrinsic matrix [K] based on 5-Point algorithm (Nistér, 2004). The essential matrix is a specialized case of fundamental matrix expressed in normalized image coordinates that describes the relation between the stereo pair in terms of epipolar constraint [${x_e^2}^T E_{ss} x_e^1 = 0$] (Hartley and Zisserman, 2004).

3D Accuracy Computation

Instead of measuring re-projection error (project the estimated 3D point on to the image plane and compare with the known ground-truth 2D values), accuracy computation in 3D space was considered to be more relevant for the application scenario being evaluated.

The feature correspondences [x_e^1, x_e^2] are used to estimate camera pose [R_e,T_e]. It is unfair to use the same correspondences as test data, which might imply a bias test. Therefore a new test data, which is the ground-truth 2D datapoints [x_g^1, x_g^2] was used.

The 3D point estimation was carried out by triangulating (Hartley and Zisserman, 2004). Now, by triangulating the 2D test points with known camera pose [R_g,T_g], ground-truth 3D point [X_g] were obtained and by triangulating the test points with estimated camera pose [R_e,T_e], 3D points [X_e] were estimated. Then the accuracy in 3D space was measured using the metric *Normalized Correlation Coefficients* (η) (equation 4.6) as a normalized accuracy value [0-low and 1-high].

Testing Schemes

The evaluation tests were carried out based on the following pointers:

1. 450 stereo pairs * 26 feature extractor combinations, i.e., 11700 datasets were tested.

2. The virtual dataset comprised of a single 3D model with empty background, which is similar to a movie shot with blue screen as background. This can also be related to a textured background scene, where the test models of the dataset represent the scene at a specific depth.

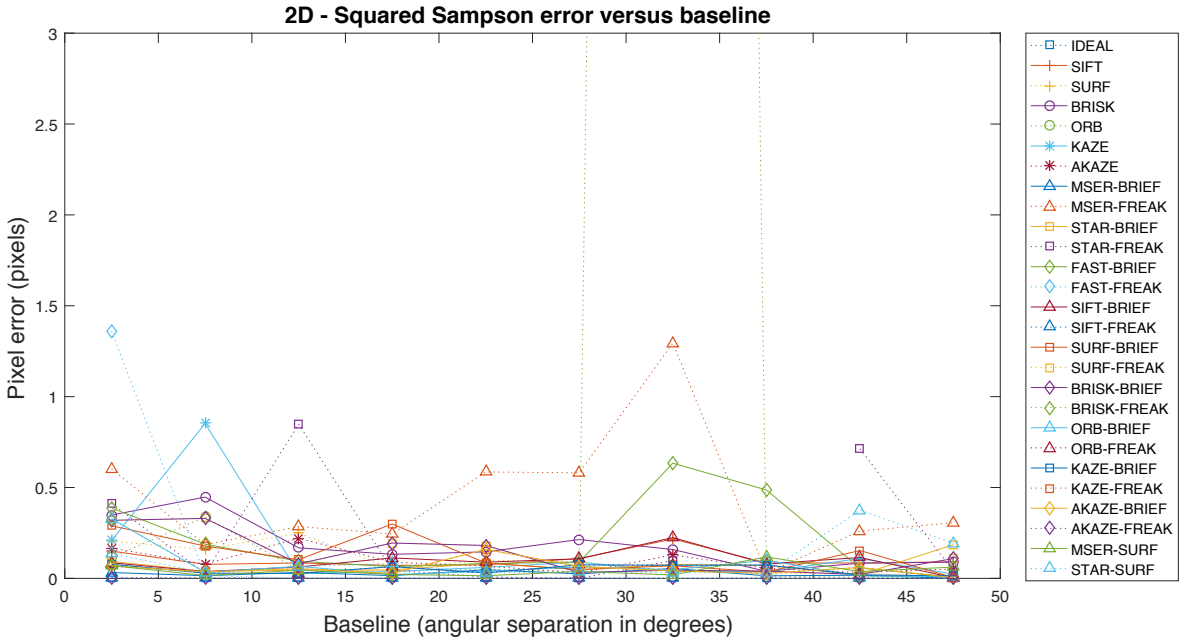


Figure 4.16: The 2D error (Squared Sampson) based on epipolar constraint over varied baselines

3. 'Camera Baseline' referred to the angular displacement of stereo pair of cameras placed on a circular configuration.

4. Ground-truth 2D features was used in the pipeline and considered it as an 'IDEAL' feature extractor, which is a noiseless feature correspondence. This was as good as a blackbox testing of the pipeline itself.

5. The estimation error was expressed as averaged error over every 5 degrees of the camera baseline. This was to gain better readability as the variability within the 5 degrees is hard to present visually.

6. The performance of all feature extractors were evaluated at all stages of the complete pipeline described in figure 4.13, i.e., 2D pixel error, camera pose error, penalty/reliability and 3D estimation error.

For all the results, the X-axis represented the baseline expressed between 1-50 degrees. Along the Y-axis, the error was averaged over every 5 degrees, to increase readability.

4.3.2 2D Pixel Error

The 2D pixel error was expressed as $EpipolarError$ (E_p) (Hartley and Zisserman, 2004) and was computed as in equation 4.5. This metric determines how close every point feature in one image is to its corresponding epipolar line in the other image of the stereo pair. For an ideal match, $E_p = 0$.

The E_p measured for 2D stereo pairs, varying in baseline is shown in figure 4.16. This error was computed for all meaningful combinations of feature extractors (described in the table 4.1).

The pixel error stayed fairly low (although fluctuating) over all camera baselines. This means that the features on one image were very close to the epipolar line in the other image. The epipolar constraint was maintained here, however, this did not guarantee a consistent accuracy of 3D estimation for all camera baselines as seen in figure 4.12(a). This is more evident in the results of camera pose error.

4.3.3 Camera Pose Error

Next, the estimated camera pose was compared with known camera extrinsic obtained from the dataset. The deviations of the estimated camera rotation and translation parameters from the ground-truth value were summed up over all three axes. The results are as shown in the figures 4.17 and 4.18. In these figures, 'Own' descriptors refer to the feature detectors that have their own descriptors, e.g., SIFT, SURF, etc.

Each figure is categorized into sub-figures based on the descriptors used. It is evident that pose errors do not follow the same pattern as in figure 4.16, which signifies that the variation in camera baseline has an effect on the camera pose estimation. As the baseline of the stereo camera increased, the pose estimation error increased (figures 4.17(a), 4.17(b), 4.18(a) and 4.18(b)) or stayed high throughout (figures 4.17(c) and 4.18(c)). This behavior is due to the following reasons:

1. When wrong feature matches between the stereo pairs exists, the estimation of fundamental matrix becomes incorrect. This is quite obvious.

2. When correct feature matches between the stereo pairs exists, and if the feature matches are confined to a small area, i.e., a set of 2D match points corresponds to only a part of the 3D model, then the estimation of fundamental matrix becomes incorrect as there is not enough information about rotation or translation covering the whole 3d model.

In both of the above cases, an incorrect fundamental matrix and thereby an incorrect estimation of essential matrix resulted in an incorrect pose estimation. The 2D pixel error seemed like a biased measure because the same number of feature points were used for both to estimate fundamental matrix and to compute pixel error based on the fundamental matrix. Due to this nature, although incorrect fundamental matrix was used, the 2D pixel error still stayed low over all baselines (figure 4.16), as an effect of using RANSAC.

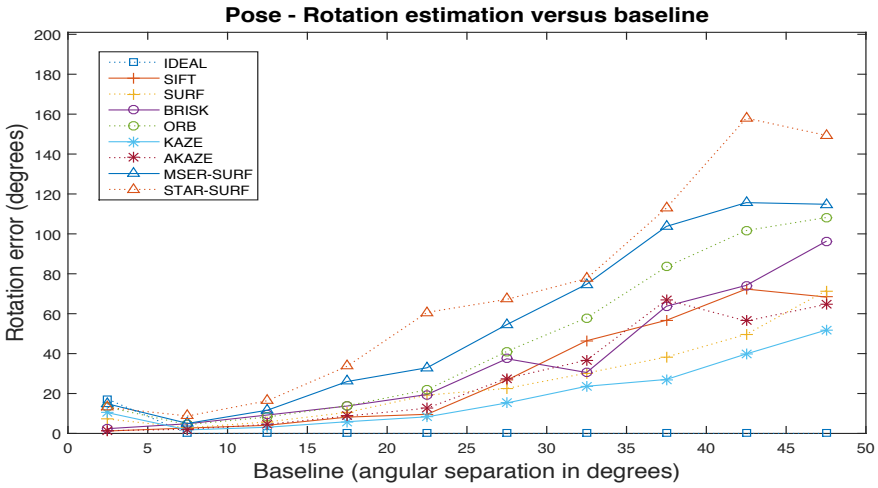
4.3.4 Penality

In the process of estimating the camera pose, three types of invalidity can occur.

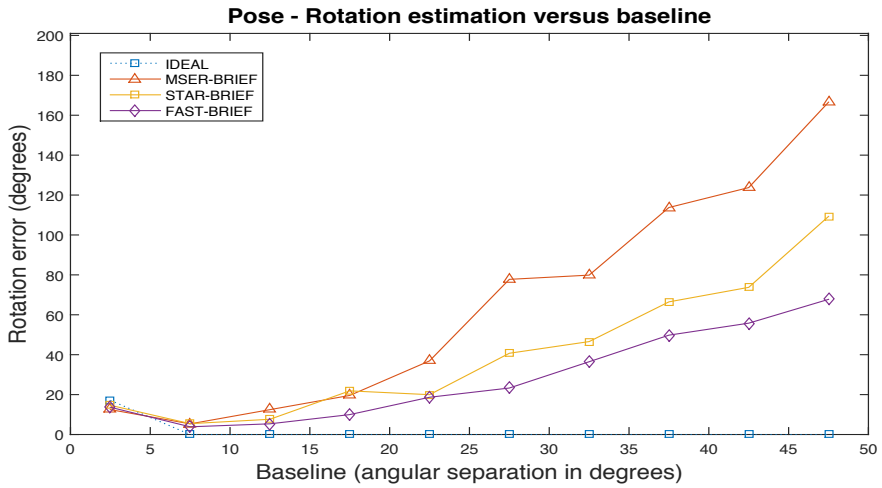
- Type 1 - when camera rotational error in either of the three directions is more than $90°$, then the camera seems to be rotated more than expected, in a true situation.

- Type 2 - if any of the translation error is more than unity, then it means that the right camera is estimated to be on the left side.

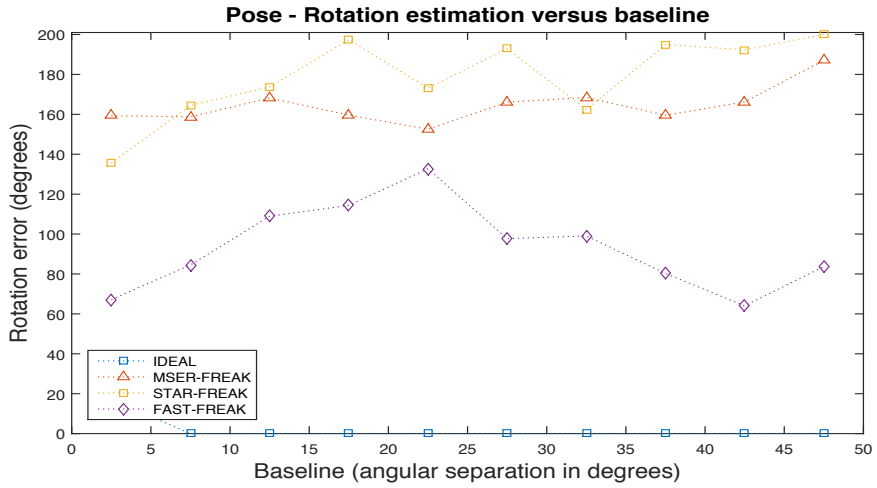
- Type 3 - though not directly related to pose estimation, this error occurs when the feature extraction gives zero matches. This error also relates to non-estimation of fundamental matrix due to very few matches.



(a) Own Descriptors

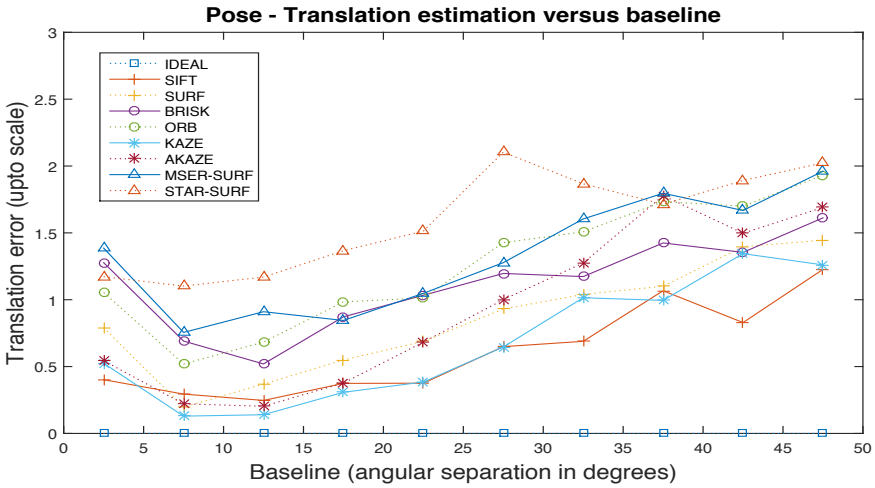


(b) Brief Descriptor

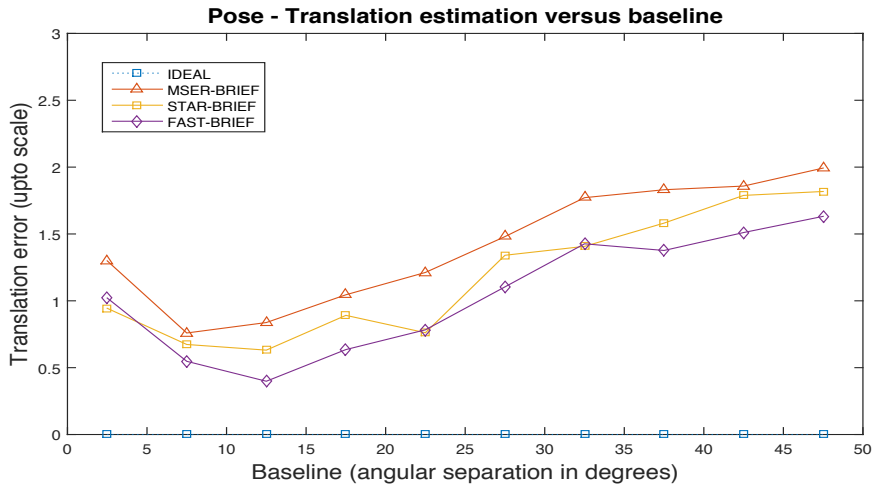


(c) Freak Descriptor

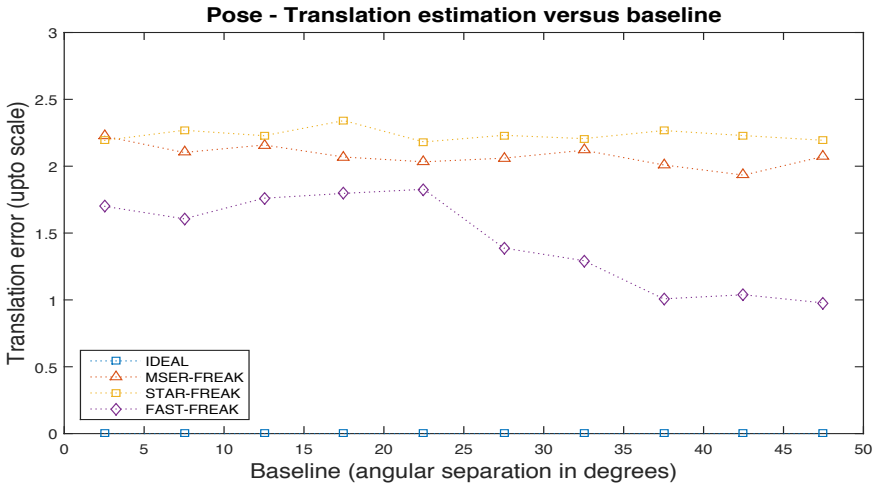
Figure 4.17: Rotation error of stereo camera over varied camera baselines.



(a) Own Descriptors



(b) Brief Descriptor



(c) Freak Descriptor

Figure 4.18: Translation error of stereo camera over varied camera baselines.

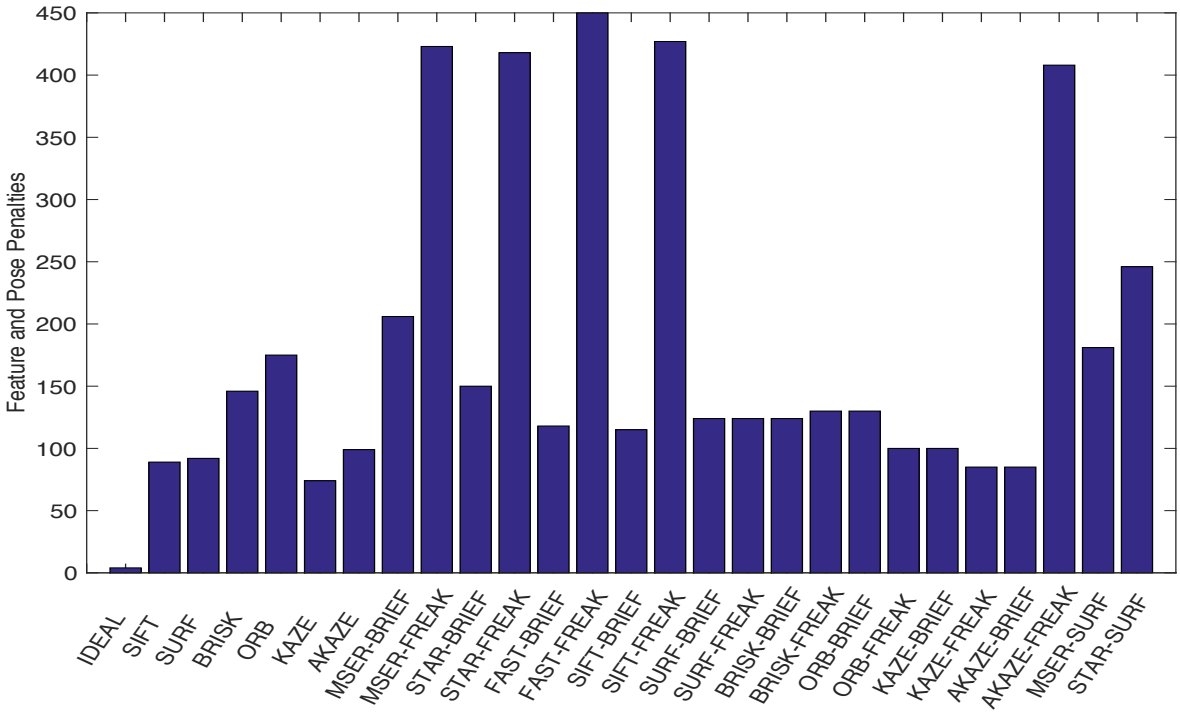


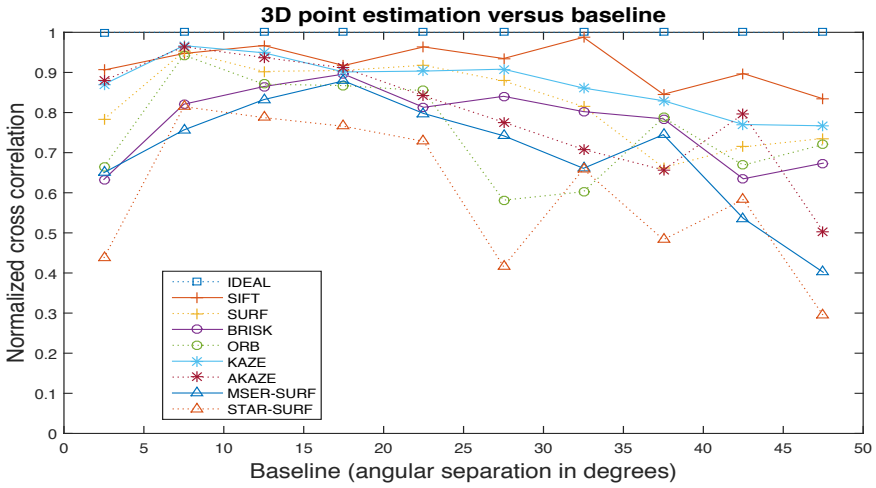
Figure 4.19: Penalty values for all feature extractors.

In the above cases, the camera pose estimation is deemed invalid. This situation can occur, when the number of feature correspondences in a stereo pair are zero or very few or wrong to a large extent. In these cases, the feature extractor was penalized, whenever any of the above types of invalidity occurs. Therefore, every feature extractor combination gets a penalty score for the invalidity.

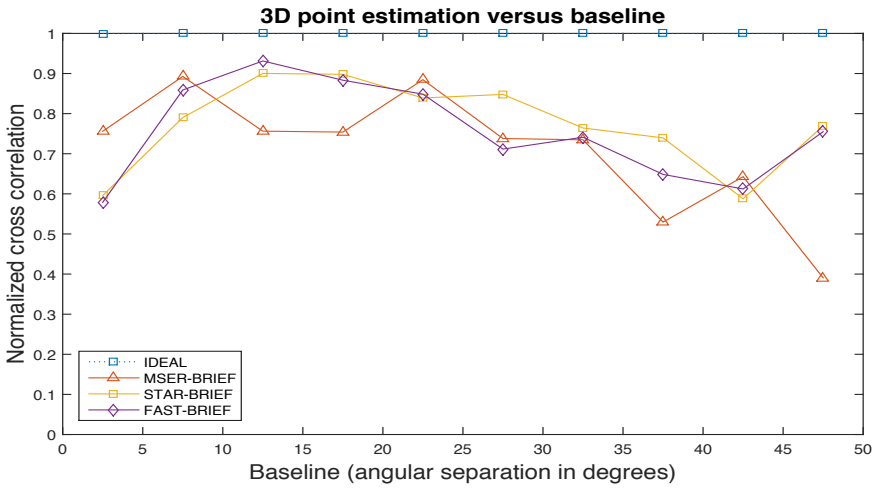
Figure 4.19 shows the penalty score of each of the feature extractors. The penalty scores are used to measure the feature extractor's reliability that is used as a factor to recommend a good feature extractor (explained in section 4.3.7). The penalty score is expressed as probability of success over 450 trials (450 stereo test data images). For example, a penalty score of 200 means that out of 450 stereo test images, the feature extractor failed to provide valid data in 200 images.

Mostly, the FREAK descriptor showed the highest penalty score, which means that the FREAK descriptors yielded wrong feature correspondence and camera pose estimation, and that affected the 3D point estimation.

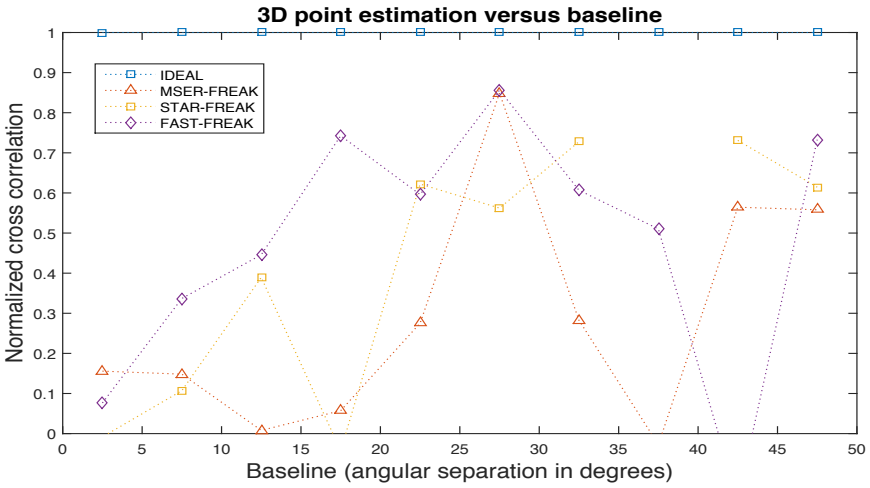
As can be seen in figure 4.19, the IDEAL feature extractor has a penalty of about 4 at very low baselines. This means pose estimation algorithm is sensitive at low baselines. This sensitivity has no effect directly on the comparative study on feature extractors, as penalty score samples are considered invalid.



(a) Own Descriptors



(b) Brief Descriptor



(c) Freak Descriptor

Figure 4.20: Mean 3D estimation error, categorized based on feature descriptors over varied camera baselines

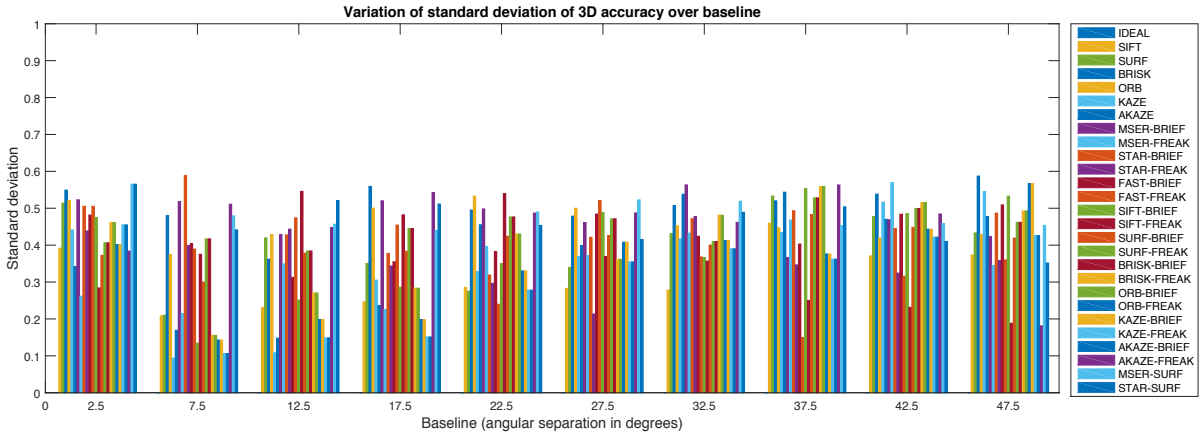


Figure 4.21: Standard deviation of 3D estimation over varied baselines.

4.3.5 3D Estimation Error

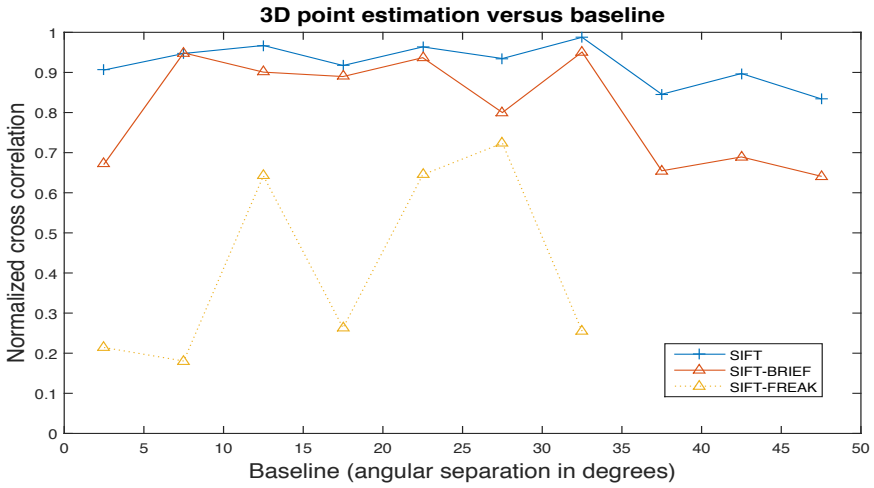
Using the feature correspondences and the recovered camera pose, the corresponding 3D points were estimated and compared to their ground-truth values. The resulting samples were filtered based on the penalty score. Only the samples that were not penalized were considered valid and used for further evaluation. The 3D accuracy expressed as η (equation 4.6) was measured against varied baselines and the results are shown in figures 4.20, 4.21, 4.22 and 4.23. Results in figures 4.20(a), 4.20(b) and 4.20(c) are categorized based on the descriptors used, such as OWN (means SIFT descriptor for SIFT detector and so on), BRIEF and FREAK. Known detectors such as SIFT, SURF, BRISK, ORB, KAZE and AKAZE were not only evaluated on their own descriptor but also with BRIEF and FREAK descriptors (figures 4.22 and 4.23). Accuracy is indicated as a mean value of η over every 5 degrees to increase the readability of the result, because the variability of η (as shown in figure 4.21) is hard to present visually.

The η value decreased over increase in camera baseline. 3D estimation was carried out using triangulation, which is conceptually a back-projection of rays originating from feature points in the image. Back-projection takes place with the help of intrinsic $\&$ extrinsic camera parameters. Since the intrinsic camera matrix was constant throughout the experiment, the camera extrinsic or camera pose was the variability that has an effect on the variability of 3D accuracy. This showed that the increase in camera pose error reduces the accuracy. This effect was evident by comparing figures 4.20 with 4.17 and 4.18. From this comparison, it is clear that low camera pose error yields high 3D accuracy. This is why markerless pose estimation becomes important in 3D applications.

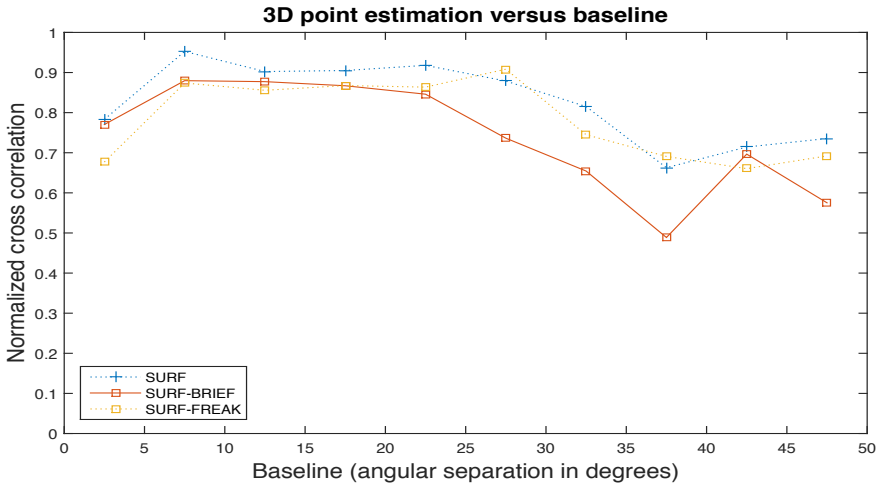
4.3.6 Comparative Performance

It is shown how the 2D pixel error cannot be used for evaluating feature extractors as it is a biased measure. It is also shown how the camera pose estimation was erroneous, which was due to the the feature correspondence error. The penalized feature extractors and invalid datasets were ignored in the calculation of the 3D accuracy.

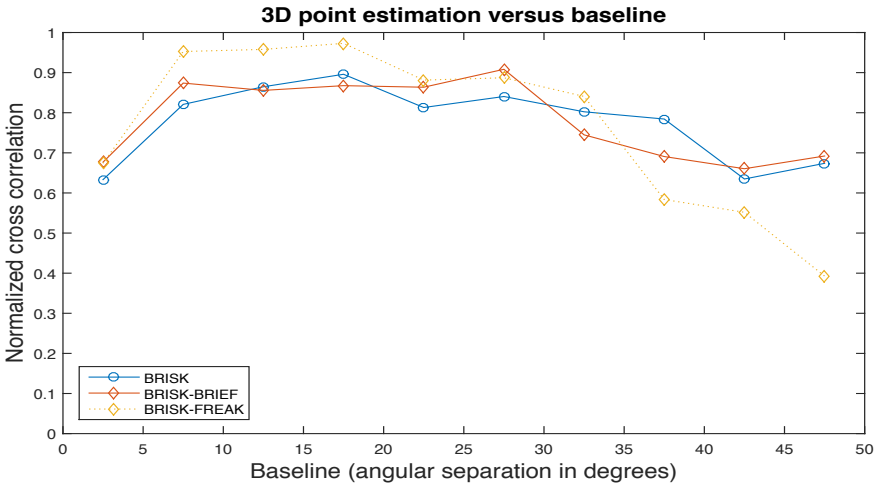
Overall, some of the feature extractors outperformed others and some of the descriptors



(a) SIFT Detector

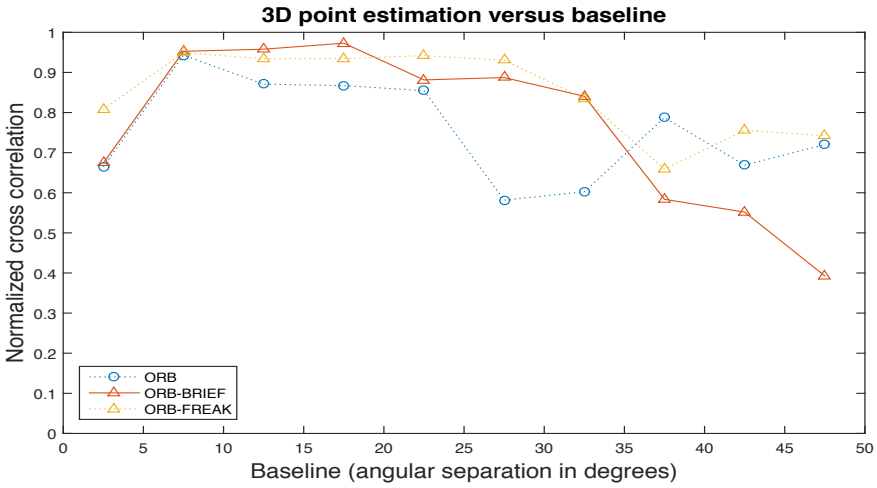


(b) SURF Detector

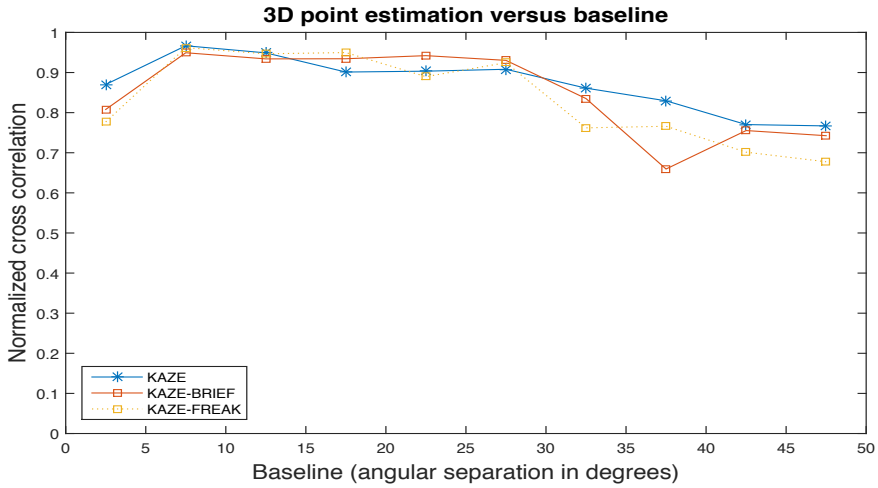


(c) BRISK Detector

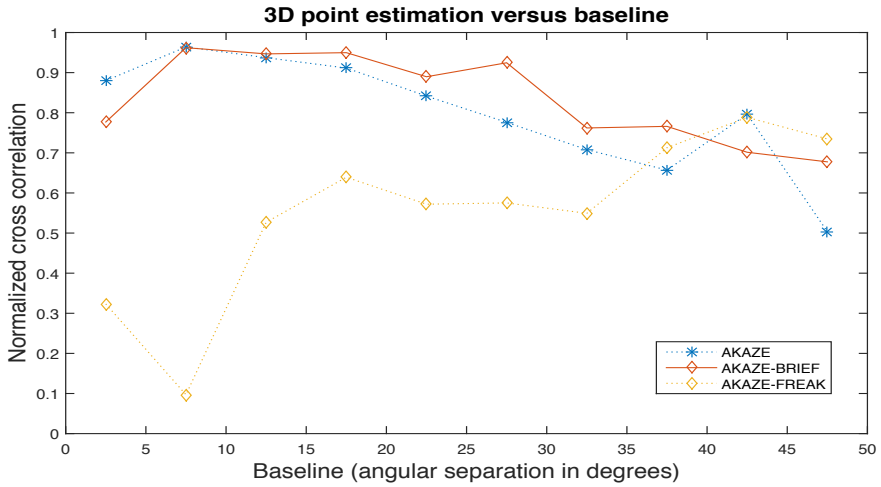
Figure 4.22: Mean 3D estimation error, categorized based on feature detectors (SIFT, SURF and BRISK) over varied camera baselines



(a) ORB Detector



(b) KAZE Detector



(c) AKAZE Detector

Figure 4.23: Mean 3D estimation error, categorized based on feature detectors (ORB, KAZE, AKAZE) over varied camera baselines

showed better performance when combined with certain detectors, over others. The important aspect to notice here is that each feature extractor performed relatively better in certain baseline ranges. For ease of explanation, the behavioral pattern of extractors over a certain range of baselines, i.e., ($< 5°$), ($5° - 30°$) and ($> 30°$) are analyzed. From the evaluation of the state-of-the-art feature extractors, the results were summarized based on the accuracy as follows (for more details please refer the publication in chapter 11):

- For baselines ($< 5°$):
  SIFT, KAZE and AKAZE were good performers (figure 4.20(a)). However, rotation-translation ambiguity existed during camera pose estimation. As explained before the sensitivity of pose estimation algorithm plays a role here. This behavior can also be seen in figure 4.19, where IDEAL penalty score was not zero. So, for other feature extractors the penalty and thus the pose error magnifies. Even a small deviation in the accuracy of feature correspondences yields a large pose estimation error (figures 4.17 and 4.18) and thereby triangulation errors.

- For baselines ($5° - 30°$):
  SIFT, SURF and KAZE with their own descriptors (figure 4.20(a)); BRIEF descriptor with all detectors except MSER, STAR and FAST (figures 4.22 and 4.23); FREAK descriptor with SURF, BRISK, ORB and KAZE were good performers (figures 4.22(b), 4.22(c), 4.23(a) and 4.23(b)).

- For baselines ($> 30°$):
  SIFT and KAZE performed better than others (figure 4.20(a)). However, SURF detector with both SURF and FREAK descriptors (figure 4.20(b)); BRIEF descriptor with BRISK, KAZE and AKAZE (figures 4.22(c), 4.23(b) and 4.23(c)) were the next candidates.

- It was claimed (Alcantarilla, Bartoli, and Davison, 2012) that KAZE performed as good as SIFT. But this holds good only upto a baseline of about $30°$.

- ORB claimed to be an alternative to SURF (Rublee, Rabaud, Konolige, and Bradski, 2011), but it was observed that, after a baseline of about $25°$, ORB did not perform better than SURF. This is probably because the modified BRIEF descriptor used in ORB is not as efficient as the SURF descriptor, which is based on Haar wavelets, in terms of rotational invariance for higher baselines.

- The MSER and STAR detectors have been evaluated using SURF descriptor in their original works (Matas, Chum, Urban, and Pajdla, 2002) and (Agrawal, Konolige, and Blas, 2008), respectively. Therefore, it was intended to use these combinations as well. It was observed that the SURF descriptor was better off with its own detector rather than MSER or STAR.

- In case of BRISK and ORB, as shown in figures 4.22(c) and 4.23(a), both the BRIEF and the FREAK descriptors performed better than their own descriptor upto $\approx 35°$. So, the BRIEF descriptor was more robust to baseline changes than the modified BRIEF (used in ORB) and the BRISK descriptor.

- Although AKAZE (Pablo Alcantarilla and Bartoli, 2013) showed a better performance over other detectors, it performed as good as KAZE upto $\approx 20^\circ$ baseline and then, the performance dropped down severely. Pose estimation error showed the same trend (figures 4.17(a) and 4.18(a)). However, AKAZE saves the computation time.
- SIFT and AKAZE, using the BRIEF descriptor (figures 4.22(a) and 4.23(c), respectively), maintained their accuracy similar to that of using their own descriptors, upto $\approx 35^\circ$ baseline. Moreover, BRIEF descriptor is advantageous in terms of computation time.

4.3.7 Discussions

Although η gives a relative performance measure of feature extractors, it is difficult to use this information directly for practical applications. For making a sensible choice of feature extractors for a specific 3D application, feature extractors need an absolute measure that gives a sense of quality of service (QoS). The QoS depends on the type of application and its requirements. Therefore the evaluation of features was extended based on QoS. QoS was represented by Mean Squared Error (MSE) of reconstructed 3D point positions and orientations, reliability and computation time of the feature extractors. As explained earlier, the penalties correspond to the success rate of the feature extractor over several samples on all baseline ranges. Therefore, the penalties are used to represent the "Reliability" of the feature extractor, which shows the probability of success over 450 samples. The ground-truth test data is represented as three unit vectors originating from the geometric center of the model. The positional and rotational changes in the 3D reconstructed points are computed as the deviations from the ground-truth 3D points. This gives an idea of how the reconstructed 3D structure would be transformed in 3D space, due to the errors in feature based calibration, i.e., camera pose estimation.

The table 4.2 provides an overview of statistics of MSE of 3D points. As the results for very low baseline (5°) was sensitive to noise, baseline ranges 5°-50° was considered. These ranges were further divided into three categories of baseline ranges - *Small* (5°-20°), *Medium* (20°-35°) and *Large* (35°-50°), which represents around 30% of scene coverage area, so that it can be analyzed with higher baseline resolution. The MSE is expressed in the 3D model units (mu) for positional deviation (P_d) and in degrees (deg) for rotational deviation (R_d). The table also specifies the computation time (T) required by the feature extractors (FEs), which is useful for real-time applications. The reliability parameter (Re) in the table 4.2 becomes very important along with accuracy and computation time, in making a choice of feature extractor.

Baseline	**($5^\circ - 20^\circ$)**		**($20^\circ - 35^\circ$)**		**($35^\circ - 50^\circ$)**			
FEs	R_d [deg]	P_d [mu]	R_d [deg]	P_d [mu]	R_d [deg]	P_d [mu]	T [sec]	Re [%]
	mean(deviance)		mean(deviance)		mean(deviance)			
IDEAL	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00	99.11

SIFT	13.09 (7.17)	8.23 (1.96)	2.14 (1.06)	2.83 (2.64)	2.64 (0.56)	4.22 (1.11)	17.34	80.22
SURF	15.58 (5.94)	12.04 (2.26)	5.59 (0.64)	6.27 (0.30)	3.63 (0.89)	5.33 (0.80)	5.47	79.56
BRISK	20.21 (8.94)	25.31 (13.48)	6.43 (2.41)	18.73 (13.03)	3.82 (0.52)	88.69 (141.00)	1.75	67.56
ORB	21.04 (9.31)	9.41 (0.47)	8.29 (1.54)	33.30 (41.34)	3.93 (0.05)	9.22 (6.91)	0.85	61.11
KAZE	12.12 (3.84)	7.76 (2.23)	4.78 (1.25)	6.34 (1.64)	2.92 (0.26)	7.27 (2.13)	27.67	83.56
AKAZE	11.68 (2.76)	6.91 (3.97)	7.51 (0.74)	12.36 (5.03)	4.61 (1.22)	14.48 (13.06)	4.96	78.00
MSER-SURF	19.95 (11.36)	261.94 (422.71)	8.12 (1.33)	20.55 (10.17)	5.10 (0.19)	16.09 (12.27)	7.55	59.78
STAR-SURF	29.67 (12.08)	53.51 (47.55)	11.34 (3.75)	16.08 (7.11)	7.08 (0.98)	22.15 (8.20)	0.75	45.33
MSER-BRIEF	24.01 (4.87)	44.64 (45.12)	7.03 (1.17)	7.86 (3.64)	6.25 (1.77)	8.88 (5.47)	2.50	54.22
STAR-BRIEF	21.30 (11.42)	154.87 (254.05)	6.52 (0.22)	7.58 (1.61)	4.99 (0.89)	6.18 (3.66)	0.65	66.67
FAST-BRIEF	18.00 (9.78)	24.70 (9.83)	7.37 (1.36)	8.65 (2.04)	5.12 (1.48)	9.53 (4.65)	4.73	73.78
SIFT-BRIEF	14.00 (2.80)	33.85 (27.00)	4.98 (1.65)	7.82 (7.56)	4.81 (0.49)	7.68 (0.54)	7.75	74.44
SURF-BRIEF	18.70 (7.20)	16.40 (3.54)	8.93 (1.07)	270.54 (446.17)	5.63 (1.25)	11.04 (3.82)	3.22	72.44
BRISK-BRIEF	20.10 (8.38)	21.75 (17.02)	6.46 (1.57)	22.14 (25.46)	4.91 (1.00)	49.14 (46.38)	3.76	72.44
ORB-BRIEF	15.70 (10.42)	4.79 (1.70)	4.84 (0.42)	7.61 (4.39)	4.41 (0.46)	59.16 (80.32)	0.80	71.11

KAZE-BRIEF	13.63 (7.43)	38.77 (47.78)	4.33 (0.43)	4.64 (0.34)	4.18 (0.96)	16.78 (10.96)	21.12	77.78
AKAZE-BRIEF	12.86 (6.91)	10.07 (6.27)	5.37 (2.04)	16.56 (10.51)	4.45 (0.20)	8.52 (1.14)	4.48	81.11
MSER-FREAK	61.67 (4.78)	60.57 (48.24)	15.67 (9.97)	2.77 (2.04)	8.72 (3.38)	11.38 (18.21)	7.29	6.00
STAR-FREAK	52.15 (29.05)	9.95 (6.09)	15.82 (6.92)	1.49 (0.29)	4.43 (0.00)	0.74 (0.11)	1.13	7.11
FAST-FREAK	52.70 (21.19)	8.42 (9.06)	9.50 (3.29)	23.62 (39.51)	6.38 (3.73)	11.14 (17.77)	6.09	0.00
SIFT-FREAK	51.65 (11.97)	5.74 (5.23)	22.41 (14.61)	30.14 (50.25)	10.78 (0.00)	3.78 (0.00)	9.29	5.11
SURF-FREAK	20.10 (8.38)	21.75 (17.02)	6.46 (1.57)	22.14 (25.46)	4.91 (1.00)	49.14 (46.38)	3.23	72.44
BRISK-FREAK	15.70 (10.42)	4.79 (1.70)	4.84 (0.42)	7.61 (4.39)	4.41 (0.46)	59.16 (80.32)	1.21	71.11
ORB-FREAK	13.63 (7.43)	38.77 (47.78)	4.33 (0.43)	4.64 (0.34)	4.18 (0.96)	16.78 (10.96)	21.10	77.78
KAZE-FREAK	12.86 (6.91)	10.07 (6.27)	5.37 (2.04)	16.56 (10.51)	4.45 (0.20)	8.52 (1.14)	7.88	81.11
AKAZE-FREAK	53.24 (25.05)	28.34 (37.48)	14.56 (8.09)	4.46 (5.23)	6.85 (0.15)	5.16 (5.38)	9.13	9.33

Table 4.2: Quality - accuracy, reliability and execution time of 24 feature extractors, which provides practical recommendation for 3D applications(section 4.3.7). Here "Rotation" is the mean 3D rotational change (expressed in degrees) and "Position" is the mean 3D positional shift (expressed in model units) of all the estimation 3D unit vectors that represent a model in 3D space.

Practical Recommendations

The result shown in table 4.2 is useful for any 3D application, which uses markerless camera pose estimation, i.e., FBC. Some applications demand real-time performance. The camera

placements vary from small to large baseline range. The table 4.2 can be used as a recommendation for practical 3D applications, where one can either choose feature extractors or estimate the camera density around the object of interest, based on the desired QoS.

Scenario I An application scenario using *Small* baseline range for which a feature extractor is required to be chosen: From the table, both KAZE and AKAZE have good accuracy in terms of 3D position and rotation, but one may choose AKAZE if the application demands fast computation time. However, this choice is at the cost of reliability, because KAZE seems to be more reliable than AKAZE. On the other hand, AKAZE+BRIEF offers accuracy similar to KAZE and is equally reliable, and also much faster than KAZE. So, in this case, the application could choose AKAZE+BRIEF.

Scenario II Another application, where number of cameras around an object needs to be determined using KAZE (assuming KAZE is chosen for its high reliability): Here, KAZE offers the best positional accuracy at *Medium* baseline range. If a baseline of about 30° is considered, then number of cameras required to capture an object in 360°, is about 12. On the other hand, if one can compromise on the positional accuracy slightly, at the same time gain higher rotational accuracy, one would choose to operate with KAZE at *Large* baseline range. In this case, for a baseline of about 45°, one could capture the same object with only 8 cameras, which is more effective for applications, in terms of calibration, storage and transmission tasks.

Hence, the study of feature extractors and their evaluation based on various baselines for 3D error in terms of position and orientation will very helpful for such applications. Experiments were carried out on a virtual dataset that mimicked the application scenarios such as VERDIONE and BAGADUS, in order to obtain ground-truth values for determining the 3D accuracy. However, the application scenario testing was limited to only a foreground model object. The situation where the captured scene involves background with a large depth of field, is more likely to occur in the application scenario such as VERDIONE, BAGADUS and POPART. Such a situation was not tested in this study.

The recommendations for feature extractors were based on the object distance of about 300 model units from the camera. Clearly if the object is farther away from the camera, every pixel covers more area in 3D space. Consequently, feature detection errors or pose estimation errors increase and thereby leads to less accurate 3D reconstruction. In such cases a re-calibration process with more number of features covering the entire image might be necessary.

All the above evaluation tests are evident enough to reject a part of the *Null Hypothesis III* (stated in section 1.4), and hence, it can be concluded that the performance of state-of-the-art feature extractors have significant differences for the change in camera extrinsic parameters represented as camera baseline. Several state-of-the-art feature extractors were evaluated and the performance was measured in terms of 3D accuracy (normalized correlation coefficient and mean squared error), computational time and reliability. This study provides a recommendation for 3D application designer that will enable them to:

1. Select the feature extractor based on an acceptable accuracy or an acceptable execution time, with a cost of reliability.

2. Decide the camera density required to capture an object of interest, for a desired QoS.

4.4 FBC using SIFT for Wide Baseline

In the previous section, the quality of feature extractors for a variation in camera baseline was explored. In this section, the focus is specifically on using SIFT feature extractor for feature based calibration in the application scenarios such as VERDIONE (detail sin section 2.2.1) and BAGADUS (details in section 2.2.2).

Several types of camera arrays are in practical use and development today (Wilburn, Joshi, Vaish, Levoy, and Horowitz, 2004; Zhang and Chen, 2004). The camera arrays differ in camera density and physical space it covers. While some image processing techniques such as light-field processing, stereoscopic and multiview video, require relatively dense camera placement, other image processing applications such as free-viewpoint rendering (Min, Kim, Yun, and Sohn, 2009), visual hull reconstruction (Matusik, Buehler, Raskar, Gortler, and McMillan, 2000) and tracking or geometrical scene reconstruction can deal with relatively sparse camera placements. In large space application scenarios such as VERDIONE or BAGADUS, it is necessary to distribute cameras around a large space to capture the entire volume of the scene from an optimal number of viewpoints. If camera array is placed at wide *baselines* (angular displacement between camera centers of the stereo camera pair), then less density of cameras will be required to capture the large space, which is absolutely suitable for cost effectiveness.

It is known that an important necessity for such applications is camera calibration. However traditional calibration techniques (Bouguet, 2008; Tsai, 1992; Zhang, 2000) cannot be used, because it is sometimes impossible to place a big sized checkerboard in the stadium (BAGADUS scenario) or the stage (VERDIONE scenario). Using markers for marker based calibration (Kurillo, Li, and Bajcsy, 2008) causes inconvenience and disturbs the scene setting. Therefore, feature based calibration (FBC) can replace the traditional checkerboard calibration in such scenarios. Again, FBC is the most important aspect defining the quality of 3D reconstruction.

SIFT has been a a very popular feature extractor which is known for scale and rotational invariance. However, according to the SIFT users, when the stereo cameras had a camera baseline more than 30°, the accuracy of calibration system on the whole degraded. The repeatability of SIFT detection started reducing after view angle of around 30° (Lowe, 2004). A similar trend was seen in the results of the previous section in figure 4.20(a). Hence, it was important to explore if the accuracy of SIFT based FBC can be maintained at an acceptable level when the camera view point is more than 30 degrees, with a motivation to decrease the number of cameras to capture the large space.

SIFT based FBC was proposed (Li and Lu, 2010; Liu, Zhang, Liu, Xia, and Hu, 2009; Yun and Park, 2006) as an improvement over the traditional CBC that uses a calibration target. Using SIFT, these systems automatically match the features between camera images, which are then used to perform the calibration. However, the 30° view-angle limitation on performance of SIFT still persisted, where feature matching performance degraded with an increase in viewing angle between two perspectives. Moreover, with a growing *baseline*, less similarities exist between images and consequently, fewer SIFT features were matched. The accuracy differences might not only manifest due to less overlap areas or an increased number of occlusions, it could also result in more false positive SIFT matches.

In earlier works (Li and Lu, 2010; Liu, Zhang, Liu, Xia, and Hu, 2009; Yun and Park, 2006), all the point correspondences obtained by SIFT feature matching have been used for calibration.

This is redundant and prone to noise due to mismatches of SIFT features. Eliminating such wrong matches was studied (Jiayuan, Yigang, and Yun, 2010), using a error canceling algorithm based on RANSAC (Random Sample Consensus - a widely used algorithm for outlier removal) (Fischler and Bolles, 1981). Alternatively, a simpler method based on the geometry of lines joining the matched points was proposed. The proposed outlier removal process executes faster than and performs as good as the RANSAC (Fischler and Bolles, 1981), in the test scenario.

Therefore, in this exploration, an investigation was carried out to find out how to minimize the false positive feature matches of SIFT and arrive at an optimal way of feature match selection for FBC. Consequently, a new algorithm was proposed.

4.4.1 Proposed Algorithm

In our paper titled "Faster and More Accurate Feature-Based Calibration for Widely Spaced Camera Pairs" [details in chapter 7], the operation of state-of-the-art SIFT was extended and proposed a new extrinsic SIFT feature based calibration method called $newSIFTcalib$.

The highlights of the proposed algorithm are as follows:

- The $newSIFTcalib$ algorithm is a SIFT based FBC algorithm proposed for a camera pair with an arbitrary baseline that works without a calibration target. Here, some of the limitations of current SIFT-based methods were addressed.
- Specifically, the novelty of the method lies in,
  - a new technique for the detection and removal of wrong SIFT matches.
  - a method for selecting a small subset of all detected SIFT features.
- The $newSIFTcalib$ particularly compensates for increased viewing angles or large baselines, making SIFT-based calibration usable for camera arrays with large baselines.

This work involved following:

- Proposal for a new algorithm based on original SIFT features extractor to maintain the quality of SIFT-based FBC for wide baseline setup.
- Evaluation of multiview real video dataset in comparison to other related algorithms.
- Performance metrics for both accuracy and execution time.
- Theoretical and practical operational limits of the algorithm, computed for allowable baseline and an acceptable accuracy.

The proposed FBC system is illustrated in figure 4.24, where a number of stereo camera pairs capture a scene of interest. For every 2D stereo images, SIFT feature points were detected in stereo images and matched them using brute-force method. The feature matches were used to estimate camera pose, with known camera intrinsic. The intrinsic camera parameters are assumed to be known or have been determined in a prior offline calibration step.

As a pre-processing step, outliers (false positives in the matching process) are detected and removed. Only a subset of stable points (referred as $FeatureVector$ in rest of the section), which are less prone to noise, were used for camera pose estimation. The cameras are assumed to be pre-calibrated for camera intrinsic.

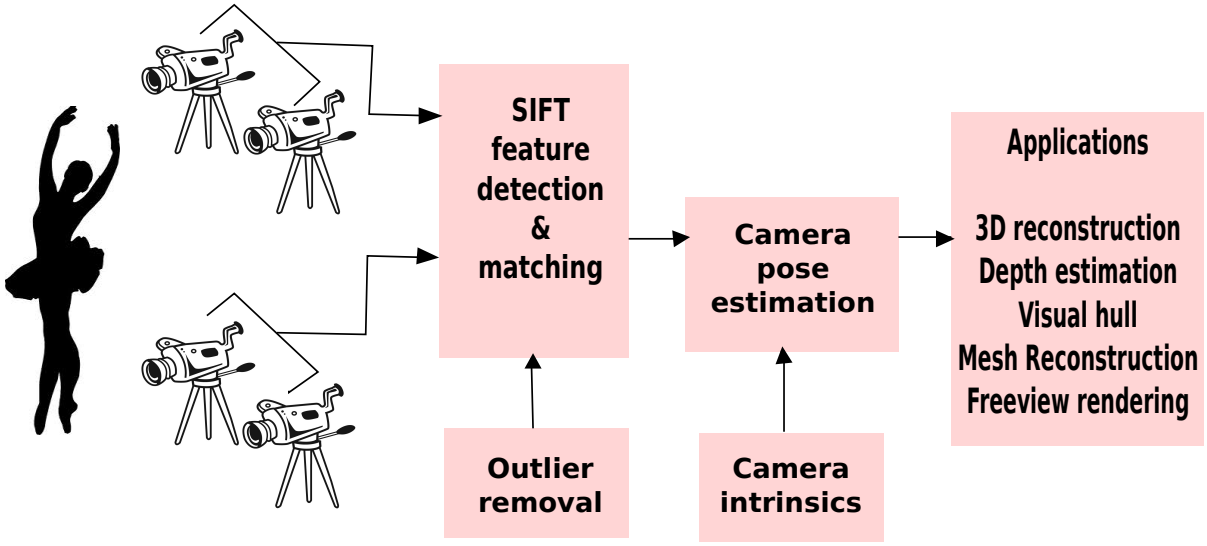


Figure 4.24: System overview.

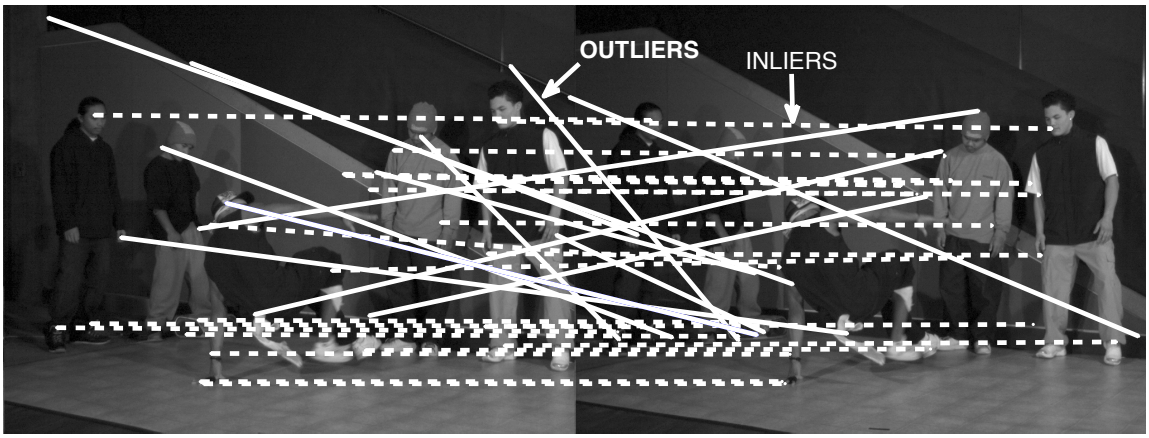


Figure 4.25: Process of outlier detection: outliers (solid), inliers (dotted)

Proposed Outlier Detection and Removal

The angular deviation of the lines connecting corresponding points in the camera pair was computed. The detection of the outliers were based on the statistics of the angular deviation. Consider two images from stereo cameras placed horizontally apart from each other. Lines are drawn from every feature point in image 1 to their respective correspondences in image 2, as in figure 4.25.

The mean (μ_θ^x) and standard deviation (σ_θ^x) of the angle between all lines and the X-axis was computed. Then, the outlier detector compared the angle between each line and the x-axis to μ_θ^x and σ_θ^x. A line l_{ij} (and thereby the point correspondence) is identified as an outlier if the angle θ_l^x differs by more than σ_θ^x, as in equation (4.7). The same is possible to be carried out based on the statistics of angle between all lines and the Y axis. So in this way, it is made sure that this algorithm can be used on images taken from both horizontally and vertically aligned cameras. In general, this algorithm can be used on any stereo pair based on the statistics (mean

and standard deviation) computed for the angular displacement of lines joining the matched feature points with respect to X and Y axes.

$$outlier = \begin{cases} l_{ij} & \text{if} \quad |\theta_l^{x/y}| > \mu_\theta^{x/y} + \sigma_\theta^{x/y} \\ 0 & \text{if} \quad |\theta_l^{x/y}| < \mu_\theta^{x/y} + \sigma_\theta^{x/y} \end{cases} \tag{4.7}$$

FeatureVector - size and selection

The feature points detected by SIFT are normally assigned a scale which can be interpreted as a representation of the stability of the feature detection. This property was exploited and the inlier point correspondences was sorted out and defined as $FeatureVector$, a vector consisting of point correspondences. The size of the vector was determined empirically.

Next, from the pool of inlier point correspondences, five candidates of subsets from highest order of stability were chosen. Every candidate vector was used to estimate the camera pose. With this information, estimated 3D points were re-projected onto the 2D image and compared with the candidate vectors. This measure is re-projection error. Out of these five candidates, the best subset was chosen as the $FeatureVector$ based on the least re-projection error.

Camera Pose Estimation

The $FeatureVector$ of point correspondences is used to estimate the essential matrix E. At the time of this experiment, normalized 8-point algorithm (Hartley, 1997) was considered to estimate the fundamental matrix and thereby estimate essential matrix using camera intrinsic matrix. In a stereo camera setup, if the world coordinates are considered to be at the center of the reference camera, the rotation matrix of reference camera is an identity matrix and translation is a zero matrix. Relative rotation R and translation t of the second camera of the camera pair represents the camera pose, and are related to essential matrix as $E = [t]_X R$, where $[t]_X$ is a skew-symmetric matrix as in equation 4.8,

$$[t]_X = \begin{bmatrix} 0 & t_x & -t_z \\ -t_x & 0 & t_y \\ t_z & -t_y & 0 \end{bmatrix} \tag{4.8}$$

The Essential matrix can be decomposed using Singular Value Decomposition (SVD) as in (Hartley, 1992), which is detailed as follows:

Let K_1 and K_2 be the intrinsic parameters of the camera pair respectively. Upon SVD of E, the equation 4.9 is obtained:

$$E = USV^T \tag{4.9}$$

where U and V are unitary matrices and S is a rectangular diagonal matrix. Accordingly, R has two solutions R_a, R_b, and t has two solution t_a, t_b, which are given by equation 4.10.

$$R_a = UWV^T, R_b = UW^TV^T, t_a = +u3, t_b = -u3, \tag{4.10}$$

where u3 is the 3rd column of matrix U and W is as follows:

$$W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

This gives a choice of four solutions to obtain the camera pose. A projection matrix of the reference camera is given as $P_1 = K_1[I|0]$. If $P_2 = K_2[R|t]$ is the projection matrix of the camera, then solution is one of the following:

$$P_2 = K_2[R_a|t_a] \text{ , } K_2[R_a|t_b] \text{ , } K_2[R_b|t_a] \text{ , } K_2[R_b|t_b]$$

The above four solutions have a geometrical meaning and one of the solution is always meaningful. For every possible solution of P_2, 3D points corresponding to the intersection of back-projected ray from 2D point correspondences are estimated through triangulation. Using cheirality constraint (Hartley and Zisserman, 2004), the 3D points obtained are checked for positive sign of depth and hence the solution for camera pose is determined.

4.4.2 Evaluation

A widely used multi-view dataset from Microsoft (Zitnick, Kang, Uyttendaele, Winder, and Szeliski, 2004) was used for evaluating the proposed algorithm with others. The dataset was captured by a setup as illustrated in the figure 4.26. All 8 cameras (separated by $\approx$ 0.3 meters distance) captured an event (taken place at $\approx$ 4.6 meters) with a resolution of 1024x728, and rate of 15fps. The calibration parameters for these cameras were computed using traditional approach (checkerboard). These known calibration parameters were used for comparing estimated calibration parameters estimated using $newSIFTcalib$ algorithm. Based on the object distance and camera separation, the angular difference between consecutive cameras in the Microsoft dataset is about $3.82°$. Since there are 8 cameras, total angular difference between first and the last cameras is about $26.74°$. So, this baseline range is very similar to the angular difference of the stereo cameras in VERDIONE or BAGADUS scenarios, although the object distances are different. Hence Microsoft data can be considered for those scenarios.

The following are the performance metrics used to evaluate the algorithms.

- *Epipolar Error (E_p):* This is computed as in equation 4.5.
- *Re-projection Error (R_p):* Given the point correspondences $\{x_1, x_2\}$ and the estimates for projection matrices $P1$, $P2$ for two cameras respectively, if the estimated 3D points are re-projected onto the 2D image plane - referred to as new point correspondences $\{\widetilde{x}_1, \widetilde{x}_2\}$ ($\widetilde{x}_1 = P1\hat{X}$, $\widetilde{x}_2 = P2\hat{X}$) then, R_p averaged over N test samples, can be computed as,

$$R_p = \frac{1}{N}\sum_{i=1}^{N}[d(x'_{1i}, \widetilde{x}'_{1i}) + d(x'_{2i}, \widetilde{x}'_{2i})] \tag{4.11}$$

$$d(x', \widetilde{x}') = ||(x' - \widetilde{x}')||_2 \tag{4.12}$$

[3]Microsoft dataset (Zitnick, Kang, Uyttendaele, Winder, and Szeliski, 2004).

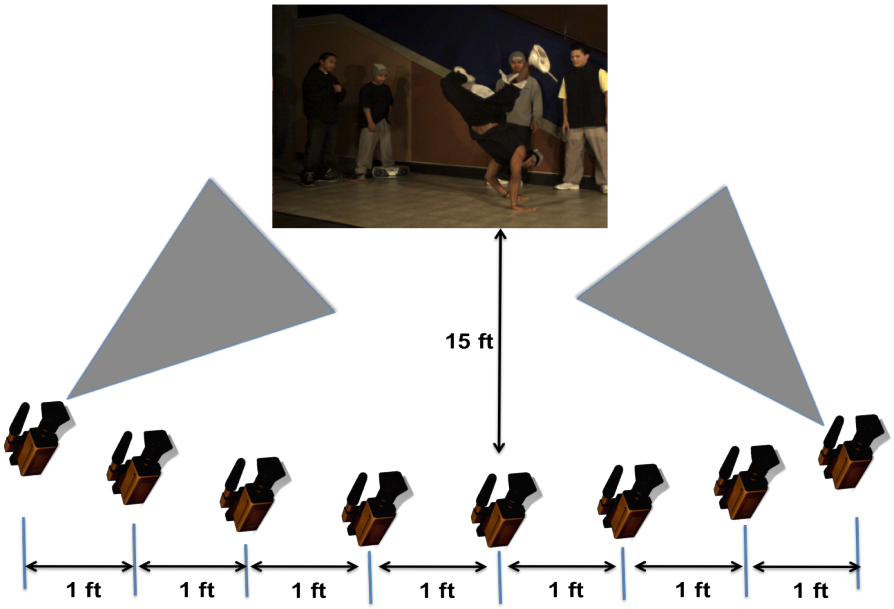


Figure 4.26: Illustration of setup used by Microsoft to produce the multiview dataset. Courtesy: Microsoft[3].

4.4.3 Performance of Proposed Algorithm

Firstly, the proposed outlier removal process was tested for its performance. The performance in this case was measured using Epipolar constraint. Secondly, the proposed algorithm was tested as a whole, which involves the $FeatureVector$ selection. The known and estimated camera pose error was computed and the execution time of the algorithm was measured in comparison to the other algorithms.

Testing the accuracy of outlier removal: The following algorithms were compared to evaluate the proposed outlier removal (solid lines in figure 4.25 are the outliers). *Epipolar Error* (as in equation 4.5) was computed for following methods:

- 8-pt algorithm without outlier detection.
- 8-pt algorithm with RANSAC.
- 8-pt algorithm with the proposed outlier removal.

The test results in figure 4.27 showed the RANSAC method performed better than 8-point algorithm without outlier removal, as expected. It is very evident that the proposed outlier removal performed as good as RANSAC, and the computation time was drastically reduced because RANSAC requires a large number of points for estimation. The minimum number of points in the *FeatureVector* required for the good performance of outlier removal was deduced. As shown in the figure 4.27, the proposed outlier removal performed similar to the popular RANSAC method at a minimum of 25 feature points. Therefore, the size of the *FeatureVector* was chosen to be 25 points. However, the proposed outlier detector performance was tested only with relative rotation around vertical axis.

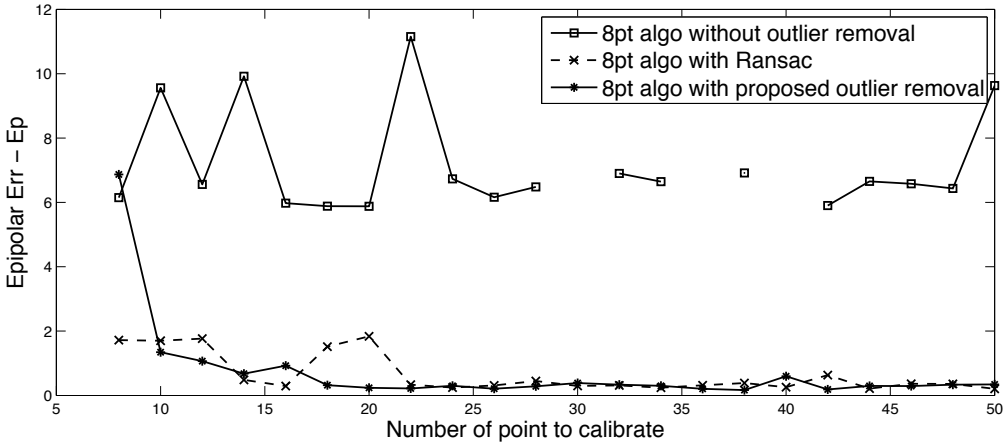


Figure 4.27: Epipolar error (E_p) computed for three different methods

Testing the accuracy of the proposed algorithm: The performance of the proposed algorithm was compared with other existing ones. The test candidates / algorithms were:

- The *Checkerboard* algorithm represents calibration using corners detected on the checkerboard.
- The *FullSift_RANSAC* algorithm represents calibration based on SIFT, using all the feature points detected and outliers removed by RANSAC.
- The *FullSift* algorithm represents calibration based on SIFT, using all the feature points detected and outliers removed by the proposed method.
- The *newSIFTcalib* / *Proposed* algorithm - represents the proposed algorithm for calibration based on SIFT, using the proposed outlier removal method and selection of stable subset (*FeatureVector*) of feature points.

The same dataset of images was used to test all algorithms for fair comparison. To evaluate the accuracy of calibration, *Re-projection Error* (R_p) was considered the performance metric. The calibration parameters for the *Checkerboard* algorithm were given by the dataset. Re-projection error was thus computed using the known calibration parameters. For the other algorithms, the camera parameters were estimated and then used them for computing the re-projection error. Usually, in 3D vision applications, subpixel accuracy is considered good quality, and therefore, $R_p \leq 1$ was chosen as an acceptable re-projection error.

The test results are as shown in figure 4.28 for R_p against various baseline distances (in meters) between neighboring cameras. Here, the baseline of maximum 2.1 meters is recorded, and this limit is as per the dataset availability. The summary of the results are as follows:

1. *FullSift_RANSAC* and *FullSift* performed very similarly. This verifies, as in the previous test, that the proposed outlier removal algorithm used in *FullSift* was as good as the RANSAC method for outlier removal while being faster.
2. At small baselines ($\approx$ 0 - 1.2 meters), the *newSIFTcalib* algorithm performed as good as other algorithms in the test, with minimal but acceptable error level of $R_p \leq 1$.

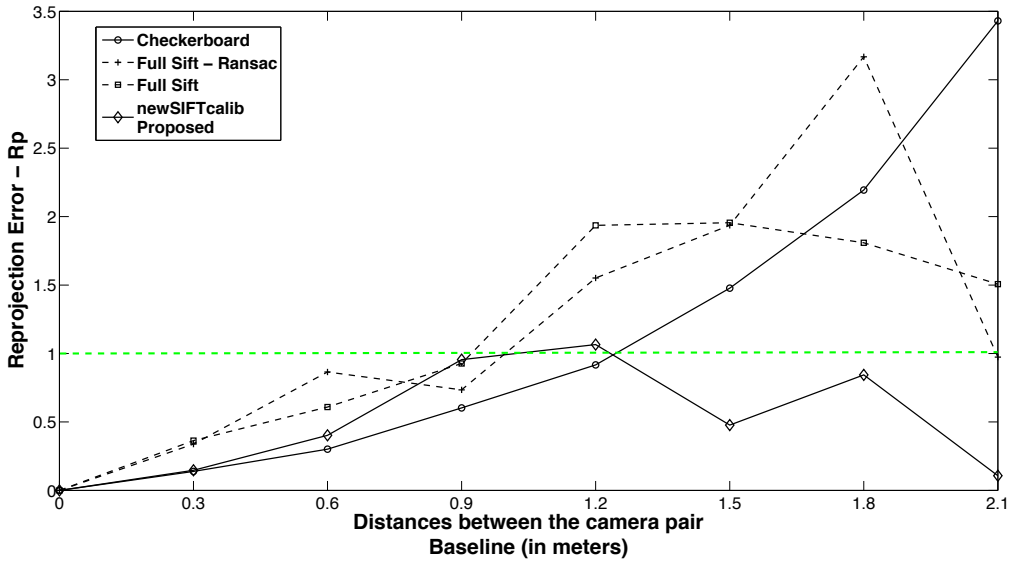


Figure 4.28: Re-projection error (R_p) computed for different algorithms

3. At large baselines ($\approx$ 1.2 - 2.1 meters), the *newSIFTcalib* outperformed all other algorithms. The performance of the other algorithms degraded because of the noise prone feature points, which was introduced due to large view-angles and baselines. On the other hand, the *newSIFTcalib* algorithm uses the *FeatureVector*, which are more stable and less prone to noise. The *newSIFTcalib* algorithm performs with high consistency at sub-pixel level and is robust to noise.

4. The error in *Checkerboard* algorithm is observed to increase rapidly as the baseline increases. This is because 2D SIFT feature points that were used for testing corresponds to non planar 3D points distributed in the scene, but this algorithm is optimized for checkerboard corner points lying on a plane. So, as the baseline increases, the effect of poorly estimated camera pose increases to produce an increase in the reprojection error.

Testing the camera pose estimation of proposed algorithm: The estimated camera pose parameters were compared in terms of rotation angles (θ, ϕ, ψ) in 3-dimension, in comparison to the given rotation angles between cameras. Table 4.3 shows the parameters known (*Checkerboard*) and parameters estimated (*newSIFTcalib*) for different baseline distances are very close to each other. This signifies that a good camera pose estimation was achieved, which is very important for 3D applications.

Testing the execution time of the proposed algorithm: The execution time was evaluated. The camera pose estimation using different algorithms for cameras separated by 1.2 meters was executed and the elapsed time was measured in seconds. The performance of the *newSIFTcalib* was reasonably measured relative to other algorithms. Figure 4.29 shows that the *newSIFTcalib* algorithm achieved 58.82% and 74.07% decrease in the execution time compared to the *FullSift* and the *FullSift_RANSAC*. One important thing to note is, at camera baseline of about 1.2 meters, the quality of *newSIFTcalib* was comparable to other algorithms (as in figure 4.28), while the execution time of *newSIFTcalib* drastically reduced.

Camera pair	Rotation		
Baseline	θ	ϕ	ψ
0.3 (known)	3.1624	-3.1100	-3.1353
0.3 (estimate)	3.1253	-3.0839	-3.1362
1.2 (known)	3.1547	-3.1015	-3.1271
1.2 (estimate)	3.1278	-2.8736	-3.1355

Table 4.3: Comparing known and estimated camera rotational parameters.

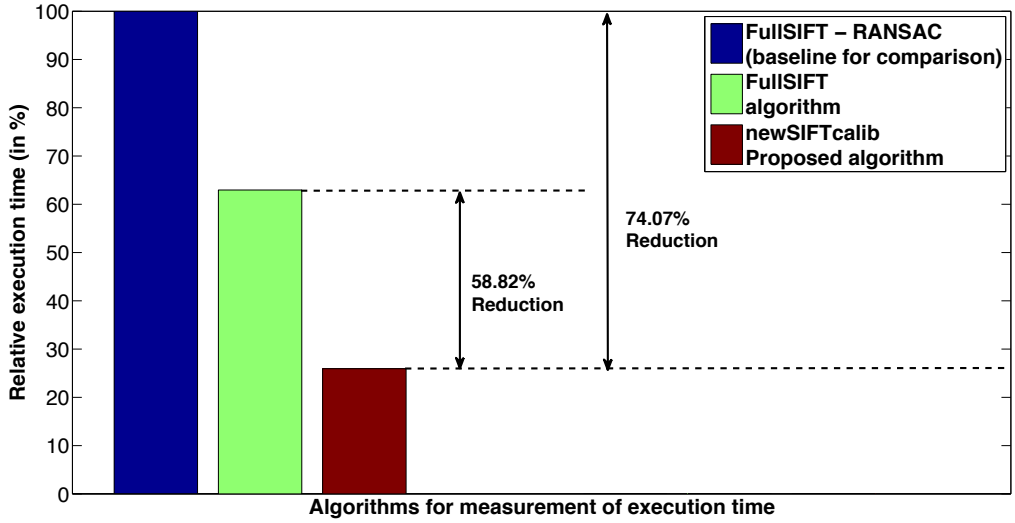


Figure 4.29: Execution time of various algorithms relative to $FullSIFT - RANSAC$.

Moreover, the execution time of SIFT should also be considered. The SIFT was executed using hardware-accelerated devices like graphics processing units (GPUs), and found that the feature detection and matching of two images takes about 700 milliseconds. The major contribution to the execution time for the proposed algorithm is from the SIFT feature detection and matching compared to the camera pose estimation process. In this sense, one could achieve closer to real-time performances, along with time-optimized SIFT implementation.

4.4.4 Discussions

It is known that for SIFT users, feature detection on images from cameras, whose view-angle differences are more than 30°, introduced matching errors and thereby degraded the accuracy of calibration system on the whole.

This condition was used, $\theta \leq 30^{\circ}$ as a constraint to develop a relationship between object distance (D) and distance between the cameras i.e., the baseline (B), based on which the performance of the algorithms evaluated and their operational limits were estimated.

Theoretical Limit

Consider figure 4.30, where D represents the object distance from the camera, B and θ represents the baseline distance and view angle between neighboring cameras, respectively. The triangle equations θ can be expressed as:

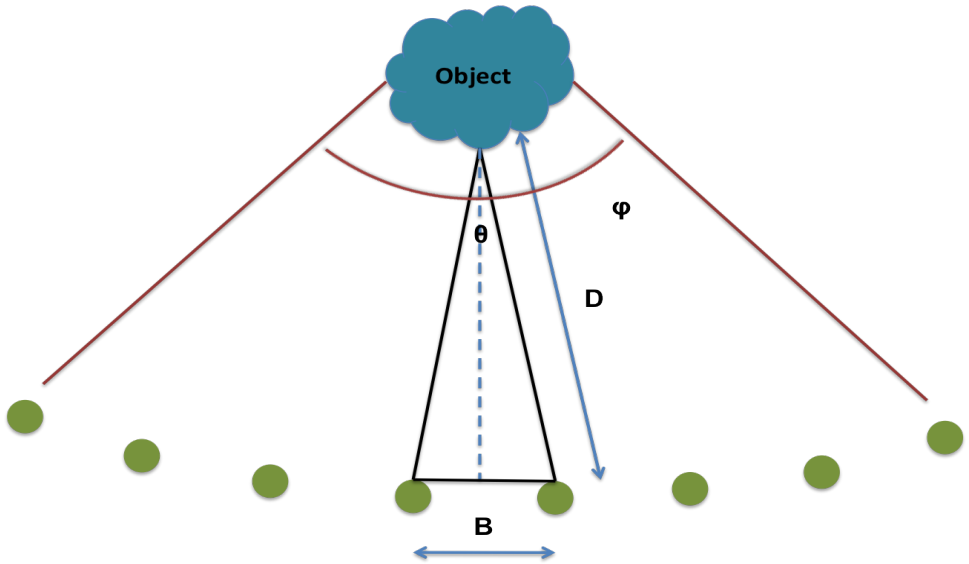


Figure 4.30: Deduction of relationship between object distance (D) and the baseline distance between the cameras (B).

$$\theta = 2 * sin^{-1}(\frac{B}{2D}) \tag{4.13}$$

Using the condition, $\theta \leq 30°$, the following was obtained:

$$2sin^{-1}(\frac{B}{2D}) \leq 30° \Rightarrow \frac{B}{2D} \leq sin(15°) \Rightarrow B \leq 0.52D \tag{4.14}$$

The relation $B \leq 0.52D$ is the theoretically defined limit for the baseline using the constraint $\theta \leq 30°$. In the test dataset, the object distance, $D = 4.6meters(15feet)$, and therefore, the theoretically set limit for baseline, deduced by equation 4.14 would then be and maximum of 2.4 meters.

Next, the practical limit for the algorithms was checked on the given dataset.

Practical Limit

From results as shown in figure 4.28, the existing algorithms performed with an acceptable error ($R_p \leq 1$) only up to a baseline separation of 1 meter. Although the theoretical limit for the baseline is up to 2.4 meters, the existing algorithms practically performed well only up to 1 meter. Hence, one could say that the existing algorithms are well suited for smaller baselines.

On the other hand, the *newSIFTcalib* algorithm extends the practical limit for the baseline up to 2.1 meters and is well suited for large baselines. The dataset used in the experiments contain stereo images separated by a maximum distance of 2.1 meters. Due to this limitation, the

newSIFTcalib algorithm was not tested for wider baselines, however, it might fail to maintain an acceptable performance, i.e., $R_p < 1$. This is merely due to the limitations posed by the SIFT feature detection for variance in view angle.

However, it is evident that the *newSIFTcalib* algorithm pushes the practical limit of the existing algorithms and reaches very close to the theoretical limit. Overall, the accuracy of the *newSIFTcalib* algorithm has been consistent at sub-pixel level over multiple baselines, while outperforming the existing algorithms, especially at large baselines. The execution time of the *newSIFTcalib* algorithm has shown a drastic reduction in comparison to other stated algorithms.

The above test results have provided sufficient evidence to reject *Null Hypothesis IV* (as stated in section 1.4). Hence, it can be definitely concluded that SIFT based FBC with few modifications like in the *newSIFTcalib*, can achieve better accuracy and faster execution time comparatively, at wide baselines of upto around 2 meters. This enhances the usability and scalability of multiview capture system especially in large volume spaces.

4.5 Conclusions for Feature Extraction

In this chapter, research *Hypotheses III* and *IV*, as stated in section 1.4, were tested. In the process of testing the research hypotheses, for each of the topics discussed, this chapter stated the proposed ideas, explained the experiments setup and discussed the results.

Accordingly, the effects of change in the internal (for the VERDIONE and BAGADUS scenarios) and the external (for the POPART scenario) camera properties on the performance of state-of-the-art feature extractors were studied and hence, it was shown that the performances of the state-of-the-art feature extractors have significant differences to the change in intrinsic and extrinsic camera parameters. The study of SIFT based FBC (for the VERDIONE and BAGADUS scenarios) was also extended, especially for wide baseline camera setup and showed that the accuracy of SIFT features for wide baseline FBC is maintained at an acceptable level for more than 30 degrees angular separation between stereo cameras.

Overall, in a practical perspective, the contributions from the study in this chapter is summarized as follows:

1. The prominent state-of-the-art feature extractors such as SIFT, SURF and ORB were evaluated against practical issues of camera internal properties, i.e., de-focus, lens distortion and noise. As a consequence, operating ranges of the feature extractors were determined for a certain perturbations limit. These results can be used by researchers or system developers for making design decisions for 3D multiple camera systems which are prone to change in internal camera properties, especially in the context of the VERDIONE and BAGADUS application scenarios.

2. About 26 combinations of various state-of-the-art feature detectors and descriptors were evaluated for varying camera baseline. Consequently, a design recommendation was made to help practical 3D application designers to make a choice of the feature extractor based on the accuracy, deformation of 3D object, execution time and reliability factor. The application designers can also use the result in order to determine the camera density required to capture the scene. The evaluation was carried out using virtual dataset

that mimicked the VERDIONE, BAGADUS or POPART scenarios, however, application scenario testing was limited.

3. A new algorithm was proposed, which modified the existing SIFT extractor for FBC methodology and a much better accuracy and a faster execution time was achieved, especially in wide baseline camera setups. This naturally increased the usability and scalability of 3D multiview capture systems. Especially for large volume scenarios like the VERDIONE or BAGADUS, the density of cameras capturing the scene can be reduced and thereby the solution becomes cost effective.

In this chapter the accuracy of feature extraction was assessed, aiming at exploring their robustness against practical issues. However, pose estimation is another part of the FBC pipeline for 3D reconstruction. Therefore, characterization of pose estimation for various attributes of matched feature correspondences between a stereo pair is described in the next chapter.

Chapter 5

Pose Estimation

It has been explained that the feature based calibration (FBC) process comprises of feature extraction and pose estimation modules (figure 2.1). In the previous chapter, the feature extraction module and their characterization for various practical issues was discussed. However, the quality of 3D reconstruction systems not only depends on the feature extraction quality, but also on the quality of pose estimation. Hence, in this chapter, the influence of accuracy of camera pose estimation on the performance of 3D systems is explored.

In this chapter, the focus is on characterizing pose estimation part of FBC. The feature correspondences in a stereo pair are extracted by feature extraction part of FBC, and are used to estimate the camera extrinsic, i.e., relative camera translation and rotation of one camera with respect to another in a stereo camera setup. This process of estimating extrinsic camera parameters is usually termed as 'extrinsic camera calibration' or 'camera pose estimation'. In the context of FBC, the pose estimation relies on the feature correspondences between a stereo pair. Hence, it is very important to understand the sensitivity of the pose estimation module, in terms of various attributes of the matched feature correspondences. The attributes that were investigated are noise, number of feature points, sparsity of feature points in 2D space and feature points from various camera baselines.

In application scenarios such as VERDIONE (details in section 2.2.1), POPART (details in section 2.2.4) and SEMRECON (details in section 2.2.5), in order to adopt FBC, it is required to have feature correspondences extracted and matched by state-of-the-art feature extractors described in section 4.1.

Using state-of-the-art feature extractors, as in the previous chapter, the resulting feature correspondences was found to be noisy. Noise can be interpreted as follows:

- feature correspondences not adhering to epipolar constraint.
- feature correspondences, when used for camera pose estimation, yield wrong camera pose.
- feature correspondences when triangulated, yield 3D points which are geometrically dislocated or deformed.

When such noisy feature correspondences are used for FBC, the noise affects the accuracy of pose estimation and thereby the 3D reconstruction. It is interesting to explore to what extent does the noise affects the pose estimation and 3D reconstruction.

When the feature correspondences are available, the question arises as to how many points to use and if not all, what subset of feature points to use for the pose estimation. In section 4.4, it is seen that the use of subset of SIFT feature points yielded good results for wide baseline scenario. This is an indication that pose estimation is sensitive to selection of feature correspondences. Therefore, it is worth investigating the dependency of sparsity of feature points on the quality of pose estimation and thereby 3D reconstruction. Here, the sparsity of feature points is explored and a selection is made based on how sparse are the feature points in 2D space.

In section 4.3, it was shown that the camera baseline influences the accuracy of feature extraction quality. Hence, it is important to characterize pose estimation algorithm based on the various camera baselines, as well.

Therefore, to study the influence of attributes of feature point correspondences (i.e., noise, number, sparsity and baseline) on the pose estimation, the *Hypothesis V* was stated, as in section 1.4. In this chapter, the aim is to test *Hypothesis V* and in order to test the hypothesis, the sensitivity of pose estimation was explored.

5.1 Sensitivity of Pose estimation

Pose estimation algorithms are evaluated by comparing the estimated pose, i.e., camera rotation and translation, with the reference camera pose (Rodehorst, Heinrichs, and Hellwich, 2008).

Some of the other evaluations considered projecting 3D reference points to 2D image plane and compare with the reference 2D points. This is a typical measure known as *Re-projection Error*. Few evaluations carried out based on re-projection error for camera pose are made in applications such as visual odometry (Alismail, Browning, and Dias, 2011), augmented reality (Maidi, Mallem, Benchikh, and Otmane, 2013), etc. Several performance metrics such as geometric error, re-projection error, algebraic error and Sampson error were used for evaluating camera pose estimation (Brückner, Bajramovic, and Denzler, 2008) on real dataset. Contrarily, in this study, the evaluatation was carried out based on the performance metric measured in 3D space, which is more meaningful for measuring accuracy and deformation of 3D reconstruction.

In another work (Kniaz, 2016), the camera pose was evaluated based on motion trajectory of a drone. Here, the position of drone in 3D space was tracked based on pose estimation and motion capture equipment separately and was compared to evaluate the pose estimation. Interestingly, the performance measure in 3D space was appropriate for their application. Similarly, the performance in 3D space was also evaluated. However, the exploration of pose estimation was extended, in terms of robustness to various attributes of the feature correspondences.

Some of the evaluations carried out in the past (Brückner, Bajramovic, and Denzler, 2008; Rodehorst, Heinrichs, and Hellwich, 2008) have evaluated several camera pose estimation and concluded that 5-Point algorithm (Nistér, 2004) performed the best. In this study, the 5-Point algorithm was considered for pose estimation and studied its sensitivity to noise, number and sparsity of feature correspondences obtained by a virtual dataset.

Most of the evaluation on camera pose estimation have considered variation of noise in the feature correspondences. In this study, the evaluation was extended with respect to change in camera baseline to investigate if noise and camera baseline together have an influence on the quality of pose estimation.

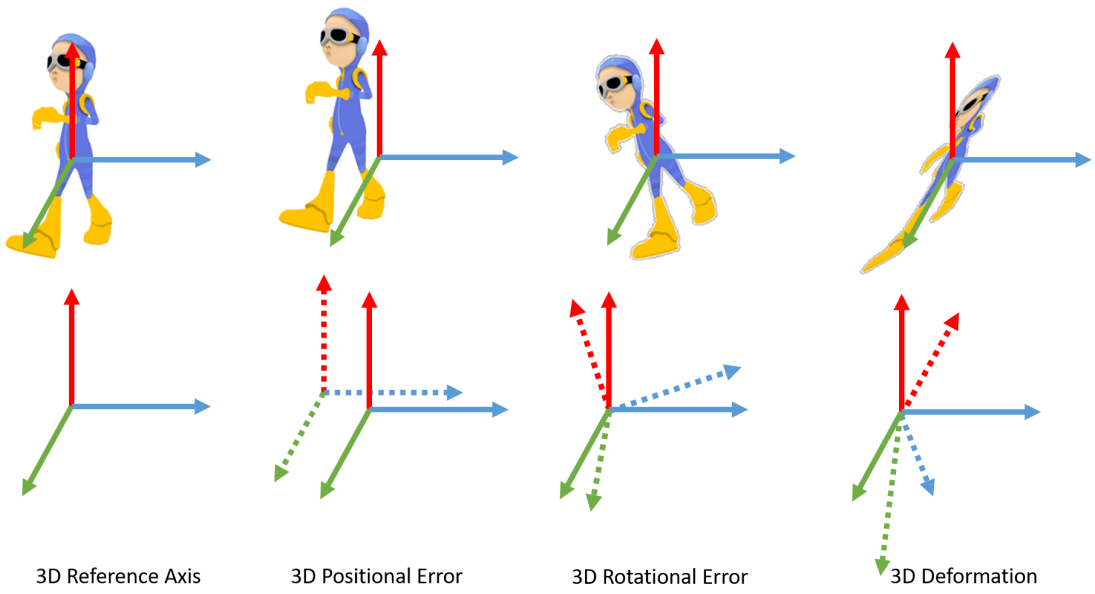


Figure 5.1: The extended 3D performance metric explained.

Most of the study mentioned above, have used real datasets for testing, but a virtual dataset was used to have a complete control on the experimental conditions by obtaining ground-truth values to perform the evaluation.

Moreover, other evaluations on the influence of sparsity of feature correspondence on quality of pose estimation, have been unknown so far.

5.1.1 Performance Metrics

A new performance metric was used, which relates to measuring quality in 3D space, i.e., 3D positional error, 3D rotational error and 3D deformation, as depicted in figure 5.1. The explanation of these metrics are as follows:

- 3D position accuracy is measured as mean squared difference between known and estimated 3D point positions.
- 3D orientation accuracy error is measured as mean squared difference between known and estimated 3D vector angles.
- 3D deformation is measured as a normalized average orthogonality. Orthogonality measures whether the angular difference between any two vectors of the test data is at right angle to each other.

Additionally, in comparison to 3D errors, even 2D re-projection error was computed to discuss about the best performance metric for the 3D reconstruction applications.

5.1.2 Evaluation

In this study regarding characterization of camera pose estimation, the sensitivity of the pose estimation algorithm was explored over various attributes of feature correspondences, which included:

- Studying the influence of noise, number and sparsity of feature point correspondences obtained by the stereo camera pair, on pose estimation.
- Studying also the influence of variation in camera baseline, in order to investigate whether the change in camera baseline makes a difference on the behavior of pose estimation.
- Evaluating based on performance metric (section 5.1.1) specifically pertaining to 3D space.

The evaluation setup is a complete pipeline of 3D reconstruction from 2D image points. The setup for evaluation is illustrated in figure 5.2.

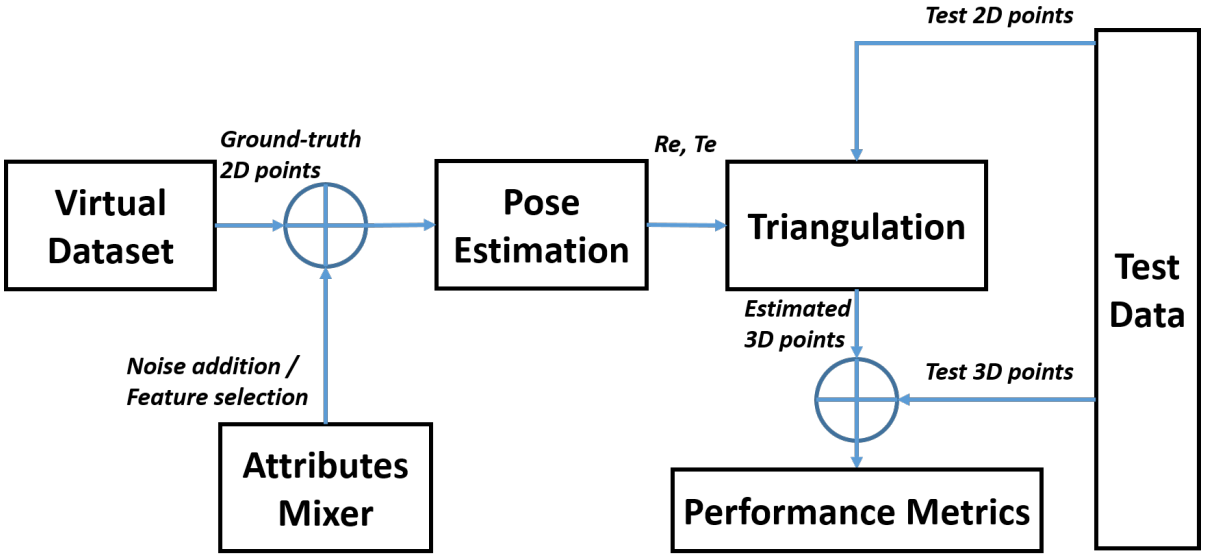


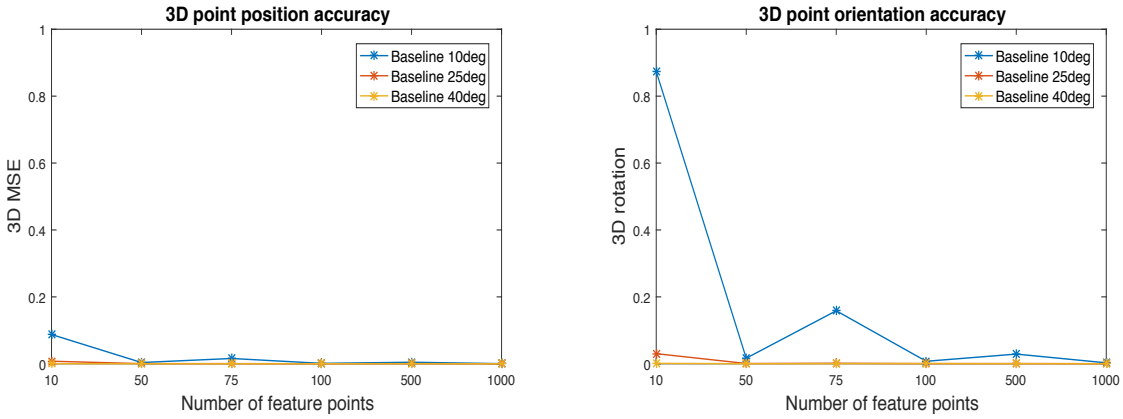
Figure 5.2: Experimental setup

The dataset was generated using virtual 3D models (figure 4.15) with a circular camera configuration (figure 4.14). From this dataset, known noise-free 2D points were obtained, which are considered ground-truth values.

Attribute mixer adds noise or selects a subset of the ground-truth 2D feature correspondence points. The selection of subset of feature points is either by selection of a specific number of points randomly or based on sparsity in 2D space. Three distinct camera baselines (angular separation between stereo) were defined, low ($10°$), medium ($25°$) and high ($40°$).

Based on a perturbed 2D feature correspondence the pose estimation computes relative camera rotation (R_e) and translation (T_e) between stereo pairs.

The aim is to measure 3D positional accuracy, 3D rotational accuracy and 3D deformation, hence a relevant test data represented by three unit vectors originating from the same origin was



(a) Mean 3D position error(measured as mean squared error)

(b) Mean 3D orientation error (measured in degrees)

Figure 5.3: Mean 3D error for different number of total feature correspondences.

generated. These unit vectors are depicted as solid three lines in figure 5.1. The figure also shows how to interpret the 3D positional, rotational and deformation errors with respect to a 3D reconstructed model.

By projecting the test 3D points onto an image, using known camera pose from the virtual dataset, the test 2D points were also generated. The test 2D points were then triangulated using estimated camera pose (R_e, T_e), which resulted in estimated 3D points. The estimated 3D points were compared with true test 3D points for measuring the 3D error.

Estimated camera pose was directly compared with known camera pose, and their average difference was measured as performance metrics: Rotation error (expressed in degrees) and Translation error (expressed upto a scale).

5.1.3 Number of Feature Correspondence

Various number of feature point (10, 50, 75, 100, 500, 1000 points) were chosen for the experiment, to estimate the camera pose. The noise level was kept at zero level to have an ideal 3D reconstruction by triangulating noiseless 2D points. Hence, this part of the study will focus independently on the effect of number of feature points.

For every set in the dataset, a subset of feature matches were chosen randomly. Each set of this experiment ran for 30 times to compensate for the randomness.

The results are shown in figures 5.3(a) for 3D position error and 5.3(b) for 3D rotation error for variation in number of feature points used for pose estimation. Here, the errors are depicted for three baselines ($10°$,$25°$,$40°$) and the 2D test points are noise-free.

For medium and high baselines ($25°$ and $40°$ respectively), the change in the number of feature matches do not significantly affect the accuracy of reconstructed 3D points. In large space application scenarios, 1 degree rotation error of the reconstructed scene would be hardly noticeable, however the significance still depends on the application.

At low baseline of $10°$, largest error is obtained by using least number of feature points, i.e., 10 points. Although the pose estimation algorithm can operate with a minimum of 5 points, it

works better with more feature points due to the use for least squares estimation (details in the algorithm description (Nistér, 2004)).

It is quite intuitive that higher the number of feature points better will be the pose estimation. However, pose estimation showed better performance at higher baselines. Therefore, pose estimation algorithm seemed to be less effective or more sensitive to low baseline. This is probably because at low baseline, there is an ambiguity in the estimation of camera rotation and translation (Hartley and Zisserman, 2004).

5.1.4 Noise in Feature Correspondence

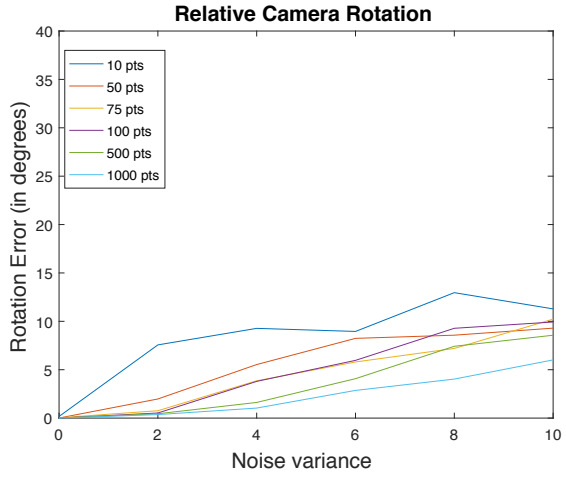
Now, the noise variation was added to the above experiment in order to study the effects of noise in the pixel coordinates of feature point correspondence. Noise was assumed to be additive Gaussian white noise, and was generated using random generator of variances σ = 0,2,4,6,8,10. The experiment was carried out for about 30 times to compensate for randomness.

The results depicting the accuracy of camera pose in presence of noise is shown in figure 5.4. Figures 5.4(a), 5.4(c) and 5.4(e) refers to 3D orientation error and 5.4(b), 5.4(d) and 5.4(f) refers to 3D position error.

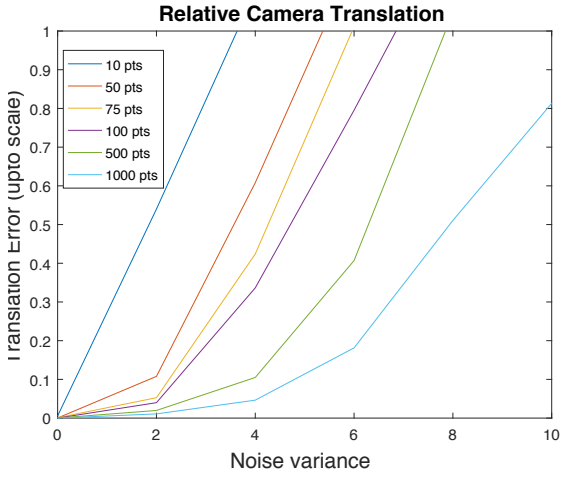
For any given baseline and any given number of feature points used (e.g., medium baseline and 50 points, in figures 5.4(c) and 5.4(d)), the increase in noise, increased the overall camera pose error. The translation error increases at a faster rate than the rotation error. This behavior was only for low and medium baselines, and for high baseline, as in figures 5.4(e) and 5.4(f), the sensitivity of translation and rotation error remained the same. Therefore, at lower baselines, pose estimator is more sensitive to noise than at higher baselines, based on the camera pose error. This is also evident in the results of 3D reconstruction, measured as 3D rotation error & 3D position error as shown in figure 5.5 and orthogonality preservation as shown in figure 5.6. Both 3D rotation and position error decreased with increase in baseline and the noise sensitivity decreases with higher baseline.

For any given baseline (e.g., medium baseline in figures 5.4(c) and 5.4(d)), it was observed that the higher the number of points used, error tends to decrease. So, at low and medium baselines where the noise sensitivity is high, using more number of points compensates for maintaining the accuracy of pose estimation.

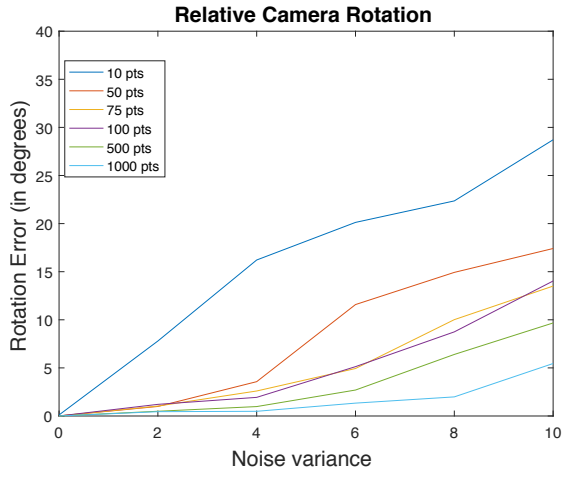
The results in figures 5.6(a), 5.6(b) and 5.6(c), showed the orthogonality preservation of angles between the test vectors that represents the deformation of 3D reconstruction quality (e.g., in figure 5.2). These plots showed orthogonality preserved value (OPV), which ranges from 0 to 1, where 1 exhibits a perfect 3D reconstruction without deformation. Highly deformed object has OPV = 0. In the figure 5.6, at high baseline, there are more OPVs at unity or close to unity compared to medium and low baselines, for any given number of feature points. And, for a given baseline (e.g., medium baseline as in figure 5.6(b)), OPVs are higher for increase in the feature points. This shows the same trend as the analysis of 3D position and orientation error discussed above. OPV represented the deformation of the 3D reconstruction object. This is comparable to the reliability of system, which is of importance to any application scenario.



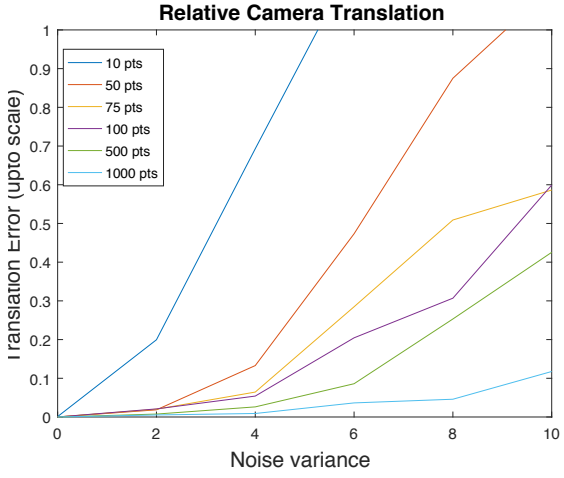
(a) Baseline 10°



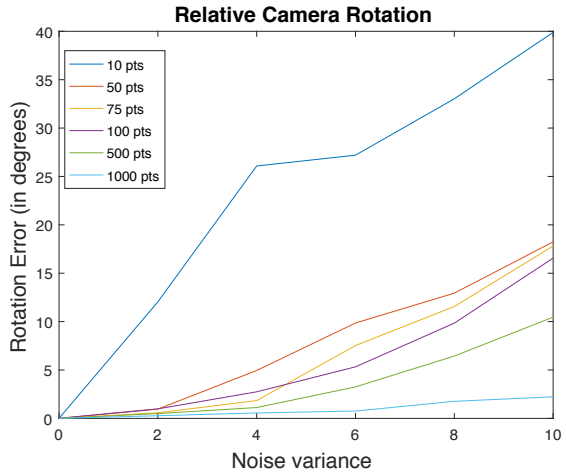
(b) Baseline 10°



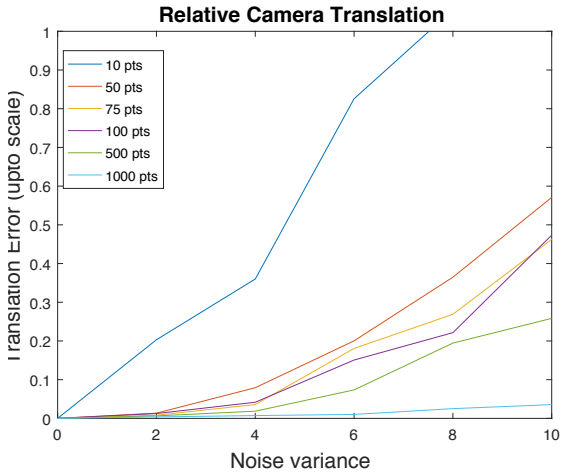
(c) Baseline 25°



(d) Baseline 25°

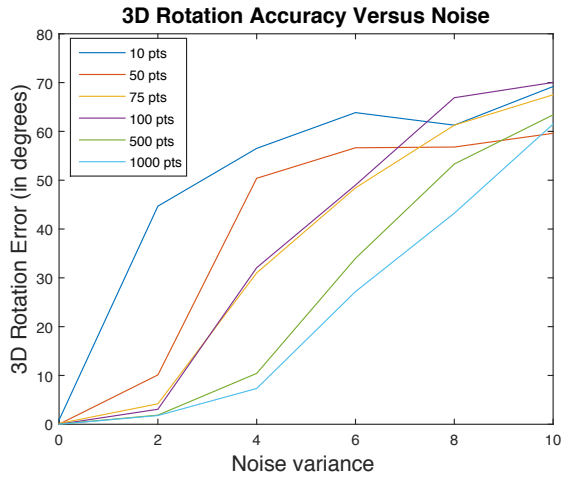


(e) Baseline 40°

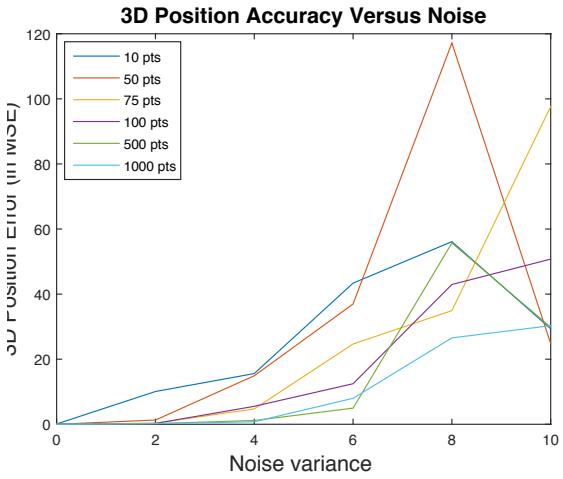


(f) Baseline 40°

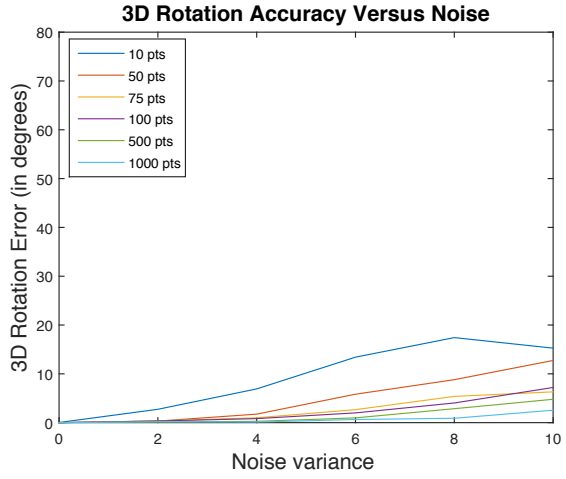
Figure 5.4: Measure of camera rotation and translation error over various noise levels with different number feature points for 3 different camera baselines.



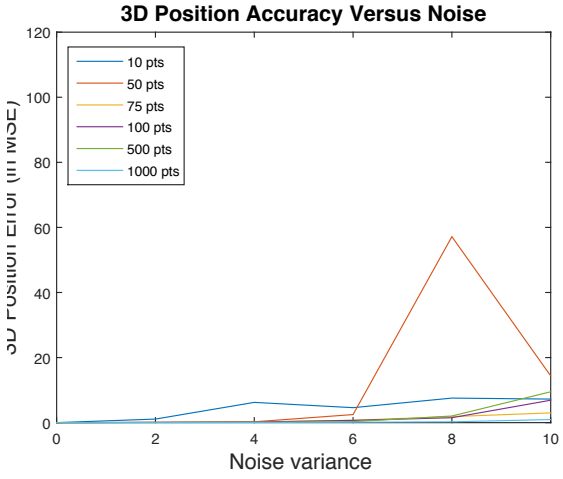
(a) Baseline 10°



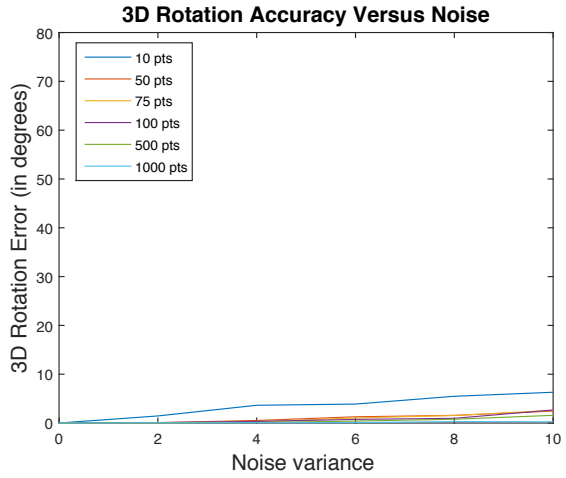
(b) Baseline 10°



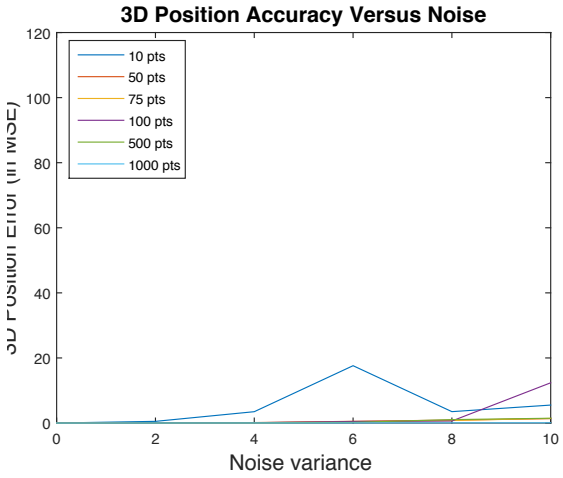
(c) Baseline 25°



(d) Baseline 25°

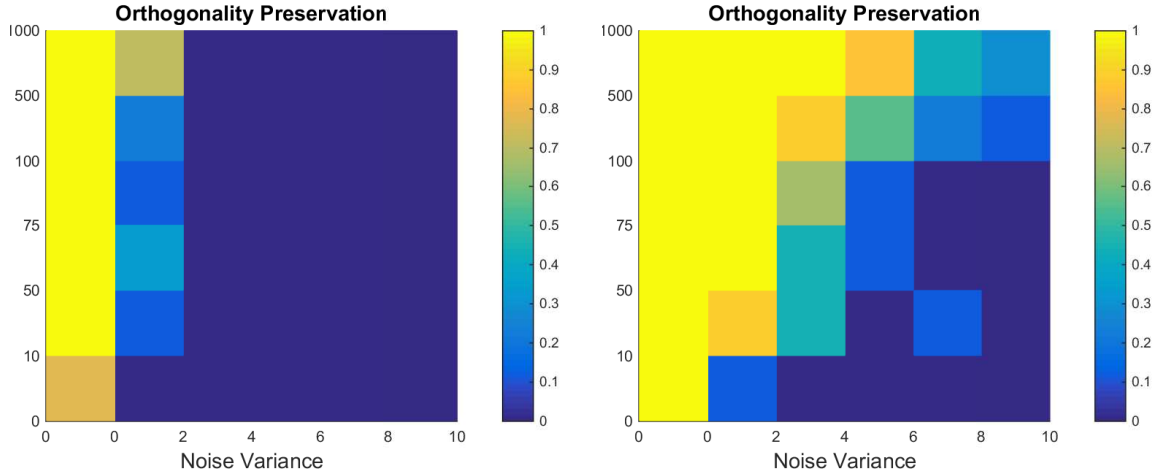


(e) Baseline 40°



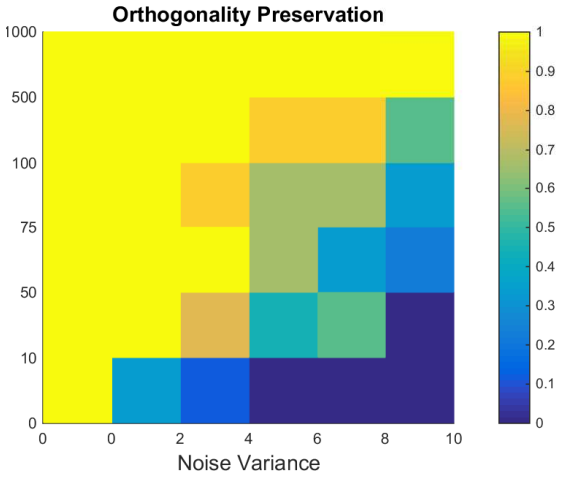
(f) Baseline 40°

Figure 5.5: Measure of 3D rotation accuracy and 3D position accuracy over various noise levels with different number feature points for 3 different camera baselines.



(a) Baseline 10°

(b) Baseline 25°



(c) Baseline 40°

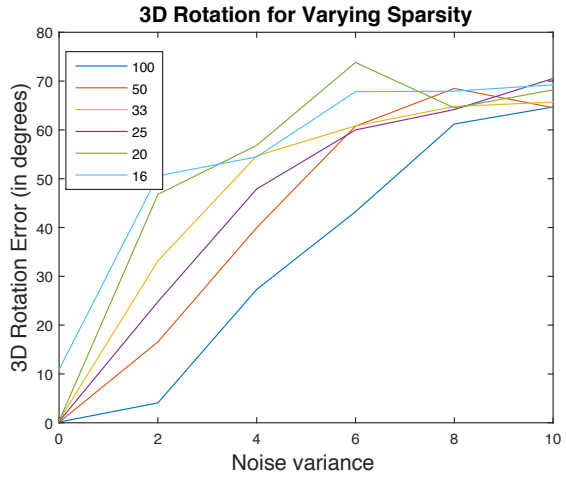
Figure 5.6: Measure of 3D orthogonality over various noise levels with different number feature points for 3 different camera baselines.

5.1.5 Sparsity of Feature Correspondence

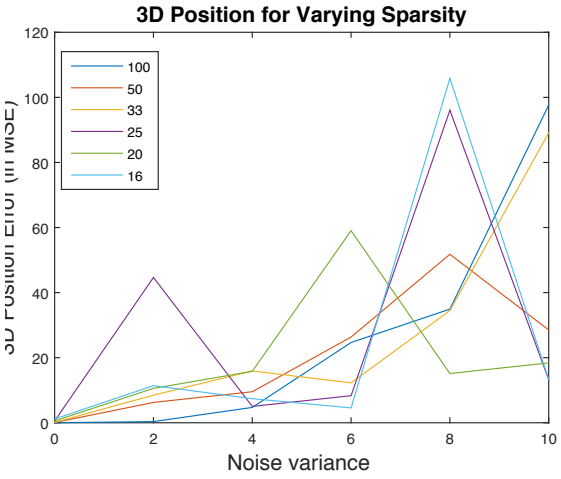
Sparsity of feature corresponding points between the stereo pair represents the spread of the matched feature points within each image. The sparsity is expressed as the percentage of area covered by the feature points within the image. For implementing sparsity, k-Means clustering methods (OpenCV implementation) was used to gather feature points with a controlled spread value. Various sparsity values used were 16,20,25,33,50,100 (expressed in percentages). Sparsity value of 100 depicts the set of feature points having the maximum spread in the 2D space.

Here, the experiments were conducted to measure 3D error and orthogonality for various sparsity values. The test combinations for varied noise levels, were as follows:

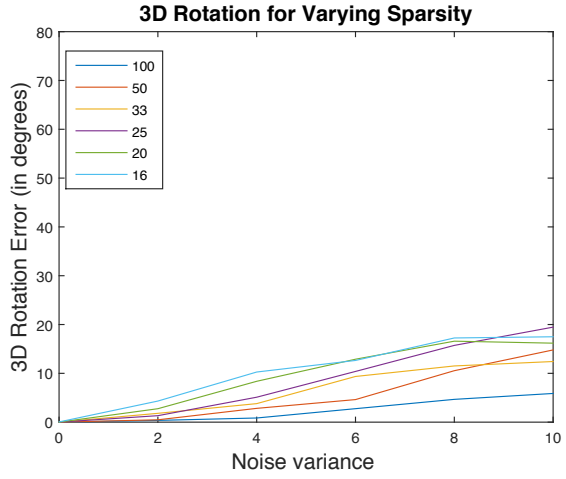
1. Low baseline - 10° and 75 feature points.
2. Medium baseline - 25° and 75 feature points.
3. Medium baseline - 25° and 500 feature points.



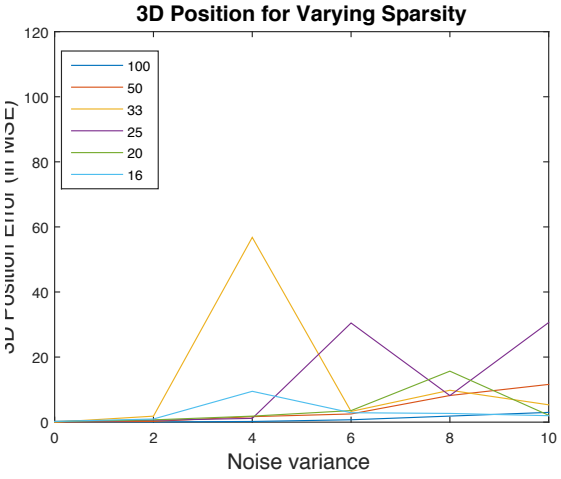
(a) Npoints 75, Baseline 10°



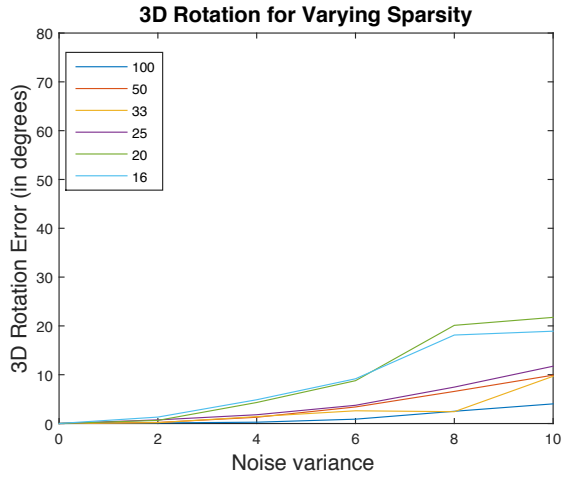
(b) Npoints 75, Baseline 10°



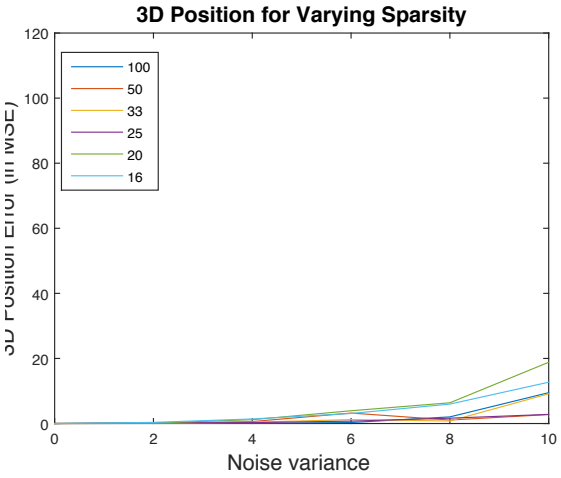
(c) Npoints 75, Baseline 25°



(d) Npoints 75, Baseline 25°

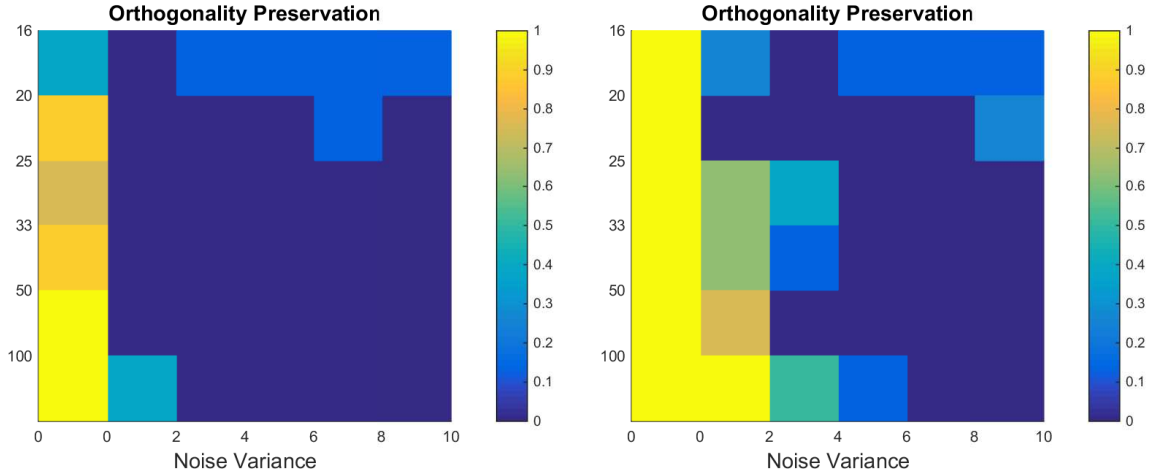


(e) Npoints 500, Baseline 25°



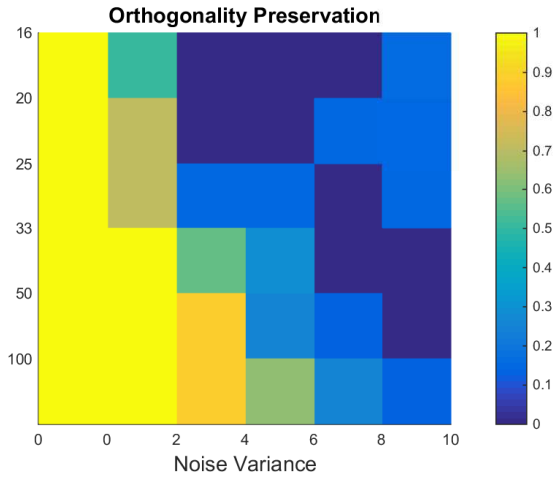
(f) Npoints 500, Baseline 25°

Figure 5.7: Measure of 3D rotation accuracy and 3D position accuracy over various noise levels with different sparsity (dispersion of points in 2D space) and various camera baselines.



(a) Npoints 75, Baseline 10°

(b) Npoints 75, Baseline 25°



(c) Npoints 500, Baseline 25°

Figure 5.8: Measure of 3D orthogonality over various noise levels with different sparsity (dispersion of points in 2D space) and various camera baselines.

In this way, the effect of sparsity can be analyzed in conjunction with various baselines and various number of feature points, for all the noise levels at the same time.

The results for using 75 points for baselines $10°$ are shown in figures 5.7(a), 5.7(b) and 5.8(a), and for baseline $25°$ are shown in 5.7(c), 5.7(d) and 5.8(b). The results for using 500 points for baselines $25°$ are shown in figures 5.7(e), 5.7(f) and 5.8(c).

The results of effects of sparsity on the accuracy of 3D reconstruction is as following:

- Overall, the accuracy of pose estimation increases with the increase in the sparsity as shown in figure 5.7. Also the reliability metric representing the orthogonality preservation, is higher for higher sparsity level. This is because, a small sparsity level represents the points confined to a small area in the image. When these feature points were used to estimate the camera extrinsic, then only a part of the pose is recovered, which is more like a localized pose. This eventually leads to error in camera pose estimation.

- At low baseline, pose estimation was noise sensitive. For higher baselines, (e.g., medium baseline in figures 5.7(c) and 5.7(e)), sparsity does not have a significant effect for small noise levels, however, when noise level increased, then the sparsity level had greater effect on the 3D accuracy. Hence, high sparsity points are also noise tolerant.
- In figures 5.7(a) and 5.7(b) in comparison with figures 5.7(c) and 5.7(d) (using same number of points, i.e., 75 points), it was observed that as the baseline increases, the accuracy increases in terms of both 3D rotation error and position error, although there are few outliers in the measurements for translation error.
- Figures 5.7(c) and 5.7(e) show that the results were similar to those seen in figure 5.5, where the accuracy of 3D error and reliability improved with increase in the number of feature points.
- The orthogonality measure for various sparsity of feature correspondences are shown in figure 5.8. The orthogonality preservations are much better as the sparsity increases.

5.1.6 3D Reconstruction Metric Evaluation

This study used the performance metric in 3D space, i.e., 3D position error, while earlier works have used the re-projection error. With the experiment it is shown how these two might be related to each other. With 75 feature points, the pose estimation is carried out for all various baselines and all noise levels. The sparsity is maintained at 100%. The result of pose estimation is as in figure 5.9.

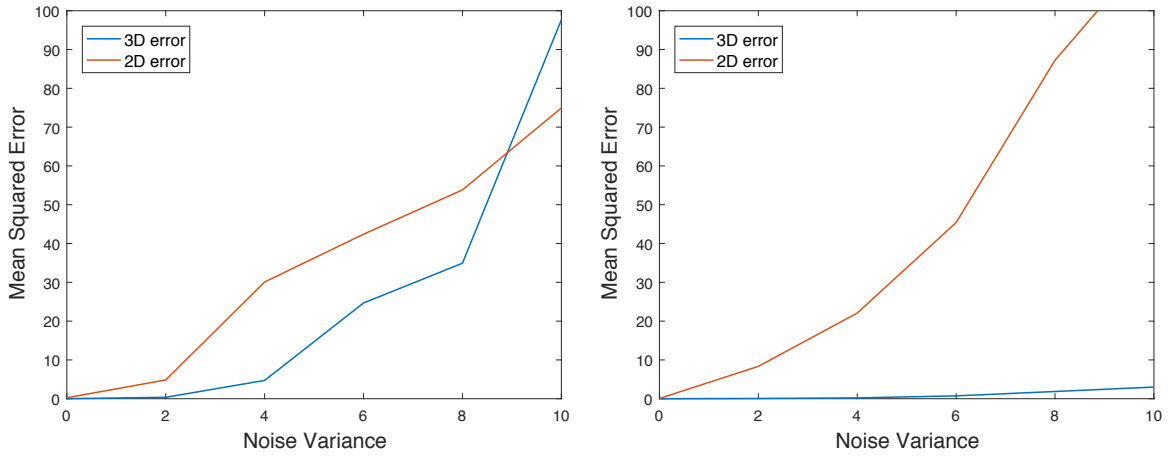
Comparing the errors in the figures 5.9(a), 5.9(b) and 5.9(c), the 2D error has lesser noise tolerance compared to 3D error at every baseline. Also the 2D error significantly increases with the increase of baseline. The 3D points estimated have a good accuracy, but the same points when re-projected back have very low accuracy. So, it is interesting to know the reason why there is a huge difference even though the re-projection is the same 3D point projected onto the 2D image.

Here, 3D error refers to only the position of reconstructed 3D points. It might be considered that the re-projection error seems like a culmination of all 3D metrics, i.e., 3D position error, 3D rotation error and orthogonality deviations. However, in earlier analysis (section 5.1.4), all 3D metrics obtained better values with increase in baseline. So, the significant increase in the reprojection error is more due to the influence of noise in the 2D points on higher baselines.

So, the re-projection error gives an idea about the inaccuracy, but not the details of the nature of error. Therefore, the 3D performance error metrics provides much better clarity to the assess the quality of a 3D reconstruction in terms of positioning, orientation and deformation compared to the re-projection error.

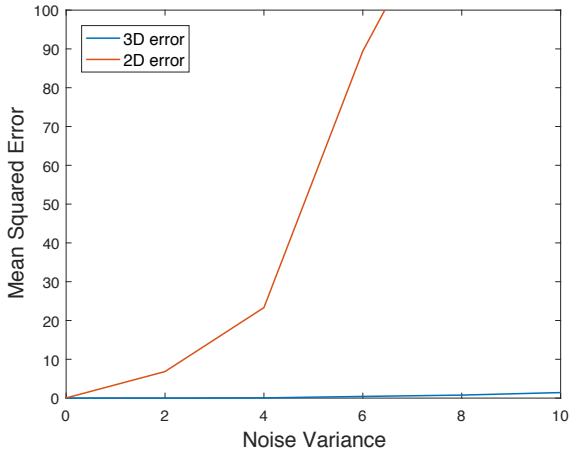
5.1.7 Discussions

The effects of noise on pose estimation was explored to understand the limitations and tolerances of pose estimation when noisy feature correspondences exist, especially when state-of-the-art feature extractors are used in FBC. From the results obtained, it was observed that for noisy feature correspondences, the pose estimation method is:



(a) Npoints 75, Baseline 10°

(b) Npoints 75, Baseline 25°



(c) Npoints 75, Baseline 40°

Figure 5.9: Comparing 2D error and 3D error variation with noise for given N=75 pts and Sparsity = 100% and three baselines.

- Sensitive to small baselines.
- Sensitive to noise in feature correspondence.
- Using more number of feature corresponding points increases the accuracy with a cost on execution time.
- Using highly sparse points gives good noise tolerance.
- Under low noise conditions, using high baseline camera setup gives more accurate pose estimation results. However, for wide baselines, the feature extraction is challenging.
- Under high noise conditions, especially for scenarios such as VERDIONE, where deformation is important, the reprojection error can become important, especially under noisy conditions.

In the context of FBC, where state-of-the-art feature extractors are used, noise in the feature correspondences are inevitable. The system designers can minimize the effect of noise based on the results of sensitivity of pose estimation algorithm. The operational limitations of pose estimation can be defined through camera baselines, number of points and feature point selection, for a minimal or an acceptable error. For instance, based on the application requirement in terms of accuracy and speed, one needs to make a good choice of number of feature correspondences to be used and the subset to use. In an instance where time is not important, one could ideally use all the feature points that are highly sparse, while the camera is setup with high baseline of around 40° apart.

Experiments were carried out on a virtual dataset that mimicked the application scenario such as VERDIONE, in order to obtain ground-truth values for determining the 3D accuracy. However, the application scenario testing was limited to only a foreground model object. The situation where the captured scene involves background with a large depth of field, is more likely to occur in the application scenario such as VERDIONE and POPART. Such a situation was not tested in this study.

The performance metrics involved measures of both 3D accuracy and 3D deformation. The importance of both of these measures depends on the application requirement. For example, the mixed reality performance scenario such as VERDIONE cares about both: accuracy - for placing a reconstructed human in the scene, and deformation - to display the reconstruction of a human subject properly. On the other hand, scenario such as POPART might care only about the accuracy, where the virtual objects need to be placed.

Validation of Sparsity Effects

The effect of sparsity on the performance of pose estimation was seen. So, the feature selection based on the sparsity has a great impact on the performance of 3D reconstruction. To validate the results, the Scanning Electron Microscopy (SEM) image reconstruction scenario was considered.

As explained in section 2.2.5, the 3D reconstruction workflow consisted of feature extraction, rectification and depth estimation. Here, the feature correspondences were used to estimate a fundamental matrix and thereafter a rectification homography matrix was computed. The homography was used to convert each of the stereo pair to rectified images. In the rectified images, the feature correspondences must lie on a straight line. It is interesting to see how the quality of rectification is affected by subset of selected features based on sparsity. The actual quality of depth estimation was ignored at this point, assuming that high quality rectification is important and necessary for a high quality depth estimation.

The feature selection procedure was implemented based on sparsity in Mountains Software[1]. The results are as shown in the figure 5.10. In figure, a stereo pair of SEM images can be seen, where the camera was tilted to capture two images of a nano-particle. When about 15 feature correspondences were manually selected with sparsity value $\approx 10\%$, a very bad rectification result was obtained as shown in figure 5.10. When the same 15 feature correspondences were selected with a larger spread $\approx 80\%$, then the rectification result was much better as shown in

[1]Digital Surf: Mountains - Surface imaging and metrology software.`http://www.digitalsurf.com/en/mntkey.html`

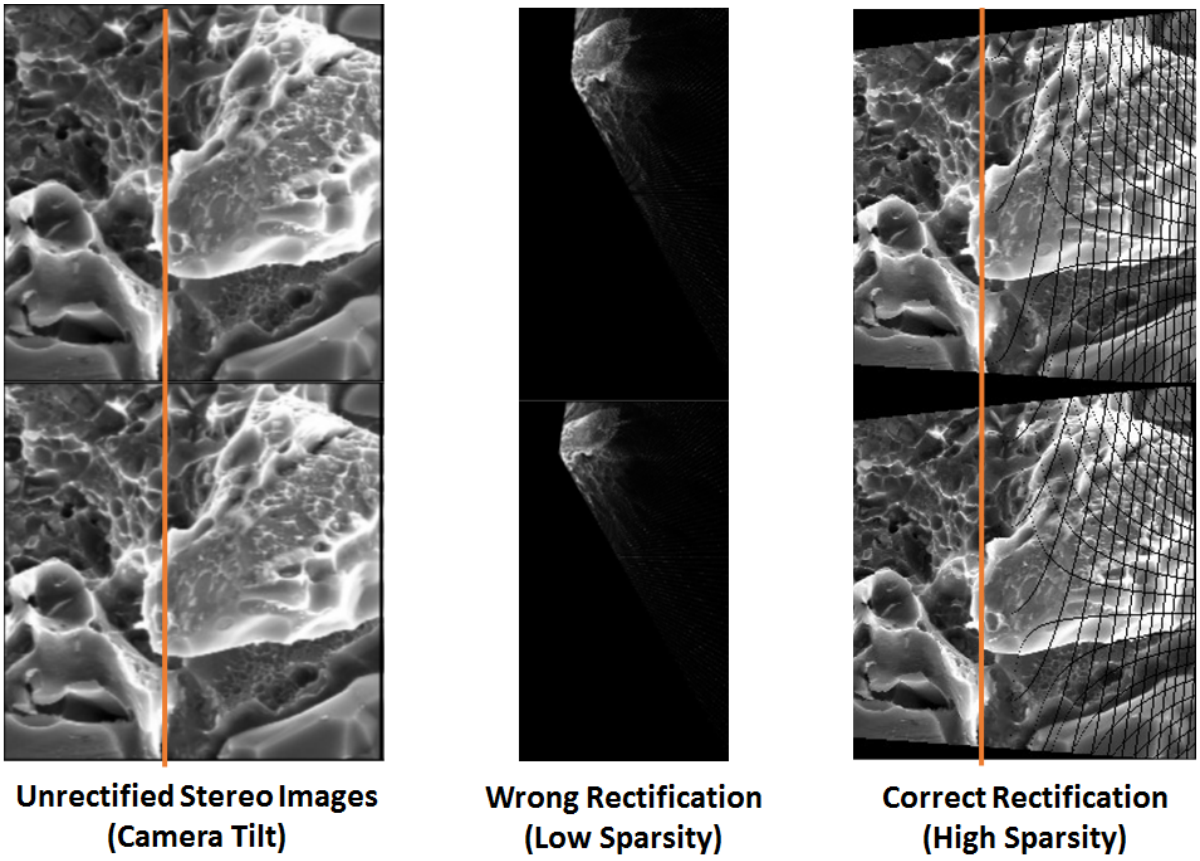


Figure 5.10: SEM reconstruction with Mountains software.

figure 5.10. The dotted line guides you to compare the corresponding points before and after rectification. This signifies the quality of rectification. Although the result using high sparsity was better than low sparsity, it is still not perfect. This is due to using very less number of manually annotated feature matches and the sparsity was not at its maximum. This example validates the effect of sparsity on the quality of rectification, i.e., as the sparsity of feature correspondences increase, the quality relating to 3D representation will be higher.

Therefore, the importance of the feature selection based on sparsity of matched feature points in a stereo pair was understood. This was validated by implementing the feature selection in the SEMRECON scenario.

In this study, results on the sensitivity of pose estimation for FBC were shown and were validated using a current 3D application. All the above evaluation tests are evident enough to reject the Null Hypothesis V (stated in section 1.4), and hence, it can be concluded that the pose estimation accuracy significantly decreases with the increase in noise and increases with increase in the sparsity of matched feature points.

5.2 Conclusions for Pose Estimation

In this chapter, the *Hypothesis V*, as stated in section 1.4, was tested. In the process of testing the research hypothesis, this chapter stated the proposed ideas, explained the experiments setup

and discussed the results.

Accordingly the effects of pixel noise, number of feature points, sparsity and camera baseline on the performance of camera pose estimation (for the VERDIONE, POPART and SEMRECON) was studied and it was shown that the accuracy of the pose estimation has significant changes with the increase in the noise and the sparsity of matched feature points between a stereo pair.

Overall, in a practical perspective, the contributions from the study in this chapter, is summarized as follows:

- Pose estimation algorithm was evaluated for several attributes of feature correspondences in stereo image pairs. Based on the sensitivity of pose estimation algorithm, it was investigated how to maximize the accuracy of pose estimation and provided operational limitations for given camera baselines, number of points and feature point selection. This helps in choosing a better set of feature points in a scene and thereby achieve a better quality 3D reconstruction.

- The performance metrics that measured the 3D deformation was introduced. This metric is helpful for certain application scenarios (e.g., in the VERDIONE scenario), where deformation of 3D reconstruction is more important than the accuracy of 3D reconstruction (e.g., in the POPART scenario).

- A virtual dataset was generated on which the evaluation was carried out. This dataset mimicked the VERDIONE or the POPART scenarios, however, application scenario testing was limited.

In this chapter, the robustness of FBC with focus to pose estimation was explored. This completes the exploration of FBC as a whole system and in terms of its building blocks, i.e., feature extraction and pose estimation. Therefore, in the next chapter, the conclusions for investigating robustness of FBC in new-age 3D multimedia systems, are drawn.

Chapter 6

Conclusions

This thesis was motivated with new age application scenarios such as VERDIONE, BAGADUS, PTMS, POPART and SEMRECON. The FBC was necessary in order to overcome few practical challenges occurring in these applications. Hence, the main research question was posed as follows:

What are the challenges in designing FBC to achieve high accuracy and robustness against practical issues in 3D multimedia systems?

Based on the main research question, the following hypotheses were formulated to conduct research:

1. H_0 : The 3D reconstruction accuracy has insignificant effect when the camera is misaligned.

2. H_0 : The accuracy and robustness of 3D reconstruction has an insignificant effect, when the 3D system replaces CBC by FBC techniques.

3. H_0 : The performances of the state-of-the-art feature extractors have insignificant differences to the change in intrinsic and extrinsic camera parameters.

4. H_0 : Accuracy of SIFT features for wide baseline FBC is maintained at an acceptable level, only upto 30 degrees angular separation between stereo cameras.

5. H_0 : The pose estimation accuracy has an insignificant change with the increase in noise and sparsity of matched feature points.

In order to test these hypotheses, relevant experiments were conducted, and the results were analyzed based on the simulations using real datasets (for the PTMS and SEMRECON scenarios) and virtual datasets (for the VERDIONE, BAGADUS and POPART scenarios).

6.1 Main Contributions

The main contributions of this thesis are as follows:

1. A statistical tool was developed for single camera 3D systems to determine the mechanical tolerances of the camera rigs that minimize the camera misalignment error in the PTMS. This helps to improve the robustness to practical error such as camera misalignment.

2. Feature based calibration was adopted in the PTMS by replacing the traditional checkerboard calibration, to improve the flexibility and maintainability of the PTMS without manual intervention. This also helps to improve robustness to practical error, such as pantograph misalignment and image analysis error of the PTMS.

3. The adverse effects of camera misalignment in stereo 3D applications were exhibited. This helps system users to build stable camera rigs to improve the accuracy of the 3D system by restricted erroneous camera misalignment in application scenarios such as VERDIONE, BAGADUS and POPART.

4. The state-of-the-art feature extractors (SIFT, SURF and ORB) were characterized and their operating limits were determined in the presence of image defocus, lens distortion and sensor noise, at various resolutions in VERDIONE and BAGADUS like application scenarios. This helps the system users to choose a feature extractor based on the requirement for accuracy, execution time and robustness.

5. The state-of-the-art feature extractors (SIFT, SURF, ORB, KAZE, AKAZE, MSER, BRISK, FAST, STAR, BRIEF, FREAK) were characterized and design considerations were recommended for using state-of-the-art feature extractors at different camera baselines (angular displacement between the stereo pair) using virtual dataset that mimics POPART, VERDIONE or BAGADUS like application scenarios. The design considerations were based on the 3D accuracy, deformation of 3D object and execution time. This helps system users to choose a feature extractor based on design parameters. This also helps system users to determine the camera density required to capture the scene.

6. A new algorithm - $NewSIFTcalib$ was proposed, which modified the existing SIFT to yield better accuracy and computation time, especially in wide baseline camera setups. This helps to improve the usability and scalability of 3D multiview capture systems. This also helps to reduce the camera density for capturing the scene and thereby is cost effective (in terms of storage, transmission and processing of multiple images) for VERDIONE and BAGADUS like application scenario.

7. The state-of-the-art pose estimation algorithm was characterized and the camera baselines and feature selection criteria were recommended to minimize noise in the feature correspondences of a stereo pair and thereby maximize the 3D accuracy. The experiments were carried out using virtual dataset that mimics VERDIONE or POPART application scenarios. The effect feature selection based on the sparsity of feature correspondences on the 3D accuracy was validated using SEMRECON scenario. This study helps system users to make a choice of camera baseline and a subset of feature correspondences and improve the robustness of pose estimation.

6.2 Practical Implications

Based on the contributions, one could have the following implications on practical aspects.

- The 3D system builders will be able to obtain exact tolerances of both single/stereo camera movement for high accuracy 3D reconstruction. This way, it is possible to manufacture more stable camera rigs that restrict the camera for specific movements, i.e., restricting the change in camera translation or rotation.
- PTMS can now obtain hassle-free, accurate and robust 3D measurements. Even if there are any misalignment causing inaccuracy in the system, automatic re-calibration is possible without manual intervention.
- Researchers and system designers can now make a choice of feature extractor for a selection of quality metrics for 3D reconstruction. These quality metrics can be categorized into 3D accuracy, 3D deformation and execution time. 3D accuracy represents precise 3D reconstruction on a metric scale. 3D deformation represents robust 3D reconstruction that matches the real object with similarity criteria. Execution time represents the real-time-ness of the feature extractors.
- A modified SIFT based algorithm can give the same or better performance compared to SIFT feature extractor at wide baselines of upto 2 meters.
- Researchers and system designer can now use various strategies to handle noisy data in order to efficiently estimate the camera pose, and eventually obtain high quality 3D reconstruction.

6.3 Practical Insight

Most of the evaluations and practical recommendations made are based on various camera baselines. As an overall insight of this thesis, the performances of feature extractors and pose estimation were examined for various camera baselines.

Feature extractors are found to perform very well at low baselines and worsens at higher baselines. On the other hand, pose estimation performs very well at high baseline and worsens at low baselines. Therefore, for an efficient implementation of FBC and thereby 3D reconstruction, one should consider to design FBC with a balanced selection of baseline.

For large spaces, cost effective solution would be to use less number of cameras, i.e., cameras setup at high baselines. Here, pose estimation works fine, but more focus is required to improve the robustness of feature extractors for higher baselines.

For small spaces, the cameras are at small baselines. Here, the feature extractors are accurate, but pose estimation is too sensitive. So, the robust pose estimation at low baselines is necessary.

Accordingly, it is recommended to avoid very low or very high camera baselines to achieve an optimal result based on the requirements of the application. However, there is room for improving the accuracy of the feature extraction at high baselines and the accuracy of pose estimation at low baselines.

6.4 Future work

In order to move towards having good solution for 3D reconstruction at extreme baselines, one has to work on improving the accuracy of feature extraction at high baselines, and the accuracy of pose estimation at low baselines. This would extend the operation range of the feature extractors and pose estimation.

A virtual dataset was good enough for evaluations, as they provided ground-truth values for performance measurement. However, considering to create the background of the scene or objects that have a higher depth of field, in the virtual dataset helps test more realistic situations. Such virtual datasets mimic 3D multiview applications in an extensive way. This does not guarantee that good reconstruction is obtained based on the current algorithms, because, for very high depths, the camera baseline becomes very small, and the reconstruction can give unstable results. However, it is worth exploring the limits in obtaining a accurate 3D reconstruction.

For outdoor scenes, sunlight plays an important role for accurate 3D reconstruction, similar to artificial light in an indoor scene. Sunlight can cause instability in the feature detection process because, a high illumination results in low contrast images. Indoors, there are artificial lights that vary in illumination, and the feature extractors are very sensitive to change in illuminations. Therefore, it would be worthwhile to study the effects of illumination on feature extraction between two stereo images.

Until now, cameras were assumed to be homogeneous, i.e., all cameras have same focal length throughout this thesis. In some situations, heterogeneous cameras might be needed, and hence, exploring 3D reconstruction when two stereo cameras differ in their camera intrinsic values would be really interesting.

It was also assumed that the intrinsic camera calibration parameter was known. But, in cases like the SEMRECON scenario, the microscope is modeled as pin hole camera, and the 3D reconstruction takes place as an uncalibrated case. This might compromise the accuracy of the 3D reconstruction. If FBC is extended to determine intrinsic parameters, then the accuracy of the 3D reconstruction in cases of unknown focal length becomes possible.

In all the studies, the cameras are assumed to be fixed. In case of moving cameras, especially in movie sets, FBC might have constraints on real-time-ness. It would be worthwhile to investigate the limitations of the speed of feature extractors by using hardware-accelerated devices like graphics processing units (GPUs).

Bibliography

Agarwal, S., Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski (2011). Building rome in a day. *Communications of ACM 54*(10).

Agrawal, M., K. Konolige, and M. R. Blas (2008). Censure: Center surround extremas for realtime feature detection and matching. In *Proceedings of the 10th European Conference on Computer Vision (ECCV)*, Volume Part IV, Berlin, Heidelberg, pp. 102–115. Springer Berlin Heidelberg.

Alahi, A., R. Ortiz, and P. Vandergheynst (2012). Freak: Fast retina keypoint. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 510–517. IEEE Computer Society.

Alcantarilla, P. F., A. Bartoli, and A. J. Davison (2012). Kaze features. In *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, Volume Part VI, Berlin, Heidelberg, pp. 214–227. Springer-Verlag.

Alismail, H. S., B. Browning, and M. B. Dias (2011). Evaluating pose estimation methods for stereo visual odometry on robots. In *Proceedings of the 11th International Conference on Intelligent Autonomous Systems (IAS)*.

Araki, N., T. Sato, Y. Konishi, and H. Ishigaki (2009). Vehicle's orientation measurement method by single-camera image using known-shaped planar object. In *Proceedings of the 4th International Conference on Innovative Computing, Information and Control (ICICIC)*, pp. 193–196.

Basso, F., R. Levorato, and E. Menegatti (2014). Online Calibration for Networks of Cameras and Depth Sensors. *The 12th Workshop on Non-classical Cameras, Camera Networks and Omnidirectional Vision (OMNIVIS)*.

Bay, H., A. Ess, T. Tuytelaars, and L. Van Gool (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding (CVIU) 110*(3), 346–359.

Bolecek, L. and V. Ricny (2015). Influence of stereoscopic camera system alignment error on the accuracy of 3D reconstruction. *Journal of Radioengineering 24*.

Bouguet, J. Y. (2008). Camera calibration toolbox for matlab. $http : //www.vision.caltech.edu/bouguetj/calib\_doc/$.

Bres, S. and B. Tellez (2009). Localisation and augmented reality for mobile applications in cultural heritage. In *Proceedings of the Workshop 3D ARCH*.

Brosnan, T. and D.-W. Sun (2004). Improving quality inspection of food products by computer vision–a review. *Journal of Food Engineering 61*(1), 3 – 16.

Brückner, M., F. Bajramovic, and J. Denzler (2008). Experimental evaluation of relative pose estimation algorithms. In *Proceedings of the 3rd International Conference on Computer Vision Theory and Applications (VISAPP*, pp. 431–438.

Calonder, M., V. Lepetit, C. Strecha, and P. Fua (2010). Brief: Binary robust independent elementary features. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*, Volume Part IV, Berlin, Heidelberg, pp. 778–792. Springer-Verlag.

Carr, P., Y. Sheikh, and I. Matthews (2012). Point-less calibration: Camera parameters from gradient-based alignment to edge images. In *Proceedings of the Workshop on Applications of Computer Vision (WACV)*.

Cesetti, A., E. Frontoni, A. Mancini, P. Zingaretti, and S. Longhi (2009). A single-camera feature-based vision system for helicopter autonomous landing. In *Proceedings of the International Conference on Advanced Robotics (ICAR)*, pp. 1–6.

Chen, X., L. Xu, Y. Wang, H. Wang, F. Wang, X. Zeng, Q. Wang, and J. Egger (2015). Development of a surgical navigation system based on augmented reality using an optical see-through head-mounted display. *Journal of Biomedical Informatics 55*, 124 – 131.

Comer, D. E., D. Gries, M. C. Mulder, A. Tucker, A. J. Turner, and P. R. Young (1989). Computing as a discipline. *Communincations of ACM 32*(1), 9–23.

Cord, A. and S. Chambon (2012). Automatic road defect detection by textural pattern recognition based on adaboost. *Journal of Computer-Aided Civil and Infrastructure Engineering 27*(4), 244–259.

Şahin IşÄśk and K. Özkan (2014). A comparative evaluation of well-known feature detectors and descriptors. *International Journal of Applied Mathematics, Electronics and Computers 3*(1).

Dosovitskiy, A., J. T. Springenberg, M. Riedmiller, and T. Brox (2014). Discriminative unsupervised feature learning with convolutional neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, Cambridge, MA, USA, pp. 766–774. MIT Press.

Drummond, T. and R. Cipolla (1999). Real-time tracking of complex structures with on-line camera calibration. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 574–583.

El-gayar, M., H. Soliman, and N. meky (2013). A comparative study of image low level feature extraction algorithms. *Egyptian Informatics Journal 14*(2), 175 – 181.

El-Mashad, S. Y. and A. Shoukry (2014). Evaluating the robustness of feature correspondence using different feature extractors. In *Proceedings of the International COnference on Methods and Models in Automation and Robotics (MMAR)*, pp. 316–321. IEEE.

Faugeras, O. (1993). *Three-dimensional Computer Vision: A Geometric Viewpoint*. Cambridge, MA, USA: MIT Press, ISBN: 0262061589.

Feng, W. C., R. Feng, P. Wyatt, and F. Liu (2016). Understanding the impact of compression on feature detection and matching in computer vision. In *Proceedings of the International Symposium on Multimedia (ISM)*, pp. 457–462. IEEE Press.

Fischler, M. A. and R. C. Bolles (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the Association for Computing Machinery (ACM) 24*(6), 381–395.

Fragoso, V., S. Gauglitz, S. Zamora, J. Kleban, and M. Turk (2011). Translatar: A mobile augmented reality translator. In *Proceedings of the Workshop on Applications of Computer Vision (WACV)*, Kona, Hawaii. IEEE Press.

G, N. V. and H. K. S (2010). Article: Quality inspection and grading of agricultural and food products by computer vision-a review. *International Journal of Computer Applications (IJCA) 2*(1), 43–65.

Gauglitz, S., T. Höllerer, and M. Turk (2011). Evaluation of interest point detectors and feature descriptors for visual tracking. *International Journal of Computer Vision (IJCV) 94*(3), 335–360.

Godding, R. (2000). 6 - geometric calibration of digital imaging systems. In *Computer Vision and Applications*, pp. 153 – 175. San Diego: Academic Press.

Goesele, M., N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz (2007). Multi-view stereo for community photo collections. In *Proceedings of the 11th International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, pp. 265–270. IEEE Press.

Halvorsen, P., S. Sægrov, A. Mortensen, D. K. C. Kristensen, A. Eichhorn, M. Stenhaug, S. Dahl, H. K. Stensland, V. R. Gaddam, C. Griwodz, and D. Johansen (2013). Bagadus: An integrated system for arena sports analytics: A soccer case study. In *Proceedings of the 4th ACM Multimedia Systems Conference (MMSys)*, New York, NY, USA, pp. 48–59. ACM.

Hartley, R. I. (1992). Estimation of relative camera positions for uncalibrated cameras. In *Proceedings of the 2nd European Conference on Computer Vision (ECCV)*, Berlin, Heidelberg, pp. 579–587. Springer Berlin Heidelberg.

Hartley, R. I. (1997). In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 19*(6), 580–593.

Hartley, R. I. and A. Zisserman (2004). *Multiple View Geometry in Computer Vision* (Second ed.). Cambridge University Press, ISBN: 0521540518.

Hayman, E. and D. W. Murray (2003). The effects of translational misalignment when self-calibrating rotating and zooming cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence 25*(8), 1015–1020.

Heikkilä, J. (2000). Geometric camera calibration using circular control points. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 22*(10), 1066–1077.

Heimonen, T., J. Hannuksela, J. Heikkila, J. Leinonen, and M. Manninen (2001). Experiments in 3D measurements by using single camera and accurate motion. In *Proceedings of the International Symposium on Assembly and Task Planning (ISATP)*, pp. 356–361.

Hijazi, A., A. Friedl, and C. Kähler (2011). Influence of camera's optical axis non-perpendicularity on measurement accuracy of two-dimensional digital image correlation. *Jordan Journal of Mechanical and Industrial Engineering (JJMIE) 5*(4).

Jain, R. (1991). The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling. *Wiley-Interscience: ISBN: 0471503363*.

Jiayuan, R., W. Yigang, and D. Yun (2010). Study on eliminating wrong match pairs of SIFT. In *Proceedings of the 10th International Conference on Signal Processing*, pp. 992–995.

Juhász, E., A. Tanács, and Z. Kato (2013). Evaluation of point matching methods for wide-baseline stereo correspondence on mobile platforms. In *Proceedings of the 8th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pp. 813–818.

Kassebaum, J., N. Bulusu, and W.-C. Peng (2010). 3DD target-based distributed smart camera network localization. *IEEE Transactions on Image Processing 19*(10), 2530–2539.

Kniaz, V. V. (2016). Robust vision-based pose estimation algorithm for an UAV with known gravity vector. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLI-B5*, 63–68.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Journal of The Annals of Mathematical Statistics 22*(1), 79–86.

Kurillo, G., Z. Li, and R. Bajcsy (2008). Wide-area external multi-camera calibration using vision graphs and virtual calibration object. In *Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, Stanford, CA, USA, pp. 1–9. IEEE Press.

Le Flohic, J., V. Parpoil, S. Bouissou, M. Poncelet, and H. Leclerc (2014). A 3D displacement control by digital image correlation for the multiaxial testing of materials with a stewart platform. *Journal on Experimental Mechanics 54*(5), 817–828.

Lepetit, V., F. Moreno-Noguer, and P. Fua (2009). Epnp: An accurate O(n) solution to the pnp problem. *International Journal of Computer Vision (IJCV) 81*(2), 155–166.

Leutenegger, S., M. Chli, and R. Y. Siegwart (2011). Brisk: Binary robust invariant scalable keypoints. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Washington, DC, USA, pp. 2548–2555. IEEE Computer Society.

Levinson, J. and S. Thrun (2013). Automatic online calibration of cameras and lasers. In *Proceedings of the Robotics: Science and Systems*, Berlin, Germany.

Li, C. and L. Lu, Pingand Ma (2010). A camera on-line recalibration framework using SIFT. *The Visual Computer 26*(3), 227–240.

Lima, J. P., F. Simões, L. Figueiredo, and J. Kelner (2010). Model based markerless 3D tracking applied to augmented reality. *Journal on 3D Interactive Systems 1*.

Liu, R., H. Zhang, M. Liu, X. Xia, and T. Hu (2009). Stereo cameras self-calibration based on sift. In *Proceedings of the International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, Volume 1, Washington, DC, USA, pp. 352–355. IEEE Computer Society.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV) 60*(2), 91–110.

Lu, C.-P., G. D. Hager, and E. Mjolsness (2000). Fast and globally convergent pose estimation from video images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 22*(6), 610–622.

Ma, Y., S. Soatto, J. Kosecka, and S. S. Sastry (2003). *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer-Verlag, ISBN: 0387008934.

Maidi, M., M. Mallem, L. Benchikh, and S. Otmane (2013). An evaluation of camera pose methods for an augmented reality system: Application to teaching industrial robots. In *Transactions on Computational Science XVII*, Berlin, Heidelberg, pp. 3–30. Springer Berlin Heidelberg.

Mar, N. S. S., C. Fookes, and P. K. Yarlagadda (2009). Design of automatic vision-based inspection system for solder joint segmentation. *Journal of Achievements in Materials and Manufacturing Engineering 34*(2), 145–151.

Matas, J., O. Chum, M. Urban, and T. Pajdla (2002). Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 36.1–36.10. BMVA Press.

Matusik, W., C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan (2000). Image-based visual hulls. In *Proceedings of the 27th International Conference on Computer Graphics and Interactive Techniques*, New York, NY, USA, pp. 369–374. ACM Press/Addison-Wesley Publishing Co.

Mavrinac, A., X. Chen, and K. Tepe (2008). Feature-based calibration of distributed smart stereo camera networks. In *Proceedings of the 2nd International Conference on Distributed Smart Cameras (ICDSC)*, pp. 1–10.

Michael Ying Yang, Yanpeng Cao, J. M. (2011). Fusion of camera images and laser scans for wide baseline 3D scene alignment in urban environments. *ISPRS - Journal of Photogrammetry and Remote Sensing 66*(6, Supplement), 52 – 61.

Min, D. B., D. Kim, S. Yun, and K. Sohn (2009). 2D/3D freeview video generation for 3DTV system. *Signal Processing: Image Communication 24*(1-2), 31–48.

Moeslund, T. B. and E. Granum (2001). A survey of computer vision-based human motion capture. *Journal on Computer Vision Image Understanding (CVIU) 81*(3), 231–268.

Moreels, P. and P. Perona (2007). Evaluation of features detectors and descriptors based on 3D objects. *International Journal of Computer Vision (IJCV) 73*(3), 263–284.

Muhanna, M. A. (2015). Virtual reality and the CAVE: Taxonomy, interaction challenges and research directions. *Journal of King Saud University - Computer and Information Sciences 27*(3), 344 – 361.

Nistér, D. (2004). An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 26*(6), 756–777.

Pablo Alcantarilla, J. N. and A. Bartoli (2013). Fast explicit diffusion for accelerated features in nonlinear scale spaces. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press.

Pedersini, F., A. Sarti, and S. Tubaro (1998). Multi-camera acquisitions for high-accuracy 3D reconstruction. In *Proceedings of the 3D Structure from Multiple Images of Large-Scale Environments (SMILE)*, Berlin, Heidelberg, pp. 124–138. Springer Berlin Heidelberg.

Ribo, M., A. Pinz, and A. L. Fuhrmann (2001). A new optical tracking system for virtual and augmented reality applications. In *Proceedings of the 18th Instrumentation and Measurement Technology Conference (IMTC)*, Volume 3, Budapest, Hungary, pp. 1932–1936. IEEE Press.

Rocchini, C., P. Cignoni, C. Montani, P. Pingi, and R. Scopigno (2001). A low cost 3D scanner based on structured light. *Journal of Computer Graphics Forum*.

Rodehorst, V., M. Heinrichs, and O. Hellwich (2008). Evaluation of relative pose estimation methods for multi-camera setups. In *Proceedings of the International Society for Photogrammetry and Remote Sensing*.

Rosten, E. and T. Drummond (2006). Machine learning for high-speed corner detection. In *Proceedings of the 9th European Conference on Computer Vision (ECCV)*, Volume Part I, Berlin, Heidelberg, pp. 430–443. Springer Berlin Heidelberg.

Rublee, E., V. Rabaud, K. Konolige, and G. Bradski (2011). Orb: An efficient alternative to sift or surf. In *Proceedings of the International Conference on Computer Vision (ICCV)*, Washington, DC, USA, pp. 2564–2571. IEEE Computer Society.

Santoro, M., G. AlRegib, and Y. Altunbasak (2012). Misalignment correction for depth estimation using stereoscopic 3-D cameras. In *Proceedings of the International Workshop on Multimedia Signal Processing (MMSP)*, pp. 19–24. IEEE.

Shkurti, F., I. Rekleitis, and G. Dudek (2011). Feature tracking evaluation for pose estimation in underwater environments. In *Proceedings of the Canadian Conference on Computer and Robot Vision (CRV)*, pp. 160–167.

Snavely, N., S. M. Seitz, and R. Szeliski (2006). Photo tourism: Exploring photo collections in 3D. In *Proceedings of the International Conference on ACM SIGGRAPH*, New York, NY, USA, pp. 835–846. ACM.

Suenaga, H., H. H. Tran, H. Liao, K. Masamune, T. Dohi, K. Hoshi, and T. Takato (2015). Vision-based markerless registration using stereo vision and an augmented reality surgical navigation system: a pilot study. *Journal of BMC Medical Imaging 15*(1), 1–11.

Sun, W. and J. R. Cooperstock (2006). An empirical evaluation of factors influencing camera calibration accuracy using three publicly available techniques. *Journal of Machine Vision and Applications 17*(1), 51–67.

Tanimoto, M. (2010). Free-viewpoint television. In R. Ronfard and G. Taubin (Eds.), *Image and Geometry Processing for 3-D Cinematography*, Berlin, Heidelberg, pp. 53–76. Springer Berlin Heidelberg.

Tsai, R. Y. (1992). Radiometry. Chapter A Versatile Camera Calibration Technique for High-accuracy 3D Machine Vision Metrology Using Off-the-shelf TV Cameras and Lenses, pp. 221–244. USA: Jones and Bartlett Publishers, Inc.

Turk, M., C. Kim, J. Yi, and J. Park (2005). Structured light based depth edge detection for object shape recovery. *Computer Vision and Pattern Recognition (CVPR) Workshops*, 106.

Ventura, J. and T. Höllerer (2012). Wide-area scene mapping for mobile visual tracking. In *Proceedings of the International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 3–12. IEEE Press.

Wheatstone, C. (1838). Contributions to the physiology of vision. part the first. on some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London (1776-1886) 128*, 371–394.

Wilburn, B., N. Joshi, V. Vaish, M. Levoy, and M. Horowitz (2004). High-speed videography using a dense camera array. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, Volume 2, pp. 294–301.

Yuan, C. (2006). Markerless pose tracking for augmented reality. In *Proceedings of the 2nd International Symposium in Visual Computing*, Volume Part I, Berlin, Heidelberg, pp. 721–730. Springer Berlin Heidelberg.

Yun, J.-H. and R.-H. Park (2006). Self-calibration with two views using the scale-invariant feature transform. In *Proceedings of the 2nd International Symposium on Visual Computing (ISVC)*, Berlin, Heidelberg, pp. 589–598. Springer Berlin Heidelberg.

Zhang, C. and T. Chen (2004). A self-reconfigurable camera array. In *Proceedings of the International COnference on ACM SIGGRAPH*, New York, NY, USA, pp. 151. ACM.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 22*(11), 1330–1334.

Zhao, D. and S. Li (2005). A 3D image processing method for manufacturing process automation. *Journal of Computers in Industry 56*(8-9), 975–985.

Zhao, W. and N. Nandhakumar (1996). Effects of camera alignment errors on stereoscopic depth estimates. *Journal of Pattern Recognition 29*(12), 2115 – 2126.

Zitnick, C. L., S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski (2004). High-quality video view interpolation using a layered representation. In *Proceedings of the International Conference on ACM SIGGRAPH*, New York, NY, USA, pp. 600–608. ACM.

Part II

Research Papers

Chapter 7

Paper I: Faster and More Accurate Feature-Based Calibration for Widely Spaced Camera Pairs

Title: Faster and More Accurate Feature-Based Calibration for Widely Spaced Camera Pairs.

Authors: Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz and Pål Halvorsen.

Published & **Presented:** In Digital Information and Communication Technology and it's Applications (DICTAP), 2012.

Abstract: The increasing demand for live multimedia systems in gaming, art and entertainment industries, has resulted in the development of multiview capturing systems that use camera arrays. We investigate sparse (widely spaced) camera arrays to capture scenes of large volume space. A vital aspect of such systems is camera calibration, which provides an understanding of the scene geometry used for 3D reconstruction.

Traditional algorithms make use of a calibration object or identifiable markers placed in the scene, but this is impractical and inconvenient for large spaces. Hence, we take the approach of features-based calibration. Existing schemes based on SIFT (Scale Invariant Feature Transform), exhibit lower accuracy than marker-based schemes due to false positives in feature matching, variations in baseline (spatial displacement between the camera pair) and changes in viewing angle.

Therefore, we propose a new method of SIFT feature based calibration, which adopts a new technique for the detection and removal of wrong SIFT matches and the selection of an optimal subset of matches. Experimental tests show that our proposed algorithm achieves higher accuracy and faster execution for larger baselines of up to ≈ 2 meters, for an object distance of $\approx$4.6 meters, and thereby enhances the usability and scalability of multi-camera capturing systems for large spaces.

Faster and more Accurate Feature-based Calibration for Widely Spaced Camera Pairs

Deepak Dwarakanath[1,2], Alexander Eichhorn[1], Carsten Griwodz[1,2], Pål Halvorsen[1,2]
[1]*Simula Research Laboratory,* [2]*University of Oslo, Norway*
{*deepakd, echa, griff, paalh*}*@simula.no*

***Abstract*—The increasing demand for live multimedia systems in gaming, art and entertainment industries, has resulted in the development of multi-view capturing systems that use camera arrays. We investigate sparse (widely spaced) camera arrays to capture scenes of large volume space. A vital aspect of such systems is *camera calibration*, which provides an understanding of the scene geometry used for 3D reconstruction.**

Traditional algorithms make use of a calibration object or identifiable markers placed in the scene, but this is impractical and inconvenient for large spaces. Hence, we take the approach of features-based calibration. Existing schemes based on SIFT (Scale Invariant Feature Transform), exhibit lower accuracy than marker-based schemes due to false positives in feature matching, variations in baseline (spatial displacement between the camera pair) and changes in viewing angle.

Therefore, we propose a new method of SIFT feature based calibration, which adopts a new technique for the detection and removal of wrong SIFT matches and the selection of an optimal subset of matches. Experimental tests show that our proposed algorithm achieves higher accuracy and faster execution for larger baselines of up to ≈2 meters, for an object distance of ≈4.6 meters, and thereby enhances the usability and scalability of multi-camera capturing systems for large spaces.

***Keywords*-Camera arrays; Multiview capture; Feature-based calibration; SIFT; Interactive multimedia for large spaces;**

I. Introduction

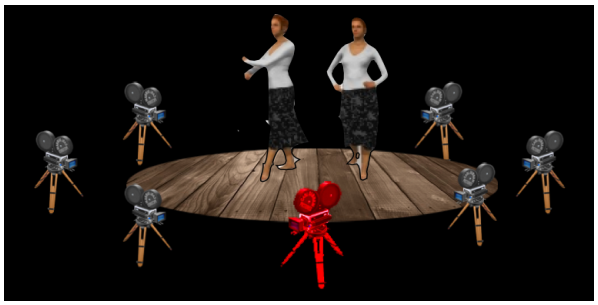
Growing computing performance and the massive parallelization in multi-core processors and specialized graphics hardware have made it possible to process complex computer graphics and computer vision algorithms in real-time. At the same time, camera sensors are becoming cheaper and improve in performance. As a consequence, new kinds of live multimedia systems based on stereoscopic and multi-view video become increasingly attractive for gaming, art and entertainment productions.

Several types of camera arrays are in practical use and development today [1], [2]. They differ in camera density and physical extent. While some image processing techniques such as light-field processing, stereoscopic and multi-view video require relatively dense camera placement, other image processing applications such as free-viewpoint rendering,visual hull reconstruction, tracking or geometrical scene reconstruction can deal with relatively sparse placement.

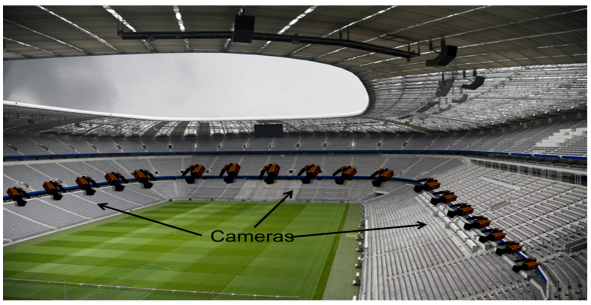
Common to all types of camera arrays is the need for geometric calibration, that is, the identification of intrinsic camera parameters (focal length, principal point and lens distortions) and extrinsic parameters (the geometrical displacement of cameras against each other). Many techniques for the calibration of low-cost camera sensors exist in the computer vision literature, with the most popular ones being methods that use a planar checkerboard pattern [3], [4] or identifiable markers [5]. The calibration accuracy that these methods achieve is sufficient for 3D image processing algorithms, but in many cases, it is inconvenient or impossible to place a measurement target like a checkerboard pattern of sufficient size in front of the cameras.

Calibration based on image feature detection, for example using SIFT [6] (Scale Invariant Feature Transform), has been proposed [7], [8], [9] as an improvement over the traditional, often manual, methods that need a calibration target. Using SIFT, these systems automatically match the features between camera images, which are then used to perform the calibration. However, a particular limitation of SIFT is the decreased feature matching performance with an increase in viewing angle between two perspectives. With a growing *baseline*, the direct distance between any two cameras in an array, less similarities exist between images and consequently, fewer SIFT features are matched. However, the difference may not only manifest in lower overlap or an increased number of occlusions. It may also result in more false positive SIFT matches.

In this work, we extend the prior state-of-art and propose an extrinsic calibration method called $newSIFTcalib$, for pairs of cameras with an arbitrary baseline that works without a calibration target. Our $newSIFTcalib$ is also based on image features obtained using SIFT, but we address some of the limitations of current SIFT-based methods. Specifically, the novelty of the method lies in (a) a new technique for the detection and removal of wrong SIFT matches and (b) a method for selecting a small subset of all detected SIFT features. Our $newSIFTcalib$ particularly compensates for increased viewing angles and large baselines, making SIFT-based calibration usable for camera arrays with large baselines. While the calibration accuracy using SIFT features depends on different factors such as camera baseline and rotation, image resolution, motion blur and external lighting, we focus on the effects of camera baselines and assume that other factors remain constant. We assume further that



(a) Mixed Reality Art Performance Stage



(b) Soccer stadium

Figure 1. Large volume application examples

intrinsic camera parameters are already known or have been determined in a prior calibration step. Based on experimental results, we show that our new method $newSIFTcalib$ can achieve higher calibration accuracy than traditional methods, works with larger baselines than existing calibration schemes and requires less execution time.

In the remainder of the article, we first introduce some example application scenarios where camera baselines are typically large. Section III presents some representative related work. Our new feature-based calibration system is introduced in section IV. Experimental setup and results are described in section V before we conclude the paper in section VI.

II. Applications with large capturing volumes

In several application scenarios, it is necessary to distribute cameras at wide baselines around a large space to capture the entire volume from an optimal number of viewpoints. Examples for such scenarios are:

Mixed Reality On-Stage Performances As in figure 1(a), a camera array is typically placed around the stage. On a remote stage the captured performers are embedded as free-viewpoint video to correct for perspective differences and achieve an aesthetically appealing result.

Sports Events in large arenas such as soccer or baseball games are captured from a large number of perspectives from around the stadium (see figure 1(b)). The video feeds obtained from multiple cameras can be used in various ways such as for silhouette extraction, video mosaicing, motion tracking of players, content analysis.

High accuracy in camera calibration is a prerequisite for high-quality processing of images from cameras at various angles. Accuracy at wide baselines and long shots that are typical in the huge volumes of arenas becomes even more important.

III. Related Work

Previously, similar work on calibration has been carried out using SIFT by, for example, Yun et al. [7], Li et al. [8] and Liu et al. [9]. However, in such algorithms, all the point correspondences obtained by SIFT feature matching have been used for calibration. This is redundant and prone to noise due to mismatches of SIFT features. Eliminating such wrong matches has been studied by Jiayuan et al. [10], using a error canceling algorithm based on RANSAC (Random Sample Consensus - a widely used algorithm for outlier removal). Alternatively, we use a simpler method based on the geometry of lines joining the matched points. Our outlier removal process is faster than and performs as good as RANSAC in our test scenario.

IV. System Description

The system overview is illustrated in figure 2, where a number of stereo camera pairs capture a scene of interest. For every 2D stereo images, we use Vedaldi's library [11] to detect SIFT feature points in stereo images and match them. As a preprocessing step, outliers (false positives in the matching process) are detected and removed. Only a subset of stable points (referred as $FeatureVector$ in rest of the paper), less prone to noise, are used for calibration. We assume the cameras are pre-calibrated for intrinsics.

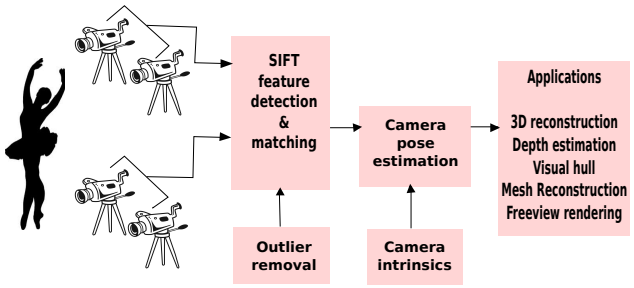


Figure 2. System Overview

A. Outlier detection

This filtering step is based on the angular deviation of the lines connecting corresponding points from, the mean direction of all the lines that connect pairs of corresponding points in two images. Consider two images from stereo

cameras placed horizontally apart from each other. Lines are drawn from every feature point in image 1 to their respective correspondences in image 2, as in figure 4.

We compute the mean (μ_θ^x) and standard deviation (σ_θ^x) of the angle between all lines and the x-axis. Now the outlier detector compares the angle between each line and the x-axis to μ_θ^x and σ_θ^x. A line l_{ij} (and thereby the point correspondence) is identified as an outlier if the angle θ_l^x differs by more than σ_θ^x, as in equation (1). The same is done for the Y-axis. In this way, we make sure that this algorithm can be used on images taken from both horizontally and vertically aligned cameras.

$$outlier = \begin{cases} l_{ij} & \text{if} \quad |\theta_l^{x/y}| > \mu_\theta^{x/y} + \sigma_\theta^{x/y} \\ 0 & \text{if} \quad |\theta_l^{x/y}| < \mu_\theta^{x/y} + \sigma_\theta^{x/y} \end{cases} \tag{1}$$

B. FeatureVector - size and selection

The feature points detected by SIFT are assigned a scale which can be interpreted as a representation of the stability of the feature detection. We exploit this property and sort the inlier point correspondences and define a $FeatureVector$, a vector consisting of point correspondences used for estimating camera pose. Tests in section V-B1 show that the dimension of $FeatureVector$ is chosen to be 25, which is the minimum number of feature points required to achieve a quality similar to the RANSAC algorithm. Next, from the pool of inlier point correspondences, five candidates of subsets from highest order of stability are chosen. Out of these five candidates, the best subset is chosen as the $FeatureVector$, based on least re-projection error, computed for the estimated camera pose.

C. Camera Pose Estimation

The $FeatureVector$ of point correspondences is used to estimate the essential matrix E using normalized 8-point algorithm [12]. Ina stereo camera setup, if the world coordinates are considered to be at the center of the reference camera, the rotation matrix of reference camera is an identity matrix and translation is a zero matrix. Relative rotation R and translation t of the second camera of the camera pair represents the camera pose, and are related to essential matrix as $E = [t]_X R$, where $[t]_X$ is a skew-symmetric matrix,

$$[t]_X = \begin{bmatrix} 0 & t_x & -t_z \\ -t_x & 0 & t_y \\ t_z & -t_y & 0 \end{bmatrix}$$

The Essential matrix can be decomposed using SVD (Singular Value Decomposition) as in [13], which is detailed as follows:

Let K_1 and K_2 be the intrinsics of the camera pair respectively. Upon SVD of E, we obtain:

$$E = USV^T \tag{2}$$

where U and V are unitary matrices and S is a rectangular diagonal matrix. Accordingly, R has two solutions R_a, R_b, and t has two solution t_a, t_b, which are given by

$$R_a = UWV^T, R_b = UW^TV^T, t_a = +u3, t_b = -u3, \tag{3}$$

where u3 is the 3rd column of matrix U and W is as follows:

$$W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

This gives a choice of four solutions to obtain the camera pose. A projection matrix of the reference camera is given as $P_1 = K_1[I|0]$. If $P_2 = K_2[R|t]$ is the projection matrix of the camera, then solution is one of the following:

$$P_2 = K_2[R_a|t_a] \ , \ K_2[R_a|t_b] \ , \ K_2[R_b|t_a] \ , \ K_2[R_b|t_b]$$

The above four solutions have a geometrical meaning and one of the solution is always meaningful. For every possible solution of P_2, 3D points corresponding to the intersection of back projected ray from 2D point correspondences are estimated through triangulation. Using cheirality constraint [14], the 3D points obtained are checked for positive sign of depth and hence the solution for camera pose is determined.

V. Experimentation

A. Dataset

We used widely accepted multi view image dataset by Microsoft Research Laboratory [15] to test our algorithm against others. The dataset was produced using a setup as illustrated in figure 3. All 8 cameras (separated by $\approx$ 0.3 meters distance) captured an event (taken place at $\approx$ 4.6 meters) with a resolution of 1024x728, and rate of 15fps. The calibration parameters for these cameras were computed using traditional approach (checkerboard). These known calibration parameters are used for comparing parameters estimated using other algorithm.

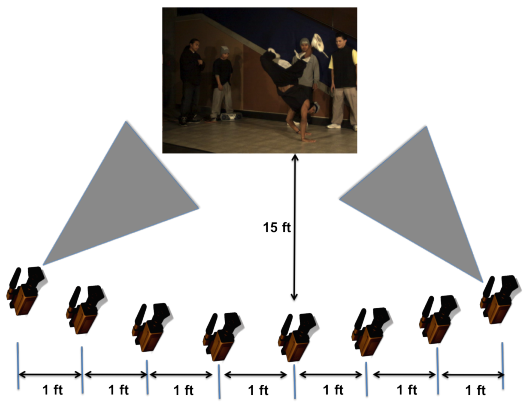


Figure 3. Illustration of setup used by Microsoft [15] to produce the multi-view dataset

B. Test results

First, we conducted an experiment to evaluate the performance of the outlier removal module, and then to evaluate our *newSIFTcalib* algorithm, which comprises of two main techniques - outlier removal and *FeatureVector* selection.

1) Testing the outlier removal performance: To evaluate the performance of outlier removal, we use a first order approximation to geometric error, referred as $EpipolarErr$ in this paper, as stated in [14] and computed as in equation 4, where F is the fundamental matrix: $F = K_2^{-T} E K_1^{-1}$.

$$E_p = \sum_{i=1}^{N} \left(\frac{x\prime_i F x_i}{(Fx_i)_1^2 + (Fx_i)_2^2 + (F^T x\prime_i)_1^2 + (F^T x\prime_i)_2^2} \right) \quad (4)$$

After outlier (solid lines in figure 4) removal, *Epipolar Err* is computed for following methods (a) 8-pt algorithm without outlier detection (b) 8-pt algorithm with RANSAC (c) 8-pt algorithm with our proposed outlier removal.

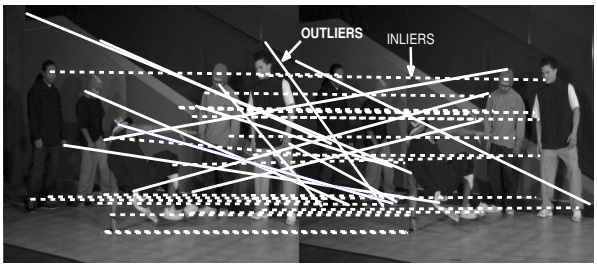


Figure 4. Process of outlier detection: outliers (solid), inliers (dotted)

The test results in fig 5 shows the RANSAC method performs better than 8-point algorithm without outlier removal, as expected. It is very evident that our proposed outlier removal performs as good as RANSAC, and the computation time is drastically reduced because RANSAC requires a large number of points for estimation. From the figure, we can deduce minimum number of points in the *FeatureVector* required for the good performance of outlier removal. Therefore we choose the size of the *FeatureVector* to be 25 points, where our outlier detection performs as good as RANSAC, while reducing the computation time. However, our outlier detector performance is tested only with relative rotation around vertical axis.

2) Testing the proposed algorithm: The performance of our proposed algorithm is compared with other existing ones. The algorithms under study are:

- *Checkerboard* algorithm represents calibration using corners detected on the checkerboard.
- *FullSift_RANSAC* algorithm represents calibration based on SIFT, using all the feature points detected and outliers removed by RANSAC.
- *FullSift* algorithm represents calibration based on SIFT, using all the feature points detected and outliers removed by our proposed method.

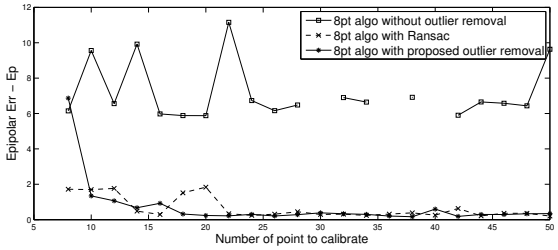


Figure 5. E_p computed for three different methods

- *newSIFTcalib / Proposed* algorithm - represents our algorithm for calibration based on SIFT, using our proposed outlier removal method and selection of stable subset (*FeatureVector*) of feature points.

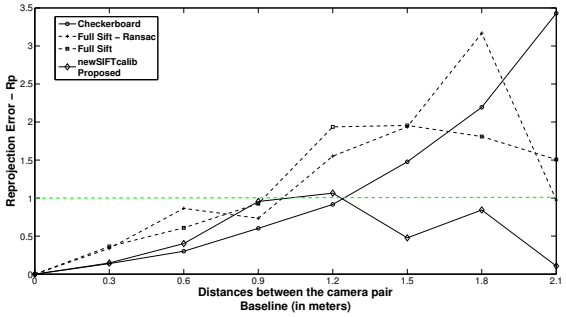


Figure 6. R_p computed for different algorithms

To evaluate the accuracy of calibration, we choose *Reprojection Error* (R_p), measured in pixels, that computes the offset between the estimated image-points using calibration parameters, with that of the measured image-points. Usually, in 3D vision applications, $R_p \leq 1$ is chosen as an acceptable re-projection error.

Given the point correspondences $\{x_1, x_2\}$ and the estimates for projection matrices P1, P2 for two cameras respectively, if we re-project estimated 3D points onto the 2D image plane - referred to as new point correspondences $\{\widetilde{x}_1, \widetilde{x}_2\}$ ($\widetilde{x}_1 = P1\hat{X}$, $\widetilde{x}_2 = P2\hat{X}$) then, re-projection error averaged over N test samples, can be computed as,

$$R_p = \frac{1}{N} \sum_{i=1}^{N} [d(x'_{1i}, \widetilde{x}'_{1i}) + d(x'_{2i}, \widetilde{x}'_{2i})] \quad (5)$$

$$d(x', \widetilde{x}') = ||(x' - \widetilde{x}')||_2 \quad (6)$$

The test result, as shown in figure 6, plots R_p against various baseline distances (in meters) between neighboring cameras. *FullSift_RANSAC* and *FullSift* perform very similarly. This verifies, as in our previous test, that our outlier

removal algorithm used in *FullSift* is as good as RANSAC method for outlier removal while being faster.

At small baselines ($\approx$ 0 - 1.2 meters), the *newSIFTcalib* algorithm performs as good as other algorithms under test, with minimal but acceptable error level of $R_p \leq 1$.

At large baselines ($\approx$ 1.2 - 2.1 meters), our *newSIFTcalib* outperforms *FullSift*, *FullSift_RANSAC* and *Checkerboard* methods. The performance of the other algorithms degrade because of the noise prone feature points, introduced due to large view-angles and baselines. On the other hand, our *newSIFTcalib* algorithm uses the *FeatureVector*, which are more stable and less prone to noise. The *newSIFTcalib* algorithm performs with high consistency at sub-pixel level and is robust to noise.

Alternatively, we compare the estimated camera pose parameters in terms of rotation angles (θ, ϕ, ψ) in 3-dimension, in comparison to the given rotation angles between cameras. Table below shows the parameters known (*Checkerboard*) and parameters estimated (*newSIFTcalib*) for different baseline distances. We can see that the estimated parameters are very close to the given values.

Camera pair	Rotation		
Baseline	θ	ϕ	ψ
0.3 (known)	3.1624	-3.1100	-3.1353
0.3 (estimate)	3.1253	-3.0839	-3.1362
1.2 (known)	3.1547	-3.1015	-3.1271
1.2 (estimate)	3.1278	-2.8736	-3.1355

Now, we evaluate the execution time. The camera pose estimation using different algorithms for cameras separated by 1.2 meters is executed and the elapsed time is measured in seconds. The performance of our *newSIFTcalib* can be reasonably measured relative to other algorithms. Figure 7 shows that our *newSIFTcalib* algorithm achieves 58.82% and 74.07% of percentage decrease in the execution time compared to the *FullSift* and the *FullSift_RANSAC*. One important thing to note is, at 1.2 meters baseline distance, the quality of *newSIFTcalib* is comparable to other algorithms (as in figure 6), while the execution time of *newSIFTcalib* has drastically reduced.

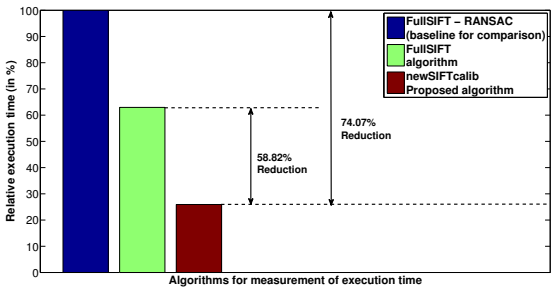


Figure 7. Execution time of various algorithms

Overall, the accuracy of our *newSIFTcalib* algorithm has been consistent at sub-pixel level over multiple baselines, while outperforming the existing algorithms, especially at large baselines. The execution time of our *newSIFTcalib* algorithm has shown a drastic reduction in comparison to other stated algorithms.

C. Operational limits

As a rule of thumb, known to SIFT users, feature detection for cameras, whose view-angle differences are more than 30°, introduces matching errors and thereby degrades the accuracy of calibration system on the whole.

which we can evaluate the performance of the algorithms under study on the operational limits.

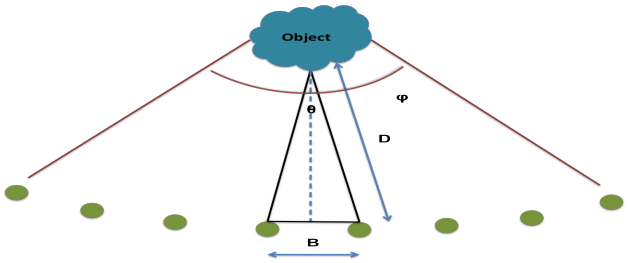


Figure 8. Deduction of relationship between object distance (D) and the baseline distance between the cameras (B)

1) Theoretical limit: Consider figure 8, where D represents the object distance from the camera, B and θ represents the baseline distance and view angle between neighboring cameras. Using triangle equations θ can be expressed as:

$$\theta = 2 * sin^{-1}(\frac{B}{2D}) \tag{7}$$

Using the condition, $\theta \leq 30°$, we have

$$2sin^{-1}(\frac{B}{2D}) \leq 30° \Rightarrow \frac{B}{2D} \leq sin(15°) \Rightarrow B \leq 0.52D$$

The relation $B \leq 0.52D$ is the theoretically defined limit for the baseline using the constraint $\theta \leq 30°$. In our dataset, the object distance is given as 4.6 meters (15 feet), and therefore the theoretically set limit for baseline would then be $\approx$2.4 meters. Let us now check the practical limit for the algorithms on the given dataset.

2) Practical limit: From results as in figure 6, the existing algorithms perform with an acceptable error ($R_p \leq 1$) only up to a baseline separation of $\approx$1 meter. Although the theoretical limit for the baseline is up to 2.4 meters, the existing algorithms practically perform well only up to $\approx$1 meter. Hence we can say that the existing algorithms are well suited for small baselines.

On the other hand, our *newSIFTcalib* algorithm extends the practical limit for the baseline up to 2.1 meters and is well suited for large baselines. The dataset used contains

stereo images separated by a maximum distance of 2.1 meters. Due to this limitation, our *newSIFTcalib* algorithm was not tested for wider baselines, however, it might fail to maintain an acceptable performance. This is merely due to the limitations posed by the SIFT feature detection for variance in view angle.

However, it is evident that our *newSIFTcalib* algorithm pushes the practical limit of the existing algorithms and reaches very close to the theoretical limit.

VI. Conclusion

In this paper, we proposed an algorithm for feature based calibration of camera pairs with application to large volume spaces such as mixed reality performances and soccer event scenarios. Our algorithm uses novel techniques for outlier removal and selection of a lower dimension feature vector consisting of stable, low noise features.

Several tests have shown that our feature based calibration algorithm performs with high consistency and accuracy even at large baselines, compared to existing algorithms. This is definitely an improvement because cameras can be widely spaced, without compromising on the calibration accuracy. Such calibration scheme can be extended to multi camera setup easily.

The execution time of our algorithm was reduced drastically and hence, can be adopted in realtime applications such as gaming, mixed / augmented reality, networked performances and is very useful for structure-from-motion applications.

Overall, our proposed algorithm has shown better performance, which makes it suitable for wide baselines of up to ≈ 2 meters, and thereby enhances the usability and scalability for multi-view capturing system in large spaces. This contribution is the first step in reaching higher accuracies in image-based rendering, especially for large volume spaces.

In our future work, we would like to work with an extensive dataset that will help us study the effects on image resolution, object distance and size, and lighting conditions on the accuracy of feature based calibration. Moreover, it is interesting and important to understand how the accuracy of calibration affects the quality of 3D representation, and thereby, image based rendering schemes.

Acknowledgement

This work is sponsored by the Norwegian Research Council under the Verdione project (project number 187828).

References

[1] B. Wilburn, N. Joshi, V. Vaish, M. Levoy, and M. Horowitz, "High-speed videography using a dense camera array," in *Proceedings of the IEEE computer society conference on Computer Vision and Pattern Recognition*, 2004, pp. 294–301.

[2] C. Zhang and T. Chen, "A self-reconfigurable camera array," in *Proceedings of the ACM SIGGRAPH 2004 Sketches*, 2004, p. 151.

[3] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330 – 1334, November 2000.

[4] R. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Journal of Robotics and Automation*, vol. 3, no. 4, pp. 323 –344, August 1987.

[5] G. Kurillo, Z. Li, and R. Bajcsy, "Wide-area external multi-camera calibration using vision graphs and virtual calibration object," in *Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras*, September 2008, pp. 1 –9.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[7] J. H. Yun and R. H. Park, *Self-Calibration with Two Views Using the Scale-Invariant Feature Transform*. Springer Berlin / Heidelberg, 2006, vol. 4291/2006, pp. 589–598.

[8] C. Li, P. Lu, and L. Ma, "A camera on-line recalibration framework using sift," *The Visual Computer: International Journal of Computer Graphics*, vol. 26, pp. 227–240, March 2010.

[9] R. Liu, H. Zhang, M. Liu, X. Xia, and T. Hu, "Stereo cameras self-calibration based on sift," *International Conference on Measuring Technology and Mechatronics Automation*, vol. 1, pp. 352–355, 2009.

[10] R. Jiayuan, W. Yigang, and D. Yun, "Study on eliminating wrong match pairs of sift," in *IEEE 10th International Conference on Signal Processing*, oct. 2010, pp. 992 –995.

[11] A. Vedaldi and B. Fulkerson, "Vlfeat – an open and portable library of computer vision algorithms," in *Proceedings of the 18th annual ACM international conference on Multimedia*, 2010.

[12] R. I. Hartley, "In defense of the eight-point algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 580–593, June 1997.

[13] R. I. Hart-ley, "Estimation of relative camera positions for uncalibrated cameras." Springer-Verlag, 1992, pp. 579–587.

[14] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[15] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM SIGGRAPH and ACM Transactions on Graphics, Los Angeles, CA*, pp. 600–608, August 2004.

Chapter 8

Paper II: Evaluating Performance of Feature Extraction Methods for Practical 3D Imaging Systems

Title: Evaluating Performance of Feature Extraction Methods for Practical 3D Imaging Systems.

Authors: Deepak Dwarakanath, Alexander Eichhorn, Pål Halvorsen and Carsten Griwodz.

Published & **Presented:** In 27th International Conference Image and Vision Computing New Zealand (IVCNZ), 2012.

Abstract: Smart cameras are extensively used for multi-view capture and 3D rendering applications. To achieve high quality, such applications are required to estimate accurate position and orientation of the cameras (called as camera calibration-pose estimation). Traditional techniques that use checkerboard or special markers, are impractical in larger spaces. Hence, feature-based calibration (auto-calibration), is necessary. Such calibration methods are carried out based on features extracted and matched between stereo pairs or multiple cameras.

Well known feature extraction methods such as SIFT (Scale Invariant Feature Transform), SURF (Speeded-Up Robust Features) and ORB (Oriented FAST and Rotated BRIEF) have been used for auto-calibration. The accuracy of auto-calibration is sensitive to the accuracy of features extracted and matched between a stereo pair or multiple cameras. In practical imaging systems, we encounter several issues such as blur, lens distortion and thermal noise that affect the accuracy of feature detectors.

In our study, we investigate the behaviour of SIFT, SURF and ORB through simulations of practical issues and evaluate their performance targeting 3D reconstruction (based on epipolar geometry of a stereo pair). Our experiments are carried out on two real-world stereo image datasets of various resolutions. Our experimental results show significant performance differences between feature extractors' performance in terms of accuracy, execution time and robustness to blur, lens distortion and thermal noise of various levels. Eventually, our study identifies suitable operating ranges that helps other researchers and developers of practical imaging solutions.

Errata 1: In section 3.3 and caption of figure 4 of this paper, the low resolution mentioned as *160x240* was typed incorrectly. This is changed to the correct value, i.e., *320x240*, in the thesis at section 4.2.1 and figure 4.5.

Errata 2: In section 5, end of bullet point 1, "...based on the resolution 320x240" in typed incorrectly. This is changed to the correct resolution value, i.e., *640x480*, in the thesis at section 4.2.6.

Evaluating Performance of Feature Extraction Methods for Practical 3D Imaging Systems

Deepak Dwarakanath[1,2], Alexander Eichhorn[1], Pål Halvorsen[1,2], Carsten Griwodz[1,2]
[1]Simula Research Laboratory, [2]University of Oslo
Oslo, Norway
{deepakd, echa, paalh, griff}@simula.no

ABSTRACT

Smart cameras are extensively used for multi-view capture and 3D rendering applications. To achieve high quality, such applications are required to estimate accurate position and orientation of the cameras (called as *camera calibration-pose estimation*). Traditional techniques that use checkerboard or special markers, are impractical in larger spaces. Hence, feature-based calibration (*auto-calibration*), is necessary. Such calibration methods are carried out based on features extracted and matched between stereo pairs or multiple cameras.

Well known feature extraction methods such as SIFT (Scale Invariant Feature Transform), SURF (Speeded-Up Robust Features) and ORB (Oriented FAST and Rotated BRIEF) have been used for auto-calibration. The accuracy of auto-calibration is sensitive to the accuracy of features extracted and matched between a stereo pair or multiple cameras. In practical imaging systems, we encounter several issues such as blur, lens distortion and thermal noise that affect the accuracy of feature detectors.

In our study, we investigate the behaviour of SIFT, SURF and ORB through simulations of practical issues and evaluate their performance targeting 3D reconstruction (based on epipolar geometry of a stereo pair). Our experiments are carried out on two real-world stereo image datasets of various resolutions. Our experimental results show significant performance differences between feature extractors' performance in terms of accuracy, execution time and robustness to blur, lens distortion and thermal noise of various levels. Eventually, our study identifies suitable operating ranges that helps other researchers and developers of practical imaging solutions.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Digitization and Image Capture-Camera calibration, Imaging geometry; I.4 [**Image Processing and Computer Vision**]: Segmentation-Edge and feature detection

IVCNZ'12, November 26 - 28 2012, Dunedin, New Zealand

Keywords

Feature detection & extraction, Stereo matching performance, SIFT, SURF, ORB, Auto calibration, 3D applications.

1. INTRODUCTION

Multi-view vision applications such as free-view rendering [3], motion tracking [11], structure from motion [16], and 3D scene reconstruction [10] require precise geometrical information about location and pose of each camera. Traditionally, camera calibration techniques use checkerboards [19, 18] or special markers [5] to find point correspondences between images. While such methods achieve sufficient accuracy they are often inconvenient and limited in practice. In some cases, it is impossible to place a measurement target like a checkerboard pattern of sufficient size in a scene. Automatically finding corresponding points based on image features alone is thus a desired goal.

To avoid dedicated calibration patterns and special markers in a scene, several auto-calibration methods have been proposed [6] [7]. They rely on matching automatically detected feature points between images from different camera perspectives to estimate geometrical calibration parameters. Feature extractors like SIFT (Scale Invariant Feature Transform) [8], SURF (Speeded-Up Robust Features) [1] and ORB (Oriented FAST and Rotated BRIEF) [15] are widely used due to their easy availability, good detection and matching performance, and a relatively small computational cost. However, little is known about their spatial accuracy and robustness to real-world distortion although these issues play a major role for precise reconstruction and scene geometry.

Sun et al. [17] have shown that the accuracy of calibration is sensitive to the quality of corresponding features. At least 7 matching points are required for an accurate estimation of calibration parameters [4] and more points will usually improve the performance.

Most feature extraction algorithms are optimized for image recognition tasks and search scenarios instead of geometry calibration. Hence, typical performance metrics such as repeatability, precision and recall or number of matches only consider the performance of matching [12], rather than the performance in terms of 3D geometry reconstruction.

Moreover, the quality of images obtained from real-world imaging sensors suffers from practical issues such as defocus and motion blur, different lens distortions, thermal noise, offsets in exposure time and white balance. Such perturbations may degrade the performance of feature extraction and matching up to a point where geometry reconstruction accuracy becomes unacceptable.

Our study provides practical insights about the robustness of existing feature extractors obtained in real-world experiments and simulations. We seek to understand the typical operation ranges of three prominent feature extraction methods; SIFT, SURF and ORB. We particularly investigate how different image distortions can impact the precision of camera pose estimation when relying on detected and matched feature points. We evaluate 3D calibration performance based on extracted image features under different levels of quality degradation. We use two real-world video data sets with a medium depth range, both captured in-doors from multiple camera perspectives. We simulate image quality degradation by introducing several levels of gaussian blur, geometrical lens distortion and sensor noise. To measure the geometrical accuracy of feature-based calibration, we use a performance metric derived from the epipolar constraint [4] which defines a precise geometrical relation between a stereo pair of images. Together with an analysis of computational costs we identify suitable operation ranges to aid researchers and developers of multi-view applications.

In our experiments, we find substantial differences in robustness and execution time between SIFT, SURF and ORB. SIFT and SURF are more robust, than ORB, to defocus, lens distortion and thermal noise. Although SURF performs similar to SIFT in terms of accuracy, SURF reduces the computational cost drastically, by almost half. Comparatively, ORB is the most computationally efficient extractor at higher resolutions and is robust to lens distortion, but accuracy is inadequate for defocused and noisy images.

2. FEATURE EXTRACTORS

In this section, we briefly explain the principle of operation of SIFT, SURF and ORB feature extractors.

SIFT detects key points in an image that are highly distinct, scale and rotation invariant, and fairly invariant to illumination. SIFT is computed as follows. First, the interesting points are searched over scale-space representation of a image, and a difference of the Gaussian function is used to identify the interesting points, which are invariant to scale and orientation. The interesting points are subjected to a 3D quadratic function to determine their location and scale. Every key-point is assigned one or more orientations depending on the direction of local gradients of the image around this key-point and a highly distinct 128-bit descriptor is computed.

SURF uses novel schemes for detection and description, which mainly focuses on reducing computational time. Integral images are computed and interesting points are obtained based on the Hessian matrix approximation. Using scale-space representation, interesting points are searched over several scales and levels. Localization: carried out using interpolation of space. This is important because number of interesting points in different layers of scales are large. The descriptor is built using the distribution of intensity content within the interesting points. SURF uses distribution of first order Haar wavelet responses in, both the x and y directions. An additional step of indexing is based on the sign of the Laplacian to increase robustness and matching speed.

ORB modifies the FAST [14] detector to detect key points by adding a fast and accurate orientation component, and uses the rotated BRIEF [2] descriptor. Corner detection using FAST is carried out and that results in N points that are sorted based on the Harris measure. A pyramid of the image is constructed, and key points are detected on every level of the pyramid. Detected corner intensity is assumed to have an offset from its center. This offset representation, as a vector, is used to compute orientation. Images are smoothened with the 31 x 31 pixel patch. Orientation of each pixel patch is then used to steer the BRIEF descriptor to obtain rotational invariance.

3. EVALUATION OVERVIEW

3.1 Simulation parameters

All imaging systems encounter practical issues such as defocus, radial lens distortion and thermal noise. *Image blur* is the loss of image sharpness caused due to defocus, shallow depth of field and motion of the camera or the scene objects and quantization process. In our study, we focus on image blur due to defocus only, because we consider multi view capture using only stationary cameras and hence motion blur is of lesser significance. *Radial lens distortion* is an optical aberration caused by spherical lens surfaces of the cameras, which produces aberrations symmetrically and radially from the image center. Barrel and pincushion are the types of radial distortions where the image aberration increases and decreases respectively as the radial distance from image center increases. *Image thermal noise* appears as random speckles in an image which is random variation in the luminosity or color information of the pixels caused by the camera sensor and its circuitry. To study the performance of the feature extractors under such practical scenarios, we simulate defocus, lens distortion and noise using the mathematical models.

Defocus $I_b(u,v)$ is accomplished by smoothing an image $I(u,v)$ with a linear 2D Gaussian filter $G(u,v)$, as in equation 1. Various defocus levels can be controlled by the variance σ_b of the Gaussian kernel, which represents blur radius.

$$I_b(u,v) = I(u,v) * G(u,v) \tag{1}$$

$$G(u,v) = \frac{1}{2\pi\sigma_b^2} e^{-\frac{u^2+v^2}{2\sigma_b^2}} \tag{2}$$

Lens distortion can be modeled as a 3rd order polynomial, as given by equation 3, where R_u and R_d is undistorted and distorted pixel radius, respectively. The distortion co-efficient k_1 can be varied to obtain various levels of distortion.

$$R_u = R_d + k_1 R_d^3 \tag{3}$$

Thermal noise is modeled as Gaussian distribution. A noisy image $I_n(u,v)$ is obtained by adding Gaussian random noise $N(u,v)$ with zero mean and variance σ_n to an image $I(u,v)$, as in equation 4. To obtain various noise levels N_l, measured in decibels, the variance σ_n is controlled as, $\sigma_n = 10^{N_l/10}$.

$$I_n(u,v) = I(u,v) + N(u,v) \tag{4}$$

3.2 Performance measure

The performance of feature extractors are measured in terms of accuracy, detectability and execution time.

Accuracy of feature extraction in stereo images is measured by deviations of measured positions of matched feature points from their ideal positions. To explain this in detail, we bring in the concept of *epipolar geometry*. Researchers [4] [9] [13] have shown that in 3D imaging systems,

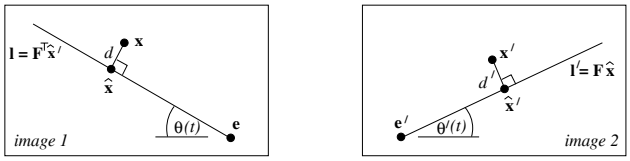


Figure 1: Illustration of Epipolar Geometry. Courtesy R. I. Hartley [4]

the geometrical relationship between the point correspondences between stereo images is important and is characterized by a mapping matrix called *Fundamental Matrix* (F).

The epipolar geometry is illustrated in figure 1. Ideally, for every point in one of the stereo images (say $\hat{x}$), a corresponding point on the other stereo image ($\hat{x'}$) should lie on a line, called *epipolar line* (l'), which is computed using the matrix (F). In practice, feature extractors estimate the corresponding point (x'), which can lie outside the line and thus producing an error (d'). Such an error averaged over all N_p feature points will be referred as *Epipolar Error* (E_p), and can be computed as in equation 5. Thus, the *Epipolar Error* aids in measuring the accuracy of feature extractors, in pixels. The sub-pixel errors, that is $Ep < 1$ pixel, is an acceptable value for good performance in most of the relevant applications.

$$E_p = \sum_{i=1}^{N_p} \frac{x_i' F x_i}{(Fx_i)_1^2 + (Fx_i)_2^2 + (F^T x\prime_i)_1^2 + (F^T x\prime_i)_2^2} \qquad (5)$$

Detectability measures the ability to obtain sufficient feature point correspondences in stereo images. A good estimation of Fundamental Matrix requires at least 7 feature corresponding points in stereo images [4]. Therefore, the percentage of trials resulting in at least 7 feature correspondences represents the detectability of a feature extractor.

Execution time measures the computational speed of the feature extractors. It is computed as time spent on the extraction step (detecting interesting points in two images and building descriptors for them) and the matching step (performing feature matching to obtain feature correspondences).

3.3 Simulation Setup

Our experimental setup, as illustrated in figure 2, comprises a database of the test stereo images, an image degradation module and a feature extraction and matching module. During our evaluation, stereo images are retrieved from the database, and the image degradation module pre-transforms the stereo images to simulate defocus, lens distortion and sensor noise, with various levels using a tuner. Then, the feature detector-descriptor-matcher operates over all stereo images that are pre-transformed. The resulting feature matches on degraded images are used to evaluate the performance of the feature extractor based on the fundamental matrix estimated for the stereo images before degradation.

In our experiments, we have used 30 stereo images from the dataset of an opera performance, captured using 8 cameras (2 camera arrays, each consisting of 4 cameras of narrow and wide angle lens respectively). A second dataset used for evaluation contains 35 images, from the popular breakdance

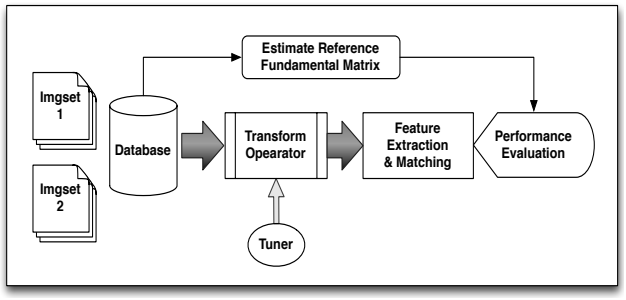


Figure 2: Evaluation Pipeline

video sequence from Microsoft [20]. The stereo images from both datasets were of HD resolution (1280x768). All these stereo images were scaled to high resolution (1280x960), medium resolution (640x480) and low resolution (160x120) images to study the behavior of feature extractors across various resolutions in conjunction to image degradation. Image degradation was carried out at different levels on every test stereo pair (equally on both images of a stereo pair). Blur radius levels ranged from values 1.5-6.0. Barrel distortion and pincushion distortion were varied as -50% to -10% and +10% to +50% respectively. Thermal noise levels were 5 - 50dB. Then, feature extraction and matching using SIFT, SURF and ORB methods are performed and the performance is evaluated. An example of feature extraction in stereo images for various datasets and the image degradation using simulation parameters are shown in figure 3.

4. EVALUATION RESULTS

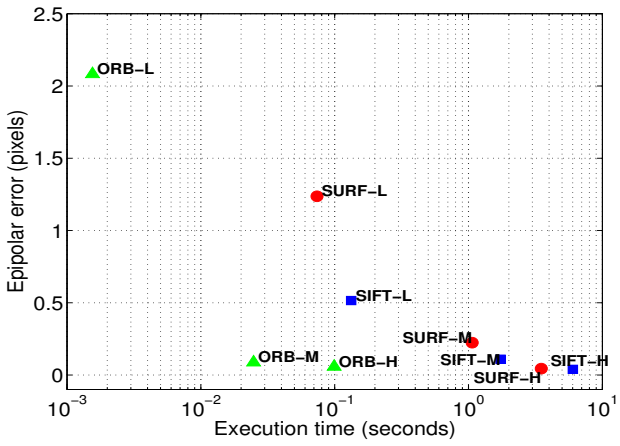
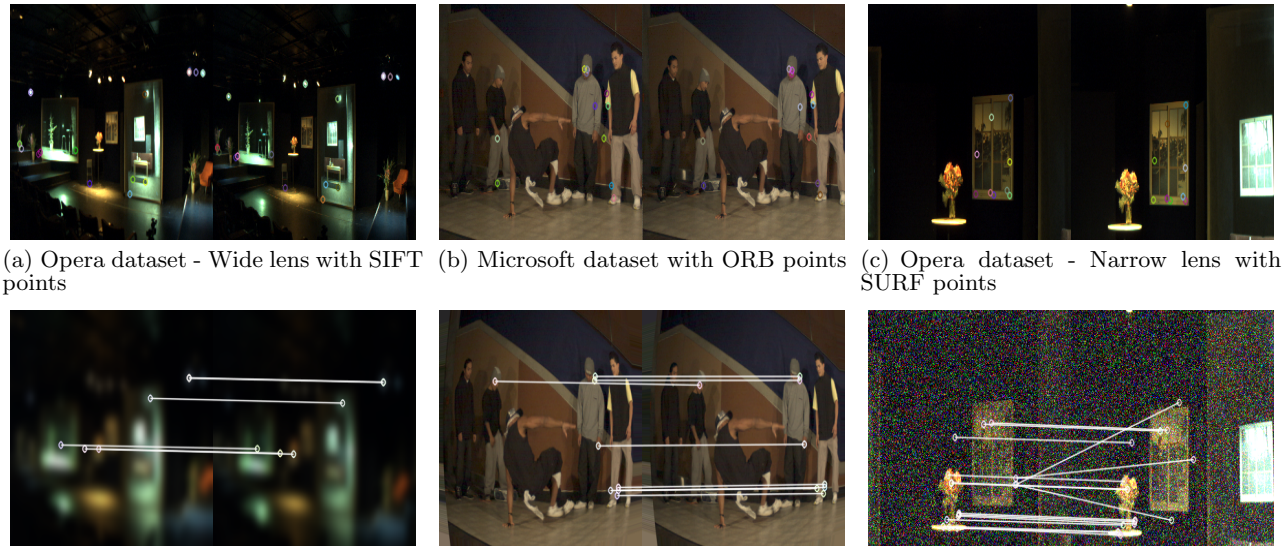


Figure 4: Accuracy Vs Computational time. L- low (160x120) resolution, M- medium resolution (640x480), H- high resolution (1280x960)

First, we ran the tests to measure accuracy and execution time of various feature extractors to comparatively analyze the performance of feature extractors at various image resolutions. Figure 4 shows the results of the test (note that the execution time is plotted in logarithmic scale). Obviously, a tradeoff exists in choosing feature extractors between achiev-



(a) Opera dataset - Wide lens with SIFT points (b) Microsoft dataset with ORB points (c) Opera dataset - Narrow lens with SURF points

(d) SIFT matching on blurred images (e) ORB matching on distorted images (f) SURF matching on noisy images

Figure 3: Stereo images from various datasets of resolution 320x240

ing higher accuracy and higher speed. Overall, ORB is computationally efficient compared to SIFT and SURF at all resolutions. A relative difference in execution time between SIFT and SURF is significant; SURF reduces the computational cost by 48% at all resolutions. SIFT, SURF and ORB results in acceptable (sub-pixel) accuracies, except for SURF-L and ORB-L. This shows that SIFT is more robust to change in scale.

Next, we conducted experiment to discuss how the defocus/image blur, lens distortion and thermal noise affects the performance of feature extractors, and the results are shown in figure 5.

4.1 Effects of blur variation

Figures 5(a), 5(b) and 5(c) show that SIFT outperforms in terms of accuracy at all resolutions. SIFT seems to be robust to blur levels probably because of its own way of finding key points, which uses scale space representation with various blur levels. SIFT operations on blurry images are equivalent to having more levels of blurs in every octave of the scale space, for an un-blurry image. Obviously, at lower resolutions blur-ness has a greater effect and hence SIFT shows an acceptable accuracy up to blur level 4.5, as in figure 5(c).

SURF performs marginally at acceptable accuracy ($E_p <= 1$, figure 5(c)) up to blur level 4.5, at low resolutions, for the same reasons mentioned for SIFT. However, the difference in accuracies between SURF and SIFT is due to the descriptor construction. SURF integrates the gradient information and loses distinctiveness when blur increases, while SIFT uses individual gradient to create the descriptor and sustains the performance to a larger extent of blur, compared to SURF.

The detectability measure (figure 5(d)) for both SIFT and SURF reduced drastically with increase in blur level at low resolution, which makes them unsuitable to use when low resolution images are blurred, especially at levels > 4.5.

Although, ORB performs good only at medium and high resolution (figures 5(b) and 5(a)) up to blur level 3.5, the detectability of ORB decreases rapidly with increase in blur level. The use of huge box filters in ORB to obtain descriptors seems to limit performance on blurry images. Additional blur worsens the efficiency of the descriptor. Hence ORB fails at low resolutions.

4.2 Effects of distortion variation

The effects on performance of the feature extractors due to different levels of barrel and pincushion distortion can be seen in figures 5(e), 5(f), 5(g) and 5(h). All the feature extractors perform well and similar at high and medium resolution. A at low resolutions, SIFT outperforms SURF, which in turn outperforms ORB; however, all of them exhibit an acceptable accuracy and a constant detectabitlity. Overall performance of SIFT, SURF and ORB at all resolution andseems to be unaffected by lens distortion. It should be noted that this result is for a homogenous stereo pair where the distortions are assumed to be of same degree in both the cameras.

4.3 Effects of noise variation

The measurements for this experiment peaked at around 10 pixels, hence the results are shown in log scale for y axis in figures 5(i), 5(j) and 5(k). Here, we show that SIFT outperforms SURF and ORB, at all resolutions and exhibited resilience to thermal noise, but becomes sensitive to noise at around 15dB for low resolution images. SURF and ORB showed resilience to noise up to 20dB and 15dB, respectively, at both high and medium resolutions. Importantly, we observe high and constant detectability rate (figure 5(l)) for SURF and ORB, suggesting that the performance of SURF and ORB are not affected by noise, but the accuracy is too low ($E_p > 10$ pixels). This behavior is because SURF and ORB detect more features which are not supposed to be, in noisy images. Hence under noisy conditions, above 15dB none of the feature extractors perform within the acceptable accuracy.

5. CONCLUSION

In this paper, we evaluated the popular and widely used feature extractors SIFT, SURF and ORB. The experiments were conducted over different datasets at various resolutions to test the resiliency of the feature extractors to defocus/blur, lens distortion and thermal noise. From the results, we can conclude that:

- At resolutions > 320x240, SIFT and SURF are the best choices. However, choosing SURF would save execution time of 48%, on an average, with a cost of around 0.10 pixels in accuracy. A choice of feature extractor should be made considering the below conclusions, which are based on the resolution 320x240.
- For blurry images, SIFT is the best choice. However, using SURF would save 48%, on an average with a cost of 0.22 pixels in accuracy.
- For lens distorted images, SIFT, SURF and ORB all are good choices. By using ORB, the execution time reduces by 98.12% and 95.27% with a cost of 0.69 pixels and 0.33 pixels in accuracy compared to SIFT and SURF, respectively.
- For noisy images, SIFT and SURF are good choice and using SURF saves 32% time with a cost of 0.67 pixels in accuracy.

Unlike other feature evaluations, we have used the Epipolar Error to measure the accuracy of the feature correspondence, which aids to selection of feature extractors for feature based calibration and other 3D applications.

6. REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110:346–359, June 2008.

[2] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary Robust Independent Elementary Features. In *European Conference on Computer Vision*, September 2010.

[3] M. Dongbo et al. 2d/3d freeview video generation for 3dtv system. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1760 –1763, October 2008.

[4] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[5] G. Kurillo et al. Wide-area external multi-camera calibration using vision graphs and virtual calibration object. In *Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras*, pages 1 –9, September 2008.

[6] C. Li et al. A camera on-line recalibration framework using sift. *The Visual Computer: International Journal of Computer Graphics*, 26:227–240, March 2010.

[7] R. Liu et al. Stereo cameras self-calibration based on sift. *International Conference on Measuring Technology and Mechatronics Automation*, 1:352–355, 2009.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[9] Y. Ma et al. *An Invitation to 3D Vision: From Images to Geometric Models*. Springer Verlag, 2003.

[10] W. Matusik et al. Image-based visual hulls. In *Proceedings of the Conference on Computer Graphics and Interactive Techniques*, pages 369–374, 2000.

[11] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, March 2001.

[12] P. Moreels and P. Perona. Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, 73:263–284, July 2007.

[13] F. Olivier. Three-Dimensional Computer Vision – A Geometric Viewpoint. *MIT Press*, 1996.

[14] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, volume 1, pages 430–443, May 2006.

[15] E. Rublee at al. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2564 –2571, nov. 2011.

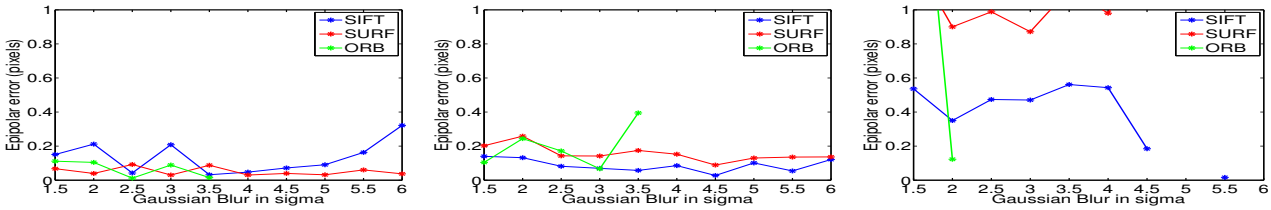
[16] A. Sameer et al. Building rome in a day. In *International Conference on Computer Vision*, pages 72–79, 2009.

[17] W. Sun and J. Cooperstock. An empirical evaluation of factors influencing camera calibration accuracy using three publicly available techniques. *Machine Vision and Applications*, 17:51–67, 2006.

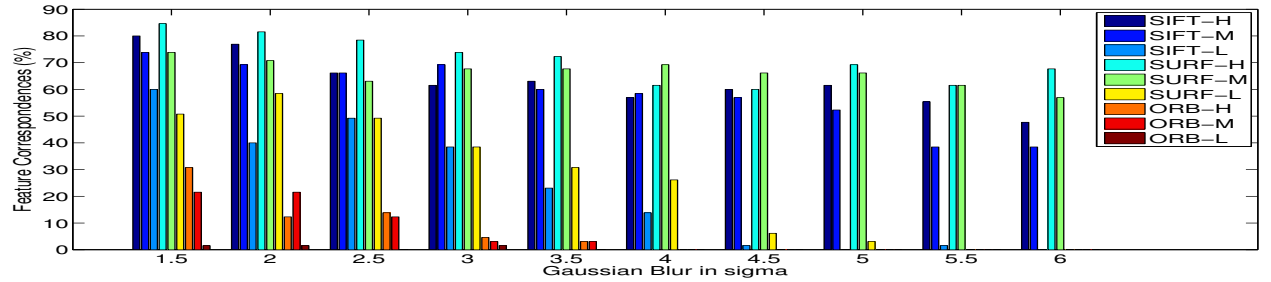
[18] R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323 –344, August 1987.

[19] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330 – 1334, November 2000.

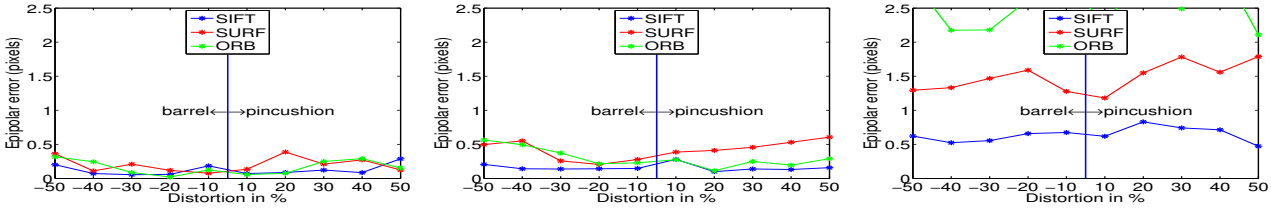
[20] C. Zitnick et al. High-quality video view interpolation using a layered representation. *ACM SIGGRAPH and ACM Transactions on Graphics, Los Angeles, CA*, pages 600–608, August 2004.



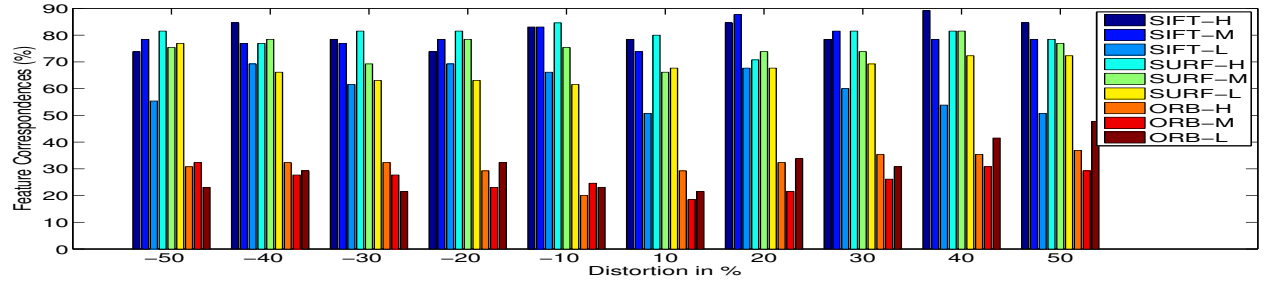
(a) Blur at High Resolution (b) Blur at Medium Resolution (c) Blur at Low Resolution



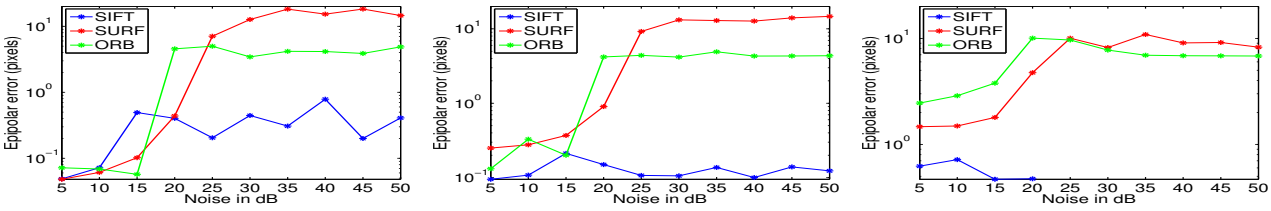
(d) Detectability for blurred images



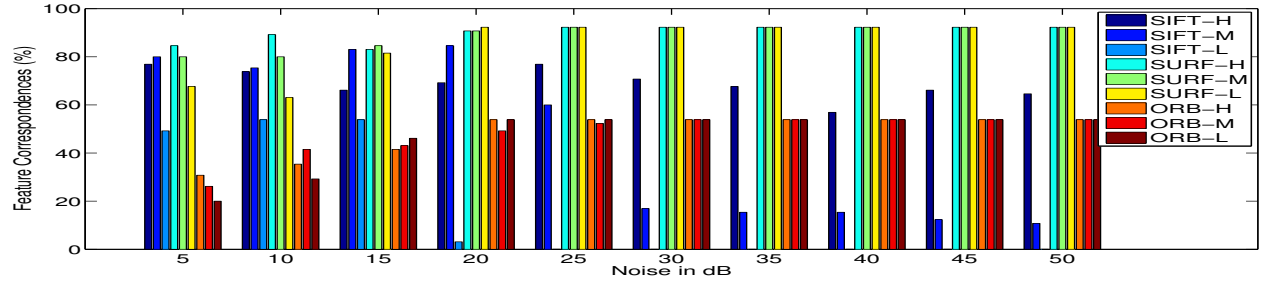
(e) Distortion at High Resolution (f) Distortion at Medium Resolution (g) Distortion at Low Resolution



(h) Detectability for Distorted images



(i) Noise at High Resolution (j) Noise at Medium Resolution (k) Noise at Low Resolution



(l) Detectability for Noisy images

Figure 5: Performance of feature extractors for simulation of blur, distortion and noise levels over various resolutions

Chapter 9

Paper III: Study the Effects of Camera Misalignment on 3D Measurements for Efficient Design of Vision-Based Inspection Systems

Title: Study the Effects of Camera Misalignment on 3D Measurements for Efficient Design of Vision-Based Inspection Systems.

Authors: Deepak Dwarakanath, Carsten Griwodz, Pål Halvorsen and Jacob Lildballe.

Published & **Presented:** In 8th Hellenic Conference on Artificial Intelligence (SETN), 2014.

Abstract: Vision based inspection systems for 3D measurements using single camera, are extensively used in several industries, today. Due to transportation and/or servicing of these systems, the camera in this system is prone to mis-alignment from its original position. In such situations, although a high quality calibration exists, the accuracy of 3D measurement is affected. In this paper, we propose a statistical tool or methodology which involves. a) Studying the significance of the effects of 3d measurements errors due to camera mis-alignment. b) Modelling the error data using regression models. c) Deducing expressions to determine tolerances of camera mis-alignment for an acceptable inaccuracy of the system. This tool can be used by any 3D measuring system using single camera. Resulting tolerances can be directly used for mechanical design of camera placement in the vision based inspection systems.

Errata 1: In table 2 of this paper, the row values with headings *width,* r_x^- and *depth,* r_x^- had been interchanged. This is corrected in the thesis, in table 3.2.

Errata 2: In equation 1 of this paper, scaling factor $\frac{1}{Z_c}$ is missing. This is corrected in the thesis, in equation 3.1.

Study the Effects of Camera Mis-alignment on 3D Measurements for Efficient Design of Vision-based Inspection Systems

Deepak Dwarakanath[1 2 3], Carsten Griwodz[1 2],
Paal Halvorsen[1 2], and Jacob Lildballe[3]

[1] Simula Research Laboratory
Martin Linges vei 17/25, Lysaker 1325, Norway
{deepakd}@simula.no
[2] University of Oslo
Gaustadallen 23 B, N-0373 Oslo, Norway
{griff,paal}@ifi.uio.no
[3] Image House PantoInspect A/S
Carsten Niebuhrs Gade 10, 2, DK-1577 Copenhagen V, Denmark
jl@pantoinspect.com

Abstract. Vision based inspection systems for 3D measurements using single camera, are extensively used in several industries, today. Due to transportation and/or servicing of these systems, the camera in this system is prone to mis-alignment from its original position. In such situations, although a high quality calibration exists, the accuracy of 3D measurement is affected. In this paper, we propose a statistical tool or methodology which involves. a) Studying the significance of the effects of 3d measurements errors due to camera mis-alignment. b) Modelling the error data using regression models. c) Deducing expressions to determine tolerances of camera mis-alignment for an acceptable inaccuracy of the system. This tool can be used by any 3D measuring system using single camera. Resulting tolerances can be directly used for mechanical design of camera placement in the vision based inspection systems.

Keywords: Camera calibration, Vision based inspection systems, Camera mis-alignment and Regression models.

1 Introduction

With the advent of automation in all types of industries, manual intervention in the operations of machines is minimized. Nowadays, automatic inspection systems are used to inspect various types of faults or defects in several application areas such as sorting and quality improvements in food industry [11], [10], inspection of cracks in roads [4], crack detection of mechanical units in manufacturing industries [8], [9] and so on. Vision based inspection systems are increasingly growing with the advance in computer vision techniques and algorithms.

Typically, vision based inspection systems that inspect objects of interest and estimate measurements, are required to know a priori information about the intrinsic (focal length, principal axes) and the extrinsic (position and orientation) parameters of the camera without any freedom of scale. These parameters are obtained by a camera calibration process [3],[5]. Usually, calibration is carried out offline, i.e., before the system is deployed and thereafter the calibrated parameters are used to recover 3D measurements from the 2D image of the camera [12], [13] [14]. The quality of the camera calibration is an important factor that determines the accuracy of the inspection system.

Although the quality of calibration might be very high, it is difficult to guarantee highly accurate measurements, if the camera is physically mis-aligned from the position assumed during calibration. However, the transportation or installation can cause mis-alignment, e.g., due to wrong mounting during installation, due to ways of handling the system during maintenance or service etc. Consequently, the performance of the inspection system degrades.

A possible correction to this problem would be to re-position the camera, physically, to its calibrated position or to run the calibration process after deployment. It is very difficult to physically re-position the camera with high precision. Alternatively, it might also be difficult to recalibrate in some situations based on the location and accessibility of the installed system.

Therefore, it becomes important to understand the effects of the offset in cameras' position and orientation on inaccuracies. The significance of the inaccuracies depends on design (acceptable inaccuracy level) and the application of the system. So, an important question is: what is the maximum tolerable camera mis-alignment for an acceptable inaccuracy of the system? By answering this question, we will be able to design and operate the system better. When the tolerance limits of the camera mis-alignment are known, the mechanical design of the camera housing and fixtures will need to adhere to these tolerances to maintain the inaccuracy below an acceptable level. Also, by using an empirical model, it is possible to estimate the camera mis-alignment and further re-calibrate the camera parameters to increase the robustness of the system.

This paper aims to enhance the design and operational aspects of vision-based inspection systems. The main contribution of this paper is to provide a simple statistical method or tool which can compute acceptable tolerance values for positions and orientations in all directions for a given accuracy requirements. This tool is useful in designing the mechanics and in increasing the robustness of the vision based inspection system. It is easily implementable and reproducible. The limitation of this tool is that the measurements are carried out on points that are assumed to be lying on a plane. However, the tool is easily extendable to measure 3D points as long as an appropriate calibration process is carried out based on known 3D points. Related work is described in section 2.

First, we identify a suitable use case for the study of effects of camera mis-alignment on 3D measurements. One such vision based inspection system that exhibits a similar purpose and problems mentioned so far, is the PantoInspect system [2]. This system is explained in detail in section 3. Details of our exper-

imental design is explained in section 4. The simulation results and the empirically obtained regression model is explained in section 5. Finally the paper is concluded by summarising the goal and evidence of the paper.

2 Related work

The effects of mis-alignment of stereoscopic cameras are studied in [15], [16], however, in our case we study the effects due to mis-alignment of single cameras. [15] focusses on the effects of calibration errors on depth errors, and provides tolerances on calibration parameters. In [16], camera mis-alignment is estimated and corrected. In both the papers, the approaches rely strongly on a second image and errors of the cameras' orientation with respect to each other. Other papers only discuss effects of camera mis-alignment on calibration parameters itself [15], [17] and [18]. In our case, where we use a single camera, we assume that calibration is of sufficiently high quality, but once calibrated, the effects of camera mis-alignment due to certain factors requires more attention in practical systems and hence, we study this in our paper. Our approach leads to an estimation of tolerances for camera mis-alignment that aims directly at the mechanical design of single camera vision systems. One major feature of our approach is that it is not specific to one application, but can be used for any application of this type.

3 The PantoInspect System

PantoInspect is a fault inspection system, which inspects pantographs and measures the dimensions of the defects in their carbon strips. PantoInspect is installed, as shown in figure 1, over railway tracks to inspect trains running with electric locomotives that are equipped with pantographs. Pantographs are mechanical components placed on one or more wagons of the train, which can be raised in height so that they touch the contact wire for electricity. Pantographs have one or more carbon strips that are actually in contact with the wire. Over time, due to constant contact of carbon strips with the wire, and probably other factors, various types of defects (cracks, edge chips etc.) are seen. Such defects are detected by the PantoInspect system.

3.1 Principle

PantoInspect is mounted right above the train tracks on bridges or other fixtures. The PantoInspect system receives a notification when the train is approaching and prepares itself. When the train passes right below the system, three line lasers are projected onto the carbon strips (depicted as green line in the figure), and the camera captures the near infrared image of the laser. When defects are present, the line deforms instead of remaining a straight line in the image. Hence, the laser line defines the geometry of the defect. The system then analyses the images, measures the dimension of the defects and notifies the user with alarms on certain measurement thresholds.

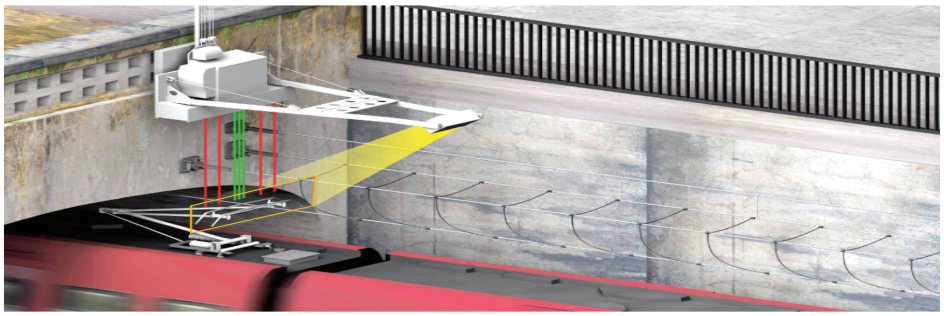


Fig. 1. PantoInspect system: inspects defects on the pantographs mounted on the trains.

The system measures various defects in the carbon strip based on the captured images. These defects are represented in figure 2, which are *(1)-thickness of carbon wear, (2)-vertical carbon cracks, (3)-carbon edge chips, (4)-missing carbon and (5)-abnormal carbon wear*. In general, all these defects are measured in terms of width and/or depth in real world metrics. Although the PantoInspect system measures various types of defects in pantographs, the common attribute in these measurements are width and depth. We therefore consider these attributes as the main 3D measurements in our scope of simulation and study of the effects of camera mis-alignment, in section 4.

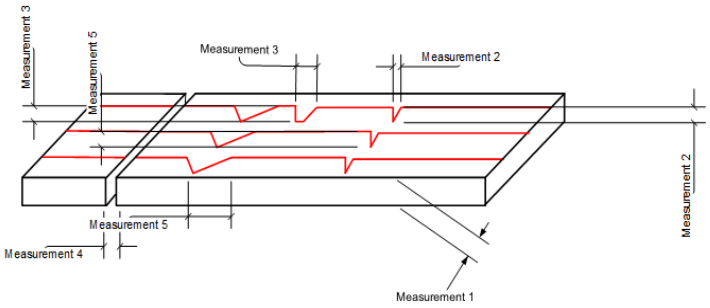


Fig. 2. Different carbon defects and the laser line deformations.

3.2 Calibration:

The system uses 2D pixel measurements in the image and estimates the real-world 3D scales. Camera calibration is an important step in obtaining such 3D measurements. For PantoInspect, this is carried out in the factory before

installing the system, using Bouguet's method [1]. A number of checkerboard images are used to estimate the intrinsic parameter K of the camera that constitutes focal length and principle axes of the camera. Next, a single image of the checkerboard that is placed exactly on the laser plane, is used to estimate the extrinsic parameter of the camera - position T and orientation R, with respect to the checkerboard coordinates.

3.3 Homography

In the scenario of PantoInspect system, we consider an imaginary plane passing vertically through the line laser as in figure 3. Then, the points representing defects are lying on a laser plane. These 2D points of the defects in the image are detected, and the conversion from 2D (p,q) to 3D (X,Y,Z) points becomes merely a ray-plane intersection [6], as shown in equation 1, where 3D and 2D points are expressed in homogeneous coordinates.

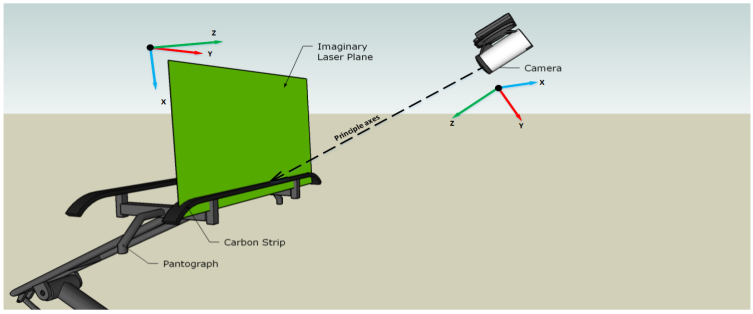


Fig. 3. Inspection scenario: world coordinates (Z=towards camera, Y=horizontal, X=vertical) and camera coordinates (Z=towards plane, Y=vertical, X=horizontal).

$$\begin{bmatrix} p \\ q \\ 1 \end{bmatrix} = K[R|T] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{1}$$

The K, R and T are obtained from the calibration process. The R matrix and the T vector are represented with their components in equation 2. Since 3D points are lying on the plane, the Z axis is zero. The rotation components in the 3rd column (r_{13}, r_{23}, r_{33}) are ignored because they are multiplied by zero.

$$\begin{bmatrix} p \\ q \\ 1 \end{bmatrix} = K * \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = \underbrace{K * \begin{bmatrix} r_{11} & r_{12} & t_1 \\ r_{21} & r_{22} & t_2 \\ r_{31} & r_{32} & t_3 \end{bmatrix}}_{H} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{2}$$

Equation 2 describes a 2D-2D mapping between points on the image and points on the laser plane. This mapping is a homography (H). Using the homography, points on the plane can be recovered and measured for width and depth of defects that corresponds to defects detected in 2D pixel points.

4 Study methodology

We have seen how the camera parameters play an important role in estimating the measurements in PantoInspect. However when the camera is mis-aligned from its original position, estimated 3D measurements incur inaccuracies in the performance of the system. To study the effects of camera mis-alignment on 3D measurements, we carry out a simulation of the PantoInspect image analysis for 3D measurements, under the conditions of camera mis-alignment.

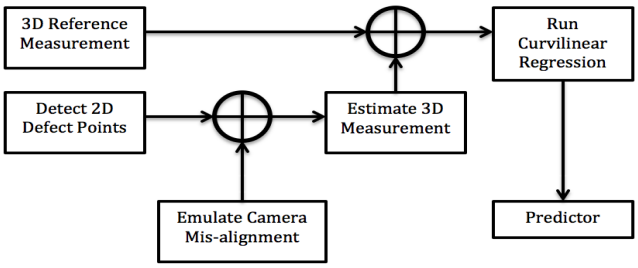


Fig. 4. Simulation procedure.

For repeatability of this simulation in any application involving 3D measurements of points lying on a plane and a single camera, a general procedure is shown in figure 4, followed by a specific explanation of the procedure for our case study.

4.1 Error computation

A set of points (P_w) that represents the crack edges on a plane are synthesized in the world coordinates. Note that the points are on the plane. Hence the Z axis is 0 for all points. These 3D points are synthesised using a random number generator. From these points, the width (W_{known}) and depth (D_{known}) of cracks are computed and recorded. These known measurements in 3D space are our baseline and used as a reference to evaluate the accuracy of the inspection.

The projection of the known set of points, P_w are computed based on the known camera parameters (K, R and T). These points represent the 2D points

(P_i) in image coordinates that are detected and further analysed by the PantoInspect system.

Typically, when the camera stays perfectly positioned and oriented, the width and depth of the cracks are measured with a reasonably good accuracy, due to high quality camera calibration process. To study the effects of camera mis-alignment on the accuracy of the measurements, the camera mis-alignment process needs to be emulated as if the camera had shifted position or orientation. Accordingly, points (P_i) are first represented in the camera coordinate system as P_{cam}, as depicted in equation 3. Next ,the rotation or translation effects are introduced, as a result of which the detected points obtain new positions, represented as $P_i^{misalign}$ in the image coordinates. Due to this emulation process that is based on changed camera orientation (R_{cam})and position (T_{cam}), the $P_i^{misalign}$ is estimated as in equation 4.

During inspection, the PantoInspect system detects measurable points (edges) of the cracks in the image and back-projects the 2D points into the 3D plane. The estimation of 3D points (P_w^{est}) of the crack is based on a pin-hole camera model and is mathematically shown in equation 5, where homography is a plane-plane projective transformation [6] as in equation 2.

$$P_{cam} = K^{-1} * P_i \tag{3}$$

$$P_i^{misalign} = K * \begin{bmatrix} R_{cam} & T_{cam} \\ 0^T & 0 \end{bmatrix} P_{cam} \tag{4}$$

$$P_w^{est} = H * P_i^{misalign} \tag{5}$$

Finally, the width (W^{est}) and depth (D^{est}) measurements are estimated and compared with the known values to compute the mean squared error, in equations 6 and 7. These errors $Error_{width}$ and $Error_{depth}$ represent the accuracy of the defect measurements.

$$Error_{width} = ||W - W^{est}||_2 \tag{6}$$

$$Error_{depth} = ||D - D^{est}||_2 \tag{7}$$

4.2 Prediction model

The simulation produces data pertaining to error in the 3D measurements with respect to camera mis-alignment in terms of three positional ($T_{cam} = [t_x, t_y, t_z]$) and three rotational ($R_{cam} = [r_x, r_y, r_z]$) mis-alignments. Considering each of these camera mis-alignment components as a variable, and the error as the response to it, the error can be modelled using appropriate regression models. Once the data fits to a model, the parameters of that model can be used for prediction purposes [7].

This is helpful to make predictions of camera mis-alignment based on the error estimated in the system. Then, given the acceptable accuracy of the system,

in terms of maximum allowable error in the measurements, one can deduce maximum limits or tolerances of camera mis-alignment to maintain an acceptable inaccuracy.

5 Simulation results

5.1 Priori

The carbon strip on each pantograph measures about 1.2meters in length and between 30-50mm in width and 30mm in thickness. For simulation purposes, we assume that there are about five defects per pantograph, and the system inspects about 200 such pantograph, i.e., 1000 measurements.

The defect width of maximum 50mm and defect depth of maximum 30mm are assumed to be present across the length of the carbon strip. The camera used for inspection is calibrated offline, and hence, a priori calibration data is available for that camera. The K, R and T matrices are as follows:

$$K = \begin{bmatrix} 4100.8633085 & 0 & 947.0315701 \\ 0 & 4104.1593558 & 554.2504842 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R = \begin{bmatrix} 0.0108693 & 0.9999407 & -0.0006319 \\ 0.7647318 & -0.0079055 & 0.6443002 \\ 0.6442570 & -0.0074863 & -0.7647724 \end{bmatrix}$$

$$T = \begin{bmatrix} -540.7246414 & -119.4815451 & 2787.2170789 \end{bmatrix}$$

5.2 Effect of camera mis-alignment

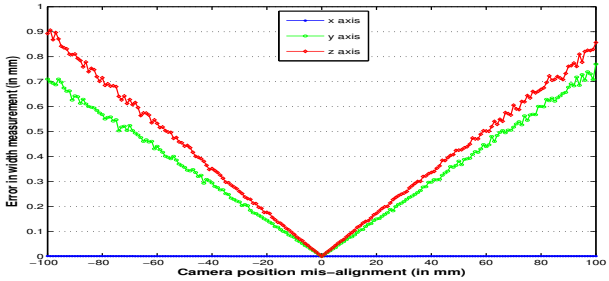
The simulation procedure explained in section 4 is for one camera-plane pair, where a single camera calibration parameter (K, R, T, as given above) is used to recover the 3D measurements. We have conducted experiments on 6 such pairs. We used two different cameras, and each camera calibration with three planes corresponding to three line lasers. Results from all the 6 configurations yields similar patterns and are explained as follows.

For every such configuration, the simulation was carried over a range of camera's positional mis-alignment between -100mm to +100mm and orientational mis-alignment between -40 and +40 degrees. For every new position and/or orientation of mis-alignment, the simulation was carried out for 1000 measurements each.

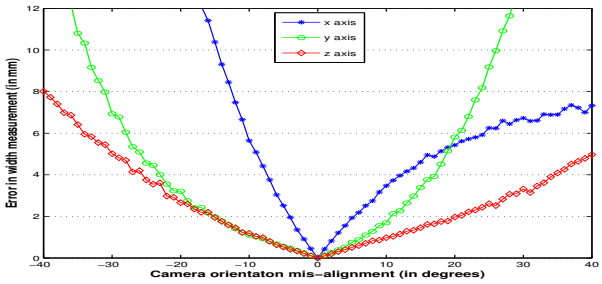
The results of the simulation as in figure 5 and figure 6, show the variation in mean squared error of both the width and depth measurements for every camera mis-aligned position and orientation. This error represents the ability of the system to measure the inspected data accurately and error is measured in millimetres.

From figures 5(a) and 6(a), it can be seen that the camera translation t_x has the least effect on the errors compared to translations t_y and t_z. For insight into

the camera axes for translation and rotation please refer figure 3. The error for translations in t_y and t_z is higher, however, not significantly higher than 1mm, which might be an acceptable inaccuracy limit for certain applications. These effects are caused by the camera position mis-alignments, which shifts the back projected points, defining the width and depth measurements proportionally, so the relative width and depth measurements remain almost unchanged.



(a) Width error Vs camera position t_x, t_y, t_z

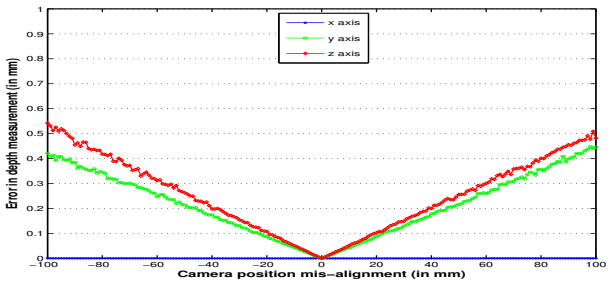


(b) Width error Vs camera orientation r_x, r_y, r_z

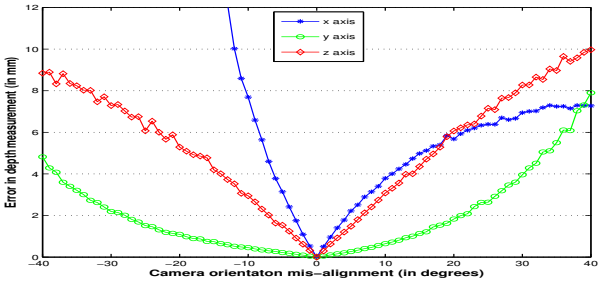
Fig. 5. Variation of error in 3D measurements (width) of the defects, due to changes in camera position and orientation about its camera centre.

Interesting effects are seen due to camera rotation, which has slightly different effects on width and depth. From figures 5(b) and 6(b), it can be seen that the camera rotations r_y and r_z, has noticeable effects on the width and depth errors. When the camera is rotated around axes y and z, the resulting 2D image point moves symmetrically within the image. Furthermore, the rate of increase of width is higher than depth for r_y, because when camera is rotated around the y axis, the horizontal component of the 2D point is changed more than the y component and width is a function of the x component. Exactly the opposite is seen when depth increases at higher rate for r_z, because depth is a function of the y component.

Special cases are the errors due to r_x. Remember that the camera is placed in a position to look down at the laser lines. The rotation around x axis will have a drastic projective effects in the image plane. The projective properties result in a non-symmetric variation of the errors around zero. One more thing to notice is that the error increases very quickly on the negative r_x than positive side. This behaviour can be explained using projective geometric properties. Consider an image capturing parallel lines and in perspective view, the parallel lines meet at vanishing (imaginary) point. It is possible to imagine that the width of parallel lines is shorter when the capturing device tilts downwards. Similarly when our camera is tilted downwards i.e. r_x in the positive (clockwise) direction, the defect points are moved upwards in the image plane, and the measurement becomes so small that the error seems to be constant. Contrarily, when the camera is tilted upwards, the detected points are moved downwards in the image plane, increasing the measurements and thereby the error.



(a) Depth error Vs camera position t_x, t_y, t_z



(b) Depth error Vs camera orientation r_x, r_y, r_z

Fig. 6. Variation of error in 3D measurements (depth) of the defects, due to changes in camera position and orientation about its camera centre.

5.3 Regression

By visual inspection of figures 5 and 6, we can say that errors are linearly varying with camera translations (t_x, t_y, t_z), and non-linear with camera rotations (r_x,

r_y, r_z). We not only model the data for every rotation and translation but also their direction (positive(+) and negative(-)). This means we separate out the error data for variables r_x^+, r_x^-, r_y^+, r_y^-, r_z^+, r_z^-, t_x^+, t_x^-, t_y^+, t_y^-, t_z^+ and t_z^-.

We model the emphirical data related to translations as a simple linear regression model and the data related to rotations are modelled as a curvilinear regression of degree 2. This results in the estimation of model parameters and gives rise to expressions for prediction.

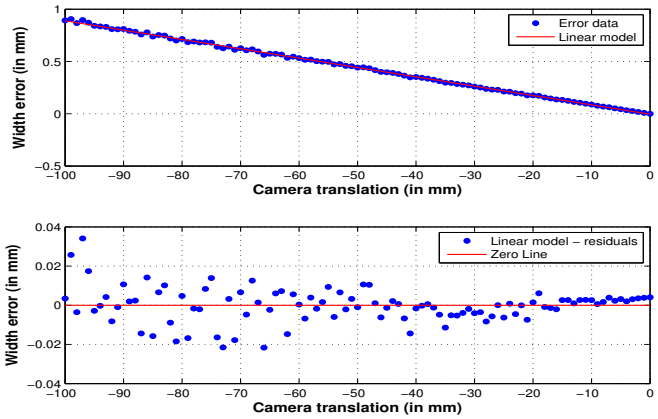


Fig. 7. Linear model fit and residual plots for width error data variation with t_z^-.

Figure 7 illustrates line fitting of variation in width due to t_z^- and figure 8 illustrates curve fitting of variation in depth due to r_y^+. Similarly all the data are modelled suitably well and the model parameters are estimated. An exhaustive list of parameters is shown in the table 1 and table 2.

Now, we have the model fitted to our data with root mean squared error (RMSE) less than unity values that implies good confidence level for estimation. The estimated model parameters are now used to deduce equations for prediction. Examples are shown in equations 8 and 9:

$$width = p0 + p1 * (t_z^-) \tag{8}$$

$$depth = p0 + p1 * (r_y^+) + p2 * (r_y^+)^2 + p3 * (r_y^+)^3 \tag{9}$$

5.4 Tolerance

Let us consider, in case of PantoInspect, the acceptable inaccuracy is 0.5mm. For this acceptable level of inaccuracy we can find the camera mis-alignment (rotation and position) based on the estimated model parameters. By solving

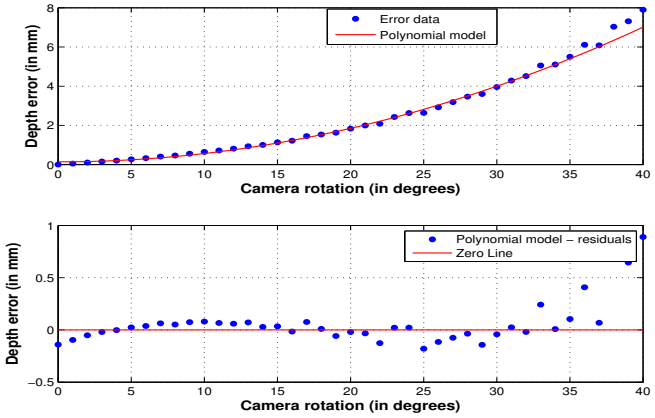


Fig. 8. Curvilinear (2 degree) model fit and residual plots for depth error data variation with r_y^+.

the equations defining the model for 0.5mm error, the maximum tolerance for the camera mis-alignments are estimated and are summarised as in table 3.

6 Conclusion

We identified the PantoInspect system as a suitable use case for measuring inspected data in 3D, using a single calibrated camera. To study the effects of camera mis-alignment on the accuracy of measurements, we emulated the camera mis-alignment in both position and orientation for several values, and obtained the width and depth error data. The resulting data was modelled using suitable regression models and we deduced expressions for prediction. Using the model parameters and expressions, we obtained tolerances for given acceptable inaccuracy limit.

Overall, our paper provided a statistical tool or a study methodology, that is easily implementable and reproducible. Our approach can be directly used by single camera vision systems to estimate tolerances of camera mis-alignment for an acceptable (defined) accuracy.

The knowledge about tolerance is helpful for mechanical design considerations of the camera placement in vision based inspection system, to achieve a desired level of confidence in the accuracy of the system. However, our approach assumes that the measurements are carried out on points that lie on a plane.

In the future, we would like to use the same model to estimate camera motion and further re-calibrate the camera on-the-fly, without the aid of the checkerboard.

Data	Linear		
f(x),x	p_0	p_1	RMSE
width, t_x^-	5.13e-07	-7.14e-06	6.36e-06
width, t_x^+	-1.73e-07	7.15e-06	5.47e-06
width, t_y^-	2.22e-03	-7.13e-03	6.47e-03
width, t_y^+	-1.28e-03	7.43e-03	6.41e-03
width, t_z^-	-4.04e-03	-8.93e-03	7.34e-03
width, t_z^+	3.84e-03	8.37e-03	7.20e-03
depth, t_x^-	5.09e-07	-4.23e-03	4.33e-06
depth, t_x^+	-6.48e-08	4.26e-06	3.51e-06
depth, t_y^-	1.54e-03	-4.23e-03	4.11e-03
depth, t_y^+	-2.7e-03	4.47e-03	4.118e-03
depth, t_z^-	-2.45e-03	-5.30e-03	5.62e-03
depth, t_z^+	1.77e-03	5.01e-03	3.83e-03

Table 1. Model parameters estimated for translation

Acknowledgement

We thank Lars Baunegaard With, Morten Langschwager and Claus Hoelgaard Olsen at Image House PantoInspect A/S, for all their valuable discussions and encouragement.

References

1. Bouguet, J.Y.: "Camera calibration toolbox for Matlab", (2008): *http : //www.vision.caltech.edu/bouguetj/calib_doc/*
2. Image House PantoInspect, Denmark: "*http : //www.pantoinspect.dk/*"
3. Zhengyou, Zhang: "A Flexible New Technique for Camera Calibration:IEEE Transactions on Pattern Analysis and Machine Intelligence":(1998): vol.22, p.1330-1334.
4. Aurélien, Cord, et. al.: "Automatic Road Defect Detection by Textural Pattern Recognition Based on AdaBoost": Comp.-Aided Civil and Infrastruct. Engineering: (2012): vol.27, No.4.
5. T Roger: "A Versatile Camera Calibration Technique for High-accuracy 3D Machine Vision Metrology using Off-the-shelf TV Cameras and Lenses": Robotics and Automation, IEEE Journal: (1987): 323-344.
6. R Hartley, Richard and Zisserman, Andrew:"Multiple View Geometry in Computer Vision":(2003):Cambridge University Press, isbn0521540518, ed.2, New York, USA.
7. J Raj: "The Art of Computer Systems Performance Analysis": John Wiley & sons inc.: (1991).
8. N S S Mar et.al.: "Design of Automatic vision Based Inspection System for Solder Joint Segementation": Journal of AMME: (2009): vol.34, issue 2.
9. Z Dongming, L Songtao: "3D Image Processing Method for Manufacturing Process Automation": Journal of Computing Industry: (2005): vol 56, no. 8.
10. T Brosnan et. al.: "Improving Quality Inspection of Food Products by Computer Vision - Review": Journal of Food Engineering 61: (2004): 3-16

Data	Polynomial			
f(x),x	p_0	p_1	p_2	RMSE
width, r_x^-	0.926	-0.014	0.064	0.912
width, r_x^+	-0.212	0.356	-0.005	0.142
width, r_y^-	0.688	0.095	0.011	0.468
width, r_y^+	0.974	-0.177	0.020	0.563
width, r_z^-	0.085	-0.072	0.003	0.102
width, r_z^+	0.127	0.065	0.001	0.076
depth, r_x^-	0.386	-0.235	0.029	0.394
depth, r_x^+	0.426	0.370	-0.005	0.141
depth, r_y^-	0.158	0.005	0.002	0.096
depth, r_y^+	0.140	-0.001	0.004	0.106
depth, r_z^-	-0.032	-0.319	-0.002	0.150
depth, r_z^+	-0.099	0.338	-0.002	0.130

Table 2. Model parameters estimated for rotations

Tolerances	X axis (deg/mm)	Y axis (deg/mm)	Z axis (deg/mm)
Rotation (width)	-0.46 to 0.82	-2.96 to 4.27	-4.73 to 5.12
Rotation (depth)	-0.11 to 0.19	-12.57 to 9.21	-1.68 to 1.79
Translation (width)	-6.97e04 to 6.98e04	-69.83 to 67.42	-56.41 to 59.20
Translation (depth)	-11.82e05 to 11.75e04	-117.93 to 112.35	-94.67 to 99.44

Table 3. Tolerances for camera mis-alignment, given the system inaccuracy limit as 0.5mm.

11. V G Narendra, K.S.Hareesh:"Quality Inspection and Grading of Agricultural and Food Products by Computer Vision - Review": IJCA: (2010).
12. T Heimonen, et. al.: "Experiments in 3D Measurements by Using Single Camera and Accurate Motion": Proceedings of the IEEE International Symposium: (2001):p.356,361.
13. S Wei, et. al.: "3D Displacement Measurement with a Single Camera based on Digital Image Correlation Technique": International Symposium on Advanced Optical Manufacturing and Testing Technologies: (2007): vol. 6723.
14. N Araki, et. al.: "Vehicle's Orientation Method by Single Camera Image Using Known-Shaped Planar Object": IJICIC: (2009): vol. 7, no. 7.
15. W Zhao and N, Nandhakumar: "Effects of Camera Alignment Errors on Stereoscopic Depth Estimates": Pattern Recognition, Elsevier: (1996): vol.29, p.2115-2126.
16. M , Santoro, et. al.:"Misalignment Correction for Depth Estimation using Stereoscopic 3-D Cameras": MMSP, IEEE 14th International Workshop: (2012): pp.19,24.
17. Godding et.al: "Geometric calibration and orientation of digital imaging systems.": Aicon 3D Systems, Braunschweig: (2002): http://www. falcon. de/falcon/eng/documentation.
18. H Eric, et.al.: "The Effects of Translational Misalignment when Self-Calibrating Rotating and Zooming Cameras.": Pattern Analysis and Machine Intelligence, IEEE Transactions: (2003): 1015-1020.

Chapter 10

Paper IV: Online Re-calibration for Robust 3D Measurement Using Single Camera-PantoInspect Train Monitoring System

Title: Online Re-calibration for Robust 3D Measurement Using Single Camera-PantoInspect Train Monitoring System.

Authors: Deepak Dwarakanath, Carsten Griwodz, Pål Halvorsen and Jacob Lildballe.

Published & **Presented:** In Proceedings of the International Conference on Computer Vision Systems (ICVS), 2015.

Abstract: Vision-based inspection systems measures defects accurately with the help of a checkerboard calibration (CBC) method. However, the 3D measurements of such systems are prone to errors, caused by physical misalignment of the object-of-interest and noisy image data. The PantoInspect Train Monitoring System (PTMS), is one such system that inspects defects on pantographs mounted on top of the electric trains. In PTMS, the measurement errors can compromise railway safety. Although this problem can be solved by re-calibrating the cameras, the process involves manual intervention leading to large servicing times.

Therefore, in this paper, we propose Feature Based Calibration (FBC) in place of CBC, to cater an obvious need for online re-calibration that enhances the usability of the system. FBC involves feature extraction, pose estimation, back-projection of defect points and estimation of 3D measurements. We explore four state-of-the-art pose estimation algorithms in FBC using very few feature points.

This paper evaluates and discusses the performance of FBC and its robustness against practical problems, in comparison to CBC. As a result, we identify the best FBC algorithm type and operational scheme for PTMS. In conclusion, we show that, by adopting FBC in PTMS and other related 3D systems, better performance and robustness can be achieved compared to CBC.

Errata: In equation 1 of this paper, scaling factor $\frac{1}{Z_c}$ is missing. This is corrected in the thesis, in equation 3.1.

Online Re-calibration for Robust 3D Measurement Using Single Camera-PantoInspect Train Monitoring System

Deepak Dwarakanath[1,2(✉)], Carsten Griwodz[1], Pål Halvorsen[1], and Jacob Lildballe[2]

[1] Simula Research Laboratory, University of Oslo, Oslo, Norway
[2] ImageHouse PantoInspect A/S, Copenhagen, Denmark
deepakdw@ifi.uio.no

Abstract. Vision-based inspection systems measures defects accurately with the help of a checkerboard calibration (CBC) method. However, the 3D measurements of such systems are prone to errors, caused by physical misalignment of the object-of-interest and noisy image data. The *PantoInspect Train Monitoring System* (PTMS), is one such system that inspects defects on pantographs mounted on top of the electric trains. In PTMS, the measurement errors can compromise railway safety. Although this problem can be solved by re-calibrating the cameras, the process involves manual intervention leading to large servicing times.

Therefore, in this paper, we propose Feature Based Calibration (FBC) in place of CBC, to cater an obvious need for online re-calibration that enhances the usability of the system. FBC involves feature extraction, pose estimation, back-projection of defect points and estimation of 3D measurements. We explore four state-of-the-art pose estimation algorithms in FBC using very few feature points.

This paper evaluates and discusses the performance of FBC and its robustness against practical problems, in comparison to CBC. As a result, we identify the best FBC algorithm type and operational scheme for PTMS. In conclusion, we show that, by adopting FBC in PTMS and other related 3D systems, better performance and robustness can be achieved compared to CBC.

1 Introduction

Nowadays, industries make extensive use of 3D measurement systems to cater for inspection applications. The *PantoInspect Train Monitoring System* (PTMS) [1] adopted by Rail Net Denmark, Sydney Trains Australia and others, is a system that inspects defects occurring on pantographs of electric locomotives (their root-mounted carbon structures that are in-contact with electric wires). PTMS makes use of lasers and a camera to measure defects with a priori knowledge of camera calibration parameters, i.e. camera intrinsic (focal length, principal axes) and extrinsic (position, orientation). The quality of this calibration governs the accuracy of its 3D measurements.

L. Nalpantidis et al. (Eds.): ICVS 2015, LNCS 9163, pp. 498–510, 2015.
DOI: 10.1007/978-3-319-20904-3_45

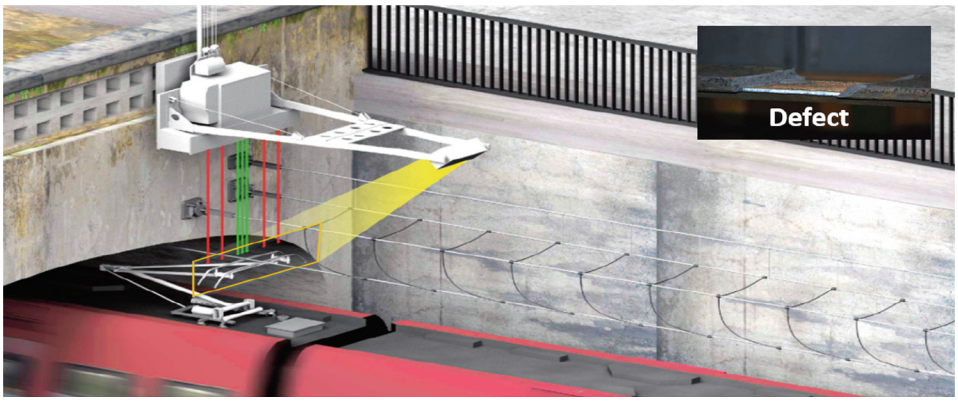


Fig. 1. PTMS system to inspect defects on the pantographs (Color figure online)

In this application scenario, practically unavoidable situations such as camera/pantograph misalignment (change in position and orientation) can occur. Camera misalignment is caused by mishandling PTMS during transportation and servicing. Pantograph misalignment is caused by the movement of train and thrust of pantograph against the catenary wire. In such cases, the calibration data is not useful anymore and therefore leads to inaccurate defect measurements. Hence, practical misalignment degrades the performance of PTMS, unless the camera is re-calibrated. Currently, PTMS adopts traditional checkerboard calibration (CBC) [8]. The typical placement of PTMS (see Fig. 1) requires that it is unmounted before CBC, leading to huge maintenance and servicing times. Therefore, PTMS is in need of an automatic camera calibration process.

In this paper, we explore *Feature Based Calibration* (FBC) methods [3–6] to provide an alternative solution for PTMS. Since FBC can be performed without unmounting, PTMS and other related applications acquire robustness against camera/pantograph misalignment effects. FBC is a calibration process that consists of feature extraction and 3D-2D pose estimation to measure the defects.

Four state-of-art 3D-2D pose estimation algorithms [7–10] were selected for FBC. Although these algorithms can work independently on arbitrary points, the challenges in adopting these algorithms for PMTS are: (a) PMTS yield only few feature points, and (b) feature points are noisy due to misalignment errors and motion blur. Thus, evaluation of the algorithms becomes important to understand the practical implications of adopting FBC. We use CBC as the reference and compare the robustness of FBC-based methods against practical problems. The evaluation was carried out on a dataset from the PTMS testing facility, which offered a variety of sample data and noise-free reference measurements.

This paper is organized as follows: Functional overview of PTMS is explained in Sect. 2. Section 3 outlines the proposed FBC methodology for PTMS. The evaluation setup and their results are discussed in Sects. 4 and 5, respectively. In Sect. 6, we conclude the results and identify a suitable FBC algorithm that performs better than CBC and is more feasible for dynamic environments.

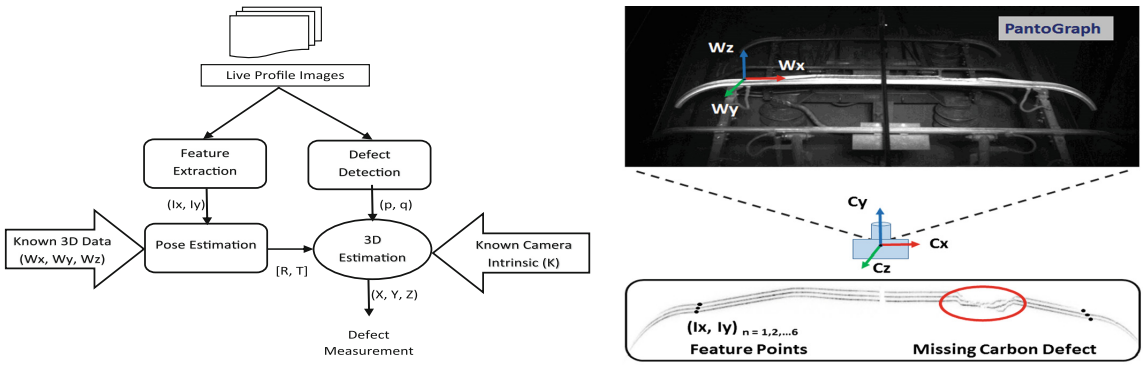
2 PTMS System

PTMS is a non-tactile fault detection system, which inspects pantographs and measures the defects in the carbon strips. Pantographs are mechanical structures with carbon strips, fixed on top of train wagons, which are raised to touch the overhead contact wire for electricity. In the course of time, due to constant contact of carbon strips with the wire, various types of defects (vertical crack, edge chip, abnormal wear and missing carbon) might occur. PTMS is meant to discover when these defects become severe, while allowing for expected wear.

PTMS is mounted over the train tracks (see Fig. 1). When the train passes right below the system, range sensors (red lines in Fig. 1) detect the pantograph and three laser lines are projected onto the carbon strips (green line in Fig. 1). The camera captures a near infrared image of the laser lines, termed as *profile image*. When defects are present, the laser lines are deformed and define the geometry of the defects. The system then detects the defects, measures their width & depth and raises an alarm if measurements are above certain thresholds.

3 Proposed Calibration Methodology

We propose a feature based calibration (FBC) as in Fig. 2(a), which involves a 2-step process consisting of (1) Feature Extraction and (2) Pose Estimation. This is done by extracting features from the same *profile images* that are taken to detect defects. These features allow the estimation of the camera pose, and subsequent 3D measurements of defects.



(a) FBC Methodology. (b) Profile Image Analysis.

Fig. 2. Feature Based Calibration for PantoInspect

3.1 Feature Extraction

PTMS casts 3 parallel laser lines, each yielding 2 distinctive, well-known points that can be detected in each *profile image* as shown in Fig. 2(b), for a total of 6 feature points. Notice that the shape of the line traverses the shape of the pantograph and bends on both ends, where the carbon strip ends. These 6

end points are extracted from the profile image using known feature extraction techniques. These points serve as 2D feature points (I_x,I_y) for calibration. The corresponding 3D points always lie on an imaginary 3D plane parallel to the surface of the pantograph.

The 3D reference axis (world coordinate system) is assumed to be on the pantograph, as shown in Fig. 2(b). Pantographs have standard dimensions of length, by which real metric measurement of each points in 3D world coordinate system (W_x, W_y, W_z) can be obtained. Thus, both 2D feature points and 3D world points are determined and further used for camera pose estimation process.

3.2 Pose Estimation

Camera Pose Estimation refers to estimating rotation (R) and translation (T) parameters of the camera with respect to the world coordinate system. As the 2D and 3D points lie on a plane, a homography between camera and world coordinate system can be found. Estimating the projective mapping and thereby extracting the camera parameters [R,T] is the goal of camera pose estimation. The parameter R is an orthonormal 3×3 matrix representing rotations in x,y,z axes. T is a 3×1 matrix representing camera translation along 3 axes.

There are several state-of-art algorithms [7–13] for estimating single camera pose. Out of those, we select few well-known good performing algorithms named, FBC-boug [8], FBC-zhang [9], FBC-gold [10] and FBC-epfl [7] as candidates for calibrating PTMS. All methods can operate with $n \geq 4$ point correspondences and they assume that the intrinsic camera parameters (K) are known.

FBC-boug method initially estimates planar homography using the Quasi-Linear algorithm and recovers [R,T] parameters, which are further optimized to minimise reprojection error through Gradient Descent. **FBC-zhang** method estimates planar homography using the Direct Linear Transformation followed by a non-linear optimization (Levenberg Marquardt) based on Maximum Likelihood criterion. Then, [R,T] are recovered using orthogonal enforcement. **FBC-gold** method estimates a projective geometric transformation using Gold Standard algorithm before recovering [R,T]. Unlike other methods, **FBC-epfl** is non-iterative approach to PnP problem. Under PnP problem, 3D points are expressed in camera coordinate system and then, the Euclidean motion that aligns both world and camera references is used to retrieve [R,T]. This method adopts the idea of expressing n 3D points as weighted sum of four virtual control points, which reduces complexity and noise sensitivity.

3.3 3D Estimation

Since the reference 3D points lie on an imaginary plane, the conversion from 2D (p,q) to 3D (X,Y,Z) points becomes merely a ray-plane intersection [10], as shown in Eq. 1, where the points are expressed in homogeneous coordinates.

$$\begin{bmatrix} p \\ q \\ 1 \end{bmatrix} = K[R|T] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{1}$$

K, R and T are obtained by the FBC process. The Z-axis of 3D points are zero for points on a plane. Now, expanding the matrix R and vector T, we have Eq. 2, and 3D points (X,Y) of the defects on the plane can be estimated.

$$\begin{bmatrix} p \\ q \\ 1 \end{bmatrix} = K * \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = \underbrace{K * \begin{bmatrix} r_{11} & r_{12} & t_1 \\ r_{21} & r_{22} & t_2 \\ r_{31} & r_{32} & t_3 \end{bmatrix}}_{H} \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{2}$$

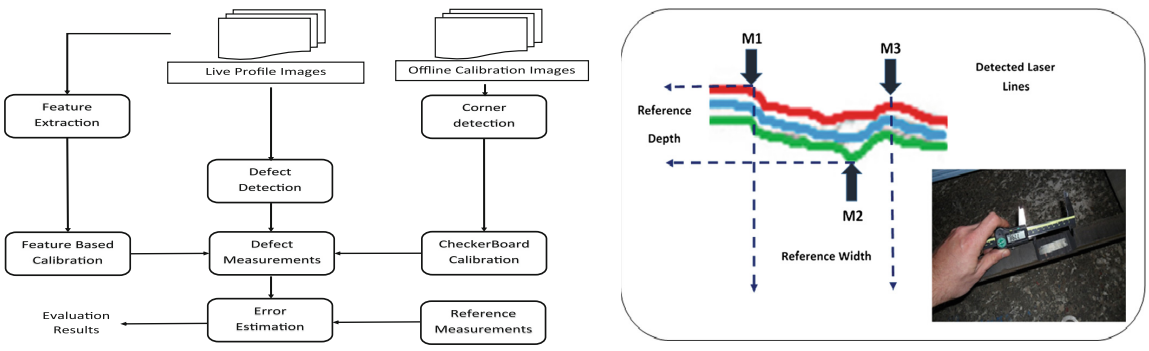
3.4 Defect Measurement

Referring to Fig. 3(b), in each of the three lines, three major points (M_1, M_2, M_3) are determined for measurement purposes. These major points are back-projected to estimate their 3D coordinates ($\widehat{M_1}$, $\widehat{M_2}$, $\widehat{M_3}$) as in Eq. 2.

$$Width = \widehat{M_3} - \widehat{M_1} \qquad and \qquad Depth = AbsMax(H1, H2) \tag{3}$$

$$where, \qquad H1 = \widehat{M_1} - \widehat{M_2} \qquad H2 = \widehat{M_3} - \widehat{M_2}$$

All defects are characterized by a width and depth, which are computed using Eq. 3. Thus, PTMS can estimate FBC parameters (Sects. 3.1 and 3.3), and use the parameters to measure the defects (Sects. 3.3 and 3.4).



(a) Experimental Setup. (b) Defect Measurement.

Fig. 3. Evaluation scenario

4 Evaluation

The evaluation determines the performance traits of PTMS in estimating defect width and depth measurements by adopting the four proposed FBC methods, in comparison with the currently used CBC method.

The evaluation methodology is illustrated in Fig. 3(a). CBC is carried out offline using around 20 images of a checkerboard pattern with the help of the Matlab Calibration Toolbox [2], whereas FBC is carried out online, with only 6 feature points that are extracted from the *profile image*.

Every *profile image* is used for defects detection. Whenever the lines in the profile image are not straight, there is potentially a defect (Fig. 3(b)). After defects are identified, their width and depth are measured based on the estimated calibration parameters of both CBC and FBC. Using reference measurements (Table 1), the error estimations are computed. These errors were compared to a ground truth and used for evaluating the performance of PTMS over various calibration methods.

4.1 Datasets

Defect measurements were carried out on Pantographs (BR and EG types) provided by BaneDanmark (Rail Net Denmark). The dataset was obtained from product testing conducted at the PTMS factory. The real width and depth of each of these defects were measured using a calibrated caliper to acquire the ground truth (Fig. 3(b) and Table 1). The experiments were conducted on 5 readings of 2 types of pantographs having 4 types of defects on them, for a total of 40 data samples. To avoid noisy features extracted and to focus on correct evaluation, every feature point was manually annotated. However, the performance with noisy data is analysed later in our study.

Table 1. Reference measurements of defects of two pantograph types

Measurement	Pantograph-type			
(in millimeters)	BR-type		EG-type	
Defects	Width	Depth	Width	Depth
Vertical crack	2.38	17	5.88	20
Missing carbon	77.98	17	39.04	20
Edge crack	24.21	6	21.18	5
Abnormal wear	19.36	6	14.78	5

4.2 Operational Schemes

FBC can be carried out on every *profile image* before conducting the measurement, but one cannot guarantee a noise-free image that is good enough to extract

features. A noisy profile image will worsen the quality of re-calibration. Alternatively, calibration can be carried out only when the accuracy of the system deteriorates (one way of measuring this is by checking if the measured depth is larger than the pantograph thickness). Alternatively, calibration can be carried out at regular interval. However, a more stable scenario can be to calibrate during servicing and maintenance periods, where full control on measurements is possible. In this paper, we have considered two operational schemes for evaluation.
Scheme 1: FBC is carried out on every *profile image* and the defect is measured on those images using its respective calibration parameter.
Scheme 2: FBC is carried out on a random *profile image* and measurement is carried out on the rest of all profile images with the same calibration parameters.

4.3 Practical Implications

Feature point mis-detection introduces noise in the feature point locations. These are caused by poorly visible images, which can be due to laser misalignment, flash under/over exposure, motion blur or sunlight.

Since the pantographs are the moving elements that are in contact with the caternary wire, there might be linear and angular displacements of the structure. Linear displacements can occur due to vertically upward movement, called *Uplift*, which is deliberately made to provide more upward thrust to the wire. And angular displacements occur due to uneven forces being exerted on the pantograph over time, when the wire in contact is off-centered. Theses displacements are called *Yaw Angle*, *Roll Angle*, and *Pitch Angle*, referring to rotation around 3 axes. We have considered these practical implications in our study.

Table 2. Absolute angular difference in degrees between CBC and FBC - scheme1

FBC-type	R_x (Tilt)	R_y (Roll)	R_z (Pan)
epfl	7.17	3.50	9.46
boug	9.99	8.21	11.93
zhang	10.86	6.7	11.30
gold	11.51	5.96	10.99

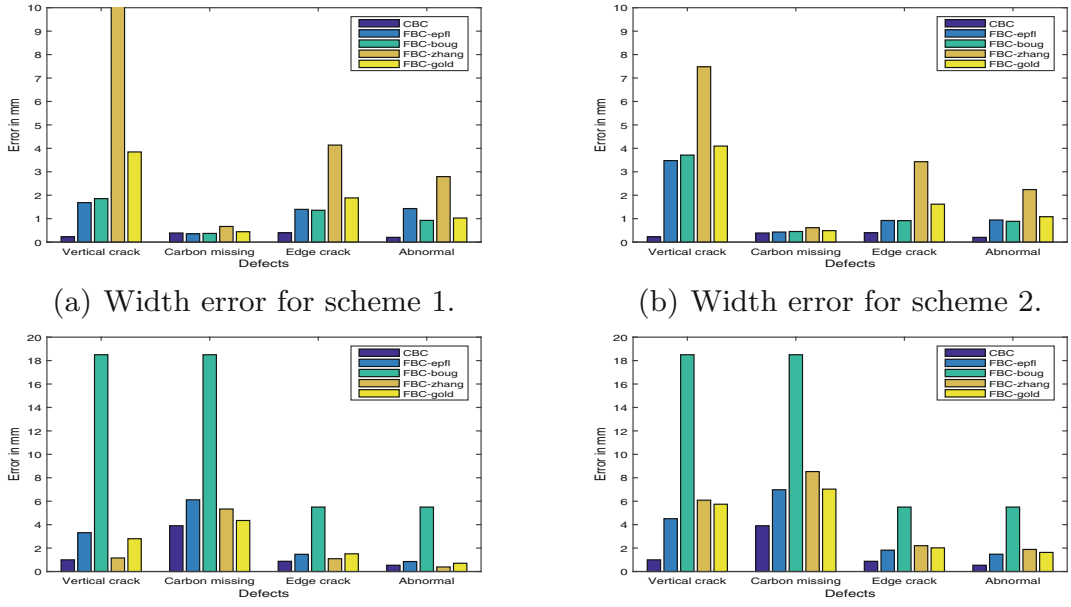
5 Results and Discussion

The FBC was carried out for 4 candidate methods, namely FBC-epfl, FBC-boug, FBC-zhang and FBC-gold (explained in Sect. 3.2) and their rotational parameter difference from that of CBC was noted (Table 2). The camera rotational parameter plays an important role in the accuracy of FBC for measurements in PTMS. In the first experiment, width and depth error of defects (in millimeters) were computed for the 4 FBC methods in comparison to CBC, yielding a mean absolute error over several samples and defects for both schemes.

5.1 Width Measurement of Defects

Intuitively, the angular deviation (Table 2), in the estimated camera pan and camera roll parameters contribute to the errors in width measurements. The Figs. 4(a) and (b) show the result of width measurement for both schemes. For *edge crack* and *abnormal wear*, FBC-epfl, FBC-boug and FBC-gold introduced only around 1–2 mm mean error compared to CBC. All FBC types performed with accuracy very close to CBC for *missing carbon* defect, which had sufficiently large width reference (see Table 1). On the lower side, the width of *vertical crack* is about 2–6 mm and the profile image was captured from 3 m distance. Hence, profile detection of a narrow width structure introduced noise, which thereby resulted in width error as seen in the figures. In this case, the accuracy of FBC-zhang degraded, whereas other FBCs showed mean errors between 2–4 mm in both schemes. However, *scheme 1* is most suitable in this case. It is observed that all FBC types are more sensitive to narrow widths (<5 mm) in *scheme 2* than in *scheme 1*, because in *scheme 2*, once FBC is carried out, the pantograph position can be misaligned, which results in wrong measurements.

Overall, FBC-epfl and FBC-boug performed best for width measurement compared to CBC, with a maximum increase in the mean error of about 1 mm for *scheme 1* and 1.5 mm for *scheme 2*.



(a) Width error for scheme 1.

(b) Width error for scheme 2.

(c) Depth error for scheme 1.

(d) Depth error for scheme 2.

Fig. 4. Mean difference of width and depth measurements for both schemes

5.2 Depth Measurement of Defects

Depth measurement results for both schemes are shown in Figs. 4(c) and (d). Depth errors were introduced due to deviations, mainly in camera pitch rotational parameter. The figures show that all FBC types except FBC-boug performed very close to the CBC with mean errors between 1–3 mm in both schemes.

FBC-boug uses the same algorithm for pose estimation as in CBC, but there is a difference in the feature point locations. The feature points in CBC assume z axis = 0 (places the checkerboard in XZ plane of world coordinate system) and both depth & width are determined along the X and Z axes, respectively. In FBC, feature points lie on the XY plane and depth is measured along the Z axis. Here, the thinness of the pantograph itself restricts the feature plane for FBC on which the feature points were detected. Hence, estimation of mean depth error gave a higher value for FBC-boug. For *edge crack* and *abnormal wear*, FBC types performed with a mean error difference of about 1 mm compared to CBC. For both *vertical crack* and *missing carbon*, FBC introduced mean errors of about 2 mm for *scheme 1* and 3 mm for *scheme 2*. The errors were higher for these two defects because the long narrow depth in *vertical crack* was detected with noise and a large depth in *missing carbon* is strongly affected even by small deviations in the camera pitch angle. Table 2 shows the rotational parameter offsets that cause such depth errors. It is observed that scheme 1 is more suitable for depth measurements in terms of accuracies.

Overall, FBC-epfl, FBC-zhang and FBC-gold performed the best for depth measurement compared to CBC, with a maximum increase in mean error of about 1.5 mm for *scheme 1* and 3 mm for *scheme 2*.

5.3 Error Distribution

To accommodate the randomness of the error, we considered to observe and compare the error distributions. We assumed an ideal error distribution as a baseline for comparison. Computed Cumulative Density Function (CDF) for all (width and depth) errors, are shown in Fig. 5, where a tendency of divergence of FBC/CBC from ideal baseline can be seen. To quantify the measure of divergence, we used Kullback-Leibler distance (KLD) [14]. For discrete PDFs P and Q, the KL divergence of Q from P is defined as in Eq. 4.

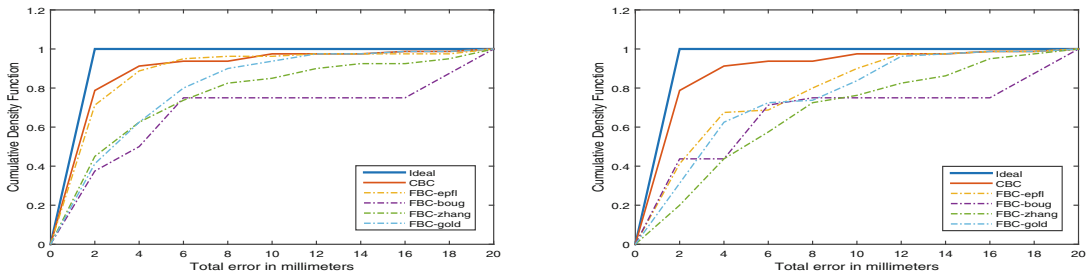


Fig. 5. Cumulative Density Function for scheme 1 and 2.

$$D_{KL}(P||Q) = \sum_i P(i) ln \frac{P(i)}{Q(i)} \tag{4}$$

The KLD values for FBC/CBC methods are given in Table 3. From the table, it is quite evident that for both *scheme 1* and *scheme 2*, FBC-boug performed the

best reduction of width error, which was summarized in the Sect. 5.1. Similarly, FBC-epfl showed the best performance for depth and overall error in *scheme 1* as summarized in Sect. 5.2. Evidently, FBC-epfl performed better than CBC for depth measurement. Unlike in mean error estimates for *scheme 2* shown in Fig. 4(d)), FBC-boug showed performance closer to FBC-epfl in terms of error distribution, as in Fig. 5. FBC-zhang and FBC-gold performed alternatively better than each other in various configurations, but still not up to FBC-epfl and FBC-boug.

Eventually, we see that FBC methods have performed with better accuracy in *scheme 1* configuration than in *scheme 2*.

Table 3. Kullback-Leibler Divergence values for total (width + depth) error

Measurement	Ideal	CBC	FBC-epfl	FBC-boug	FBC-zhang	FBC-gold
Scheme-1						
Width	0	1.29	1.39	**0.92**	1.39	1.20
Depth	0	0.51	**0.39**	0.69	0.52	0.80
Total	0	0.24	**0.34**	0.98	0.80	0.88
Scheme-2						
Width	0	1.29	1.29	**0.80**	1.61	1.39
Depth	0	0.52	0.92	**0.70**	1.39	1.12
Total	0	0.24	0.88	**0.83**	1.61	1.16

5.4 Resilience

Next, both the FBC and CBC were evaluated for resilience to practical disturbances, which are feature detection error (*pixel noise*) and pantograph misalignment (*uplift*, *yaw*, *roll*, *pitch*), as explained in Sect. 4.3. Variations of these parameters were emulated within a practical range of values and a new set of observations and feature points were obtained. Using new sets of data, FBC was carried out for each perturbation of the parameter and the KLD was computed. Only *scheme 1* operation, which had shown better performance so far, was considered to evaluate resilience of FBC and CBC.

For pixel noise variation, gaussian noise with a variance between 10 and +10 was added to the signal. The uplift was emulated by varying the vertical axes from −0.5 m to +0.5 m. All rotations (yaw, roll and pitch) were allowed to vary between −10 and +10 degrees. All subfigures in the first column of Fig. 6 show resilience of width error estimation to all five practical implications. Similarly, resilience of depth error is shown in subfigures of the second column.

For pantograph misalignment disturbances, CBC errors are higher than one or more of the FBC-types. In CBC, the reference world coordinate axis is fixed in space based on the position of the checkerboard. For FBC, on the other hand, the reference world axis is located on the pantograph itself. Hence, misalignment

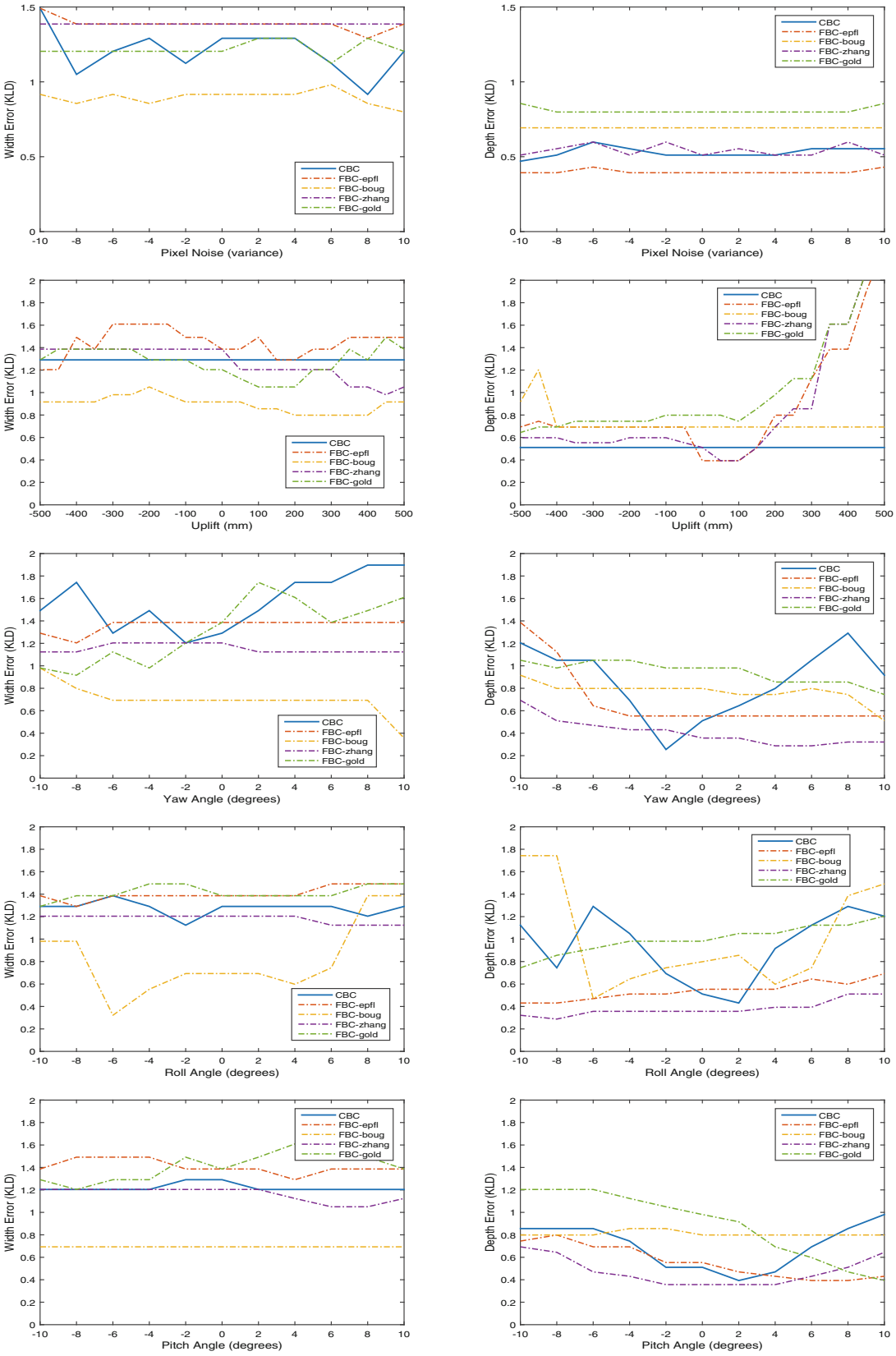


Fig. 6. Experimenting resilience over physical implications

of the pantograph does not affect the measurements using FBC, and several FBC methods are more robust to pantograph misalignment than CBC.

For feature detection errors, FBC-types will be obviously more sensitive than CBC, because FBC relies on noise-free feature points for high quality calibration. However, FBC-boug and FBC-epfl showed better resilience than CBC in terms of handling noisy feature points, because their optimization yielded better results with the localization of world coordinate system on the pantograph.

Width error of FBC-boug consistently showed the best resilience over CBC and other FBC types. Although FBC-boug is similar to CBC in terms of pose estimation procedure, FBC-boug used only 6 points compared to CBC, which used more than 200 points for bundle optimized solution for the pose.

In most of the pantograph rotational disturbances, the depth error for FBC-types showed a consistently flatter response compared to CBC, which was very sensitive to rotational disturbances. However, FBC-zhang and FBC-epfl showed the best resilience.

6 Conclusion

Considering the PantoInspect Train Monitoring System as a usecase, we have outlined the specific practical problem underlining the usage of vision based inspection systems. The paper is motivated with the need for online-recalibration and how CBC fails to fulfill the need.

We proposed FBC methods for PTMS, which uses very few points in an image. We have evaluated four state-of-art algorithms for camera pose estimation. The results have shown that FBC has outperformed CBC in many cases. The FBC-epfl and FBC-boug methods have shown best results in terms of accuracy and robustness for depth and width error, respectively. Carrying out FBC on every profile image before analysing the defect (scheme 1), is found to be more accurate. However, if the image is too noisy to extract features, recent FBC parameter needs to be re-used. All FBC methods can be executed in real-time, without relevant penalty to the system speed.

Hence, we conclude that online re-calibration for error-sensitive 3D measurement systems (such as PTMS), is possible using FBC methods that give effectively a better performance and robustness than CBC. This tremendously increases the usability of 3D vision inspection systems with greater flexibility of using online re-calibration without any manual intervention.

References

1. Imagehouse PantoInspect A/S, Denmark. http://www.pantoinspect.dk/
2. Bouguet, J.Y.: Camera calibration toolbox for Matlab (2008). http://www.vision.caltech.edu/bouguetj/calib_doc/
3. Li, C., Lu, P., Ma, L.: A camera on-line recalibration framework using SIFT. Vis. Comput. **26**(3), 227–240 (2010). Springer-Verlag

4. Dwarakanath, D., Eichhorn, A., Griwodz, C., Halvorsen, P.: Faster and more accurate feature-based calibration for widely spaced camera pairs. In: Second International Conference on Digital Information and Communication Technology and it's Applications (DICTAP), pp. 87–92 (2012)
5. Mavrinac, A., Xiang, C., Tepe, K.: Feature-based calibration of distributed smart stereo camera networks. In: Second ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC 2008, pp. 1–10 (2008)
6. Carr, P., Sheikh, Y., Matthews, I.: Point-less calibration: camera parameters from gradient-based alignment to edge images. In: IEEE Workshop on the Applications of Computer Vision (WACV) (2012)
7. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: an accurate O(n) solution to the PnP problem. Int. J. Comput. Vis. **81**(2), 155–166 (2009)
8. Bouguet, J.Y.: Visual methods for three-dimensional modeling Ph.D. Thesis, California Institute of Technology (1999). http://www.vision.caltech.edu/bouguetj/thesis/thesis.html
9. Zhengyou, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1330–1334 (1998)
10. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, New York (2003). ISBN: 0521540518
11. Faugeras, O.: Three-Dimensional Computer Vision: AGV. MIT Press, Cambridge (1993)
12. Tsai, R.Y.: A versatile camea calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV camera and lenses. IEEE J. Robot. Autom. **3**, 323–344 (1987)
13. Lu, C.-P., Hager, G.D., Mjolsness, E.: Fast and globally convergent pose estimation from video images. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 610–622 (2000)
14. Kullback, S., Leibler, R.A.: On information and sufficiency. J. Ann. Math. Stat. **22**(1), 79–86 (1951)

Chapter 11

Paper V: Robustness of 3D Point Positions to Camera Baselines in Markerless AR Systems

Title: Robustness of 3D Point Positions to Camera Baselines in Markerless AR Systems.

Authors: Deepak Dwarakanath, Carsten Griwodz and Pål Halvorsen.

Published & **Presented:** In Proceedings of the 7th International Conference on Multimedia Systems (MMSys), 2016.

Abstract: In the Augmented Reality (AR) applications, high quality relates to an accurate augmentation of virtual objects in the real scene. This can be accomplished only if the position of the observer is accurately known. This boils down to solving image-based location problem by an accurate camera pose (relative position and orientation) estimation, when a stereo or multiple camera setup is used. Consider a relevant application scenario as in a movie production set, where the director is able to preview a scene as an integrated view of the real scene augmented with animated 3D models. The main camera shoots the scene, where as secondary stereo camera pair is used for image registration and localization. The director can view the integrated preview from any viewpoint perfectly, as long as the camera pose estimation is accurate.

Moreover, in the case of a markerless AR system, the challenge for camera pose estimation, is strongly influenced by the precision of detected feature correspondences between the images. Unfortunately, several of the state-of-art feature extractors (detectors and descriptors) cannot guarantee a consistent accuracy of camera pose estimation, especially at varied camera baselines (viewpoints). As a consequence, the precise augmentation of objects, as desired in an AR application, is compromised. Hence, it becomes necessary to understand the magnitude of this error in relation to the camera baseline depending on the chosen feature extractors.

We, therefore, assess the quality of the position and the orientation of 3D reconstruction by evaluating 26 feature extractor combinations over 50 different camera baselines. To be directly relevant for AR applications, we evaluate by measuring the reconstruction

error in 3D space, instead of re-projection error in 2D space. After the experiment, we have found the SIFT and KAZE feature extractors to be highly accurate and more robust to large camera baselines than others. Importantly, as a result of our study, we provide a recommendation for system builders to help them make a better choice of the feature extractor and/or the camera density required for their application.

Robustness of 3D Point Positions to Camera Baselines in Markerless AR Systems

Deepak Dwarakanath[1,2], Carsten Griwodz[1,3], Pål Halvorsen[1,3]
[1]University of Oslo, Oslo, Norway
[2]Image Metrology A/S, Horsholm, Denmark
[3]Simula Research Laboratory, Lysaker, Norway
deepakdw@ifi.uio.no, {griff,paalh}@simula.no

ABSTRACT

In the Augmented Reality (AR) applications, high quality relates to an accurate augmentation of virtual objects in the real scene. This can be accomplished only if the position of the observer is accurately known. This boils down to solving image-based location problem by an accurate camera pose (relative position and orientation) estimation, when a stereo or multiple camera setup is used. Consider a relevant application scenario as in a movie production set, where the director is able to preview a scene as an integrated view of the real scene augmented with animated 3D models. The main camera shoots the scene, where as secondary stereo camera pair is used for image registration and localization. The director can view the integrated preview from any viewpoint perfectly, as long as the camera pose estimation is accurate.

Moreover, in the case of a markerless AR system, the challenge for camera pose estimation, is strongly influenced by the precision of detected feature correspondences between the images. Unfortunately, several of the state-of-art feature extractors (detectors and descriptors) cannot guarantee a consistent accuracy of camera pose estimation, especially at varied camera baselines (viewpoints). As a consequence, the precise augmentation of objects, as desired in an AR application, is compromised. Hence, it becomes necessary to understand the magnitude of this error in relation to the camera baseline depending on the chosen feature extractors.

We, therefore, assess the quality of the position and the orientation of 3D reconstruction by evaluating 26 feature extractor combinations over 50 different camera baselines. To be directly relevant for AR applications, we evaluate by measuring the reconstruction error in 3D space, instead of re-projection error in 2D space. After the experiment, we have found the SIFT and KAZE feature extractors to be highly accurate and more robust to large camera baselines than others. Importantly, as a result of our study, we provide a recommendation for system builders to help them make a better choice of the feature extractor and/or the camera density required for their application.

MMSys'16, May 10-13, 2016, Klagenfurt, Austria

DOI: http://dx.doi.org/10.1145/2910017.2910611

CCS Concepts

•Information systems → Multimedia information systems;

Keywords

Augmented Reality, Feature Extraction, Pose Estimation, 3D Accuracy

1. INTRODUCTION

The multimedia industry has paid quite a lot of attention to 3D imaging as in head mount virtual reality systems [1, 2], augmented reality mobile applications [3, 4, 5], interactive augmented reality systems [6, 7], free-viewpoint rendering [8], etc. These applications use two or more cameras to perform tasks such as augmenting 3D models in video sequences, depth estimation, virtual view synthesis, etc. The underlying principle of such multi-camera systems is the estimation of camera pose, i.e., relative camera position and orientation with respect to other cameras.

A central theme in Augmented Reality (AR) research is the enhancement of the human senses by changing what human observers see with their eyes, or annotating it. Of these, modification is more challenging because accurate knowledge of the images that the observers see is required before changes can be made. This knowledge may be derived by augmenting the observers with cameras mounted on their heads [9], and perhaps reconstructing their entire view. Our project goal in POPART[1], however, is to provide an augmented, accurate preview of a film set. This is meant to provide an integrated view of real-life actors with prototype animated 3D models in real-time to director and photographer, weeks or months before post-production is finished. This implies that we augment the image that is seen by the main film camera, and that we have one or two static cameras to estimate the dynamic objects. The static film set itself is, in our case, reconstructed in advance of the filming.

The accuracy of the camera pose estimation plays an important role in order to determine the quality of these applications. Cameras are usually pre-calibrated offline (often, focal length and principal axis are determined using a checkerboard - Matlab Toolbox[2]). When the system is deployed, the camera pose is estimated automatically based on sparse feature points extracted from the images that are

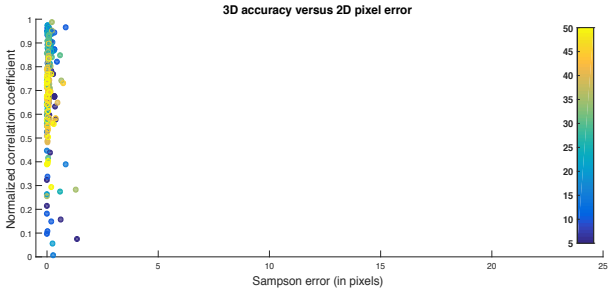
[1]http://www.popartproject.eu, EU Horizon2020 project number 644874
[2]http://www.vision.caltech.edu/bouguetj/calib_doc

captured by these cameras. This is also known as Feature-Based Calibration (FBC).

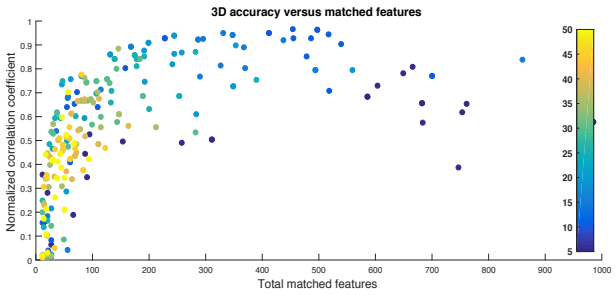
In multi-camera systems, the following statements are commonly accepted:

- A high number of matched feature points in a stereo pair results in a better camera pose estimation.
- Minimizing 2D pixel error calculated between matched pairs results in higher accuracy of 3D estimation, based on epipolar geometry [10].

The first point holds good for iteration-based estimation algorithms (e.g., RANSAC [11]). The second point, however, is not always true. We illustrate this in figure 1, which represents a scatter plot of 3D accuracy versus 2D pixel error and number of matched feature points extracted from images of stereo pair at various baselines (relative displacement between the stereo cameras). Figure 1(a) illustrates that low pixel error does not guarantee high 3D accuracy and, similarly, figure 1(b), that high 3D accuracy is not always obtained by a larger number of feature matches.



(a)



(b)

Figure 1: Scatterplots of matched feature points and 2D pixel error with 3D accuracy.

In this paper, we explore one of the important factors determining the accuracy of camera pose estimation and thereby 3D estimation, i.e., change in the camera baseline, which breaks the common assumptions made above. This paper also casts light upon the quality of current state-of-art feature extractors (combination of detectors & descriptors) [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] used for FBC or camera pose estimation today.

Each of these feature extractors has its own behavioral traits. Some of them claim invariance to change in camera baseline, but the extent of their tolerance is uncertain. Therefore, we evaluate various combinations of feature extractors with a brute-force matcher to determine their robustness to change in the camera baseline. Our study is meant to provide system builders with a better understanding of the operational limits of the state-of-art feature detectors and descriptors. It will help them to make better choices in designing 3D multimedia applications using multi-camera systems. Besides choice of algorithm, it may be helpful in estimating the number and position of cameras that are required for reconstructing rigid structures in a well-known space, with a desired accuracy.

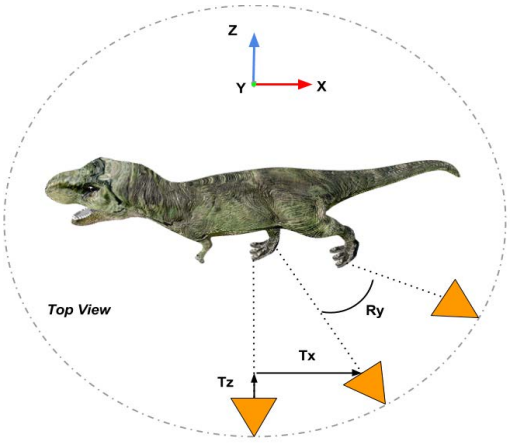


Figure 2: Cameras arranged in a circular configuration around the 3D model.

We have considered a multi-camera scenario as in figure 2, where a number of cameras are placed in a circular configuration around and looking at an object of interest, equidistant from the object's geometric center. We have chosen this configuration to concentrate on changes in baseline, and avoid changing either the objects' size in the frames or the camera's focal length between baseline configurations. This would be unavoidable, if we changed camera baselines along a line. So, with these configurations, we study the performance of feature extractors on stereo pairs. Furthermore, we have chosen to work on pure virtual scenes, which guarantees that we know the exact ground truth of 3D points position and their corresponding pixel positions, and use it for the assessment of reconstruction quality. The quality of AR applications is determined by the observer's relative position in 3D space. We, therefore assess the quality in the reconstructed 3D space, which seems more realistic for our scenario, than the usual re-projection error in 2D space.

The rest of the paper is organized as follows: section 2 describes other related feature evaluation studies. The evaluation system is explained in section 3 and the results are discussed in detail in section 4, along with the recommendations for designing 3D applications. Finally, we conclude by stating the usefulness of the evaluation study and outline the scope for future work, in section 5.

2. RELATED WORK

Previously, we have seen that the evaluation of most of the state-of-art feature extractors, i.e., detectors or descriptors, use various evaluation criteria. The feature detector KAZE [16] and feature descriptors FREAK [22] and BRIEF [21] evaluate themselves with other known feature

detectors using recall and precision metrics, which relates to a total number of correct feature matches found. Along with recall and precision, BRISK [15], STAR [19], FAST [20] and AKAZE [17], evaluate themselves in comparison to others, by the metric repeatability, which measures the extent of overlap between the detected regions in an image pair. In both SIFT [12] and SURF [13], the evaluation is carried out on various viewpoints, but not in comparison to other features. However, the performance criteria is still repeatability. Sometimes, the distance between the descriptors is considered to be an evaluation metric, as in ORB [14]. In all the above cases, the evaluation criteria focuses only on the correctness of the feature matches and this may not be enough to evaluate the feature extractors for accuracy in 3D applications and robustness to camera baseline changes.

Point feature matching algorithms for stereo were evaluated by Juhász et al. [23], but only for a particular baseline based on the re-projection error metric. In our paper, we evaluate a range of baselines to study their effects. Interest point detectors and descriptors were evaluated for tracking applications by Steffen et al. [24], where detectors were tested on various conditions such as scale, rotation, baseline, light, etc., using repeatability metric. Further, feature detectors were compared based on tracking success rate, which was computed based on the re-projection error. However, KAZE, AKAZE, BRISK, BRIEF and FREAK are not included in their study, unlike ours. Moreover, instead of measuring the re-projection error in 2D, we measure the accuracy in 3D space directly, relying on a dataset consisting of known 3D models. We believe that 3D space metrics are more suitable for AR related applications.

Michael et al. [25] evaluated SIFT feature extractors for viewpoint invariance, by comparing the descriptor properties over various baselines. Their evaluation basically outlines the quality of obtaining correct matches, but it does not guarantee high 3D accuracy.

Florian et al. [26] evaluated feature tracking for pose estimation in underwater environment. However, their evaluation is limited to very few feature detectors and descriptors with a very specific testing condition.

Pierre et al. [27] evaluated feature extractors for 3D object recognition applications over various viewpoints and lighting conditions, but with a limited number of candidates for evaluation.

Comparatively, in our paper, we evaluate a wide range of feature extractor combinations, to describe its capability for 3D applications directly, over various camera viewpoints.

3. EVALUATION SYSTEM

Our setup for evaluating feature extractors is depicted in figure 3. It comprises of the following steps: dataset generation, feature extraction, pose estimation, 3D estimation and 3D accuracy computation. The evaluation is carried out based on the accuracy of the 3D points that are estimated using the 2D test points, in comparison with the ground truth derived from the 3D model. Our experiment is implemented in C++ using the OpenCV (Open Source Computer Vision) library and results are presented using Matlab.

3.1 Dataset Generation

Ground truth data is generated based on the application scenario illustrated in figure 2. Here, we consider a number of possible positions where cameras can be placed around the

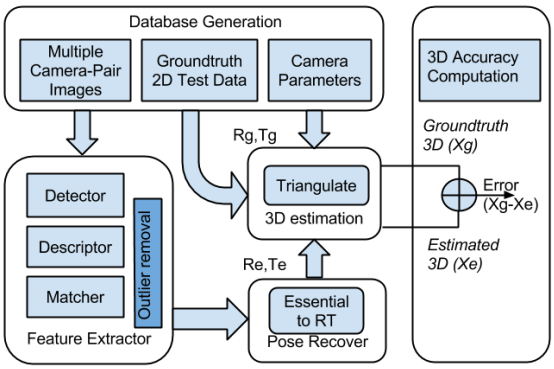


Figure 3: Experimental setup

3D model. Subsequently, we considered that many stereo camera pairs to capture images of a 3D model at various baselines (refers to relative displacement of stereo cameras). For every subsequent stereo pair, the camera motion is circularly displaced, maintaining equal distance from geometric center of the 3D object. This configuration is deliberately chosen so that scaling effects on feature extractors can be nullified and the focus stays on evaluating only baseline variation. Using 3D models is an advantage, in terms of having full control over the dataset being generated. The dataset is generated using a total of 9 3D models (depicted in figure 4 and obtained from CG Trader[3]), for baselines varying from 1 to 50 degrees angular displacement. This results in the necessary ground truth values as follows:

- Totally, 450 stereo pair images with 1 degree resolution, are generated for 9 models. Images are of resolution 600x600 with 24 bit depth.
- The ground truth 3D points are generated using four points, representing an origin and three points of unit length in three axes direction. These 3D points $[X_g]$ are sufficient to represent a model measured in world co-ordinate system, with the geometric center of the model as the origin. This type of 3D data is well suited as ground truth data, which is compared with the estimated 3D data, to compute the changes in the position and rotation in 3D space.
- The ground truth 2D feature points $[x_g^1$ and $x_g^2]$ in stereo pairs corresponds to the projection of true 3D points onto the image plane. This is considered as the 2D test data, which is used in the experiment to evaluate the feature extractors.
- The camera intrinsic parameters $[K]$ comprises camera's focal lengths (f_x,f_y) and principal axes (p_x,p_y). All cameras have identical intrinsics in all tests. In our experiment, the focal length is 520 pixels and the principal axes are 300 pixels.

$$K = \begin{bmatrix} f_x & 0 & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{bmatrix}$$

[3]http://www.cgtrader.com

Figure 4: 3D models used for the experiment. From each model, 50 stereo image pairs are generated, corresponding to various baselines.

- The camera extrinsic parameters represents relative rotation and translation of stereo pair ($[R_g,T_g]$).

$$R_{g_{3X3}} = \begin{bmatrix} r_{x1} & r_{y1} & r_{z1} \\ r_{x2} & r_{y2} & r_{z2} \\ r_{x1} & r_{y3} & r_{z3} \end{bmatrix}, T_{g_{3X1}} = \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}$$

3.2 Feature Extractors

The term feature extractor refers to a combination of state-of-art detector and descriptor. After feature extraction, the features are matched and outliers are removed.

We have tested feature extractors by combining the detectors SIFT, SURF, BRISK, KAZE, AKAZE, ORB, MSER, STAR and FAST, with their own descriptors, and combined with BRIEF and FREAK descriptors. In total, we evaluated 26 feature extractor combinations. To compute feature correspondences in a stereo pair, we applied a brute-force matcher on the descriptors, combined with Random Sample Consensus (RANSAC) [11] for removal of outliers. Each feature extractor was applied to every camera pair configuration to extract feature correspondences $[x_e^1, x_e^2]$ between the stereo images. All the state-of-art feature detectors and descriptors used for the evaluation in this paper are briefly explained with their properties in table 1.

3.3 Pose Recovery

In our tests, pose recovery estimates the pose (camera position and orientation) of the right camera with respect to the left camera in a stereo pair. Feature correspondences from the feature extractors on every stereo pair are used to estimate the camera pose $[R_e,T_e]$. Feature correspondences $[x_e^1, x_e^2]$ are used to estimate the essential matrix $[E_{ss}]$ directly, given the camera instrinsics $[K]$, by applying the 5-Point algorithm [28]. The essential matrix is a specialized case of fundamental matrix expressed in normalized image coordinates that describes the relation between the stereo pair in terms of epipolar constraint $[x_e^{2^T} E_{ss} x_e^1 = 0]$.

Finally, the camera pose is recovered using a single value decomposition, $E_{ss} = [T_e]R_e$, and selection of the optimal solution using the cheirality constraint [10]. Thereby, the estimated camera position is always upto scale expressed in model coordinates.

3.4 3D Estimation and Accuracy Computation

Usually, an estimated 3D point is projected onto a 2D image and compared with a known value to compute reprojection error, which represents the accuracy of the estimation. Instead of following this approach, we estimate the error in 3D space that is more comparable to real-time applications, using Normalized Correlation Coefficient (η).

For feature based calibration, in our tests, the feature extracted correspondences are consumed in estimating the camera pose. Using the same feature correspondences to estimate the 3D points is not a fair experiment to evaluate feature extractors for 3D applications.

Therefore, to evaluate feature extractors for feature based calibration, we compute 3D accuracy as a difference between the experimental data and the ground truth data. The ground truth 3D data (X_g) is obtained as a result of back projecting their corresponding ground truth 2D test data (x_g^1, x_g^2), using the ground truth camera pose (R_g,T_g). Similarly, experimental 3D data (X_e) is estimated from the same 2D ground truth test data (x_g^1, x_g^2) using the estimated camera pose (R_e,T_e). The back projection of feature corresponding points of two stereo pair is accomplished by triangulation [10]. Here, T_g & T_e are expressed upto scale, and all distances are always expressed in the model coordinates.

Thus, the 3D accuracy can be quantified as η, a measure over all three axes components between X_e and X_g. η provides a similarity measure of estimated 3D points with the ground truth 3D points, which is represented as a normalized accuracy value [0-low and 1-high].

$$\eta^\dagger = \frac{\sum(X_e^\dagger - mean(X_e^\dagger)) * (X_g^\dagger - mean(X_g^\dagger))}{\sqrt{\sum(X_e^\dagger - mean(X_e^\dagger))^2 * \sum(X_g^\dagger - mean(X_g^\dagger))^2}}$$

$$\eta = \sum_{\dagger=x,y,z} \frac{\eta^\dagger}{3}$$

where † represents 3D axes components x, y and z.

4. RESULTS AND DISCUSSION

The experiment described in section 3 is carried out on a total of 450(stereo pairs) * 26(feature extractor combinations), i.e., 11700 datasets. Our test results, which are based on virtual models in an empty scene, can be compared directly to a film scenario that applies blue screen, i.e. where the background consists of large, artificial, untextured surfaces. In other cases where textured background provides depth to the scene, our tests are relevant only for objects at certain depth. Other factors in real scenes, such as blur or challenging lighting conditions, are considered future work.

In our results, the "baseline" of the stereo camera pair is represented in terms of relative angular separation between the cameras, where both cameras are directly facing the 3D model and the camera movement with respect to each other is as in a turn-table configuration.

All combinations of feature extractors are evaluated at every stage in the pipeline (described in figure 3), i.e., 2D pixel error, camera pose error and 3D estimation error. As

Feature Extractor	Properties	Detection	Description
SIFT [12]	Scale and rotation invariant. Robust to change in illumination, 3D viewpoint and noise.	Interesting points are identified using Difference of Gaussian (DoG) over several linear scales of images. Then, the location and scale of keypoints are accurately computed using neighbor pixels.	The descriptor is represented by histograms of image gradients that are computed at every image point around the keypoints detected.
SURF [12]	Scale and rotation invariant. Features are distinctive, robust to noise, geometric and photometric deformations. It can be computed quickly.	Using integral images makes the image convolution faster. The detector is based on Hessian-matrix based approximation of blob-like interesting points using Gaussian scale space.	The descriptor is based on distribution of interesting points in its neighborhood. This is similar to SIFT but instead of using gradients, distribution of first order Haar Wavelets responses are used.
ORB [14]	Designed to perform two magnitudes faster than SIFT.	This is a FAST detector with addition of an accurate orientation component using intensity centroid.	"Rotation-Aware" binary descriptor based on the BRIEF descriptor. Computed by introducing a learning method for de-correlating the BRIEF features under rotational invariance.
BRISK [15]	Adaptive feature detector designed to lower computational complexity compared to SURF.	It is a combination of FAST detector in scale space and identifier of keypoints by fitting a quadratic function.	The descriptor is a bit-string assembly from intensity comparisons, retrieved by dedicated sampling of each keypoint neighborhood.
KAZE [16]	Scale and rotation invariant. Attains high accuracy in object boundaries. Robust to noise.	Similar to SIFT, except that the keypoints are detected in nonlinear scale space using "Additive Operator Splitting" techniques and variable conductance diffusion.	Uses a modified SURF descriptor, which adds a two-stage Gaussian weighting scheme.
AKAZE [17]	Accelerated KAZE - motivated to compute faster with similar scale and rotational invariance and lower storage requirement properties, compared to KAZE.	Instead of using non-linear scale space as in KAZE, a numerical scheme called "Fast Explicit Diffusion" in a pyramid framework is used.	A "Modified-Local Difference" binary descriptor, which exploits gradient and intensity information from nonlinear scale space.
MSER [18]	Affine-invariant feature extractor suitable for wide baselines in stereo. Robust to change in scale, illumination, out-of-plane rotation, occlusion and viewpoints.	Distinguished regions are detected and affine invariant procedure is carried out to estimate the stable invariant regions, from which the keypoints are measured.	n/a
STAR [19]	A suite of scale invariant center-surround detectors focused on visual odometry applications. Stable and repeatable in viewpoint changes. (CenSurE)	The CenSurE features are computed at the extrema over multiple scales using full image resolution using center-surround filters. There is an approximation to scale space based on Laplacian of Gaussian.	n/a
FAST [20]	High Speed corner detector extensively used in machine learning methods and is suitable for real-time applications.	Considers a circle comprising of 16 pixels in an image. Then every pixel is compared with only 4 neighbors to classify if it is a corner or not.	n/a
BRIEF [21]	A highly distinct binary descriptor designed to compute faster. Invariant to large in-plane rotation.	n/a	Binary string descriptor relying on image patches-pairwise intensity comparisons. A classifier is trained with image patches form various viewpoints.
FREAK [22]	Inspired by the human visual system - retina, this descriptor is a cascade of binary strings aimed at faster computation.	n/a	Computed by efficiently comparing image intensities over a retinal sampling pattern containing Gaussian kernel information.

Table 1: Brief overview of feature extractors that are used for feature based calibration.

a reference, 2D ground truth feature correspondences are passed through the pipeline with known camera parameters and reference plots at every stage are generated. These are referred to as "IDEAL" feature extractor combination, throughout the experiment. In this way, every step in the pipeline is tested as a black box to operate correctly.

The estimation error is expressed as averaged over every 5 degrees of the camera baseline. The variability of the error data within every 5 degrees is shown in figure 9. This variability makes it hard to present the comparison of the feature extractors visually. Therefore, the mean value was chosen to gain better readability.

4.1 2D pixel error

The 2D pixel error (P_{error}) is expressed as the squared Sampson error, which is the first-order approximation to the geometric error [10]. The P_{error} between feature points in a stereo pair is computed as in equation 1, where F is the fundamental matrix computed using N feature correspondences (x, x'). This metric determines how close every point in one image is to its corresponding epipolar line in the other image of the stereo pair. For an ideal match, $P_{error} = 0$.

$$P_{error} = \sum_{i=1}^{N_p} \frac{(x_i' F x_i)^2}{(Fx_i)_1^2 + (Fx_i)_2^2 + (F^T x\prime_i)_1^2 + (F^T x\prime_i)_2^2} \quad (1)$$

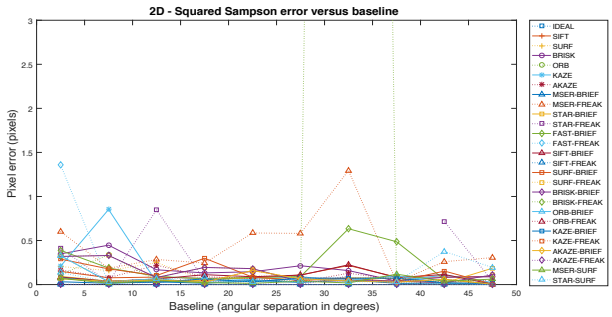


Figure 5: 2D error (Squared Sampson) based on epipolar constraint over varied baselines[4].

The P_{error} measured in 2D for stereo pairs varying in baseline is shown in figure 5. This error is computed for all meaningful combinations of feature extractors (described in the table 1). We can observe that the pixel error stays fairly low (although fluctuating) over all camera baselines. However, this does not guarantee a consistent accuracy of 3D estimation for all camera baselines as seen in figure 1(a). This is evident when we observe the effect of baseline variation on camera pose and 3D estimation error.

4.2 Camera pose error

Based on the estimated feature correspondences, the camera pose of stereo cameras are estimated. The pose estimated is compared with known camera extrinsics from the dataset (section 3.1), and thereby, the deviations of the estimated camera rotation and translation parameters from the ground truth value are computed. These deviations are the sum of deviations in all three axes, for both rotation and translation and are plotted in figures 6 and 7, respectively. Each figure is categorized into sub-figures based on the descriptors used, i.e., (i) figures 6(a) and 7(a) depict detectors having their own descriptors (with an exception for MSER and STAR, which uses SURF descriptor as in their original contribution), (ii) figures 6(b) and 7(b) depict detectors with BRIEF descriptor and (iii) figures 6(c) and 7(c) depict detectors with FREAK descriptor.

It is noticeable from the figures 6 and 7 that pose errors do not follow the same pattern as in figure 5. As the baseline of the stereo camera increases, the pose estimation error increases (figures 6(a), 6(b), 7(a) and 7(b)) or stays high throughout (figures 6(c) and 7(c)). This is observed to be due to the following reasons:

1. When wrong feature matches between the stereo pairs exist, the estimation of fundamental matrix becomes incorrect. This is quite obvious.
2. When correct feature matches between the stereo pairs exists, and if the feature matches are confined to a small area, i.e., a set of 2D match points corresponds to only a part of the 3D model, then the estimation of fundamental matrix becomes incorrect as there is not enough information about rotation or translation covering the whole 3d model.

In both of the above cases, an incorrect fundamental matrix and thereby an incorrect estimation of essential matrix results in an incorrect pose estimation. The 2D pixel error seems like a biased measure because the same number of feature points are used to both estimate fundamental matrix and to compute pixel error based on the fundamental matrix. Due to this nature, although we have an incorrect fundamental matrix, the 2D pixel error still stays low over all baselines (figure 5), as an effect of using RANSAC.

4.2.1 Penalty for invalidity

In the process of estimating camera pose, three types of invalidity can occur.

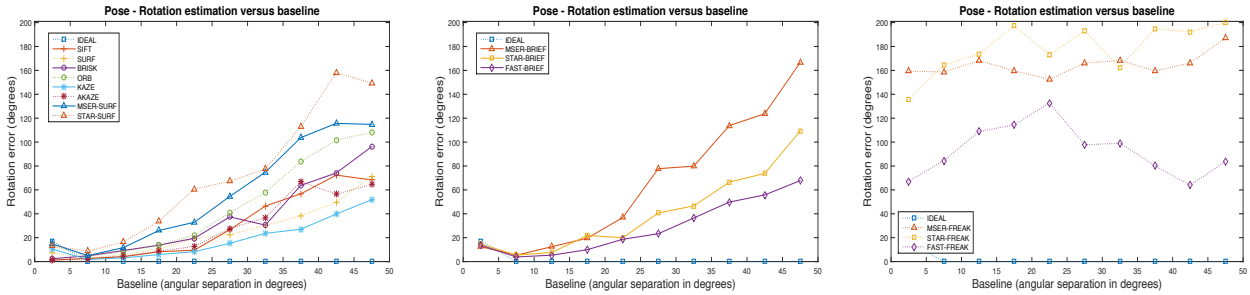
- Type 1 - when rotation error in either of the three directions is more than 90^o (as in figure 6(c)), then the camera seems to be rotated more than expected, in a true situation.
- Type 2 - as in figure 7(c), if any of the translation error is more than unity, then it means that the right camera is estimated to be on the left side.
- Type 3 - this is not directly related to pose estimation, but this error occurs when the feature extraction gives zero matches. This error also relates to non-estimation of fundamental matrix due to very few matches.

In the above cases, the camera pose estimation is deemed invalid. This situation can occur, when the number of feature correspondences in a stereo pair are zero or very few or wrong to a large extent. In these cases, we penalize the feature extractor, whenever any of the above types of invalidity occurs. Therefore, every feature extractor combination gets a penalty score for the invalidity.

In our tests, the penalties for every feature combination is given in figure 11. The maximum penalty score is 450, which represents samples that constitutes 9 models of 50 baselines each. It is clearly observable that most of the combinations with FREAK descriptor have higher penalty score.

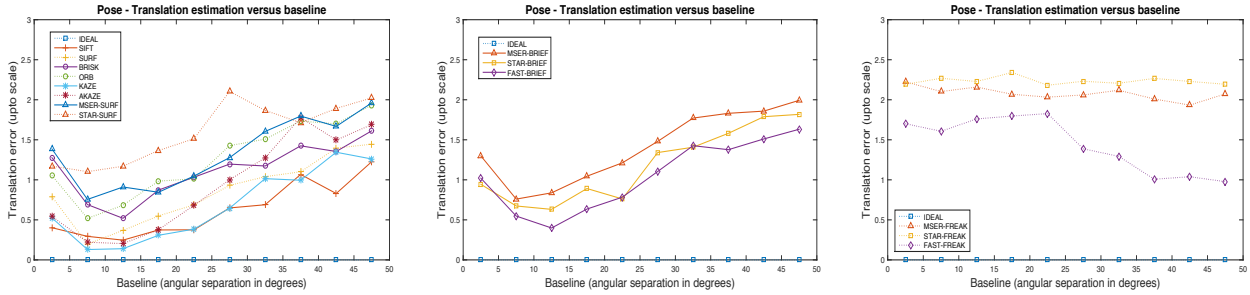
The sensitivity of the pose estimation can be observed by IDEAL features. The pose estimation seems to be sensitive

[4]The X-axis depicts baseline expressed between 1-50 degrees. Along the Y-axis, the error is averaged over every 5 degrees, to increase readability. Details in section 4.



(a) Own Descriptors (b) Brief Descriptor (c) Freak Descriptor

Figure 6: Mean estimation error of relative stereo camera rotation over varied camera baselines[4].



(a) Own Descriptors (b) Brief Descriptor (c) Freak Descriptor

Figure 7: Mean estimation error of relative stereo camera position over varied camera baselines[4].

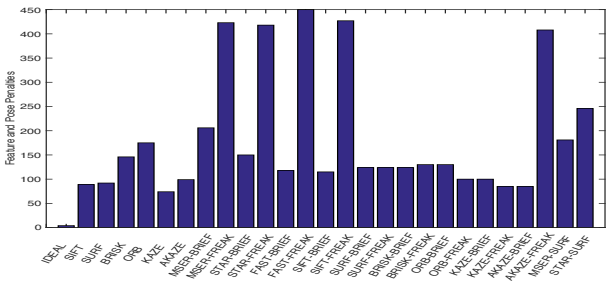


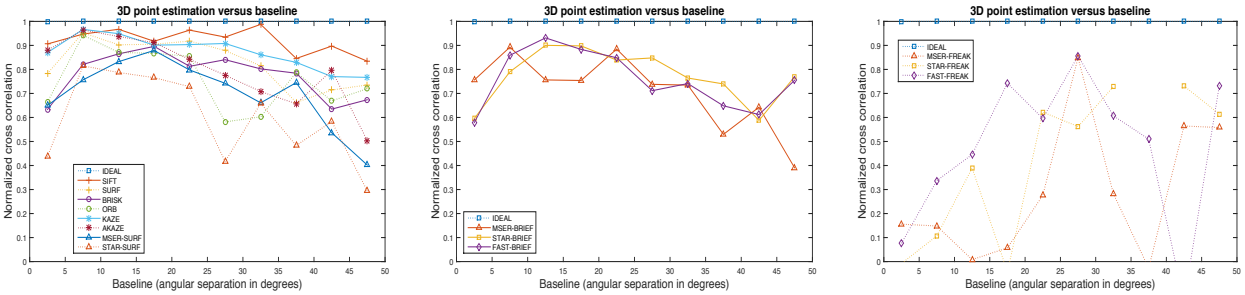
Figure 11: Penalty values for all feature extractors.

to rotation at low baselines (figure 6(a)). In figure 11, we see that IDEAL has about 4 penalties, and these are at very low baseline. This confirms that pose estimation algortihm has limitation at very low baselines. This sensitivity does not affect our comparative study on feature extractors as the penalty scored sample is considered invalid. However, we will use the penalty score to define the success rate or reliability of the feature extractor in further sections.

4.3 3D estimation error

Using the feature correspondences and the recovered camera pose, the corresponding 3D points are estimated and are compared to their ground truth values. The resulting samples are filtered based on the penalty score (described in section 4.2.1). Only the samples that are not penalized are considered valid and are used for further evaluation. The resulting 3D estimation error is plotted against varied baselines as shown in figure 8. In this figure, the 3D accuracy, expressed as normalized correlation coefficient (η), tends to reduce as the baseline of the camera increases. The error in camera pose propagates to 3D accuracy. 3D estimation is conceptually, the point of intersection of two rays back projected from a pair of feature matches. The back projection is carried out using the camera instrinsic and extrinsic (position and orientation) parameters. While camera instrinsics are maintained the same for the stereo pairs, the change in pose affects the 3D accuracy, i.e., lower the camera pose error, higher is the 3D accuracy. This is why, markerless pose estimation becomes important in 3D applications.

Figure 8(a) shows the performance of feature detectors using their own descriptors (SIFT, SURF, BRISK, ORB, KAZE, AKAZE). To compare the performance when other type of descriptors are used, we have evaluated each of these detectors with BRIEF and FREAK descriptors and the results are shown in figure 10. We have also evaluated other detectors such as MSER, STAR and FAST, which do not have their own descriptors, but using BRIEF and FREAK descriptors as shown in figures 8(b) and 8(c). In figure 8(a), we also include MSER and STAR detectors but with SURF descriptor, because they are evaluated based on SURF descriptor in [18] and [19], respectively. All the above mentioned feature extractor combinations are evaluated based of mean value of η over every 5 degrees, and there respective variances are shown in figure 9.



(a) Own Descriptors (b) Brief Descriptor (c) Freak Descriptor

Figure 8: Mean 3D estimation error (normalized correlation co-efficient) over varied camera baselines[4].

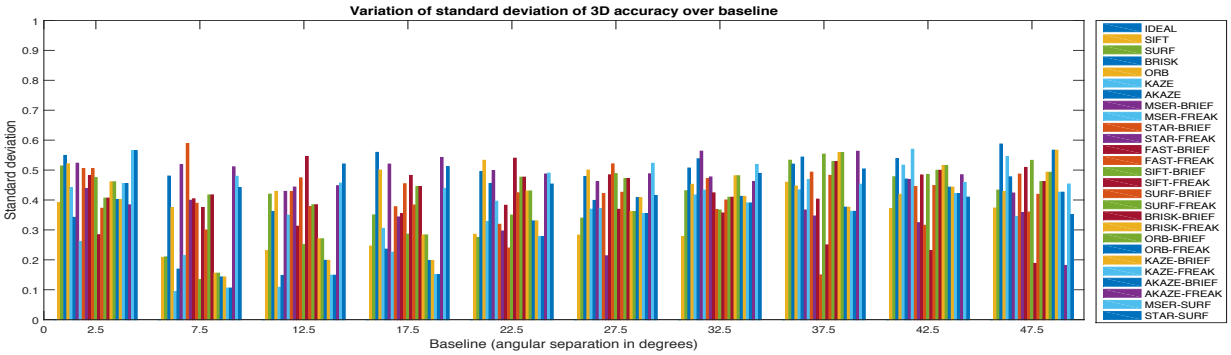


Figure 9: Standard deviation of 3D estimation over varied baselines[4].

The quality of feature extractors affect the 3D accuracy, but to what extent and how robust are they to large baselines, is what needs to be evaluated. Hence, we study the behavior and limitations of various feature extractors, especially for varied baselines. We shall now evaluate the performance of the feature extractors based on normalized cross correlation and discuss their application traits in terms of 3D mean squared errors, computation time and reliability.

4.4 Performance evaluation

At a very low baselines (less than $\approx 5^o$), the feature extractors seems to not perform very well. As explained before the sensitivity of pose estimation algorithm plays a role here. However, at very small baselines, even a small deviation in the accuracy of feature correspondences yields a large pose estimation error and thereby triangulation errors.

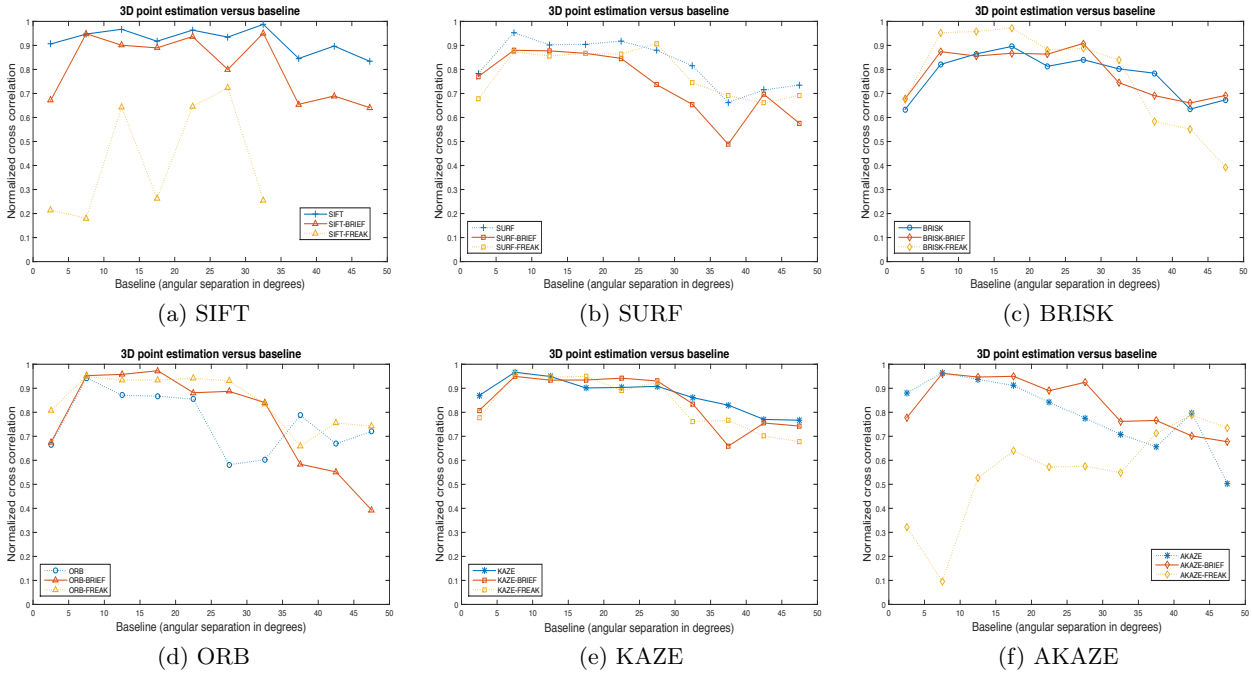
From figure 8(a), we can observe that KAZE (detector with its own descriptor), outperforms all other feature extractor combinations upto $\approx 20^o$. SIFT performs close to KAZE upto $\approx 20^o$, and thereafter outperforms KAZE at higher baselines. However, KAZE and SIFT both perform better than other feature extractors. The detectors KAZE and SIFT differ in the scale space representation, while the descriptor remains the same. As it is claimed in [16], KAZE performs as good as SIFT. However, it holds good only upto a limit specified.

The SURF and the ORB perform with almost equal accuracy upto $\approx 20^o$ baseline, and then SURF maintains the accuracy much better than ORB. Correspondingly, figures 6(a) and 7(a) show how the rotational and translational error of ORB increases after $\approx 20^o$ baseline and stays higher than SURF. This is probably because the modified BRIEF descriptor used in ORB is not as efficient as SURF descriptor, which is based on Haar wavelets, in terms of rotational invariance for higher baselines. ORB claims to be an alternative to SURF in [14], but we see that after the specified baseline limit, ORB cannot perform better than SURF.

Although AKAZE is shown to have better performance over other detectors (in [17]), we see that AKAZE performs as good as KAZE upto $\approx 20^o$ baseline and then, the performance drops down severely. Pose estimation error shows the same trend (figures 6(a) and 7(a)). However, by using AKAZE the computation time reduces comparatively.

The BRISK performs as good as ORB upto $\approx 20^o$ baseline, then seems to outperform ORB thereafter. The detectors BRISK and ORB are designed with a motivation to reduce computation time, but we notice that it is at the cost of reduction in 3D accuracy.

The MSER and STAR detectors have been evaluated using SURF descriptor in their original work. Therefore we intended to use these combinations as well. However, it seems that SURF descriptor is better off with its own detector rather than MSER or STAR. From figure 6(a), we can see that rotational errors are more prominent for MSER and STAR in combination with SURF descriptor. So, comparatively, SURF detector seems better than MSER and STAR.



(a) SIFT (b) SURF (c) BRISK

(d) ORB (e) KAZE (f) AKAZE

Figure 10: Mean 3D estimation error with varied baselines[4].

From figures 8(b) and 8(c), we can observe that MSER, STAR and FAST detectors perform with almost similar accuracy with two individual descriptors, BRIEF and FREAK. However, BRIEF descriptor seems to be well suited for these detectors compared to the FREAK descriptor. With BRIEF descriptor, STAR and FAST seem to perform with a similar pattern as in SURF atleast upto $\approx 25^o$, while BRIEF with MSER detector seems to match ORB, especially between $\approx 25^o - 40^o$ baselines and thereafter degrades. BRIEF descriptor is claimed to be as good as SURF descriptor in [21] and a modified BRIEF is used in ORB, and hence the similar performance pattern. The STAR detector seems to be better with BRIEF than SURF descriptor. The MSER detector with BRIEF and SURF descriptor shows similar performance pattern, however, SURF descriptor creation is faster than BRIEF.

On the other hand, all three detectors with FREAK descriptor in figure 8(c) seems to perform worse compared to the rest. From these observations, it is hard to generalize the behavior of BRIEF and FREAK descriptors when it is combined only with MSER, STAR and FAST detectors. So, we extended the descriptor evaluation with other detectors which basically have their own descriptors defined. Consequently, the respective results are shown in figure 10.

Feature extractors, such as SIFT and AKAZE, using the BRIEF descriptor (figures 10(a) & 10(f), respectively), maintain their accuracy similar to that of using their own descriptors, upto $\approx 35^o$ baseline. Moreover, using BRIEF descriptor is advantageous in terms of computation time.

The accuracy of SURF and KAZE stays almost the same when used with both BRIEF and FREAK descriptors as shown in figures 10(b) and 10(e), but again only upto $\approx 25^o$ baseline. So the possibility of making a choice of descriptor is higher for these detectors.

In case of BRISK and ORB, as shown in figures 10(c) and 10(d), both BRIEF and FREAK descriptors performs better than their own descriptor upto $\approx 35^o$. So, the BRIEF descriptor seems more robust to baseline changes than the modified BRIEF (used in ORB) and the BRISK descriptor.

So, the BRIEF descriptor seems to be a good choice in combination with BRISK, ORB, KAZE and AKAZE detectors for upto $\approx 35^o$ baseline. And, the FREAK descriptor is seemingly a good choice for BRISK and ORB for upto $\approx 35^o$. Moreover, FREAK descriptor could be the best choice for SURF and KAZE detectors, whose performance is comparable to SIFT and KAZE with their own descriptors.

Overall, some of the feature extractors have outperformed others and some of the descriptors have shown better performance when combined with certain detectors over others. The important aspect to notice here is that each feature extractor has performance relatively better in certain baseline range. From the evaluation of the state-of-art feature extractors, we can summarize the observed as follows:

- For baselines ($<5^o$):
  $SIFT$, $KAZE$ and $AKAZE$ seem to be good performers, however rotation-translation ambiguity exists.
- For baselines ($5^o - 30^o$):
  $SIFT$, $SURF$ and $KAZE$ with their own descriptors; $BRIEF$ descriptor with all detectors except $MSER$, $STAR$ and $FAST$; $FREAK$ descriptor with $SURF$, $BRISK$, ORB and $KAZE$ are good performers.
- For baselines ($>30^o$):
  $SIFT$ and $KAZE$ perform better than others. However, $SURF$ detector with both $SURF$ and $FREAK$ descriptors; $BRIEF$ descriptor with $BRISK$, $KAZE$ and $AKAZE$ are the next candidates.

4.5 Design recommendation

Although η gives a relative performance measure of feature extractors, it is difficult to use this information directly for practical applications. For making a sensible choice of feature extractors for a specific 3D application, feature extractors need an absolute measure that gives a sense of quality of service (QoS). The QoS depends of the type of application and its requirements. We therefore provide an extension to our evaluation of features based on QoS. We represent QoS in terms Mean Squared Error (MSE) of reconstructed 3D point positions & orientations, and reliability & computation time of the feature extractors.

The comparative observation of accuracy between feature evaluation based on η also holds good in the case of MSEs in most of the cases. However, one should not expect a direct relation because η measures the similarity and MSE measures euclidean distance between estimated and ground truth 3D points, at different baseline ranges.

Our ground truth data is represented as three unit vectors originating from the geometric center of the model. The positional and rotational changes in the 3D reconstructed points are computed as the deviations from the ground truth 3D points. This gives an idea of how the reconstructed 3D structure would be transformed in 3D space, due to the errors in feature based calibration, i.e., camera pose estimation. The reconstructed 3D points are observed to maintain the orthogonality of the 3D unit vectors randomly over various models tested under various baselines. This is because pose estimation algorithm [28] along with singular value decomposition does not yield perfect solution when singularities are present. However, this limitation of the pose estimator has a potential for further investigation, and is not in the scope of this paper.

The table 2 provides an overview of statistics of MSE of 3D points, for three categories of baseline ranges - *Small* (5^o-20^o), *Medium* (20^o-35^o) and *Large* (35^o-50^o). The MSE is expressed in the 3D model units for positional deviation and in degrees for rotational deviation. The table also specifies the computation time required by the feature extractors, which is relevant information for real-time applications.

As explained in section 4.2.1, we have filtered the invalid data occurred during pose estimation and noted down the penalties. These penalties correspond to the success rate of the feature extractor over several samples on all baseline ranges. Therefore, we use the penalties to represent the "Reliability" of the feature extractor, which shows the probability of success over 450 samples. This parameter is also reflected in table 2. The comparisons made so far in relation to η or MSE is at the cost of reliability of every feature extractor. Hence, the reliability parameter in the table becomes very important apart from accuracy and computation time, in making a choice of feature extractor.

The result shown in the table is useful for any 3D application, which uses markerless camera pose estimation. Some relevant applications for discussion are the AR applications such as head mount display systems [1, 2], mobile applications [3, 4, 5], interactive systems [6, 7] and free view rendering application such as [8]. All these applications rely on markerless camera pose estimation, where the accuracy of the camera pose estimated becomes really important. Some applications demand real-time performance as well. The camera placements vary from small to large baseline range in these applications. Hence, our study of feature extractors and their evaluation based on various baselines for 3D error in terms of position and orientation is very helpful for such applications.

Let us consider an application scenario using *Small* baseline range and a feature extractor is required to be chosen. From the table, both KAZE and AKAZE have good accuracy in terms of 3D position and rotation, but one may choose AKAZE if the application demands fast computation time. However, this choice is at the cost of reliability, because KAZE seems to be more reliable than AKAZE. On the other hand, AKAZE+BRIEF offers accuracy similar to KAZE and is equally reliable, moreover, much faster than KAZE. So, in this case, the application could choose AKAZE+BRIEF.

Now, let us consider another application, where number of cameras around an object needs to be determined using KAZE (assuming KAZE is chosen for its high reliability). Here, KAZE offers the best positional accuracy at *Medium* baseline range. Say, if we consider a baseline of about 30^o, then number of cameras required to capture an object in 360^o, is about 12. On the other hand, if one can compromise on the positional accuracy slightly, at the same time gain higher rotational accuracy, one would choose to operate with KAZE at *Large* baseline range. In this case, for a baseline of about 45^o, one could capture the same object with only 8 cameras, which is more cost effective for applications.

In this way, table 2 can be used as a recommendation for practical 3D applications, where one can either choose feature extractors or estimate the camera density around the object of interest, based on the desired quality of service.

5. CONCLUSION

In this paper, we focused on stereo vision for 3D applications such as AR and free-view rendering, where the accuracy of position and orientation of 3D points play an important role. This paper is motivated by claiming that low 2D pixel error does not guarantee good 3D accuracy, however, 3D accuracy is dependent on the quality of feature based calibration (FBC). One of the major factors determining the quality of FBC is the camera baseline.

We designed an experiment to evaluate 26 feature extractor combination and discussed the comparative study of feature extractors over 50 camera baselines. We observed that each of the feature extractors had a certain operating range for various baseline range. However, the performance of SIFT and KAZE seemed promising, in terms of accuracy and robustness to large camera baselines.

Finally, we provided a recommendation for practical 3D applications, as in table 2, which specifies quality of service in terms of 3D position & orientation accuracy of reconstructed 3D points and computation time & reliability of feature extractors. This information is very useful for the 3D application designers, which will enable them to:

1. Select the feature extractor based on an acceptable accuracy or an acceptable execution time, with a cost of reliability.

2. Decide the camera density required to capture an object of interest, for a desired quality of service.

We believe that the system built for the movie production scenario (POPART), will benefit from our recommendations, by gaining the ability to preview integrated scene

Feature extractors	**Baseline**($5^o - 20^o$)		**Baseline**($20^o - 35^o$)		**Baseline**($35^o - 50^o$)		**Time**	**Relia-bility**
	Rotation [degrees]	***Position*** [model]	***Rotation*** [degrees]	***Position*** [model]	***Rotation*** [degrees]	***Position*** [model]	[seconds]	[percent]
	mean(deviance)		mean(deviance)		mean(deviance)			
IDEAL	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00	99.11
SIFT	13.09 (7.17)	8.23 (1.96)	2.14 (1.06)	2.83 (2.64)	2.64 (0.56)	4.22 (1.11)	17.34	80.22
SURF	15.58 (5.94)	12.04 (2.26)	5.59 (0.64)	6.27 (0.30)	3.63 (0.89)	5.33 (0.80)	5.47	79.56
BRISK	20.21 (8.94)	25.31 (13.48)	6.43 (2.41)	18.73 (13.03)	3.82 (0.52)	88.69 (141.00)	1.75	67.56
ORB	21.04 (9.31)	9.41 (0.47)	8.29 (1.54)	33.30 (41.34)	3.93 (0.05)	9.22 (6.91)	0.85	61.11
KAZE	12.12 (3.84)	7.76 (2.23)	4.78 (1.25)	6.34 (1.64)	2.92 (0.26)	7.27 (2.13)	27.67	83.56
AKAZE	11.68 (2.76)	6.91 (3.97)	7.51 (0.74)	12.36 (5.03)	4.61 (1.22)	14.48 (13.06)	4.96	78.00
MSER-SURF	19.95 (11.36)	261.94 (422.71)	8.12 (1.33)	20.55 (10.17)	5.10 (0.19)	16.09 (12.27)	7.55	59.78
STAR-SURF	29.67 (12.08)	53.51 (47.55)	11.34 (3.75)	16.08 (7.11)	7.08 (0.98)	22.15 (8.20)	0.75	45.33
MSER-BRIEF	24.01 (4.87)	44.64 (45.12)	7.03 (1.17)	7.86 (3.64)	6.25 (1.77)	8.88 (5.47)	2.50	54.22
STAR-BRIEF	21.30 (11.42)	154.87 (254.05)	6.52 (0.22)	7.58 (1.61)	4.99 (0.89)	6.18 (3.66)	0.65	66.67
FAST-BRIEF	18.00 (9.78)	24.70 (9.83)	7.37 (1.36)	8.65 (2.04)	5.12 (1.48)	9.53 (4.65)	4.73	73.78
SIFT-BRIEF	14.00 (2.80)	33.85 (27.00)	4.98 (1.65)	7.82 (7.56)	4.81 (0.49)	7.68 (0.54)	7.75	74.44
SURF-BRIEF	18.70 (7.20)	16.40 (3.54)	8.93 (1.07)	270.54 (446.17)	5.63 (1.25)	11.04 (3.82)	3.22	72.44
BRISK-BRIEF	20.10 (8.38)	21.75 (17.02)	6.46 (1.57)	22.14 (25.46)	4.91 (1.00)	49.14 (46.38)	3.76	72.44
ORB-BRIEF	15.70 (10.42)	4.79 (1.70)	4.84 (0.42)	7.61 (4.39)	4.41 (0.46)	59.16 (80.32)	0.80	71.11
KAZE-BRIEF	13.63 (7.43)	38.77 (47.78)	4.33 (0.43)	4.64 (0.34)	4.18 (0.96)	16.78 (10.96)	21.12	77.78
AKAZE-BRIEF	12.86 (6.91)	10.07 (6.27)	5.37 (2.04)	16.56 (10.51)	4.45 (0.20)	8.52 (1.14)	4.48	81.11
MSER-FREAK	61.67 (4.78)	60.57 (48.24)	15.67 (9.97)	2.77 (2.04)	8.72 (3.38)	11.38 (18.21)	7.29	6.00
STAR-FREAK	52.15 (29.05)	9.95 (6.09)	15.82 (6.92)	1.49 (0.29)	4.43 (0.00)	0.74 (0.11)	1.13	7.11
FAST-FREAK	52.70 (21.19)	8.42 (9.06)	9.50 (3.29)	23.62 (39.51)	6.38 (3.73)	11.14 (17.77)	6.09	0.00
SIFT-FREAK	51.65 (11.97)	5.74 (5.23)	22.41 (14.61)	30.14 (50.25)	10.78 (0.00)	3.78 (0.00)	9.29	5.11
SURF-FREAK	20.10 (8.38)	21.75 (17.02)	6.46 (1.57)	22.14 (25.46)	4.91 (1.00)	49.14 (46.38)	3.23	72.44
BRISK-FREAK	15.70 (10.42)	4.79 (1.70)	4.84 (0.42)	7.61 (4.39)	4.41 (0.46)	59.16 (80.32)	1.21	71.11
ORB-FREAK	13.63 (7.43)	38.77 (47.78)	4.33 (0.43)	4.64 (0.34)	4.18 (0.96)	16.78 (10.96)	21.10	77.78
KAZE-FREAK	12.86 (6.91)	10.07 (6.27)	5.37 (2.04)	16.56 (10.51)	4.45 (0.20)	8.52 (1.14)	7.88	81.11
AKAZE-FREAK	53.24 (25.05)	28.34 (37.48)	14.56 (8.09)	4.46 (5.23)	6.85 (0.15)	5.16 (5.38)	9.13	9.33

Table 2: Practical recommendation for 3D applications. [Here "Rotation" is the mean 3D rotational change (expressed in degrees) and "Position" is the mean 3D positional shift (expressed in model units), of all the estimation 3D unit vectors that represent a model in 3D space.]

more accurately in real time or decide better camera positions, and thereby ease their post-production tasks.

In the future, we would like to continue to explore the factors affecting the quality of camera pose estimation, especially the spatial distribution of feature correspondences in the stereo pair and also, evaluate the feature extractors for their invariance to illumination changes. It could also be interesting to study the limitations of the pose estimation algorithms, in general.

6. REFERENCES

[1] Chunrong Yuan. Markerless pose tracking for augmented reality. In *Proc. of ISVC*, pages 721–730, 2006.

[2] Miguel Ribo, Axel Pinz, and Anton L. Fuhrmann. A new optical tracking system for virtual and augmented reality applications. In *Proc. of IEEE - IMTC*, pages 1932–1936, 2001.

[3] Stéphane Bres and Bruno Tellez. Localisation and augmented reality for mobile applications in cultural heritage . In *Proc. of Workshop 3D ARCH*, 2009.

[4] Victor Fragoso, Steffen Gauglitz, Shane Zamora, Jim Kleban, and Matthew Turk. TranslatAR: A mobile augmented reality translator. In *Proc. of IEEE - WACV*, pages 497–502, 2011.

[5] Jonathan Ventura and Tobias Höllerer. Wide-area scene mapping for mobile visual tracking. In *Proc. of ISMAR*, pages 3–12, 2012.

[6] João Paulo Lima, Francisco Simões, Lucas Figueiredo, and Judith Kelner. Model based markerless 3d tracking applied to augmented reality. *Journal on 3D Interactive Systems*, 1, 2010.

[7] Hideyuki Suenaga, Huy Hoang Tran, Hongen Liao, Ken Masamune, Takeyoshi Dohi, Kazuto Hoshi, and Tsuyoshi Takato. Vision-based markerless registration using stereo vision and an augmented reality surgical navigation system: a pilot study. *BMC Medical Imaging*, 15(1):1–11, 2015.

[8] DongBo Min, Donghyun Kim, SangUn Yun, and Kwanghoon Sohn. 2d/3d freeview video generation for 3dtv system. *Signal Processing: Image Communication*, 24(1-2):31–48, 2009.

[9] Chris Aimone, James Fung, and Steve Mann. An eyetap video-based featureless projective motion estimation assisted by gyroscopic tracking for wearable computer mediated reality. *Springer - Personal and Ubiquitous Computing*, 7(5):236–248, 2003.

[10] Andrew Hartley and Andrew Zisserman. *Multiple view geometry in computer vision (2. ed.)*. Cambridge University Press, 2006.

[11] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *ACM - Communications*, 24(6):381–395, 1981.

[12] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc J. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[14] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: an efficient alternative to SIFT or SURF. In *Proc. of IEEE - ICCV*, pages 2564–2571, 2011.

[15] Stefan Leutenegger, Margarita Chli, and Roland Siegwart. BRISK: binary robust invariant scalable keypoints. In *Proc. of IEEE - ICCV*, pages 2548–2555, 2011.

[16] Pablo Fernández Alcantarilla, Adrien Bartoli, and Andrew J. Davison. KAZE features. In *Proc. of ECCV (Part VI)*, pages 214–227, 2012.

[17] Pablo Fernández Alcantarilla, Jesus Nuevo, and Adrien Bartoli. Fast explicit diffusion for accelerated features in nonlinear scale spaces. In *Proc. of BMVC*, 2013.

[18] Jiri Matas, Ondrej Chum, Martin Urban, and Tomás Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. of BMVC*, pages 1–10, 2002.

[19] Motilal Agrawal, Kurt Konolige, and Morten Rufus Blas. Censure: Center surround extremas for realtime feature detection and matching. In *Proc. of ECCV (Part IV)*, pages 102–115, 2008.

[20] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In *Proc. of ECCV (part I)*, pages 430–443, 2006.

[21] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: binary robust independent elementary features. In *Proc. of ECCV (Part IV)*, pages 778–792, 2010.

[22] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. FREAK: fast retina keypoint. In *Proc. of IEEE - CVPR*, pages 510–517, 2012.

[23] Endre Juhász, Attila Tanács, and Zoltan Kato. Evaluation of point matching methods for wide-baseline stereo correspondence on mobile platforms. In *Proc. of ISPA*, pages 813–818, 2013.

[24] Steffen Gauglitz, Tobias Höllerer, and Matthew Turk. Evaluation of interest point detectors and feature descriptors for visual tracking. *Int. J. Comput. Vision*, 94(3):335–360, 2011.

[25] Michael Yin Yang, Yanpeng Cao, and John McDonald. Fusion of camera images and laser scans for wide baseline 3d scene alignment in urban environments. *Journal of Photogrammetry and Remote Sensing*, 66(6):S52–S61, 2011.

[26] Florian Shkurti, Ioannis Rekleitis, and Gregory Dudek. Feature tracking evaluation for pose estimation in underwater environments. In *Proc. of CRV*, pages 160–167, 2011.

[27] Pierre Moreels and Pietro Perona. Evaluation of features detectors and descriptors based on 3d objects. *Interntaional Journla of Computer Vision*, 73(3):263–284, 2007.

[28] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions - Pattern Analysis and Machine Intelligence*, 26(6):756–777, 2004.

Chapter 12

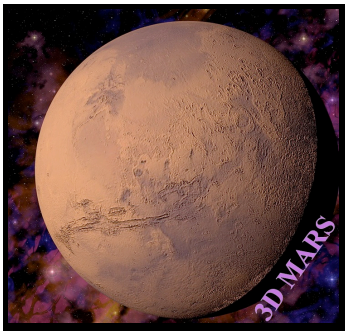
Poster I: 3-D Video Processing for Mixed Reality Art Performances

Title: 3-D Video Processing for Mixed Reality Art Performances.

Authors: Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz and Pål Halvorsen.

Presented: In VERDIKT 2009, Norwegian Research Council, Oslo, Norway (BEST POSTER AWARD 2009).

Demo: A live image of the audience was project onto the poster. Then we tracked the head movements of the audience using the webcam. Based on the head movement, we modified the live image using perspective transform to exhibit the effect of looking outside a window. We used short throw projector to display the image on the poster.



3-D Video Processing for Mixed Reality Art Performances

Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz, Pål Halvorsen

[simula ▪ research laboratory]

Motivation

World Opera envisions a distributed stage performance (Real & Virtual artists perform on a single stage). Project 'Verdione – Virtually Enhanced Reallife synchroniseD Interaction On the Edge' is motivated in constructing a suitable platform to provide realistic experience with high video quality, in real-time. The visual part of this project is designated to the research topic '3D Video Processing for Mixed Reality Art Performances', which focuses on providing real experience of a physical activity at remote location.

Research Goal & Challenges

The research aims at developing a 3D Multiview Acquisition and Rendering System (3DMARS) with low latency, high resolution and robustness. Accordingly, the objective is to develop algorithms, techniques and methods for calibration & depth estimation of multi camera system and for 3D reconstruction & rendering free viewpoint video for display system. Some of the challenges identified in this research are large volume of 3D space, synchronization of cameras, backdrop estimation, marker-less tracking, illumination & shadow effects, parallax and occlusions.

D E M O

D E M O

Problems & Methodology

Region of Interest (ROI): very important step in determining good features to extract and process. This process also involves segmentation, shape estimation and tracking of ROI.

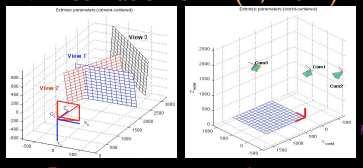
Calibration: accurate geometric camera calibration is necessary for efficient reconstruction. This process determines parameters:
- Intrinsic - focal length (f), distortion (s), principal axes (o_x,o_y)
- Extrinsic - relative positions (T) and orientations (R)

Depth estimation: is expected to be accurate in order to obtain robustness in the rendering of captured video.
- Disparity between multiple views provide depth estimates
- Variations in depth estimation results in structural artifacts

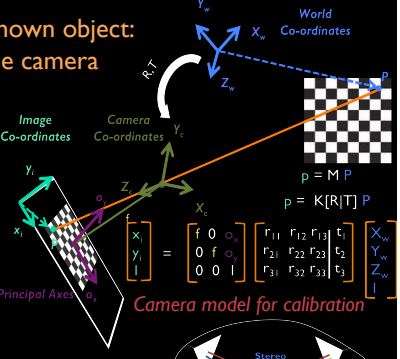
Rendering: accounts for high quality view synthesis, which involves
- 3D reconstruction using shape and depth information
- Free viewpoint rendering via interpolation & warping
- Suitable display is used for rendering video streams.

Experimentation & Initial results

Single Camera Calibration with known object:
- Three views are shown to the camera
- P (X_w , Y_w , Z_w) are known
- Corner detection p (x_i, y_i)
- Estimation of M (K, R & T)

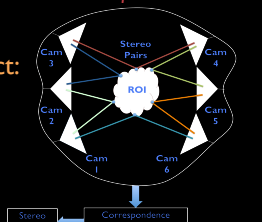


Extrinsic parameter visualization

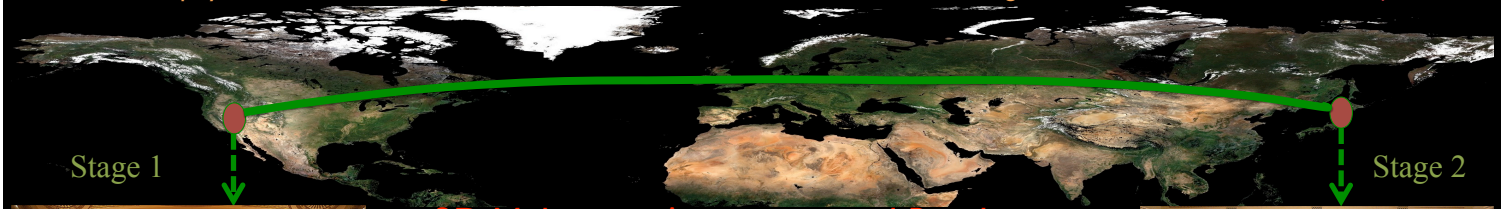


Camera model for calibration

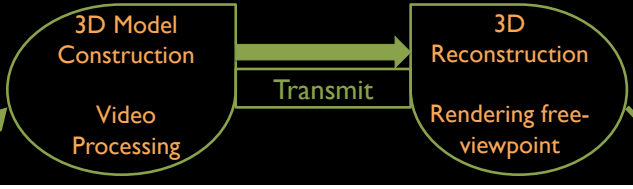
Multi Camera Calibration with unknown object: This is the proposed step forward.
- Two cameras look at the same scene
- Matching correspondences
- Computation of fundamental matrix
- Estimation of each camera's extrinsic relative to their neighbors

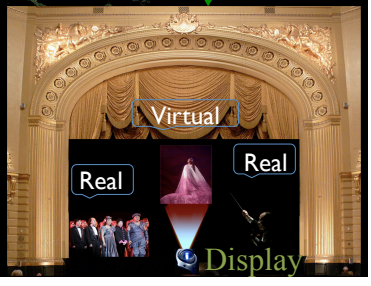


Multi Array Calibration



3D Multi-view Acquisition and Rendering System (3DMARS)





Chapter 13

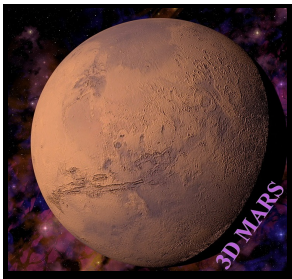
Poster II: 3D Multi-view Acquisition and Rendering System

Title: 3D Multi-view Acquisition and Rendering System.

Authors: Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz and Pål Halvorsen.

Presented: In VERDIKT 2010, Norwegian Research Council, Oslo, Norway (BEST POSTER AWARD 2010).

Demo: Interactive poster. When the user pointed his hand at a certain diagram on the poster, then the diagram was projected on a display wall. This was achieved using the webcam, by which, the hand of the user was detected by simple hand detection algorithm. A short throw projector was used to display the images.



3D Multi-view Acquisition and Rendering System

Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz, Pål Halvorsen

Project: Verdione [simula ▪ research laboratory]

Application

The research topic is a part of realizing the project 'Verdione – Virtually Enhanced Reallife synchroniseD Interaction On the Edge', which provides the technology for mixed reality performances. World Opera envisions such networked performance in a distributed fashion over various physical stages/locations and their interactions.

Highlights:

- Distributed performance using the existing network technology overcomes limitations of physical presence.
- Seamless projection of remote artists on physical stages gives a realistic experience.
- Interaction between real and virtual artists on various physical spaces creates new dimension of experience.

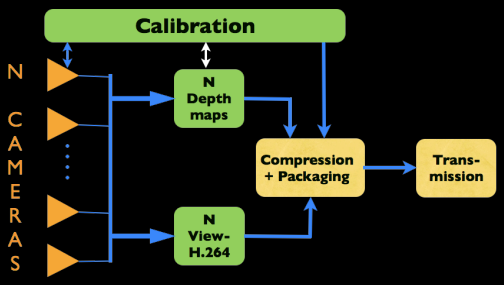
Research

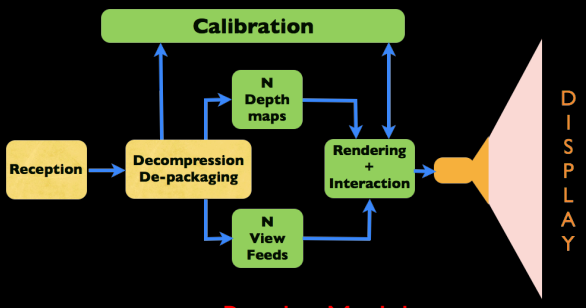
The research aims at developing a 3D Multiview Acquisition and Rendering System (3DMARS) and the current challenges are:

System Context
Low latency, High resolution and Robustness

Multi-view Acquisition Context
Accurate calibration, Precise Depth Estimation, Large Volume spaces, Ambient light, Occlusions, Shadows

Free-view Rendering Context
Precise 3D reconstruction, Seamless multiple viewpoint rendering, Low delay interactive cues

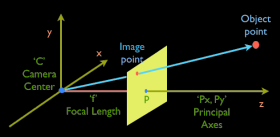




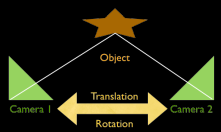
Capture Model

Calibration: Accurate estimate of camera parameters

Single Camera



Stereo



Multi Camera

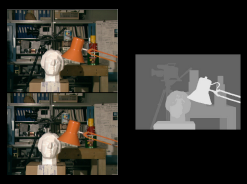
'N' number of cameras stereo calibrated to estimate relative positions of all cameras

Camera Projector Pair

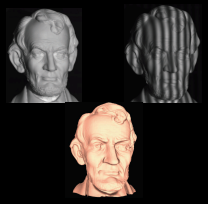
Projector used as inverted camera to estimate relative positions of projector & camera

Depth estimation: Accurate estimate of 3D from 2D images

Stereo Disparity



Structured Light



Render Model

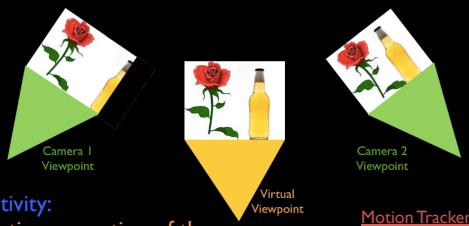
Free-View Rendering: Virtual viewpoint is obtained using 'N' views

3D Reconstruction (implicit geometry)

known Correspondences of 'N' views – to estimate virtual viewpoint
Example: Interpolation and Morphing techniques

3D Reconstruction (explicit geometry)

known Depth maps of 'N' views – to estimate virtual viewpoint
Example: 3D Warping, Layered Depth Images



Interactivity:
Perspective correction of the virtual (projected) performer changes according to motion of real performer on the stage.

Motion Trackers

Head pose detection
Face detection
Eyes tracking
Gaze tracking

Chapter 14

Poster III: Multiple Camera Arrays for Real-time 3D Rendering Systems

Title: Multiple Camera Arrays for Real-time 3D Rendering Systems.

Authors: Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz and Pål Halvorsen.

Presented: In VERDIKT 2012, Norwegian Research Council, Oslo, Norway (BEST POSTER AWARD 2012).

Demo: Video poster presentation[1].

[1] 3D Video Poster - `https://www.youtube.com/watch?v=DteUb-chUqk`

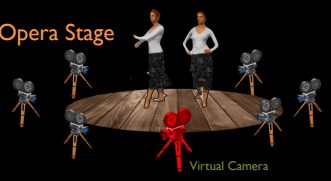
Multiple Camera Arrays for Real-time 3D Rendering Systems

Deepak Dwarakanath, Alexander Eichhorn, Carsten Griwodz, Pål Halvorsen

Project:Verdione [Simula Research Laboratory]

Camera Array Setup

Opera Stage



2D Images to 3D objects

Problem Statement

Such 3D systems are:
Highly Accurate?
Real-time?
Robust?

Research Problem

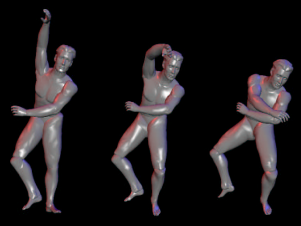
Feature Matches → Camera Calibration

3D Applications

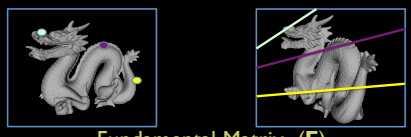
Free-view or Virtual-view synthesis

Visual Hull Reconstruction

3D reconstruction



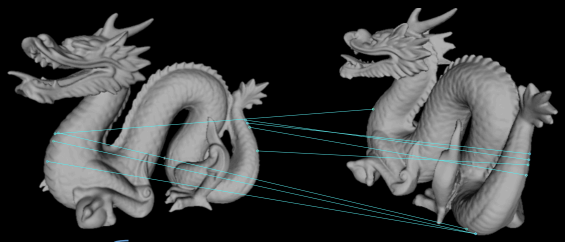
Feature Correspondence



Fundamental Matrix (**F**)

Point (image 1) — mapping — Line(Image 2)

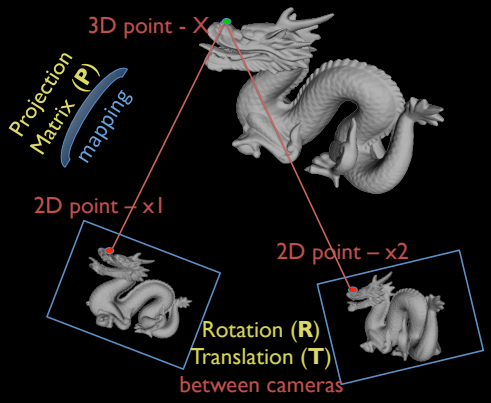
Feature - distinguishable keypoints detected in an image
Descriptor - representation of features detected
Matching - feature correspondence between two images



Popular Feature Extractors:
- SIFT – Scale Invariant Feature Transform
- SURF – Speeded Up Robust Feature
- MSER – Maximally Stable Extremal Region

Camera Calibration

Calibration – estimation of Projection matrix P = K[R|T]

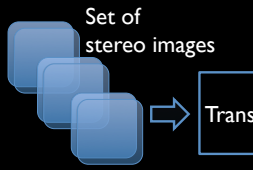


Fundamental matrix can be decomposed to estimate R & T

Therefore, $x = PX \Rightarrow x = K[R|T]X$

$$\begin{bmatrix} x_{il} \\ y_{il} \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & o_x \\ 0 & f & o_y \\ 0 & 0 & 1 \end{bmatrix} \left[\begin{array}{ccc|c} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{array}\right] \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}$$

Experimentation and Results



Set of stereo images → Transformation → Feature Based Calibration

Performance Evaluation
- -Accuracy
- -Computation Time
- -Robustmness

Internal parameters
- Resolution
- Distortion
- Noise
- Blur

External variation
- Position b/n cameras
- Orientation b/n cameras
- Object Distance and Size
- View angle