FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Evaluation of a 3D Reconstruction System Comprising Multiple Stereo Cameras

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Computational Intelligence

eingereicht von

Christian Kapeller, BSc.
Matrikelnummer 0225408

an der Fakultät für Informatik
der Technischen Universität Wien
Betreuung: Ao.Univ.-Prof. Dipl.-Ing. Mag. Dr. Margrit Gelautz

Wien, 13. November 2018

Christian Kapeller	Margrit Gelautz

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Evaluation of a 3D Reconstruction System Comprising Multiple Stereo Cameras

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Computational Intelligence

by

Christian Kapeller, BSc.
Registration Number 0225408

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.-Prof. Dipl.-Ing. Mag. Dr. Margrit Gelautz

Vienna, 13th November, 2018

Christian Kapeller	Margrit Gelautz

Erklärung zur Verfassung der Arbeit

Christian Kapeller, BSc.

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 13. November 2018

Christian Kapeller

Acknowledgements

I would like to express my deep gratitude to my supervisor, Professor Margrit Gelautz, for her generous support and her insightful feedback throughout the creation of this work.

I also thank my colleagues Braulio Sespede and Christine Mauric at Vienna University of Technology for their assistance. Braulio participated in numerous fruitful discussions and enthusiastically helped with the code for novel view evaluation and the user study. Christine was brave enough to help me in correcting the present text. I would further like to thank my colleagues at emotion3D GmbH, Matej Nezveda and Dr. Florian Seitner, for their feedback, providing hardware and assistance in acquiring the data sets. Also big thanks to my colleagues at Rechenraum e.U., Matthias Labschütz and Dr. Simon Flöry, for their input and help in generating the 3D models used in this work, and for fitting spheres into the validation object data.

I gratefully acknowledge the funding provided by the Austrian Research Promotion Agency (FFG) and the Austrian Ministry BMVIT. This work has been carried out as part of the project *Precise3D* (grant no. 6905496) under the program ICT of the Future.

A big thank you to all friends and other people, who participated in the user study and bore the trials.

Above all, I would like to thank my parents Gerhild Kapeller and Marcus Hartmann for my existence, for making it possible for me to study and for supporting me in becoming the person that I am. Thanks for your kindness and patience. You did a great job.

Kurzfassung

Jüngste Fortschritte in den Bereichen Medienproduktion und Mixed/Virtual-Reality erzeugen zunehmenden Bedarf nach qualitativ hochwertigen 3D Modellen realer Szenen. Mehrere 3D Rekonstruktionsmethoden inklusive Stereo Vision können zur Berechnung von Bildtiefe angewendet werden. Generell kann die Genauigkeit von Stereo Matching Algorithmen mit etablierten Benchmarks und öffentlich zugänglichen Referenzlösungen ermittelt werden. Im Gegensatz zu üblichen Bildsensorkonfigurationen, bedarf die Evaluierung von Daten aus 3D Rekonstruktionssystemen mit einem speziellen Aufbau der Entwicklung neuer oder adaptierter, auf das jeweilige System zugeschnittener, Bewertungsstrategien. Diese Arbeit befasst sich mit der Bewertung von Qualität und Genauigkeit von 3D Modellen, die mit einem 3D Rekonstruktionssystem bestehend aus drei Stereokameras erzeugt wurden. Dazu werden drei verschiedene Evaluierungsmethoden vorgeschlagen und umgesetzt. Zunächst wird die Genauigkeit von 3D Modellen mittels geometrisch einfacher, speziell zu diesem Zweck erstellter, Körper (Kugel, Quader) ermittelt. Entsprechende ideale 3D Objekte werden in rekonstruierte Punktwolken eingepasst und mit den echten Maßen verglichen. Weiters bestimmt eine bildbasierte Novel View Evaluierung die Genauigkeit verschiedener Rekonstruktionsmethoden bei Punktwolken und finalen 3D Netzmodellen. Zuletzt ermittelt eine paarvergleichsbasierende Studie die subjektive Qualität verschiedener Rekonstruktionsverfahren anhand selbst erstellter texturierter 3D Netzmodelle. Wir demonstrieren die drei Evaluierungsverfahren anhand selbst erstellter Daten. In diesem Kontext beobachten wir, dass sich die Genauigkeit der betrachteten Methoden in der Novel-View Evaluierung nur leicht voneinander unterscheidet, die Resultate der Benutzerstudie jedoch eindeutige Präferenzen zeigen. Dies bestätigt die Notwendigkeit quantitative mit qualitativen Evaluierungsmethoden zu verbinden.

Abstract

Recent advances in the fields of media production and mixed/virtual reality have generated an increasing demand for high-quality 3D models obtained from real scenes. A variety of 3D reconstruction methods including stereo vision techniques can be employed to compute the scene depth. Generally, the accuracy of stereo matching algorithms can be evaluated using well-established benchmarks with publicly available test data and reference solutions. As opposed to standard imaging configurations, the quality assessment of data delivered by customized 3D reconstruction systems may require the development of novel or adapted evaluation strategies tailored to the specific set-up. This work is concerned with evaluating the quality and accuracy of 3D models acquired with a 3D reconstruction system consisting of three stereo cameras. To this end, three different evaluation strategies are proposed and implemented. First, the 3D model accuracy is determined by acquiring reconstructions of geometrically simple validation objects (sphere, cuboid) that were specifically created for this purpose. Corresponding ideal 3D objects are fitted into the reconstructed point clouds and are compared to their real measurements. Second, an image-based novel view evaluation determines the accuracy of multiple reconstruction approaches on intermediate point clouds and final 3D mesh models. Finally, a pair comparison-based user study determines the subjective quality of different depth reconstruction approaches on acquired textured 3D mesh models. We demonstrate the three evaluation approaches on a set of self-recorded data. In this context, we also observe that the performance of the examined approaches varies only slightly in the novel view evaluation, while the user study results show clear preferences, which confirms the necessity to combine both quantitative and qualitative evaluation.

Contents

Kurzfassung **ix**

Abstract **xi**

Contents **xiii**

1 Introduction **1**
1.1 Objectives and Contributions 2
1.2 Organisation of this Work 3

2 Fundamentals of 3D Reconstruction **5**
2.1 Transformations in 3D 5
2.2 Camera Models and Calibration 7
2.2.1 Pinhole Camera Model 7
2.2.2 Lens Distortion 9
2.2.3 Stereo Cameras 10
2.2.4 Stereo Camera Calibration 11

3 3D Reconstruction Using Multiple Depth Sensors **15**
3.1 Data Acquisition 15
3.2 Depth Reconstruction 17
3.2.1 Image Similarity 17
3.2.2 Stereo Matching 19
3.2.3 Scene Representations 22
3.3 View Fusion 23
3.4 Summary 24

4 Evaluation Methods **25**
4.1 Overview of Evaluation Methods 25
4.2 Image-based Novel View Evaluation 27
4.2.1 Third-Eye Technique 27
4.2.2 Two-View Evaluation 27
4.3 Subjective Quality Assessment 29
4.3.1 Subjective Assessment of 3D Models 29

4.3.2 Study Design and Environmental Conditions 30
4.3.3 Testing Methodologies 31

5 System and Evaluation Framework **33**
5.1 System Description 33
5.1.1 System Overview 34
5.1.2 Hardware and Data Acquisition 34
5.1.3 Calibration, Registration and Image Rectification 40
5.1.4 Depth Reconstruction and 3D Model Generation 43
5.1.5 Summary 54
5.2 Evaluation Strategies 54
5.2.1 Evaluation on Validation Objects 54
5.2.2 Novel View Evaluation 55
5.2.3 Subjective User Study 56
5.3 Summary 59

6 Evaluation Results **61**
6.1 Data Set and Evaluated Approaches 61
6.1.1 Data Set 61
6.1.2 Evaluated Approaches 62
6.2 Results of Evaluation on Validation Objects 63
6.2.1 Data Set and Validation Objects 64
6.2.2 Results for Spherical Objects 65
6.2.3 Results for Cuboid Objects 66
6.2.4 Discussion 66
6.3 Results of Novel View Evaluation 68
6.3.1 Data Set 69
6.3.2 Results of Depth Sensors 69
6.3.3 Results for Evaluated Approaches 70
6.3.4 Discussion 72
6.4 User Study Results 73
6.4.1 Study Design 74
6.4.2 Compared Approaches 75
6.4.3 Discussion 77

7 Conclusion **79**

List of Figures **81**

List of Tables **83**

Acronyms **85**

Bibliography **87**

Appendix A - System Ground Truth Measurements **97**

Appendix B - User Study **99**
User Instructions 99
User Questionnaire 99
User Screening 99
Detailed User Information 106

CHAPTER 1

Introduction

Reconstruction of 3D object models has been a long tackled problem in computer vision research. Applications requiring such models include quality control of manufactured items [BKH10], cultural heritage preservation [VCB15], and urban reconstruction [KHSM17]. Another area demanding high quality 3D models is media content generation, such as mixing real world objects with synthetic content.

The generation of dynamic 3D model content can be divided into five principal processing steps: acquisition and preprocessing, 3D point generation, meshing and texturing, temporal mesh processing and mesh post-processing (see Figure 1.1). A common method is to capture the scene from multiple view-points with *calibrated and registered stereo cameras*. The resulting image pairs are then used to recover scene depth by means of *stereo matching* (e.g. [SNG+15, WFR+16]). The results of matching process are *disparity maps*, images whose pixel values encode the scene depth in terms of the horizontal displacement between a scene point's location the input image pairs. Using known geometric camera properties, disparity maps can then be turned into *3D point clouds*. Surface reconstruction algorithms (e.g. [DTK+16, GG07]) then transfer point clouds into surface meshes [GG07] or volumetric grids [DTK+16]. In the case of dynamic scenes, the task of temporary tracking merging models is often performed by non-rigid registration (e.g. [DTK+16]).

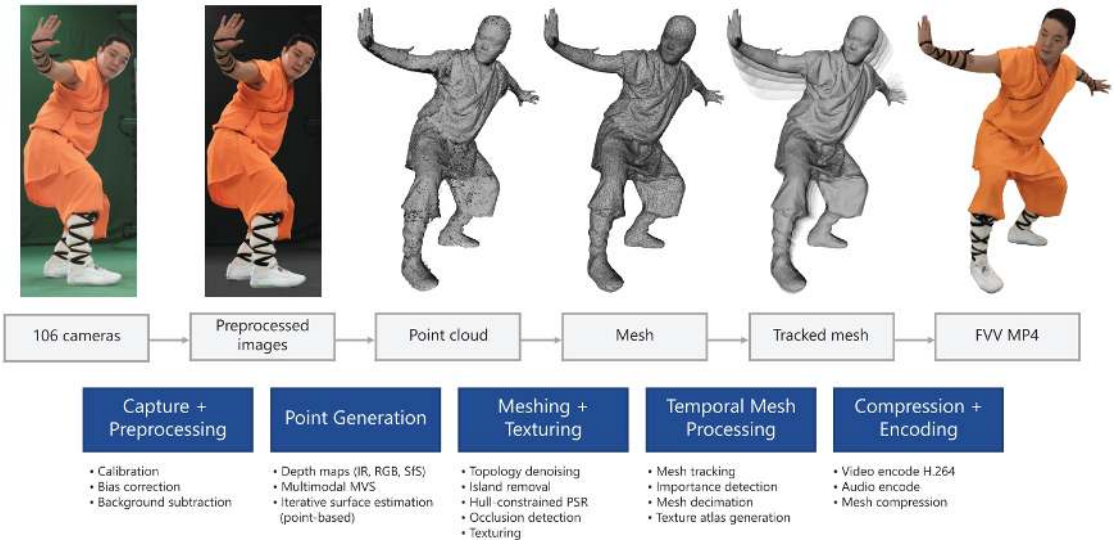


Figure 1.1: 3D model reconstruction processing pipeline. Figure taken from [CCS+15].

3D model reconstruction is a challenging process. Errors diminishing the result's accuracy may be introduced at every stage. Inexact stereo camera rig calibration introduces geometric errors [BR15] affecting shape and reconstructed depth of the computed point clouds. Specific scene properties of captured scene objects such as untextured, reflective or translucent regions, often cause stereo matching algorithms to recover incorrect or invalid depth values. Depth value quantisation in stereo matching can further limit the accuracy of the reconstructed point clouds. Said issues lead to "noisy" point clouds, i.e. they contain erroneous points which pose a problem for subsequent point cloud registration and surface reconstruction algorithms, thus making error detection and removal necessary [WKZ+16] in order to achieve accurate results.

1.1 Objectives and Contributions

This master thesis is concerned with evaluating a multi stereo camera based 3D reconstruction system consisting of three stereo cameras. The goal is to determine the system's accuracy, and model quality. To this end three types of evaluation will be conducted on models acquired for this purpose. First, we seek to determine the accuracy of the geometric reproduction by comparing models of primitive validation objects (sphere, cuboid) against their true known properties by means of shape fitting. Second, the model reconstruction quality of several model generation methods will be analysed using image-based similarity, by comparing novel views of intermediate and final products against the original image input (see Figure 1.2). Third, the subjective model quality will be assessed by conducting a pair-based subjective user study showing coloured mesh-models.

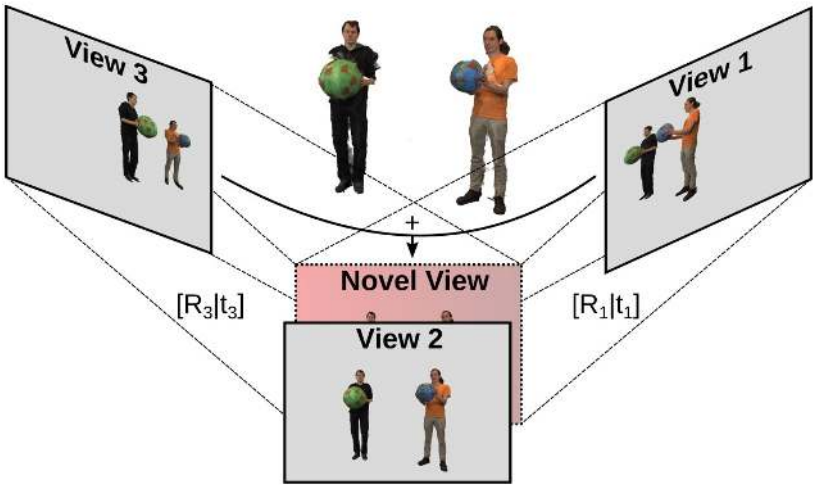


Figure 1.2: Illustration of the employed novel view evaluation method. Point clouds and coloured meshes fused from views 1 and 3 are transformed into the view point of view 2 by application of rigid-body transformations $[R_1|t_1]$ and $[R_3|t_3]$. View 2 serves as independent source of validation.

Several depth-image based fusion 3D reconstruction systems employ real-time capable stereo matching algorithms, e.g. [WFR+16, EFR+17, OEDT+16], for acquiring 3D points. A thorough comparison of the impact of a particular chosen algorithm on model quality in the mentioned systems, however, has not been undertaken to the author's knowledge. A primary research question is to ask for the influence of different stereo matching algorithms on the model quality. Three methods that deliver disparity maps of substantially different characteristics will be examined. In particular, (1) integer valued disparity maps computed with a cost volume filtering algorithm [SNG+15] will be used as a base line algorithm. Point clouds computed from such maps exhibit low depth resolution, their points are positioned along discrete planes in space. (2) Floating point valued disparity maps computed with the same cost volume filtering algorithm [SNG+15] using different parameter settings result in point clouds of high depth resolution. Points are located near object surfaces, however they are locally noisy. (3) Disparity maps computed with a PatchMatch based algorithm [LZYZ18] result in point clouds exhibiting very smooth object surfaces.

A second question we address is how fusion of individual view point clouds at different processing steps alters the model quality. Specifically, view fusion of point clouds before and after model generation will be compared.

1.2 Organisation of this Work

The chapters of this work are structured in the following way:

Chapter 2 introduces selected fundamental concepts of 3D vision used throughout this work. First, basic concepts of 3D projective space and 3D transformations will be described. Second, the pinhole camera model will be introduced and extended to the stereo camera case. Third, the camera calibration procedure is shortly outlined. The chapter focuses on the most important concepts that are used throughout this work.

Chapter 3 presents the state-of-the-art in 3D reconstruction of dynamic scenes. First, different methods of depth acquisition hardware and their characteristics summarised. Then, fundamental concepts of the employed stereo matching algorithms will be presented. This includes a summary of popular image similarity measures, stereo matching algorithms, as well different kinds of scene representation. Lastly, model generation algorithms commonly used for reconstruction will be shown.

Chapter 4 presents the state-of-the-art of evaluation methods relevant in this work. It starts with an overview of applicable quantitative and qualitative methods. Next, a summary of image-based novel view evaluation follows. Lastly, subjective quality assessment is explained.

Chapter 5 describes the system under examination and the methods applied for its evaluation. This chapter comprises two parts. In the first part, the examined system and its processing pipeline are described in detail. Specifically, hardware and data acquisition, camera calibration and registration, and depth reconstruction will be discussed. The chapter also contains discussions of failure cases that can arise for each of the described pipeline stages. The chapter's second part is outlines the concrete application of evaluation methods laid out in Chapter 4.

Chapter 6 presents the results of our evaluation. First, the used data set will be presented, and the approaches we compare will be explained in detail. Second, the results of the evaluation on validation objects will be shown. Third, the results of the novel view evaluation will be presented. Fourth, the results of the user study carried out within this work will be discussed.

Chapter 7 summarises the covered topics and discusses conclusions that can be drawn. Further, it shows possible future work.

CHAPTER 2

Fundamentals of 3D Reconstruction

This chapter recapitulates selected fundamental topics of 3D vision that are employed within this work. First, transformations in three-dimensional space are presented. Next, we introduce the pinhole camera, and the stereo camera model, and show how points in 3D space are projected onto 2D coordinates of a camera image.

2.1 Transformations in 3D

Intuitively, we can imagine that each object in 3D space has its own coordinate system attached to it. Transformations formalise the notion of how we can arrive from the *coordinate frame* of one object to that of another. The content in this section summarises selected topics presented in Hartley and Zisserman's book Multiple View Geometry [HZ04].

Translation. Translation of an object is presented by shifting it from its attached coordinate frame into another coordinate frame that is displaced by the translation vector. Translation can be seen as shifting the origin of object coordinate frame into another one given by $T = C_o - C_c$, where T is a three-dimensional vector, C_o is the origin of the object's original coordinate frame, and C_c is the origin of the object's new coordinate frame after translation.

A point $P = (X, Y, Z)$ is translated by a vector $t = (t_x, t_y, t_z)$ in homogeneous coordinates

with a *translation matrix* T_t given as

$$T = \begin{pmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} = \begin{bmatrix} X + t_x \\ Y + t_y \\ Z + t_z \\ 1 \end{bmatrix} = P + t \tag{2.1}$$

The direction of the translation can be reversed, $T_t^{-1} = T_{-t}$. The product of two translation matrices is given by addition of the translation vectors: $T_r T_s = T_{r+s}$, where T_r and T_s are translation matrices of the 3D vectors r and s.

Rotation. Any rotation can be expressed as a sequence of rotations around different coordinate axes, as shown in the Euler theorem. In the case of three dimensions it can be expressed as a sequence of 2D rotations around each of the three coordinate axes where the pivot axis remains constant. Rotating counter-clockwise around the coordinates z, y and x by angles α, β and γ, respectively, results in a combined rotation matrix R that is the product of the three single axis rotations $R_x(\gamma)$, $R_y(\beta)$ and $R_z(\alpha)$:

$$R_z = \begin{pmatrix} cos\gamma & -sin\gamma & 0 \\ sin\gamma & cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix} R_y = \begin{pmatrix} cos\beta & 0 & sin\beta \\ 0 & 1 & 0 \\ -sin\beta & 0 & cos\beta \end{pmatrix} R_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & cos\alpha & -sin\alpha \\ 0 & sin\alpha & cos\alpha \end{pmatrix}$$

The combined matrix is then $R = R_z(\alpha)R_y(\beta)R_x(\gamma)$. Since matrix multiplication is not commutative, the order of rotation is important. Rotation is first performed around the z-axis, then around the new position of the y-axis and finally around the new position of the x-axis. The rotation matrix R has the property that its inverse is its transposition, that is $R^\top R = RR^\top = I$, where I is the identity matrix.

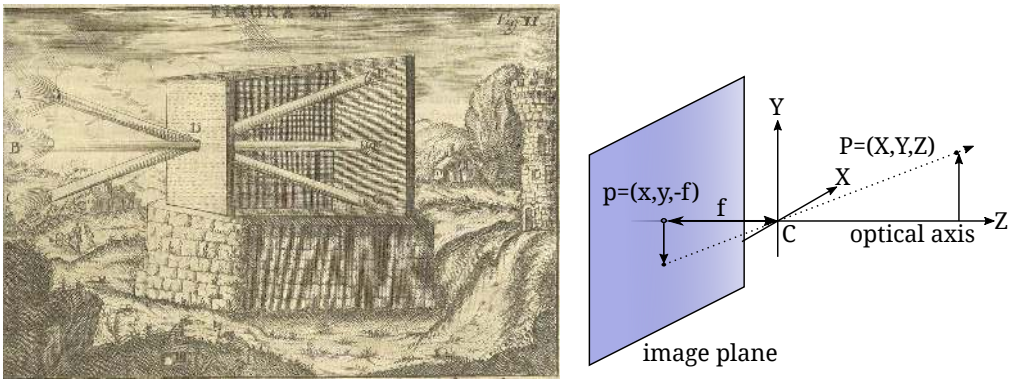
Euclidean Transformation. Euclidean transformations, also known as *isometries*, are the composition of a translation and rotation. It models the motion of a rigid object and is thus often referred to as *rigid body* transformation. It is given by

$$X' = \begin{pmatrix} R & t \\ 0^T & 1 \end{pmatrix} \tag{2.2}$$

where R is a 3×3 rotation matrix and t is a 3D translation vector and 0 is a three-dimensional null vector. The inverse of a Euclidean transformation is given by

$$T^{-1} = \begin{pmatrix} R^T & -R^T t \\ 0^T & 1 \end{pmatrix} \tag{2.3}$$

The Euclidean transformation preserves the geometric properties of length, angles and area.



(a) Camera obscura [Zah85] (b) Pinhole camera model

Figure 2.1: The pinhole camera model.

2.2 Camera Models and Calibration

2.2.1 Pinhole Camera Model

A common way to represent a camera is the *pinhole camera model*. Its operating principle is that of the *camera obscura* (see Figure 2.1a). It consists of a light enclosed compartment with a small hole, the "pinhole", or aperture, in its front. Rays of light may only enter the camera by passing the hole. The backside of the compartment, the *image plane*, contains a photosensitive surface, which detects incoming light. Analogue cameras use a sheet of photosensitive paper for this purpose, while with digital cameras, an electronic chip is used.

Figure 2.1b shows a ray originating from the world point P located at the coordinates (X, Y, Z) entering the camera at the point C and hitting the image plane at coordinates $(x, y, -f)$. The point C is the optical centre. Any ray has to pass it. The distance between aperture and image plane is called the *focal length*. The imaged point p lies on the opposite side of the optical axis as the originating point P, which makes images appear upside down.

To avoid the flipped images, the equivalent *central projection model* can be used [Bra00]. Here, the image plane is located in front of the optical center (see Figure 2.2). A 3D world point $P = (X, Y, Z)$ is mapped in the central projection model onto a pixel coordinate $p = (x_u, y_u)$ in the image with the equation

$$(x_u, y_u) = \left(\frac{fX}{Z}, \frac{fY}{Z}\right) \tag{2.4}$$

The camera's *principal point* (c_x, c_y) lies at the intersection between optical axis and image plane. It denotes the origin of the camera's coordinate system. In any practical camera, however, the principal point may not lie exactly at this position and needs to be determined by camera calibration (see Section 2.2.4). Deviations of the principal point

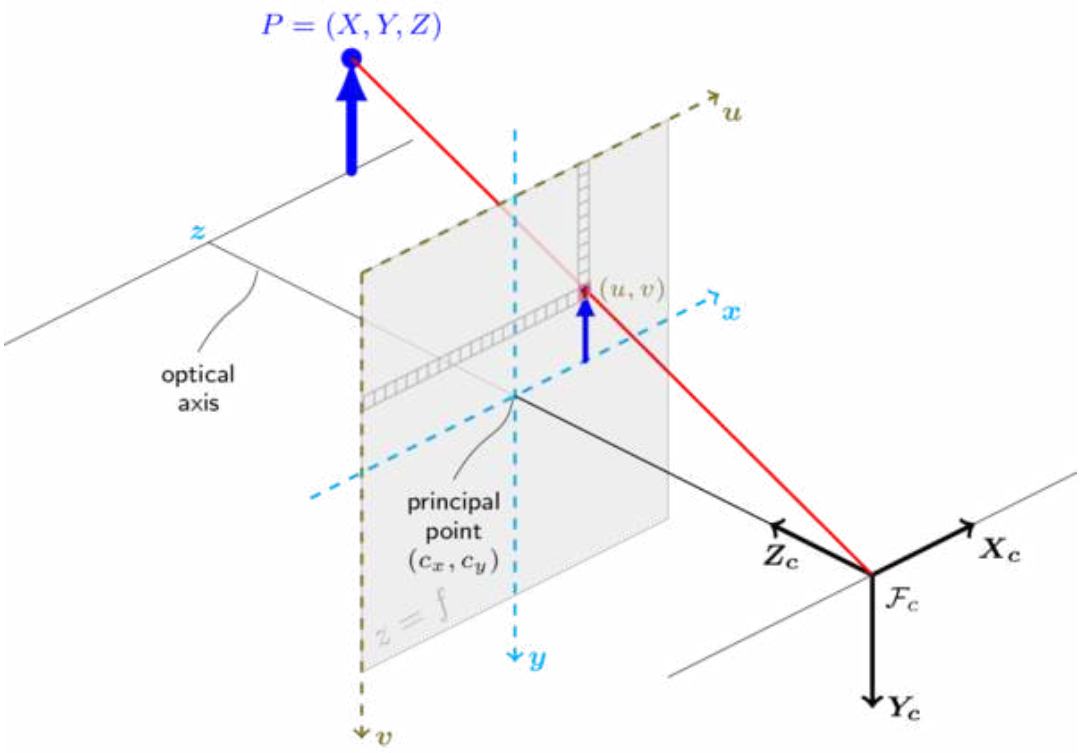


Figure 2.2: Frontal projection model. Taken from [Bra00].

from the camera center change the location of mapped image points and needs to be compensated:

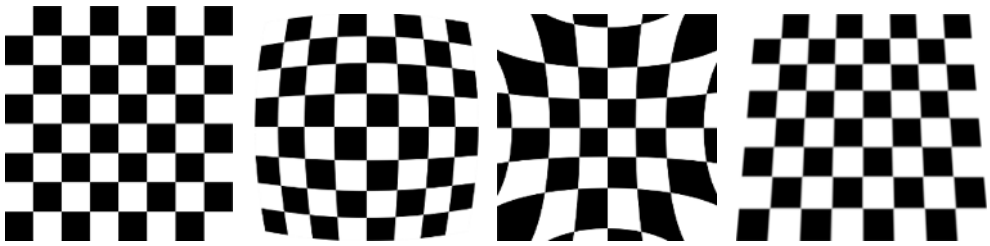
$$(x, y) = (x + c_x, y + c_y) = (\frac{fX}{Z} + c_x, \frac{fY}{Z} + cy) \tag{2.5}$$

The camera is positioned within the 3D space. To relate the camera's coordinate system with the world coordinate system the following perspective projection is used.

$$qm^\top = K[R|t]M^\top \tag{2.6}$$

$$q\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2.7}$$

where $m^\top$ is an image point and $M^\top$ is a world point. q denotes a scaling factor, f_x, f_y are vertical and horizontal focal lengths, c_x, c_y are the image coordinates of the principal



(a) No Distortion (b) Barrel (c) Pincushion (d) Tangential

Figure 2.3: Typical types of lens distortion. A checkerboard pattern (a) when imaged by a camera, exhibits different types of distortion. Barrel distortion (b) and pincushion distortion (c) are caused by lens manufacturing inaccuracies, whereas tangential distortion (d) is caused by misalignment of the image sensor relative to the image plane. Source: [Bra00]

point and s is the camera's skew. r_{ij} for $i, j \in \{1, 2, 3\}$ is a rotation matrix and t_1, t_2, t_3 is a translation vector.

The matrix K denotes the *intrinsic parameters* of the camera. It determines the camera's internal projection completely and is independent of its position in the world coordinate system. The focal length is given both horizontally and vertically and are often assumed to be the same for both directions, $f_x = f_y$. This amounts to stating that pixels are squares. Another common restriction is that the camera does not exhibit skew, that is $s = 0$.

The joint rotation-translation matrix $[R|t]$ are called *extrinsic parameters* or its *pose*. It relates the world coordinate system to the camera coordinate system.

2.2.2 Lens Distortion

A physical pinhole camera has significant drawbacks. The small aperture of the pinhole limits the amount of light that can pass onto the image plane in a given time. Further, the focal length is determined by the camera's physical size. For this reason, practical cameras use lenses. They allow the adaptation of focal length and correspondingly the field of view. Moreover, they feature apertures of varying size, which allows controlling the amount of light that can reach the image plane.

A drawback of using lenses is that they introduce distortion to the images by changing pixel locations of imaged world points. A common way for modelling lens distortion is the *Brown-Conrady* [Bro66] model. It allows compensating for radial and tangential distortion.

Radial Distortion Due to manufacturing inaccuracies, lenses distort the location of imaged pixels of light rays entering near the lens' outer rim. Straight lines, for example those of a rectangular pattern imaged facing parallel to the image plane (see Figure 2.3a),

become increasingly curved near the image border. This effect is often called *barrel* or *fish-eye* distortion (see Figure 2.3b). It is especially noticeable in wide-angle lenses, more so in those of low quality. The opposite effect, *pincushion* distortion, bends straight lines into inwards direction (Figure 2.3c).

The amount of radial distortion is small near the image centre and increases with the distance from it. It can be modelled in terms of a Taylor series expansion around the camera's principle point. Let r be the radius of a circle with the center in the principle point, then

$$x_{rad} = x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6)$$
$$y_{rad} = y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6)$$

where (x_{rad}, y_{rad}) is the image coordinate of the radially distorted pixel and (x, y) is the location of the pixel corrected for radial distortion. k_1, k_2, k_3 are the *radial distortion coefficients*.

Tangential Distortion *Tangential distortion* occurs due to the camera's sensor not being aligned parallel with the image plane (Figure 2.3d). It can be characterised as follows:

$$x_{tang} = x + [2p_1 y + p_2(r^2 + 2x^2)]$$
$$y_{tang} = y + [p_1(r^2 + 2y^2) + 2p_2 x]$$

where (x_{tang}, y_{tang}) is the image coordinate of the radially distorted pixel and (x, y) is the location of the pixel corrected for tangential distortion. p_1, p_2 are the *tangential distortion coefficients*.

The tuple of coefficients $dist = (k_1, k_2, p_1, p_2, k_3)$ determines the combined radial and tangential distortion an imaged pixel. Distortion is independent from image resolution and only depends on the distance of a pixel from the distortion center, that is assumed to be equal to the camera's principal point.

2.2.3 Stereo Cameras

Two pinhole cameras placed closely next to each other pointing in the same direction constitute a *stereo camera*. The coordinates of a 3D point P seen by both cameras can be reconstructed, when it is observed by the two cameras if intrinsic and extrinsic parameters are known. As shown in Figure 2.4a, principal points of both cameras, C_l and C_r are connected with a line, called the stereo baseline B. A 3D point P together with C_l and C_r form the epipolar plane, which intersects the cameras' image planes at the epipolar lines e_l and e_r. When P is known to be projected to the image coordinate x_l on the left camera, the corresponding image coordinate x_r in the right camera has to

lie on the epipolar line e_r. If x_r it unknown, it is sufficient to restrict the search for it to image coordinates of e_r.

The search for an unknown image point x_r can be further simplified by virtually moving the image planes of both cameras in a way that aligns the epipolar lines e_l and e_r with horizontal coordinates of the cameras' image planes, as shown in Figure 2.4b. The epipolar lines become parallel to the baseline B. This configuration is said to meet the *epipolar constraint*. Rectification of a stereo camera involves modifying both cameras' relative pose (R, t) so that the epipolar constraint is met. Determining suitable relative poses of both views of a stereo camera is a task of stereo camera calibration. If the epipolar constraint is met, the search for an unknown corresponding image point x_r for a known image point x_l reduces to a one dimensional scan for x coordinate at the y coordinate of x_l in the right camera's image plane. The difference between x_l and x_r is called the disparity d of point P. With known point correspondences x_l and x_r, the distance of P to the stereo camera can be determined by the method of similar triangles (see Figure 2.4c), resulting in the following formula for the distance Z of point P

$$Z = \frac{f * B}{x_l - x_r} = \frac{f * B}{d} \tag{2.8}$$

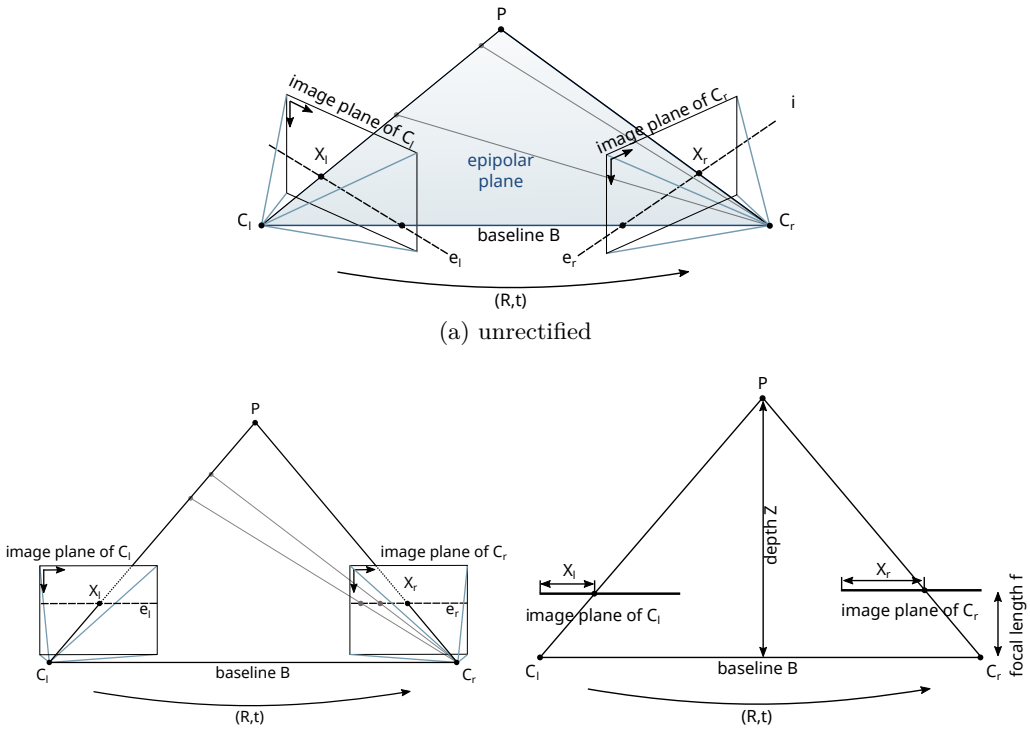
where f is the camera's focal length in pixels. B is the base line in meters. x_l and x_r denote horizontal pixel coordinates in the left and right camera image, respectively. Their difference is the disparity denoted by d.

2.2.4 Stereo Camera Calibration

Camera calibration refers to determining the intrinsic and extrinsic camera parameters. When these are known, 3D world points can be projected to 2D image coordinates and vice versa. Calibration involves acquiring images of objects with well-known dimensions. Camera parameters are then determined by relating object properties with their imaged counterparts.

Several calibration methods have been proposed. Objects used for calibration include spheres [SBMM15], wands with multiple collinear points [Zha04] and point-like [SMP05] objects.

A particularly popular approach is that of Zhang [Zha00], which uses planar checkerboard patterns. The first step is to acquire images of a checkerboard pattern with the stereo camera from several poses (see Figure 2.6). Next, the interior checkerboard corner coordinates are extracted from the captured images. They are then used to estimate the four cameras' intrinsic parameters (f_x, f_y, c_x, c_y) and the orientation of each checkerboard view. Next, the five lens distortion parameters, three radial parameters k_1, k_2, k_3 and two tangential parameters p_1, p_2, are estimated by minimizing the reprojection error between the estimated parameters and the observed checkerboard corners. Once the intrinsic parameters of both cameras are known, their rectifying transformation, which makes the epipolar lines of both cameras parallel, can be determined with the method of Hartley [HZ04].

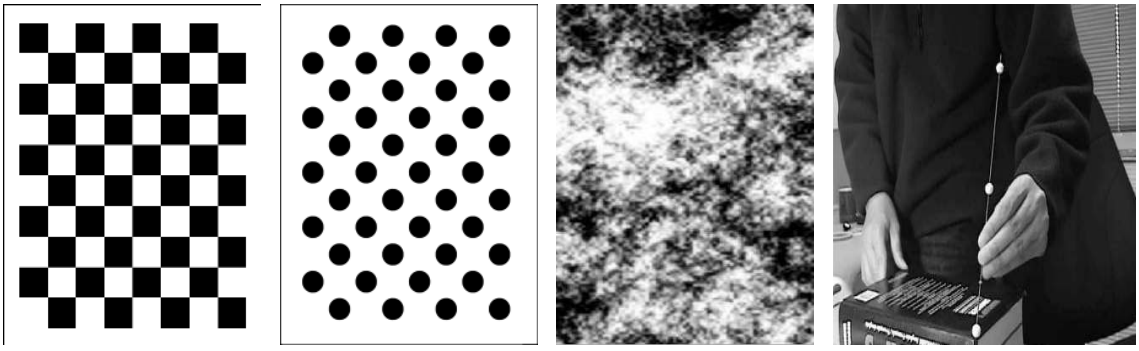


(a) unrectified

(b) rectified, frontal view

(c) rectified, top view

Figure 2.4: Epipolar geometry of a stereo camera. (a) Unrectified stereo camera. Image planes of both cameras do not lie on a common plane. (b) Rectified stereo camera seen from the backside. Epipolar lines are parallel and lie on the same image y coordinate (c) Rectified stereo camera shown from above. Depth of P can be triangulated. Figure adapted from [NBG13].



(a) [Zha00] (b) [DKK09] (c) [LHKP13] (d) [Zha04]

Figure 2.5: Illustration of various calibration objects. (a)-(c) Examples of planar patterns; (d) Example of a one-dimensional calibration object.

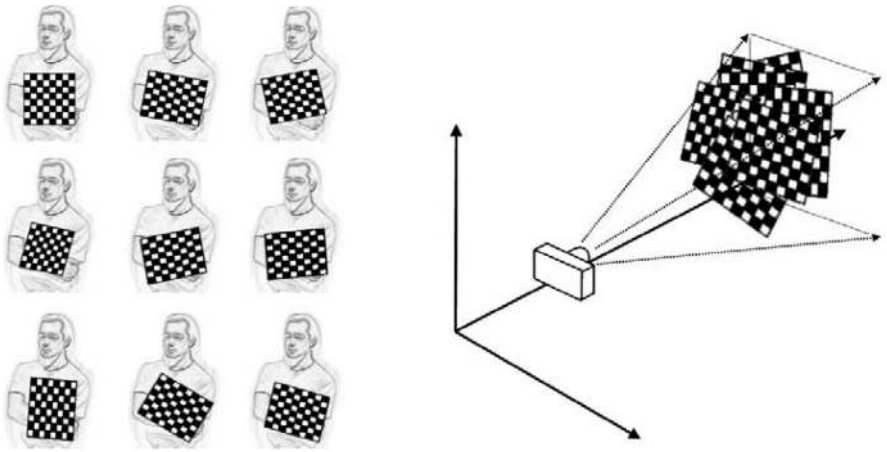


Figure 2.6: Camera calibration with planar patterns. Figure taken from [BKB08].

CHAPTER 3

3D Reconstruction Using Multiple Depth Sensors

In this chapter, we present the state-of-the-art related to our approach to 3D reconstruction that is suited for acquiring dynamic scenes in a controlled environment. Here, depth sensors are positioned statically around an acquisition area. Each sensor acquires 3D information from its own viewpoint. Individual views of the scene are then fused to combined 3D models.

This chapter is structured as follows. In Section 3.1, we give an overview of employed data acquisition techniques. Next, we discuss how depth information can be reconstructed from image pairs in Section 3.2. Lastly, the fusion of views into combined models are explained in Section 3.3.

3.1 Data Acquisition

3D reconstruction data is commonly captured by photogrammetric-, laser-scan- or range-image-based acquisition devices. Each method has its own characteristic advantages and drawbacks determining their suitability for specific reconstruction applications. Figure 3.1 gives an overview of the mentioned methods and their resulting output data.

Photogrammetry Photogrammetry refers to the reconstruction of three-dimensional information from images (see Figure 3.1a). Ubiquitous and cheap availability of digital cameras has led to photometry being a highly popular approach to depth acquisition. Images do not convey three-dimensional information directly. The 3D information needs to be inferred from image data by subsequent processing steps. Usually multiple images of a scene are used to perform this task. They are acquired by either a moving single camera and capturing multiple images subsequently, or by simultaneous acquisition

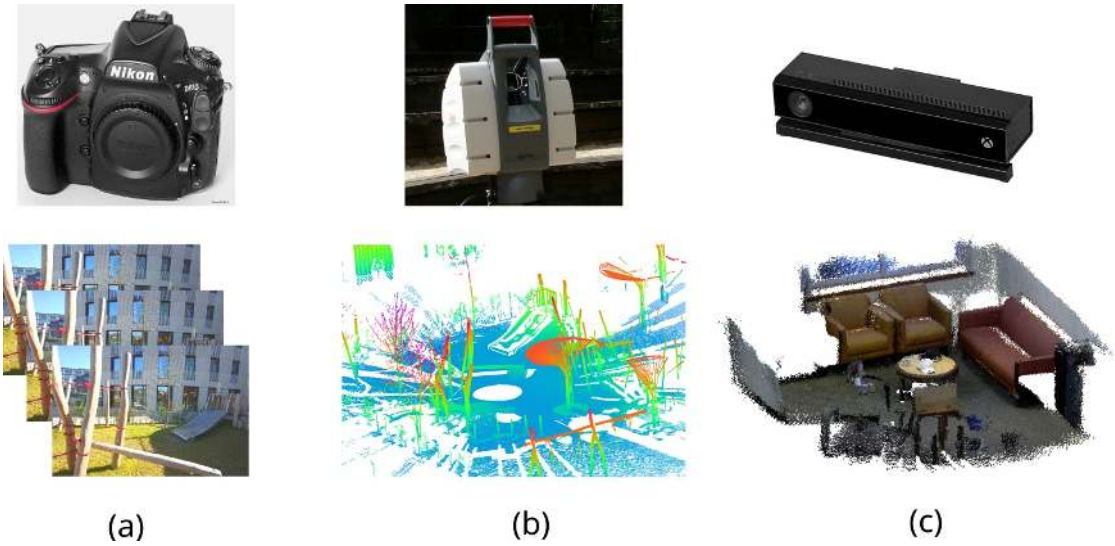


Figure 3.1: Data acquisition methods and their results. (a) Photogrammetric acquisition yields RGB images. Source: [Nym17]; (b) Laser scanning devices produce uncoloured point clouds, illustrated as colour-mapped. Source: [Mon07]; (c) Time-of-flight cameras may produce RGB-D images. Source: [EA14].

with multiple cameras at different positions. Depth information can then be recovered by identifying corresponding image points and triangulating them between multiple images (see Section 2.2.3). Although this way of depth acquisition is prone to errors and requires a significant amount of computation, it is also the most flexible depth reconstruction method. Applications of photogrammetry range from highly precise depth measurements of small objects [BKH10, SCD+06], over large-scale reconstruction of terrain [ZTDVAL14] and urban areas [LNSW16] to the acquisition of highly dynamic scenes [DTK+16, EFR+17, CCS+15].

Range Images Range imaging is an active method of depth acquisition. A signal created by the depth sensor interacts with the scene and is then measured by the sensor. Two predominant technologies delivering range images are structured-light [HLCH12, SLK15] and time-of-flight (TOF) [HLCH12, SAB+07]. Figure 3.1c illustrates the popular Kinect One sensor and an acquired point cloud that exhibits depth as well as RGB colour information (i.e. RGB-D). In structured light techniques, a pattern is projected onto a scene, usually in the near infra-red spectrum invisible to the human eye. The pattern is distorted by the scene and is again captured by a monochrome charge-coupled-device (CCD) image sensor. Examples of recent high performance real-time 3D acquisition systems employing structured light are e.g. [DTK+16, OEDT+16]. There, additional RGB cameras are used to acquire colour information for model texturing. TOF cameras, on the other hand, achieve similar results by measuring runtime differences of light sent by the sensor and reflected by scene objects. Range image cameras are able to provide

depth maps and colour images simultaneously. The spatial resolution of depth maps is usually limited, as is the maximal achievable measurement distance. The popular Kinect One sensor has a depth image resolution of 512×424 pixels and can measure distances of up to 4.5 meters with a frame rate of up to 30 Hz [SLK15].

3.2 Depth Reconstruction

We will now discuss how 3D models can be recovered from images of a scene. We start by presenting image-based measures that can be used to find corresponding regions among multiple images. Next, depth reconstruction from stereo image pairs is elaborated. Lastly, common ways of scene representation are explained.

3.2.1 Image Similarity

A key task for any reconstruction algorithm is to identify corresponding points or features within two or more images. *Image similarity*, also called *photo-consistency*, captures this concept. An object's illumination and colour can change significantly when viewed from different positions due to directional light sources, or object material. It is desirable for similarity measures to be invariant to such changes.

We can distinguish sparse and dense approaches. In the first category, we have feature descriptors (e.g. SIFT [Low04] or SURF [BTV06]) that identify prominent image regions. They are considered *sparse*, as they only track prominent image regions, such as edges or contours. The second are *dense* similarity functions that assign a numerical value to every pixel of an image. Given a set of N images and a 3D point p seen in every , we can define photo-consistency [FH15] between pairs of images I_i and I_j, $i, j \in (1, \ldots, N)$ as

$$C_{ij}(p) = \rho(I_i(\Omega(\pi_i(p))), I_j(\Omega(\pi_j(p)))), \tag{3.1}$$

where ρ is a similarity function, $\pi_i(p)$ is the projection that maps the 3D point p into image i, $\Omega(x)$ defines a support region, also called domain, around point p and $I_i(x)$ denotes the intensity or colour values of pixels within the domain. The choice of ρ and Ω describes a particular similarity measure. Note that image coordinates are integral, whereas $\pi_i(p)$ is real valued. To accommodate 3D points that are projected onto real-valued coordinates some interpolation scheme is needed. Further, the described similarity measures operate on single channel images only. RGB images require preprocessing in order to determine similarity. One way is to perform computation on grey-scale versions. Another way is to compute similarity on each of the three colour channels separately, and then combine the results by pixel-wise averaging [FH15].

We give details on three commonly used similarity measures.

- **Sum of Absolute Differences (SAD)** is defined as the L^1 norm between two vectors of pixel intensity values in support regions f and g around the image

coordinates to which a 3D point p is projected. More formally,

$$\rho_{SAD}(f,g) = ||f-g||_1 \tag{3.2}$$

SAD is sensitive to brightness and contrast changes. It is useful mainly for images of similar illumination characteristics. On the other hand, SAD is computationally cheap, which makes it useful for real-time applications that can guarantee similar image illumination.

- **Normalised Cross Correlation (NCC).** The normalised cross coefficient is an established tool for determining image similarity in presence of illumination and exposure changes. It is a statistical measure defined as

$$\rho_{NCC}(f,g) = \frac{(f-\bar{f})\cdot(g-\bar{g})}{\sigma_f \sigma_g} \in [-1,1] \tag{3.3}$$

where $\bar{f}$, $\bar{g}$ denote the mean values and σ_f and σ_g the standard deviations of pixel intensity values within the domains around the pixels projected into I_i and I_j, respectively.

Its invariance to illumination changes makes NCC one of the most commonly used similarity measures in two-view and multi-view stereo. NCC, however, fails to detect pixels in untextured regions.

- **Census** [ZW94] Census is one of the best performing similarity measures for stereo correspondences [HS07]. In contrast to other presented measures, it does not use intensity values themselves, but first computes a bit string describing whether pixels within the support domain of a pixel p are lighter or darker than p and then computes the Hamming distance between the two resulting bit strings.

Formally, a comparison operator is defined that determines whether a pixel a is brighter than a pixel b

$$\xi(a,b) = 1 \; if \; a < b, 0 \; otherwise, \tag{3.4}$$

A bit string describing brighter and darker pixels in Ω is computed as

$$census(f) = \oplus_{q\in\Omega}\xi(f(p), f(q)), \tag{3.5}$$

where $\oplus$ is the concatenation operator and p and q are image pixels. The Census score is then the Hamming distance of the two bit strings.

$$\rho_{Census}(f,g) = |census(f) - census(g)|_1 \in [0,N] \tag{3.6}$$

where N is the size of the support region Ω. Census is especially robust against image brightness and contrast changes, as well as around depth boundaries.

3.2.2 Stereo Matching

Stereo matching is the problem of recovering depth from image pairs that are slightly displaced akin to human eyes (see Figure 3.2a). After appropriate rectification, the imaged objects exhibit a horizontal displacement, called *disparity*, depending on their distance to the camera. It is measured in terms of the number of displaced pixels. Near objects have high disparity, whereas objects that are more distant have low disparity. For example, the chimney edge shown in Figure 3.2 appears in the left view at point P_l and in the right view as P_r. The disparity between these points is denoted as d_P. The Teddy's ear, on the other hand, is marked as Q_l and Q_r respectively, and is located further in the backside of the scene. The chimney is nearer than the ear, and we have $d_P > d_Q$. A stereo matching algorithm computes depth in form of a *disparity map*, which is a single channel image whose pixel intensity values correspond to the scene disparity at pixels in the corresponding (left) input image.

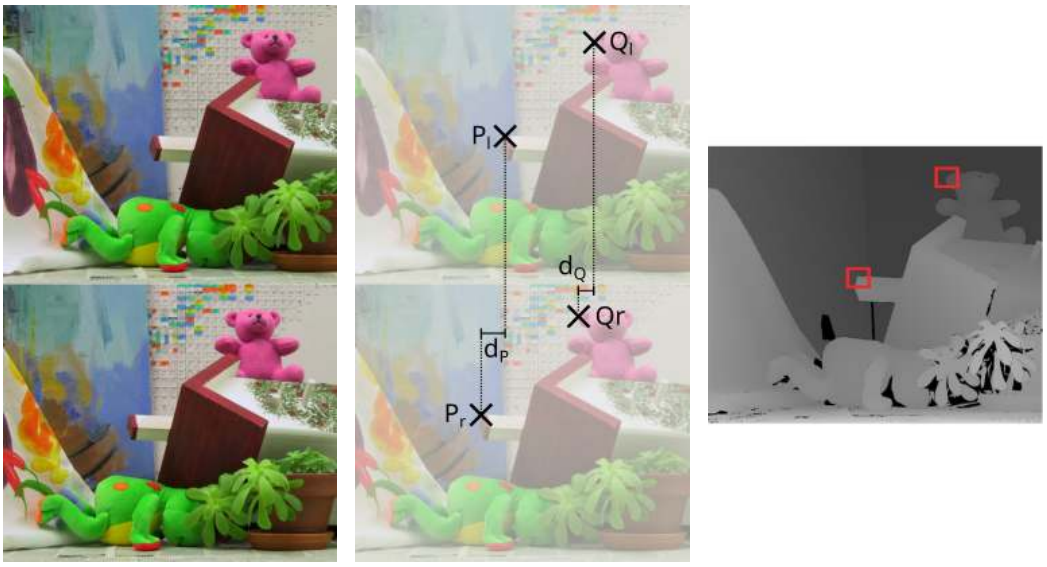
Assumptions. Stereo matching algorithms commonly make some assumptions to compute disparity maps. The first one is the *photo consistency assumption*. It demands that corresponding pixels in the left and right view have the same colour values. Next, there is the *epipolar assumption*, requiring that corresponding pixels in left and right view always appear on a horizontal line. This is ensured by image rectification explained in Section 2.2.3. The *smoothness assumption* states that spatially close pixels have similar disparity values. The smoothness assumption holds in most image regions, except for object borders.

Types of Stereo Matching Algorithms. Stereo Matching algorithms can be broadly categorised by the way they determine point correspondences. Local methods find disparity values by searching for each pixel in the left view a corresponding pixel in the right view by sliding local windows along horizontal image lines. A typical example for local stereo matching is the cost volume filtering technique [HRB+13, SNG+15]. Global methods minimise an explicit energy function over all image pixels. A typical energy function [BB13] has the form

$$E(p) = E_{data}(p) + \lambda E_{smooth}(p)$$

where $E(p)$ is the total energy value of an image pixel p. The data term E_{data} accounts for colour similarity. E_{smooth}, the smoothness term, captures local object smoothness. The parameter $\lambda \in \mathbb{R}$ balances is the relative influence of the terms E_{data} and E_{smooth}. Common examples for global stereo matching algorithms are dynamic programming [BT99] and belief propagation [FH06]. An elaborate discussion on global stereo methods can be found in [BB13].

Learning based methods rely on machine learning methods for depth estimation. There, first a model is trained with ground truth data, and then the trained model can be used to estimate depth. An example for a learning based algorithm is the Global Patch Collider [WFR+16]. It uses random forests to represent the model.



(a) Left and Right View (b) Disparities (c) Disparity Map

Figure 3.2: Stereo matching. (a) Left input view (top), right input view (bottom). (b) Disparity d_P of a nearer object, marked as P_l and P_r in left and right input is large, while objects father away, e.g. Q_l, and Q_r, exhibit a lower displacement d_Q. (c) Disparity map: intensities encode the disparity value.

Stereo Matching Pipeline. Stereo matching algorithms generally have to perform the same processing steps [SS02], regardless whether they are local or global algorithms. In the following, the conceptual processing steps involved are summarised. Specific methods may skip or aggregate some steps. As the 3D reconstruction system addressed in this work employs primarily a local stereo matching algorithm, the focus lies on local methods. Local algorithms compute disparity by searching for each pixel potentially matching candidates in the other image in a pre-defined search range (Figure 3.3b). The complete search space is often called Disparity Space Image [SS02], or cost volume [BRR11]. Candidate search can be limited to pixels on the same horizontal line when relying on rectified input images (Figure 3.3a).

Cost Computation (CC). Pixels are compared using some dissimilarity- or cost-measure that takes into account a small support region around pixels. Measures that are typically employed include SAD, NCC or Census. They have been discussed in Section 3.2.1.

Cost Aggregation (CA). Local methods usually enforce spatial consistency implicitly in the Cost Aggregation stage (Figure 3.3c). This process can be viewed as filtering of the cost volume [HRB+13]. Cost aggregation has a significant impact

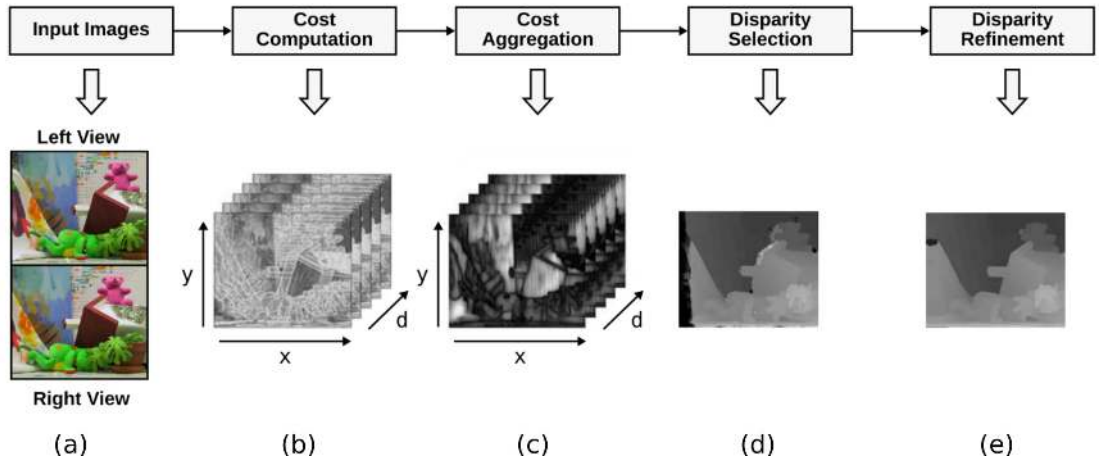


Figure 3.3: Outline of the basic steps of a typical local stereo matching processing pipeline. (a) Rectified input images. (b) Cost Computation (CC) determines the dissimilarity between two pixels for specific disparity levels, and gives rise to a cost volume. (c) In the Cost Aggregation (CA) stage, matching costs are filtered, enforcing a local consistency assumption. (d) In the Disparity Selection (DS) step, disparities are selected from the cost volume. (e) In the depth refinement (DR) stage, disparity inconsistencies are detected, and resulting holes are filled. Figure inspired by [BRR11].

on the quality of local stereo matching algorithms as noted by several authors (e.g. [BRR11, HRB+13, HBG13, LZYZ18, ZFM+17]).

Disparity Selection (DS). Next, disparity values are determined as illustrated in Figure 3.3d. Local stereo methods usually achieve this by selecting the disparity value with minimum costs within the cost volume for each pixel. This procedure is often called the winner-takes-it-all (WTA) strategy.

Disparity Refinement (DR). Disparity selection already yields a rough disparity map that will usually contain wrongly computed disparity values (e.g. Figure 3.3d). The goal of disparity refinement is twofold. The first is to eliminate pixels with wrong disparity values by means of a consistency metric. Typically, two disparity maps are computed, one from left to right, and one from right to left. Subsequently, pixel disparity values that do not agree up to a threshold t in both maps are rejected. Common values of t are between 0.5 and 4 pixels [Midb]. Note that consistent disparity values do not necessarily imply correctly estimated depth values. The second task of disparity refinement is to fill in holes caused by occlusions and eliminated disparity map inconsistencies. Smaller holes are often closed by filtering techniques, such as median filtering [SS03].

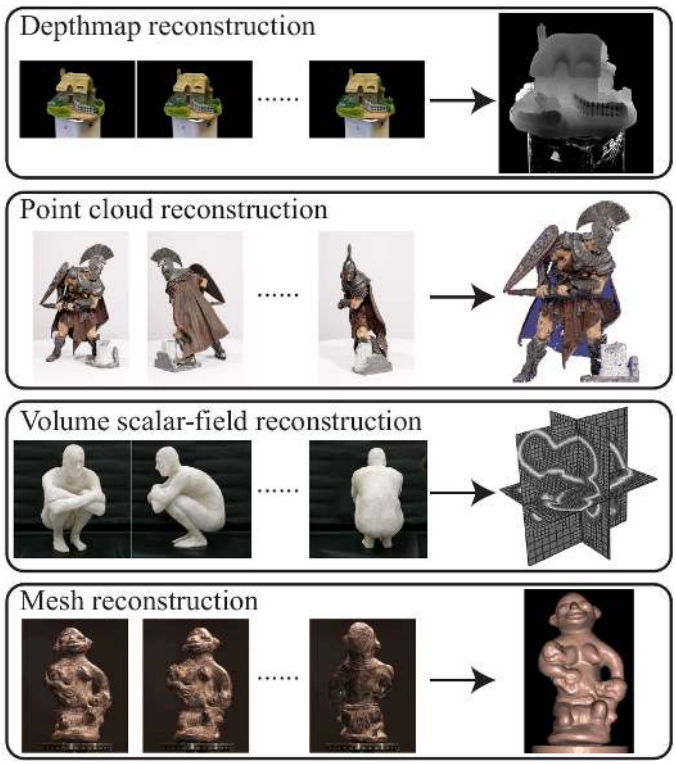


Figure 3.4: Common scene representations in 3D reconstruction. Figure taken from [FH15].

3.2.3 Scene Representations

3D objects can be represented in several ways, each of which has its own advantages and drawbacks. A specific 3D reconstruction system may use multiple representations at different processing stages. Figure 3.4 shows four popular representations, namely depth maps, point clouds, meshes, and volumetric representations.

Depth Map Depth maps are single-channel images whose intensity values correspond to scene depth in terms of distance to the camera. They are closely related to disparity maps, which have been discussed in Section 3.2.2. Depth maps can be obtained from disparity maps, when the camera focal length and stereo baseline are known, by employing equation 2.8. Depth maps are a compact way for representing 3D scenes. As they are essentially images, they are well suited for employing 2D filtering techniques.

Point Cloud A point cloud is a collection of points in space. Points are represented as 3D coordinates and may have additional attributes attached. Point colour or surface normal vectors are common attributes present in point clouds. Point clouds can be obtained from disparity maps or depth maps when the camera focal length,

principal point and stereo baseline are known, as shown in Section 2.2.1. Depth maps, in result, are often treated as geometric proxy for point clouds.

Polygon Mesh Meshes, specifically polygonal, 3D meshes, are collections of vertices, edges and faces. Vertices correspond to points of a 3D point cloud. Edges define (typically triangular) faces attached to mesh vertices, and approximate an object's surface. Algorithms commonly used to reconstruct meshes from point clouds are the Screened Poisson surface reconstruction [KH13], and Algebraic Point Set Surfaces [GG07].

Volumetric Representations Volumetric representations were originally proposed in [CL96] for range images. Here, the scene space is divided into a three-dimensional regular grid of cells, also called voxels (e.g. volumetric pixels). Each voxel stores the value of a signed distance function (SDF) that describes the distance of a voxel's center to an object's surface. Voxels also have a weight that accounts for the measurement reliability. Positive SDF values denote voxels in front of a surface point p. Voxels behind p have negative values. Often, the SDF is truncated (TSDF) to some threshold $\pm t$ to allow a compact representation of small distances. An object's surface is implicitly represented within the volume by the TSDF's zero crossings. The potential accuracy of regular voxel grids is determined by volume size and grid resolution. A high memory demand of $\Theta(n^3)$ complexity in the grid size n, makes them impractical for precise real-time applications. Hierarchical volumetric grids lower the memory consumption by storing voxels in an octree-like data structure that provides high spatial resolution only for regions containing close points (e.g. [DTK$^+$16]). Volumetric representations are popular in 3D reconstruction systems for dynamic scenes (e.g [DTK$^+$16, OEDT$^+$16, YGX$^+$17, CCS$^+$15]) because they are well suited for spatial fusion of multiple views, as well as temporal fusion of subsequent sensor readings.

3.3 View Fusion

Once a scene has been acquired and reconstructed as seen by each depth sensor, their individual views are fused into a combined model. A common approach is to employ *non-rigid registration* [DTK$^+$16] for fusion, as illustrated in Figure 3.5. There, a reference model is stored within a volumetric grid, the key volume. View data of a new frame is fused in a separate data volume. The key volume is then deformed non-rigidly. First, an embedded deformation (ED) graph [SSP07] is extracted from the implicit object surface in the key volume by uniform sampling. The ED graph is then deformed to align with the data volume, yielding local affine transformations for each ED node. Voxels in the key volume are then blended to the data volume according to transformations of ED nodes in their vicinity. A 3D mesh of the deformed frame can then be extracted using the Marching Cubes algorithm [LC87].

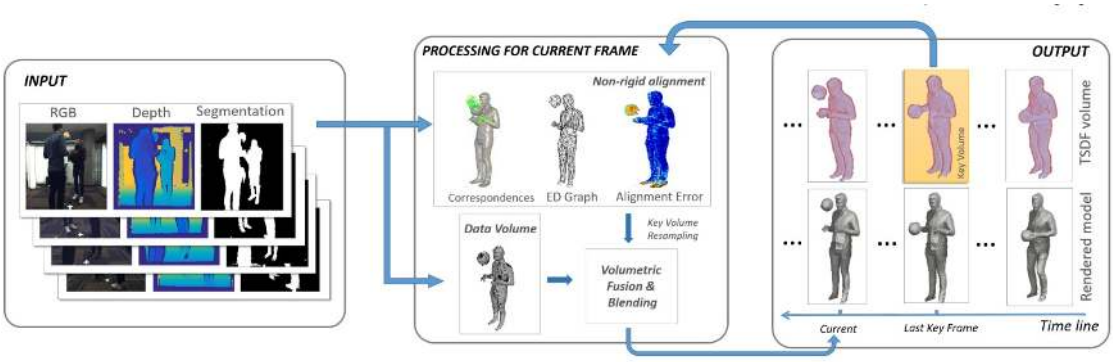


Figure 3.5: View fusion with non-rigid alignment. Figure taken from [DTK+16].

3.4 Summary

In this chapter, we have presented the principles and state of the art of 3D reconstruction using multiple sensors. The processing pipeline acquires dynamic 3D models from statically positioned stereo cameras, and computes point clouds with stereo matching. Point clouds of individual viewpoints are fused into a combined object model from which polygonal meshes are extracted. We have presented the two main methods of acquiring scene data, namely photogrammetry and range imaging. Further, we have discussed how scene depth is computed from stereo camera images by means of stereo matching. Lastly, we have outlined how 3D models from single views are combined into 3D mesh models with non-rigid registration.

CHAPTER 4

Evaluation Methods

This chapter introduces the state-of-the-art methods for evaluation 3D reconstruction system.

The following chapter is structured as follows. Section 4.1 provides an overview of available evaluation methods. Next, Section 4.2 focuses on image-based novel view evaluation. Finally, Section 4.3 discusses subjective evaluation in more detail.

4.1 Overview of Evaluation Methods

3D reconstruction systems can be evaluated in several ways. Figure 4.1 provides an overview. In particular, we can distinguish between quantitative and qualitative evaluation. Quantitative evaluation analyses properties of observations numerically. Qualitative evaluation, on the other hand, is concerned with comparing the subjective impression of an observation or product. Here, we focus on quantitative methods, while Section 4.3 is concerned with qualitative subjective evaluation.

Quantitative evaluation can further be categorised into methods that are ground truth based and those without ground truth. An overview of 3D visual content datasets in the context of 3D video quality evaluation can be found in [FBC+18].

Ground Truth-based Methods. In the context of 3D reconstruction, ground truth is data set containing highly accurate reference solutions, such as disparity maps [Midb] or point cloud [SSG+17]. Creation of ground truth data sets (e.g. [Mida, Midb, KIT]) for 3D reconstruction is often performed with structured light [SS02, SS03, SHK+14, SCD+06] or laser scanners [GLU12, SSG+17]. Given a ground truth data set, the deviation of the result of a compared method can be determined using an error measure. A typical measure often employed in the field of stereo matching is the Bad Matched Pixel (BMP) error (e.g. [CTF12]). It is defined as the ratio of disparity map pixels, whose values

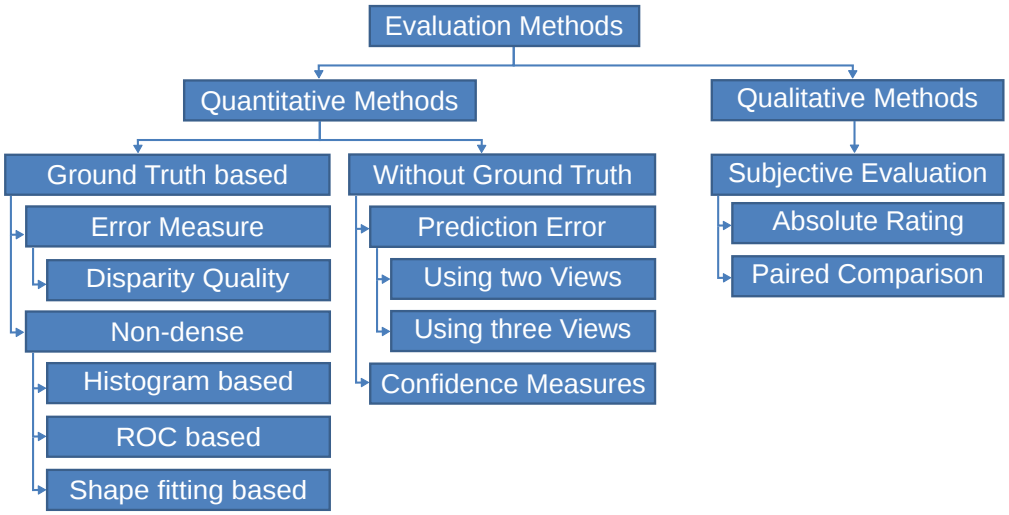


Figure 4.1: Taxonomy of evaluation methods. Figure adapted from [VCB15]

deviate from a ground truth disparity map more than a threshold value. BMP is defined over every valid pixel and is said to be *dense*. In contrast, *non-dense* methods, such as histogram- or Receiver Operator Characteristic (ROC), measure ground truth deviations only for certain regions of the ground truth. Further, shape-fitting relies on the measuring features of reconstructed models and comparing them with real-world counter parts, such as edge lengths of a complex manufactured product [BKH10].

Methods without Ground Truth. In cases where no direct comparison to ground truth data is available or viable, sources of validation can be obtained directly from the input data. This approach can be distinguished into two categories, confidence based and prediction error based methods. Confidence-measures can be computed from the reconstruction input alone. High confidence values correlate with high reconstruction quality. An example for a confidence measure in stereo vision is the Left-Right consistency check that is computed from two corresponding disparity maps. Numerous other measures have been proposed (e.g. [HM12]). Prediction error-based evaluation methods are image-based methods that employ image warping to align input data to the reconstruction output. One way, often seen in stereo vision, is to acquire data with a third camera and to use the additional sensor data as source of validation [Sze99, CGK14, MK09, SCSK13, SCK15]. Another way, which is often used in image-based rendering, is to directly employ input images for validation [WBF+17, VV14]. Both methods will be discussed in the following Section 4.2.

4.2 Image-based Novel View Evaluation

Image-based novel view methods allow evaluating a 3D reconstruction system in absence of ground truth data provided by a ready-made data set, or a high-precision depth sensor, such as a measurement laser. They validate against an image that is acquired or computed in the course of the acquisition process. We can distinguish methods that rely on an additional sensor, and those that do not.

4.2.1 Third-Eye Technique

A stereo setup is extended and registered with another camera. The additional view then acts as independent source of validation. Figure 4.2 shows an example of such a setup. This method was originally proposed by Szeliski [Sze99] and was termed *prediction error*. Today it is often called *third eye Technique* for its use of a third camera. The procedure works as follows:

1. Acquire image sequences with a stereo camera pair (c_1, c_2) and an additional third camera c_3.
2. Compute a disparity maps from (c_1, c_2).
3. Map the recorded images of the (e.g. left) camera into the image plane of c_3, thus creating a virtual novel view image.
4. Compare the novel view image with that recorded by the third camera.

If both the images of cameras basically coincide, then the disparity maps are of good quality. Root Mean Squared Error (RMS) (in [Sze99, MK09] and Normalized Cross Coefficient (NCC) (in [SCK15, MK09, SCSK13] often serve as similarity measure.

4.2.2 Two-View Evaluation

Another approach to novel view evaluation is to use camera input images as validation source. Here, the necessity of an additional sensors is lifted. Only two views are needed, the input image, and a reconstruction product warped into the view point of the input image. This type of evaluation is often performed in the field of depth-image based rendering (e.g. [VV14]) but has been recently also proposed for general 3D reconstruction systems [WBF+17].

Figure 4.3 shows an example of the method. Evaluation is performed by using a set of input images (a), corresponding reconstruction results (disparity map, point cloud, mesh model), exact intrinsic, and extrinsic camera parameters, and an optional image mask (c) that defines the area of comparison. The results are rendered into the viewpoint of the input images (b) using the reconstruction methods native rendering system. Both input image and reconstruction result are then compared in the valid mask region by a

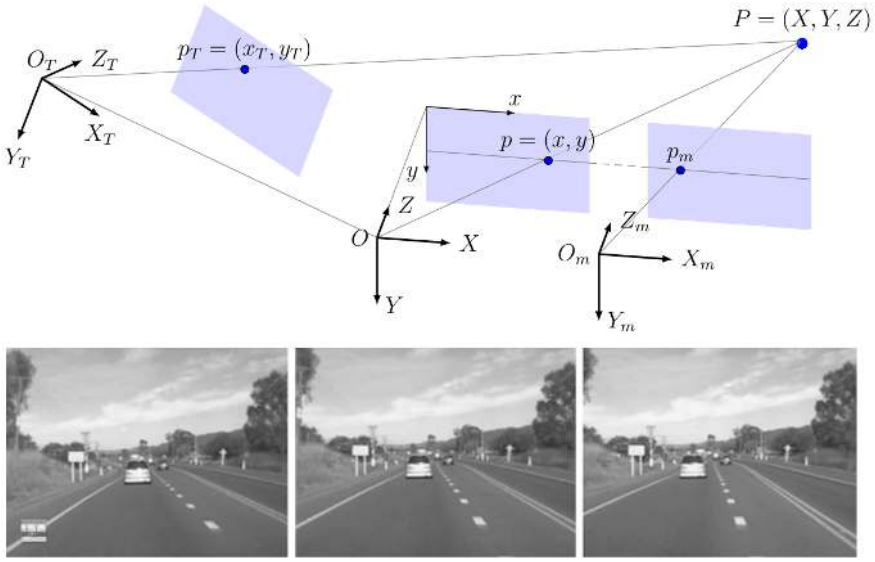


Figure 4.2: Illustration of the third eye technique. A stereo camera pair (middle, right) is extended with a third camera (left). The top shows involved camera coordinate systems. Figure taken from [Kle14].

Figure 4.3: Virtual rephotography evaluation. (a) Input Image; (b) Reconstruction result rendered in the same view port as the input image. (c) A completeness mask defines the area of comparison. (d) Error Image; (e) Visualisation of error image using the jet colour map. Figure taken from [WBF+17].

luminance invariant image metric such as (1-NCC) or Census (d). The comparison can be visualised as an error image (e) shown colour mapped.

Evaluation yields three results, namely accuracy, completeness and an error map. Accuracy of a reconstruction result is measured with respect to the used image metric. Completeness measures the amount of image pixels that contribute to the evaluation. Both results give a single value per evaluated view. The error map aids in error localisation reconstruction area locations. A rendered 3D model will usually diverge from the input image. The choice of the used rendering technique influences accuracy. Results obtained by the same technique are still comparable, though. The completeness ratio depends on the examined system. It is important, however, that the completeness value is included in interpretation of the results, as it can give valuable insights, especially for temporal comparisons.

4.3 Subjective Quality Assessment

The quality of 3D point clouds and meshes can be determined by means of subjective quality assessment. Here, the observer's perception constitutes a ground truth value. The International Telecommunication Union (ITU) has published recommendations on how to conduct subjective quality assessment in the context of television system. The aim is to arrive at meaningful, unbiased and reproducible evaluation results. A thorough review of different methods for measuring the quality of experience related to 3D video content can be found in [BÁBB+18]

This section is structured as follows. We start by introducing state-of-the-art in subjective assessment of 3D meshes and specific considerations when doing so in Section 4.3.1. Then, we continue by elaborating on study design and procedures to conduct subjective assessments in Section 4.3.2. Finally, commonly applied testing methodologies are discussed in Section 4.3.2.

4.3.1 Subjective Assessment of 3D Models

Subjective assessment of mesh models is often performed to determine the influence of algorithms modifying them, such as compression [GVC+16] and watermarking [CGEB07]. Another area is performance testing of quality measures [VSKL17, AUE17, TWC15]. Traditionally these studies use high quality benchmark data sets such as [SCD+06]. Subjective evaluation of dynamically captured 3D models, on the other hand, is an area of active research. Only few publications are concerned with quality assessment of dynamic [TWC15], or captured mesh models [DZC+18].

For subjective assessment of 3D models, no specific recommendation has been proposed by ITU, so authors usually adopt one of the test procedures of ITU BT-500.13 [ITU12] or ITU P.910 [ITU08]. These recommendations target evaluation of images, and videos in the context of television systems. Contrary to images and videos, point clouds and mesh models allow an observer to regard them from multiple viewpoints. Fixed view interaction just shows one preselected view [TWC15]. Free view interaction allows the observer to view the test material by his choice. This includes free rotation, translation and zooming. [TWC15] This procedure has two shortcomings. The first is cognitive overload of the observer. The other issue is, that each observer will have a different impression of the test material, which can bias assessment results. Another approach to user interaction is adopted in [GVC+16], where animated renderings of the material is shown. This hybrid technique allows observers to regard the material from multiple view points, while guaranteeing reproducible impressions. Torkhani et. al [TWC15] observe a significant difference between mean objective scores given by viewers of free view and fixed view setting.

Additional factors to consider when assessing quality of 3D models are the type of shading method and scene illumination. Guo et.al [GVC+16] notes that choice of position and type of illumination has as strong impact on the observer's perception of mesh models.

Characteristic	Condition
Maximum observation angle	30 deg
Ratio of luminance background behind picture monitor to peak picture luminance	≈ 0.15
Background Chromacity	D65
Room illumination	low

Table 4.1: Viewing conditions for subjective assessment as defined in ITU Recommendation ITU-R BT.500 [ITU12]

Corsini et. al suggest the use of a non-uniform background [CGEB07] and extensively comment on rendering conditions.

4.3.2 Study Design and Environmental Conditions

Extensive recommendations exist concerning the testing environment and the study design.

Environment. To ensure meaningful results, ITU recommends a controlled environment in which study participants perform their task. Table 4.1 summarises the most important aspects of a suitable environment. Observers should be seated in a distraction-free room of neutral colour and low light conditions. They should sit orthogonally to the evaluation screen.

Study Design. At least fifteen observers are recommended [ITU08]. For preliminary studies, 4 to 8 persons are sufficient. It is important to report supplemental information on the observers in order to be able to put observations into context. Particulars to report are, for example, whether the observers are naïve (non-experts) or experts, their level of expertise, occupation category (student, professor), as well as gender and age range. It is advised to include as much detail as possible in the assessment. A trial starts with observers being introduced to the task. Next, they need to be screened for visual acuity to ensure they are able to make sensible judgements. A training phase makes sure participants have understood the task at hand. Then, a testing phase starts, to allow observers to familiarise themselves with the task. Test sessions are expected to last up to 30 minutes, to avoid subject exhaustion. The exact mode and sequence of shown material depends on the goals of the study. A number of testing methodologies often used for 3D mesh model evaluation is discussed in Section 4.3.3. For image content, the recommended show time of a single stimulus is approximately 4 seconds. Dynamic content, such as video needs to be presented for a longer time, approximately 10 seconds, these times can be adapted in a particular study.

4.3.3 Testing Methodologies

In order to meet the needs of varying assessment purposes and contexts a number of testing protocols have been proposed. Two broad categories can be distinguished, impairment and quality. Impairment protocols assume comparison of a degraded, or somehow altered *test signal* against a *references signal* of known characteristics. Quality protocols, on the other hand, assess quality of either a single or multiple stimuli.

- **Absolute categorical rating** (ACR) [ITU08] is also called single stimulus method. It is especially useful, to assess quality in absence of a reference signal. One stimulus is presented at a time, after that observers judge on a five grade scale.
- **Double-Stimulus continuous quality-scale** (DSQS) [ITU12] is appropriate if a new system tested, or impairment parameters cannot be varied. Observers are shown a series of picture pairs, one item shown at a time, and can freely switch among the two. Each of the pairs, shown in randomised order, consists of an unmodified, and an impaired stimulus, both of which are themselves shown randomly ordered. Each pair is typically shown two to three times. Voting happens at the last time on a five grade scale for both images.
- **Pair comparison** (PC) [ITU08] This method is advised if the difference of original and modified stimulus is small in terms of perceived quality. In PC all possible combinations of original and changed stimuli are presented. Given n different stimuli $\binom{n}{2} = \frac{n(n-1)}{2}$ impressions are shown. Elements of pairs should be displayed in both the possible order. That is, for two stimuli A,B, both (A,B) and (B,A) are shown. Test subjects vote by expressing their preference of one stimulus over the other. Possible judgements can be “A is better than B” or “B is better than A”. Depending on the study design, observers also may vote for a tie between A and B when the presented stimuli are only slightly different in terms of perceived quality. A major drawback of the PC methodology is the high number of pairs that need to be shown. It limits the number of comparable stimuli in a trial session.

CHAPTER 5

System and Evaluation Framework

This chapter contains a detailed exposition of the examined reconstruction system, and then lays down the framework of methods used for its evaluation. First, the system is introduced and discussed in some detail. In particular, this includes hardware components and physical setup, as well as an overview of the processing pipeline. Next, the individual stages, data acquisition, calibration and registration, and depth reconstruction are described. Further, failure cases and challenges that can occur at the respective processing steps are discussed. In the second part of this chapter, methods that are used to evaluate the system are introduced and discussed.

The chapter is structured as follows. In Section 5.1, the evaluated 3D reconstruction system is presented. Section 5.2 describes the framework of evaluation methods used in the course this work. Finally, Section 5.3 provides a summary.

5.1 System Description

This section gives a detailed exposition of the reconstruction system under examination. In particular, we start with an overview of the system's processing pipeline in Section 5.1.1. Next, a description of the processing pipeline follows. Then, we give details of each step and challenges that can arise in the course of processing. Hardware and data acquisition are discussed in Section 5.1.2. Next, calibration and registration of the capturing units are detailed in Section 5.1.3. Finally, the process of depth reconstruction and model generation follows in Section 5.1.4.

5.1.1 System Overview

The system presented in this work acquires 3D models in the fashion described in Section 3.2, that is, 3D information is computed for each depth sensor first, subsequently the individual views are fused to acquire combined models. The system consists of three depth sensor units, each of them is a stereo camera. To facilitate the distinction of a single camera that is part of a unit and the depth sensor unit itself, the latter is referred to as 3D measurement unit ($3DMU$), "depth sensor" or "view" in the following. The term "camera" denotes a single camera that is part of a $3DMU$.

An overview of the system's processing pipeline is illustrated in Figure 5.1. The first step is calibration of the individual $3DMU$s. It yields intra-3DMU intrinsics, that is, all parameters necessary to faithfully compute individual viewpoint clouds. Next, the units' relative poses are determined in the inter-$3DMU$ registration step. Once the whole rig has been calibrated and registrated, image sequences of dynamic scenes are acquired. As a preparatory step for the following depth reconstruction, resulting image sequences are rectified using the intra-3DMU calibration parameters. Next, stereo matching based depth reconstruction yields individual viewpoint clouds. They are then segmented and refined in a semi-automatic post-correction step. To achieve optimal model quality, point clouds are again registered before the 3D mesh generation step merges individual views and computes 3D models in the form of combined point clouds and corresponding untextured 3D meshes.

5.1.2 Hardware and Data Acquisition

The system consists of three stereo sensors ($3DMUs$) of identical physical properties. This setup is insufficient to perform full 360-degree reconstructions. Generation of complete 3D models is not an intended goal of this architecture. Rather, it has been chosen to facilitate the system's evaluation that follows in Section 5.2.

Camera Hardware. Each 3DMU, illustrated in Figure 5.2, is assembled from two industrial-grade RGB cameras. The unit's characteristics are summarised in Table 5.1. Ximea MC050CG-SY [XIM] acts as cameras for each view. They can acquire images in a resolution of 2464×2056 pixels, respectively 5 Mega-pixels. The camera's RGB sensor has a size of 8.5×7.1 mm, and an image diagonal of 2/3 inch [1]. A PC connected to the camera with USB3.1 interface triggers frame acquisition at a frame rate of 25 Hz, while the camera's maximum frame rate is 100 Hz. Attached to each camera is a C mount lens with a nominal focal length f of 6 mm. Left and right camera comprising a $3DMU$ are mounted 70 mm horizontally apart forming the stereo baseline.

Physical System Setup. For data acquisition, the three $3DMU$s were placed in a studio environment, depicted in Figure 5.4. They faced a central spot in front of

[1] Sensor size is customarily defined in metric units, while sensor diagonal is given in imperial inch units.

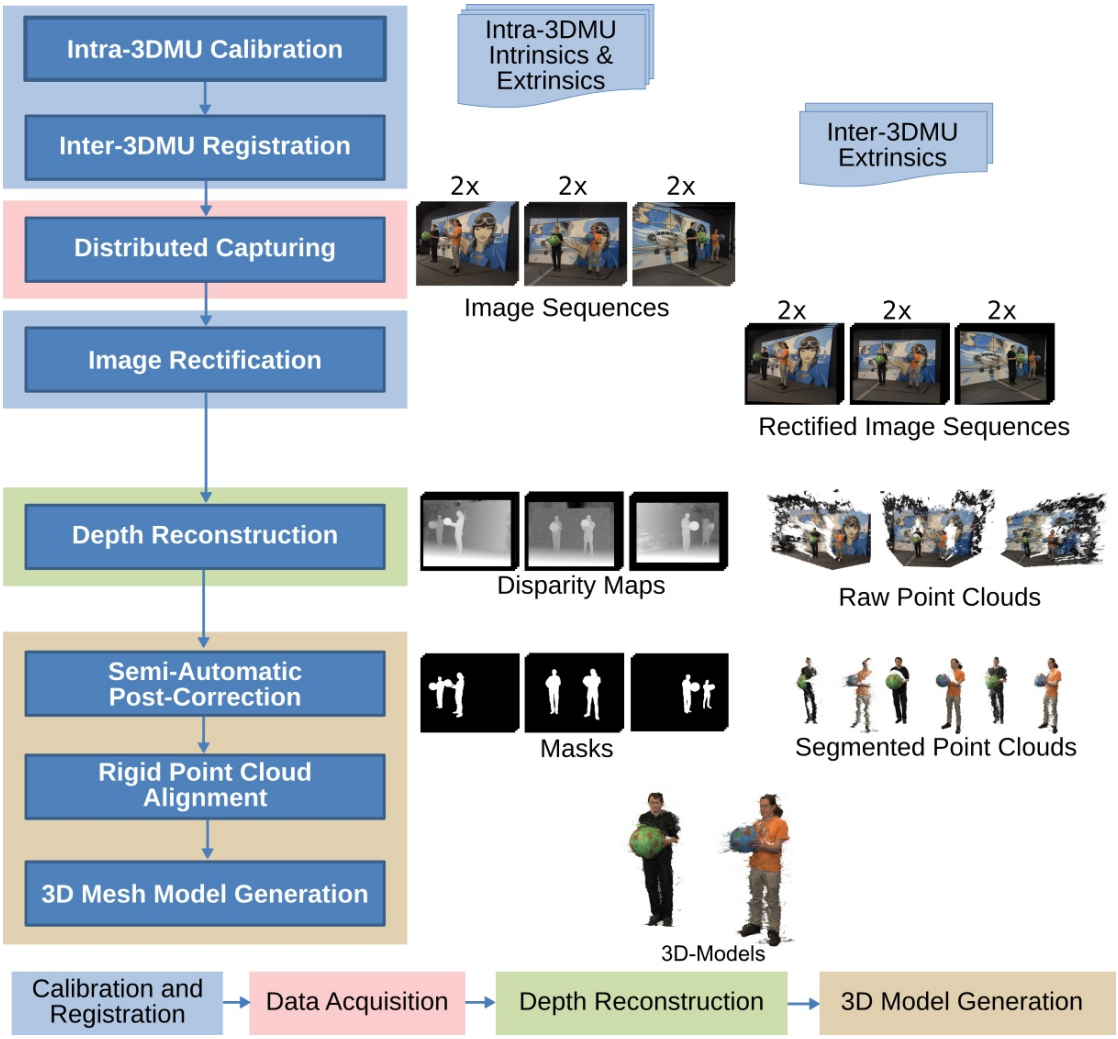


Figure 5.1: Overview of the processing pipeline and created intermediate products. We can broadly distinguish four stages, namely calibration and registration, data acquisition, depth reconstruction and 3D model generation.

Figure 5.2: Image of a 3D Measurement Unit (3DMU). It is a stereo camera comprising two Ximea MC050CG-SY [XIM] industrial-grade cameras.

Camera properties	**Unit**	**Value**
Camera model		Ximea MC050CG-SY [XIM]
Lens mount		C
Interface		USB 3.1
Sensor dynamic range (used/max)	bits per pixel	8/12
Sensor size/diagonal	mm	$8.5 \times 7.1/11.1$
Sensor size	inch	2/3
Camera resolution	pix / Mpix	2464×2056 / 5.0
Frame rate (used/max)	Hz	25/100
Lens focal length f	mm	6
Stereo baseline B	mm	70

Table 5.1: System hardware characteristics.

the background that was at roughly 4 meter distance away. The main unit, $3DMU_2$, was placed approximately orthogonally to the planar scene background. $3DMU_1$ was situated 2.2 m rightwards, and $3DMU_3$ 2.4 m leftwards to the main unit. Ground truth measurements of the relative unit placement where taken with a measurement tape (see Appendix A). In addition to ambient light from three windows, additional studio lights illuminated the scene. Camera lenses were adjusted with the help of a preview and image histogram. Lens apertures were set such that the image intensity values covered the whole histogram, while simultaneously avoiding overexposed image regions. Next, the lens focal points were jointly adjusted to have the scene's center spot in focus. Finally, the camera's white balance was carefully set to have the same chromatic properties in all views.

Data Acquisition. Once the system is calibrated and rectified (see Section 5.1.3), image sequences of dynamic scenes can be acquired. A single controlling PC is connected via Ethernet to three capturing PCs, each of which is handling acquisition for a single $3DMU$. The system time of all PCs is synchronised via the Network Time Protocol (NTP).

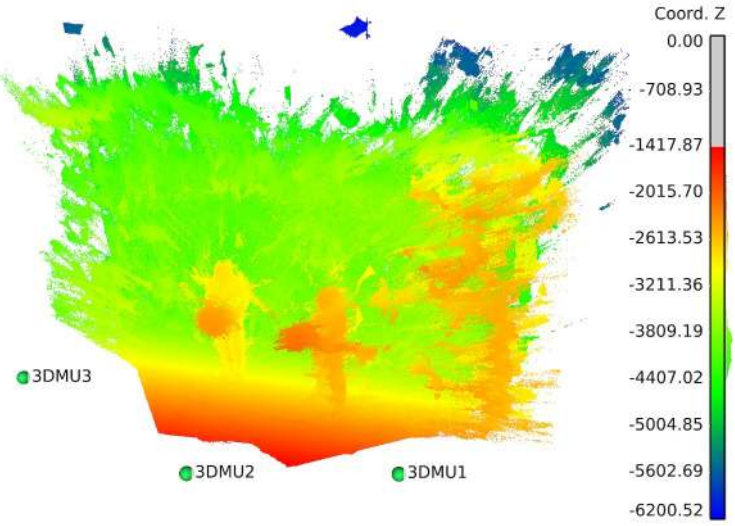


Figure 5.3: Illustration of the physical camera setup and scene distance. The green circles denote the physical positions of the depth sensor units. $3DMU_1$ is positioned at $(2214.64, -61.71, -481.90)$, $3DMU_2$ at $(0, 0, 0)$ and $3DMU_3$ at $(-2471.13, 85.6284, -1228.66)$ in millimeter units. The reconstructed scene is depicted as point cloud coloured by distance to the origin of the coordinate system in $3DMU_2$.

Figure 5.4: Physical system setup. The three $3DMU$s were placed approximately 2 meters apart facing the scene in roughly 4 meters camera distance. Controlling and capturing PCs are shown in the back. Additional to the light provided by the windows, three studio lights with diffuser boxes were used to illuminate the captured scenes.

The controlling PC starts and stops image acquisition by sending respective messages via the User Datagram Protocol (UDP) to the capturing PCs. Upon simultaneous reception of the start message, capturing PCs start a local timer and periodically trigger frame acquisition of the attached cameras at a frequency of 25 frames per second. Images that arrive at the PCs in raw sensor format are first converted into the RGB colour space, and are then saved to hard disk as uncompressed bitmap files. This format was chosen for lossless storage, and to keep processing time low. At the end of acquisition, the controller sends a stop signal causing the capturing PCs to end capturing. Data acquisition of a scene yields sequences of RGB image pairs for each of the three units.

Failure Cases. Two major failure cases with respect to data acquisition can be identified, namely image synchronisation and motion blur.

Image Synchronisation. One challenge in acquiring image sequences for depth reconstruction is image synchronisation. Stereo matching algorithms that recover depth information from dynamic scenes assume that images are acquired at the same time. This assumption does not always hold for the acquired data used in this work. An error in the camera control software caused image sequences captured by individual views to be of diverging lengths. For the same scene, one camera would output a different number of frames than others. An attempt to recover synchronous image sequences has been made according to the following procedure. First, one camera, namely the left, of each $3DMU$ was chosen as the main image source. Then for each frame of the main camera, a corresponding frame in the right view with minimum time difference to the main camera's frame is chosen. This yields a "repaired" image sequence for a single $3DMU$. The same procedure was repeated to temporally align image sequences among several $3DMUs$. Looking at subsequent frames of each camera, we can observe relative motion. Due to the fact that one camera captured the scene at a different point in time than its corresponding camera, we can see different amounts of motion in one view of a camera pair than the other (see Figure 5.5 top left and mid). The restoration procedure introduced another case of temporal misalignment. There, motion is present in one view, but none occurs in the other (Figure 5.5 top right). Disparity maps computed from insufficiently synchronised image pairs exhibit temporal flickering, depth discontinuities within objects, as well as missing disparity values (see Figure 5.5 bottom). Point clouds computed from such disparity maps show distortions in moving regions (see Figure 5.5 bottom). Subsequently, the system evaluation in this work is restricted to the first frame of each acquired scene only. There, strict camera synchronisation has been found to be present, whereas it is usually lost in later scene sections. Consequently, the evaluation of dynamic motion is not part of this work.

Motion Blur. Another factor determining the quality of acquired models is motion blur. A video camera captures a certain number of frames per fixed units of time. It is measured in frames per second (FPS). Images are acquired by integrating photons hitting the image sensor for a certain amount of time, the exposure time. The camera

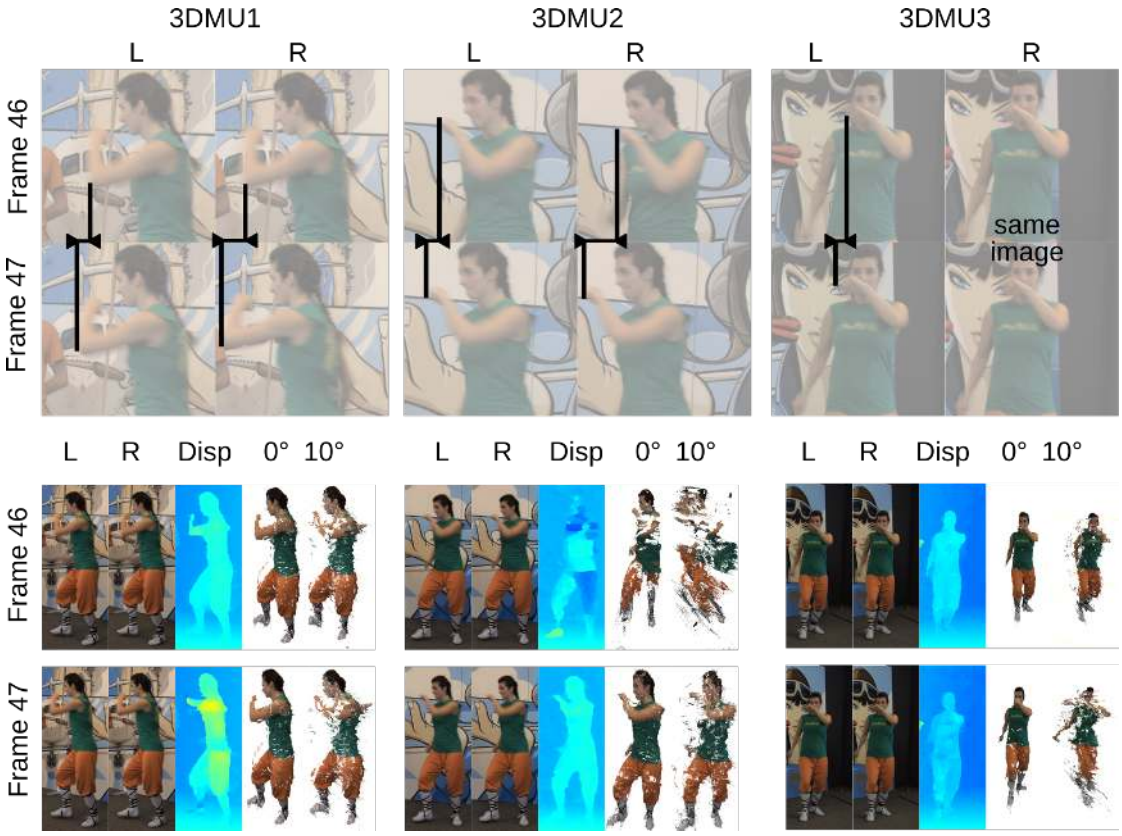
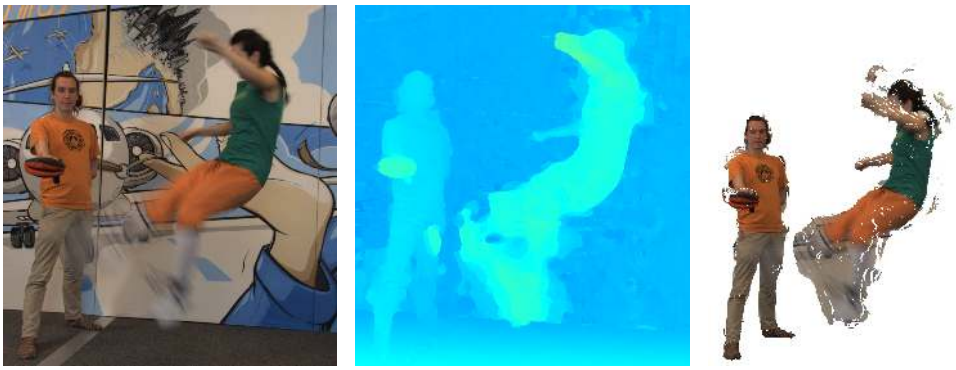


Figure 5.5: Effects of slightly unsynchronised image acquisition for two frames captured approximately 5 ms apart. The top row shows movement of different magnitude in corresponding stereo image pairs. The length of the horizontal line shows the amount of relative motion of the marked hand and elbow between subsequent frames. In a synchronized sequence, the lines are of the same length. The bottom row shows disparity maps computed from unsynchronised image pairs exhibit temporal flickering among subsequent frames. The corresponding point clouds are distorted in 3D space. Note that the temporal difference is hardly visible in the stereo image pairs.



(a) Input image (b) Disparity map (c) Point cloud

Figure 5.6: Illustration of motion blur in a scene containing fast movement. (a) Left input image; (b) Computed disparity map; (c) Point cloud.

FPS setting determines the maximum exposure time. For example, a camera capturing at 30 FPS frame rate has a maximum shutter speed of 33.3 ms. Any motion happening seen by the camera in this time period is averaged in the same image. Borders of moving objects appear blurred (i.e. *motion blur*). An example of this effect is illustrated in Figure 5.6. Borders of feet moving are almost completely blurred (Figure 5.6a). In the computed disparity map blurred areas are often assigned to the foreground, causing object fattening, (Figure 5.6b). Corresponding point clouds (Figure 5.6c) contain erroneous points not belonging to the person's foot, but rather to the background. At 25 fps, the feet of a slowly walking person appear blurred. It has been shown that moving persons can be captured at 60 fps by state-of-the-art systems [DTK+16], whereas other sources [YGX+17] report failures already at 40 fps. Increasing the capture frame rate is one measure to cope with motion blur. Another approach is to keep the frame rate fixed, while decreasing the camera's exposure time. A "staccato" effect can be observed in videos captured in this way. 3D reconstruction systems, however, use captured images not as final output, but rather as an intermediate product for model generation. This effect is not expected to play an important role in this application.

5.1.3 Calibration, Registration and Image Rectification

In order to perform depth reconstruction, first, individual camera parameters and depth sensor relative poses need to be determined. In our project, the method of Zhang [Zha00] is used to determine these parameters. Calibration and registration are performed before the actual image acquisition. The workflow is illustrated in Figure 5.7. It starts with estimating individual camera parameters. Next, relative poses of camera pairs within individual $3DMU$s are determined, yielding the intra-$3DMU$ calibration. It contains all parameters necessary to reconstruct models for each unit. Then camera poses of $3DMU$s relative to the main unit $3DMU_2$ are obtained, denoted as inter-3DMU registration. They facilitate later view fusion. Lastly, captured image sequences are rectified to allow

depth reconstruction by a stereo matching algorithm.

Intra-3DMU Calibration. We start by determining intra-3DMU calibration, and acquire all parameters necessary to reconstruct depth for a single view. For calibration, we use the camera model of the OpenCV [Bra00] library. Its camera model assumes that no camera skew is present and uses the lens distortion model of Brown [Bro66].

Single Camera Calibration. Calibration images of a circle-grid pattern (see Figure 2.5b) are acquired while it is held in multiple poses. The asymmetric pattern grid has 4 rows and 11 columns of circles. Circle center points are spread 50 mm apart between rows and 25 mm between columns. Image coordinates of the circle centers are detected from the captured calibration images. Together with the known planar 3D coordinates of the circle grid, a system of equations is formulated, whose solution yields the camera focal lengths f^i_{xv}, $f^i_{y_v}$ and principle points c^i_{xv}, $c^i_{y_v}$ for each camera $v \in \{l, r\}$ of $3DMU_i$. They can be represented as camera matrix:

$$K^i_v = \begin{bmatrix} f^i_x & 0 & c^i_x \\ 0 & f^i_y & c^i_y \\ 0 & 0 & 1 \end{bmatrix} \tag{5.1}$$

The skew parameter of K^i_v is assumed to be 0. Lens distortion coefficients are given as $dist^i_v = (k^i_{1v}, k^i_{2v}, k^i_{3v}, p^i_{1v}, p^i_{2v})$ as defined in Section 2.2.2 for each camera. In the project, the calibration process is initialised with a known nominal lens focal length of 6 mm (see Table 5.1) for both x and y dimensions to assist the calibration procedure in the determination of focal lengths, principal points and lens distortion. Said parameters are then computed by invoking the function `cv::calibrateCamera()`.

Stereo Camera Calibration. Next, relative poses of cameras within individual 3DMUs are obtained. Here, in the project OpenCV's `cv::stereoRectify()` function is used. It yields rotation- and translation-vectors R^i and T^i, respectively. They describe the rectifying transformation aligning image planes of the left view cam^l_i and right view cam^r_i belonging to $3DMU_i$, and put corresponding epipolar lines onto the same horizontal image lines. Figure 5.8 illustrates the involved transformations. Note that the translation vectors point from cam^r_i to cam^l_i. After image rectification, a 3D world point at distance infinity will have a value of 0 in the computed disparity maps. To keep the same image size along the whole processing pipeline, and to avoid aggressive image interpolation, applying the rectifying transformation warps captured image content into a sub-region of the original image. Stereo camera calibration describes these regions as bounding boxes for the left, ROI^i_l and the right, ROI^i_r views of $3DMU_i$.

Finally, we have fully parameterised the individual camera geometry. For each $3DMU_i$, $i \in \{1, 2, 3\}$, the set $Intr^i = (K^i_l \cup dist^i_l) \cup (K^i_r \cup dist^i_r) \cup R^i \cup T^i \cup ROI^i_l \cup ROI^i_r$ is then called the *intra-3DMU* calibration of the depth sensor $3DMU_i$.

Camera Registration. Once we know the intra-3DMU calibration, we can determine relative poses between individual 3DMUs. We choose the middle depth sensor, $3DMU_2$,

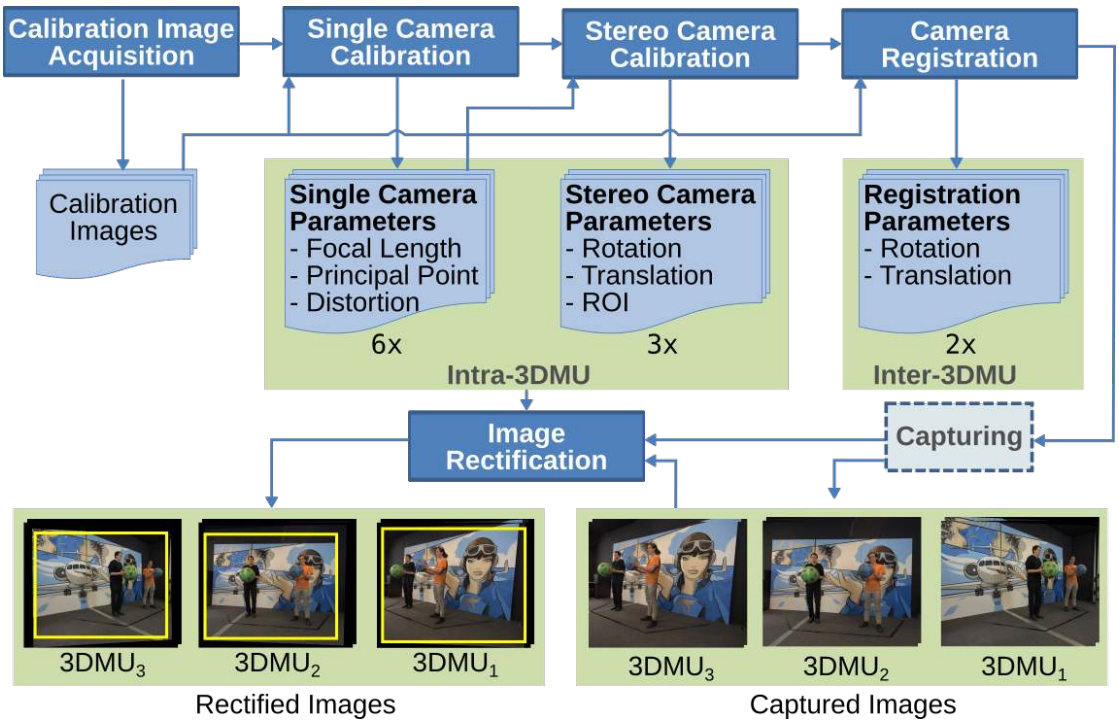


Figure 5.7: Illustration of camera calibration, registration and image rectification process. First, individual cameras are calibrated, then stereo camera calibration is determined. Camera registration yields relative camera poses. Lastly, captured images are rectified.

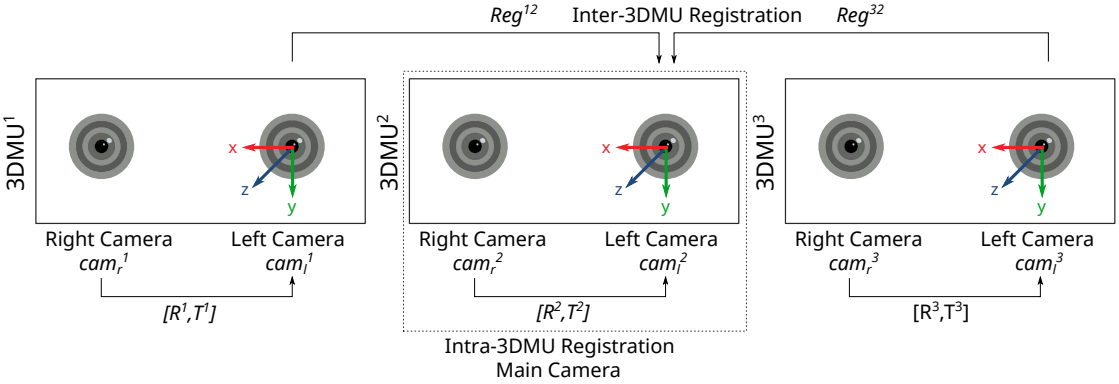


Figure 5.8: Illustration of transformations involved in intra-3DMU calibration and inter-3DMU registration. (3DMUs are viewed from the front). Transformations Reg^{12} and Reg^{32} point from the left cameras of $3DMU_1$ and $3DMU_3$ respectively to the left camera of $3DMU_2$, the origin of the 3D world coordinate system.

as the main unit. The origin of the 3D world coordinate system is rooted in its left camera. For two 3DMUs i and j, the relative pose is determined from a subset of calibration images in which the pattern is visible in both $3DMU_i$ and $3DMU_j$. The poses are given as rigid-body transformations denoted as $Reg^{ij} = [R^{ij}|t^{ij}]$. The transformation comprises a rotational part R^{ij} and a translation part t^{ij}. In the project, we determine the poses by invoking OpenCV's `cv::stereoRectify` function while keeping the intrinsic camera parameters fixed.

As a result, *inter-3DMU* registration parameters are then the set $Extr = Reg^{12} \cup Reg^{32}$. For $3DMU_2$ no transformation is necessary, as it is the coordinate system's origin.

Image Rectification. Image rectification removes distortions from captured image pairs, and warps them, so that corresponding epipolar lines align with horizontal image lines. We apply image rectification on image pairs captured by $3DMU_i$ using the intra-3DMU calibration parameters $Intr^i$ obtained in the previous step. To achieve this, we first determine rectifying transformations for the left and right images of a depth sensor with OpenCV's `cv::initUndistortRectifyMap()` function, and subsequently warp images with *cv::remap()* with bicubic interpolation.

Failure Cases. We can identify two major issues arising from insufficient calibration and registration.

Inaccurate Camera Calibration. Camera calibration influences depth reconstruction on two levels. First, images rectified with inaccurate calibration parameters can cause corresponding images to not lie exactly on the same horizontal lines. That is, the epipolar constraint is not met. Depth reconstruction with a stereo matching algorithm applied on such images can result in disparity maps where objects have incorrect depth values. Further, disparity maps may exhibit a distorted geometry. Planes, such as walls, can appear curved in 3D space. The second case is that images are correctly rectified, but the camera's focal length and principal point are determined insufficiently. Here, computed disparity maps may be accurate, but a projection of points from disparity maps into 3D space can cause the corresponding 3D points to be placed onto wrong spatial coordinates. We observed that the point clouds may then be either distorted, or of incorrect size.

Inaccurate Depth Sensor Registration. Insufficient depth sensor registration can cause reconstructed point clouds to be placed at the wrong position in 3D space. In that case point clouds of individual views may not form a coherent combined view, but rather the point clouds can diverge or appear rotated relative to each other.

5.1.4 Depth Reconstruction and 3D Model Generation

Depth reconstruction and model generation involves first recovering depth information from rectified image pairs. Depth is represented in the form of disparity maps. Next,

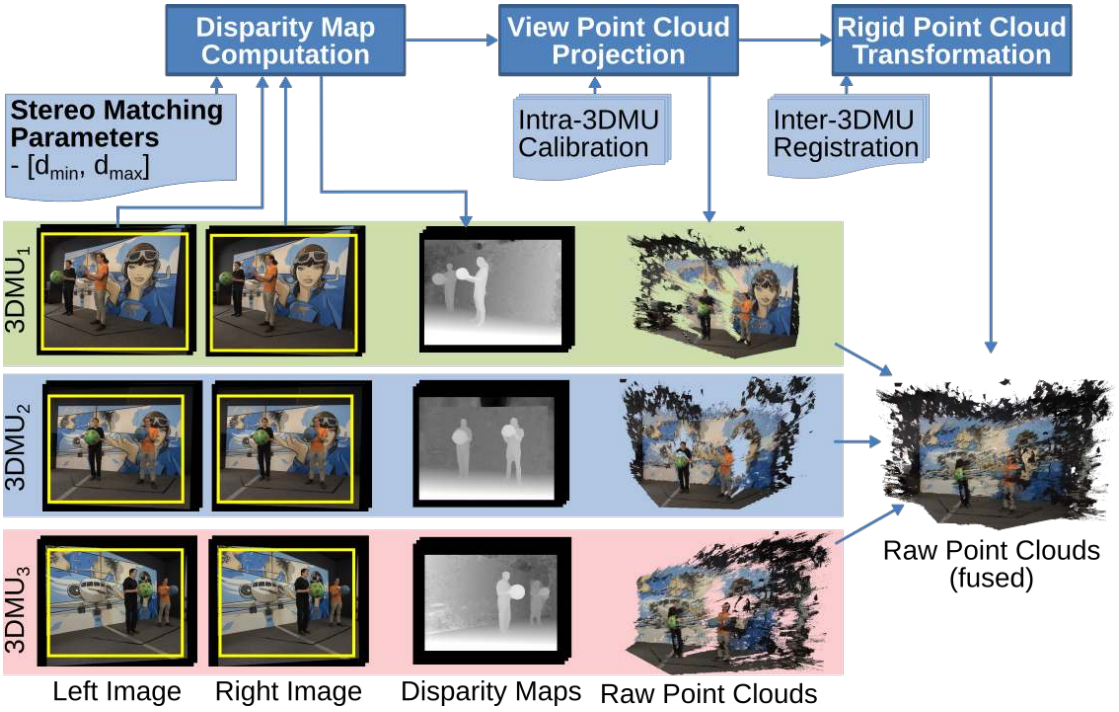


Figure 5.9: Illustration of the depth reconstruction process. Disparity maps are computed from rectified image pairs. From disparity maps, first, point clouds are projected into 3D space, and are then transformed into a shared coordinate system.

disparity maps are turned into a 3D point cloud representation whose points contain colour information obtained from the corresponding input images. The point clouds are then segmented, refined, and again registered to prepare them for the model generation. Our processing pipeline first turns point clouds into 3D meshes of individual views and then fuses views into combined 3D meshes.

Point Cloud Generation

Point cloud generation is performed in two stages. First, we compute disparity maps from rectified image pairs for each single $3DMU$. Then, in the second stage, we convert the disparity maps into 3D point clouds, and then transform them into in a shared coordinate system. The process is illustrated in Figure 5.9.

Disparity Map Computation. We compute disparity maps from image pairs of each $3DMU$ by stereo matching. An algorithm based on cost volume filtering [SNG+15] is employed for this purpose. Specific parameters and algorithm stages are summarised in Table 5.2, and are described in the course of the following paragraphs. To increase performance, and to a lesser degree to avoid disparity errors, we restrict stereo matching

Algorithm Stage	Method	Parameter Value
Cost computation	Census [ZW94]	W=5 × 5
Cost aggregation	Permeability Filter [ÇığlaAA13]	$\sigma = 12$
Disparity post-processing	Left-right consistency check	Threshold: 0
Hierarchical Matching	[SNG+15]	Levels: 0-2
Temporal filtering	Permeability Filter [ÇığlaAA13]	$\sigma = 12$, $\sigma^t = 1$

Table 5.2: Stereo matching parameters used for disparity map computation.

to image areas containing active valid pixels. They are given as part of the intra-3DMU calibration (see Section 5.1.3) and are denoted by ROI_l^i and ROI_r^i.

Cost Computation. Matching costs are computed with the Census [ZW94] similarity measure (see Section 3.2.1). Here, it measures the dissimilarity or costs between two potentially corresponding pixels in left and right images of a pair. A local rectangular window of 5×5 size centered around pixels is taking into account for matching cost computation.

Cost Volume. The per-pixel matching costs are stored in a three-dimensional matrix, the cost volume, of dimensions $W \times H \times D$ in pixel units. W and H denote image width and height, and $D = |d_{max} - d_{min}|$ is the size of the disparity search range (d_{min}, d_{max}). The disparity search range is given to the algorithm as input parameter, and has to be chosen the corresponding to minimum and maximum disparity values occurring in the rectified image pairs. Note that disparity map values computed by the system are negative numbers, and consequently d_{max} denotes the disparity value of the farthest point considered, while d_{min} denotes the value of the nearest point.

Cost Aggregation and Disparity Selection. The cost volume is then filtered in the cost aggregation step with the edge preserving permeability filter [ÇığlaAA13]. The disparity value for a pixel p is selected with the common Winner-takes-it-all (WTA) strategy. Out of all D possible disparity values in the cost volume at position of pixel p, the one with minimum matching costs $C(p, d)$ is chosen:

$$d(p) = argmin_{d \in D} C(p, d) \tag{5.2}$$

Hierarchical Matching Scheme. Disparity estimation is embedded in a hierarchical matching scheme to improve run-time and quality of the result. For each image pair, a Gaussian image pyramid with $k = 3$ layers is built. Stereo matching is first performed on the coarsest layer l_2. Images in this layer have a resolution of $\frac{W}{2^2} \times \frac{H}{2^2}$ pixels. Based on this initial disparity map, an offset map is computed that guides disparity estimation on the next finer layer l_1 on images with $\frac{W}{2^1} \times \frac{H}{2^1}$ pixels resolution. The process

is then repeated for the finest layer l_0 of the Gaussian pyramid with images of native resolution.

Consistency Checking. Next, inconsistent disparity map values are removed with a left-right consistency check. To do so, in addition to the original disparity map $D_l r$, a second disparity map $D_r l$ is computed with the right view as the reference image. Any pixels whose corresponding disparity values differ between D_{lr} and D_{rl} by more than a threshold value τ are marked as invalid. Here, τ is set to 0, to minimise the number of erroneously computed pixels.

Sub-pixel Refinement. In the cost volume approach possible disparity values are represented as slices of the cost volume. WTA disparity selection chooses as disparity the index of the slice with minimum costs, yielding integral values. This leads to points being arranged along discrete planes in the corresponding point clouds. An enhancement step recovers sub-pixel accuracy for each pixel p with disparity d by fitting a parabolic curve through neighbouring cost values at p's position in the cost volume [YYDN07].

$$d_{sp}(p) = d - \frac{C(p, d+1) - C(p, d-1)}{2 \times (C(p, d-1) + C(p, d+1) - 2 \times C(p, d))} \tag{5.3}$$

where d_{sp} denotes the interpolated disparity value of pixel p at sub-pixel precision.

Temporal Filtering. Slightest changes in capturing conditions can lead to temporal noise between subsequent frames of a video sequence. In order to reduce it, temporal filtering with the permeability filter is applied. The cost penalties approach of [ÇığlaAA12] is used. Specifically, the cost volume is updated with a temporal consistency term. First, the temporal change of each pixel in terms of its RGB value is determined by computing its permeability weight μ^t:

$$\mu^t = min(e^{-\frac{\Delta R}{\sigma}}, e^{-\frac{\Delta G}{\sigma}}, e^{-\frac{\Delta B}{\sigma}}) \tag{5.4}$$

where ΔR, ΔG, ΔB denote the absolute difference between R, G and B channel values of a pixel in two subsequent image frames, and σ is the permeability filter smoothing factor. A pixel's permeability weight μ^t determines the ratio of pixels to be transferred from the previous to the next frame. μ^t is high for pixels with similar RGB values, those in non-moving areas, and low for pixels in fast moving areas, as they have diverging RGB values. Next, the cost volume is updated. For each pixel $p = (x, y)$ at disparity value d and costs $C(p, d)$, updated costs $C^t(p, d)$ are computed as:

$$C^t(p, d) + \mu^t \times |d - d^{t-1}(p)| \times \sigma^t \tag{5.5}$$

where d^{t-1} is the disparity value of p at the previous frame. σ^t is a temporal smoothing factor. For pixels in fast moving regions, those with high permeability weight, the new disparity value is favoured, while for pixels in non-moving areas disparity values change smoothly from one frame to the next. The temporal smoothing factor σ^t is set to 1.

Disparity Map Format. Computed disparity maps are single channel images stored in EXR [EXR] formatted files. The format has been chosen for its capability to store the disparity map values as floating-point numbers. Note that the representation differs from other implementations that often store disparity values scaled as positive integer numbers (e.g. [SS03]).

Individual Point Cloud Projection. Next, disparity maps are projected into 3D space using the known intra-3DMU calibration, and are put into the coordinate system of the main unit, $3DMU_2$. This step is illustrated in Figure 5.9. Its input constitute disparity maps D^i and rectified RGB images I^i_{rect} of each individual $3DMU_i$, with $i \in \{1, 2, 3\}$. The intra-3DMU calibration parameters $Intr^i$ are used to project point clouds of individual views. Specifically, first the disparity map D^i is projected into 3D using the following reprojection matrix Q [Bra00] in homogeneous coordinates:

$$Q = \begin{pmatrix} 1 & 0 & 0 & -c_x \\ 0 & 1 & 0 & -c_y \\ 0 & 0 & 0 & f \\ 0 & 0 & -1/T_x & (c_x - c'_x)/T_x \end{pmatrix} \tag{5.6}$$

where $c = (c_x, c_y)$ is the left camera's principle point, f is the focal length. T_x is the stereo baseline between left and right camera, and c'_x is the x coordinate of the principle point of the right camera. Points of the disparity map of $p = (x, y) \in D^i$ of $3DMU_i$ with disparity value d are then projected to 3D point clouds $P^i = (X, Y, Z)$ as follows:

$$Q \begin{bmatrix} x \\ y \\ d \\ 1 \end{bmatrix} = \begin{bmatrix} X \\ Y \\ Z \\ W \end{bmatrix} \tag{5.7}$$

3D point coordinates are then given as $P^i = (X/W, Y/W, Z/W)$, where W is the fourth dimension of the homogeneous coordinate presentation, and can be chosen as $W = 1$. To allow convenient viewing in off-the-shelf 3D viewing software, the point clouds P_i undergo an additional transformation that flips both the Y and Z axis of points in P_i and yields point clouds P_{view}. This transformation causes Z coordinates to be negative values. Smaller Z coordinates then denote points that are farther away from the camera.

Shared Coordinate System. Finally, point clouds are mapped into a shared world coordinate system rooted in the left camera of the middle depth sensor, $3DMU_2$, using inter-$3DMU$ registration $Extr^i$.

$$P^i_{raw} = [R^i|t^i]P^i_{view} \tag{5.8}$$

where P^i_{view} denotes the point cloud computed in the previous step, and $[R^i|t^i] \in Extr^i$ denotes a 4×4 rigid transformation matrix that maps points of $3DMU_i$ into the coordinate system of the left camera of $3DMU_2$.

The output of the point cloud projection constitute point clouds $P^i_{raw} = (X, Y, Z, R, G, B)$ where X,Y, and Z are coordinates in the shared coordinate system, and R,G,B are red, green and blue colour channel values obtained from the corresponding input images I^i_{rect}. The point clouds P^i_{raw} are stored in the PLY [PLY] format.

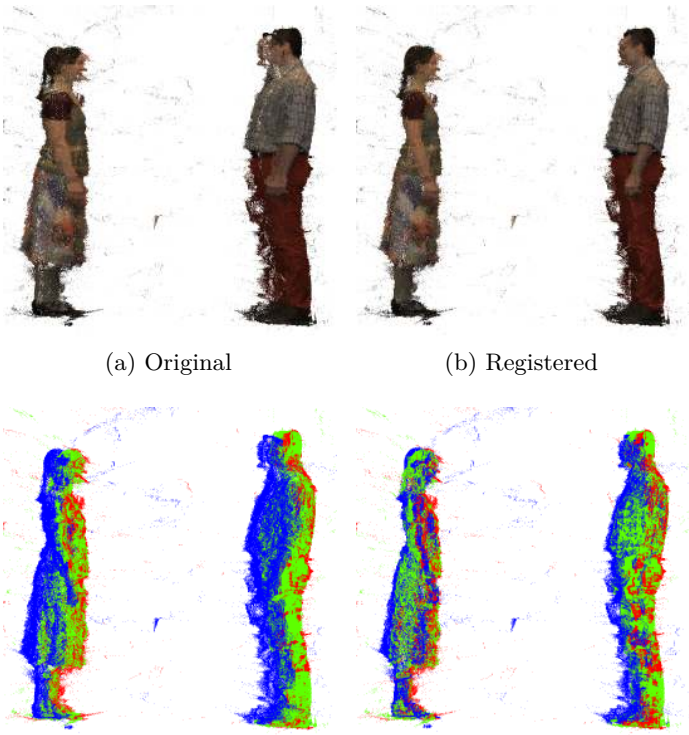
Rigid Point Cloud Registration

The relative pose between individual views (see Section 5.1.3) is refined in the 3D domain on a per scene basis to achieve optimal view overlap. An example of the updated registration and its effect on the scene is shown in Figure 5.10. The refined registration shifts points of $3DMU_3$ to the right, and removes duplication of the right person's front. As input for rigid point cloud registration, a segmented point cloud containing only the object P^i_{raw} reconstructed from $3DMU_1$ and $3DMU_2$ is aligned to the static $3DMU_2$. The point clouds are registered by our project partners using a variant of the Super4PCS [MAM14] algorithm that takes point coordinates, as well as points' RGB colour information, into account. This process results in two rigid body transformations T^{12}_{corr} and T^{32}_{corr}. Applying these transformations to $3DMU_1$ and $3DMU_3$ brings them into optimal alignment with $3DMU_2$.

Semi-automatic Post-Correction

A custom interactive software application, developed outside of the scope of this work, is used to process and examine point clouds based on previously computed disparity maps (see Section 5.1.4). The application serves multiple purposes. First, it enables the creation of segmentation masks that identify scene regions that serve as input for model generation. Next, it creates registered and corrected individual viewpoint clouds. Lastly, it contains several tools for point cloud refinement, such as cleaning point clouds from outliers.

Image Segmentation Although models for the whole captured scene can be generated, usually, only parts of the scene are of interest. Identification of these areas is done by creating segmentation masks. The user draws coarse annotations (i.e. "scribbles") over target objects (Figure 5.11a). Two kinds of scribbles can be drawn, namely foreground- and background-scribbles. They identify object- and background-regions, respectively. Scribbles are interactively expanded using a model [Bro16] that incorporates colour-information from the view's input image, as well as depth-cues from disparity maps. To assist the user, scribbles can be propagated onto the views of other 3DMUs, employing inter-3DMU registration. The same model [Bro16] allows the propagation of drawn scribbles temporally over several image frames, and facilitates video segmentation. Once the user is satisfied with the annotation, scribbles are then converted into object masks (see Figure 5.11b). In another step, the computed object masks are then propagated temporally to subsequent frames of a temporal sequence. The application contains several tools, to facilitate accurate mask generations. Among them are morphological operations,



(a) Original (b) Registered

(c) Original coloured (d) Registered coloured

Figure 5.10: Example of point cloud registration. (a) Original point clouds; (b) Point clouds registered; (c) Original points of 3DMUs 1,2, and 3 coloured in red, green, blue, respectively; (d) Coloured registered point clouds. Notice that the transformation shifts $3DMU_3$'s points into the direction of $3DMU_2$.

such as image dilation and erosion, and 2D filters, like a Guided Filter [HBG13] and Weighted Median Filter [MHW+13] to further refine segmentation masks.

Point Cloud Enhancement and Projection Either a full (Figure 5.11c), or a segmented scene (Figure 5.11d) can be projected into 3D space, to create registered point clouds. These contain spurious point errors, called outliers. An outlier filter [WKZ+16] is used to remove them.

Mesh Model Generation

After generation and refinement of the individual viewpoint clouds, they are processed into reconstructed models in both point clouds, as well as an untextured mesh representation (see Section 3.2.3). The algorithm is based on Algebraic Point Set Surface reconstruction (APSS)[GG07]), and has been developed by the project partner company Rechenraum. The model generation process is illustrated in Figure 5.12.

(a) Scribble (b) Mask

(c) Projected point cloud (d) Segmented persons

Figure 5.11: Illustration of the semi-automatic post-correction application. (a) Creation of a segmentation mask with foreground- and background-scribbles; (b) Segmentation mask obtained from scribbles in (a); (c) Full scene point cloud; (d) Segmented point cloud after applying the segmentation mask.

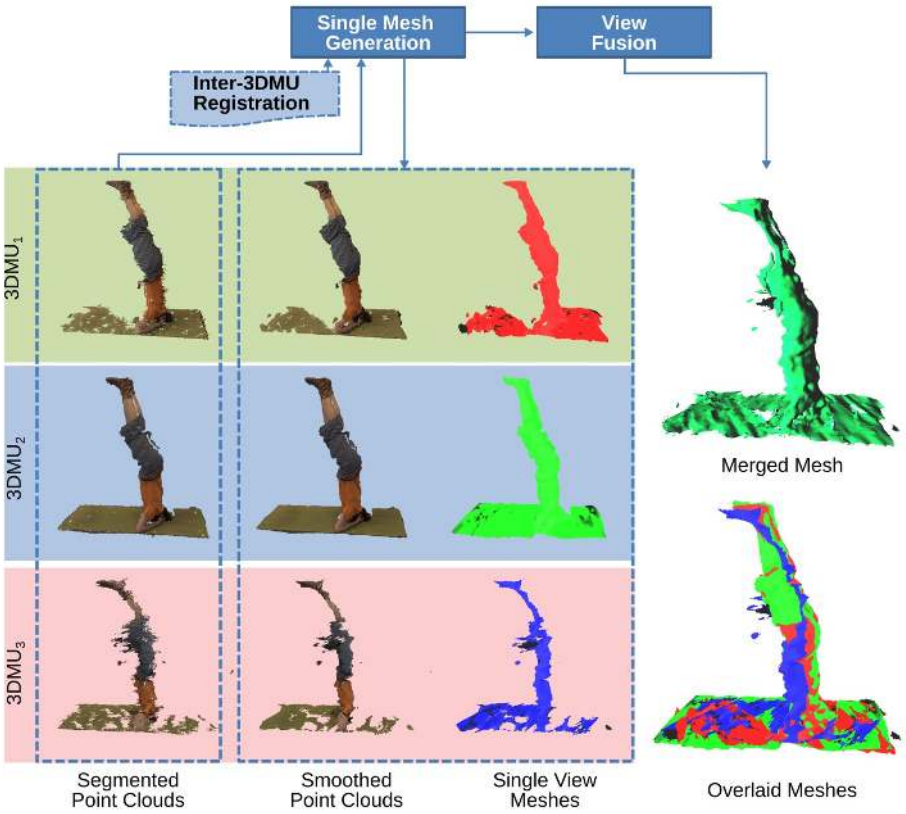


Figure 5.12: Illustration of mesh generation. First, point clouds of individual views are processed into smoothed point clouds, and mesh models. Then, they are fused into combined meshes. A resulting merged mesh is shown on top right. For comparison, meshes from individual views are shown overlaid on the bottom right.

Single-view Mesh Generation. Point clouds of single views serve as input for 3D model generation. The generation process delivers as output smoothed coloured point clouds, as well as untextured meshes. Further, model generation performs per-scene rigid registration (see 5.1.4) among different views in the course of its computation.

View Fusion. Smoothed point clouds are fused in a volumetric grid (see Section 3.2.3). Model fusion is performed both spatially to integrate individual views, and temporally to fuse models of subsequent points in time. The volumetric grid representation divides 3D space into grid cells covering 3D space. Points that are put into the grid are stored as signed distance to the actual model surface. A combined surface model is extracted from the grid, by iterating through cells, and extracting from the cells points that lie at a zero distance to the actual surface. Mesh extraction is performed with the marching cubes [LC87] algorithm.

Failure Cases

Depth reconstruction is an ill-posed problem, and is consequently prone to errors. We can broadly identify two major areas that can cause inaccurate results. The first has to do with artefacts introduced in disparity map computation. The second area lies in mesh model generation and subsequent fusion of the single views.

Depth Reconstruction Artefacts. Stereo matching recovers scene depth by identifying pixels belonging to the same imaged objects in image pairs. In addition to occlusions, that is pixels, which are only visible in one of the two images, untextured regions, such as evenly white walls, and specular or transparent image regions often cause stereo matching algorithms produce erroneous matches. When disparity maps are projected into 3D space as point clouds, inaccuracies become more prominent. Disparity map noise within an object is translated to ill-defined surfaces in point clouds. This makes reconstruction of fine object features challenging. In the example of the dancer's head depicted in Figure 5.13, facial features like nose and cheeks are hard to distinguish in the point cloud. Subsequent 3D model generation also applies outlier detection and smoothing to fit smooth surfaces onto the remaining points. In our example, this causes the nose to appear flattened in the computed mesh model. The reverse effect can be observed on the chin. It appears more pronounced than it is in reality. For larger objects, the amount of point cloud noise poses a lesser problem, as can be seen on the torso in Figure 5.14. The high amount of smoothing necessary to recover other, more delicate areas, however, leads to spurious bumps, as can be seen at the arms.

View Fusion Artefacts. View fusion is performed to achieve complete models. The fusion method has to cope with overlapping, and slightly misaligned regions. The person's feet in the overlaid mesh shown in Figure 5.12 give an example of such a misalignment occurring despite careful unit pose registration, and per-scene re-registration. Misalignment can lead to degraded models that exhibit duplication of misaligned features. This issue can be mitigated by iteratively updating pose registration on a frame-by-frame basis from reconstructed model data [ZSG+18]. Another type of artefact occurring while fusing models are chromatic misalignments. Model colours diverge slightly in the individual views due to differing illumination conditions at respective depth sensor positions. Careful chromatic alignment of the capturing cameras is not sufficient to eliminate this effect. Some state-of-the-art systems, such as [DCC+18], additionally perform chromatic alignment by creating linear colour mapping among the views offline, and applying them to the captured images. The generated meshes are untextured. They are textured by simple colour interpolation from the corresponding smoothed point cloud. In order to acquire high resolution texture and to eliminate ridges in fused model textures, more sophisticated texture stitching approaches need to be applied, e.g. [EFR+17, DCC+18]. Accurate texture reproduction, however, was outside of the scope of the examined system.

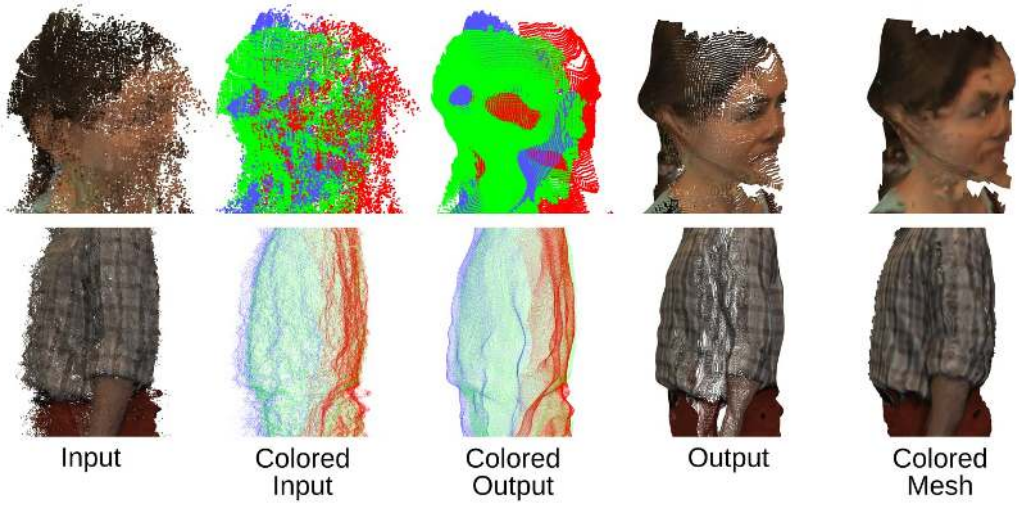
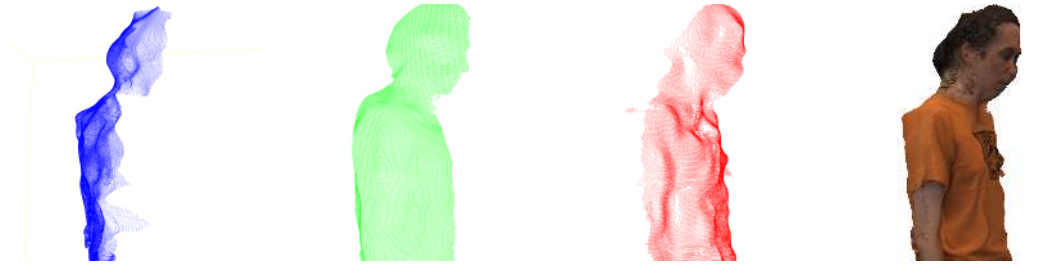


Figure 5.13: Detailed views of the `dancers1` data set. Input point clouds, shown in colour and colour-coded by capturing $3DMU$, are noisy in small local regions, like the head. Output point clouds are aggressively smoothed (middle). The mesh output is initially untextured. Texture information is added by colour interpolation from the output point cloud.



(a) $3DMU_3$ (b) $3DMU_2$ (c) $3DMU_1$ (d) Merged Points

Figure 5.14: Illustration of model artefacts introduced by views. $3DUM_3$ (a) introduces bumps on the back. While $3DMU_2$ (b) captures the persons nose accurately, $3DMU_1$ captures no distinct nose. Chromatic artefacts in the final mesh (d) are introduced by simple colour interpolation.

5.1.5 Summary

In this section, the examined system has been described in detail along the processing pipeline's major building blocks. First a high-level overview of the processing pipeline and processing steps has been given. Then, the individual stages depth acquisition, calibration and registration, and depth reconstruction have been discussed. Further, challenges and failure cases that can arise at the respective steps have been discussed.

5.2 Evaluation Strategies

This section describes the evaluation methods used to evaluate the system described in Section 5.1. It is structured as follows. Section 5.2.1 lays out the evaluation on validation objects. Next, in Section 5.2.2 the novel view evaluation method is described. Lastly, Section 5.2.3 covers the design of the user study conducted in this work.

5.2.1 Evaluation on Validation Objects

Ideally, 3D reconstruction recovers metric properties of the captured scene. Lengths and angles measured on real objects are the same as on their reconstructed counterparts. Any given real system inevitably exhibits deviations from the ideal situation. Goal of the ground truth evaluation is to quantify these deviations, and to characterise the system's ability to recover metric properties of captured objects. We consider two different types of objects.

The first type are spherical objects. From a geometric perspective, spheres are simple bodies that can be defined only by their position in 3D space and their radius. Reconstruction of spheres, however, can be challenging, since, apart from their outline, they do not exhibit any distinct geometric features. An important measure is the distribution of deviation of sampled points compared to their ideal position. Depth- as well as surface reconstruction algorithms often perform filtering that can lead to sphere models not being round but rather "bumpy".

The second type of object examined are cuboid bodies, that is boxes. Adjacent faces of a box join in a 90 deg angle. Deviations from this angle allow us to identify spatial skew in the reconstruction.

The evaluation on validation objects will be carried out in the following manner:

1. **Creation of validation objects** Spherical and cuboid objects will be created. Special care will be put into the quality of their surfaces, as they need do be properly acquired by the system. Another consideration is their size. They will be made large enough, so they that their reconstructions comprise sufficiently many sample points. For spheres, the measured diameter will be used as ground truth value.

2. **Validation object acquisition** The objects will then be acquired with the system. Samples from various object positions will be taken, in order to get diverse data set.

3. **Object segmentation** A semi-automatic post-correction application (see Section 5.1.4) will then be used to segment out only points constituting the validation objects. For spheres, the whole visible sphere area is considered. For boxes, faces will be segmented separately to allow plane fitting on each face. After segmentation, point clouds only consisting of the object areas will be generated.

4. **Shape Fitting** Validation object point cloud samples will undergo a shape fitting process. For spherical objects, an ideal sphere with measured true sphere radius will be fitted into the point cloud samples, determining the position in space that best fits the sample. For cuboid objects, plane fitting will determine the plane optimally describing the orientation of each face. Shape fitting is performed for sphere and planes by means of the non-linear least squares optimisation [Sha98]. The sphere fitting procedure is performed by our project partner.

5. **Analysis of shape fitting results** Once the ground truth data set is available, deviations from ground truth measurements will be determined. The signed error between the distance of a reconstructed point to the fitted sphere center and the sphere's true radius is determined. Another area of interest is the relation between the signed distance error and the distance of the object to the camera.

5.2.2 Novel View Evaluation

For novel view evaluation, we adopt the image-based virtual rephotography method proposed in [WBF+17] that has been described in Section 4.2. The goal of this evaluation is to determine the quality of intermediate and final reconstruction products in terms of similarity to the input images of the reconstruction process. Selected input images serve as ground truth against which the reconstruction products are compared.

1. **Data set acquisition** Representative scenes of persons are acquired and processed into 3D models. Several variants are created to examine their performance on different variants of model generation. A full description of the considered variants is given in Section 6.1.2. For each variant and model, *original point clouds*, *smoothed point clouds*, and *textured meshes* are prepared for examination. The use of these products allows to track the results of individual processing steps over the course of the reconstruction process.

2. **Novel view generation** Next, novel view images of the created products are rendered into the viewpoint of each depth sensor, such that the novel view images ideally show the exact same content as the respective rectified input image. View generation takes into account intra-3DMU calibration parameters to determine

the corresponding field-of-view, as well as inter-$3DMU$ registration to replicate the pose of the capturing depth sensor camera.

3. **Accuracy computation** Then, the accuracy of all products and variants is determined. Accuracy is defined as the similarity between the (rectified) image of the depth sensor's left camera and the novel-view image of the corresponding product. For similarity computation, the normalised cross coefficient (NCC) measure is chosen due to its robustness to illumination and exposure changes. Further, the NCC measure has been shown to correlate to user opinions in subjective evaluations [WBF+17] on 3D reconstructed models.

   Similarity is computed in the masked region used for segmentation of the foreground models from the scene background. NCC values are linearly transformed into the $[0, 1]$ value range, so that a value of 0 indicates total dissimilarity, and 1 identity of compared images. A window size of 30 pixels is used for similarity computation.

   Model completeness is determined for all comparisons. It is defined as the ratio of valid mask pixels with respect to the total numbers of image pixels. Waechter et al. [WBF+17] deem reporting of completeness ratios as necessary in conjunction with accuracy, for fair comparison.

4. **Analysis of similarity results** Model completeness is reported for each model. For accuracy, two specific cases are considered, reconstruction at the original camera position and reconstruction at a novel view. In the first case, the accuracy of each depth sensor is determined with views corresponding the original sensor's position. In the second case, models that have been fused from the two outermost views $3DMU_1$ and $3DMU_3$ are compared against the reference view $3DMU_2$. Since data of the reference view is not part of the compared models, it can be used as an independent source of validation.

5.2.3 Subjective User Study

This section describes the experimental setup and design of the user study conducted for this work. It serves two aims. The first is to assess how the participants rate the design of the proposed study. Its results will be used as feedback for the design of a subsequent larger study with more participants, which is outside the scope of this diploma thesis. The second goal is to compare different 3D model reconstruction methods in terms of perceived model quality (see Section 6.1). The participants' judgement is taken as measure for the quality of the selected reconstruction methods.

Testing Environment. A room was prepared to allow users to regard the testing material without visual distraction. While the trial was in progress, the room was kept darkened, so to achieve uniform and controlled illumination conditions. We set up a testing environment, shown in Figure 5.15, according to the requirements of [ITU12]. A laptop was running a software developed for this purpose. The material was presented on an uncalibrated monitor (Asus VG278) with a diagonal of 27 inch and a display

Figure 5.15: Illustration of the lab environment for subjective evaluation.

resolution of 1920×1080 pixels. The monitor's white balance was set to 5000 Kelvin, its brightness was adjusted to a minimum. Participants were seated at approximately 0.9 meter distance to the screen.

Testing Material. We rendered videos of textured 3D meshes in front of a non-uniform (violet to black) background. Each video was shown for a fixed time of 20 seconds. To allow viewers to concentrate on the models, the scene background was removed as described in Section 5.1.4 with the same segmentation masks as for the novel-view evaluation (see Section 5.2.2). In the course of the video the camera's viewpoint moves slowly on a path, and stops at certain positions for one second. The positions are those of the depth sensors capturing the models (see Section 5.1.2). A fourth, additional, viewpoint above $3DMU_2$ is introduced as a novel viewpoint. This procedure is adopted for two reasons. First, it allows viewers to regard the material from different perspectives. Second, it avoids mental overload. Participants can concentrate on regarding the presented material without having to navigate themselves within the scene. This aspect is especially helpful for naïve participants not familiar with navigation in 3D scenes.

Study Design. The Pair Comparison (PC) methodology described in Section 4.3 is adopted for this assessment. The study design is similar to that of [Nez14, NBSG14]. In PC, pairs of stimuli (e.g 3D models, videos, images, etc.) are displayed to the participants. They express their preference for one item of a pair with an "A is better than B" or "B is better than A" choice. Participants are also allowed to vote for "no preference" [LGE13], as the compared approaches can be similar. The trial sessions lasted for approximately 40 minutes. One participant at a time was admitted to the trial.

A session comprises five stages, with the steps being (1) introduction, (2) screening, (3) practice, (4) experimental trial and (5) interview. Participants where briefly introduced orally, and where given written instructions in either English (see Figure B.1) or German

language (see Figure B.2) explaining their task (1). In the following user screening (2), participants performed a visual acuity and a colour vision test. Visual acuity was tested with a Snellen chart (see Figure B.5) printed on an A4 format sheet of paper. Persons with glasses or contact lenses were allowed to wear these visual aids. They were positioned at 2.8 meters distance to the chart, and where instructed to read letters starting those in the first line. Participants able to read letters of the 8th line of the Snellen chart passed the visual acuity test. Colour vision was tested with pseudoisochromatic plates (see Figure B.6) printed on an A4 format sheet of paper. The tested persons were instructed to identify the numbers depicted on the plates. They could come as close to the sheet as they liked. Persons that could read all numbers from the plates passed the colour vision test. Next followed a short practice session (3) that introduced the participants to the task at hand. They were shown three video pairs selected by an expert to exemplify all three possible judgements, "A is better than B", "B is better than A" and "No preference". Practice videos were not counted as part of the evaluation. The actual evaluation process (4) followed. As mentioned above, an application developed for this purpose showed a single video of a pair (A,B) in full-screen. The users could freely switch between video A and video B by pressing the mouse button. After 20 seconds, a voting screen appeared, and the user was asked to make a judgement. After finishing the evaluation procedure, participants filled out a questionnaire (see Figure B.3, and B.4).

The evaluation procedure is illustrated in Figure 5.16). A welcome screen states the participant's task in a short statement, and calls the user to proceed. After confirmation, the first comparison set consisting of stimulus A and B is shown, with A initially displayed in full-screen. The user can freely switch between stimuli A and B by pressing the left mouse button. To assist the users in remembering which stimulus they prefer, the name of the currently shown stimulus is shown at the top of the screen. The name of stimulus A is shown on the left, and that of stimulus B is shown on the right side for easy distinction. After 20 seconds the screen turns dark, and the user is asked to vote for one of the three options: "A better than B", "B better than A", and "no preference". After voting, the next comparison set is shown.

The PC method calls for presenting both stimuli in both orders, that is AB and BA for a pair of stimuli (A,B) [ITU08]. We modify this approach in showing one model at a time in full-screen, while allowing the participants to switch between stimuli A and B as they choose. This is done for two reasons. First, evaluation of videos takes time. Showing each comparison set first in AB, and then in BA order further limits the number of content that can be shown in a single session. Second, and more importantly, in this way participants are able to compare the two models without visual effort.

Processing of the Study Results. After the trial, we first detect and remove outliers, and then compute an opinion score that allows us to compare the evaluated approaches.

When multiple stimuli only slightly diverge in their perceived quality, inconsistent ratings are likely to be made by the participants. Depending on the study design and shown material, small inconsistencies in ratings may be not avoidable. Ratings containing many

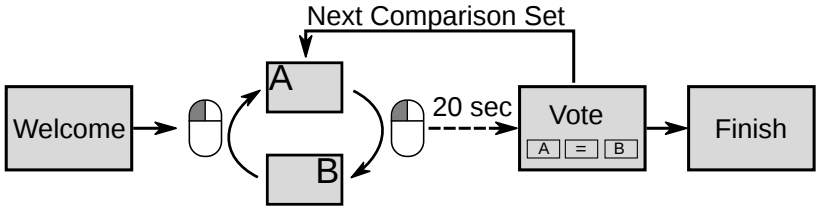


Figure 5.16: Illustration of the pair-based comparison scheme. The user is briefly introduced to the task on a welcome screen. Next, the first comparison set is shown, with one stimulus of a pair in full-screen. Users can freely switch between two stimuli by pressing the left mouse button. After 20 seconds, a gray screen appears calling to vote.

contradicting, or even random, judgements, however, are considered as outliers, and need to be excluded from consideration, so not to bias the evaluation results. We apply the algorithm of [LGE13] to detect and reject outliers. It detects inconsistent judgements by counting *circular triads*, that is inconsistent triples of stimuli preference. An example for a circular triad are three stimuli A, B and C together with the preferences $A > B$, $B > C$ and $C > A$, where the relation symbols $<$ and $>$ denote preference. A few inconsistent judgements per participant are allowed. If the number of circular triads rises beyond a threshold, however, the judgement is to be excluded from statistical processing. The ratio of inconsistent judgements to the total number of judgements is referred to as the *transitivity satisfaction rate* [LGE13].

For statistical processing, we first extract separate comparison sets for depth reconstruction and view-fusion approaches from the combined comparison sets obtained from the trials. To this end, we group judgements according to the respective type of compared approach. Following the approach of [NBSG14], we then transform the comparison sets into continuous quality scores using the Bradley-Terry model [BT52], by maximum likelihood estimation of a log-likelihood function. The converted quality scores and their standard deviation are then reported.

5.3 Summary

This chapter presented a description of the examined system and the applied evaluation methods. The system consists of three RGB stereo-camera based depth sensors termed $3DMU$ (3D measurement unit). Prior to model acquisition, the depth sensors are calibrated and their poses relative to the middle (reference) view are determined. Stereo-image pairs are acquired in a synchronised and distributed fashion on multiple computers. A stereo matching algorithm recovers point clouds from image pairs of individual depth sensors. In a semi-automatic approach, point clouds of individual views are segmented and cleaned from outliers. An APSS based surface reconstruction algorithm first creates refined point clouds for each view and finally fuses them into combined mesh models.

The plan for evaluating the regarded multi-view 3D reconstruction system in terms of

accuracy and model quality comprises two quantitative and one qualitative evaluation. The first approach uses acquisitions of specially crafted validation objects (sphere, box), to determine the geometric reproduction capabilities. Second, a novel view evaluation determines model accuracy by measuring image-based similarity on intermediate and final system products. Lastly, a pair-based user study conducted in the course of this work determines the quality of several model reconstruction methods in terms of subjective user opinions.

CHAPTER 6

Evaluation Results

This chapter discusses the evaluation results in detail. First, Section 6.1.1 presents the data set used for novel view evaluation and the user study. Next, Section 6.3 contains the results of the novel view evaluation. Finally, Section 6.4 details the results of the user study conducted for this work.

6.1 Data Set and Evaluated Approaches

6.1.1 Data Set

We have chosen five acquired models for evaluation (see Figure 6.1). Due to some acquisition limitations described in Section 5.1.2, only static models are considered in the current evaluation. Several intermediate and final products, summarised in Table 6.1, are examined. From the three available views, we generate combined models consisting of views $3DMU_1$ and $3DMU_3$, while ignoring for now the view generated from $3DMU_2$. Combined versions are created for smoothed points and coloured meshes and are denoted in the following as view $3DMU_{1+3}$.

Product	**Description**
Original points	Input point cloud for model generation
Smoothed points	Intermediate point cloud created in model generation
Coloured meshes	Created from final mesh output of model generation

Table 6.1: Products used for novel-view evaluation and user study. For a more detailed description see Section 5.1.4.

Figure 6.1: Models used for novel view evaluation and user study.

6.1.2 Evaluated Approaches

We divide the evaluation into two steps, point cloud generation, and view fusion. For the former we use three algorithms to generate point clouds, for the latter we define two strategies of view fusion. All evaluated approaches are summarised in Table 6.2. We ask two questions:

- How do different point cloud generation techniques influence the model quality?
- How does the model generation algorithm cope with different modes of view fusion?

Point Cloud Generation Approaches. We compare point cloud generation results obtained with three different stereo matching algorithms. In particular, the approaches are cost volume filtering (CVF) with "integer disparities" (ID), cost volume filtering with depth refinement (DR), and patch match (PM).

CVF with integer disparities (ID) [SNG+15] Disparity maps are computed with a cost volume filtering algorithm. It is set up to output integer valued disparity maps. Other algorithm parameters are listed in Table 5.2. The generation of a single disparity map at 2564x2056 pixel resolution takes 25.6 seconds to compute

a disparity map [1], and is the fastest of all considered methods. Integer valued disparity maps exhibit discrete spatial resolution. In the derived point clouds, the 3D points are grouped along few planes in Z axis direction. The model generation algorithm has to cope with relatively few cues with respect to available depth information.

CVF with depth refinement (DR) [SNG+15] Disparity maps are computed with the same algorithm as for the ID variant. Here, the disparity map is additionally refined using sub-pixel enhancement [YYDN07], and filtered with a weighted median filter [MHW+13]. Other algorithm parameters are listed in Table 5.2. Generation of a disparity map takes 71.37 seconds on average. DR disparity maps and derived point clouds are dense, and locally noisy. The model generation algorithm has to correctly derive smooth surfaces from many points.

PatchMatch (PM) [LZYZ18] In contrast to CVF, PatchMatch derives disparity values by estimating slanted planes per pixel and then minimising a global energy function. Models are generated with the parameters set for the Middlebury 2006 benchmark as described in [LZYZ18]. A notable difference to the paper is that we let the algorithm run for 30 iterations, instead of the originally suggested 5. Disparity map computation takes on average 25.8 minutes per view. PM disparity maps and corresponding point clouds are already very smooth. The model generation algorithm input data is optimal in terms of input noise.

View Fusion Approaches. We assess the quality of the model generation algorithm described in Section 5.1.4) by comparing two different methods of view fusion:

Fuse views after model generation (FA). Here, we first generate single view models, and afterwards fuse them. This method effectively performs a union operation on the regarded single views.

Fuse views before model generation (FB). In contrast to the previous approach, we fuse point clouds before model generation. Fused point clouds are treated as a single view by the model generation algorithm.

6.2 Results of Evaluation on Validation Objects

In this evaluation, we want to determine the accuracy of reconstructed objects. Geometrically simple-shaped validation objects created for this purpose are acquired by the system. Then, ideal 3D models of same geometric properties are fitted into the reconstructed object point clouds. We then obtain deviations of the acquired point clouds to the fitted

[1] A desktop PC with a 12 thread Intel i7-3930K CPU clocked at 3.2 GHz with 32 GB of RAM was used for determining the computation times. The active area of the input images had a resolution of approximately 2200×1800 pixels. All methods were using multi-threaded CPU implementations.

Abbr.	Description
ID	CVF [SNG+15] integer disparities. No subpixel refinement, no depth refinement
DR	CVF [SNG+15] floating point disparity. Subpixel refinement [YYDN07] and depth refinement [MHW+13]
PM	Patch Match [LZYZ18] floating point disparities
FB	View point clouds are fused before model generation
FA	View point clouds are fused after model generation

Table 6.2: Evaluated approaches for novel view evaluation and user study.

models that serve as ground truth measurements. The method has been described in Section 5.2.1.

In particular we pose the following questions:

1. How accurate is the reconstruction of spherical objects in terms of outliers?
2. Does accuracy depend on the camera to object distance?
3. How accurate is the plane reconstruction of a cuboid?

6.2.1 Data Set and Validation Objects

Validation Objects Two spheres and one box were created for this evaluation. Their surfaces were designed to allow a robust reconstruction. Specifically, a primarily blue sphere of 150 mm radius and a green sphere of 185 mm radius serve as validation object (see the `twinsphere` model in Figure 6.1). A cardboard box serves as cuboid object (e.g. Figure 6.5).

Data Set We use raw point clouds that are produced by the depth-reconstruction stage of the system (see Section 5.1.4). The point clouds were generated with the DR approach described in Section 6.1.2.

The sphere data set comprises 29 reconstructions of blue and green spheres. Out of these, 26 samples are used. 3 samples had to be dropped from the data set due to the sphere fitting algorithm not converging. Sphere point cloud samples ranged from 47500 to 217872 points, with an average point count of 108189. The distance of the spheres to the respective depth sensor ranged from 2.3 to 2.9 meters.

The data set for the evaluation of cuboid objects comprises in total 40 point clouds from 5 frames acquired by all 3 depth sensors. In each frame the box is turned slightly. The visible sides are first segmented individually from the point cloud. The number of points per side ranges from 931 to 22836 points with an average of 7646 points. Planes are fitted into each side by least squares plane fitting. We determine the angle between adjacent box sides using plane normal vectors.

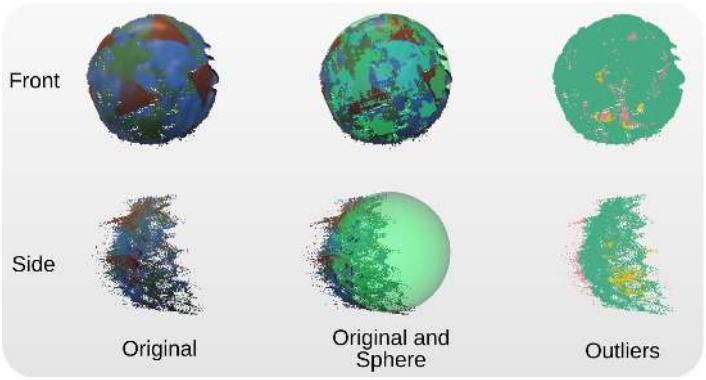


Figure 6.2: Qualitative sphere fitting results. Left: Point cloud of the blue sphere (frame 5, $3DMU_2$); Middle: An ideal sphere is fitted into the point data; Right: Outliers at an outlier threshold of 25mm. Outliers outside the sphere are coloured red, Outliers inside the sphere are coloured yellow; Inliers are marked green.

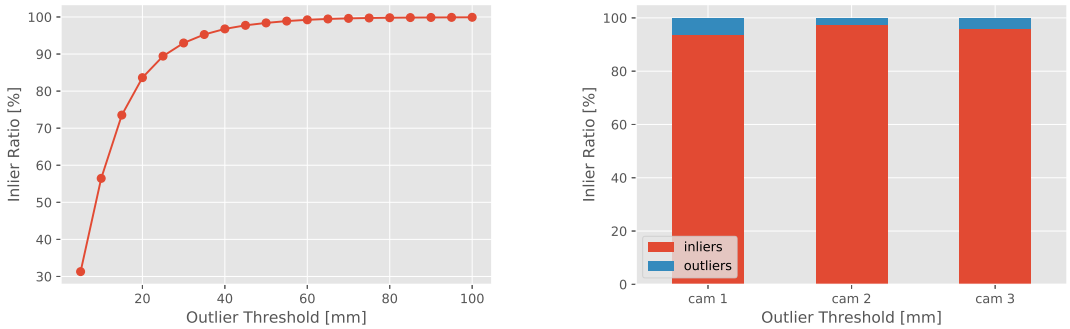
	3DMU$_1$	**3DMU$_2$**	**3DMU$_3$**
Inliers (%)	93.76	97.41	96.09
Outliers (%)	6.24	2.59	3.91

Table 6.3: Outliers by camera.

6.2.2 Results for Spherical Objects

We measure the distance of each reconstructed point to the fitted sphere center, and measure the signed error to the known radius. Points that deviate from the sphere radius by more than a threshold value τ in millimeter units are considered outliers. An example of a sphere fitted into the point cloud of the blue sphere in frame 5 acquired by depth sensor 2 is shown in Figure 6.2. The ideal sphere is shown overlaid with reconstructed points. Red points indicate outliers outside the sphere, yellow points are outliers inside of the sphere.

Accuracy with Respect to Outliers. The outlier ratio for τ in the range from 5 to 100 mm is shown in Figure 6.3a. At a threshold of 5 mm only 31.3 percent of the points are counted as inliers. The inlier ratio reaches almost 90 percent at a 25 mm outlier threshold, and reaches 99 percent at a 55 mm threshold. The depth sensor accuracy at an outlier threshold of $\tau = 25$ is shown in Figure 6.3b and Table 6.3. Here, $3DMU_2$ is the most accurate sensor with 97.41 percent, the least accurate is $3DMU_1$ with 93.76 percent and $3DMU_3$ has an inlier ratio of 96.06.



(a) Inlier ratio vs. outlier threshold

(b) Outlier statistics at 25 mm outlier threshold

Figure 6.3: Outlier statistics for spherical objects.

Accuracy and Object Distance. Figure 6.4 illustrates the error distribution of sampled point clouds in the upper part, and shows the sphere distance in the lower part. The Pearson correlation coefficient between camera to sphere distance and outlier ratio is 0.009, indicating that no correlation between sphere distance and outlier ratio can be observed in the present data.

6.2.3 Results for Cuboid Objects

Figure 6.5 shows an example of planes fitted onto the sides of the box. From two to three sides, denoted as front (f), top (t), left (l) or right (r), can be seen by a single depth sensor. Front and top sides are visible in all frames. The middle depth sensor $3DMU_2$ sees the left and right box sides in some samples, while the outer depth sensors $3DMU_1$ and $3DMU_3$ can only see either the left or the right side.

The median deviation from 90 deg between two adjacent box sides is reported. The results are shown in Table 6.4. Deviation values range from -3.44 to 9.47, with a median value of -0.06, and an average value of 0.80. Deviations lower than 1 deg are present for $3DMU_2$. The highest deviation of 9.47 between front ant top side of $3DMU_3$ is an outlier that results from an unfavourable point distribution due to this side being partially occluded by another object.

6.2.4 Discussion

Our first question is concerned with system accuracy. Outliers are a major obstacle for faithful model reconstruction. Even a small number can cause unnatural deformations in reconstructed models. The outlier ratio at a specific threshold can give insights on how small an object can be, in order to be reconstructed in a meaningful quality. From the results in Figure 6.3a we see that reconstructions reach a 99 percent inlier ratio at a threshold of 50 mm. This result is acceptable for large objects without fine surface

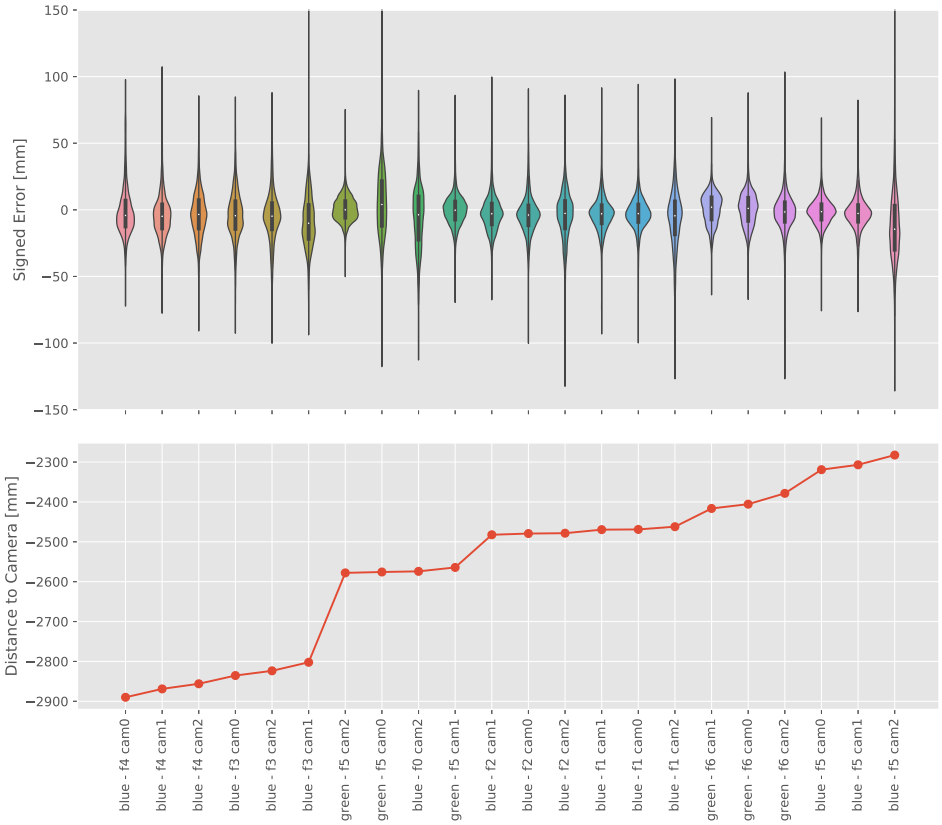


Figure 6.4: Sphere point distribution and distance. Top: Point distribution of sphere samples. A white dot at the bar center of each coloured area denotes the respective median signed error value, the thick black bar within the coloured area denotes upper and lower quartile. The coloured areas indicate the error distribution. Bottom: Distance of the sphere centre to the camera.

Depth Sensor	$\mathbf{\Delta_{ft}}$	$\mathbf{\Delta_{fl}}$	$\mathbf{\Delta_{tl}}$	$\mathbf{\Delta_{fr}}$	$\mathbf{\Delta_{tr}}$
$3DMU_1$	2.44	–	–	2.96	-3.44
$3DMU_2$	2.18	-0.06	-0.14	-0.72	-2.72
$3DMU_3$	9.47	-2.95	1.83	–	–

Table 6.4: Box reconstruction accuracy. Box sides are denoted as front (f), top (t), left (l) and right (r). Rows denote values per depth sensor. Columns 2-6 contain median deviations from 90 deg between two adjacent box sides.

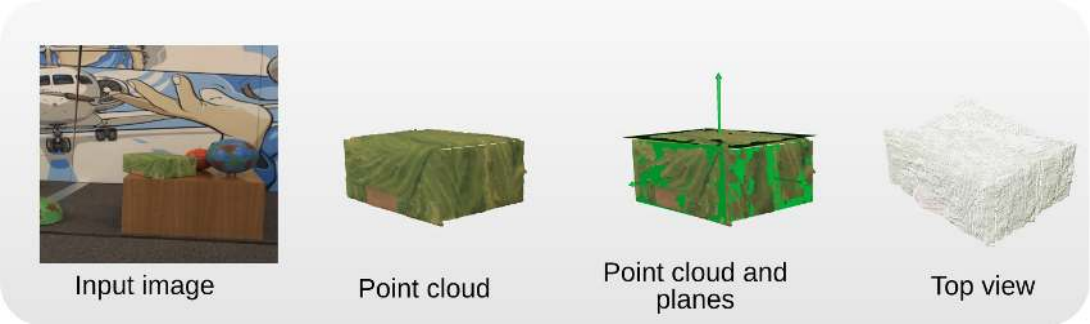


Figure 6.5: Illustration of plane fitting for a cuboid object. From left to right: The input image shows the box as placed in the scene; Reconstructed point cloud; Point cloud with overlaid fitted planes and plane normal vectors; Point cloud of the box shown from a different viewing angle.

details. To reconstruct smaller objects like a human face, a higher accuracy is needed. In Figure 5.13 in Section 5.1, an example of a person's face has been presented, where a chin was enlarged, whereas the nose was flattened. A possible solution to improve the reconstruction would be to increase the disparity range of the acquired disparity maps, which can be achieved by either increasing the depth sensor's stereo baseline or to decrease the object to camera distance. Regarding the accuracy of the depth sensors results of the sphere, our evaluation indicates that $3DMU_1$ is the worst performing.

The second question of this evaluation is on the influence of the object-to-camera distance on the reconstruction accuracy. In the present data, spheres where placed at distances from approximately 2.3 to 2.9 meters. No correlation at this distance range could be shown.

The evaluation on a cuboid validation object has shown that orthogonal angles of a box can be reconstructed to a high accuracy of up to 0.06 deg deviation from orthogonality. Other measurements, on the other hand, show significantly lower accuracy, for which two reasons can be pointed out. First, outliers affect the plane fitting process based on least squares regression. The second reason has to do with the acquired data set. In some samples only a low number of plane points represent a side, as they were frequently obstructed by other objects.

6.3 Results of Novel View Evaluation

In this section, novel view evaluation results are presented and discussed. The general procedure has been laid out in Section 5.2.2. We first discuss the creation and preparation of the data set in Section 6.3.1. Next, results in terms of accuracy for the individual depth sensors are presented in Section 6.3.2. Then, we look into the results for novel view accuracy in Section 6.3.3. Finally, the results are discussed in more depth in Section 6.3.4.

6.3.1 Data Set

For novel view evaluation, we use the 3D models and approaches variations described in Section 5.2.2. Five different 3D models, and in total six approaches of model generation are examined, which are divided into three point cloud generation, and two view fusion methods. Point cloud generation methods are denoted as *ID* for point clouds computed from CVF integer-valued disparity maps, *DR* for point clouds generated from CVF disparity maps with sub-pixel values and disparity filtering, and *PM* for point clouds computed from a PatchMatch algorithm. The examined fusion methods are view fusion before model generation (*FB*) and view fusion after model generation (*FA*). In the following, we examine various system products resulting from different processing stages. In particular, `pnts-raw` denotes original per view point clouds that serve as input to model generation. `pnts-fa`/`pnts-fb` are smoothed point clouds, created by model generation, and `mesh-fa`/`mesh-fb` are coloured meshes created by model generation. Variants with the `fb` suffix are only available for the novel view $3DMU_{1+3}$, as they are generated from the two outward views.

6.3.2 Results of Depth Sensors

Completeness Results. Completeness describes the ratio of image pixels subject to similarity computation to the total number of image pixels. Figure 6.6 shows our model completeness. The average is 5.7 percent in terms of the total input image size. Three factors contribute to this rather low value. First, only a fraction of the original input image size is actually used. Image rectification reduces the active image region for $3DMU_1$, $3DMU_2$ and $3DMU_3$ to 80.6, 74.1, and 68.8 percent of the original image area. Second, during acquisition the sensors were positioned such that a large person was able to jump, while still staying within the field of view of each camera. Third, as we examine segmented models, only regions belonging to the persons are counted. Each selected model can be fit into a bounding box of 1300×1300 pixels size.

Depth Sensor Accuracy. We analyse the individual depth sensor accuracies by comparing single view products against the original, rectified, image as seen by the left camera of the respective sensor. The graphs in Figure 6.7 show the mean model accuracy of examined reconstruction products for each of the three depth sensors. Mean accuracy values for each view can be seen in Table 6.5. Figure 6.8 shows qualitative accuracy results.

The accuracy of the middle sensor $3DMU_2$ is the rated highest with an average score of 0.985. Accuracy ratings among models are largely consistent, as the low deviations from the mean value show. The second best sensor $3DMU_1$ has a slightly lower average of 0.97 and exhibits higher accuracy deviations. Lastly, $3DMU_3$ shows the lowest accuracy with a mean score of 0.92. A reason for the lower performance of $3DMU_3$ can be seen in the third column of Figure 6.8. Dark regions at the left edge of the persons indicate a global offset of the view of $3DMU_3$, caused by the rigid registration step.

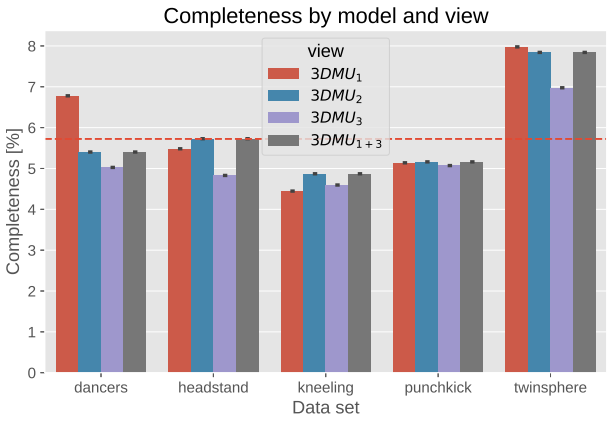


Figure 6.6: Data set completeness results with respect to the total image area grouped by model. Bars denoted by $3DMU_1$, $3DMU_2$, and $3DMU_3$ correspond to model completeness as seen by the respective depth sensor. $3DMU_{1+3}$ denotes the novel view. The horizontal line indicates the average model completeness.

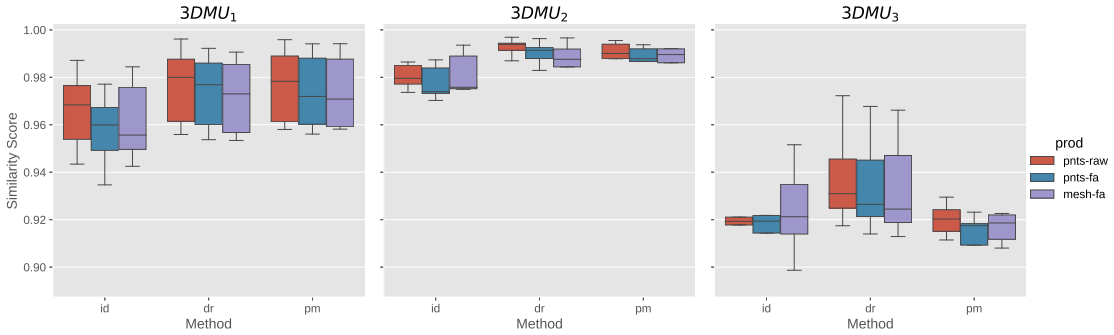


Figure 6.7: Accuracy results for units $3DMU_1$, $3DMU_2$, and $3DMU_3$ on original point clouds (pnts-raw), smoothed point clouds fused before model generation (pnts-fa) and smoothed points fused after model generation (pnts-fb).

6.3.3 Results for Evaluated Approaches

Next, we compare the performance of the different model generation approaches in terms of accuracy. Here, models fused from views $3DMU_1$ and $3DMU_3$ and the rectified original image captured by $3DMU_2$ are compared. The fused models are denoted by $3DMU_{1+3}$. Figure 6.9 shows the accuracy results. Mean accuracies of the novel view $3DMU_{1+3}$ are shown in Table 6.5. Figure 6.10 shows qualitative accuracy results for the `dancers` model.

Point Cloud Generation Approaches. Out of the three compared point cloud generation approaches ID shows the smallest accuracy values. While PM achieves the highest rating, it is comparable with DR. ID models are computed from integer-valued

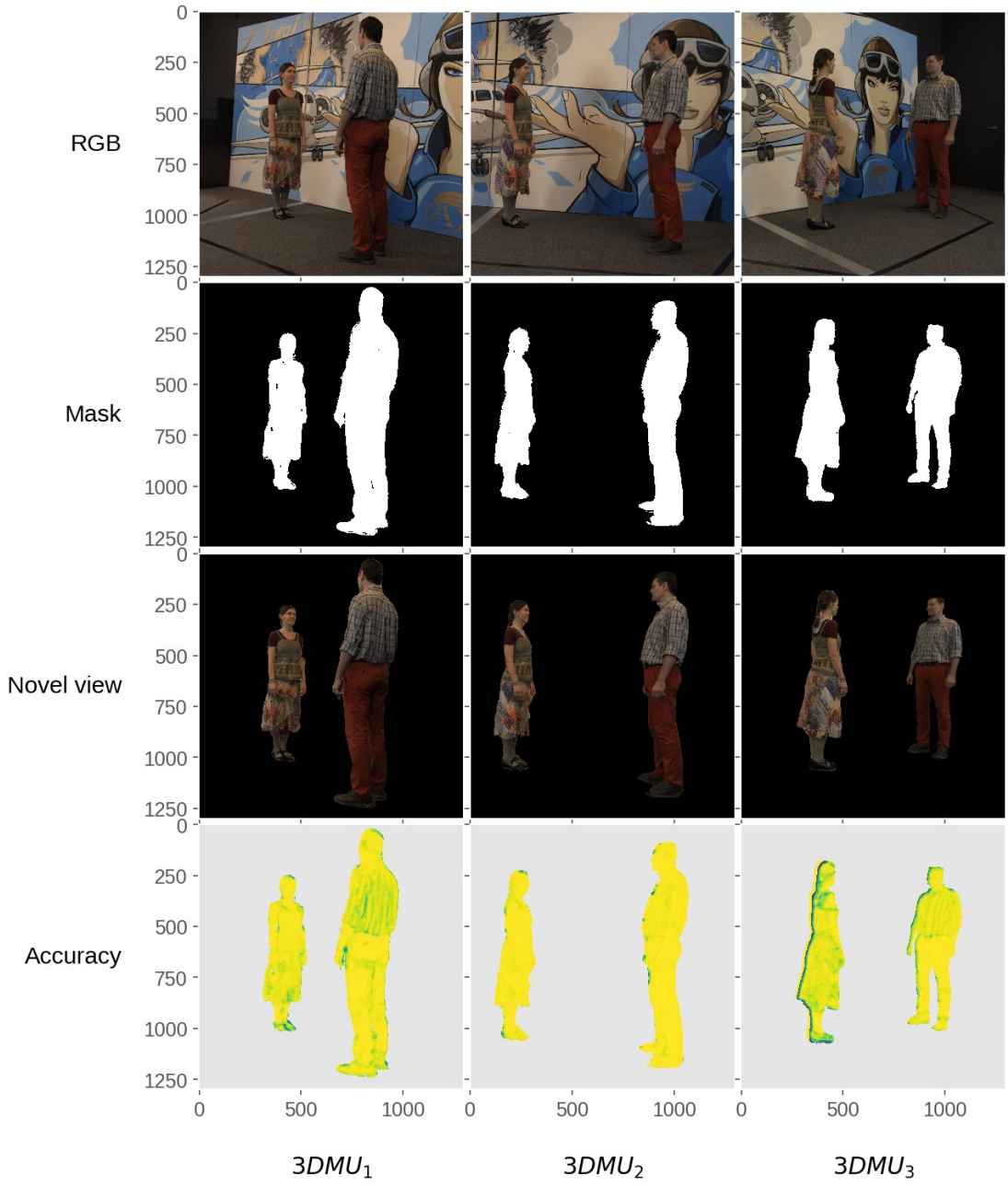


Figure 6.8: Qualitative accuracy results for the `dancers` model. Image regions are cropped to fit the model. Rows from top to bottom: Original image; Mask of evaluated image area; Novel view image generated from point clouds as captured from the sensor's viewpoint; The accuracy map visualises dissimilar regions in false colour. Dark colours denote non-similar model regions. Columns show views corresponding to the respective depth sensor.

Appr.	Product	View			
		$3DMU_1$	$3DMU_2$	$3DMU_3$	$3DMU_{1+3}$
ID	mesh-fa	0.962	0.982	0.924	0.934
	mesh-fb	—	—	—	0.918
	pnts-fa	0.958	0.978	0.920	0.928
	pnts-fb	—	—	—	0.917
DR	mesh-fa	0.972	0.989	0.934	0.945
	mesh-fb	—	—	—	0.936
	pnts-fa	0.974	0.990	0.935	0.945
	pnts-fb	—	—	—	0.937
PM	mesh-fa	0.974	0.986	0.917	0.947
	mesh-fb	—	—	—	0.936
	pnts-fa	0.974	0.986	0.912	0.944
	pnts-fb	—	—	—	0.932

Table 6.5: Mean accuracy scores of compared approaches by product and view.

disparity maps, however, they perform still reasonably well against models that were generated from far more elaborate stereo matching algorithms. A shortcoming of the ID method are holes in the models. Examples of such holes can be seen in Figure 6.10 at the legs of the person on the right. There, holes are indicated by dark circle shaped spots. PM models are similar to DR models despite the PatchMatch-based algorithm producing very smooth input for model generation. However, there are artefacts present in these models as a result of the PatchMatch algorithm not being able to further the optimise respective regions. DR models suffer less from holes or artefacts compared to the other two methods. Thus, they provide better results than ID and are comparable with PM in terms of quality while taking approximately 71 seconds for disparity map computation of a single view. Compared to PM's 25.6 minutes per disparity map, DR is approximately 25 times faster.

View Fusion Approaches. FA can produce models of higher accuracy than FB, as seen in Figure 6.9. This result holds for both smoothed point clouds and for coloured meshes. The results are consistent with those of the conducted user study, where FA was found to be the better method in terms of perceived quality (see Figure 6.11).

6.3.4 Discussion

As can be seen in Figure 6.7, single view comparison identified that $3DMU_3$ has comparatively lower accuracy than the other depth sensors. Its impact on accuracy is higher that of any examined model generation approach. This result is in line with our qualitative results (e.g. Figure 5.14) that suggest $3DMU_3$ supplies the data with the most noticeable reconstruction artefacts.

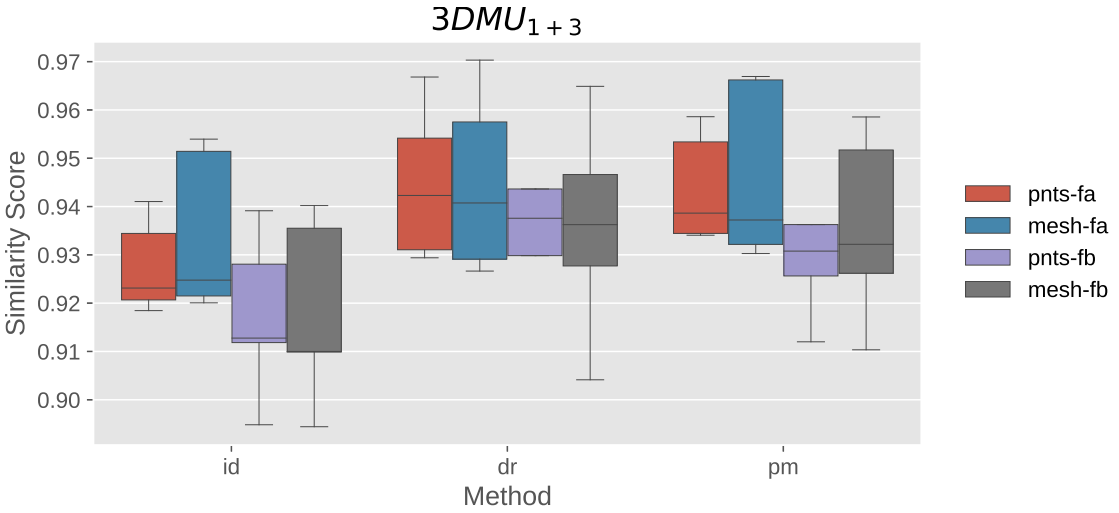


Figure 6.9: Similarity results for novel view $3DMU_{1+3}$.

The investigated point cloud generation methods can be ranked with respect to the novel view evaluation from worst to best as ID, DR and PM. The accuracy differences between the examined methods, however, were found to be of lesser impact than the degradation in terms of accuracy of $3DMU_3$.

It has to be noted that for this evaluation input image masks are used. Another option, not included here, would be to use as mask valid pixels of the created novel view images.

The data presented herein exhibits large error margins. For this work, only five models have been available. Examination of a larger data set of models could give a more stable basis for accuracy analysis.

6.4 User Study Results

This section discusses the results of our user study that is conducted as preliminary work to an up-following main study outside the scope of this diploma thesis. The outcome of this study will be used as input to the design of the main study as part of an ongoing research project. It aims to answer the following questions:

1. How do the test subjects rate the study design?

2. How do the depth reconstruction methods perform in terms of model shape and colour quality?

3. How does the model generation perform in terms of shape and colour quality for different view fusion approaches?

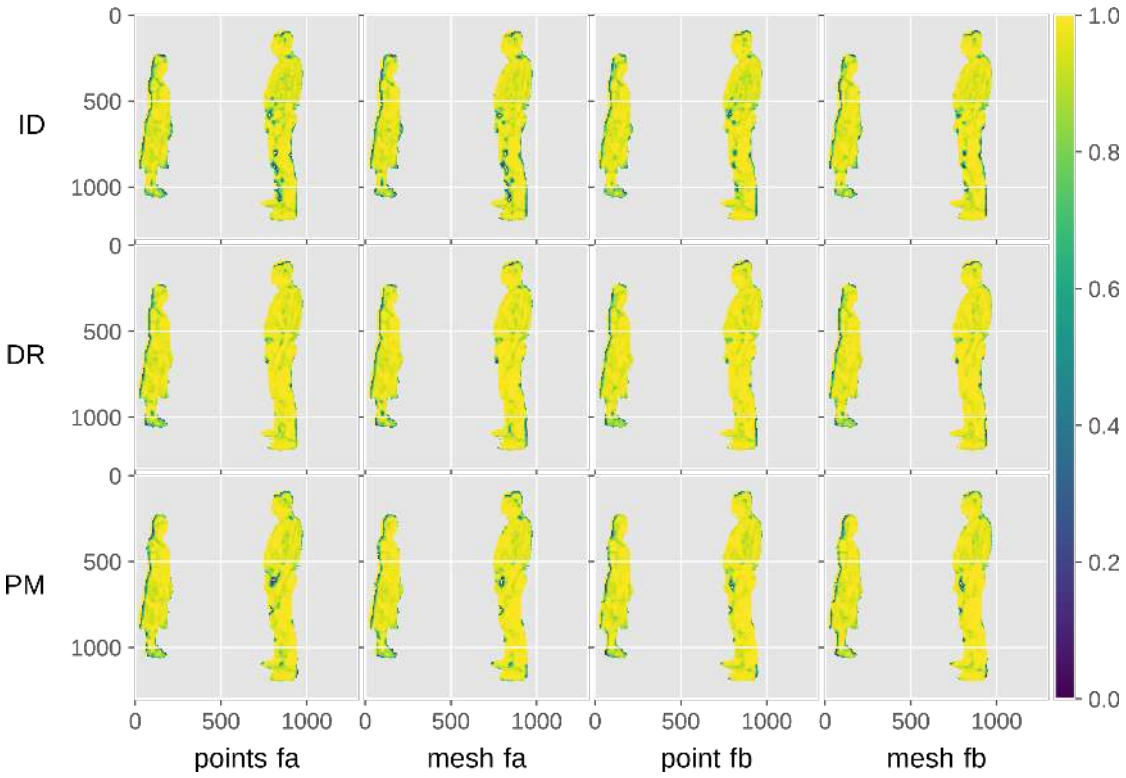


Figure 6.10: Qualitative accuracy results for compared approaches. Image regions have been cropped to fit the model. Rows from top to bottom: Comparison of input image and point clouds computed from integer disparity maps (ID), refined disparity maps (DR), and PatchMatch disparity maps (PM). Columns from left to right: Accuracy maps for smoothed point clouds, coloured meshes for fusion after mesh generation (FA) method, smoothed point clouds and coloured meshes for fusion before mesh generation (FB) method.

This section is structured as follows. Section 6.4.1 describes characteristics of the participants and testing material. Section 6.4.2 presents results for the compared approaches. Section 6.4.3 closes with a discussion.

6.4.1 Study Design

Participants. In total ten persons (six male, four female) participated in the subjective evaluation. Table B.1 contains supplementary information on the participants. Five persons were considered experts in evaluating image material or 3D model generation, five others were counted as naïve viewers. The participants' age ranged from 26 to 58 years with an average of 39 years. While the number of participants is low for a full-fledged user study, it is appropriate for the purpose as preliminary study according to [ITU08]. A major goal of this subjective evaluation is to assess the testing procedure adequacy. All

persons passed the screening procedure comprising a visual acuity and colour vision test, and could perform the subjective evaluation. Additionally, the participants' opinions were screened by subsequent statistical processing for biased or incoherent judgements with the algorithm for determining the transitivity satisfaction rate described in Section 5.2.3. No circular triads were detected, consequently all judgements could be used for statistical analysis.

Testing Material. In total six different model generation approaches divided into three methods of point cloud generation and two view fusion methods where compared. Five acquired 3D models were selected for evaluation. The evaluated approaches and used models have been discussed in Section 6.1. For each model and approach we rendered 20 second long videos showing the model as seen from along a camera path that passed four predefined positions. Three of positions corresponded to those of $3DMU_1$, $3DMU_2$ and $3DMU_3$. An additional novel fourth viewpoint was inserted slightly above the middle 3DMU. In pair comparison any approach is compared against all others. Thus, a single comparison set consists of $\frac{6\times5}{2} = 15$ comparisons. Each comparison set is shown for all 5 models, resulting in $5 \times 15 = 75$ comparisons. The total show time was $75 \times 20 = 1500$ seconds or 25 minutes per person. Participants deemed the evaluation time as challenging, but appropriate.

6.4.2 Compared Approaches

Mean opinion scores were obtained as described in Section 5.2.3 and are shown in Figure 6.11, grouped by model.

Point Cloud Generation Approaches. The subjective evaluation results in a clear preference for DR, with ID second, and PM last (see Figure 6.11a). Ratings are consistent among the majority of models, namely "dancers", "headstand" and "kneeling". Models generated with DR exhibit comparatively round surface features, especially in faces. Participants' comments indicate this fact as the major factor for their preference. ID models have more holes than the other methods. Further, faces tend to be rather flat, due to low spatial resolution of the point clouds that serve as input for model generation. While PM models have fewer holes and rounder surface features compared to the other methods, their quality was perceived as inferior to the others. A reason for this is the large artefacts that appear in them. Although observers were instructed to dismiss them when performing their judgement, feedback comments indicate that they still had a significant impact on the given opinions.

View Fusion Approaches. Results of the view fusion evaluation can be seen in Figure 6.11b. Overall, the fusion of single views after model generation (FA) was clearly preferred by observers over view fusion before model generation (FB). In FA view fusion, no colour interpolation among individual views is performed resulting in visible ridges where views meet, as can be seen in Figure 6.12. On the other hand, colour interpolation

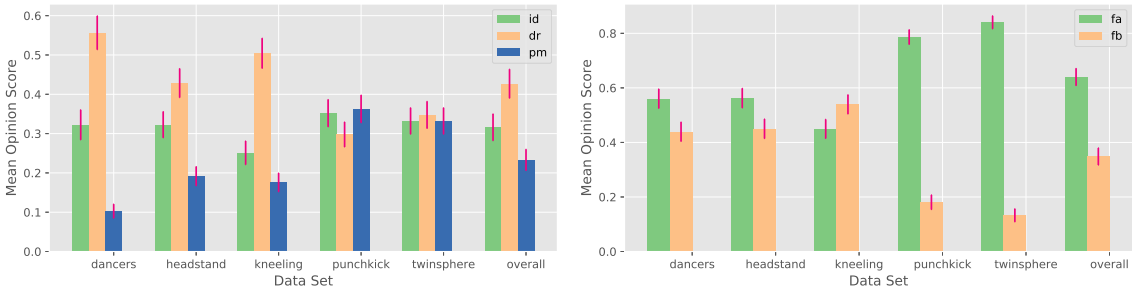


Figure 6.11: Subjective evaluation results. (a) Results for depth generation methods; (b) Results for view fusion methods.

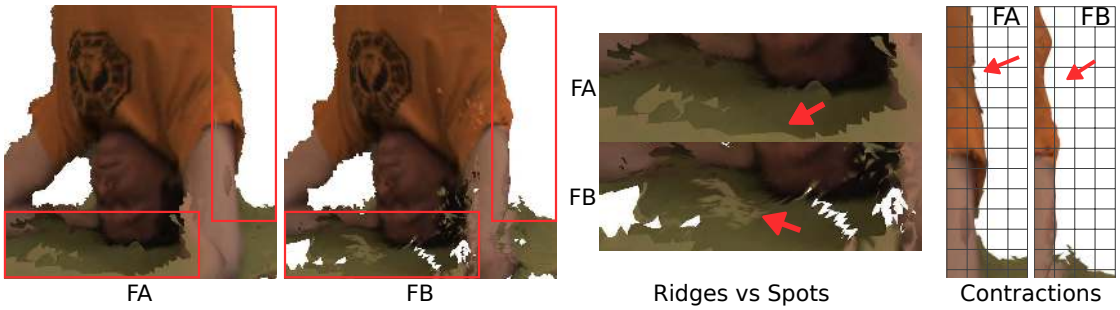


Figure 6.12: Comparison of view fusion results.

in FB leads to spotty areas on view borders. One participant noted that the model colour representation appears "very smoothed", referring to the low model texture resolution. Further, FB models appear contracted compared to FA and contain fewer geometric ridges (see Figure 6.12), as all individual partial views are smoothed in the model generation phase.

Perceived Quality of Different Viewpoints. In addition to the comparison task, study participants were asked to rate the general model appearance from one of four predefined view points. The results are shown in Figure 6.13. Models were only composed of views of $3DMU_1$ and $3DMU_3$, and these positions where rated best. The viewpoint of $3DMU_2$, that was not part of the model, was rated third, but not conclusively with respect to $3DMU_1$ and $3DMU_3$. The newly introduced novel viewpoint "top" was decisively ranked last. It has to be noted that three out of ten participants were not able to answer the question regarding view positions. Further, the high error margin contained in the voting data suggests that participants were overwhelmed with this rating task.

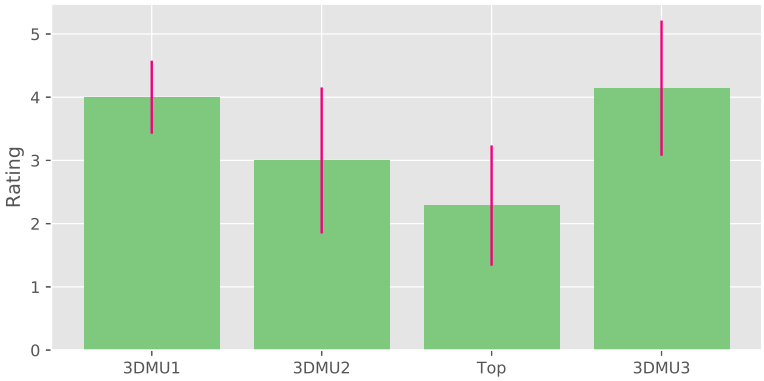


Figure 6.13: Subjective ratings of view point positions.

6.4.3 Discussion

Regarding the questions asked in this study, the following conclusions can be drawn. Overall, the study design was perceived as suitable by the observers. 25 minutes of model show time was adequate, but should not be increased further. Voting was performed on a separate dark screen showing only buttons for the judgement according to the recommendation of [ITU08]. Multiple participants reported difficulty of identifying videos A and B, as they could not see the models on the voting screen any more. The situation could be improved by showing both models while the viewer votes. Showing models from different perspectives was appreciated. The position rating task, however, was too much to ask. Three observers noted that the last position was the most dominant in their memory. This suggests, that the viewpoint order of the camera path can bias the position-rating task.

All participants noted geometric issues on models, especially the faces were of concern. Fusion of the models from two far apart views leads to faces being duplicated or outright missing. Single view meshes of ID models are to "flat" to achieve an accurate reconstruction of features, further they have too many holes due to locally too sparse point data for the model generation. Both the DR and PM methods can reproduce round features, like heads, to a better degree, but ultimately also struggle in accurate reconstruction. PM generated models exhibited large free-floating artefacts that are especially visible when showing models from view points other than the capturing sensor.

In the FB method, persons faces tend to disappear in the meshes due to smoothing applied in the model generation phase and the relatively low overlap between the fused views. With the available data, no model was convincingly reproduced. Geometric bodies had noticeable shape distortions that can be traced to outlier points present in the original point clouds, and aggressive smoothing applied for model generation. Simple colour interpolation can reproduce texture from RGB images only with limited quality. Other available methods, such as [EFR+17, DCC+18], need be to used to recover model texture more accurately.

CHAPTER 7

Conclusion

In this thesis, we have presented an evaluation framework for assessing the quality of point clouds and textured meshes acquired by a 3D reconstruction system that comprises multiple stereo sensors. Our evaluation determines two important properties of reconstructed 3D models: (i) the geometric fidelity of reconstructed objects, and (ii) the subjective model quality perceived by the user.

After providing some background information and a literature review on 3D reconstruction and related evaluation methodologies, we presented our proposed evaluation framework along with selected aspects of its implementation. An important part of our study demonstrates the application of our implemented evaluation modules on a 3D reconstruction system that obtains textured mesh models of dynamic scenes with three stereo cameras. One of the goals of our evaluation was to compare the quality achieved by different stereo matching and 3D model merging algorithms included in the examined multi-view stereo system. To this end, a set of intermediate and final system products were examined using three complementary quantitative and qualitative evaluation strategies. First, an evaluation with geometrically simple validation objects (sphere, cuboid) of known dimensions determined the geometric reconstruction quality by comparing object reconstructions against ideal models by means of shape fitting. The results of this evaluation were used to determine the geometric accuracy of point clouds obtained with the system in terms of reconstructed shape, size and angles between planar surfaces. The results show that the examined 3D imaging system is capable of reconstructing spherical point clouds with 90 percent accuracy at an outlier threshold of 25 mm and angles between adjacent faces of cuboid objects with a deviation of as low as 0.06 degree. Second, a quantitative image-based novel view evaluation assessed the quality of the acquired point clouds and textured mesh models in terms of image similarity. More precisely, we found that the examined CVF and PatchMatch stereo correspondence algorithms delivered comparable results in terms of textured mesh model

accuracy, when their output was used to generate mesh models, while CVF was faster than PatchMatch by a factor of 25.

Third, a subjective user study determined the perceived quality of textured mesh models of several depth reconstruction approaches. We also observed that the subjective rankings were inconsistent with those determined by our novel view evaluation, which confirms the need for a combined approach of quantitative and qualitative evaluation. In this context we noticed some model imperfections localized in small areas close to object boundaries that influenced the image-based similarity results only by a small amount, while the same imperfections dominated the subjective appearance, when shown in a 3D view of the scene.

The present evaluation methodology assumed reasonably accurate stereo sensor calibration. A possible topic for future work would be to include the effects of non-perfect sensor calibration into the analysis of our 3D reconstruction system. Furthermore, the incorporation of quality aspects that are specific to dynamic scenes (for example, to measure flickering artefacts) would be a valuable future extension.

List of Figures

1.1 3D model reconstruction processing pipeline. 2
1.2 Illustration of the employed novel view evaluation method 3

2.1 The pinhole camera model 7
2.2 Frontal projection model 8
2.3 Typical types of lens distortion 9
2.4 Epipolar geometry of a stereo camera 12
2.5 Illustration of various calibration objects 12
2.6 Camera calibration with planar patterns 13

3.1 Data acquisition methods and their results 16
3.2 Stereo matching 20
3.3 Outline of the basic steps of a typical local stereo matching processing pipeline 21
3.4 Common scene representations in 3D reconstruction 22
3.5 View fusion with non-rigid alignment 24

4.1 Taxonomy of evaluation methods 26
4.2 Illustration of the third eye technique 28
4.3 Virtual rephotography evaluation 28

5.1 Overview of the processing pipeline 35
5.2 Image of a 3D Measurement Unit (3DMU) 36
5.3 Illustration of the physical setup and scene distance 37
5.4 Physical system setup 37
5.5 Effects of slightly unsynchronised image acquisition 39
5.6 Illustration of motion blur in a scene containing fast movement 40
5.7 Illustration of camera calibration, registration and image rectification process 42
5.8 Illustration of transformations involved in intra-3DMU calibration and inter-3DMU registration 42
5.9 Illustration of the depth reconstruction process 44
5.10 Example of point cloud registration 49
5.11 Illustration of the semi-automatic post-correction application 50
5.12 Illustration of mesh generation 51
5.13 Detailed views of the `dancers1` data set 53

5.14 Illustration of model artefacts introduced by views 53
5.15 Illustration of the lab environment for subjective evaluation 57
5.16 Illustration of the pair-based comparison scheme 59

6.1 Models used for novel view evaluation and user study 62
6.2 Qualitative sphere fitting results 65
6.3 Outlier statistics for spherical objects 66
6.4 Sphere point distribution and distance 67
6.5 Illustration of plane fitting for a cuboid object 68
6.6 Data set completeness results 70
6.7 Accuracy results for units $3DMU_1$, $3DMU_2$, and $3DMU_3$ 70
6.8 Qualitative accuracy results for the `dancers` model 71
6.9 Similarity results for novel view $3DMU_{1+3}$ 73
6.10 Qualitative accuracy results for compared approaches 74
6.11 Subjective evaluation results 76
6.12 Comparison of view fusion results 76
6.13 Subjective ratings of view point positions 77

B.1 User study instructions in English language 100
B.2 User study instructions in German language 101
B.3 Questionnaire page 1 102
B.4 Questionnaire page 2 103
B.5 Snellen chart 104
B.6 Pseudoisochromatic plates 105

List of Tables

4.1 Viewing conditions for subjective assessment 30

5.1 System hardware characteristics 36
5.2 Stereo matching parameters 45

6.1 Products used for novel-view evaluation and the user study 61
6.2 Evaluated approaches for novel view evaluation and user study 64
6.3 Outliers by camera 65
6.4 Box reconstruction accuracy 67
6.5 Mean accuracy scores of compared approaches by product and view . . . 72

A.1 Ground truth measurements taken of cameras, objects and scene features 98

B.1 Detailed information on participants of the user study. 106

Acronyms

2D Two-dimensional 6

3D Three-dimensional 1

3DMU 3D Measurement Unit 34, 36

ACR Absolute Category Rating 31

APSS Algebraic Point Set Surfaces 49

CVF Cost Volume Filtering 62, 79

DSQS Double Stimulus Continuous Quality-Scale 31

FPS Frames Per Second 38

ITU International Telecommunication Union 29

NCC Normalized Cross Correlation 18, 56

PC Pair-based Comparison 31

RGB Red, Green, Blue 16

RGB-D Red, Green, Blue plus Depth 16

SAD Sum of Absolute Differences 17

TOF Time-Of-Flight 16

WTA Winner-Takes-All 21, 45

Bibliography

[AUE17] Evangelos Alexiou, Evgeniy Upenik, and Touradj Ebrahimi. Towards subjective quality assessment of point cloud imaging in augmented reality. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–6, 2017.

[BÁBB+18] Miguel Barreda-Ángeles, Federica Battisti, Giulia Boato, Marco Carli, Emil Dumic, Margrit Gelautz, Chaminda Hewage, Dragan Kukolj, Patrick Le-Callet, Antonio Liotta, Cecilia Pasquini, Alexandre Pereda-Baños, Christos Politis, Dragana Sandic, Murat Tekalp, María Torres-Vega, and Vladimir Zlokolica. Quality of experience and quality of service metrics for 3D content. In P. Assunção and A. Gotchev, editors, *3D Visual Content Creation, Coding and Delivery*, pages 267–297. Springer, Cham, 2018.

[BB13] Michael Bleyer and Christian Breiteneder. *Stereo Matching - State-of-the-Art and Research Challenges*, pages 143–179. Springer London, 2013.

[BKB08] Gary Bradski, Adrian Kaehler, and Gary Bradski. *Learning OpenCV - Computer Vision with the OpenCV Library.* O'Reilly Media, 2008.

[BKH10] Abdelkrim Belhaoua, Sophie Kohler, and Ernest Hirsch. Error evaluation in a stereovision-based 3D reconstruction system. *EURASIP Journal on Image and Video Processing*, 2010(1):Article ID 539836, 12 pages, 2010.

[BR15] Libor Bolecek and Vaclav Ricny. Influence of stereoscopic camera system alignment error on the accuracy of 3D reconstruction. *Radioengineering*, 24(2):610–620, 2015.

[Bra00] G. Bradski. The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 2000.

[Bro66] Duane C. Brown. Decentering distortion of lenses. *Photometric Engineering*, 32(3):444–462, 1966.

[Bro16] Nicole Brosch. *Spatio-temporal Video Analysis for Semi-automatic 2D-to-3D Conversion by.* PhD thesis, Vienna University of Technology, 2016.

[BRR11] Michael Bleyer, Christoph Rhemann, and Carsten Rother. PatchMatch stereo - stereo matching with slanted support windows. In *Proceedings of the British Machine Vision Conference*, pages 14.1–14.11, 2011.

[BT52] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324, 1952.

[BT99] Stan Birchfield and Carlo Tomasi. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293, 1999.

[BTV06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417, 2006.

[CCS+15] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 34(4):69:1–69:13, 2015.

[CGEB07] Massimiliano Corsini, Elisa D. Gelasca, Touradj Ebrahimi, and Mauro Barni. Watermarked 3-D mesh quality assessment. *IEEE Transactions on Multimedia*, 9(2):247–255, 2007.

[CGK14] Hsiang-Jen Chien, Haokun Geng, and Reinhard Klette. Improved visual odometry based on transitivity error in disparity space: A third-eye approach. In *International Conference on Image and Vision Computing New Zealand*, pages 72–77, 2014.

[ÇığlaAA12] Cevahir Çığla and A. Aydın Alatan. An improved stereo matching algorithm with ground plane and temporal smoothness constraints. In *European Conference on Computer Vision. Workshops and Demonstrations*, pages 134–147, 2012.

[ÇığlaAA13] Cevahir Çığla and A. Aydın Alatan. Information permeability for stereo matching. *Signal Processing: Image Communication*, 28(9):1072–1088, 2013.

[CL96] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *ACM Conference on Computer graphics and interactive techniques*, pages 303–312, 1996.

[CTF12] Ivan Cabezas, Maria Trujillo, and Margaret Florian. An evaluation methodology for stereo correspondence algorithms. In *International Conference on Computer Vision Theory and Applications*, pages 154–163, 2012.

[Dah] Jeff Dahl. Typical snellen chart to estimate visual acuity. https://commons.wikimedia.org/wiki/File:Snellen_chart.svg. Accessed: 2018-07-31.

[DCC+18] Ruofei Du, Ming Chuang, Wayne Chang, Hugues Hoppe, and Amitabh Varshney. Montage4D: Interactive seamless fusion of multiview video textures. In *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 1–11, 2018.

[DKK09] Ankur Datta, Jun-Sik Kim, and Takeo Kanade. Accurate camera calibration using iterative refinement of control points. In *International Conference on Computer Vision Workshops*, pages 1201–1208, 2009.

[DTK+16] Mingsong Dou, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, et al. Fusion4D: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics*, 35(4):114:1–114:13, 2016.

[DZC+18] Alexandros Doumanoglou, Nikolaos Zioulis, Emmanouil Christakis, Dimitrios Zarpalas, and Petros Daras. Subjective quality assessment of textured human full-body 3D-reconstructions. In *International Conference on Quality of Multimedia Experience*, pages 1–6, 2018.

[EA14] Evan-Amos. Xbox One Kinect Sensor. https://en.wikipedia.org/wiki/File:Microsoft-Xbox-One-Console-Set-wKinect.jpg, 2014. Accessed: 2018-07-31.

[EFR+17] Thomas Ebner, Ingo Feldmann, Sylvain Renault, Oliver Schreer, and Peter Eisert. Multi-view reconstruction of dynamic real-world objects and their integration in augmented and virtual reality applications. *Journal of the Society for Information Display*, 25(3):151–157, 2017.

[EXR] EXR file format. Retrieved from https://www.openexr.com/. Accessed: 2018-05-13.

[FBC+18] Karel Fliegel, Federica Battisti, Marco Carli, Margrit Gelautz, Lukáš Krasula, Patrick Le Callet, and Vladimir Zlokolica. 3D Visual Content Datasets. In P. Assunção and A. Gotchev, editors, *3D Visual Content Creation, Coding and Delivery*, pages 299–325. Springer, Cham, 2018.

[FH06] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient belief propagation for early vision. *International Journal of Computer Vision*, 70(1):41–54, 2006.

[FH15] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

[GG07] Gaël Guennebaud and Markus Gross. Algebraic point set surfaces. *ACM Transactions on Graphics*, 26(3):23, 2007.

[GLU12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[GVC+16] Jinjiang Guo, Vincent Vidal, Irene Cheng, Anup Basu, Atilla Baskurt, and Guillaume Lavoue. Subjective and objective visual quality assessment of textured 3D meshes. *ACM Transactions on Applied Perception*, 14(2):1–20, 2016.

[HBG13] Asmaa Hosni, Michael Bleyer, and Margrit Gelautz. Secrets of adaptive support weight techniques for local stereo matching. *Computer Vision and Image Understanding*, 117(6):620–632, 2013.

[HLCH12] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Horaud. *Time-of-Flight Cameras: Principles, Methods and Applications*. Springer, 2012.

[HM12] Xiaoyan Hu and Philippos Mordohai. A quantitative evaluation of confidence measures for stereo vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2121–2133, 2012.

[HRB+13] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2013.

[HS07] Heiko Hirschmüller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[HZ04] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[Ish] Pseudoisochomatic plate color vision test. `http://www.colorvisiontesting.com/ishihara`. Accessed: 2018-07-31.

[ITU08] ITU. Recommendation ITU-R P.910: Subjective video quality assessment methods for multimedia applications, 2008.

[ITU12] ITU. Recommendation ITU-R BT.500: Methodology for the subjective assessment of the quality of television pictures, 2012.

[KH13] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics*, 32(3):1–13, 2013.

[KHSM17] Andreas Kuhn, Heiko Hirschmüller, Daniel Scharstein, and Helmut Mayer. A TV prior for high-quality scalable multi-view stereo reconstruction. *International Journal of Computer Vision*, 124(1):2–17, 2017.

[KIT] The KITTI Vision Benchmark Suite. `http://www.cvlibs.net/datasets/kitti/`. Accessed: 2018-04-06.

[Kle14] Reinhard Klette. *Concise Computer Vision*. Springer London, 2014.

[LC87] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *Computer graphics and interactive techniques*, pages 163–169, 1987.

[LGE13] Jong Seok Lee, Lutz Goldmann, and Touradj Ebrahimi. Paired comparison-based subjective quality assessment of stereoscopic images. *Multimedia Tools and Applications*, 67(1):31–48, 2013.

[LHKP13] Bo Li, Lionel Heng, Kevin Koser, and Marc Pollefeys. A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern. In *IEEE International Conference on Intelligent Robots and Systems*, pages 1301–1307, 2013.

[LNSW16] Minglei Li, Liangliang Nan, Neil Smith, and Peter Wonka. Reconstructing building mass models from UAV images. *Computers and Graphics*, 54:84–93, 2016.

[Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[LZYZ18] Lincheng Li, Shunli Zhang, Xin Yu, and Li Zhang. PMSC: PatchMatch-based superpixel cut for accurate stereo matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(3):679–692, 2018.

[MAM14] Nicolas Mellado, Dror Aiger, and Niloy J. Mitra. Super 4PCS fast global pointcloud registration via smart indexing. *Computer Graphics Forum*, 33(5):205–215, 2014.

[MHW+13] Ziyang Ma, Kaiming He, Yichen Wei, Jian Sun, and Enhua Wu. Constant time weighted median filtering for stereo matching and beyond. In *IEEE International Conference on Computer Vision*, pages 49–56, 2013.

[Mida] Middlebury Multi-View Stereo - Online Evaluation. `http://vision.middlebury.edu/mview/eval/`. Accessed: 2018-04-18.

[Midb] Middlebury Stereo - Online Evaluation. `http://vision.middlebury.edu/stereo/`. Accessed: 2018-04-18.

[MK09] Sandino Morales and Reinhard Klette. A third eye for performance evaluation in stereo sequence analysis. In *International Conference on Computer Analysis of Images and Patterns*, pages 1078–1086, 2009.

[Mon07] David Monniaux. Leica terrestrial lidar scanner. `https://commons.wikimedia.org/wiki/File:Lidar_P1270901.jpg`, 2007. Accessed: 2018-07-31.

[NBG13] Matej Nezveda, Nicole Brosch, and Margrit Gelautz. Hyperion 3D - intelligent workflow design for low-cost 3D film production working package 4. Technical report, Vienna University of Technology, 2013.

[NBSG14] Matej Nezveda, Nicole Brosch, Florian Seitner, and Margrit Gelautz. Depth map post-processing for depth-image-based rendering: a user study. In *SPIE 9011, Stereoscopic Displays and Applications XXV*, page 90110K, 2014.

[Nez14] Matej Nezveda. Evaluation of depth map post-processing techniques for novel view generation. Master's thesis, Vienna University of Technology, 2014.

[Nym17] Bengt Nyman. Digital single-lens reflex camera - Nikon D810. `https://commons.wikimedia.org/wiki/File:Nikon_D810_EM1B6357-2.jpg`, 2017. Accessed: 2018-07-31.

[OEDT+16] Sergio Orts-Escolano, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Christoph Rhemann, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, Shahram Izadi, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, and Sameh Khamis. Holoportation: Vitual 3D teleportation in real-time. In *Symposium on User Interface Software and Technology*, pages 741–754, 2016.

[PLY] PLY-polygon file format. Retrieved from `http://paulbourke.net/dataformats/ply/`. Accessed: 2018-05-13.

[SAB+07] Elena Stoykova, A. Aydin Alatan, Philip Benzie, Nikos Grammalidis, Souris Malassiotis, Joern Ostermann, Sergej Piekh, Ventseslav Sainov, Christian Theobalt, Thangavel Thevar, and Xenophon Zabulis. 3-D time-varying scene capture technologies - a survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11):1568–1586, 2007.

[SBMM15] Aaron N Staranowicz, Garrett R Brown, Fabio Morbidi, and Gian-Luca Mariottini. Practical and accurate calibration of RGB-D cameras using spheres. *Computer Vision and Image Understanding*, 137:102–114, 2015.

[SCD+06] S.M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 519–528, 2006.

[SCK15] Bok Suk Shin, Diego Caudillo, and Reinhard Klette. Evaluation of two stereo matchers on long real-world video sequences. *Pattern Recognition*, 48(4):1109–1120, 2015.

[SCSK13] Verónica Suaste, Diego Caudillo, Bok-Suk Shin, and Reinhard Klette. Third-eye stereo analysis evaluation enhanced by data measures. In *Mexican Conference on Pattern Recognition*, pages 74–83, 2013.

[Sha98] C.M. Shakarji. Least-squares fitting algorithms of the NIST algorithm testing system. *Journal of Research of the National Institute of Standards and Technology*, 103(6):633, 1998.

[SHK+14] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, volume 8753, pages 31–42, 2014.

[SLK15] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. Kinect range sensing: Structured-light versus Time-of-Flight Kinect. *Computer Vision and Image Understanding*, 139:1–20, 2015.

[SMP05] Tomáš Svoboda, Daniel Martinec, and Tomáš Pajdla. A convenient multi-camera self-calibration for virtual environments. *Presence: Teleoperators and Virtual Environments*, 14(4):407–422, 2005.

[SNG+15] Florian Seitner, Matej Nezveda, Margrit Gelautz, Georg Braun, Christian Kapeller, Werner Zellinger, and Bernhard Moser. Trifocal system for high-quality inter-camera mapping and virtual view synthesis. In *International Conference on 3D Imaging*, pages 1–8, 2015.

[SS02] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[SS03] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 195–202, 2003.

[SSG+17] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2538–2547, 2017.

[SSP07] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics*, 26(3):80, 2007.

[Sze99] Richard Szeliski. Prediction error as a quality metric for motion and stereo. In *IEEE International Conference on Computer Vision*, pages 781–788, 1999.

[TWC15] Fakhri Torkhani, Kai Wang, and Jean-Marc Chassery. Perceptual quality assessment of 3D dynamic meshes: Subjective and objective studies. *Signal Processing: Image Communication*, 31:185–204, 2015.

[VCB15] Camilo Vargas, Ivan Cabezas, and John W Branch. Stereo correspondence evaluation methods: A systematic review. In *Advances in Visual Computing*, pages 102–111, 2015.

[VSKL17] K. Vanhoey, B. Sauvage, P. Kraemer, and G. Lavoué. Visual quality assessment of 3D models: On the influence of light-material interaction. *ACM Transactions on Applied Perception*, 15(1), 2017.

[VV14] Patrick Vandewalle and Chris Varekamp. Disparity map quality for image-based rendering based on multiple metrics. In *International Conference on 3D Imaging*, pages 1–5, 2014.

[WBF+17] Michael Waechter, Mate Beljan, Simon Fuhrmann, Nils Moehrle, Johannes Kopf, and Michael Goesele. Virtual rephotography. *ACM Transactions on Graphics*, 36(1):1–11, 2017.

[WFR+16] Shenlong Wang, Sean Ryan Fanello, Christoph Rhemann, Shahram Izadi, and Pushmeet Kohli. The Global Patch Collider. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 127–135, 2016.

[WKZ+16] Katja Wolff, Changil Kim, Henning Zimmer, Christopher Schroers, Mario Botsch, Olga Sorkine-Hornung, and Alexander Sorkine-Hornung. Point cloud noise and outlier removal for image-based 3D reconstruction. In *International Conference on 3D Vision*, pages 118–127, 2016.

[XIM] XIMEA GmbH. Ximea MC050CG-SY product specification brochure. `https://www.ximea.com/files/brochures/xiC-USB3_1-Sony-CMOS-Pregius-cameras-brochure-HQ.pdf`. Accessed: 2018-02-18.

[YGX+17] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. BodyFusion: Real-time capture of human motion and surface geometry using a single depth camera. *IEEE International Conference on Computer Vision*, pages 910–919, 2017.

[YYDN07] Qingxiong Yang, Ruigang Yang, James Davis, and David Nister. Spatial-depth super resolution for range images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[Zah85] Johann Zahn. *Oculus artificialis teledioptricus sive Telescopium, ex abditis rerum naturalium & artificialium principiis protractum nova methodo, eaque solida explicatum ac comprimis e triplici fundamento physico seu naturali, mathematico dioptrico et mechanico, seu practico stabilitum.* No publisher, 1685.

[ZFM+17] Kang Zhang, Yuqiang Fang, Dongbo Min, Lifeng Sun, Shiqiang Yang, and Shuicheng Yan. Cross-scale cost aggregation for stereo matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(5):965–976, 2017.

[Zha00] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[Zha04] Zhengyou Zhang. Camera calibration with one-dimensional objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):892–899, 2004.

[ZSG+18] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3D reconstruction with RGB-D cameras. *Computer Graphics Forum*, 37(2):625–652, 2018.

[ZTDVAL14] P. J. Zarco-Tejada, R. Diaz-Varela, V. Angileri, and P. Loudjani. Tree height quantification using very high resolution imagery acquired from an unmanned aerial vehicle (UAV) and automatic 3D photo-reconstruction methods. *European Journal of Agronomy*, 55:89–99, 2014.

[ZW94] Ramin Zabih and John Woodfill. Non-parametric local transforms for computing visual correspondence. In *European Conference on Computer Vision*, pages 151–158, 1994.

Appendix A - System Ground Truth Measurements

Table A.1 lists ground truth measurements that where taken of cameras, prominent scene features, and objects.

Id	Measurement	Value [mm]
1	Background height	2500
2	Background width (all 3 segments)	4980
3	Background segment width	1660
4	Black floor mark front	2030
5	Black floor mark left	1640
6	Black floor mark right	1695
7	Silver stripe floor	1715
8	Sphere green diameter	397
9	Sphere blue diameter	300
10	Brown table width	745
11	Brown table height	450
12	Brown table depth	750
13	Box green width	367
14	Box green height	157
15	Box green depth	285
16	3DMU Baseline	700
17	3DMU1 to 3DMU2; left camera to left camera	2308
18	3DMU2 to 3DMU3; left camera to left camera	2940
19	3DMU1 (left camera) to background	4050
20	3DMU2 (left camera) to background	4100
21	3DMU3 (left camera) to background	4200
22	Box brown width	570
23	Box brown height	238
24	Box brown depth	205
25	Plane green width	500
26	Plane green height	700

Table A.1: Ground truth measurements taken of cameras, objects and scene features. Measurements have been acquired by using a tape measure.

Appendix B - User Study

User Instructions

All test subjects where given written instructions introducing them to their task. User instructions where created in English (see Figure B.1) and German (see Figure B.2) language.

User Questionnaire

All test subjects had to answer a questionnaire after performing the trial. It comprises two pages. The top section of page 1 (see Figure B.3) contains trial information to be filled out by the test operator. It identifies the test number, date, start and stop time of the trial, as well as results of visual acuity and colour vision test. The lower section contains general information about the test subject. Page 2 (see Figure B.4) contains questions about subjective impressions and the test procedure.

User Screening

All participants of the user study where screened for visual acuity using a Snellen chart, and for colour vision using pseudoisochromatic (Ishihara) plates. In the following, the tests are described in more detail.

Visual acuity was tested using a Snellen chart (see Figure B.5) printed on an A4 format sheet of paper. Participants with glasses or contact lenses where allowed to wear these visual aids while undergoing the test. Test subjects where positioned 2.8 meters away from the chart. Both eyes where tested individually. Subjects would gently cover the untested eye. They were instructed to read numbers on the chart that were indicated by the trial operator. Subjects with normal visual acuity are able to read letters up to line 8.

Colour vision was tested with pseudoisochromatic plates depicted in Figure B.6. The plates where printed on an A4 format sheet of paper. The subjects were instructed to read the numbers. They could step as close to the sheet as they saw fit. Subjects with normal colour vision are able to read all numbers on the plate.

Subjective Evaluation „Precise3D“

Welcome to the research group „*Interactive Media Systems*“ at the Vienna University of Technology. Thank you for participating in this subjective evaluation that is part of my master thesis. Here, we try to investigate different approaches to 3D model reconstruction. The results are very important to us, so we are asking you for your full attention during the evaluation.

What do I have to do?

Start by carefully reading these instructions. They explain the whole procedure. We start by assessing your visual acuity, and your ability to perceive colour vision. Then, you will perform the evaluation. Afterwards, we will ask you to fill out an anonymized questionnaire about general information on yourself and the experiences you have made.

How does the evaluation look like?

You will see pairs of videos (video A and video B) showing 3D models on a computer screen. Five different models are used here. The models where generated with different model generation methods, and may differ only slightly. You can switch between video A and video B by pressing the left mouse button. Use it to make an opinion which video looks better to you. After 20 seconds the video will stop, and you are asked to judge, by selecting one of three options:

A is better than B No preference B is better than A

Please ignore *free floating artefacts* and pay special attention to *shape and colour of* the *models* when making your judgement.

Further, take note of how the models generally look from the view points where the camera stops for a short time. You will be asked rate them after the evaluation.

We start by practising the task with test videos that will not be counted as results. Once you know your task, the real evaluation will start. After half of the video pairs you will be given the opportunity to have a short break. Use it as you see fit.

Once you have finished, we ask you to fill out an anonymized questionnaire. It will contain questions on your person, and the experiences you have made.

Thank you for your participation!

Figure B.1: User study instructions in English language.

Subjektive Studie „Precise3D“

Willkommen bei der Forschungsgruppe „*Interactive Media Systems*“ der Technischen Universität Wien. Danke, dass sie an dieser Studie, die ich im Zuge meiner Masterarbeit durchführe, teilnehmen. Wir untersuchen verschiedene Ansätze der 3D Modellerstellung. Die Ergebnisse sind wichtig für unsere Arbeit. Deswegen bitten wir sie um ihre volle Aufmerksamkeit während der Studie.

Was muss ich tun?

Beginnen sie mit dem sorgfältigen Lesen dieser Anweisungen. Sie erklären den gesamten Vorgang. Wir beginnen mit dem Feststellen ihrer Sehschärfe, und ihrem Vermögen Farben wahrzunehmen. Danach werden sie die Bewertung durchführen. Am Ende, bitten wir sie einen anonymisierten Fragebogen auszufüllen, der Fragen zu ihrer Person, und zu den Erfahrungen, die sie hier gemacht haben, beinhaltet.

Wie läuft die Studie ab?

Sie werden Videopaare (Video A und Video B) mit 3D Modellen auf einem Computerbildschirm sehen. Wir verwenden fünf verschiedene Modelle. Diese wurden mit verschiedenen Methoden erstellt, und können sich nur ein wenig voneinander unterscheiden. Sie können per Mausklick zwischen Video A und Video B wechseln. Nutzen sie die Maus, um sich ein Urteil zu bilden, welches ihnen besser gefällt. Nach 20 Sekunden endet die Wiedergabe eines Videopaars, und sie werden gefragt, dieses zu beurteilen indem sie aus drei Möglichkeiten auswählen:

A is besser als B | Keine Präferenz | B ist besser als A

Bitte ignorieren sie *frei stehende Artefakte* und achten sie insbesondere auf die *Form und Farbe der Modelle* wenn sie ihr Urteil fällen.

Achten sie weiters auf das allgemeine Aussehen der Modelle aus Blickrichtungen, an denen die Kamera kurz anhält. Sie werden nach der Evaluierung danach gefragt

Wir beginnen, indem wir ihnen Beispielvideos zeigen, die nicht in das Ergebnis einfließen. Wenn sie sich mit der Aufgabe vertraut gemacht haben, startet die Bewertung. Nach der Hälfte der Videopaare erhalten sie die Gelegenheit zu einer Pause. Nutzen sie diese, wie es ihnen beliebt.

Sobald sie fertig sind, bitten wir sie einen anonymisierten Fragebogen auszufüllen. Er enthält Fragen zu ihrer Person, und den Erfahrungen, die sie gemacht haben.

Vielen Dank für ihre Teilnahme!

Figure B.2: User study instructions in German language.

Trial Number: _____

Snellen: 20 /___

Colour: ___ / 8

Date: _____

Trial Start: ____:____

Trial End: ____:____

- **What is your age**: _____
- **What is you sex**:

  ☐ male ☐ female ☐ other

- **Do you wear glasses or contact lenses?**

  ☐ yes ☐ no

- **What is your highest educational degree?**

  ☐ mid-school ☐ undergraduate ☐ graduate ☐ doctoral

- **What is your occupational status?**

  ☐ student ☐ employed ☐ unemployed ☐ retired

- **Do you have experience in photo editing or image processing (Photoshop, Python, ...)?**

  ☐ yes ☐ no

Figure B.3: Page 1 of the questionnaire to be filled out by the test subjects.

- **You saw the video stop at four positions. How do you rate your overall impression of the model quality at the respective position? 1 .. lowest, 5 ..highest**

Position 1 (right)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Position 2 (middle)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Position 3 (top)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
Position 4 (left)	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

- **Did you notice any unpleasent effects, apart from free floating artefacts?**

☐ yes ☐ no

If „yes“, which exactly?

_____

_____

_____

- **What do you think of the trial setup?**

_____

_____

_____

- **Do you have any additional comments?**

_____

_____

_____

Figure B.4: Page 2 of the questionnaire to be filled out by the test subjects.

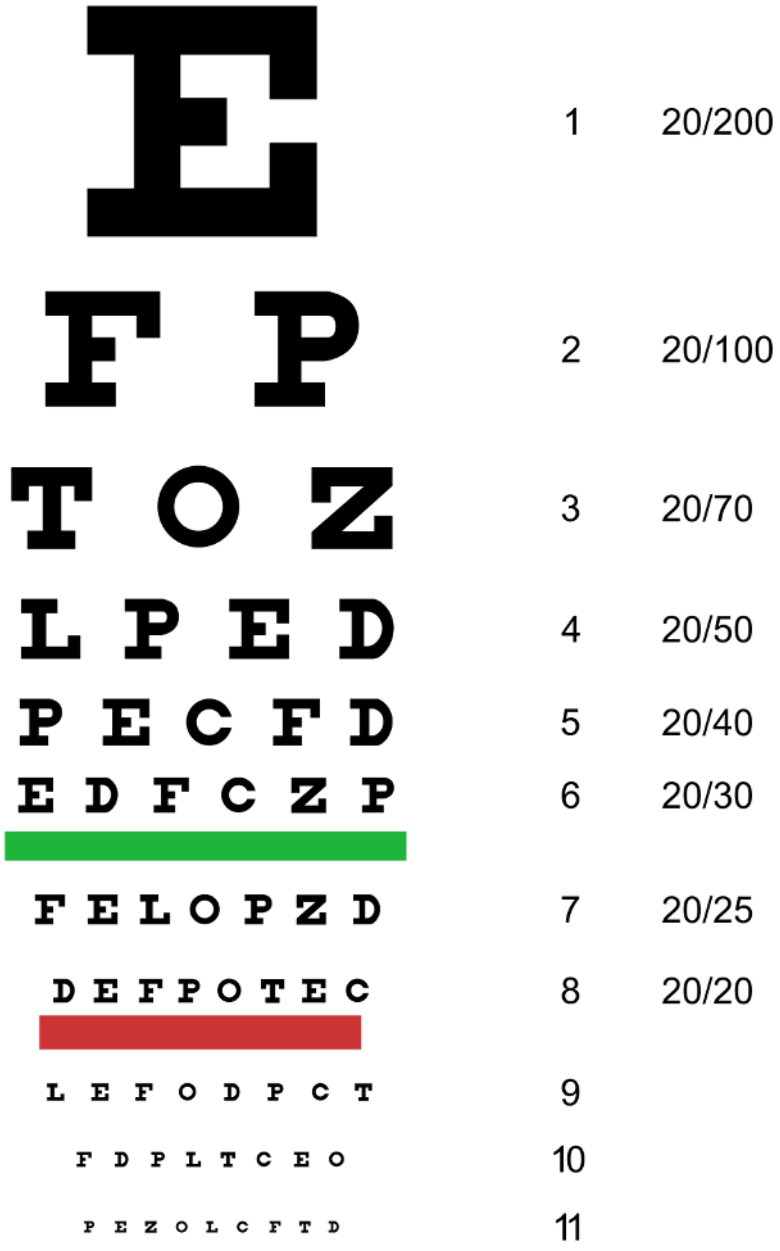
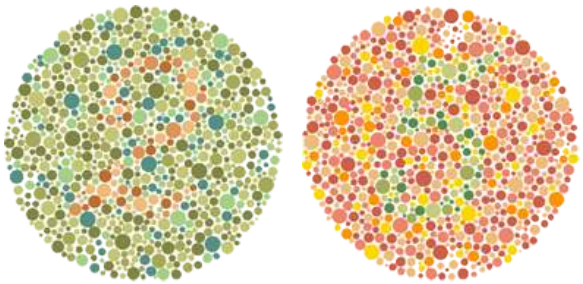
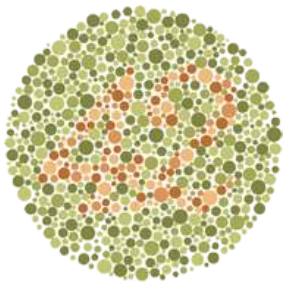


Figure B.5: Snellen chart. Figure taken from [Dah].

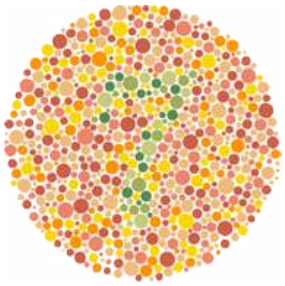


(a) 2

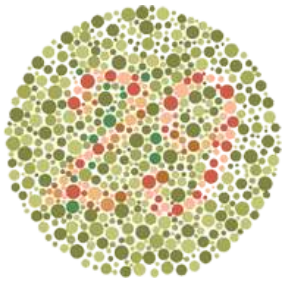
(b) 5



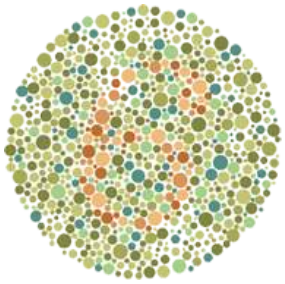
(c) 42



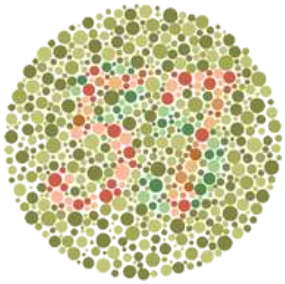
(d) 7



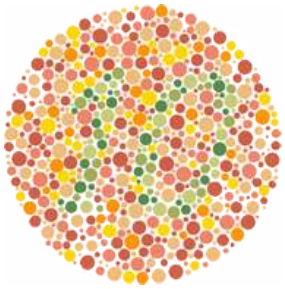
(e) 29



(f) 6



(g) 57



(h) 10

Figure B.6: Pseudoisochromatic plates used to determine participants' colour vision. Figures taken from [Ish].

Detailed User Information

ID	Age	Sex	Education	Occupation	Experienced	Optical Aids	Snellen	PiP
P1	36	m	Undergraduate	Employed	yes	yes	20/20	8/8
P2	34	f	Graduate	Employed	no	no	20/20	8/8
P3	38	m	Graduate	Student	yes	yes	20/20	8/8
P4	26	m	Undergraduate	Employed	yes	no	20/20	8/8
P5	58	m	Graduate	Employed	no	no	10th	8/8
P6	46	m	Doctoral	Employed	yes	yes	20/20	8/8
P7	36	f	Graduate	Employed	yes	yes	20/20	8/8
P8	32	m	Graduate	Employed	no	no	20/20	8/8
P9	30	f	Graduate	Employed	no	no	20/20	8/8
P10	58	f	Graduate	Retired	no	yes	20/20	8/8

Table B.1: Detailed information on participants of the user study.