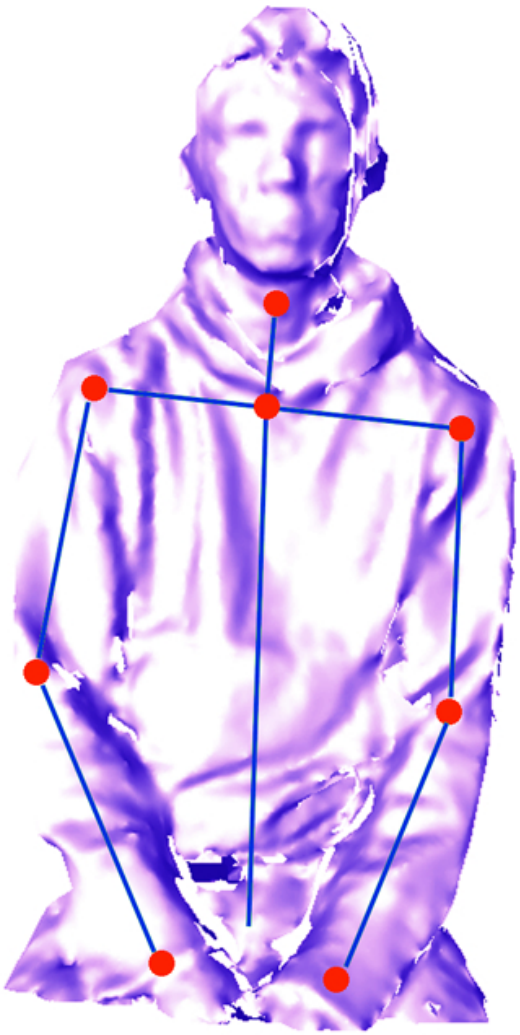
Copenhagen May 2011

Markerless motion capture for biomechanical applications

Master Thesis



By Martin Sandau Christiansen

E-mail: msc.22@hotmail.com

University of Copenhagen
Faculty Of Health Sciences
Department of Neuroscience and Pharmacology
Division of Biomechanics
Blegdamsvej 3B,
DK-2200 Copenhagen N
Denmark
TEL +45 35327900

Preface

This master thesis presents an approach to obtain a full 3D model of a human, using a photogrammetric approach with the same precision as a full body laser scanner. It also describes a pose estimation approach to markerless motion capture. The approach is based on 3D registration to track points in the reconstructed mesh and segmentation of the mesh using the tracked points to divide the reconstructed models into rigid segments. The purpose for the pose estimation approach is to develop a 3D motion capture system that is capable of meeting the criterions for biomechanical applications, without synthesizing subject specific articulated models.

The thesis consists of two parts: Part I concerning 3D modeling and Part II concerning pose estimation.

Part I introduces the basics of stereo vision and how to acquire 3D information using multiple camera views. It furthermore brings the reader deeper into applied 3D modeling algorithms. Tests based on the theory are performed to examine PhotoModeler® Scanner's abilities to create reconstructions of a full body.

Part II presents an approach for pose estimation. It begins with a review of previous work of limb segmentation and joint center localization, followed by introducing the biomechanical aspects in gait analysis. The aspects of developing a model free motion capture system is presented with a suggestion of how to perform pose estimation without using articulated models. Tests of the pose estimation approach will provide insight the processing time and the functionality.

The master thesis corresponds to 30 ECTS and is supervised by Associate Professor Erik Simonsen at Division of Biomechanics, University of Copenhagen and Associate Professor Henrik Aanæs at Informatics, DTU.

Acknowledgements

I would like to thank DBI for the economic aid to materials and software used in the project. Also a special thanks to Peter K. Larsen, Academic officer at the University of Copenhagen, for his supervision regarding the PhotoModeler® Scanner software and to Rasmus R. Paulsen, Associate Professor at DTU, for supervision regarding 3D modeling.

I would like to thank Tron A. Darvann, Research Engineer at University of Copenhagen for assistance to gain 3D images for the project.

I would also like to thank Rasmus Larsen, professor at DTU and Thomas H. Mosbech Ph. d. student at DTU for their permission to modify their source codes and m-files for this project.

At last I would thank Kim Amhild, R&D Engineer at CLAAS Agrosystems, Anders N. Christensen and Louise M. Jeppesen, M.Sc.Eng students at DTU for supervision and grammar correction.

Nomenclature

Accuracy is a measure of the closeness of an estimate to the true value. Bias is related to accuracy.

Baseline. A direct line between the focal points of two cameras.

Disparity map is a gray scale representation of depth perception provided by stereovision. See section 1.4.1.2 for more details.

DoG (Difference of Gaussian) is a function of kernel provided by subtractions of two Gaussian functions.

DoF (Degrees of Freedom) is the set of independent variables in an equation.

Epipole. The point where the base line crosses an image plane is called an epipole.

Epipolar line. The line where the epipolar plane intersects the image plane.

Epipolar plane. A plane defined by the focal points of two cameras and a given point in 3D space.

Essential matrix. A matrix describing the epipolar geometry in terms of camera coordinates, see section 1.4.1.1.

Focal length is the distance between the focal point and the image plane, often measured in terms of mm or cm, see section 3.1.2 for further details.

Focal point, also called the camera center. It defines the point 'O' in which all arbitrary point sources and the corresponding points in the image plane will intersect.

Fps (frames per second) is a unit for temporal resolution in a video sequence.

Frame is an image in a video sequence.

Fundamental matrix. A matrix describing the epipolar geometry in terms of pixel coordinates, see 1.4.1.1.

LoG (Laplacian of Gaussian). LoG is equal to the Laplacian of a Gaussian function. The Laplacian is the second derivative of a function on Euclidian space denoted by ∇^2.

Mocap is an acronym for motion capture. Motion capture is the practice of tracking objects and translating the movements into a model or measureable data.

Object plane is an imaginary plane in which the measured target(s) is located and which is used to describe the geometry of camera configuration. The term object plane that refers to a two dimensional space is quite misleading, since an object often has three dimensions. However the object depth is often negligible in relation to the distance from the camera to the object.

Optical center equals the focal point in the pinhole model. In the lens model the optical center equals the center of the lens where incident rays are passing through without bending, see Figure 3.7 in appendix.

Patch A patch can be defined in either a 3D space or an image. In 3D space it defines a point with an associated normal calculated by using the nearby points. In an image a patch is defined as a center pixel and the surrounding pixels increased by a window. Patches is used for correlation purposes since they provide more unique information than a single pixel.

Pin-hole model is a model used to describe geometric relations of the 2D image plane and a 3D object. This model is commonly used for cameras with CCD or CMOS sensors, where the pin-hole corresponds to the focal point. The algebraic relations of the pin-hole model are described in section 3.1.2.

Photogrammetry is the practice of performing geometric measurements of objects from images.

PMVS (Patch based multi-view stereo algorithm). An algorithm presented by J. Ponce and Y. Furukawa in (Furukawa, et al., 2010). The algorithm is aimed to create a 3D reconstruction of objects from multi-view camera images.

Precision is related to error estimation and is a measure of variability of en estimation.

Reference image. An image used in image registration, as reference for transformation of the template image to obtain maximal similarity between those images.

Registration. Image registration is a process where a template image is warped into a reference image to optimize the similarity between the two images.

RMS (Root Mean Square) refers to the square root of the mean of the squared of the values:

$$x_{\mathrm{rms}} = \sqrt{\frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{n}}$$

Equation 1.1

Skew-symmetric matrix has the form:

$$B = \begin{bmatrix} 0 & -b_z & b_y \\ b_z & 0 & -b_x \\ -b_y & b_x & 0 \end{bmatrix}$$

Equation 1.2

Shape contexts are a log-polar histogram representation of 2D or 3D points. A detailed description of the technique exists in section 2.6.1.1.

Template image. An image in that is transformed into a reference image in image registration, to optimize the similarity between the images.

Visual hull model (VH). A model obtained from the silhouettes of an object from multiple views. See section 3.5 in appendix.

Warp. Image warping is image manipulation in which the appearance is distorted as a result of a transformation of the pixel coordinates.

Abstract

Motion capture is widely used for gait analysis. Today most motion capture systems are marker based, which is both time consuming and imprecise. Many markerless approaches have been presented in the past years and the majority of the approaches are based on articulated models. One of the most promising approaches is presented in (Corazza, et al., 2006), in which a subject specific articulated models is applied. However the subject specific model is obtained by a laser scanner that is expensive to purchase. This master thesis presents an approach to obtain a full body model by using a photogrammetric approach that costs a fraction of a laser scanner.

The proposed approach is based on the PhotoModeler® Scanner software and eight cameras. Tests of the precision and the impact of varying spatial resolutions have been performed. The results of the precision tests showed that a 3±1 mm resolution can be obtained with 10 Mega pixel SLR cameras, if ideal illumination and texture is obtained. The precision obtained is comparable to the laser scanned models used in (Corazza, et al., 2006).

In addition, a pose estimation approach for markerless motion capture system is proposed to replace the subject specific articulated model approach. The advantage of replacing the subject specific articulated models is that the expenses and time consumption of synthesizing the models are avoided.

The key elements of the approach are as follows:

- Acquire detailed 3D models of the test subject by using a Patch Based Multi-view stereo algorithm proposed in (Furukawa, et al., 2010).
- Track the 3D points in the model over time and perform full body segmentation into rigid parts.
- Use the rigid segments to calculate the joint centers.

Since the novelty of the approach is the pose estimation represented in the second item, this thesis has focused on finding a method to solve this problem. The approach is tested on various 3D models of flexing limbs. The registrations were promising for small angular differences, but failed when the differences became large. Improvement of the registration algorithm would therefore be an obvious objective for a future work. However the final results of the test illustrated satisfying segmentations.

Contents

Introduction

The characteristics of human gait are as unique as a fingerprint (Li, 2009). The Department of Forensic Medicine at the University of Copenhagen has therefore focused on finding a markerless motion capture system that is capable of tracking joint centers and measuring joint angles. Also gait labs at hospitals and biomechanical institutes are interested in markerless systems because markers are time consuming and induce errors due to inexact positioning and movements of the skin.

The movie and gaming industry also use motion capture. Their primary interest is to make the movements of the animated characters as human as possible. The animated characters in “Avatar” are probably the most glorious example of this. The grown interest for 3D markerless systems within the past ten years has resulted in an enormous amount of approaches. Sylvia Yang, Ph. D. student at the University of Copenhagen has performed one of the latest studies of the state of the art (Yang, 2011). Thomas Moeslund, associate professor at the University of Aalborg has also performed a comprehensive study of state of the art in the past years (Moeslund, et al., 2001) and (Moeslund, et al., 2006). All of the approaches presented in these studies have in common that they are based on cameras, being either “time of flight” cameras or conventional cameras with sensitivity of light within the visual spectrum. According to (Yang, 2011) and (Moeslund, et al., 2006) it seems natural to classify the markerless approaches into single-view and multi-view approaches. Single-view approaches embrace the use of a single camera, whereas multi-view approaches embrace the use of multiple cameras. In addition these approaches can be subdivided into model based and model free approaches.

Single-view approaches are not valid for 3D motion capture, since the geometry does not allow estimation of a 3D position with the information acquired from a single camera only. Systems like the so called Stereo cameras and Microsoft’s Kinect can be classified as multi-view approaches, since they consist of at least two conventional cameras or a camera and a projector respectively (Microsoft, 2010). In theory the projector can count as a camera, since the geometrical equations for 3D point estimation can be used for the projector as well.

The Kinect seems to have a promising future for motion tracking. The price and simplicity makes the product attractive for the computer vision research area as a whole. It might be possible to perform

a setup covering all angles, based on Kinects instead of using conventional cameras, but the limited VGA resolution makes it less desirable for this study.

The most promising approach for markerless motion capture for biomechanical applications seems to be the model based. Yang claims that no model free approaches have yet been developed to measure internal and external rotations of limbs (Yang, 2011). Corazza et al. has developed a model based system, which is able to estimate the joint locations with an accuracy that competes with the marker based approach (Corazza, et al., 2007).

The approach consists roughly of following steps:

1. Obtain a subject specific articulated model
2. Automatic segmentation and localization of joint centers on the articulated model
3. Record data of the subject with 8 color VGA cameras
4. Perform a Visual hull model for each frame in data set
5. Register the articulated model to the visual hull models
6. Extract the positions of limbs segments and joint centers

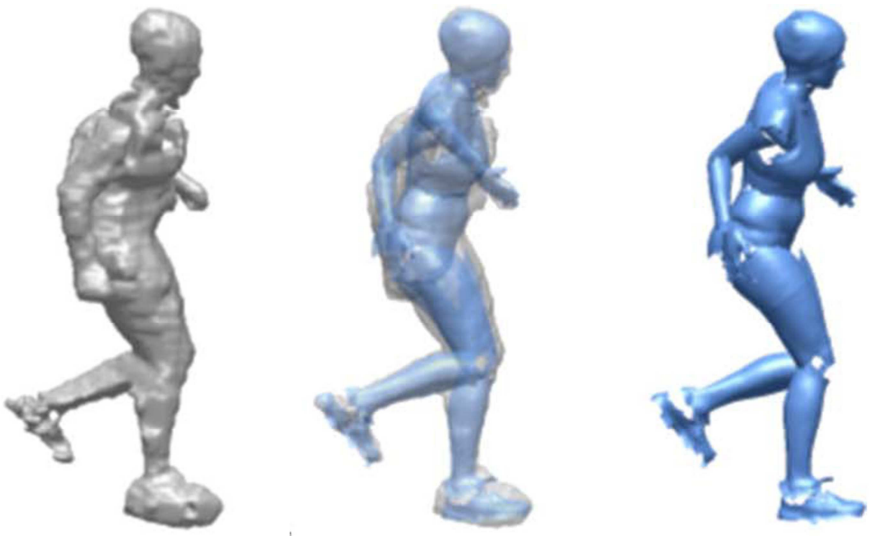


Figure 1.1: From left to right: Visual hull model, articulated model registered to the visual hull model, articulated model Reference: (Corazza, et al., 2009)

Unfortunately a subject specific articulated model has to be obtained by a laser scan of the subject. This makes the approach quite expensive.

The problem statements of this study are therefore to investigate the possibilities of replacing the laser scanned model with a model obtained with a photogrammetric approach to decrease the costs of integrating such a markerless system in a gait lab.

Providing an alternate approach for pose estimation without the use of subject specific articulated models and without compromising the accuracy will be tested as well.

A deeper introduction to the problem statements are presented in the beginning of Part I and II, concerning each to the two problems.

Part I: 3D modeling

1.1 Introduction

As mentioned in the introduction, Stefano Corazzas proposed an approach using a subject specific articulated model. Until now the subject specific model has been obtained with a laser scan of the subject. Since a full body laser scanner expensive to purchase, this part of the thesis have focused on how to replace the laser scanned model with a model obtained with photogrammetric methods.

PhotoModeler (PM) Scanner is a commercial product by EOS systems Inc. The photogrammetric software is developed to create 3D models from photos and is therefore well suited for the purpose of this study. Many other photogrammetric software products seem to be as suitable for the purpose as PM. However PM is well known and widely used product amongst those working with forensic medicine in Denmark. Because this group is primary user of the motion capture system, it might be profitable for them to keep using the software that they are already familiar with, unless there are significant advantages of changing it.

The quality of the 3D models created by any software is highly dependent on the hardware and the way it is used. It is therefore important to investigate the limitations of hardware/software combinations to find the optimal results with minimal resources.

1.2 Problem statement

The purpose of this part of the thesis is to test how to obtain an optimal hardware configuration in combination with the PM Scanner software to create a full 3D model of a person.

The term “optimal”, according to this problem, has to be weighted between economy and performance. Minimizing the amount of equipment and the quality as well whilst still being able to obtain a model with acceptable precision for the purpose, will be the key to success.

In (Mündermann, et al., 2005), it is illustrated how the visual hull (VH) model deviates from a laser scanned model for various numbers of cameras. However the results presented are based on simulations, not measurements of practical experiments. It is assumed that there will be a significant difference between theory and practice since noise in the images is not considered.

Figure 1.1 illustrates the deviations between the VH models and the scanned model. The deviations are measured by calculating the shortest distance between the two models.

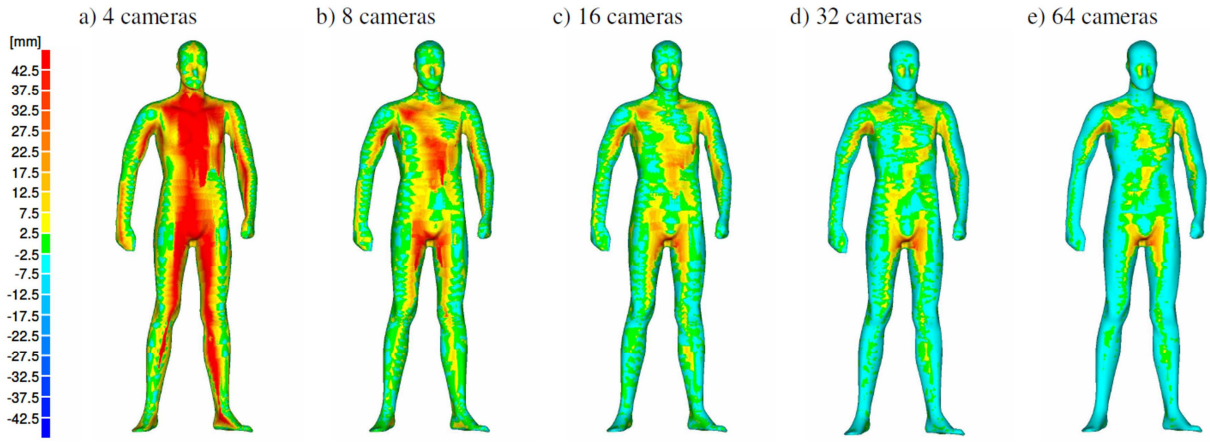


Figure 1.1: Surface deviations between the VH models and the laser scanned model for different camera configurations. A color ranging from cyan to blue indicates under-approximations whereas a color ranging from yellow to red indicates over-approximations. Reference: (Mündermann, et al., 2005)

The average of the deviations is plotted into a graph in (Mündermann, et al., 2005) as illustrated in Figure 1.2

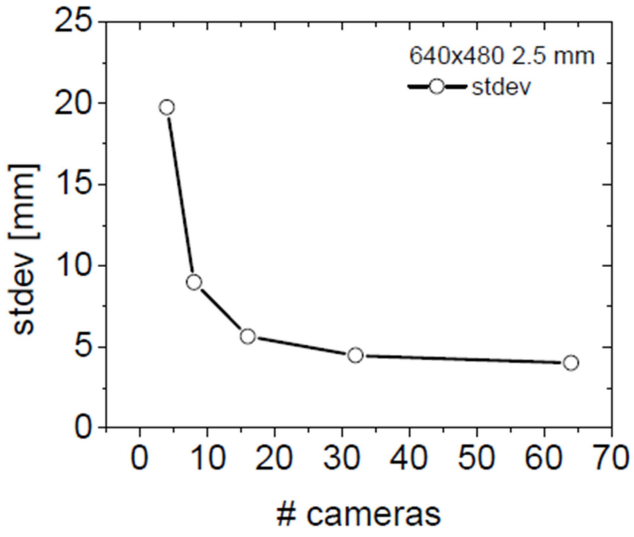


Figure 1.2: Standard deviations as function of the number of cameras. Reference: (Mündermann, et al., 2005)

From this, among other arguments, Mündermann et al. concludes that eight cameras are sufficient to obtain the VH models. Figure 1.2 shows the standard deviation to be approximately 8 mm using a configuration with eight cameras. It is therefore considered acceptable to obtain a subject specific model with a standard deviation less than 2.5 mm, since this will contribute with roughly 10% of the total variance from a statistical point of view. However the standard deviation of the VH models is expected to be significantly higher in practice.

1.3 Previous work

Using laser scans to perform 3D reconstructions are popular, because they have high accuracy for motionless objects. Corazza et al. claims in (Corazza, et al., 2010) that they provide a body model with a resolution at 4±1 mm using a Cyperware laser scanner. However a laser scan is often disturbed by movement artifacts, since the subject has to remain still throughout the scan time that typically takes ten to fifteen seconds (Istook, et al., 2000).

Full body models have already been obtained using photogrammetric approaches. J. Paul Siebert and Stephen J. Marshall claim that they can obtain a body model with accuracy at 2 mm Root Mean Square (RMS) using monochrome VGA cameras in (Siebert, et al., 2000). To achieve these results eight stereoscopic systems (sixteen cameras) are used with projection of a random pattern to ensure uniqueness in the texture.

Sixteen cameras and several projectors that project random patterns is a quite comprehensive setup. It is therefore reasonable to test how to minimize the technical equipment without compromising the precision and resolution of the reconstructed models significantly.

1.4 Introductory theory to 3D Modeling

This section presents a review of the theory that enables the reader to understand the techniques behind photogrammetry and how a 3D model can be obtained using multiple cameras. The review provides a good understanding of the problems regarding optimization of the configurations to obtain the 3D reconstructions.

Introductory optics and a presentation of the pinhole model are also described in section 3.1 in appendix.

1.4.1 Stereo vision

In this section contains a brief review of stereo vision. Derivations and algebraic approaches are presented in former work (Christiansen, 2010), is located in the “articles” directory in the attached CD.

1.4.1.1 Epipolar geometry

A stereoscopic system consists of two cameras. Using the pinhole model to describe the geometry of the cameras, it is possible to obtain depth perception by triangulation. According to the pinhole model, each camera consists of a focal point O_i and an image plane Π_i, where *i* is either *l* for left or *r* for right camera respectively. The line connecting the projective centers of the two cameras is called the **baseline**. In most common circumstances the baseline will intersect the image planes. Such a point of intersection is called an **epipole** denoted by e_i. In case of parallel image planes, the epipoles are located at infinity. The plane spanned between the focal points and an arbitrary point in space is called the epipolar plane. Since the epipolar plane intersects the image planes, each camera will observe the plane as a line. These lines are called **epipolar lines** denoted by u_i.

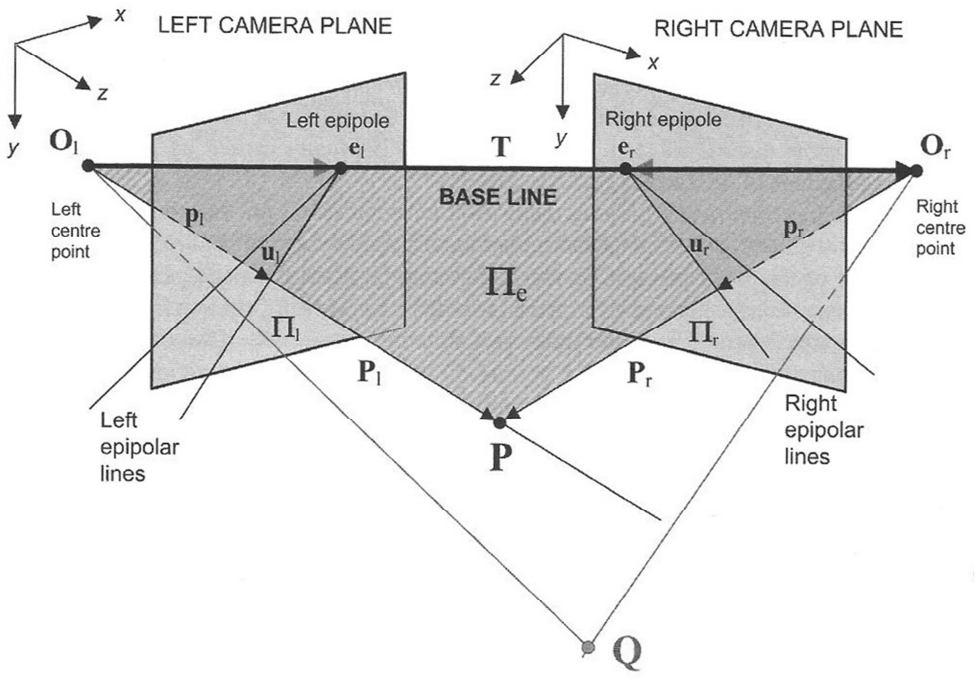


Figure 1.3: Epipolar geometry. Reference: (Cyganek, et al., 2009)

Epipolar lines play an important role in stereo vision, because a line intersecting the focal point and an arbitrary point on the image plane in one camera corresponds to an epipolar line in the image plan on the other camera.

A point correspondence appears when a point in space is reflected in both images. A point correspondence makes two vectors $\mathbf{p}_i$ defined by the focal point and the point in the image plane for each camera. The point in space P can be found by triangulation where the intersection between the lines P_i along the vectors p_i is found.

Unfortunately, whereas the lines provided by $\mathbf{p}_l$ and $\mathbf{p}_r$ intersect in the ideal situation, this is not the case in practice. Due to noise, distortions and errors in the calibration the intersection problem becomes a minimization problem that can be solved by the **Direct Linear Transformation** algorithm (DLT).

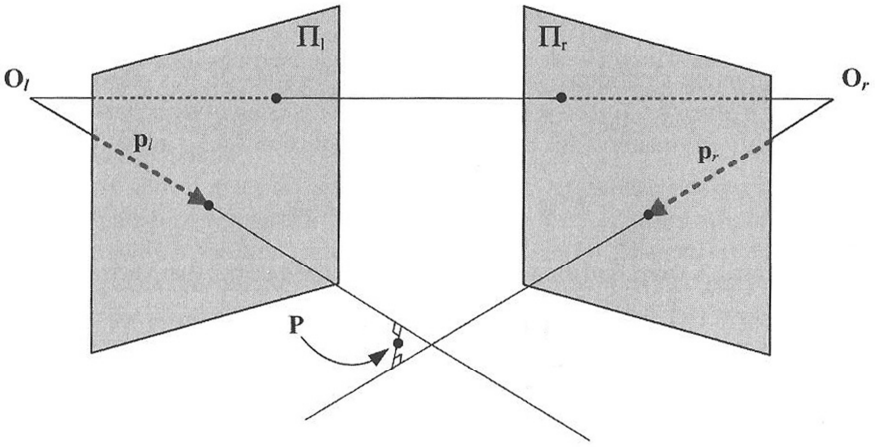


Figure: 1.4: Illustration of the practical case where p_l and p_r are not intersecting. Reference: (Cyganek, et al., 2009)

The error due to the uncertainty of the intersection is dependent on the length of the baseline. Figure 1.5 illustrates that uncertainty of point estimation becomes remarkably larger when the two cameras are positioned in parallel with a short baseline in contrast to an angled position with a wide baseline. According to the triangulation problem it is therefore always preferred to get a wide baseline between the cameras.

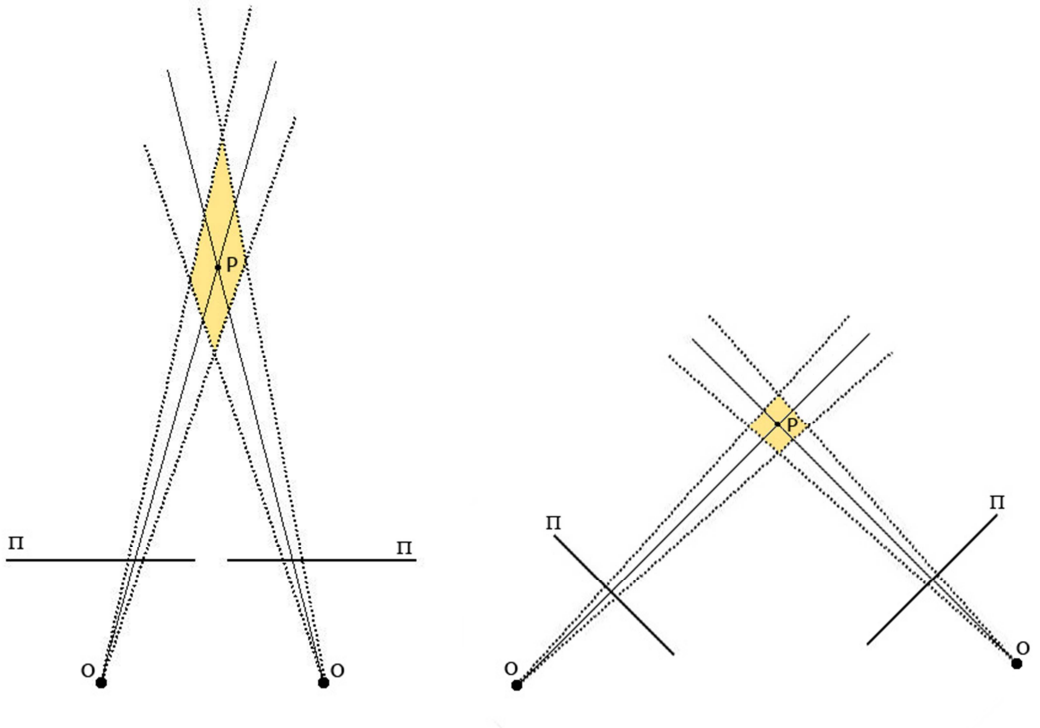


Figure 1.5: This illustration shows that the uncertainty, labeled with orange, is remarkably larger the closer the camera positions are to each other.

The vectors $\mathbf{p}_l$ and $\mathbf{p}_r$, defined by the focal points and the point correspondences in the image plane, are related by the **essential matrix** E as:

$$\mathbf{p}_r^T E \mathbf{p}_l = 0$$

Equation 1.1

By Equation 1.1 it follows that an epipolar line can be defined by:

$$u_r = E\mathbf{p}_l$$

Equation 1.2

The essential matrix is related to the extrinsic parameters (the rotation and translation from one camera to another) of the cameras. Similar to Equation 1.1 the relation between a 2D point correspondence can be written by:

$$p_r^T F p_l = 0$$

Equation 1.3

Here p_l and p_r is a 2D point in the image plane and F is called the **fundamental matrix**. In general the fundamental matrix describes the epipolar geometry in terms of 2D pixel coordinates in contrast to the essential matrix that describes the geometry in terms of camera coordinates. Like the essential matrix, the fundamental matrix is related to the extrinsic parameters but in addition it also relates to the intrinsic parameters. It is therefore possible to estimate the extrinsic parameters with a prior knowledge of the intrinsic parameters using the fundamental matrix.

1.4.1.2 Canonical Stereoscopic System

Previously it was mentioned that a wide baseline and angle is preferred to minimize the uncertainty in triangulation. Despite this argument it is common to position the cameras parallel and close to each other as illustrated in Figure 1.6.

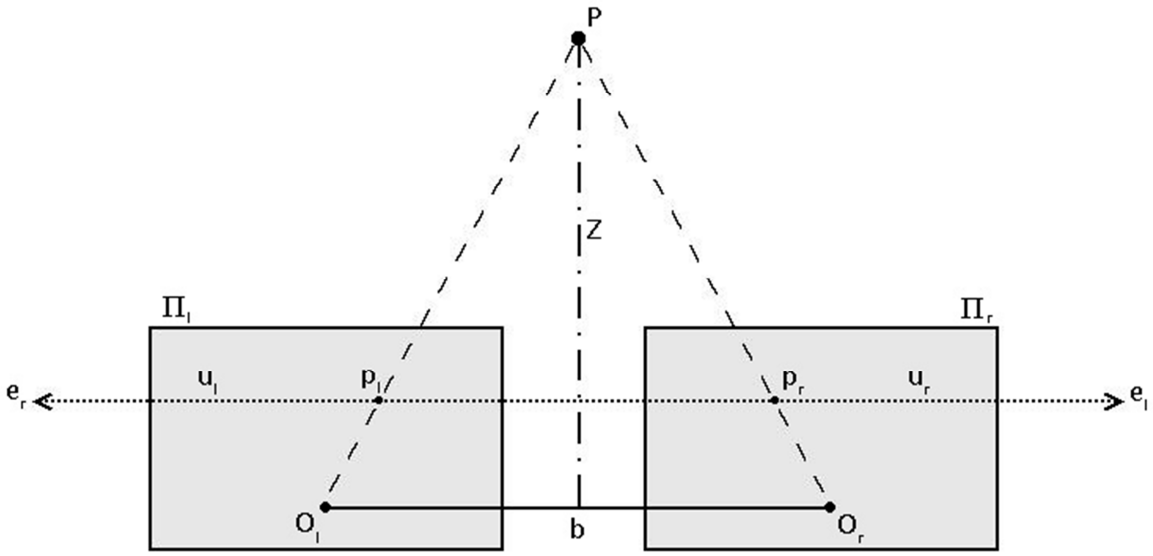


Figure 1.6: Geometry of a canonical stereoscopic system

This configuration is often denoted as a standard (canonical) stereoscopic system. By then the epipolar lines will be horizontal in the image planes. This is beneficial in sense of point matching, because the search line for a point match is along a pixel row, which limits the search space to one dimension only. Most importantly it also limits the shear of the region of interest due to the perspective projection into the image plane. Point matching is further described in section 1.4.2.

Figure 1.6, illustrates a horizontal disparity between the two images from the left side of the image plane to the point p_i. In general the disparity can be written as:

$$D(p_l, p_r) = \sqrt{D_x^2(p_l, p_r) + D_y^2(p_l, p_r)}$$

Equation 1.4

Here $D_x(p_l, p_r)$and $D_y(p_l, p_r)$is the disparity in x- and y direction, respectively. The disparities are calculated by:

$$D_x(p_l, p_r) = x_{pl} - x_{pr}$$

Equation 1.5

$$D_y(p_l, p_r) = y_{pl} - y_{pr}$$

Equation 1.6

Here 'x_i' and 'y_i' is the x- and y-coordinates of the point correspondences, respectively.

Using a stereoscopic configuration, the vertical disparity is eliminated. Equation 1.4 can therefore be rewritten as:

$$D(p_l, p_r) = |D_x(p_l, p_r)|$$

Equation 1.7

By using the canonical stereoscopic system, the measure of disparity in the horizontal direction can easily be translated to a measure of depth, since the epipolar geometry allows following statement (Cyganek, et al., 2009):

$$D_x(p_l, p_r) = b\frac{f}{Z}$$

Equation 1.8

Referring to Figure 1.6, *b* is the length of the baseline, *Z* is the orthogonal distance from the baseline to the point in space P. Note that b, f and Z are all positive parameters. From this it follows that $x_{pl} \geq x_{pr}$, which limits the search range further.

Using the disparity, it is possible to perform a transformation from one point in a correspondence to another by:

$$p_r(x, y) = p_l\left(x + D_x(p_l, p_r), y + D_y(p_l, p_r)\right)$$

Equation 1.9

Disparity can be used to perform a so called disparity map. A disparity map is an image where the depth is mapped with a gray scale in accordance to one of the two stereo viewpoints. Large disparities are labeled with high intensity values. Figure 1.7 shows an example of a stereo image pair from which a disparity map is created (see Figure 1.8) from the left camera view point.

Figure 1.7: Left: Left stereo image; Right: Right stereo image

Figure 1.8: Disparity map from the left stereo view point

In disparity maps, white spots represent areas where disparity is undefined. As illustrated in Figure 1.8, the disparities in points near depth edges or discontinuities (e.g. near the left arm, the cowl and the right side of the face) are typically hard to define. The reason is discussed in more detail in section 1.4.2.1 concerning area based point matching algorithms.

1.4.1.3 Dependence on resolution

Resolution is related to sampling frequency when speaking about signals. The resolution can be measured in both the image plane and the object plane. In the image plane it is denoted "pixel resolution" and is often denoted in number of pixels. For stereoscopic systems the resolution obtained in space is denoted as "spatial resolution". The spatial resolution defines how closely points can be in space and still be resolved in an image. The spatial resolution is also affected by optical blur. In the following presentation it is assumed that the optical blur can be neglected.

The spatial resolution can be further divided into horizontal-, vertical- and depth resolution. The geometrical relation between focal length *f*, distance from the cameras *H*, pixel size *r*, baseline *b* and depth resolution R_d is illustrated in Figure 1.9.

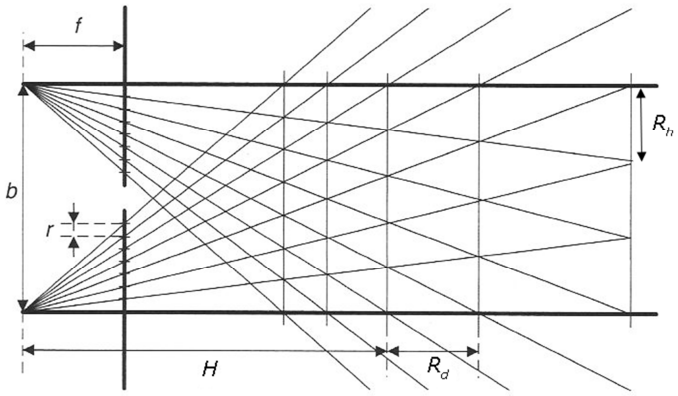


Figure 1.9: Modified sketch from (Cyganek, et al., 2009)

The depth resolution R_d can be expressed by:

$$R_d = \frac{rH^2}{fb - rH}$$

Equation 1.10

When *r* is sufficiently small, the equation can be simplified to:

$$R_d \approx \frac{rH^2}{fb}$$

Equation 1.11

The horizontal spatial resolution is the resolution obtained in the horizontal direction in the object plane. It is proportional to the pixel size and focal length, expressed by:

$$R_h = H\frac{r}{f}$$

Equation 1.12

Considering the depth- and the horizontal resolutions in Equation 1.11 and Equation 1.12 it can be seen that R_d is proportional to H^2 and R_h is proportional to *H*. This means R_d is very sensitive to *H* and is always larger than R_h as long *H* is larger than *b*. The vertical spatial resolution is similar to the horizontal spatial resolution, but in the vertical direction.

A high spatial resolution is therefore of big importance in order to minimize the reconstruction error. A matching algorithm will never be more accurate than the hardware used to perform the measurement. Anyhow a denser mesh can be achieved by interpolation. It is important to emphasize that no additional information is added by interpolation and aliasing cannot be prevented by interpolation.

1.4.2 Point matching algorithms

Point matching is a process where a unique point in an image is extracted by comparing different images of the same object. The uniqueness of a point is commonly based on either intensities or features in the original image, but can also rely on extractions in transformed versions of the images (typically log-polar transformed) or structural tensors that are discussed in section 3.2.3.

Point matching is a stepwise process that embraces the following:

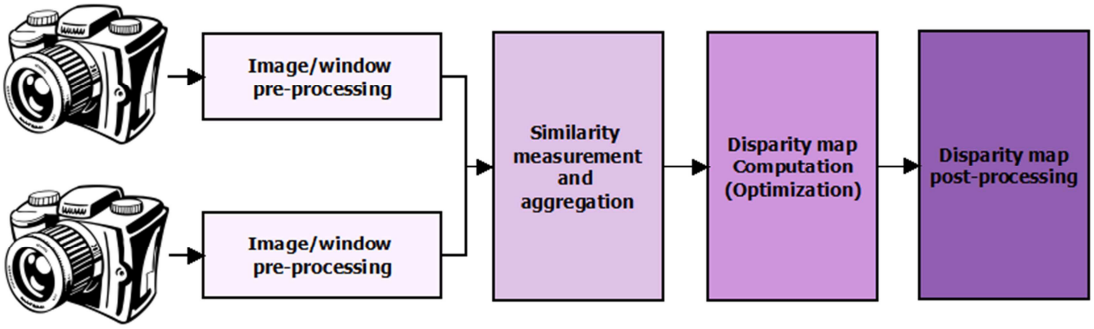


Figure 1.10: Diagram of a point matching process

Image/window pre-processing. Pre-processing consists of filtering, transformation or warping. Filtration is performed to enhance unique features such as lines and corners and suppress the noise. Transformation is an important process to make the region of interest invariant for certain parameters. As an example, a log-polar transformation obtains invariance to scaling and rotation. Finally image warping is performed to decrease the disparity search space and reduce the amount of outliers due to projection differences.

Similarity measurement and aggregation. Comparison of pixels is frequently performed with integration of the local neighborhood (often denoted as patches), since local pixels has only limited discriminative power. The different comparison approaches is discussed in section 1.4.2.1 and 1.4.2.2. Depending on the approach, all pixels or features in a template image is compared within a constrained search range of the reference image. The similarity is measured according to the measures discussed in section 1.4.2.2.

Disparity map computation. The procedure for computation of the disparity is organized by either local or global optimization methods. The optimization is based on similarity measurements.

Local optimization is used in both feature based matching and area based matching, both discussed in the next sections. The strategy is often called “Winner-Takes-All” (WTA), because it is the maximal correlation within a constrained match region that becomes the disparity value.

Global optimization is used in other techniques such as dynamic programming techniques that are not discussed in this report. This method is often more powerful, since all correspondences is simultaneously evaluated. The Hungarian algorithm, described in section 3.17 in appendix is a classic example of global optimization.

Disparity map post processing. The computed disparity map is often dominated by spikes and missing values. Therefore the following post processing is obtained on disparity maps.

1. Sub-pixel disparity estimation is performed by polynomial interpolation.
2. Verification by cross-checking. By using the disparity map from both pictures it is possible to perform a cross-check of the values.
3. Filtration of disparity map to get rid of spikes.
4. Interpolation of missing disparity values.

The next section introduces some basic approaches for point matching, corresponding to the first three out of the four steps in Figure 1.10.

1.4.2.1 Area based point match

Area based point match is a general term for matching approaches where patches in one image are compared with patches in another image. As opposed to the feature based point match, with area based point match, all pixels in the template image are compared to pixels within a restricted search space in the reference image, constrained by the epipolar line among other constraints related to the individual point match algorithm. The constraints have a big importance to avoid outliers and speed up the computation time. Due to the dense sampling, these algorithms generally perform dense disparity maps.

One approach of area based point match is the use of correlation techniques that is presented in the following sub section. According to (Inc., 2010), the PhotoModeler Scanner® is based on this specific technique.

1.4.2.1.1 The Correlation technique

The correlation technique can briefly be described as follows: Referring to Π_l as the template image and Π_r as the reference image, a patch from the template image is used as a mask. The best match for the mask is found by correlation with the reference image along the image scan line corresponding to the epipolar line as illustrated in Figure 1.11.

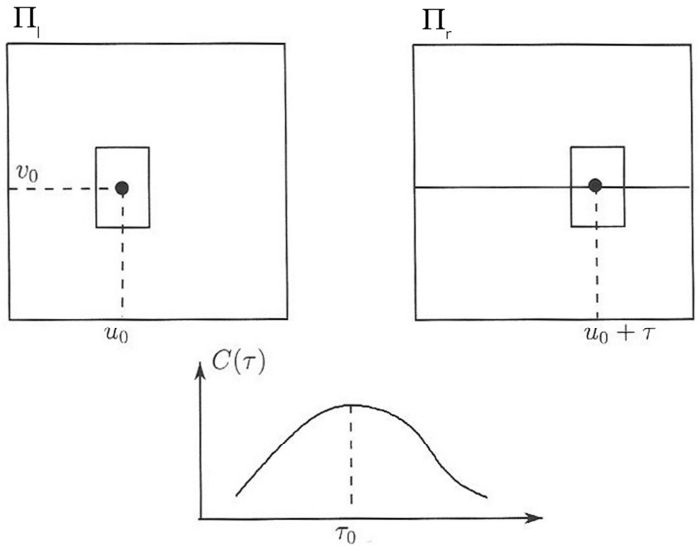


Figure 1.11: Principle of the correlation technique (Faugeras, 1993)

The best match is considered as the point with the maximal correlation $C(\tau_0)$.

The correlation technique implies following problems:

1. If no canonical stereoscopic system is applied, then rectification has to be performed to obtain the correlation along the epipolar lines.

2. If the contrast in the texture is weak and noisy, then a correlation might have several weak optimums.

3. The gradient of the disparity is assumed to be zero within the correlation window. This means that the modeled surface has to be locally continuous and locally fronto-parallel to the image planes. Figure 1.12 shows that the disparity deviates between two point correspondences, due to projection differences between the images.

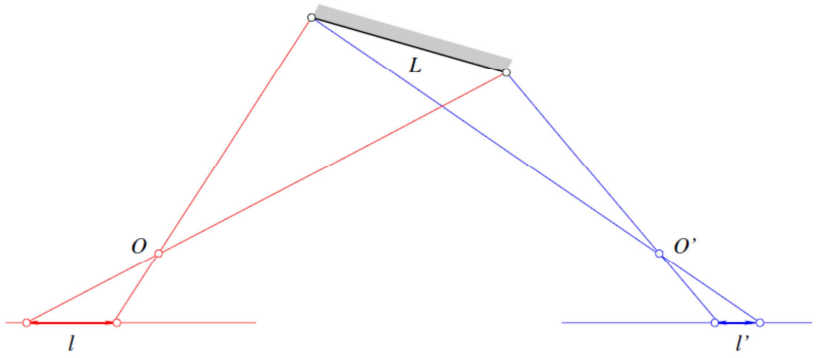


Figure 1.12: The distances *l* and *l* ' between the extracted points is different for the two images (Forsyth, et al., 2003)

The Coarse-to-fine matching approach, discussed in section 3.4, is coping with the first two problems. Warping has shown to be effective in coping with the third problem, as presented in (Devernay, et al., 1994). The warping approach is briefly described in section 3.7.1 in appendix.

In (Faugeras, 1993) it is proposed to convolve the stereo images with Laplacian of Gaussian (LoG) filters before the cross correlation. The LoG filtration has two purposes: Subtracting the mean from the signal and removal of high-frequency components. Subtraction of the mean before cross-correlating leads to results that are practically the same as finding the covariance.

$$cov_{lr} = \frac{1}{N}\sum_{x=1}^{N}(p_l(x-\mu_l))(p_r(x-\mu_r))$$

Equation 1.13: Covariance

The correlation coefficient is then achieved by dividing the covariance with the product of the standard deviations of the two patch signals. The correlation score is well suited for similarity measurement, since the scores lies in the range [-1; 1], where 1 indicates maximal correlation. The best feature of the correlation coefficient is that it is invariant to differences in gain and exposure of the cameras.

$$corr_{lr} = \frac{\frac{1}{N}\sum_{x=1}^{N}(p_l(x-\mu_l))(p_r(x-\mu_r))}{\sigma_l\sigma_r}$$

Equation 1.14: Correlation coefficient

The correlation technique can be performed on the pixel intensities, the nonparametric image spaces (such as Log-polar space) and the structural tensors as well.

1.4.2.2 Similarity measures for image regions

The similarity between two image regions is used to measure the cost of the disparity estimation. By optimizing the cost, with respect to the shift between the template window and the reference image, the best disparity approximation can be found. In the following sections, the most commonly used similarity measures for image regions are discussed.

1.4.2.2.1 Sum of squared differences (SSD)

SSD is probably the simplest similarity measure for image regions. The relative small amount of computations makes SSD beneficial to real time matching.

The sum of squared differences is defined by:

$$D_{SSD} = \sum_{(i,j)\in U} I_l\big(x + d_x + i, y + d_y + j\big) - I_r(x + i, y + j)$$

Equation 1.15: SSD of the intensities in the image regions

Here U is the windowed region centered at $\big(x + d_x, y + d_y\big)$ and d_m is the disparity along the m'th dimension. I_l and I_r is the intensity in the particular pixel coordinate for the left and the right image, respectively.

In SSD it is assumed that equal mean values and a constant variance are obtained in the two images. SSD is therefore well suited for comparing images obtained from identical camera models with identical configurations.

1.4.2.2.2 Correlation coefficient

If the camera configurations are not identical between the two stereo cameras or if non uniform illumination appears in the images, it is likely that mean and variance vary differently in the acquired images. In such a situation it would often be preferred to use the correlation coefficient instead of SSD.

Denoting the intensity value of a single pixel within a window:

$$I_l = I_l\big(x + d_x + i, y + d_y + j\big)$$

$$I_r = I_r(x + i, y + j)$$

Equation 1.16

And the mean of the window:

$$\overline{I_l} = \overline{I_l(x + d_x, y + d_y)}$$

$$\overline{I_r} = \overline{I_r(x, y)}$$

Equation 1.17

The correlation coefficient is defined by:

$$D_{SSD} = \frac{\sum_{(i,j)\in U}(I_l - \overline{I_l}) \cdot (I_r - \overline{I_r})}{\sqrt{\sum_{(i,j)\in U}(I_l - \overline{I_l})^2 \cdot \sum_{(i,j)\in U}(I_r - \overline{I_r})^2}}$$

Equation 1.18: The correlation coefficient

Equation 1.18 shows that the correlation coefficient is equal to SSD normalized with respect to both variance and mean.

If the image intensities are modified differently, e.g. by using gamma correction, it might be better to compare the gradients of the image regions instead of the direct measure of intensities. To do this, the SSD is rewritten to:

$$D_{SSD} = \sum_{(i,j)\in U} \nabla I_l(x + d_x + i, y + d_y + j) - \nabla I_r(x + i, y + j)$$

Equation 1.19: SSD of the gradients in the image regions

Mutual Information and structural tensors are even better alternatives to the gradient based SSD, if the computational power is not a problem. Structural tensors are discussed in 3.2.3 and Mutual Information is discussed in both (Cyganek, et al., 2009) and (Larsen, 2008).

Roughly all dense point match algorithms use correlation to some degree. Some state of the art approaches are presented in section 3.7 in appendix.

1.4.3 Summary

In this section the techniques to acquire a 3D model from a photogrammetric approach have been presented. The pinhole model is setting the frames for the triangulation that is used to find a 3D point in space. The triangulation can only be performed using the minimum of two cameras with known camera parameters, which are found by camera calibration.

Setting up the cameras in parallel to each other with a small baseline in relation to the object is called a stereoscopic system. With such a system it is a relatively easy and fast to obtain a disparity map, since the search space to find the point correspondences are decreased to the epipolar lines that goes along with the scan lines of the images. The spatial resolutions of the disparity maps are dependent on the pixel resolution, the baseline and the distance to the object plane. Dense point matching can be obtained using correlation based techniques, which can be optimized by multiple approaches. The most common optimization approaches are using scale space to decrease the search space for the point correspondences. The correlation coefficient is often used as similarity measure for estimating the point correspondences.

1.5 Problem analysis

Previous section provided a review of the basic concepts in stereo vision and photogrammetry. This section will use this theory to analyze the problem defined in the problem statement.

As also mentioned in section 1.4.2.1, (Hullo, et al., 2009) claims that PM Scanner's dense surface modeling algorithm (DSM) is based on the correlation technique.

The results provided by this technique are, according to (Hullo, et al., 2009), dependent on three things: Geometry, recovery and correlation. Geometry refers to the fact that a broad baseline provides better depth resolution and small errors as discussed in section 1.4.1.1 and 1.4.1.3. Recovery refers to the ability to have full insight on the whole target with both cameras i.e. occlusions for one of the cameras provide holes in the disparity map. Correlation refers to all the problems and constraints related to the correlation based technique. Among the problems related to the correlation factor is the fronto-parallel assumption, which is described in section 1.4.2.1.1, the most prominent factor.

According to the PhotoModeler Scanner Manual, the DSM algorithm requires a "Base-to-height ratio" (BH ratio) between 0.2-1.0 to create dense surfaces. A BH ratio below 0.5 should be optimal. This information is not sufficient to optimize the performance by changing the BH ratio. In (Hullo, et al., 2009) they present the table shown in Figure 1.13 that illustrates the performance. The performance is divided into four gray scales. Dark gray indicates best performance and white indicates poor performance.

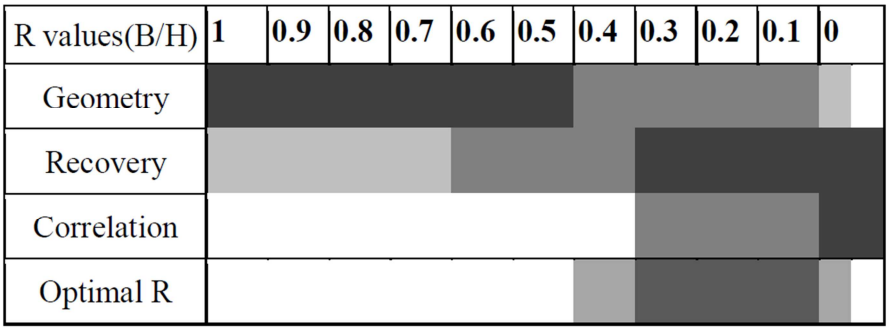


Figure 1.13: Performance of the DSM algorithm for various BH-ratios. Reference: (Hullo, et al., 2009)

The table seems to provide a proper insight into the performance of the DSM algorithm. However the references and the arguments for the claims are not clear and well documented in the article. It

is therefore necessary to investigate how the performance changes according to the geometry and the correlation parameters by two precision tests:

1. Precision dependence on BH-ratio
2. Precision dependence on surface angle

The first test concerns the impact of the geometry- and the correlation parameter on the precision of the model. The second test concerns primarily the correlation parameter and investigates how large angles the DSM algorithm is capable to model with a standard deviation less than 2.5 mm as stated in the problem statement. Both tests are presented in the following section.

The recovery parameter will not be tested since the impact of this parameter seems obvious.

Based on the knowledge acquired by the tests, an investigation will be performed on how to create a setup that is capable of obtaining a full body model.

1.6 Precision tests

1.6.1 Introduction

To measure the precision, the following hypothesis to describe a model for a plane surface is used:

$$\text{Modeled surface} = \text{Plane surface} + \text{Noise}$$

Equation 1.20

Here the precision is dependent on the amount of noise. The noise mainly consists of the quantization error caused by finite depth resolution and the noise provided by the correlation technique. The power of the noise provided by the depth resolution is dependent on pixel resolution, focal length and the baseline. Since pixel resolution and focal length is constant, it is possible to find the correlation between the precision and the geometry/baseline. The power of the noise provided by the correlation technique is, among other factors that are discussed in section 1.4.2.1.1, dependent on contrast and optical blur. Here texture has a big impact, since a good contrast in the texture enhances the signal and therefore also the SNR ratio if the image is perceived as a two dimensional signal. Similar conditions are aimed as expected for a full body model acquisition. A test board is therefore coated with textiles and used as a plane to achieve a surface similar to a dressed test subject. Anyhow the clothes of the test subject are constricted to have some degree of small randomized pattern with high contrast to minimize the noise. Both the noise provided by the quantization error and by the correlation are assumed to consist of high-frequency components.

Using Equation 1.20, the amount of noise/precision is quantified by the standard deviation of the plane surface along the direction of the normal, since it is assumed to be zero for a plane surface. The standard deviation can be obtained by Principal Component Analysis (PCA), where the square root of the third component will reflect the standard deviation along the normal direction.

Unfortunately the board used in the test has a small curvature that can be seen with the naked eye. Equation 1.20 is therefore rewritten as follows:

$$\text{Modeled surface} = \text{Plane surface} + \text{Curvature} + \text{Noise}$$

Equation 1.21

Since the goal is to quantify the noise, the curvature in the model must be removed. The suggestion is to create a model of the curvature using Thin Plate Spline (TPS), and subtract it from the modeled surface, which can legally be done because the curvature of the plane is assumed to be low-frequency whereas the noise was assumed to be high-frequency. TPS is basically a spline algorithm that behaves like a metal plate, where the stiffness can be regulated by a smoothing parameter λ. More details about TPS can be found in section 3.6 in appendix.

TPS is a computationally heavy algorithm and a TPS on a point cloud with $1 \cdot 10^6$ points will provide $1 \cdot 10^{12}$ equations with $1 \cdot 10^{12}$ unknowns. Desktops will normally run out of memory when trying to solve such an equation. It is therefore necessary to decimate the point cloud used to model the TPS. Decimation consists of smoothing the point cloud to avoid aliasing and down sampling. Both operations can be performed in PhotoModeler. Since the smoothing will suppress the noise, the smoothed point cloud can be used to model the TPS. However a heavy regularization needs to be performed to avoid overestimation of the curvature.

1.6.2 Test setup

1.6.2.1 Precision dependence on BH-ratio

To quantify the error contributed by the geometry parameter, a test is performed where stereo images are grabbed of a plane surface with varying BH-ratios. The surface will be located fronto-parallel to the image planes, as illustrated at Figure 1.14, to minimize the errors provided by the correlation factor. The BH ratio of the stereo setup is tested between 0.1-1 with 10 spatially equidistant samples. Figure 1.14 illustrates the setup of the plane and the cameras.

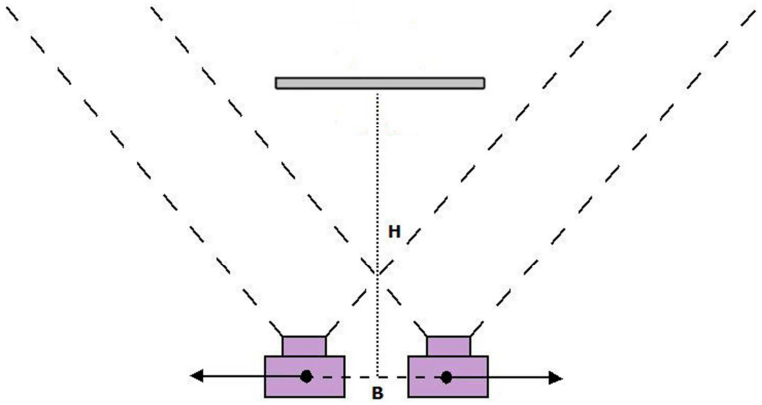


Figure 1.14: Camera configuration for the test of the geometry factor

In addition, the same test with the surface rotated 45 degrees in relation to the rectified image planes is performed as illustrated in Figure 1.15. By then the fronto-parallel assumption cannot be sustained. This implies that the error contribution from the correlation factor will be enhanced such that it can be estimated as a function of the BH-ratio.

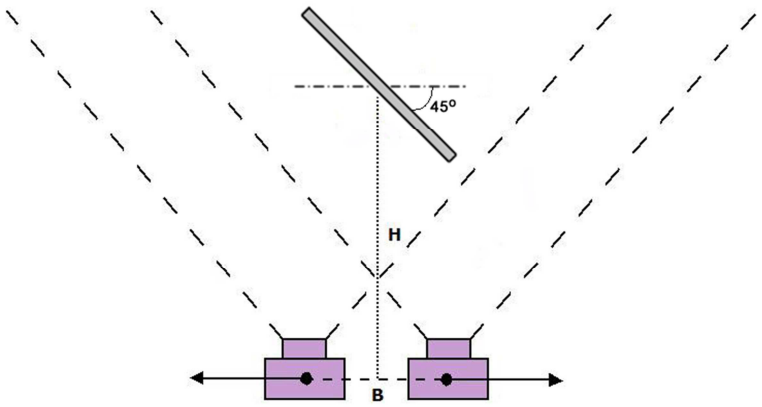


Figure 1.15: Camera configuration for the test of the correlation factor

The error will be estimated in terms of the standard deviation.

1.6.2.2 Precision dependence on surface angle

To get an estimation of the error at different surface angles, the BH-ratio was kept constant at 0.3 and the standard deviation for 10 equidistant angles between 0 and 90 degrees was measured in relation to the rectified image planes. Figure 1.16 illustrates the setup.

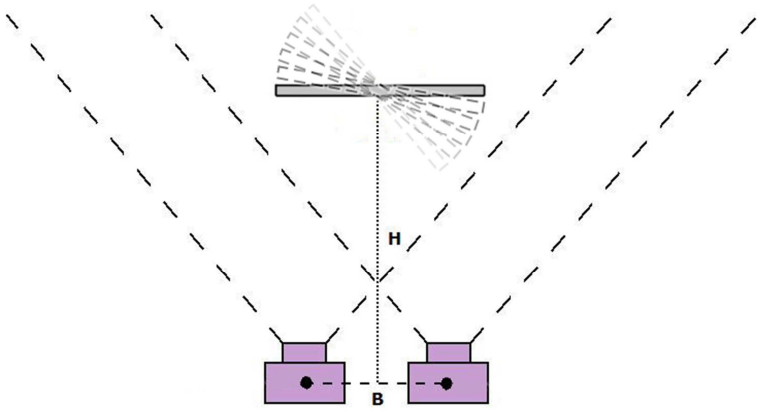


Figure 1.16

The next sections describing hardware and data processing are similar for both tests.

1.6.2.3 Hardware

The hardware used to create the setups is listed below:

1. 2 x Cameras
2. 2 x Lenses
3. 2 x Tripods
4. Test board: 120 cm x 80 cm

The specifications of lenses and cameras are listed in Table 3.1 in appendix.

The test board is illustrated in Figure 1.17.

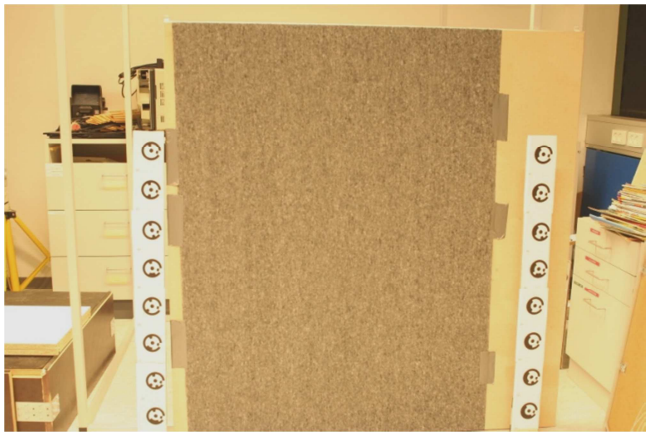


Figure 1.17: Test board with coded targets

1.6.2.4 *PhotoModeler processing*

The data processing in PhotoModeler is performed according to the processing description appended in appendix 3.8. The values for the DSM parameters are listed in Table 1.1.

Category	Value
Sampling interval	3 mm
Matching region radius	5
Texture type	3
Sub-sampling factor	2

Table 1.1: DSM parameters

The point cloud used for TPS was further processed in PM, by point decimation. Following values was used:

Category	Value
Smooth points:	
Smoothing amount	100
Iterations	100
Point decimation	99%

Table 1.2: Additional parameters for TPS data

1.6.2.5 MATLAB processing

All data was exported to MATLAB in txt-format, where the further processing was performed. According to TPS, a Degree of Freedom (DoF) at 20-25 was aimed, after a visual analysis of an optimal value. The DoF seemed to be somewhat underestimated. This probably provided a larger variance than the actual. Alternatively there would be a potential chance of overestimation, by which some of the noise would be modeled and provide better results than actually achieved.

As previously described, the standard deviation along the normal of the plane was found using the square root of the length of the third principal axis.

1.6.3 Results

1.6.3.1 Thin plate spline

The following figures illustrate the results of a TPS fit to data from an arbitrary BH ratio (BH ratio = 0.2).

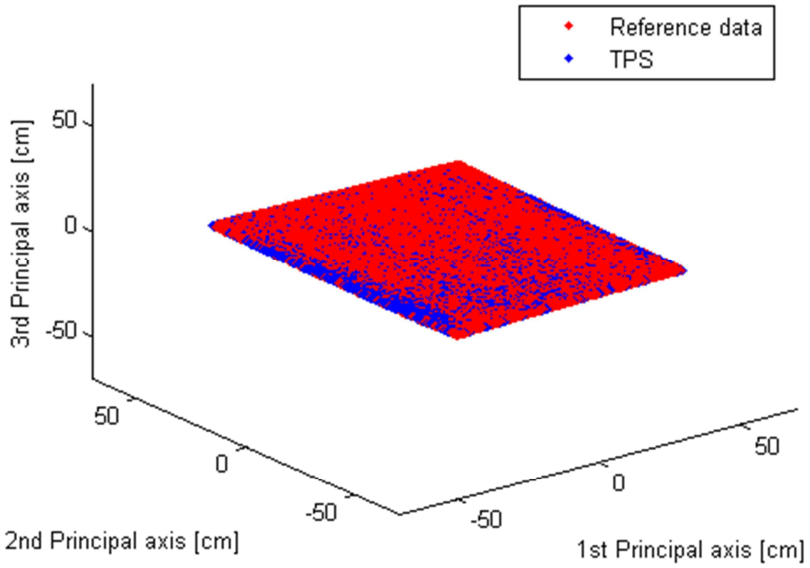


Figure 1.18: TPS model and reference data plotted with isometric axes

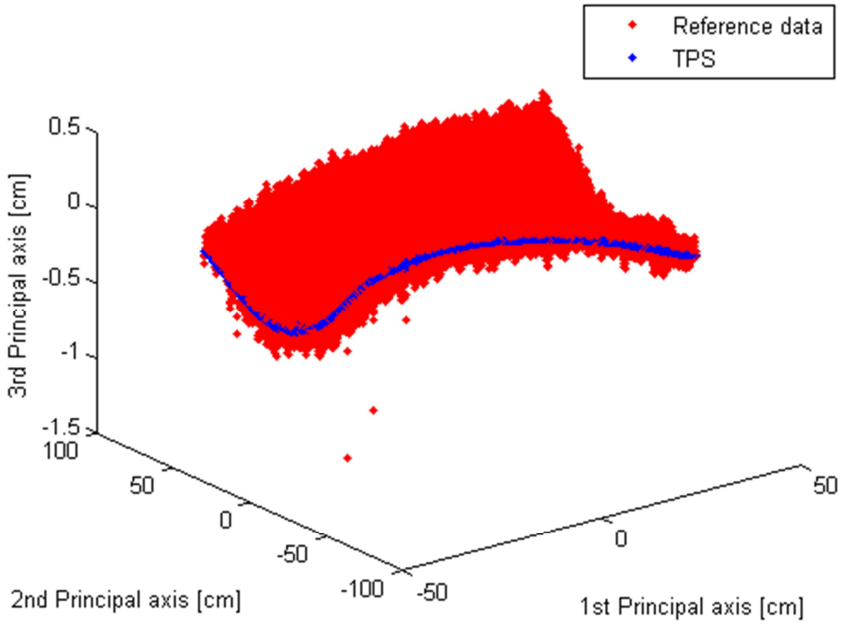


Figure 1.19: TPS model and reference data plotted with non-isometric axes

Figure 1.18 and Figure 1.19 show a successful fit of the TPS. The dominant blue color at the borders on the model in Figure 1.18 indicates a slight underestimation of the curvature of the surface.

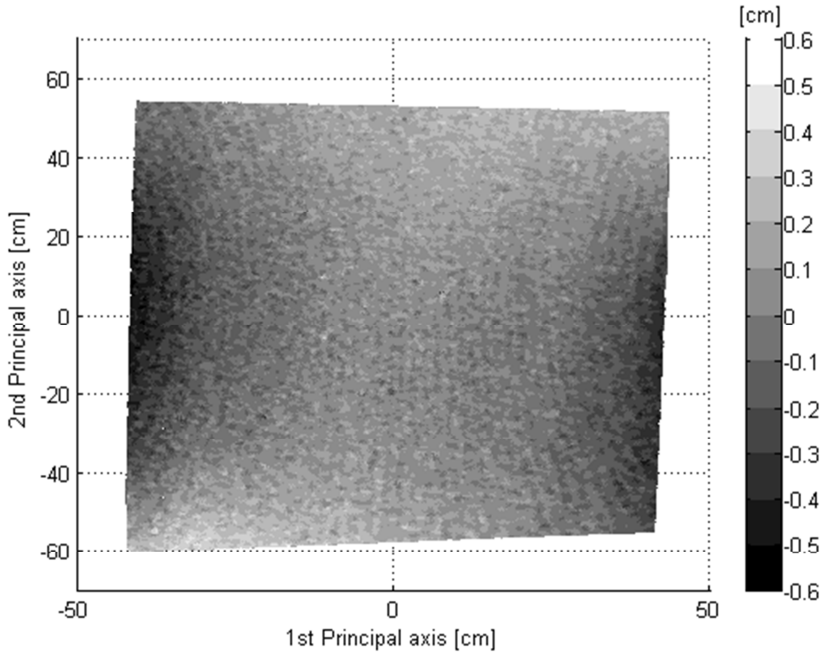


Figure 1.20: Triangulated surface plot of arbitrary data before the curvature correction.

Figure 1.20 shows that the curvature of the surface is dominating the variation along the 3rd principal axis. Some ringing effect seems to appear in the surface. This artifact has unknown appearance and might originate in some aliasing or truncation error in the lens distortion correction.

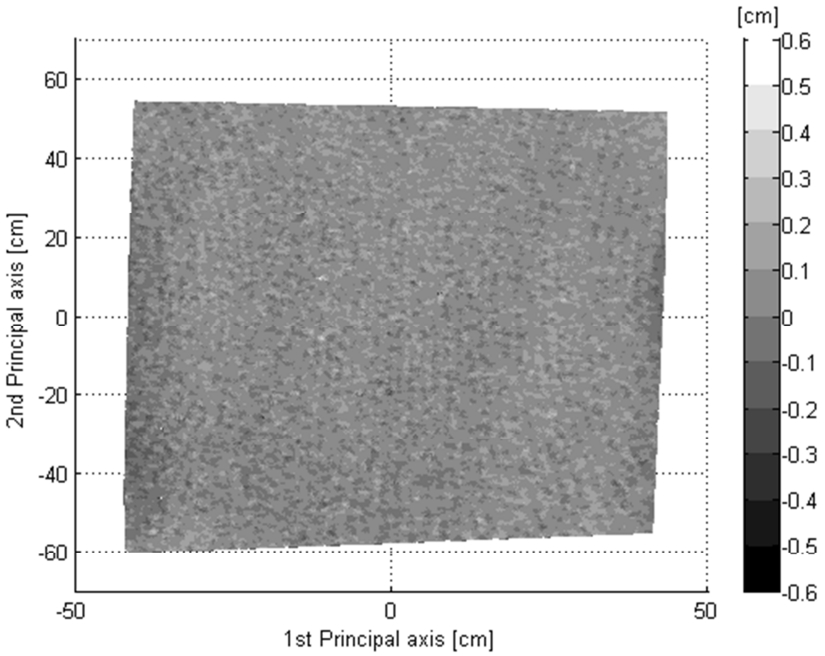


Figure 1.21: Triangulated surface plot of arbitrary data after the curvature correction.

The final result of the curvature correction, illustrated in Figure 1.21 shows that roughly all the low-frequency components are removed as aimed.

1.6.3.2 Precision dependence on BH-ratio, curvature corrected

The graph in Figure 1.22 shows the standard deviation of the surface measurements as a function of the BH-ratio for the parallel and the rotated surface.

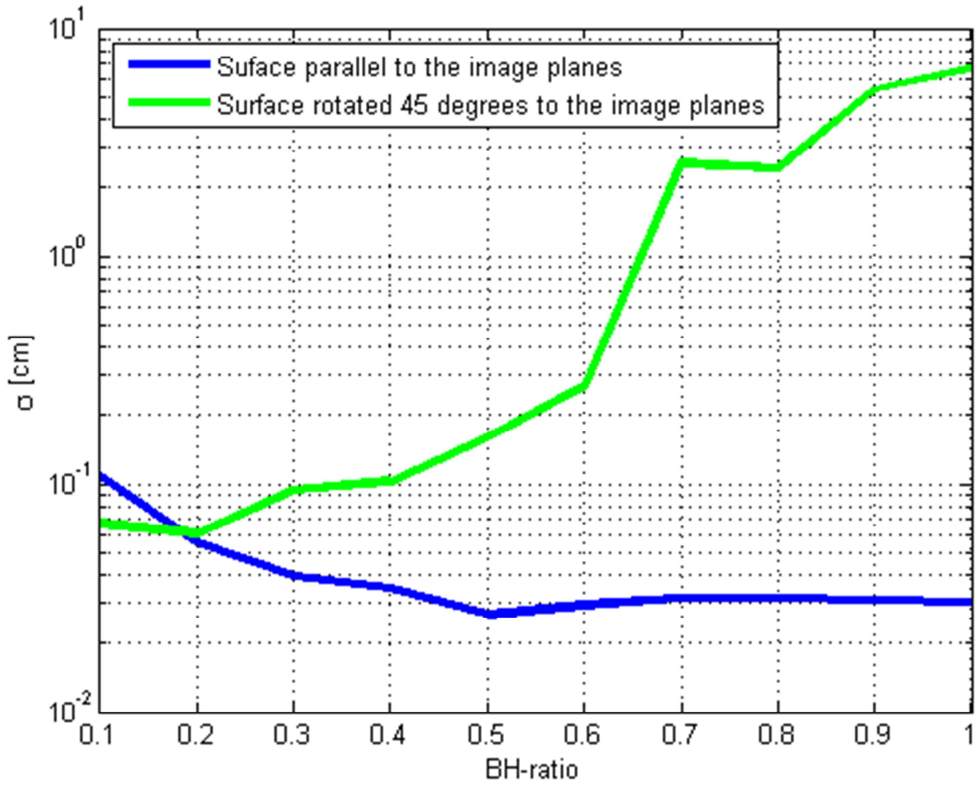


Figure 1.22: The graph shows the standard deviation of the surface measurements as function of the BH-ratio. The results are acquired using a curvature corrected surface, located parallel and with 45 degrees rotation to the rectified image planes.

Figure 1.22 shows that the precision of the surface is slightly improved as the BH-ratio grows. On the other hand the results from the rotated surface indicate moderate impairment of the precision as the BH-ratio grows.

1.6.3.3 Precision dependence on BH-ratio, not curvature corrected

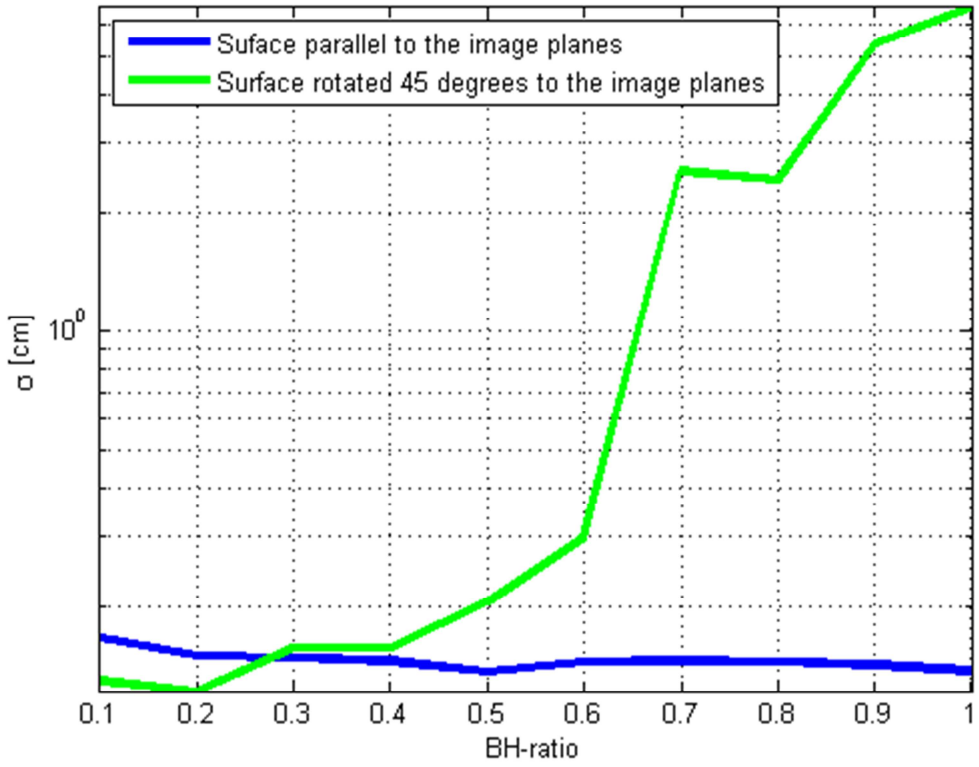


Figure 1.23: The graph shows the standard deviation of the surface measurements as function of the BH-ratio. The results are acquired using a surface that is not curvature corrected, located parallel and with 45 degrees rotation to the rectified image planes.

The most remarkable changes from the corrected results are that the standard deviation for the parallel surface is converging towards 0.1 cm, whereas the corrected version tends to converge towards 0.025 cm. From a relative point of view this is a factor four in difference, so the changes seem to be significant. On the other hand, the rotated surface is not affected as much from a relative point of view, because the deviations in general are larger than those for the parallel surface.

1.6.3.4 Precision dependence on surface angle, curvature corrected

Figure 1.24 shows the standard deviation of the surface measurements as a function of the angle between the surface and the rectified image planes.

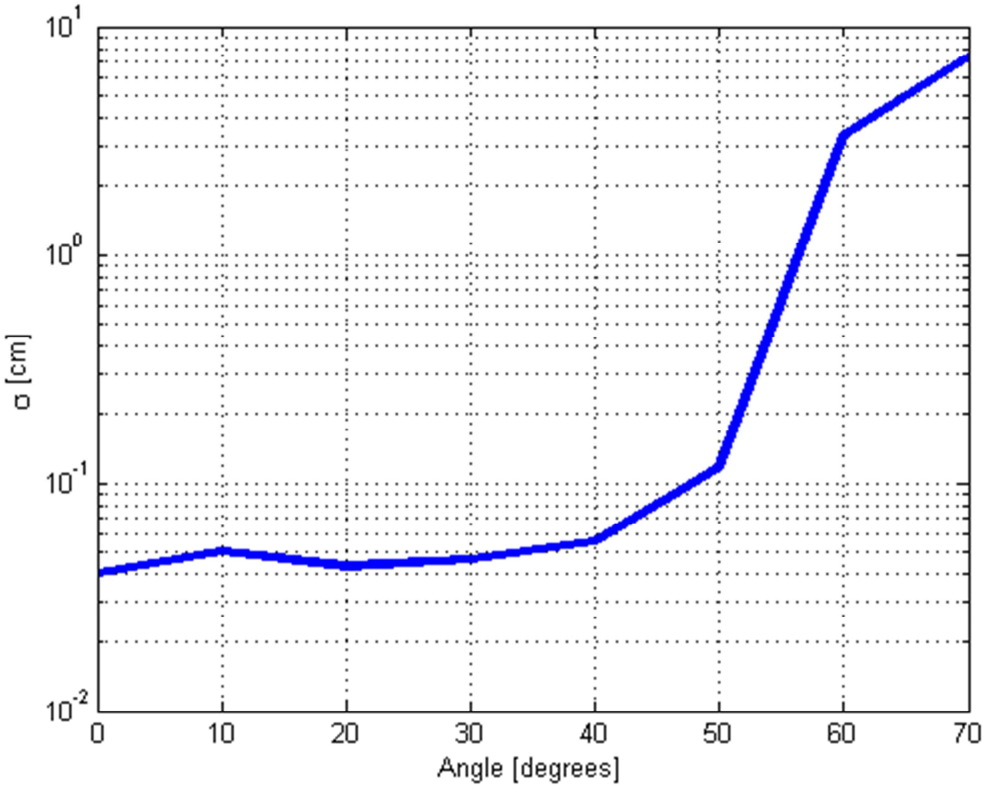


Figure 1.24: The graph shows the standard deviation of the surface measurements as function of the surface rotation. The surface, which was curvature corrected, was located with varying angles to the rectified image planes

The standard deviation is nearly 0.04 cm for surface rotations below 50 degrees. From 50 degrees and above, the standard deviation diverges significantly.

1.6.3.5 *Precision dependence on surface angle, not curvature corrected*

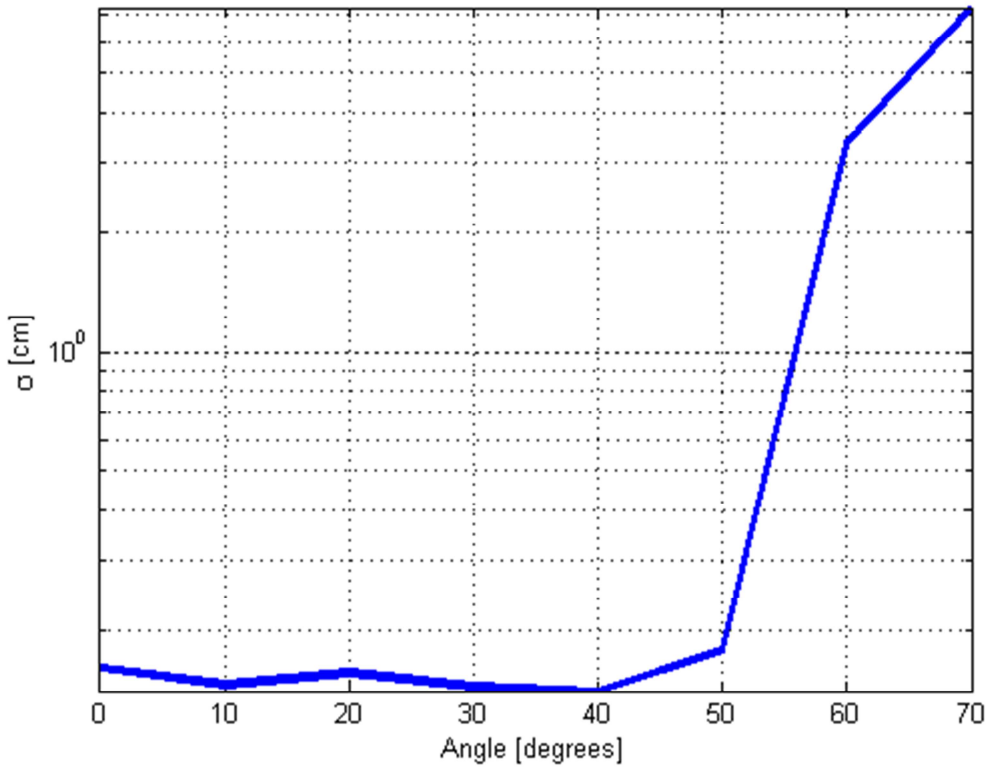


Figure 1.25: The graph shows the standard deviation of the surface measurements as function of the surface rotation. The surface, which was not curvature corrected, was located with varying angles to the rectified image planes

The standard deviation seems to be constant at 0.1 cm, until it diverges at surface angles above 50 degrees. For comparison, the standard deviation of the corrected surface was nearly 0.04 at angles below 50 degrees. The differences between the two standard deviations are significant from a relative point of view. Anyhow, from an absolute point of view, this difference would not provide drastic changes in these conclusions.

1.6.4 Discussion

1.6.4.1 Precision dependence on BH-ratio

Comparing the results from the corrected models with those that are not corrected, it seems of big importance to model out the curvature if the standard deviation should be reasonable quantified.

Considering Figure 1.22 and Figure 1.23 that illustrates the standard deviation as a function of the BH-ratio corrected and uncorrected, respectively, the deviation caused by the curvature is more than three times larger than those caused by the geometry (blue line) for large BH-ratios. However the impact of the curvature is negligible compared to the large errors caused by the correlation errors (green line) for the large BH-ratios and will not bias the fact that the best results are achieved with a BH-ratio between 0.1-0.4 as also concluded in (Hullo, et al., 2009). The standard deviation of the modeled surface within this interval is roughly 1 mm.

1.6.4.2 Precision dependence on surface angle

Considering Figure 1.24, the standard deviation seems to be constant for angles below 50 degrees at about 0.4 mm. From 50 degrees and above, the standard deviation grows up to several centimeters. Surface angles therefore have to be below 50 degrees, which means that a configuration to obtain full body models have to consist of at least four stereoscopic systems to cover 360 degrees.

1.6.5 Conclusion

The graph in Figure 1.22 clearly shows a large correlation between the standard deviation of the model and the surface angle. A standard deviation at roughly 1 mm was achieved for surface angles below 45 degrees and BH ratios below 0.5. This result is competitive to the laser scanner and the camera setup used in (Siebert, et al., 2000) that was discussed in “Previous work”, section 1.3.

However this conclusion is only valid for flat surfaces without either edges or high disparity gradients. In addition the surface has to have remarkable textures, where skin is not among those. Since the curves of the human surface in general are smooth and organic and the texture of clothes for the test subjects can be managed in a gait lab, these tests provide a good approximation of the precision for such model acquisitions.

From the test regarding surface angles, it is concluded that surface angles below 50 degrees in relation to the rectified image planes can be reconstructed with a standard deviation at 1 mm. All considered, when using four stereoscopic systems with a BH ratio between 0.2-0.3, it is possible to obtain a full body model with a precision at 1 mm with optimal texture and illumination.

1.7 Full 3D model acquisition test

1.7.1 Introduction

In the previous tests, it is concluded that four stereoscopic systems are needed to obtain full coverage for an object in 360 degrees. This test will therefore investigate the practical result of such a setup. In addition textiles with different textures are tested as well, to find out how sensitive the setup is to the contrast in the texture.

1.7.2 Test setup

The setup is illustrated below.

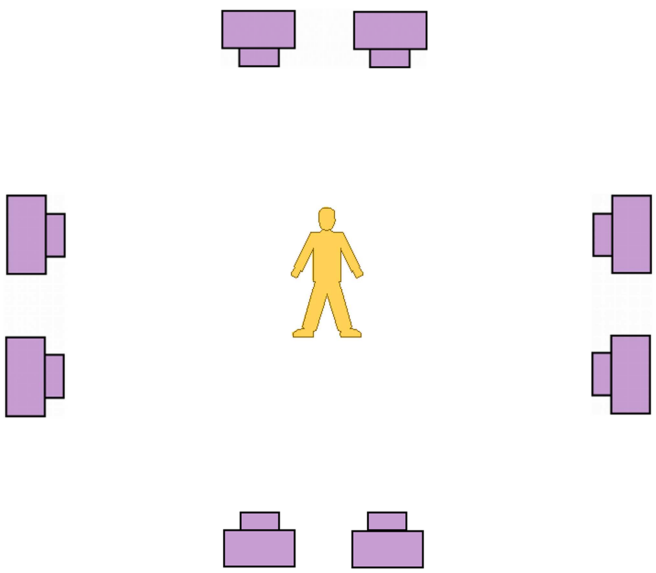


Figure 1.26: Configuration with four stereoscopic systems

1.7.2.1 Materials

- 4 x canonical stereoscopic systems, BH-ratio is 0.2, cameras are mounted on tripods
- 1 x Box coated with coded targets
- 4 x boards with a dimension at 10x80 cm, coated with coded targets
- 1 x Dummy wearing various clothes:
    a. Military uniform representing large random pattern
    b. Checkered shirt with black jeans representing structured pattern and weak random pattern, respectively
    c. Mottled sweater representing small random pattern
    d. Black sport clothes representing no texture

1.7.2.2 PhotoModeler Processing

The data processing in PhotoModeler is performed according to the process description appended in appendix 3.8. The values for the DSM parameters are listed in Table 1.3.

Category	Value
Sampling interval	3.5 mm
Matching region radius	5-7 [1]
Texture type	2-3[2]
Sub-sampling factor	2

Table 1.3: DSM parameters

[1] Matching region radius is adjusted to optimize the result according to the applied texture in the cloth
[2] Texture type is adjusted according to the applied texture in the cloth

Triangulation is performed to create a shaded surface. This is performed to make the visual inspection of the result easier. In addition the results is obtained using PM's automatic removal of outliers and smoothing operation with default options, which is presumed to be fully legal, since the curvature of the body is relatively low-frequency. No hole filling is applied in the models, to make it easier to spot poorly recovered areas.

1.7.3 Results

The head of the dummy is not included in the models since the texture of the dummy and the light conditions was too weak to get a usable result. However a test of human face models with varying BH-ratios is performed and evaluated in appendix 3.11.3.

Exact extrinsic calibration has shown to be very important. Insignificant calibration results in unsatisfactory fits of the shells obtained from each stereoscopic system. An example of insignificant extrinsic camera calibration is illustrated in section 3.14 in appendix.

An example with a naked human upper body is illustrated in appendix 3.13.2, to illustrate how skin performs as texture.

1.7.3.1 *Dummy textured with large random pattern*

Figure 1.27: Left: Photo; Center: Triangulated mesh without processing[3]; Right: Processed triangulated mesh

Figure 1.27 shows a good recovery of the uniform. Only small holes are found on the smoothed model. The holes are mostly observed at the largest angles from the stereoscopic systems, which indicate that fewer stereoscopic systems would not have been able to handle the task. The folds in the jacket and the pants are also modeled. Visual hull models would not be able such details, since visual hull models cannot cope with concave structures.

More angles of the processed model can be found in section 3.15.1 in appendix.

[3] Processing constitutes of smoothing and removal of outliers

1.7.3.2 *Dummy textured with structured pattern and weak random pattern*

Figure 1.28: Left: Photo; Center: Triangulated mesh without processing[4]; Right: Processed triangulated mesh

Figure 1.28 also shows a convincing result. The degree of details is as good as that of the uniform. The large hole on the shoulder is difficult to explain, but might be a consequence of a correlation with multiple optimums due to the structured pattern. The jeans are marked by large holes, especially in regions where the shadow falls. This clearly illustrates the importance of a powerful and uniform illumination of the object. A powerful illumination might even compensate for a weak texture.

More angles of the processed model can be found in section 3.15.2 in appendix.

[4] Processing constitutes of smoothing and removal of outliers

1.7.3.3 *Dummy textured with small random pattern*

Figure 1.29: Left: Photo; Center: Triangulated mesh without processing[5]; Right: Processed triangulated mesh

Figure 1.29 shows a satisfactory result of the sweater and a poor result of the jeans. Again the large holes are found in the shadow regions. The folds in the sweater, especially around the hood, are finely modeled.

[5] Processing constitutes of smoothing and removal of outliers

1.7.3.4 Dummy textured in black (no texture)

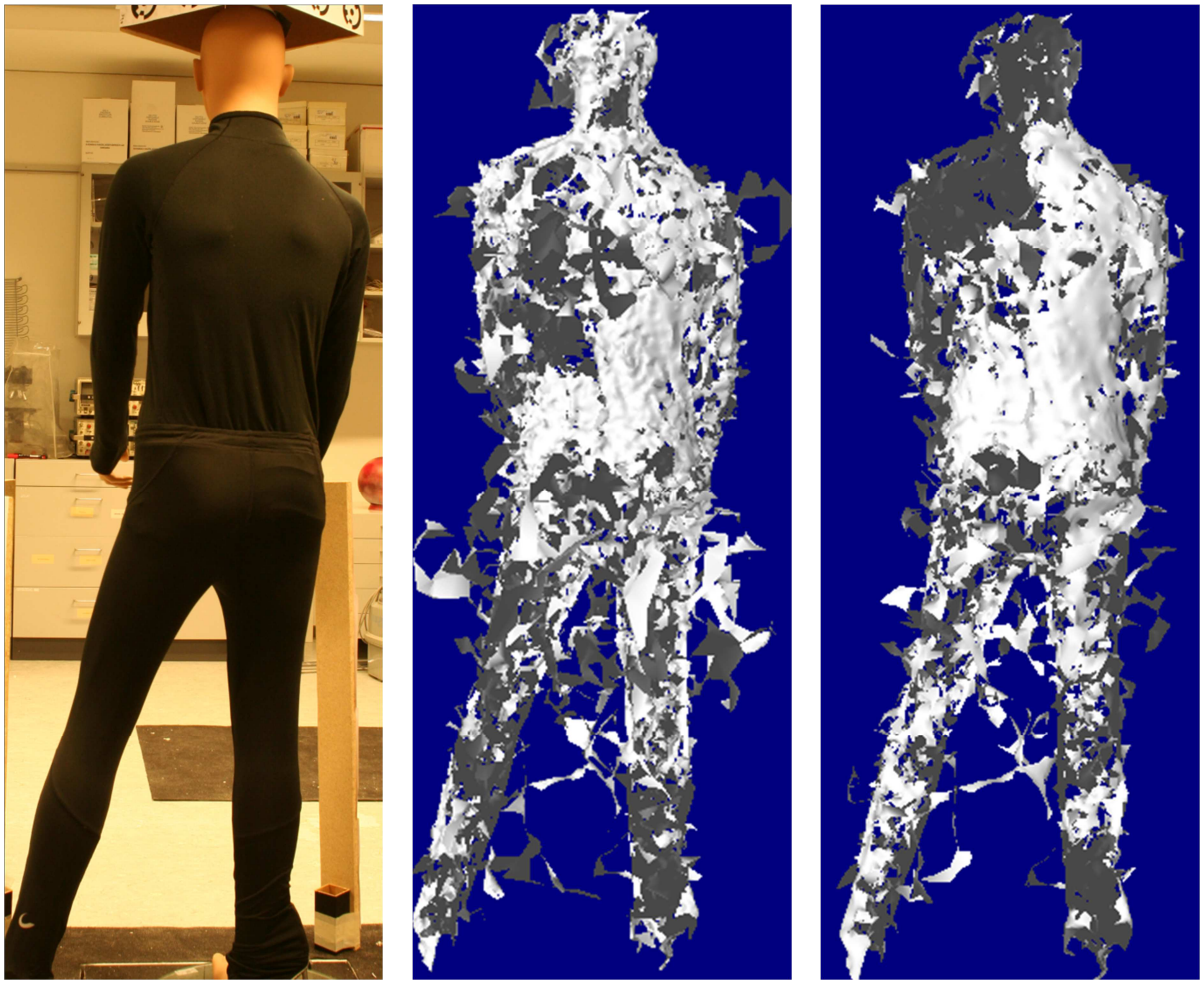


Figure 1.30: Left: Photo; Center: Triangulated mesh without processing[6]; Right: Processed triangulated mesh

Figure 1.30 clearly shows the worst result of all the models. The textureless clothes are very a bad choice when it comes to DSM models.

[6] Processing constitutes of smoothing and removal of outliers

1.7.4 Discussion

It is difficult to conclude which of the clothes that gives the best result. The mottled sweater, providing small random patter with high contrast seems to work as good at the structured pattern and the uniform. However it is clear that textureless clothes is unusable for the purpose and weak texture in general should be avoided even though the texture can be enhanced by powerful uniform illumination.

The solution with four stereoscopic systems seemed to work well for the purpose. Small holes in the regions with wide angles indicated that fewer cameras could not achieve a satisfactory result.

1.7.5 Conclusion

The results have shown that it is possible to make a full body model with a spatial resolution at 3 mm. The precision of the model is highly dependent on the light conditions and the texture on the object. Recalling the precision tests it was concluded that a standard deviation down to 1 mm could be achieved with a proper choice of illumination and texture. However large holes due to occlusions and large angles might contribute to a significantly larger standard deviation when “hole filling” is applied.

1.8 Summary

The purpose of Part I, was to test whether or not it is possible to replace a model, acquired by a 3D laser scanner, with a model created by this software in combination with a proper camera setup.

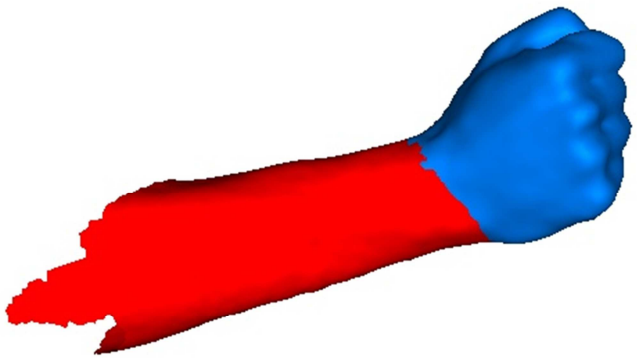
The standard deviation of a modeled surface has been tested with various BH-ratios and surface angles, to find out the optimal BH-ratio and the precision of the reconstructed models. By this it was concluded that a surface could be reconstructed with a sampling density at 3 mm and a standard deviation at 1 mm. This result is just as precise as the 3D laser scan used to create the articulated models in (Corazza, et al., 2007). However this is only valid for flat surfaces, providing a constant distortion gradient, as assumed in most correlation based point matching algorithms. Surface angles above 45 degree were too noisy to be modeled. This led to a suggestion of four stereoscopic systems (8 cameras) to recover 360 degrees in case of a full body model.

In the second test a configuration of four stereoscopic systems was tested to observe if such a configuration was sufficient to recover a full body model. Different textures were also tested to find the limits of good and poor texture. The test showed that both texture and illumination are of great importance in order to get satisfactory results. Both structured and random patterns seemed to work well as long the contrast in the pattern was strong. A quantified estimation of the standard deviation was not achieved in this test, so a visual inspection of the models combined with the previous precision test of the surfaces are considered as documentation for the claim that this configuration is suitable for a full body model acquisition with a precision close to a laser scanner.

Other tests were performed, concerning the spatial resolution and reconstruction of the face. These tests are appended in section 3.11 in appendix. The results of the tests illustrated in appendix showed that human skin is very difficult to model. Based on the face reconstructions it was concluded that the precision was insufficient to reach the criterions set in the problem statement, due to texture and the non-fronto-parallel surfaces. The requirements of a high spatial resolution obstruct the use of VGA cameras, which are often applied in gait lab configurations.

8 x 10 mega pixel SLR cameras or higher pixel resolution is necessary to get a result that can compete with a laser scanner.

However the results of the visual hull models presented in (Mündermann, et al., 2005) is assumed to be very optimistic in practice. In that case it might be sufficient to use cameras with 5 megapixel or maybe even 2 megapixel cameras as well.



Part II: Pose estimation

2.1 Introduction

Moeslund et al. presents four primary functionalities of motion capture processing in (Moeslund, et al., 2001): Initialization, tracking, pose estimation and recognition.

Initialization embraces the establishment of an articulated model if the approach is model based. Tracking implies segmenting the subject by background subtraction and reconstruction of the subject in each frame. Pose estimation corresponds to extracting the poses of the joint centers and limb segments that often serves as output to the system. Finally, recognition refers to the analysis and post processing of the output parameters, such as identifying a subject by gait.

Moeslund et al. also presents three approaches for pose estimation: Model free, indirect- and direct model use. The model free approach covers methods where there is no priori model. The indirect models approaches use a priori model in pose estimations as a reference or look-up table to guide the interpretation of measured data. The direct models approach uses articulated models of human shape to reconstruct pose.

As mentioned in the introduction section of this master thesis, Sylvia Yang claims that model free approach does not show significant results for biomechanical approaches. Even though the approaches using articulated models directly have shown the most promising results, there are several problems related to this method. First of all, fitting one standard shape to all kind of human bodies, leads to errors, since the human shapes differs significantly between individuals. Secondly, providing a subject specific articulated model can be time consuming to obtain. In addition it might also be expensive to purchase the equipment to obtain the full body model with such as a laser scanner.

To solve these problems it might be possible to develop a system that uses the tracking data to develop an articulated model or even skip the articulated model by estimating the joint centers directly on the tracking data.

PM is not an option for the tracking step, since PM is based on a graphical user interface, which requires several manual processing steps. However many other algorithms are developed to create dense surface models. (Scharstein, et al., 2009) have compared some of the most promising

approaches where (Esteban, et al., 2004) and (Furukawa, et al., 2010) seems to be among the most accurate. Both approaches are reviewed in section 3.7 in appendix.

Y. Furukawa and J. Ponce have used their own Patch based Multi-View Stereo algorithm (PMVS) to provide a markerless motion capture system themselves in (Furukawa, et al., 2007). The novelty of their work is that they track the surface of a polyhedral mesh model (provided by a robust 3D reconstruction algorithm) that is capable of estimating the motion of the points in the mesh as illustrated in Figure 2.1. These results are obtained with eight color VGA cameras.

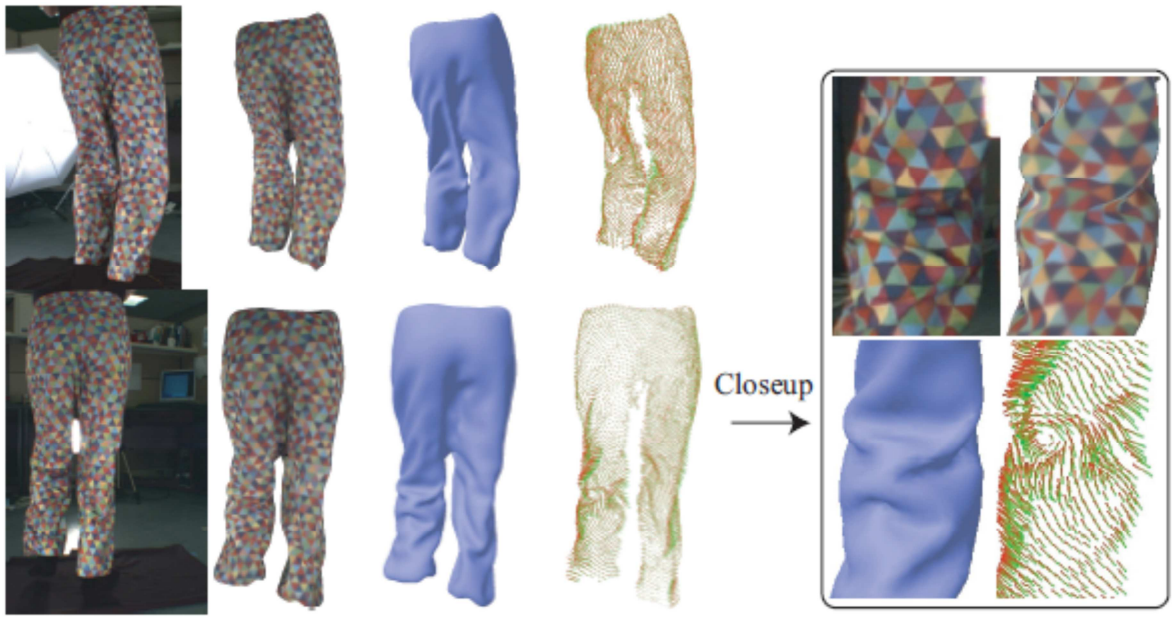


Figure 2.1: From left to right: Images acquired from the cameras, textured model, shaded triangulated mesh, motion field. Reference: (Furukawa, et al., 2007)

In the approach a model is synthesized from the first frame using the PMVS algorithm. The patches of this model are then tracked over the following frames by an iterative process that extrapolates and deforms the patches from one frame to another.

Unfortunately this approach has no biomechanical applications since it is not capable of locating joint centers and measuring joint angles. The source code does not seem to be publicly available as the PMVS algorithm is. The PMVS algorithm is protected by GNU General Public License that is intended to guarantee the freedom to share and change all versions of the program for all users.

2.2 Problem statement

This part of the report will focus on the pose estimation step of the four functionality steps presented in (Moeslund, et al., 2001).

Since the PMVS algorithm seemed to provide some of the best results among those who are presented in (Scharstein, et al., 2009), it is reasonable to perform the tracking step using this algorithm and to find an approach for pose estimation that fits the suggested tracking method.

It is aimed to find a suitable method to perform the pose estimation without the establishment of an articulated model prior to the pose estimation.

The following section reviews the previous work for estimation of joint centers on a model. The previous work provides inspiration to find a pose estimation technique that is well suited for this approach.

2.3 Previous work

As mentioned previously the majority of the motion capture systems use articulated models to register to the tracking data, represented by a point cloud interpretation of the subject or similar. The articulated model proposed in (Corazza, et al., 2006), used the joint constraints as described in section 3.16 in appendix to decrease the degrees of freedom of the model and achieved promising results by this. This is also the case for the articulated model proposed in (Cheung, et al., 2005). Anyhow the articulated model in this approach is obtained by performing a refined visual hull (VH) model of the subject using eight color cameras with VGA pixel resolution, instead of a full body laser scan. The articulated model acquired by this approach is illustrated in Figure 2.2.

Figure 2.2: Articulated model obtained in (Cheung, et al., 2005)

The approach used by (Cheung, et al., 2005) to estimate the joint centers of the articulated model is quite interesting. Using the temporal information about the relative positions of the points in the VH model, they find the rotation center between two limb segments as illustrated in Figure 2.3.

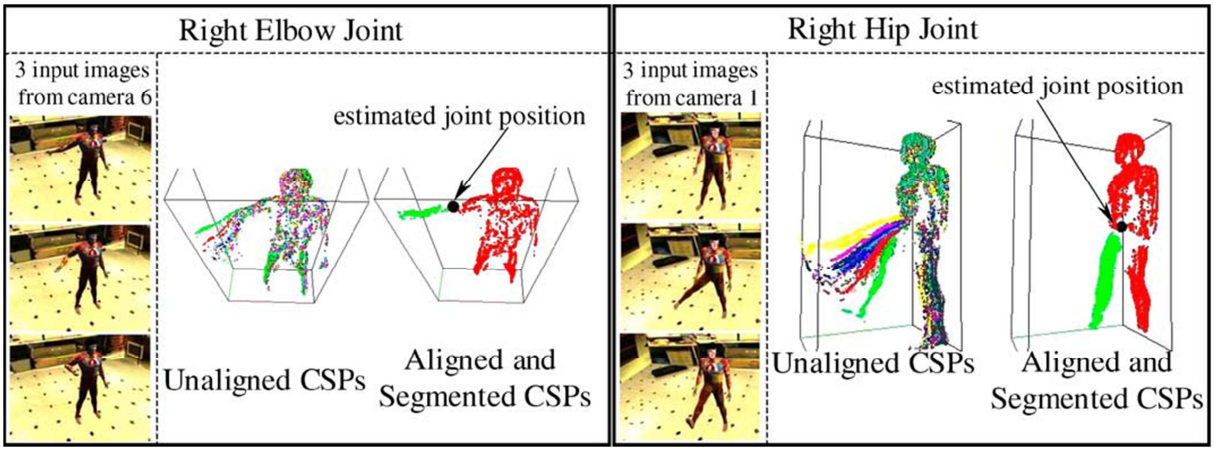


Figure 2.3: Joint center estimation proposed in (Cheung, et al., 2005)

The disadvantages of this method is that only one joint can be considered at a time, which makes the method comprehensive and time consuming, and that the errors provided by this method has an average at 26 mm with maximum error at 47.21 mm.

However, the concept regarding localization of the joint centers by using the temporal information is interesting to consider for a model free approach.

Another approach to extract the joint centers is presented in (Lien, et al., 2005). This proposed method is based on skeletonization using a convex approximation segmentation approach. Based on Figure 3.36 of the human skeleton, one can observe superficial convexities and concavities that might provide sufficient information to make a segmentation of the body into the limbs of interest. However skeletonization and extraction of joint centers based on such methods solely, does not seems to provide satisfactory results for the ankles and hips in particular. (Lien, et al., 2005) and (Katz, et al., 2005) seem to describe some of the most promising methods to provide segmentation using convex approximation. However the segmentation algorithms seem to be sensitive to varying body positions and might give even worse results for adipose people or people with muscular diseases such as muscular dystrophy. Problems might also appear if the subject wears loose cloth that covers the anatomical curvatures.

The following sections start with a brief review on introductory biomechanics to get an understanding of the practical use of a motion capture system for biomechanical purposes. Then the proposed method will be presented and validated to test how the method performs in practice.

2.4 Introductory biomechanics

2.4.1 Human gait

The human gait cycle is divided into two sub phases: A stance phase and a swing phase for each leg. The stance phase embraces approximately 60% of the cycle where the particular leg is loaded. It starts with the heel strike and ends when the toe is off the floor. The swing phase embraces the remaining 40% of the gait cycle lasting from toe off floor to heel strike. One gait cycle is illustrated in Figure 2.4.

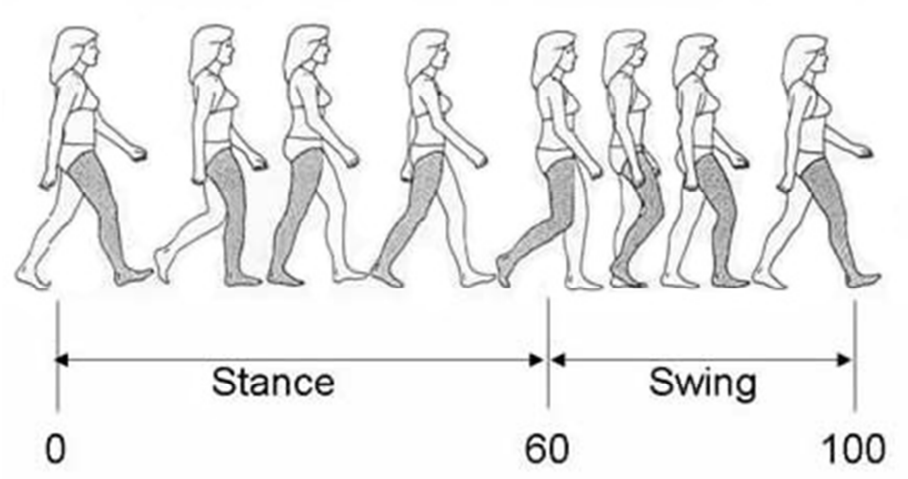


Figure 2.4: Gait cycle Reference: (Wikinoticia.com, 2010)

The features of interest in the gait are, for a forensic point of view, the kinematics of the limbs concerning: the hips, knees and ankles primarily, but also elbow and shoulder could be of interest.

2.4.2 Kinematics

Motion capture systems provide information that can be translated to kinematics for the whole body. To translate the information, it is necessary to segment the reconstructed model into individual limb segments.

The approach when segmenting a body is not standardized. However it is common to divide the body into 12 segments:

- Feet (2 segments)
- Shanks (2 segments)
- Thighs (2 segments)
- Trunk (1 segment)
- Head (1 segment)
- Upper arms (1 segment)
- Forearms and hands (2 segments)

In most common circumstances the arm movements are ignored. By then the head, arms and trunk are considered as one segment (the HAT segment).

A full description of the kinematics of a single segment in 3D requires 15 variables:

- Position of “Center of Mass” (CM) (3 DoF)
- Linear velocity of CM (3 DoF)
- Linear Acceleration of CM (3 DoF)
- Angle (2 DoF)
- Angular velocity (2DoF)
- Angular acceleration (2 DoF)

The model is often simplified to 2D by analyzing the movements in the sagittal plane only. This decreases the DoF from 15 to 9. The angles of interest are the angles spanned by the segments, located in the joint centers. Velocity and acceleration are only relevant if estimation of joint forces is desired. These parameters are not the focus of this project, but it is aimed to develop a system that has the ability to extract such information as well.

The position of CM or center of gravity (CG) is usually calculated using anthropometric table values, describing the location in relation to the joint centers proximal or distal to the segment (Vaughan, et al., 1999). Figure 2.5 shows an example of CG estimation of the thigh using the joint centers.

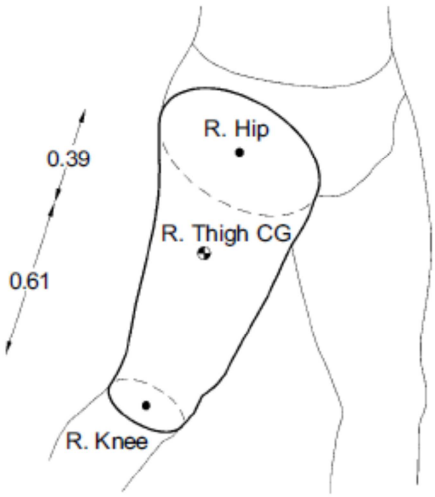


Figure 2.5: Center of Gravity estimation by using the location of the joint centers. Reference: (Vaughan, et al., 1999)

The linear velocity and acceleration is calculated by the first and second derivatives of the position, respectively.

Force platforms are often integrated in gait labs. By using the platforms it is possible to combine the action reaction forces from the floor and the measured kinematic variables for the segments, to calculate the forces in each joint using inverse dynamics.

2.5 Problem analysis

In the problem statement it is stated that this thesis will be restricted to concern the pose estimation step regarding Thomas Moeslunds four primary functionality steps. Inspired by the previous work, it was concluded that limb segmentation could be performed either by using the curvature of the body or by using the temporal information about the model deformation. Among the evaluated approaches, the approach using the temporal information seems to be most promising, since it would be more applicable to handle subjects with clothes that are not slim. To perform such segmentation, the following problems needs to be solved:

- How to track points or segments of the model from one frame to another?
- Assuming the limbs are rigid, how can points be divided into clusters of rigid segments?

The rest of this thesis is reserved to answer the two questions. Answering the questions will formulate a new approach that can be considered as part of pose estimation.

The approach will consist of two steps:

- Point tracking
- Limb segments

Whether the method can be used in practice or not will be tested through the above mentioned two steps, using modeled parts of the body.

From the segmented models, it is possible to find the rotation center between two segments using the method presented in (Cheung, et al., 2005). However joint estimation will not be considered any further in this report.

2.6 Point tracking

Tracking one point of a reconstructed model obtained from frame n to a point in another reconstructed model obtained from frame n+1, can be considered as a model registration problem. The objective of registration is to find a transformation function that describes how one point in a template model is transformed into a reference model. This statement can also be written as a least squares minimization problem, where the following statements have to be minimized:

$$D_{SSD} = \sum_{n \in \Omega} \left(T\left(y(x_n)\right) - R(x_n) \right)^2$$

Equation 2.1: Sum of squared differences

Here $y(i)$ is the transformation function, $T(j)$ is the template model, $R(j)$ is the reference model and x_n is the n[th] point that belongs to the space Ω, in which the dissimilarity is calculated.

Several problems are related to point matching between the template model and the reference model. First of all it cannot be assumed that the template and the reference have the same number of points and points might even not have a correspondence, since holes in the models can appear. Second the deformations caused by the body movements will provide nonlinearities that cannot be approximated sufficiently by linear approaches.

The nonlinearities prevent a direct implementation of the Iterative Closest Point algorithm (ICP), since it is only capable of handling six degrees of freedom, corresponding to a rigid transformation.

The first problem can be solved by setting regularizations on the deformations. (Belongie, et al., 2002) proposed a method to cope with such problems in 2D by using shape contexts and smoothing TPS. This approach has since then been expanded to 3D by Di Xiao in (Xiao, et al., 2009). Here he used his proposed method for small animal skeletons that provided promising results.

2.6.1 Point matching using shape contexts

2.6.1.1 Creating shape contexts

Shape context has many similarities to point matching using log polar space as described in section 3.2.4 in appendix. For each point in a point cloud a log polar histogram with a limited number of bins is created for each point. The histogram is illustrated in Figure 2.6 (c). Each bin represents a region in space. The number of points represented in a region are summed and mapped into the particular bin. This is illustrated for three points in Figure 2.6 (d), (e) and (f), where the histogram is a squared interpretation the circular histogram in (c).

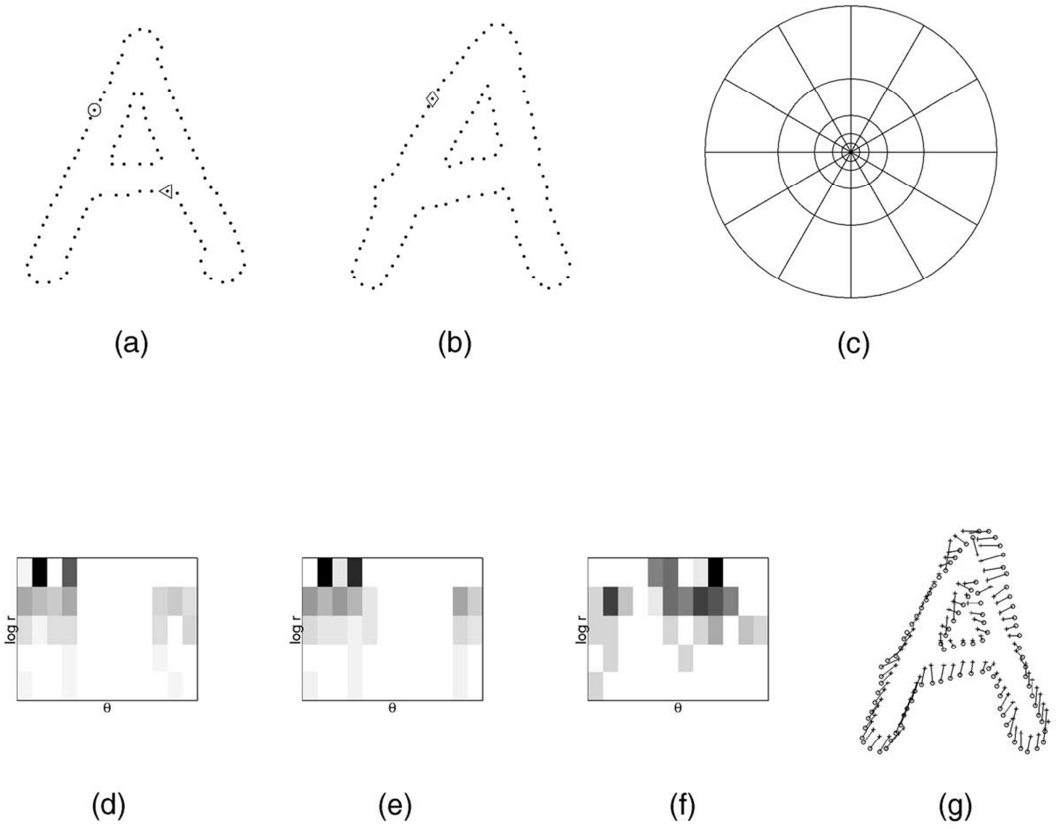


Figure 2.6: (a): Reference shape; (b): Template shape; (c): Diagram of log-polar histogram bins used in computing the shape contexts; (d), (e) and (f): Histogram for the points encircled by a circle, a square and a triangle respectively in (a) and (b); (g): The reference shape and template shape on top of each other. The lines indicate the calculated point correspondences. Reference: (Belongie, et al., 2002)

Since the first presentation of shape contexts in (Belongie, et al., 2002), the approach has been developed to 3D first by (Kortgen, et al., 2003) and later by (Xiao, et al., 2009). In 3D the shape contexts are termed in spherical coordinates as illustrated in Figure 2.7.

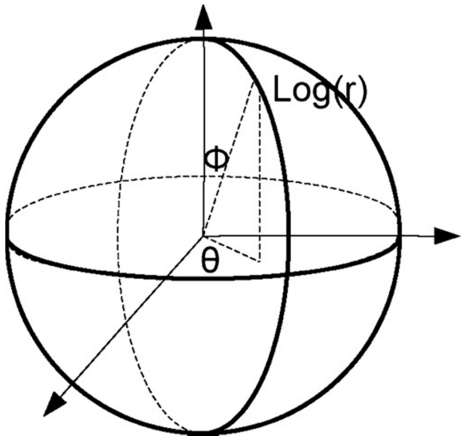


Figure 2.7: 3D spherical coordinates for 3D Shape context

The corresponding histogram is defined by a 3D matrix, where each dimension represents Log(r), θ and ϕ respectively.

2.6.1.2 Finding point correspondences using shape contexts

A cost matrix is created, in which the columns represent the points of the template model and the rows represent the points in the reference model. Each element represents the difference of the shape context histograms between the points. Denoting $h_i(l)$ the histogram function of the i'th point and l'th represents the bins of the histogram, the cost is traditionally calculated as:

$$C_{mn} = \frac{1}{2}\sum_{l=1}^{L}\frac{(h_m(l) - h_n(l))^2}{h_m(l) + h_n(l)}$$

Equation 2.2: Cost function proposed by (Belongie, et al., 2002)

However (Xiao, et al., 2009) has proposed an expansion of the cost function, taking the curvature in the points into account:

$$C^e_{mn} = \alpha C_{mn} + (1-\alpha)C^c_{mn}$$

Equation 2.3: Expansion of the cost function proposed by (Xiao, et al., 2009)

Here C^c_{mn} is the cost function of the difference between the curvatures of the 3D points, and α is a weight constant.

The mean curvature in an arbitrary point is calculated numerically by the approach presented in (Meyer, et al., 2000). For this purpose a MATLAB function has been created and tested to see if the differences resulted in significant improvements. Results and discussion of integrating the mean curvature into the cost function are located in section 3.16 in appendix. Unfortunately the expansion

was found to have a negative impact on the registration, due to pronounced fluctuations in the surfaces of the acquired models. Due to that, α was set to one in the experiments.

Using the Hungarian algorithm that is described in section 3.17 in appendix, the point correspondences that that minimizes the total cost can be found.

Both the calculation of the costs matrix and the Hungarian algorithm are computational heavy. The obtained models usually consist of 10^6 points that implies cost matrices with 10^{12} elements. Assuming that each element in the matrix is of type double in MATLAB (64 bit), memory in sizes of terabytes are needed to be dedicated, which only custom made desktops are able to support today. However the largest problem is related to the computation time of the optimization using the Hungarian algorithm. Down sampling is therefore needed.

The numbers of points are reduced to around 10^3 to keep a computation time below one hour and the data types are converted into single (32 bit) to reduce the amount of data. The conversion can be performed without any significant loss of precision, since single types makes it possible to operate with precision at 10^{-7} meters in MATLAB.

Once the point correspondences are found, the smoothing TPS can be used to register the entire template shape into the reference shape.

2.6.2 3D Registration using smoothing TPS

Since mismatches might appear from the shape context matching, it is common to regularize the transformation by a smoothing TPS. The basics of the smoothing TPS is presented in section 3.6 in appendix.

Just alike the 2D implementation of the TPS, the 3D interpretation creates a function that calculates the intensity in an arbitrary 3D point, regularized by a smoothing term. To use the smoothing TPS for point registration, a smoothing TPS for each dimension has to be created. By then the intensity parameter will reflect the dislocation between the point correspondences along each dimension.

To obtain a better approximation of the point correspondences, it is possible to iterate between point matching and registration. Most commonly there will be a similarity measure to control the number of iterations. The number of iterations is hardcoded to obtain full control of the process. A diagram of the full implementation of the registration algorithm is illustrated in Figure 2.8.

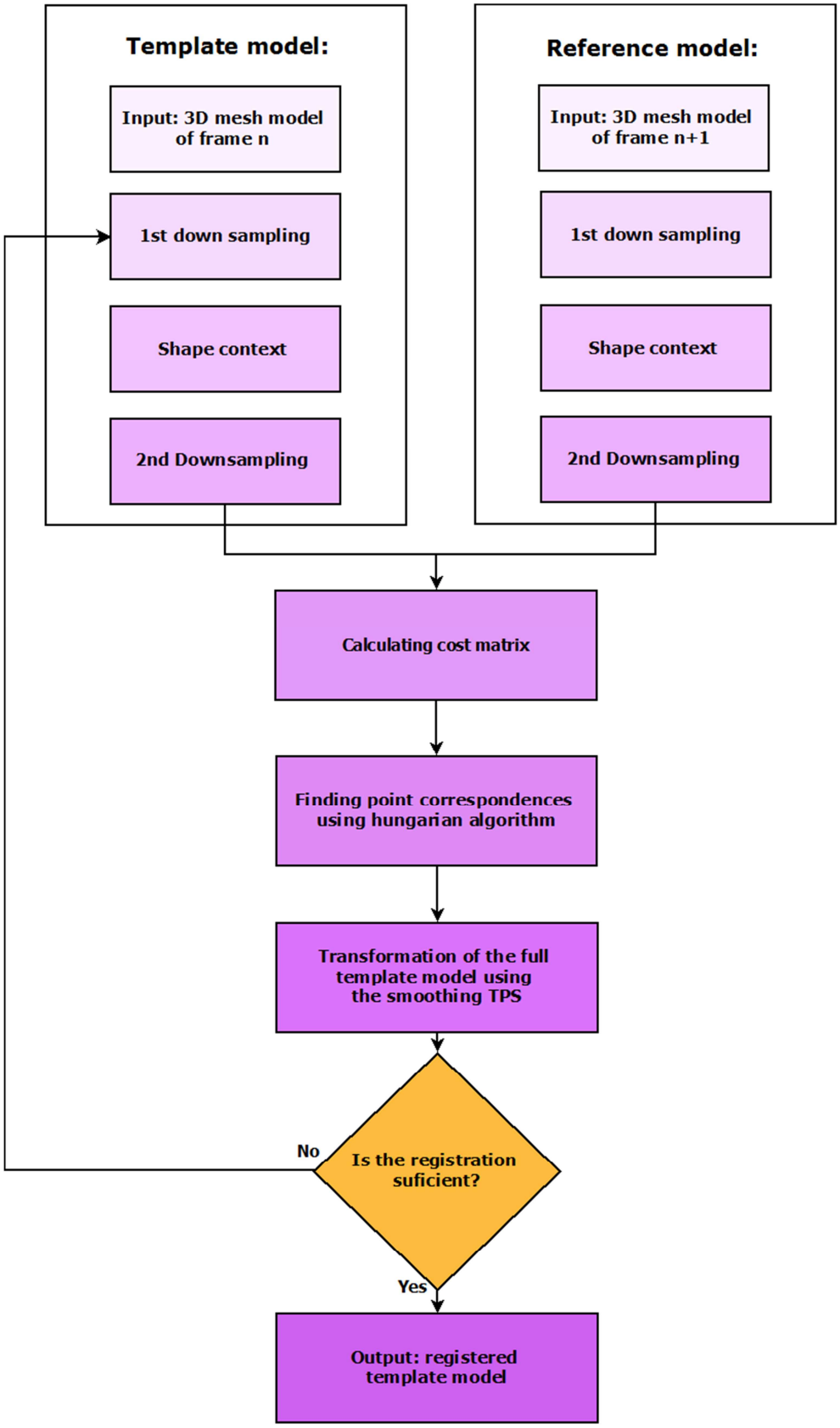


Figure 2.8: Diagram illustrating the registration algorithm

2.6.3 Quantitative error estimation

Recalling the problem statement of part I, (Mündermann, et al., 2005) uses the shortest distance between two models to estimate the error. This is sufficient when the models are very similar, but the similarity measure is sensitive to large variations. As an example, a model of the arm flexed into two positions is illustrated in Figure 2.9.

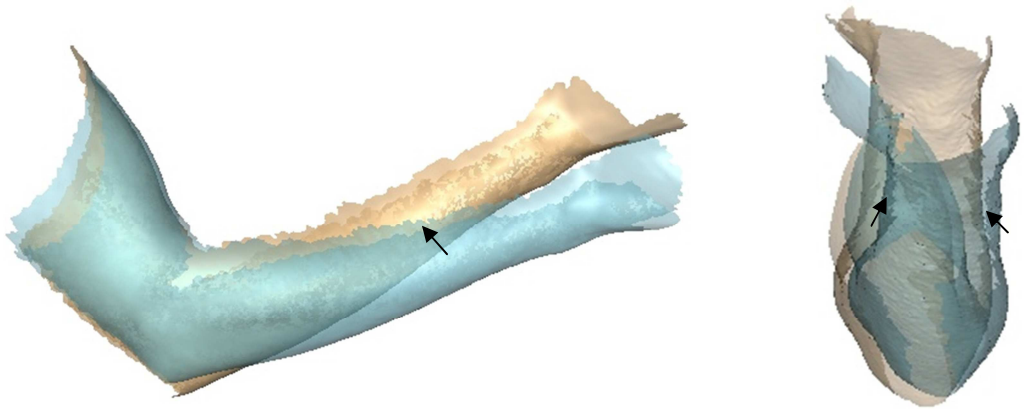


Figure 2.9: Model of an arm, flexed into two positions. Lateral view

When mapping the shortest distance from a point on the golden model to a point on the cyan colored model, the shortest distances of the points on the distal part of the arm implies a match to the points along the edge on the cyan model labeled with arrows (see Figure 2.9). As illustrated in Figure 2.10, the cyan stripes along the forearm, indicating a low error, are clearly caused by the mismatching to the upper edges of the other model.

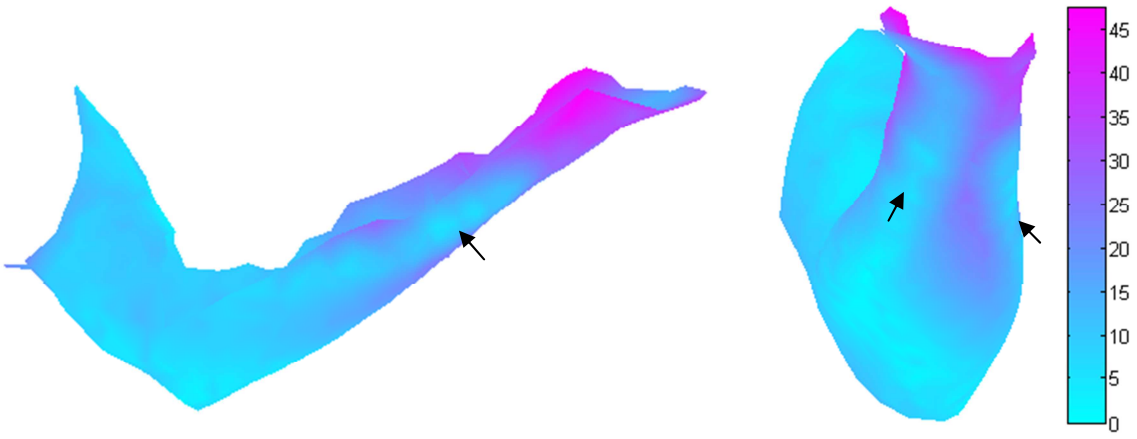


Figure 2.10: Deviation mapping of the golden model, based on the shortest distance. Cyan indicates small distance to the corresponding model, whereas magenta indicates large distance. Distance values on the color bar are labeled in millimeters.

Due to this kind of mismatching, this error estimation is not found to be optimal for our purpose. It is therefore proposed to use following approach:

Using a cost matrix where the cost of a point match is equal to the Euclidean distance, it is possible to use the Hungarian algorithm to find better matches and still use the distance between the points to estimate the error. However due to the limitations of memory and the large computation time, the points have been down sampled to thousand points per model.

This approach assumes that all points in one model have corresponding points in the other model. This might not be true, especially when down sampling to a thousand points. In case of holes in one of the models, this approach might provide even worse error estimations.

The results of estimating the distances to their correspondences and thereby the error is illustrated in Figure 2.11.

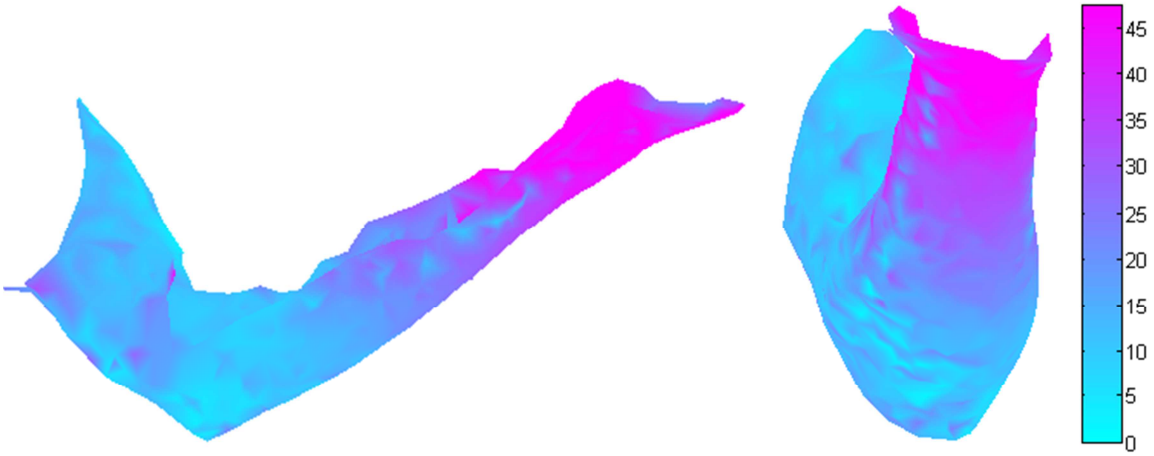


Figure 2.11: Deviation mapping of the golden model, based on the Hungarian algorithm. Cyan indicates small distance to the corresponding model, whereas magenta indicates large distance. Distance values on the color bar are labeled in millimeters.

According to the new deviation map, the cyan stripes along the forearm are not as significant as before. The error in general in the distal end of the arm is significantly higher as well. However it seems like the mapping has become more fluctuating. These fluctuations are caused by mismatches and provide in general a higher error than the shortest distance approach.

It is presumed that the shortest distance approach will provide the best estimation of the error in most circumstances. But when the differences of the two models are large, the approach using the Hungarian algorithm might provide the best estimation.

The deviation mapping is also illustrated for a case where the registration of a hand was failed in section 3.21.1.5 in appendix.

Due to the different strengths of the two approaches, the error is quantified by the Root Mean Square (RMS) value for both estimations. These estimation approaches will be denoted as the shortest distance- and the Hungarian approach throughout this thesis.

RMS is well suited to estimate the error, since it is a measure of the variability of the difference as the standard deviation.

When the error between the registered temple- and the reference model is sufficiently small, it should be possible to use the template model and the registered template model to find the rigid segments by segmentation. The segmentation process will be presented in the following section.

2.6.4 Experimental results

A setup with minimum eight cameras has not been available to record the data for the 3D reconstruction. The 3D reconstructions are therefore obtained in the 3D scanner in 3D lab at the Panum institute. The field of view of the camera configuration was decreased to the size of a head, which limited the choice of limb to be reconstructed in this thesis. An arm in various positions has therefore been used to test the approach. Just as the golden standard scans from the experimental results in part I, the scan technique is based on a stereoscopic system supported with random pattern projection. All equipment used for this process is provided by the company: 3DMD.

Three recorded examples are evaluated in this results section while an insignificant registration is evaluated in section 3.21.1 in appendix. However the insignificant registration is still considered as part of the total results.

2.6.4.1 Registration of a flexing arm

The arm is obtained in two different positions to simulate movements from one frame to another. Since the scanner was configured with only two stereoscopic systems, a whole model of the arm could not be obtained. The acquired models of the arm are illustrated in Figure 2.12 and Figure 2.13 in a shaded format. The representations are created using Sumatra, a software provided by Rasmus R. Paulsen, Associate Professor at DTU.

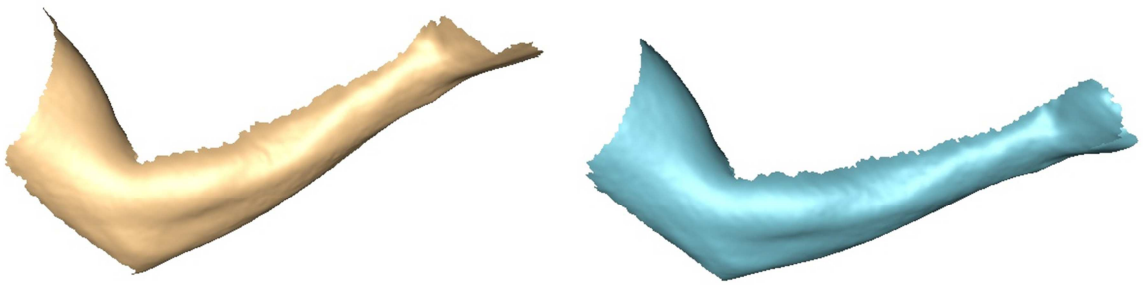


Figure 2.12: Left: Reference model, lateral view; Right: Template model lateral view

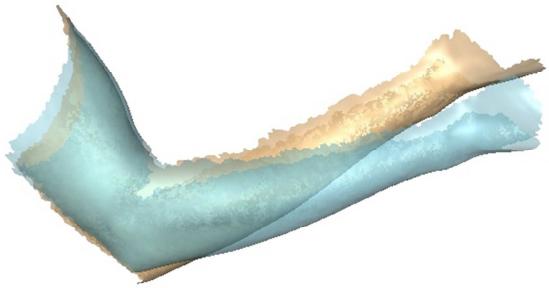


Figure 2.13: Template (cyan)- and reference (golden) models plotted on top of each other, lateral view

The registration was performed using MATLAB. The total processing time to achieve the results shown below through one iteration was 41 min and 24 sec, for a HP pavilion, 4 GB RAM and Intel core i5 2. 53 GHz CPU in windows 7.

2.6.4.1.1 Point matching using shape contexts

To obtain flexibility in the point match, the template was down sampled to 362 points out of total 45,212 points, where the reference model was down sampled to 1798 points out of total 44,939 points. This implies more freedom in the point match than if the template- and the reference model consisted of the same number of points.

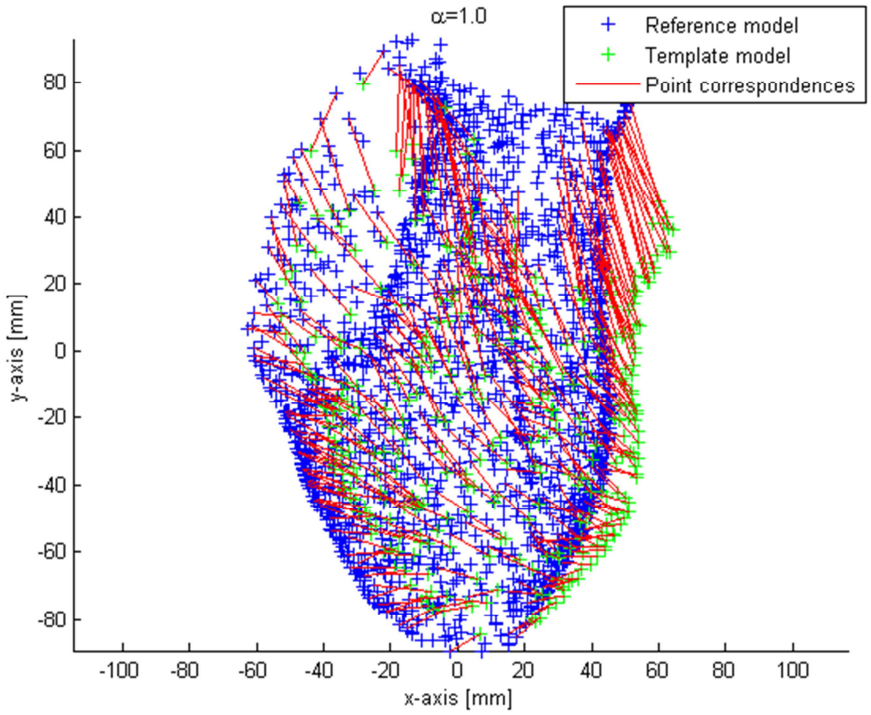


Figure 2.14: Plot of the point correspondences between the reference model plotted and the template model, frontal view

The 3D point correspondences are shown in 2D in Figure 2.14, which seems quite confusing. However the important features in Figure 2.14 are the red lines, representing the point correspondences. The lines should be nearly parallel to the neighboring correspondences. Crossing lines might indicate mismatch.

Figure 2.14 shows that the majority of the correspondences are nearly parallel to their neighbors, indicating a fairly good registration. However small mismatches seem to appear, which will disturb the registration.

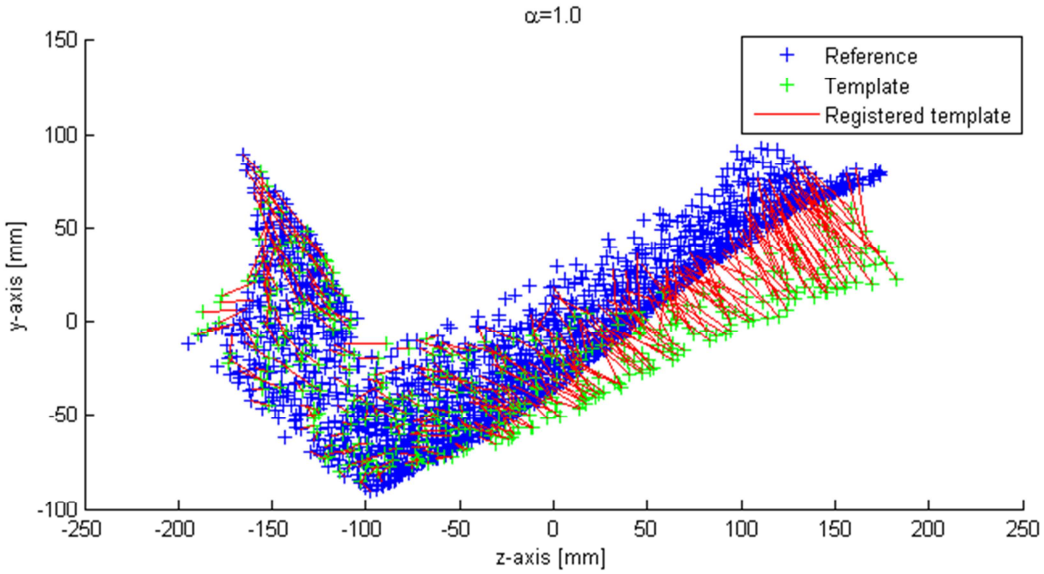


Figure 2.15: Plot of the point correspondences between the reference model plotted and the template model, lateral view

The same indications seem to appear in Figure 2.15, representing a lateral view of the arm. More crossovers can be observed from the lateral view than from the frontal view. To inhibit unnatural folds provided by the mismatches in the registered template model, the Degrees of Freedom (DoF) of the TPS have been regularized.

2.6.4.1.2 Registration using smoothing TPS

For the smoothing TPS registration, a smoothing parameter λ equal to a DoF of N-1 has been used. N corresponds to the number of knots (DoF=361 in our case). Since the number of knots corresponds to the maximal allowable degrees of freedom for the spline, the matching achieves nearly maximal flexibility. However a DoF at N-1 has shown to be sufficient to inhibit the unnatural folds in the model as illustrated in Figure 2.16

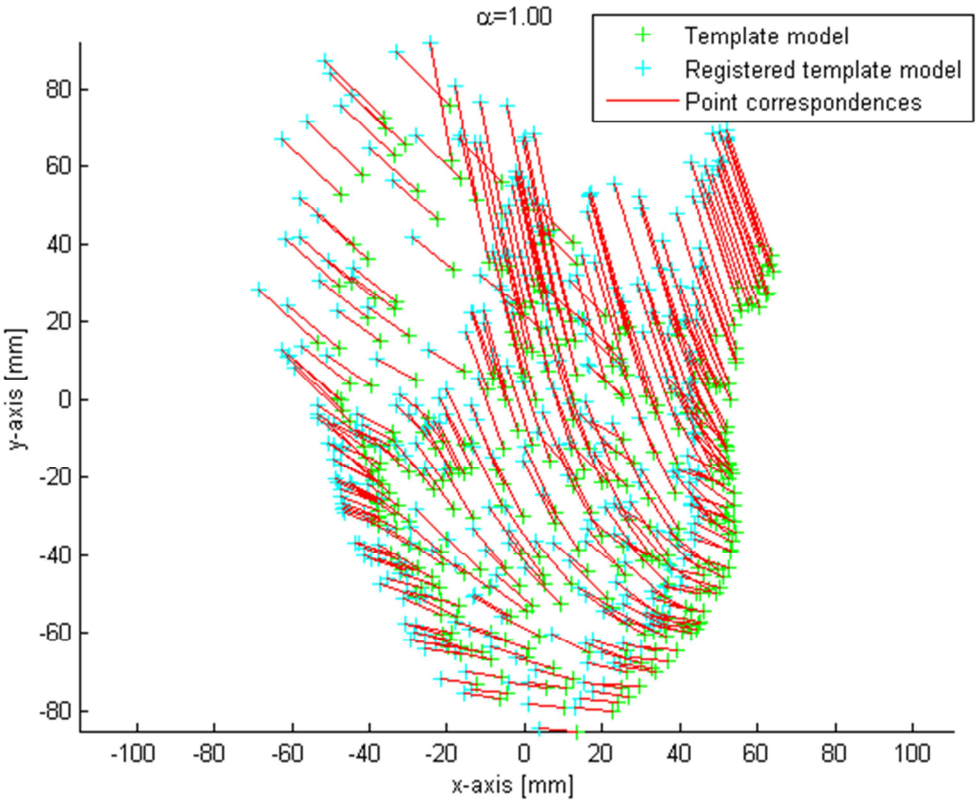


Figure 2.16: Template model plotted with the registered template model, frontal view.

As opposed to the figures of the point correspondences between the reference model and the template model, the point correspondences between the template and the registered template seem to have no crossovers. Both Figure 2.16 and Figure 2.17 confirm this observation.

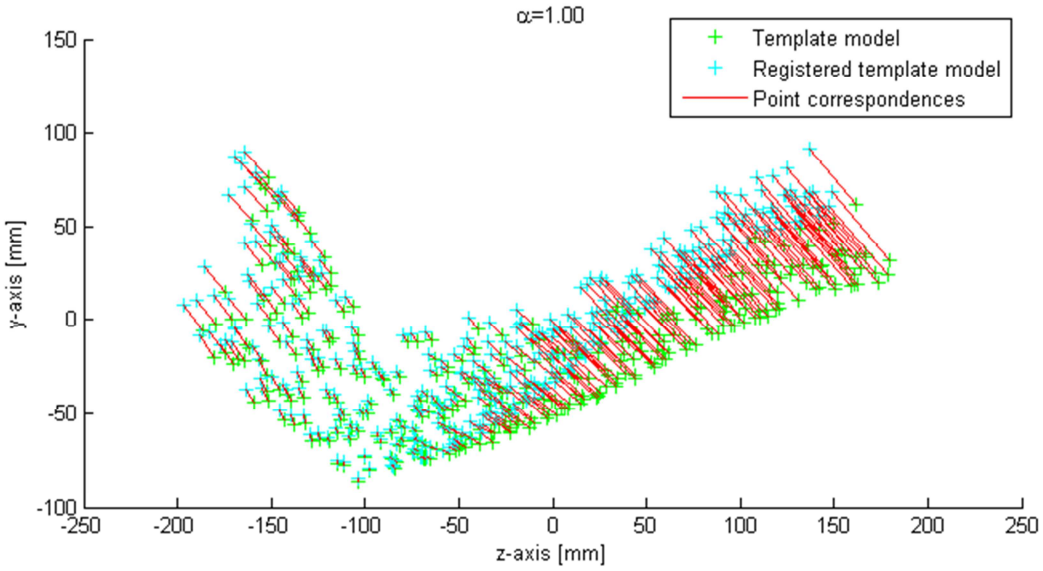


Figure 2.17: Template model plotted with the registered template model, lateral view.

Recalling Figure 2.8 (illustrating the diagram of the registration algorithm), the registration is performed through an iterative process. Only one iteration was performed to achieve these results.

An unexpected elongation of the brachial (arm) seems to appear Figure 2.17. This might be due to different cutoffs of the template model and the reference model. Such problems might not appear in full body models. The problem will therefore be ignored in this project. However the problem has shown to be a major issue for the segmentation presented in the next section and the recording had to be recorded over multiple times to achieve results with a negligible elongation.

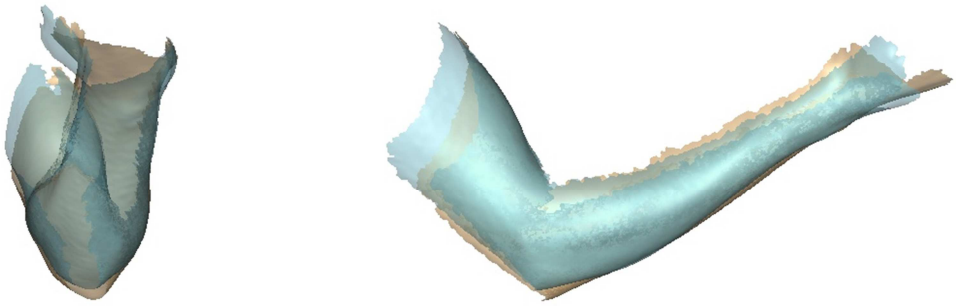


Figure 2.18: Reference model (golden) and registered template model (cyan) plotted on top of each other in frontal and lateral view respectively

Figure 2.18 shows that the registration has succeeded. Only minor derivations appear in the hand and elbow region, which is likely to be modeled out through more iterations. The largest deviations seems to be nearby the elongated part of the upper arm, that is suspected to be caused by variations in the cutoffs of the template- and reference models.

Approach for Error estimation	Value [mm]
Shortest distance RMS	6.2
Hungarian RMS	11.6

Table 2.1: Error estimation

Regarding Table 2.1, the errors seem to be quite large. Since significant differences between the lengths of the upper arms seem to provide the largest differences between the two models, it is presumed that this is the main reason for the large error. A significant difference between the two error estimations appears as well. This is most likely also related to the error provided by the elongated arm.

The results of two other registrations are listed in brief format below.

2.6.4.2 Registration of a flexing clenched fist

Images have been acquired of a flexing clenched fist. The registration was performed through three iterations. Figure 2.19 illustrates the reconstructed models. The template model is represented in cyan and the reference is represented in golden colors.

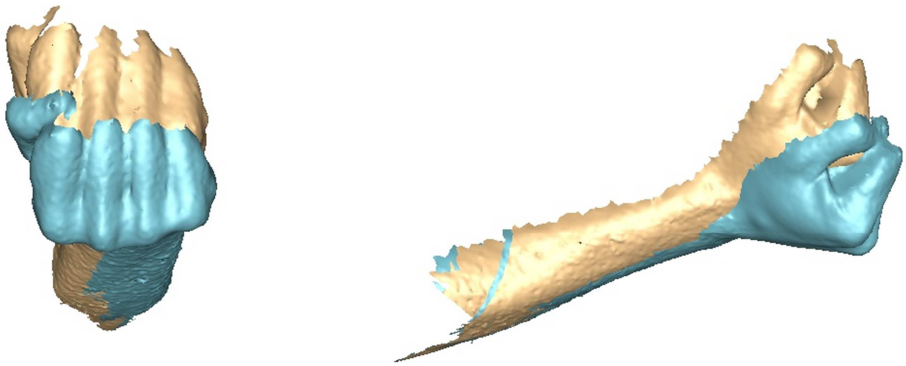


Figure 2.19: Reference- (golden) and template (cyan) model from dorsal and lateral view respectively

The registrations of the models are illustrated below.

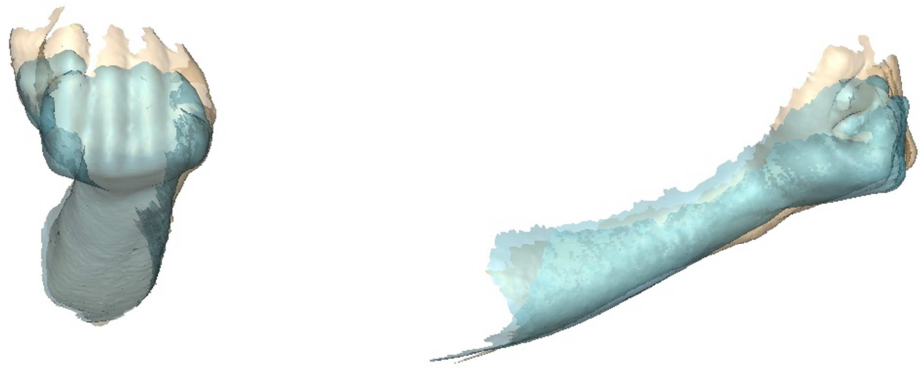


Figure 2.20: Reference- (golden) and registered template (cyan) model from distal and lateral view respectively

The registration seems not to be as good as expected. There is still a significant difference between the flexing positions. The holes between the fingers in the reference model (golden) are suspected to prevent a proper match of the finger regions.

Approach for Error estimation	Value [mm]
Shortest distance RMS	4.5
Hungarian RMS	9.6

Table 2.2: Error estimation

Regarding Table 2.2, the errors are reduced significantly in relation to the flexing arm. However the RMS of both approaches seem to be quite high. Anyhow the errors are reasonable, taking the poor registration into consideration.

2.6.4.3 Registration of a flexing stretched hand

These models illustrate a stretched hand flexing in the wrist. Three iterations have been performed in the registration as well. Figure 2.21 shows the acquired models just as in the former results.

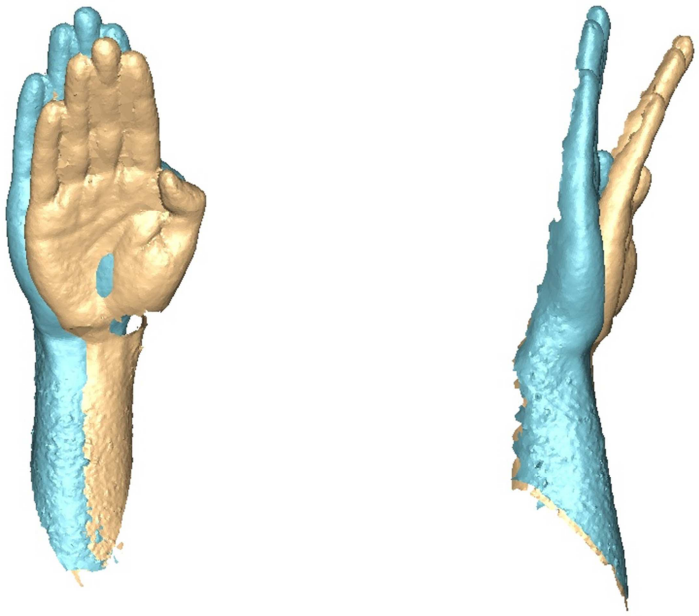


Figure 2.21: Cyan: template model; Golden: reference model. The models are viewed from planter and lateral view respectively.

The registration of the template to the reference model is illustrated below.

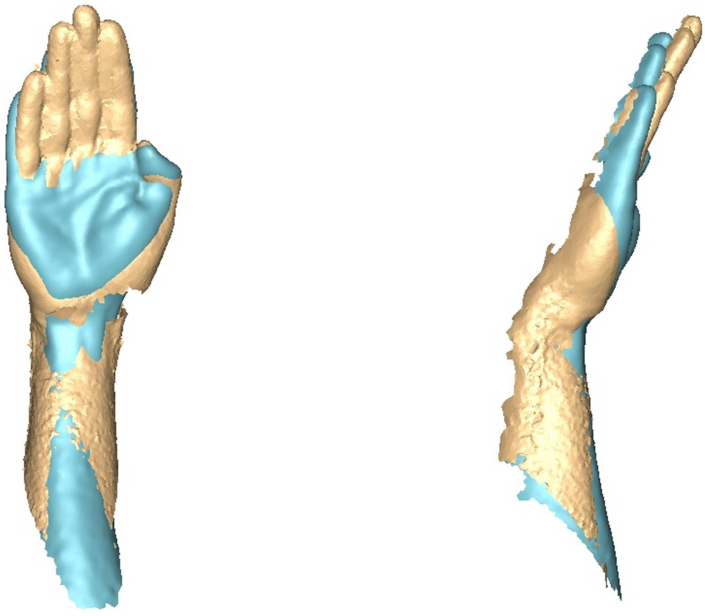


Figure 2.22: Cyan: Registered template model; Golden: reference model. The models are viewed from planter and lateral view respectively.

When looking at the registered hand from the lateral view, it is seen that the model has been shortened at the fingertips. Comparing the result with the results with the other registered hand in section 3.21.1 in appendix, this seems to be a problem that appears when the flexed angle becomes too large to achieve a satisfactory point match.

Approach for Error estimation	Value [mm]
Shortest distance RMS	4.6
Hungarian RMS	11.8

Table 2.3: Error estimation

Just like the registration of the clenched fist, regarding Table 2.3, the RMS values are considered to be quite high, but reasonable in relation to the fit that does not seems to be optimal.

2.6.5 Discussion

As described in the results, the acquisitions of the models were replicated two times before applicable registrations were achieved. The registrations were often infected by either elongations of the proximal end of the arm or shortening of the distal end. The elongation of the proximal end is a consequence of differences in the cutoffs of the reconstructed models. It would therefore not be a problem in practice. However the shortening problem seemed to appear when the angles between the reference- and the template models became too large.

The registrations of small deformations seem to work very well. But too small deformations might be insignificant for the segmentation if the erroneous registrations are dominating. The problem is comparable to a signal versus noise problem. A large deformation provides a large signal and errors in the registration contribute noise to the signal.

Considering the registrations of the stretched hands flexing in both the “Experimental results” section and appendix 3.21.1, it is observed that the registration algorithm is sensitive to large deformations. This indicates that it might be necessary to find a method to make the shape context invariant to rotations of the model. Finding a way to integrate the curvature term in the cost function, might help to improve the registration as well. Considering the figures from the curvature test in section 3.16 in appendix, it seems like the surface of the achieved models are too fluctuating. The significant down sampling of the points makes very local curvature features redundant. Smoothing the models before down sampling will improve the approach.

After all, the result of the test shows that this approach is capable of achieving a satisfactory registration to perform a successive segmentation, as presented in the next section.

2.7 Limb segmentation

According to the problem statement, the intension is to segment body parts into rigid segments, by tracking the points of the 3D mesh over time. A simplified approach will be presented to show how this can be achieved.

In this approach, following assumptions should be met:

1. All limb segments are rigid
2. Prior knowledge about the number of segments has to be acquired
3. All limbs are rotating relatively to each other in the frames

When these assumptions are fulfilled, it is possible to perform the segmentation, using the template model and the registered template model from the formerly described registration. The template model will be divided into clusters of equal sizes using k-means clustering on the 3D points, illustrated by the red and blue segments in Figure 2.23. Once the clusters are defined, a local coordinate system is generated for each cluster using PCA, illustrated with cyan and magenta colors in Figure 2.23. The first axis of the local coordinate system will be along the direction of the cluster with the largest deviation. The centers of the coordinate systems are located in the center of mass of the point cloud.

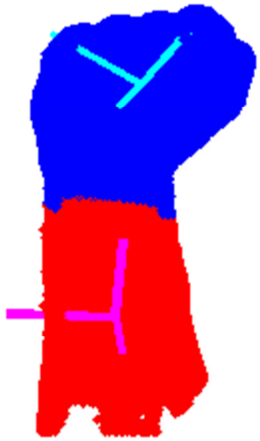


Figure 2.23: Plantar view of right hand. The hand is divided into two segments: a red segment and a blue one. A Local coordinate system is defined for each segment labeled with magenta and cyan colors respectively.

The local coordinate systems are generated as well for the registered template model on basis of the same points that are clustered in the template model.

Since the local coordinate systems will be moving with the corresponding clustered points from one frame to another, the relative movements of the points should be least in the coordinate system, defined by the cluster it belongs to. If not, the point will be reclassified through iterative clustering. In the iterative process the relative movements of the points are recalculated in each coordinate system. The points will hereby be reclassified according to which coordinate systems the moving distance is least. Figure 2.24 illustrates the reclassification of the points and a recalculation of the new coordinate systems.

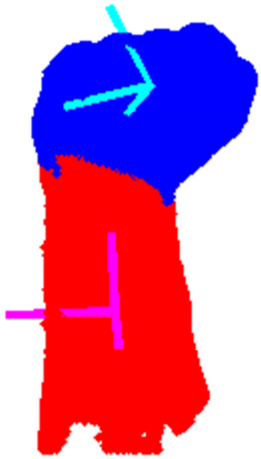


Figure 2.24: Reclassification of the hand illustrated in Figure 2.23, with recalculated coordinate systems.

2.7.1 Experimental results

2.7.1.1 Flexing arm

This result represents the segmentation of the flexing arm from the point tracking section.

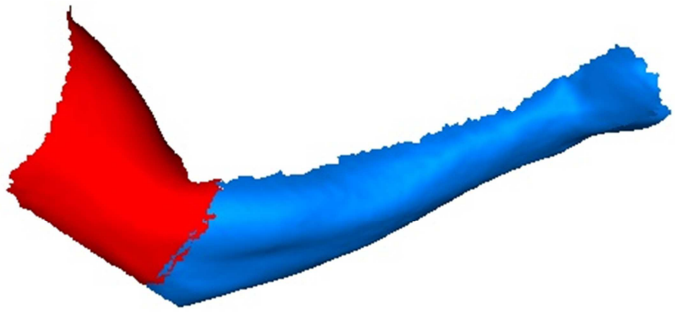


Figure 2.25: Segmentation of the flexing arm

The registration has been sufficient to provide good conditions for the segmentation. The features of interest are the rotation and translation of the segments which will be used to calculate the rotation center. It is hard to say whether the segmentation is sufficiently accurate or not for this purpose. This is a question that would be interested to get answered through a future work.

2.7.1.2 Flexing clenched fist

The figure below shows the segmentation of the dorsal view point. The segmentation converged fine towards the result shown in Figure 2.26.

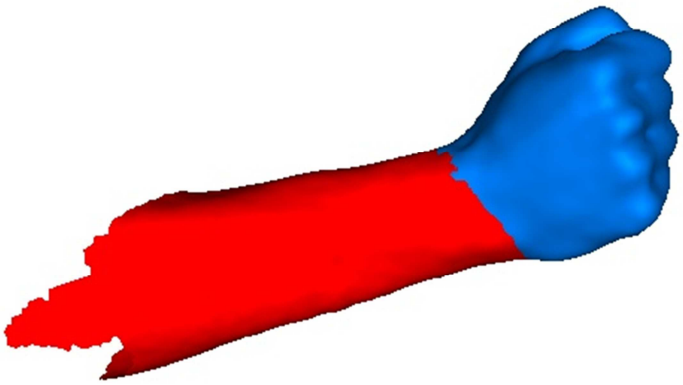


Figure 2.26: Segmentation of the flexing clenched fist

Again the segmentation does not seem to go straight through the wrist, but it is presumed that it will be sufficient to find the joint centers through their relative rotations.

2.7.1.3 Flexing stretched hand

The segmentation didn't converge through the iterations for this registration. Instead it was widely fluctuating around the wrist. The registration error illustrated Figure 2.22 in the former section seems to have a major impact on this. However in most of the iterations the segmentation ended up as illustrated in Figure 2.27.

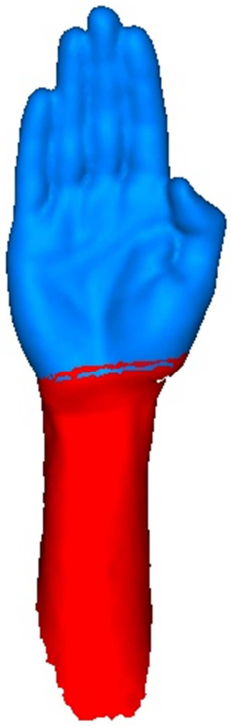


Figure 2.27: Segmentation of the flexing stretched hand

In section 3.21.2 in appendix, a test of a hand flexing in two joints is presented. The clustering was performed with both two and three segments.

2.7.2 Discussion

The segmentation algorithm seems to be very sensitive to errors in the registration. As soon as the elongation errors or the shortening errors appeared in the registered templates, the segmentation failed completely.

The results that are illustrated in the 'Experimental results' section seems to be quite promising. However the accuracy of the joint center estimation would still be highly dependent on the magnitude of the angular movement through the time the points are tracked. If it is possible to expand the approach such the points are properly tracked through several frames or even a whole gait cycle, a high accuracy of the joint center estimation would be expected.

Different approaches to improve the algorithm are discussed in the future work section.

2.8 Conclusion

The general purpose of part II of this master thesis was to find a method for pose estimation that could be used in accordance with a tracking approach using the PMVS algorithm. It was aimed to find an approach that did not require any generation of subject specific articulated models, but still have a potential to obtain accuracy similar to the approaches using such models.

There is evidence that the registration and segmentation approach can be used to segment limbs in a proper manner. It is most likely that the segmentation can be used to estimate the joint centers through the relative motions of the segments.

More tests has to be obtained to find out whether the proposed approach would be able to match the state of the art approaches that uses articulated models.

A high temporal camera resolution at approximately 100 fps seems to be necessary to use the approach for human gait. However improving the registration algorithm to be more functional for large movements might be an even better solution.

2.9 Summary

The main objective of gait analysis is to extract the joint centers of a tracked test subject and calculate the kinematics of the motion. Part II if this thesis has worked towards a pose estimation approach to markerless motion capture in which the joint centers are extracted from a test subject in motion.

To perform the pose estimation it was proposed to estimate the joint centers from the relative motions of the limb segments. The segmentation was performed by classifying the 3D points into clusters with minimal relative motion. The motion of the 3D points was tracked using 3D registration.

3D registration was performed using 3D shape contexts for point matching and calculating the minimal costs by using the Hungarian algorithm. The cost function was tested with an expansion integrating the mean curvature of the surface. The results of the test were not satisfactory, which resulted in cancelation of the expansion. However smoothing the reconstructed models might be sufficient to achieve satisfactory results with the mean curvature term.

Registration using shape contexts with no constraints, showed promising results for small deformations. However the algorithm seemed to be too sensitive for large deformations. Defining constraints for the cost function might improve the registration technique significantly.

The segmentation seemed to be very sensitive to errors in the registration. Finding a method to perform registrations over larger angular movements is therefore a natural step in the future work.

Conclusions and reviews

Through this master thesis it has been tested whether it is possible to replace a laser scan of a full body with a model obtained with a photogrammetric approach. Conclusion was that using eight ten mega pixel SLR cameras made it possible to obtain a model with a resolution at 3±1 mm, if optimal illumination and texture were obtained. This result is as precise as the laser scans obtained in (Corazza, et al., 2006).

However it might be beneficial to replace the PhotoModeler® Scanner software with the PMVS software presented in (Furukawa, et al., 2010), to gain a more cost efficient solution and less time consuming processing as well.

Alternative methods to segment the human body have also been tested, as a step in pose estimation for a markerless motion capture approach. The segmentation should be used to extract the joint centers in the reconstructed models obtained for each frame in a video sequence. First step of the pose estimation was to track the points over time using 3D point registration and hereby segment the limbs into rigid parts.

The 3D registration seemed to be well suited for registrations of small movements. In practice the magnitude of the movements between the frames is dependent on the velocity of the movements and the frame rate of the cameras. More tests have to be obtained to find out whether the registration is sufficient for normal gait recorded by 50 Hz cameras. The registration can be improved with constraints as proposed in the next section about future work.

The segmentation worked between two segments when the registration algorithm succeeded to perform a proper registration. Including more than two frames in the segmentation approach would most likely improve it and reduce the high sensitivity to errors in the registration.

The classification approach using k-means clustering to perform the segmentation worked for two segments as well. It was proven that proper segmentations can be achieved with this approach as long proper initial segmentation was performed. Splitting the limbs into two segments of equal size was sufficient for these models. However it might be necessary to perform a more intelligent initial segmentation on a full body. Using simple articulated models for this purpose might be a solution.

The processing of the pose estimation is very time consuming. Processing a ten seconds video sequence with a temporal resolution at 50 fps would take several days, even if it is calculated on a quad core processor. The time consuming part of the algorithm is by far the point matching for registration. Reducing the number of points to be matched would reduce the processing time significantly and might not provide significant losses of the precision of the segmentation if the matches are properly performed. It should be possible to reduce the processing time to less than one day with simple methods. However reducing the processing time to a couple of hours is not likely yet.

Future work

The photogrammetric approach obtains full 3D models with similar precision as a laser scan as long the conditions for illumination and texture are optimal. However using PhotoModeler® Scanner is a time consuming process, because the object has to be trimmed manually in the images. This takes approximately as long time as putting markers on a test subject. The benefit would therefore not be as large as expected. In addition there are expenses for licensing the product as well. It would therefore be of interest to replace PM with the PMVS algorithm proposed in (Furukawa, et al., 2010). PMVS is free and protected by GNU General Public License. This modeling process is fully automatic when background subtraction is applied on the input images and about the same camera configuration can be applied with this approach. However camera calibration has to be performed before use.

Regarding the pose estimation approach for markerless motion capture, it is obvious to improve the registration algorithm as discussed. There are large benefits in such improvement since the robustness of the algorithm is strongly dependent on the registration.

In (Xiao, et al., 2009) they propose to make a constraint to the point matching algorithm, such that the topological relationships are preserved when the point matching are performed. The results seem to be promising and might be helpful to integrate into future work as well.

In future work it would be beneficial to get more statistical evidence of for the segmentation by including several reconstructed models as in the test presented in section 3.21.2 in appendix.

Another future work would be to find a substitute for the k-means clustering approach. The k-means clustering was only integrated in the algorithm as a temporal solution, to show the basic concepts of the segmentation through an iterative process. It is most likely that k-means clustering will not work for a segmentation of a full body unless an initial estimation of the cluster centers is applied. Fitting an articulated model to the template model could provide a good initial estimation of the cluster centers.

References

Belongie Serge, Malik Jitendra and Puzicha Jan Shape Matching and Object Recognition Using Shape Contexts [Article] // IEEE Transactions on pattern analysis and machine intelligence. - [s.l.] : IEEE, april 2002. - VOL. 24. - 24. - 0162-8828/02.

Castello Beryl The Hungarian Algorithm [Online] // hungarian.pdf. - Johns Hopkins University, January 24, 2007. - April 20, 2011. - http://www.ams.jhu.edu/~castello/362/Handouts/hungarian.pdf.

Cheung Kong-man, Baker Simon and Kanade Takeo Shape-From-Silhouette Across Time Part I: Theory and Algorithms [Journal]. - [s.l.] : International Journal of Computer Vision, 2005. - 3 : Vol. 62.

Cheung Kong-Man, Baker Simon and Kanade Takeo Shape-From-Silhuette Across Time Part II: Application to Human Modeling and markerless Motion Tracking [Journal]. - [s.l.] : International Journal of Computer Vision, 2005. - 3 : Vol. 63.

Christiansen Martin S. 3D Motion capture for medical applications [Report]. - Copenhagen : [s.n.], 2010.

Corazza S. [et al.] A Markerless Motion Capture System to Study Musculoskeletal Biomechanics: Visual Hull and Simulated Annealing Approach [Journal]. - [s.l.] : Annals of Biomedical Engineering, 2006. - 6 : Vol. 34.

Corazza Stefano [et al.] Automatic Generation of a Subject-Specific Model for Accurate Markerless Motion Capture and Biomechanical Applications [Journal] // IEEE transactions on biomedical engineering. - [s.l.] : IEEE, 2010. - 4 : Vol. 57. - pp. 806-811.

Corazza Stefano [et al.] Markerless Motion Capture through Visual Hull, Articulated ICP and Subject Specific Model Generation [Journal]. - [s.l.] : Springer Science, 2009. - 11263-009-0284-3.

Corazza Stefano, Mündermann Lars and Andriacchi Tom A framework for the functional identification of joint centers using markerless mortin capture, validation for the hip joint [Article] // Journal of Biomechanics. - 2007. - 40.

Cyganek Boguslaw and Siebert J. Paul An introduction to 3D computer vision techniques and algorithms [Book]. - [s.l.] : Wiley, 2009. - 978-0-470-01704-3.

Devernay F and Faugeras O Computing Differential Properties of 3-D Shapes from Stereoscopic Images without 3-D Models [Book]. - Seattle : In Proc. IEEE Conf. Comp. Vision Patt. Recog., 1994.

Esteban Carlos Hernández and Schmitt Francis Silhouette and stereo fusion for 3D object modeling [Article] // Computer Vision and Image Understanding. - 2004. - 96.

Faugeras Olivier Three-Dimensional Computer Vision [Book]. - London : The MIT Press, 1993. - 0-262-06158-9.

Fletcher Luke, Dr. Course homepage from The Australian National University [Online] // An Introduction to Computer vision. - 2003. - http://users.cecs.anu.edu.au/~luke/cvcourse.htm.

Forsyth David and Ponce Jean Computer vision: a modern approach [Book]. - [s.l.] : Prentice Hall, 2003. - 0131911937, 9780131911932.

Franco Jean-Sébastien home: EPVH Visual Hull Library [Online] // EPVH Visual Hull Library. - 4D View Solutions, October 9, 2007. - January 11, 2011. - http://perception.inrialpes.fr/~Franco/EPVH/#History.

Furukawa Y and Ponce J Accurate, dense and robust multi-view stereopsis [Journal]. - [s.l.] : TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2010. - 8 : Vol. VOL. 32.

Furukawa Yasutaka and Ponce Jean Dense Patch Models for Motion Capture from Synchronized Video Streams [Article] // Willow Tech. Report. - 2007.

Harris C. and Stephens M. A Combined Corner and Cdge Cetector [Article] // Proceedings of the 4th Alvey Vision Conference. - 1988.

Hartley Richard and Zisserman Andrew Multiple View Geometry [Book]. - New York : Cambridge University Press, 2003. - 978-0-521-54051-3.

Hecht Eugene Optics [Book]. - San Francisco : Addison Wesley, 2002.

Hullo J.F, Grussenmeyer P and Fares S PHOTOGRAMMETRY AND DENSE STEREO MATCHING APPROACH APPLIED TO THE DOCUMENTATION OF THE CULTURAL HERITAGE SITE OF KILWA (SAUDI ARABIA) [Article] // CIPA Symposium. - 2009. - 22.

Inc. Eos Systems Core Technology [Online] // PhotoModeler - Photogrammetric Technology. - Eos Systems Inc., 2010. - January 5, 2011. - http://www.photomodeler.com/about_us/coretechnology.htm.

Istook Cynthia L and Hwang Su-Jeong 3D body scanning systems with application to theapparel industry [Article] // Journal of Fashion Marketing and Management. - 2000. - 5.

Joel Mitchelson Multiple Camera Studio Methods for Visual Capture of Human Motion [Report]. - Guildford : Centre for Vision, Speech and Signal Processing, University of Surrey, 2003.

Katz Sagi, George Leifman and Tal Ayellet Mesh segmentation using feature point and core extraction [Journal]. - [s.l.] : Springer, 2005.

Kortgen M [et al.] 3D Shape matching with 3D shape context [Article] // Proceedings of the seventh central european seminar on computer graphics. - 2003.

Larsen Rasmus Medical Image Analysis [Book]. - Kongens Lyngby : [s.n.], 2008.

Li Stan Z. Enclyclopedia of Biometrics [Book]. - [s.l.] : Springer, 2009. - Vol. 2.

Lien Lyh-Ming and Amato Nancy M. Simultaneous Shape Decomposition and Skeletonization [Journal]. - Texas A&M University : [s.n.], 2005.

Luhmann T [et al.] Close Range Photommetry [Book]. - Caithness : Whittles Publishing, 2006. - 1-870325-50-8.

Marr D and Poggio T A computational theory of human stereo vision [Article] // Proceedings of the Royal Society of London. - 1978.

Mc Graw Hill Anatomy and Physiology Revealed [Internet software]. - [s.l.] : Mc Graw Hill, Medical College of Ohio, 2011.

Meyer Mark [et al.] Discrete Differential-Geometry Operators for Triangulated 2-Manifolds [Report]. - 2000.

Microsoft Kinect Fact Sheet [Word document]. - 2010.

Moeslund Thomas B. and Granum Erik A Survey of Computer Vision-Based Human Motion Capture [Journal]. - [s.l.] : Computer Vision and Image Understanding, 2001. - Vol. 81.

Moeslund Thomas B., Hilton Adrian and Krüger Volker A survey of advances in vision-based human motion capture and analysis [Journal]. - [s.l.] : Comptuer vision and image understanding, 2006. - Vol. 104.

Mündermann Lars [et al.] Conditions that influence the accuracy of anthropometric parameter estimation for human body segments using shape-from-silhouette [Article] // Proceedings of the SPIE - The International Society for Optical Engineering. - [s.l.] : SPIE - The International Society for Optical Engineering, 2005. - 0277786x.

Mündermann Lars [et al.] Most favorable camera configuration for a shape-from-silhouette markerless motion capture system for biomechanical analysis [Journal]. - [s.l.] : SPIE, 2005. - Vol. 5665.

Scharstein Daniel and Szeliski Richard vision.middlebury.edu [Online] // http://vision.middlebury.edu/mview/eval/. - Support by Middlebury College, Microsoft Research, and the National Science Foundation, August 15, 2009. - March 18, 2011. - http://vision.middlebury.edu/mview/eval/.

Schneider David Visual Hull [Article]. - 2010.

Seeley Rod R, Stephens Trent D and Tate Philip Anatomy & Physiology [Book]. - New York : McGraw-Hill, 2006. - Vol. Seventh edition.

Siebert J. Paul and Marshall Stephen J. Human body 3D imaging by speckle texture projection photogrammtry [Article] // Sensor Review. - 2000. - Volume 20. - 3.

Tang Zheng-Zong [et al.] Three-dimensional digital image correlation system for deformation measurement in wxperimental mechanics [Article] // Optical Engineering. - 2010. - 49.

The Center for Orthopaedics & Sports Medicine Meniscal Repair [Online]. - 3 4, 2003. - April 7, 2011. - http://www.arthroscopy.com/sp05026.htm.

Vaughan Christopher L, Davis Brian L and O'Connor Jeremy C Dynamics of Human Gait [Book]. - Cape Town, South Africa : Kiboho Publishers, 1999. - 2nd edition.

Wikinoticia.com Gait analysis as a method af identification [Online] // Technology - Genaral. - Wiki notica, October 9, 2010. - March 15, 2011. - http://en.wikinoticia.com/Technology/general-technology/62722-gait-analysis-as-a-method-of-identification.

Xiao Di [et al.] An improved 3D shape context based non-rigid registration method and its application to small animal skeletons registration [Article] // Computerized Medical Imaging and Graphics. - [s.l.] : Elsevier Ltd., 2009. - 0895-6111.

Xiao Di [et al.] Non-rigid registration of small animal skeletons from micro-CT using 3D shape context [Article] // Medical Imaging. - 2009. - 1605-7422/09.

Yang Sylvia Markerless motion-capture systems for estima-tion of pose and tracking of human objects: A review [Report]. - Copenhagen : [s.n.], 2011.

Contents of the data CD

A data CD is attached to the back of the master thesis. The directories on the data CD are listed below with comments of their contents:

- **Articles** –All articles that are referred to in this master thesis and other articles that are used for inspiration.
- **Grabber** –VS2010 project and Source code for the grabber that is used to grab images with the chameleon cameras.
- **Master thesis** –Master thesis in PDF-format.
- **MATLAB** –MATLAB source codes and functions that are used in this master thesis.
    - **PM** –Files related to Part I concerning PhotoModeler
    - **Pose Estimation** –Files related to Part II concerning pose estimation
- **PhotoModeler photos** –image data acquired from the tests in part I
    - **Face models** –Face models presented in appendix
    - **Full models** –Data from 'Full 3D model acquisition test'
    - **Surface** –Data from the precision tests
- **Pose estimation models** –Models obtained with the 3D scanner at the Panum Institute for part II
    - **Clenched fist**
    - **Flexing arm**
    - **Flexing fingers (appendix)**
    - **Stretched fingers**
    - **Test with multiple hand poses (appendix)**
- **Time plan** –Gantt chart for the thesis

2.10 Software overview

Following flow charts illustrates the MATLAB files and –functions used in the thesis.

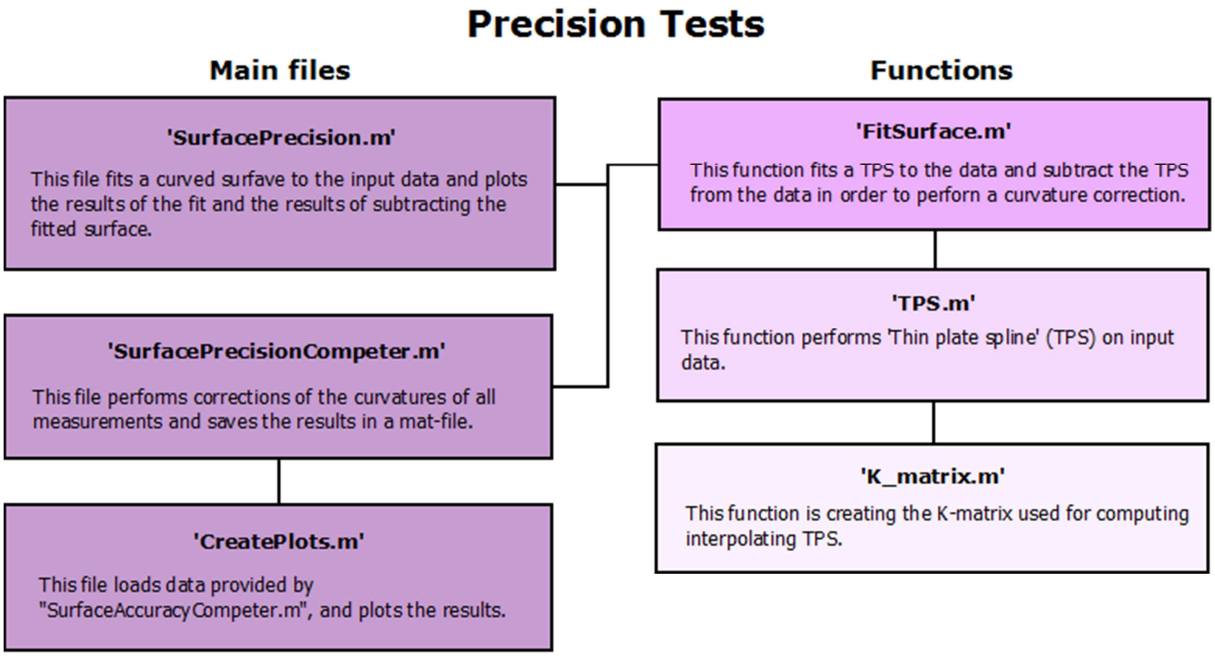


Figure 2.28: Flow chart of the MATLAB files used in the precision tests in Part I. Sub functions are labeled with light colors.

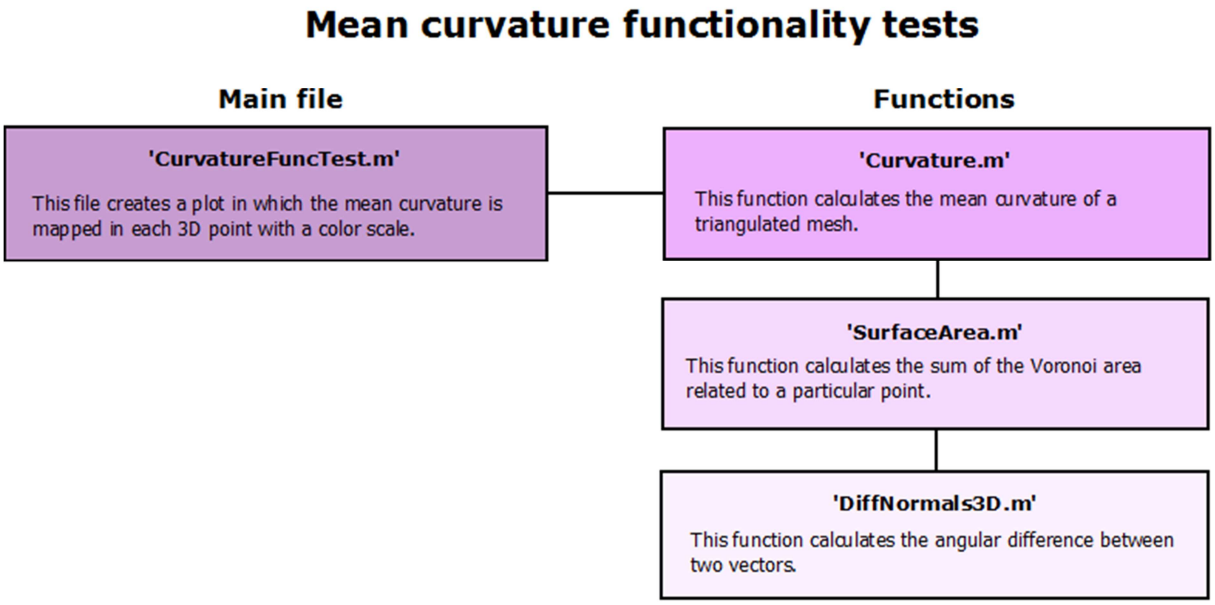


Figure 2.29: Flow chart of the MATLAB files used in the Mean curvature functionality test in Part II. Sub functions are labeled with light colors.

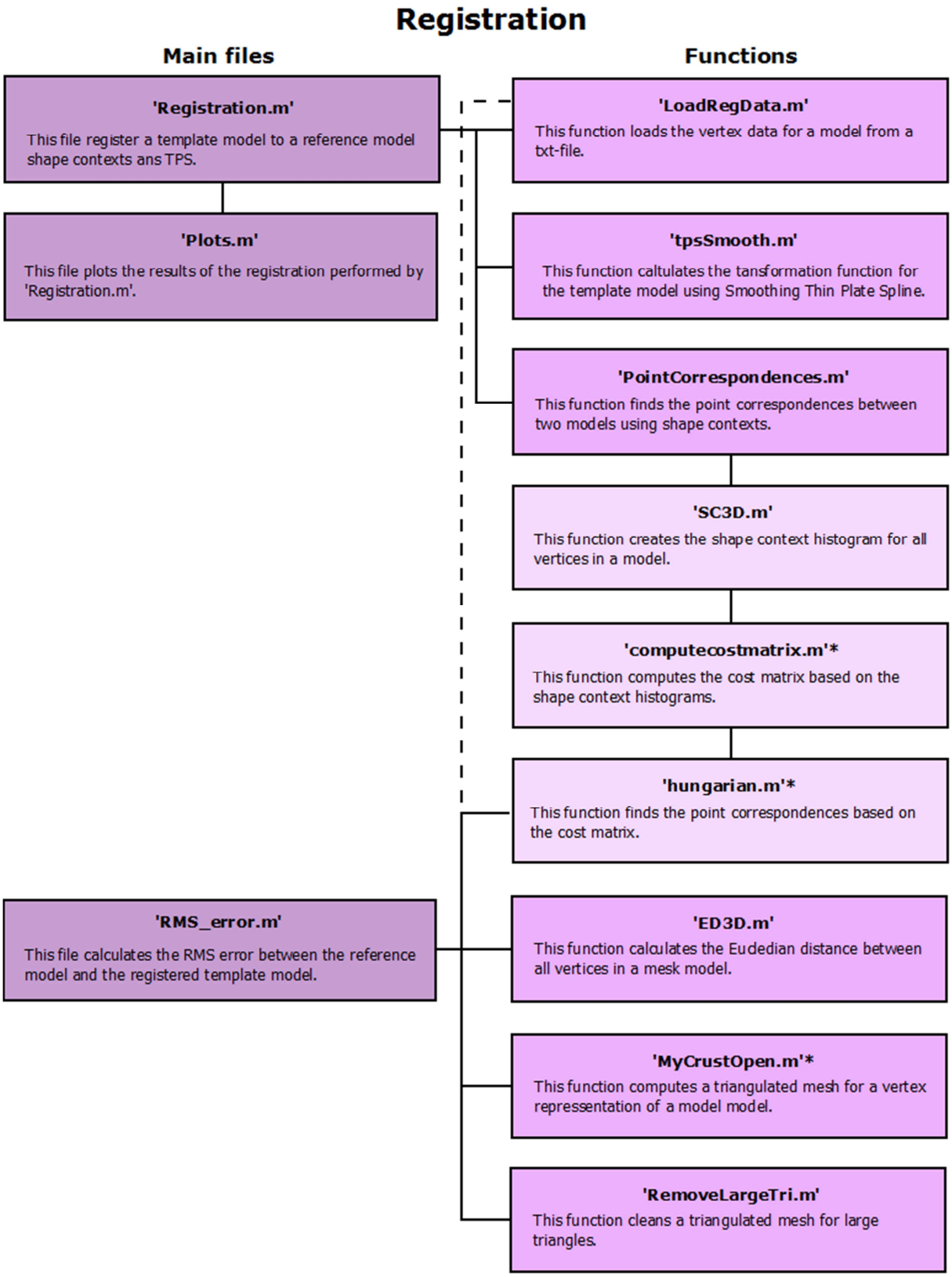


Figure 2.30: Flow chart of the MATLAB files used in the registration in Part II. Sub functions are labeled with light colors. Functions with asterisk are made by others.

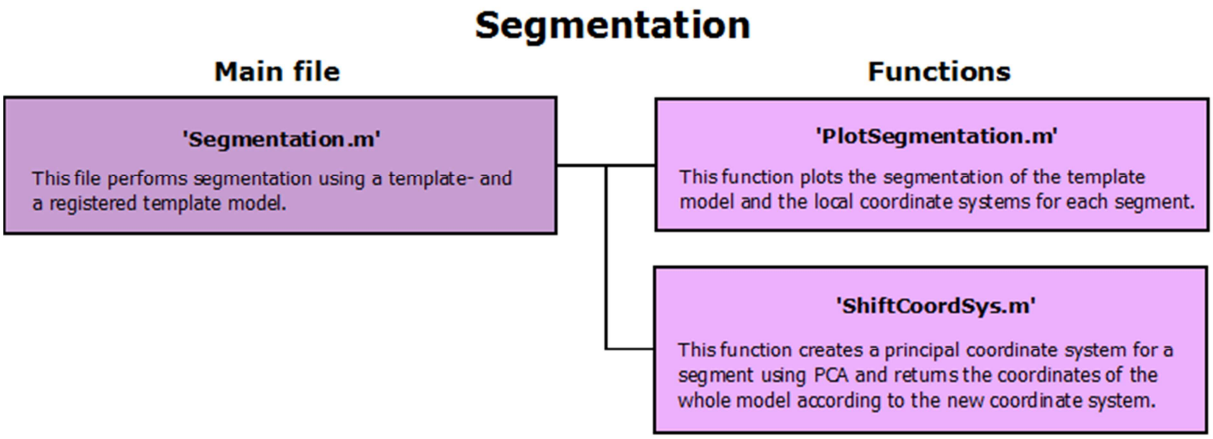
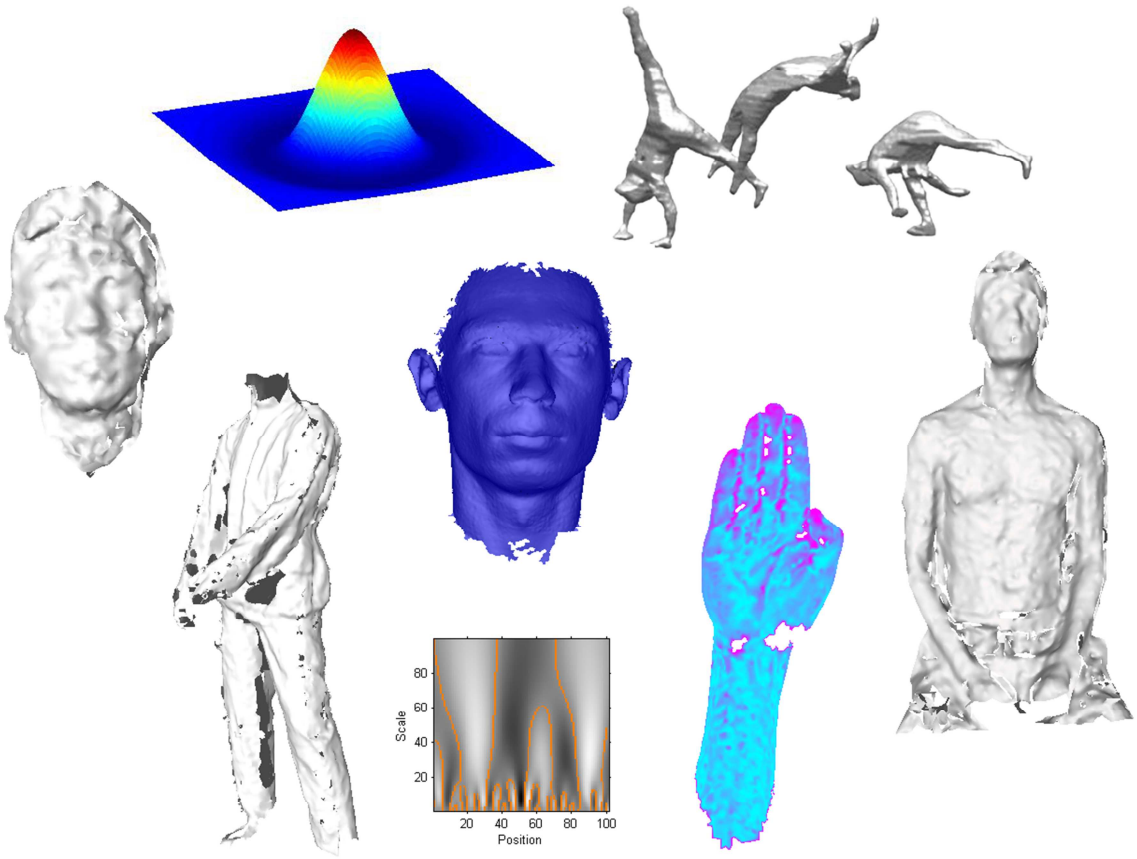


Figure 2.31: Flow chart of the MATLAB files used in the segmentation in Part II.



Appendix

3.1 Optics

3.1.1 Light propagation

Photons or electromagnetic radiation behaves like both waves and particles. When we think about photons as waves, the law of reflection and Huygens principle allows us to consider an illuminating or illuminated surface as a large number of point sources radiating hemispherical waves or rays emanating in the direction of the energy flow. If we establish a line of point sources (with a distance less than the emitted wave lengths to avoid diffraction) and an image plane parallel to the line of point sources, the waves from all point sources will contribute with energy to the whole image plane. As a result, all we would see on the image plane would be a mesh of diffuse light.

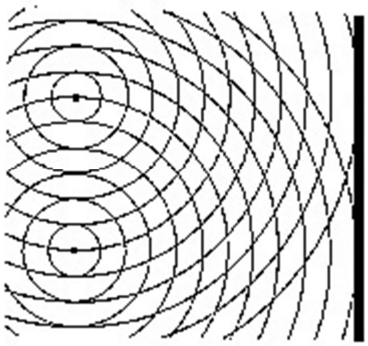


Figure 3.1: Two point sources contribute with energy to the whole image plane

To acquire an image of point sources, the light waves emitted from the point sources has to be restricted to expose a local area of the image plane. To do so one can either refocus the wave fronts on the image plane by using a lens or permit only a narrow part of the wave from each point source to expose the image plane using a pinhole. These methods is essential to understand the geometry used in stereo vision and will briefly be described through the next sections. A more detailed presentation of the concepts can be found in (Hecht, 2002), from which all optical theory, in this thesis, is based.

3.1.2 Pinhole model

A pinhole model allows only a limited part of a wave front to pass through. Ignoring the diffraction phenomena a wave front from a point source will be converted to a beam through the pinhole that hits the image plane depending on the direction of the beam. Ideally the smaller the pinhole is, the sharper would the image be as illustrated on Figure 3.2.

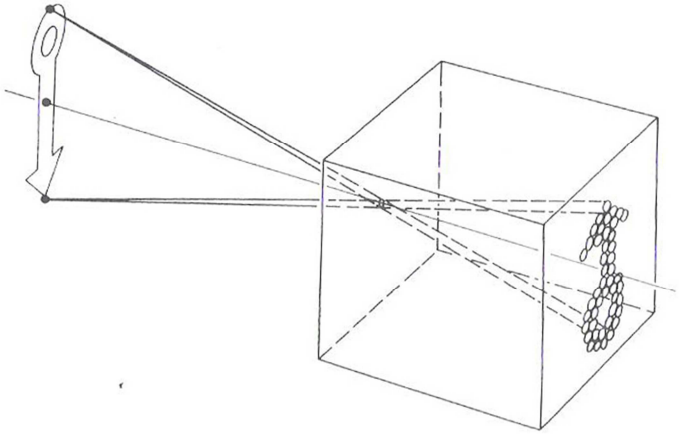


Figure 3.2: Sketch of the pinhole model (Hecht, 2002)

Unfortunately this is not the case since diffraction will bend the light, and makes the image blurry when the pinhole becomes smaller than a certain threshold. The most inconvenient about the pinhole model is that only a very limited amount of the light reflected from an object will expose the image plane and contribute to the image. In the pinhole model we use the term 'focal point' for the pinhole and the distance between the focal point and the image plane is called the focal length. This is illustrated in Figure 3.3.

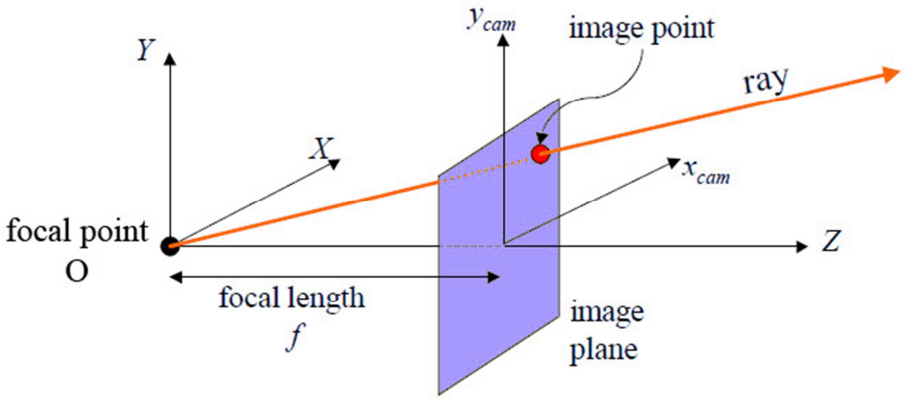


Figure 3.3: Geometry of the pinhole model (Fletcher, 2003)

3.1.3 Lens optics

To understand the optical properties of a lens, it is important to know Snell's law and Fermat's Principle. Snell's law explains the correlation between the refraction indexes and the angles of incident and refracted rays:

$$n_i \sin(\theta_i) = n_o \sin(\theta_o)$$

Equation 3.1: Snell's law

Where θ_i and θ_o represents the angles of incidence and refraction respectively and n_o and n_i the refraction indexes for the surroundings and the optical system respectively.

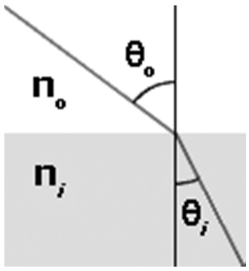


Figure 3.4: Snell's law

Fermat's Principle maintains that the optical path length is equal to a constant for a photon:

$$OPL = n_o l_o + n_i l_i$$

Equation 3.2: Fermat's Prinsiple

Here OPL represents the optical path length and l is the physical path length.

Using Snell's law (the law of refraction) and Fermat's Principle, we are able to explain how an optical system is able to refocus wave fronts, such that waves emitted from a point source is refocused in a focus point. In other words, a convex medium with a refraction index larger than the surroundings will change the wave front from diverging to converging as shown on the figure below.

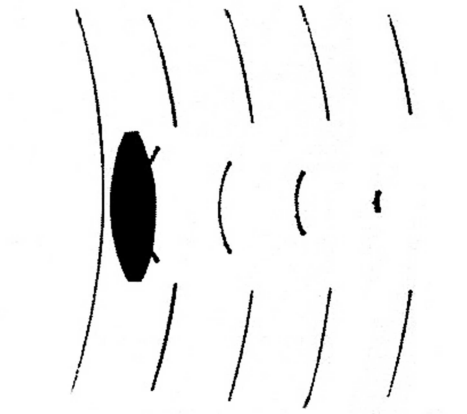


Figure 3.5: Wave front passing a lens that refocuses the wave

Denoting the **object distance** s spanning from a point source S in the **object plane** to the optical center O and the **image distance** p spanning from *O* to the corresponding focus point P, we can define the relationship as follows:

$$\frac{1}{f} = \frac{1}{s} + \frac{1}{p}$$

Equation 3.3

Here f is the focal length defined by the distance between the optical center and the focal point of the lens. Note that when s is reaching infinity, p is reaching f.

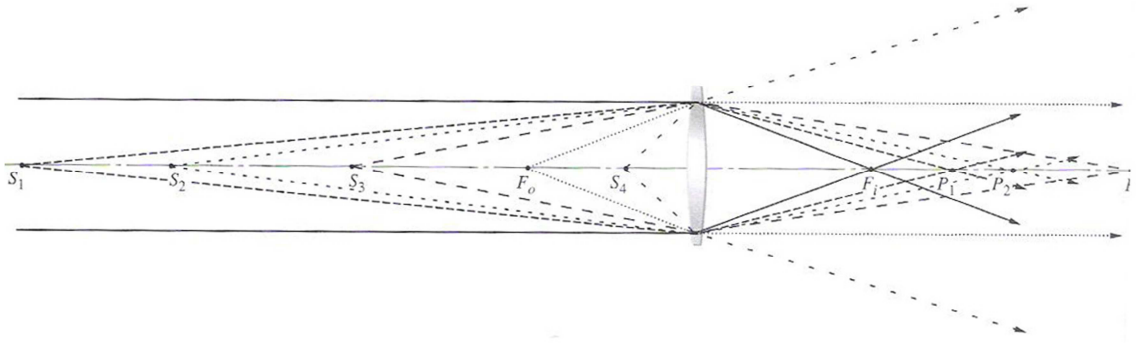


Figure 3.6: Illustration of how the location of the focus point 'P' is related to the location of the point source 'S' (Hecht, 2002)

Considering Equation 3.3 we see that if the point source is moved towards the lens then s become smaller) and the focus point the moves away from the lens which means that p becomes larger. This means we are likely to move the image plane away from the focal plane if we like to focus on a closer object. If the point source's distance to the lens equals or becomes smaller than the focal length, the rays will no longer be focused on the other side of the lens and we would no longer be able to acquire an image, no matter how far away we localize the image plane.

The term focal point is confusingly used different when it is used in accordance to either the pinhole model or the lens model. Just to remind the focal point in the pinhole model is defined by the pinhole itself. Note also that the focal length is fixed in the lens model whereas it is varying according to the location of the image plane in the pinhole model. However it is important to emphasize that the focal length estimated by calibration varies when focus or zoom changes, so these parameters has to be constant after a calibration.

Since we assume an object to consist of huge amount of point sources, we might use the term focal plane instead of a single focal point, because the location where a wave front is focused is dependent on the location of the point source.

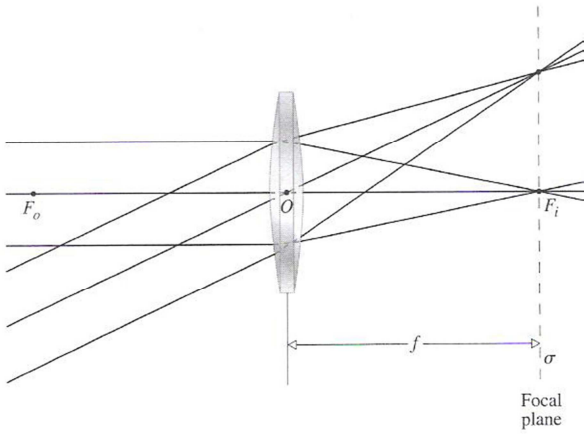


Figure 3.7: Light refocused on the focal plane. Reference: (Hecht, 2002)

If we assume the image plane is aligned with the focal plane we are able to use the geometry for the pinhole model to describe the geometrical relationships between the lens and the sensor chip in a digital camera, since the geometry between the optical center in a lens and the image plane is practically the same as between the pinhole and the image plane in the pinhole model. By this we ignore the lens distortion, which are explained in the next section.

3.1.4 Camera parameters

The pinhole model can be described by several parameters divided into the intrinsic parameters describing the internal geometry of the camera and the extrinsic parameters describing the location and orientation of the camera with reference to a world coordinate system.

The intrinsic parameters consist of:

1. Focal length
2. Principal point
3. Aspect ratio
4. Geometric distortion

These parameters provide four degrees of freedom.

As explained in section 3.1.2, regarding the pinhole model, the focal length defines the length between the focal point and the image plane as illustrated in Figure 3.8. Translated to a digital camera, the distance equals the image distance P. however P equals the focal length of the lens unless the target is significantly close to the camera. Note that the image plane often is set in front of the focal point as illustrated in Figure 3.8 instead of behind as the original pinhole model. By keeping the geometrical relations between focal point, object and image plane, this avoids turning the image upside down.

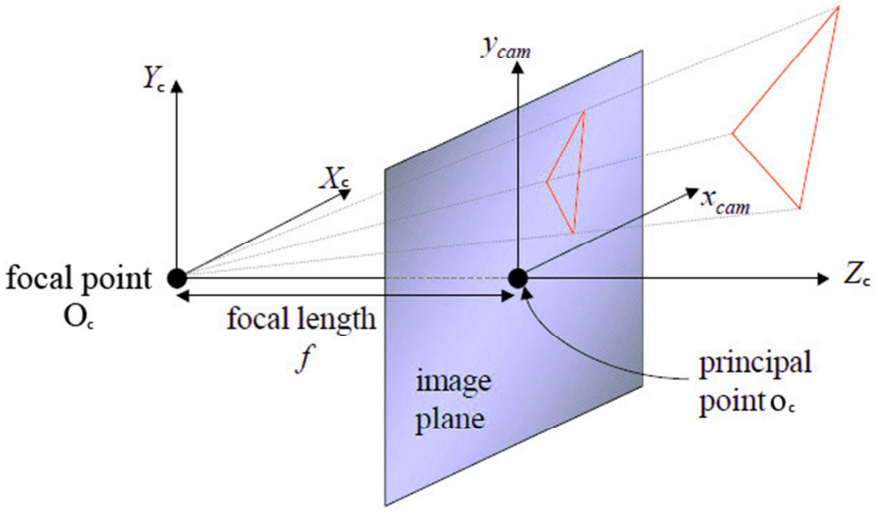


Figure 3.8: Intrinsic parameters describing the internal geometry of the camera

The principal point is defined as the point where the principal axis intersects the image plane also illustrated in Figure 3.8.

Aspect ratio is the relation between the width and height of a single pixel, defined by:

$$h = \frac{\alpha_y}{\alpha_x}$$

Equation: 3.1

Where h, α_y andα_x denotes the aspect ratio, height and width of a pixel respectively.

Due to the shape of the lens and the incidence of light, an image will be affected by geometric distortions. These distortions are dependent on the radial distance from the optical center where the distortion is assumed to be zero.

Figure 3.9: Effect of distortion correction

According to the tolerance of the accuracy, the radial distortion is often approximated by a polynomial of second or fourth order.

The extrinsic parameters consist of following rigid transformation parameters, providing six degrees of freedom:

1. Translation in three dimensions
2. Rotation in three dimensions

The rigid transformation describes the location and orientation of the focal point or optical center of a camera with reference to a world coordinate system.

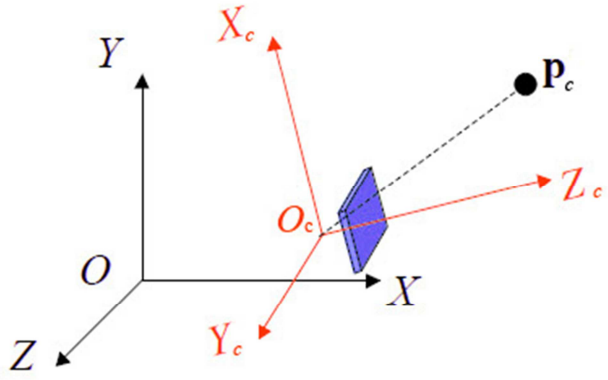


Figure 3.10: Extrinsic parameters describing translation and rotation of the camera view

Together the intrinsic and extrinsic parameters represent the camera parameters also called the **projection parameters**.

3.2 Feature based point match

Most frequent features used for point matching is edges and corners. The drawback to feature based methods is the sparse localization of points whereas the benefit is the highly reliable matches despite of noise or geometrical transformations.

Like any other data processing, it is commonly used to enhance the features of interest. Since edges can be considered as high frequent components, they can be enhanced by filtering the image by a high pass filter. On the other hand it is not desirable to enhance the noise, that is high frequent as well. A band pass filter is therefore often preferred. The most common of such filters is the gradient filters (first order derivatives) and the Laplacian of Gaussian filters (LoG) (second order derivatives) (Cyganek, et al., 2009).

3.2.1 Gradient filters

The gradient in an image is a measure of intensity variation from pixel to pixel. Assumed the contrast in the image is high, sudden changes in textures or edges of objects will therefore contribute to a large gradient. The magnitude of a gradient is typically estimated by summing the contributions for a vertical and a horizontal first order derivative filtration. Prewitt and Sobel are typical examples of derivative filters.

Figure 3.11: Upper left: original image, upper right: Gradient magnitude,
lower left: Vertical derivative using Prewitt filter, lower right: Horizontal derivative using Prewitt filter

Gradient filters are not optimal for blurred images, since the blurring makes a wide area with a relative large gradient.

3.2.2 Laplacian of Gaussian

Like the gradient filters equals the 1st order derivative of the image, the Laplacian equals the 2nd order derivative.

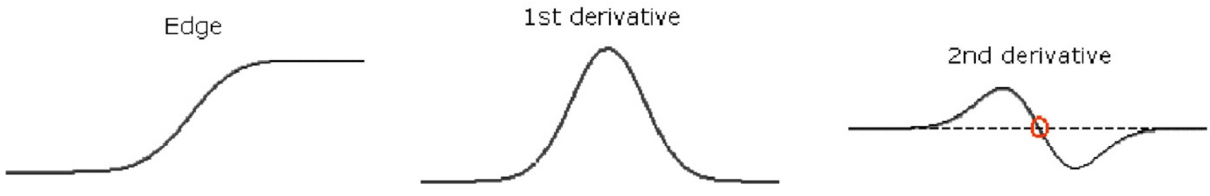


Figure 3.12: Illustration of the derivatives of a 1D signal simulating an edge

The LoG based edge detection method includes an additional step where the zero crossings of the 2nd derivative is about to be found. This makes LoG more insensitive to blur in relation to the gradient filters.

Using the Laplacian of Gaussian (LoG) instead of the Laplacian only, it prevents significant amplifications of high-frequency noise components, since the Gaussian function operates like a low pass filter. The LoG is expressed in Equation 3.4.

$$\nabla^2 G(x,y,\sigma) = LoG(x,y,\sigma) = \frac{1}{2\pi\sigma^4}\left(2 - \frac{x^2+y^2}{\sigma^4}\right) e^{-\frac{x^2+y^2}{2\sigma^2}}$$

Equation 3.4: The expression of LoG

Where σ is equal to the standard deviation of the Gaussian function. A graphical interpretation of the LoG function is shown in Figure 3.13

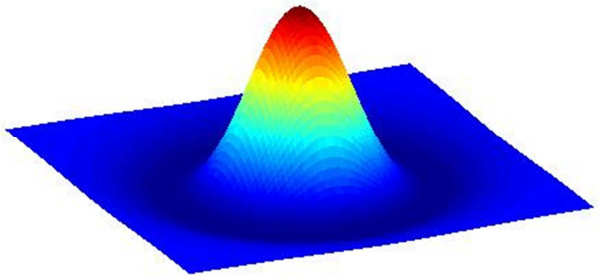


Figure 3.13: Shape of LoG filter

As oppose to the gradient filters, the LoG filter is sensitive to edges in all directions. In addition it is also more sensitive to noise. The results of edge detection using LoG is shown in Figure 3.14

Figure 3.14: Results of an edge detection using a LoG filter. Left: original photo; Right Photo Labeling of zero crossings in the LoG filtered photo.

3.2.3 Harris corners (Harris, et al., 1988)

Corners are very characteristic points in images. The intensity in those points varies remarkably often in several directions. They are therefore well suited for point matching (Cyganek, et al., 2009). Harris corners are one of the most commonly used algorithms to find corners. Harris corners are based on 2D structure tensors.

The 2D structure tensor $T(x_0, y_0)$, also called the Harris matrix, is expressed as:

$$T(x_0, y_0) = \int \int_{-\infty}^{\infty} w(x_0, y_0, x, y)\, T_0(x, y)\, dxdy$$

Equation 3.5

In Equation 3.5 is 'T_0' describing the intensity gradients defined by:

$$T_0(x, y) = \begin{bmatrix} \left(\frac{dI}{dx}\right)^2 & \frac{dI}{dx}\frac{dI}{dy} \\ \frac{dI}{dx}\frac{dI}{dy} & \left(\frac{dI}{dy}\right)^2 \end{bmatrix}$$

Equation 3.6

$w(x_0, y_0, x, y)$ is a window function. Assuming we use a Gaussian window (preferred in Harris corners), the expression of $w(x_0, y_0, x, y)$ is:

$$w(x_0, y_0, x, y) = e^{\frac{-(x-x_0)^2+(y-y_0)^2}{2\sigma^2}}$$

Equation 3.7

Note that this function equals a convolution between the window function *'w'* and the intensity gradients 'T_0'.

A Principal Component Analysis (PCA) of the covariance matrix of '*T*', leads to an interpretation of the gradients in terms of the eigenvalues λ_1 and λ_2 and the corresponding principal axis $\mathbf{v_1}$ and $\mathbf{v_2}$. The following cases of the eigenvalues indicate structures in images:

Indication of lines:

$$\lambda_1 \gg \lambda_2 \approx 0$$

Equation 3.8

Indication of corners:

$$\lambda_1 \geq \lambda_2 \gg 0$$

Equation 3.9

A more commonly used interpretation that gives the same results is using the determinant and the trace of '*T*' instead of PCA:

$$R = \det(T) - k \cdot trace(T)$$

Equation 3.10

'*R*' is the corner response and '*k*' is a parameter ranged between 0 and 0.25 (default value is 0.04 according to MATLAB and (Cyganek, et al., 2009)).

Figure 3.15 shows how a Harris corner detection looks like when using the MATLAB function 'cornermetric.m' to perform the corner detection. To reduce the number of detected corners a drastically smoothing of the image using a Gaussian kernel with σ equal to 8.5 has been made.

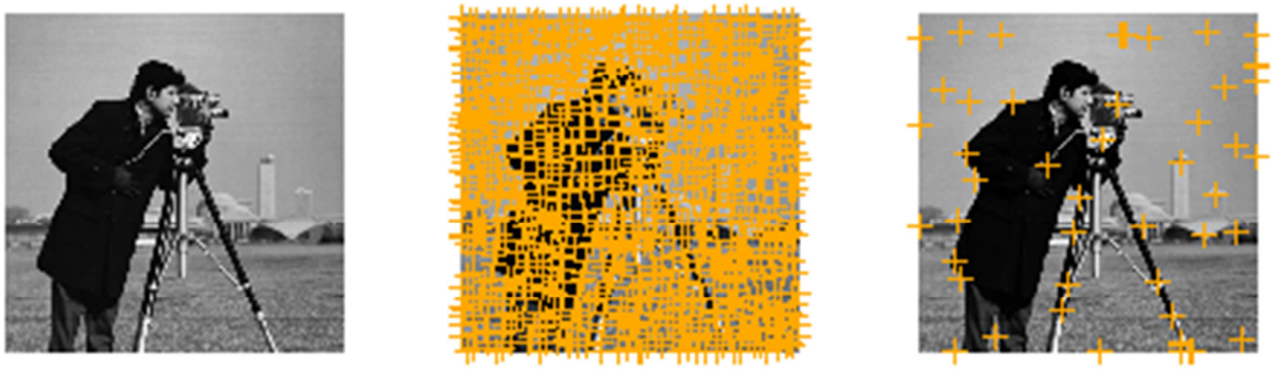


Figure 3.15: Harris corner estimation using MATLAB. Left: original image, middle: Corner detection on a lightly smoothed version of the original image, right: Corner detection on a heavily smoothed version of the original image

3.2.4 Log polar transformation

Log-polar transformation is commonly used for corner matching, since this approach makes the matching invariant to scaling and rotation. In this case it will also be well suited for recognition of coded targets used for calibration in Photo modeler in particular. The methodology of calibration and coded targets described in (Hartley, et al., 2003) and (Luhmann, et al., 2006).

The log-polar transformation is performed by:

$$\log(r) = \log(\sqrt{(x - x_0)^2 + (y - y_0)^2}$$

Equation 3.11

$$\theta = \arctan\left(\frac{y - y_0}{x - x_0}\right), for\ x \neq 0$$

Equation 3.12

A graphical interpretation of the transformation is illustrated in Figure 3.16

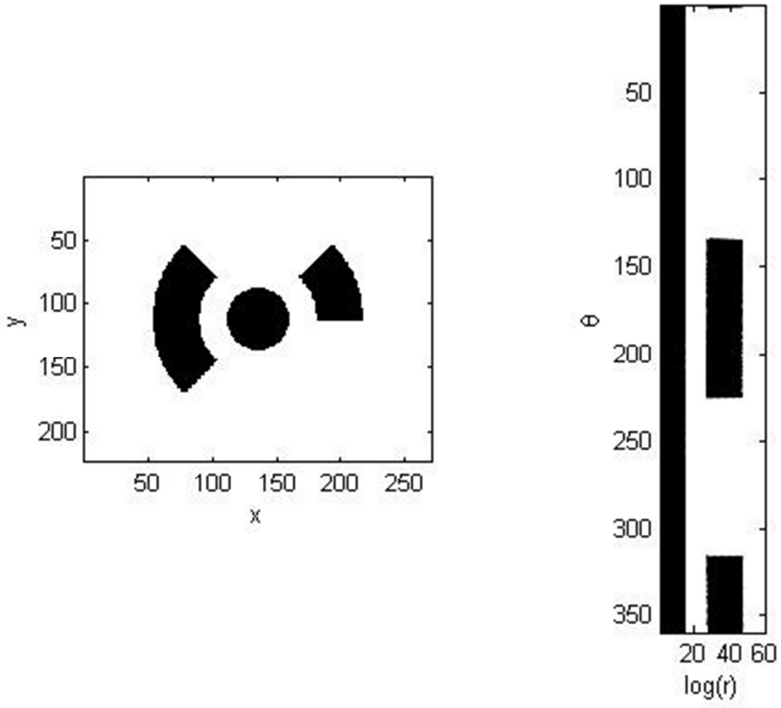


Figure 3.16: Left: Coded target in Cartesian coordinates; Right: Coded target transformed into log-polar coordinates

A periodically expansion of the reference patch along θ is performed in the point matching algorithms, since a displacement between the template and the reference patches along θ reflects a rotation. A displacement along log(r) reflects difference in scaling of the patches. The best fit between the reference patch and the template is achieved by cross-correlation.

3.3 Scale Space

A Scale space is like a new dimension provided by filtering a signal (an image is equal to a 2D signal) with a Gaussian or a LoG function with increasing σ followed by down sampling. Take a look on a 1D signal. By low-pass filtering the signal using a Gaussian kernel multiple times with increasing σ, we can make a 2D interpretation of the result as shown in Figure: 3.17.

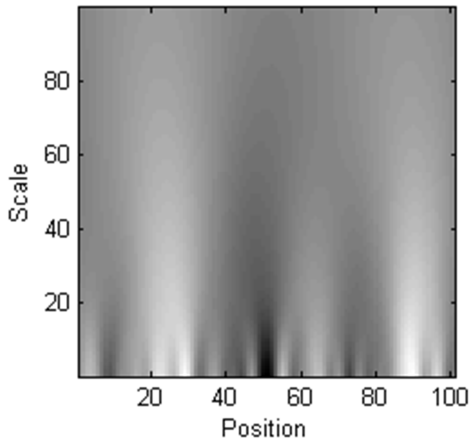


Figure: 3.17 Gaussian Scale space. Scale axis is proportional to σ

Each row in Figure: 3.17 represents the filtered 1D signal which could correspond to a single row in an image. σ of the Gaussian filter increases proportional with the Scale axis.

To enhance the new information between two levels of filtration with respect to σ, one can simply subtract the two images pixel wise from each other. A subtraction of a signal filtered with Gaussian kernels with different σ is equally the same as a filtration of the image with a Difference of Gaussian (DoG) filter. A DoG filter is approximately the same as a LoG filter that is explained in section 3.2.2. Since LoG is a bans-pass filter, a variation of sigma in the LoG filter corresponds to a displacement of the center frequency. This is illustrated for a one dimensional signal in Figure 3.18.

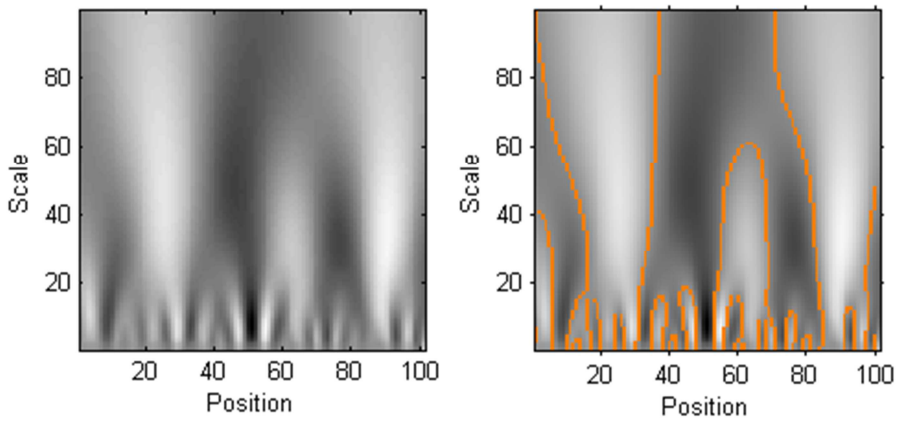


Figure 3.18: Left: LoG Scale space for a 1D signal. The Scale-axis indicates an increase of σ. Right: Labeling of Zero-crossings in Scale space

This can be directly transferred to 2D signals (images), as illustrated in Figure 3.19.

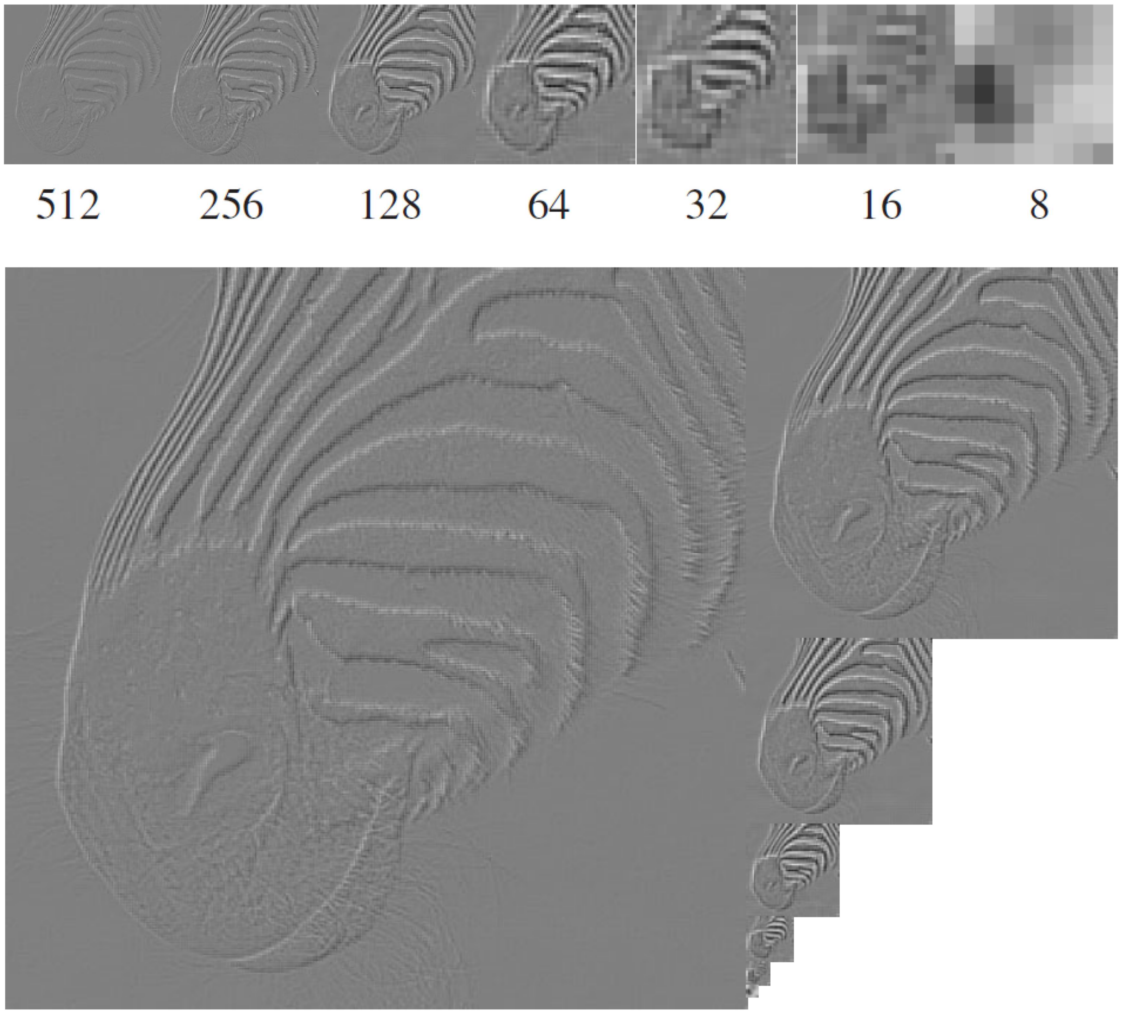


Figure 3.19: LoG scale space of a photo of a zebra muzzle running from 512x512 to 8x8 (Forsyth, et al., 2003). The numbers below the images indicates the pixel resolution of the images. Note the enhancement of the stripes in the coarser levels due to the correspondence between the frequency of the pattern and the center frequency of the filter.

3.4 Coarse-to-fine matching

LoG scale space is often used in point matching in combination with the correlation based matching technique. In (Marr, et al., 1978)it is proposed to find correspondences through a variety of scale spaces. By then one can limit the search range for the patches, which results in significantly improved disparity mapping. In (Marr, et al., 1978) they found that the correspondences of the zero-crossings of the Laplacian can be found within a disparity range equal to $\pm 2\sqrt{2}\sigma$. Once matches has been found the corresponding disparities is stored in a buffer, called the 2½-dimensional sketch. The disparities for matches in the 2½-dimensional sketch are used to register the matches in a finer scale. This can also be formulated by following stages[7]:

1. Convolve the two (rectified) images with LoG filters of increasing standard deviations: $\sigma_1 < \sigma_2 < \sigma_3 < \sigma_4$
2. Find the Zero-crossings of the Laplacian along the horizontal scan lines of the filtered images.
3. For each filter scale σ, match the zero-crossings with the same parity and roughly equal orientations in a $\pm 2\sqrt{2}\sigma$ disparity search range.
4. Use the disparities found at larger scales to get unmatched regions at smaller scales into correspondence.

Step 4 is illustrated in Figure 3.20.

[7] Cited from (Marr, et al., 1978)

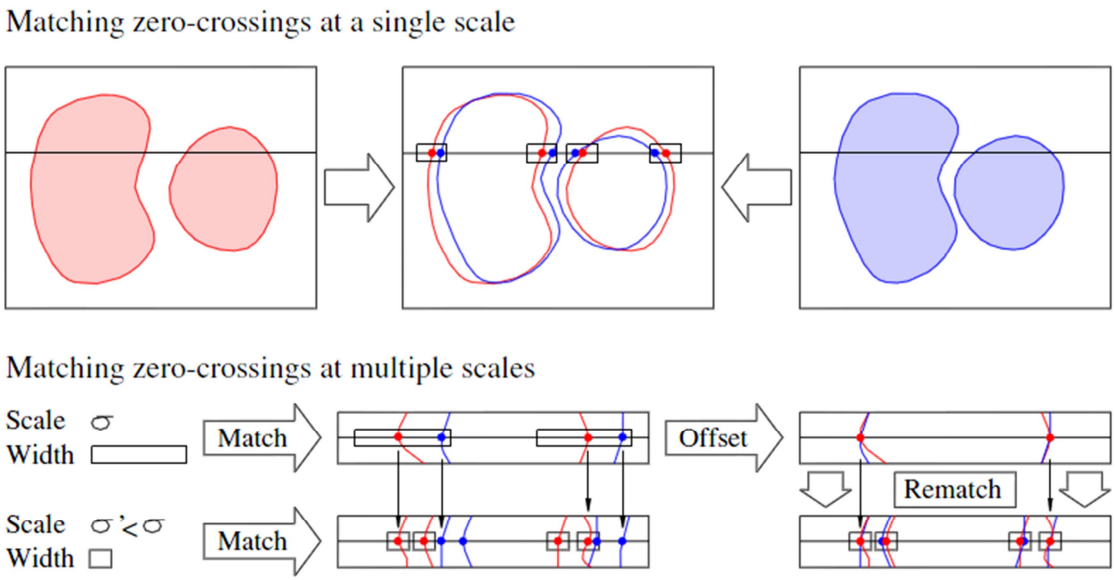


Figure 3.20: The illustration shows, that point match through Scale space is strongly dependent to perform an offset using the 2½-dimensional sketch. The term 'width' is equal to the disparity search space. Reference: (Forsyth, et al., 2003)

Constraints

Since it is possible to make a 2D kernel of LoG it is possible to perform a Scale space of an image instead of a 1 dimensional signal only. The disparity search range will then be deceased in both x- and y direction, which makes rectification and stereoscopic configurations redundant.

3.5 Visual Hull (VH)

Visual hull is a simple principle to reconstruct a 3D object. The simplicity of the algorithm makes VH to be computational fast and is therefore often used in relation to real time video tracing.

The object is reconstructed using the 2D silhouettes from multiple views. Each camera view forms a cone in 3D space is formed by the optical center and the silhouette in the image. The cone encloses a volume in which the 3D object is restricted to. With multiple camera views, the object volume is encapsulated to the common volume or envelope for all the cones as illustrated in Figure 3.21. From this it follows that concave shapes cannot be modeled using VH regardless of the amount of camera views.

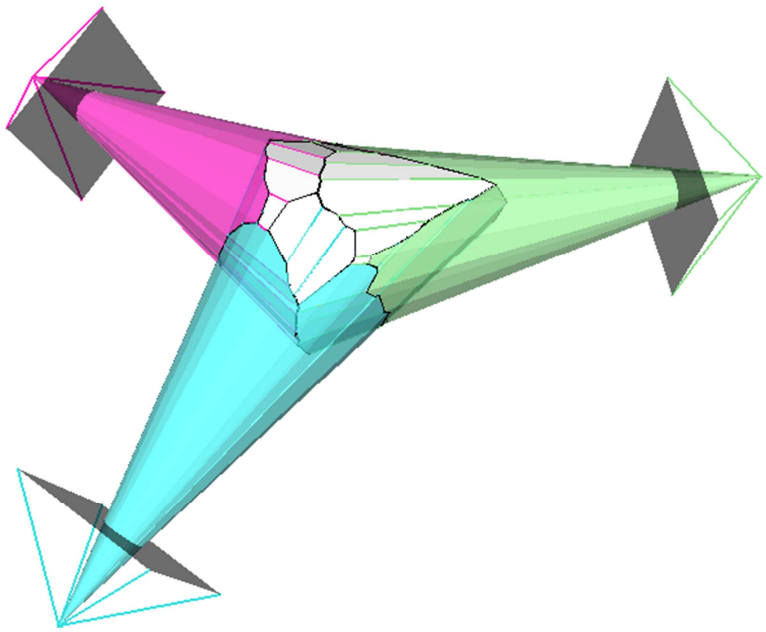


Figure 3.21: The Illustration shows the principle of visual hull. Cones defied by the optical center and silhouettes the image planes from multiple camera views are enclosing a volume that we are able to reconstruct. Reference: (Franco, 2007)

According to (Schneider, 2010) a pseudo code for a visual hull model can be expressed as:

1. Subdivide the 3D space of interest into voxels
2. Initialize all voxels to be part of the 3D object
3. **For each** voxel **in** 3D space
4. **For each** Camera **in** the configuration
5. Project the voxel into the image plane of the camera using the projection matrix $\mathbf{P_c}$
6. **If** the projection lies outside the silhouette, then classify the voxel 'v_n' as 'outside' the
7. 3D object

Even though the VH algorithm is fast and simple to implement it has a big disadvantage. The output is quite sensitive noise and it requires many cameras to make good approximations of curved shapes. Figure 3.22 shows res results of a rendered VH model of a woman, obtained with 4 and 8 cameras respectively.

Figure 3.22: Rendered Visual Hull models obtained by: (a) 4 cameras; (b) and (c): 8 cameras. Reference: (Corazza, et al., 2009)

Marching cubes is a typical approach to obtain the VH model.

3.6 Thin Plate splines

3.6.1 Interpolating Thin Plate spline

Thin Plate splines (TPS) are a conventional tool for interpolating surfaces. The interpolation function $f(x)$ is defined by:

$$f(x) = \beta_0 + \beta_1^T x + \sum_{i=1}^{N} \alpha_i \eta_m(\|x - x_i\|)$$

Equation 3.13

Where β_0 and β_1 are constants defining the linear surface parameters and α_i is a weighting constants to $\eta_m(\|x - x_i\|)$, defining the non-linear components of the spline. x_i defines an arbitrary point of the total N points that is already known. Denoting $\|x - x_i\| = r$, we can write $\eta_m(r)$ as:

$$\eta_2(r) = \begin{cases} r^2 \log r & \text{if } r > 0 \\ 0 & \text{if } r = 0 \end{cases}$$
$$\eta_3(r) = \|r\|^3$$

Equation 3.14

Where m is the number of dimensions of the space where the spine has to fit in.

According to Equation 3.13, this implies one spline knot per point in the input data, which makes TPS quite comprehensive for large numbers of input data. We are able to find the function $f(x)$ by solving the minimization problem:

$$R_{ss} = \sum_{i=1}^{N} \big(y_i - f(x_i)\big)^2$$

Equation 3.15

3.6.2 Smoothing TPS

By expanding Equation 3.15 with a regularization term, constraining the bending energy of the interpolating TPS.

$$R_{ss} = \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 + \lambda J_m(f)$$

Equation 3.16

Where λ is a smoothing factor, inverse proportional to the degrees of freedom of the spline.

$J_m(f)$ is corresponding to the Laplacian of $f(\boldsymbol{u})$, defined by:

$$J_m(f) = \int \sum_{i=1}^{m} \sum_{j=1}^{m} \left[\frac{\partial^2 f}{\partial u_i u_j}\right]^2 d\boldsymbol{u}$$

Equation 3.17

This implies for $m = 2$:

$$J_2(f) = \int \left[\frac{\partial^2 f}{\partial u_1}\right]^2 + \left[\frac{\partial^2 f}{\partial u_1 u_2}\right]^2 + \left[\frac{\partial^2 f}{\partial u_2}\right]^2 d\boldsymbol{u}$$

Equation 3.18

The Laplacian is simply the second derivative of the function, which is equal to the curvature. A large value of λ will therefore lead to a minimization of the curvature of the surface, where a small λ will make the TPS coverage towards a simple interpolation.

3.7 State of the art in 3D modeling

3.7.1 Improvement of the correlation based technique

In (Devernay, et al., 1994) they propose a correlation based technique were they integrate the derivatives of the disparity to cope with the non-parallel surface problems described in section 1.4.2.1.1. The technique is not significantly different to (Tang, et al., 2010) published in 2010 which indicates the method is still up to date and widely used.

The basic idea of the techniques described in the article is:

1. Perform the traditional disparity map using the standard correlation technique described in section 1.4.2.1.1.

2. Use Gauss-Newton optimization to warp a template window centered at pixel coordinate (x,y) of the left image into the reference image, such that SSD between the two regions are minimized. The warp is illustrated in Figure 3.23. Gauss-Newton requires a qualified initial guess that is set to the estimates found in step 1, with all derivatives set to zero. The Gauss-Newton algorithm can be regularized by the ordering constraint or putting a limit to the disparity gradient.

3. A new disparity map is computed from the optimized disparity estimates.

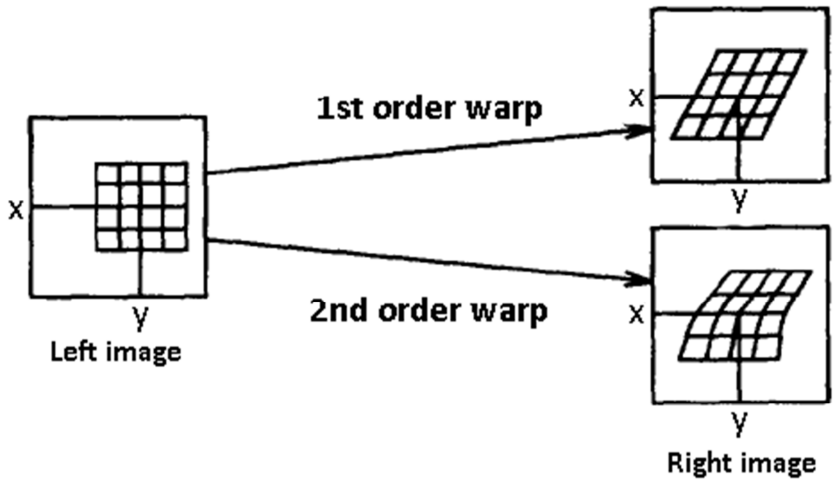


Figure 3.23: Polynomial warp optimized by Gauss-Newton method. Modified reference: (Devernay, et al., 1994)

Figure 3.24 shows the final result of the algorithm. It seem like a fairly improved reconstruction is attained by this approach

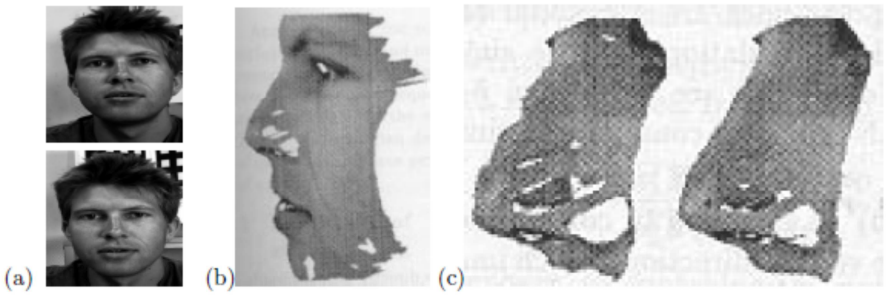


Figure 3.24: Reference: (Devernay, et al., 1994) Correlation based stereo matching: (a) a pair of stereo pictures; (b) a texture-mapped view of the reconstructed face; (c) comparison of the regular (left) and refined (right) correlation methods in the nose region.

3.7.2 A combined VH and correlation based approach

Another modern approach of 3D modeling is to create a VH model refined by photogrammetric point matching algorithms. This approach is presented in (Esteban, et al., 2004) among others. In (Esteban, et al., 2004) they fuse VH with the correlation technique described in section 1.4.2.1.1. The basics of this approach are to fit a deformable model such following minimization criterion is fulfilled:

$$\nabla E_{Mod}(S) = \nabla E_{Cor}(S) + \nabla E_{VH}(S) + \nabla E_{Reg}(S) = 0$$

Where $E_{Mod}(S)$ is the energy of the total model, $E_{Cor}(S)$ is the energy contribution from the correlation based model, $E_{VH}(S)$ is the energy contribution from the Visual Hull model and $E_{Reg}(S)$ is a regularization term. The optimization is provided by the classical snake approach.

The result of this approach with 12 equally spaced cameras is shown in Figure 3.25.

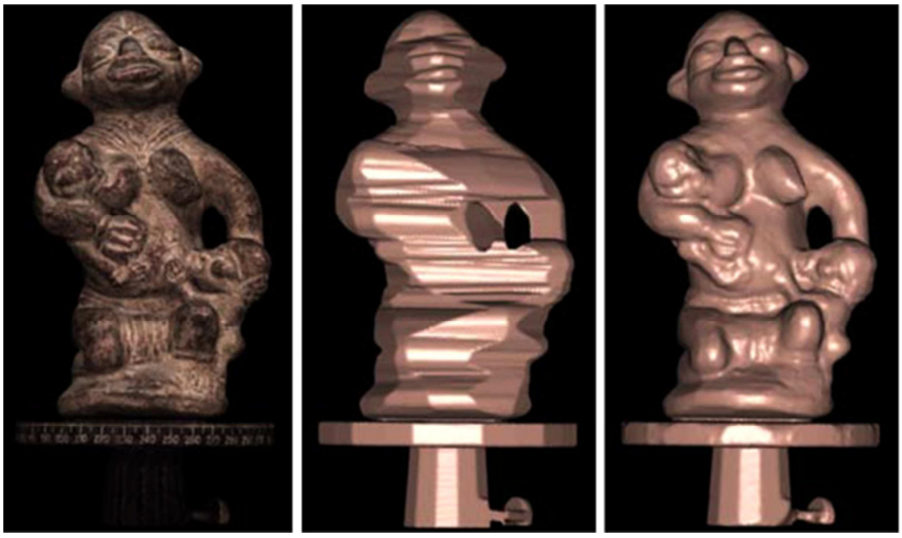


Figure 3.25: Left: Original image; Center: Visual Hull model; Right: Refined model. Reference: (Esteban, et al., 2004)

Due to the relatively low amount of cameras and the non-stereoscopic configuration, this approach seems to achieve quite promising results.

3.7.3 Patch based multi-view stereo algorithm

The 'Patch based multi-view stereo algorithm' (PMVS) is proposed by (Furukawa, et al., 2010). The algorithm creates polyhedral mesh models from images acquired from multiple camera views. As oppose to PM that is based on a correlation based technique, this method is matching features provided by Harris corners and 'Difference of Gaussian', that is described in details in section 3.2.2. An iterative expansion/filtration process is applied afterwards in order to create a denser point cloud and to extract outliers using constraints concerning visibility consistency and spread. Figure 3.26 illustrates how well a reconstruction is performed by a data set of 16 images of a dinosaur provided by Middlebury College.

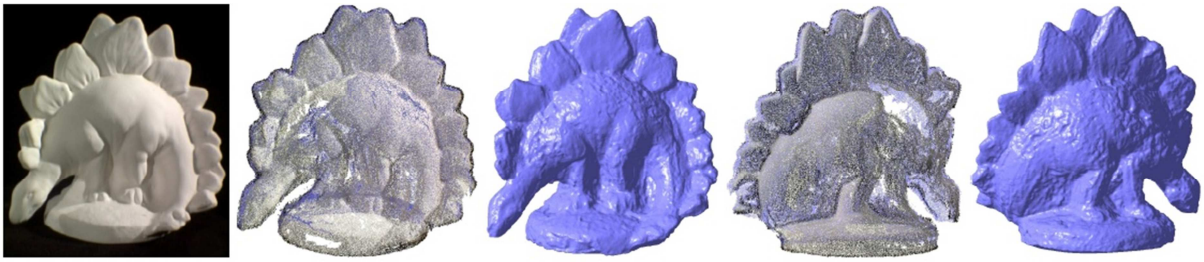


Figure 3.26: From left: Frontal photo of a total set of 16 photos, textured point cloud from front, shaded triangulated mesh from the back, textured point cloud from back and shaded triangulated mesh from the back

A comparison of the most promising methods can be found in (Scharstein, et al., 2009).

3.8 PhotoModeler processing

Following data processing was performed for all results in the thesis provided in PhotoModeler.

All cameras were calibrated (intrinsic calibration) according to the PM calibration tutorials. Likewise all processing was performed according to the tutorial: “Dense Surface Modeling - Coded Targets setup”. Essentially following steps was performed:

1. Automatic marking of the coded targets.
2. Process the coded targets (extrinsic camera calibration)
3. Idealizing the project (distortion correction)
4. Reprocess
5. Scaling
6. Trimming the region of interest in all the images
7. Dense surface modeling (DSM)

3.9 Camera specifications for Canon 350D 8M

Camera and lens Specifications	
Type	EOS 350D 8M (Canon)
Pixel resolution	3,456 (H) × 2,304 (V) pixels
Temporal resolution	N/A
Lens optic	**25 mm**
Focal length	2.5 cm
Pixel size	6.42 µm (H) x 6.42 µm (V)
F-number	10
Shutter time	0.3 s

Table 3.1: Camera and lens specifications

3.10 Camera specifications for Point Gray Chameleon 1M

Camera and lens specifications	
Type	Chameleon (Point Gray)
Pixel resolution	1,280 (H) x 960 (V) pixels
Temporal resolution	15 fps
Lens optic	Cinegon 1.4/12-0906 (Schneider)
Focal length	1.2 cm
Pixel size	3.75µm (H) x 3.75µm (V)
F-number	1.7
Shutter time	25 ms

Table 3.2: Camera and lens specifications

3.11 PhotoModeler test of a face

3.11.1 Introduction

A human face is reconstructed for various BH-ratios to illustrate the performance on a more complex shape with a weak texture. The test is performed on distances ‘H’ equal to two and four meters to observe the impact of a reduction in the spatial resolution.

The purpose of the test is to find out how low spatial resolution we can accept and still get a usable result. By this we can find out whether it is sufficient to use the VGA cameras mounted in a gait lab or not. The performance will be evaluated by a visual inspection of the modeled results.

To achieve results that would be more similar to what we could expect from a video camera, we have changed the Canon SLR eight megapixel cameras out with more conventional industrial cameras in one megapixel. However the spatial resolution is about the same, due to the focal lengths and pixel sizes are different as well.

Using Equation 1.11 and Equation 1.12 on page 33, describing the relations between the spatial resolution and the intrinsic parameters of the camera, we can estimate the resolution corresponding to an arbitrary ‘H’. The intrinsic parameters are listed in the table with camera specifications in section 3.10. For ‘H’ equal to 4 m we get a vertical and horizontal resolution at 1.3 mm and a depth resolution within an interval between 1.3-16.7 mm according to the length of baseline. For ‘H’ equal to two meters, we get a vertical resolution equal to 0.65 mm and a depth resolution within an interval between 0.65-6.2 mm.

3.11.2 Test setup

The setup is the same as described in section 1.6.2.1.

Materials

1. 2 x 1 mega pixel cameras
2. 2 x Lenses
3. 2 x Tripods
4. 2 x USB mini cables, 2 meters
5. 1 x spot light

The camera specifications are listed in appendix 3.10.

3.11.2.1 PhotoModeler processing

The setup and the data processing were performed just as described in section 1.6.2.1 about the precision dependence on the BH-ratio.

The values for the DSM parameters are listed in Table 1.1.

Category	Value
Sampling interval	3.5 mm
Matching region radius	5
Texture type	1
Sub-sampling factor	2

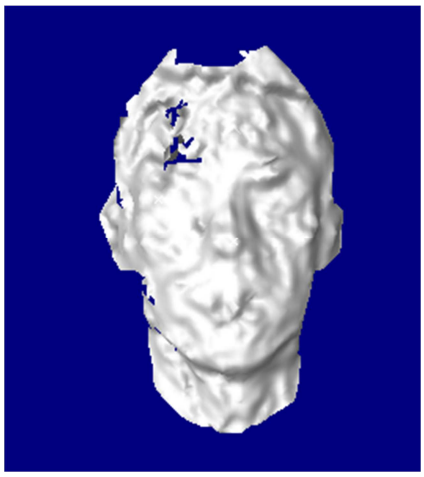
Table 3.3: DSM parameters

Only triangulation is performed on the mesh to perform a shaded model to give a better idea of the 3D shape. No smoothing or removal of outliers is obtained.

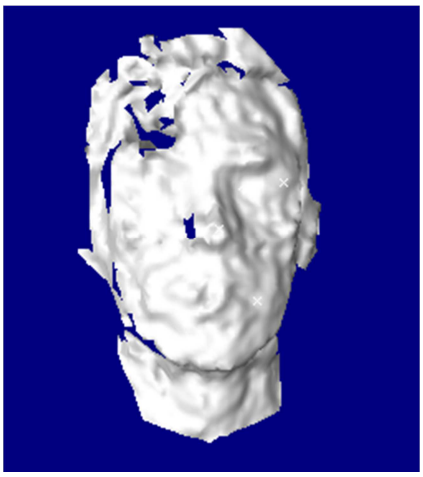
A 3D scan is performed as a golden standard to compare with the results acquired by the PM Scanners DSM algorithm. The Golden standard model was performed in the 3D lab at the Panum institute with equipment provided by 3DMD. The scan technique is based on a stereoscopic system supported with random pattern projection.

3.11.3 Results

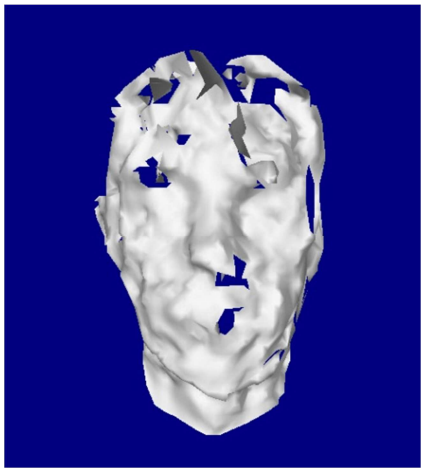
3.11.3.1 Face at 2 m



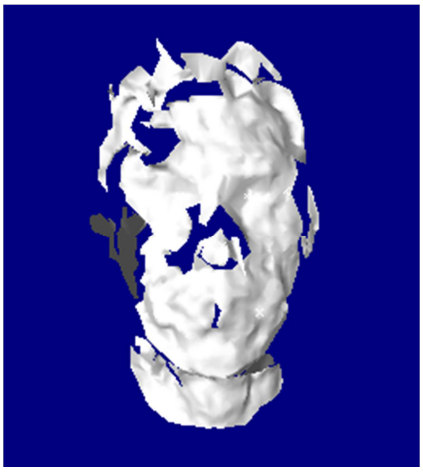
BH ratio = 0.1



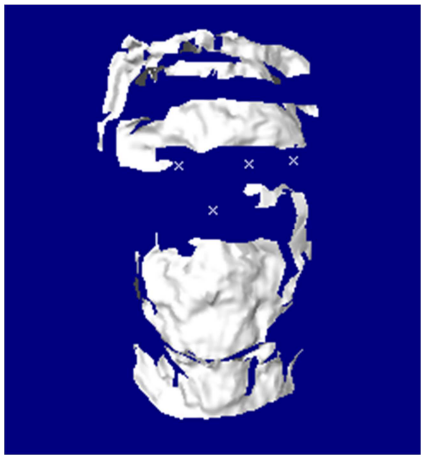
BH ratio = 0.2



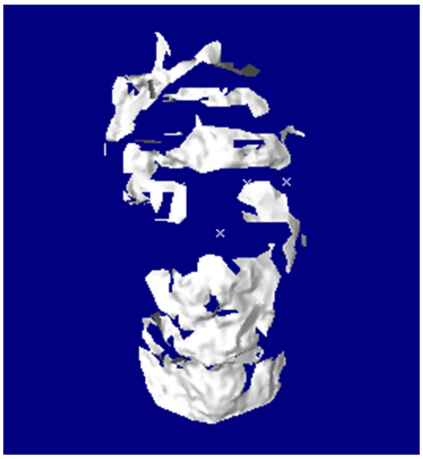
BH ratio = 0.3



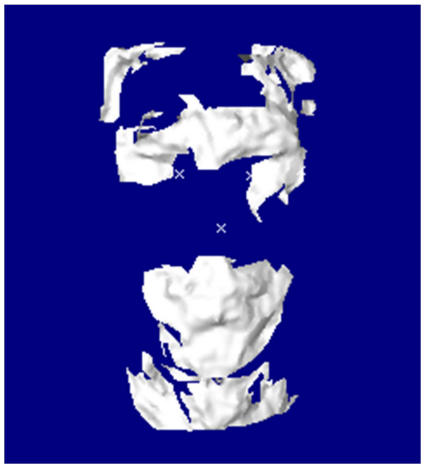
BH ratio = 0.4



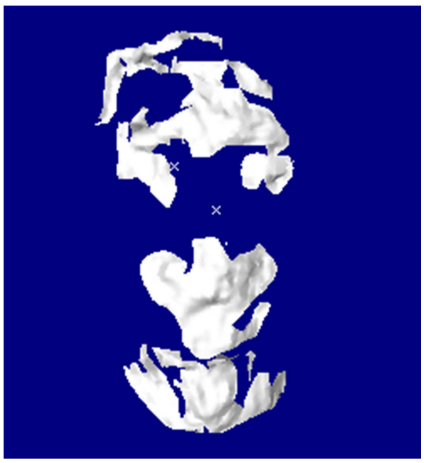
BH ratio = 0.5



BH ratio = 0.6



BH ratio = 0.7

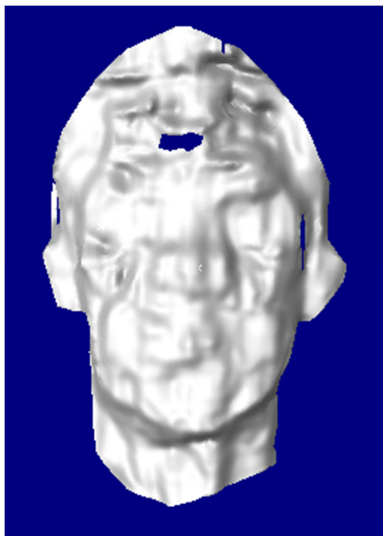


BH ratio = 0.8

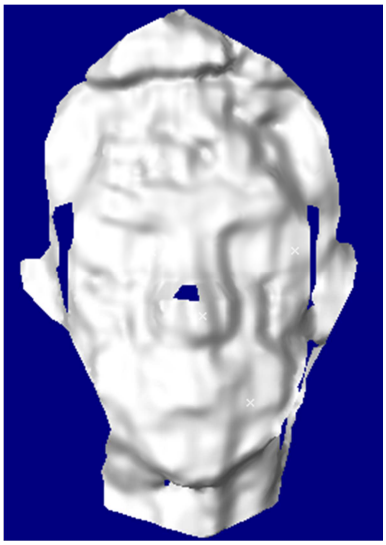
BH ratio = 0.9

BH ratio = 1.0

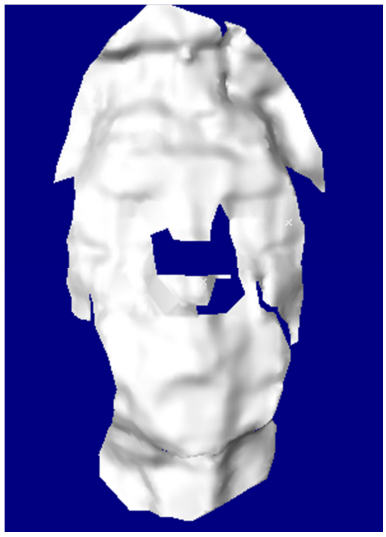
3.11.3.2 Face at 4 m



BH ratio = 0.08

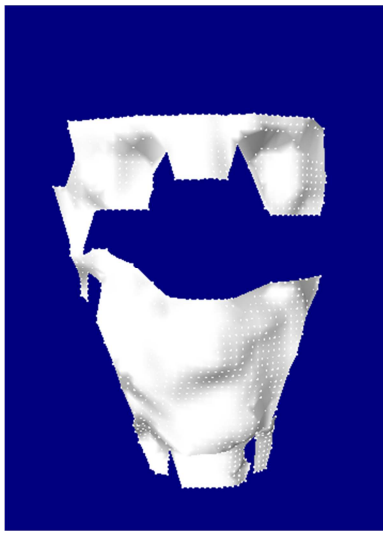


BH ratio = 0.13



BH ratio = 0.25

BH ratio = 0.38

 BH ratio = 0.50	N/A BH ratio = 0.63

3.11.4 Discussion

The results from 'H' equal to both two and four meters seems not to be acceptable at all, according to the criterions of a standard deviation less than 2.5 mm from the problem statement.

The illumination and the size of the photo sensor seem to have a major effect of the precision of the model. The size of the photo sensor can has an impact since the small surface area of the sensor implies less light sensitivity and is therefore inducing more noise.

Most of the uncovered surfaces seem to be either occluded to one of the cameras or far from fronto-parallel to the rectified image planes discussed in 1.4.2.1.1. The problems appear mostly on the sides of the face and in the nose region. Such problems cannot be fixed by improved resolution. Instead the problem might be solved by improving the number of stereoscopic systems. It is likely that locating the cameras with a vertical baseline instead of a horizontal baseline as performed in this test would improve the results as well, since the curvatures in the face are more prominent in the horizontal direction than the vertical direction.

BH ratios lower than 0.3 are quite noisy. The fluctuations seem to be smoothed as the BH ratio increases. The depth resolutions calculated in section 3.11.1, showed that small baselines provides serious errors even for small outliers. Decreasing the field of view of the cameras to the region of

interest is therefore of big importance. Some of the noise in the model could also be explained by the fact that the cameras are not completely synchronized. Since the cameras is recording with a frame rate at 15 fps, we might get a temporal difference up to 1/15 s between the acquired images. Small movements within this time interval will contribute to errors in the model.

Comparing the models acquired from two meters with the models acquired from four meters, we see the noise becomes smaller in the two meter models in the brow region in particular. This is primarily due to the difference of the spatial resolution. However even for two meters the errors seems to be too pronounced to provide a standard deviation below 2.5 mm.

The Field of View (FoV) who was roughly calculated to 1.2x1.4 meters with the 1M cameras at 4 meters. A VGA camera implies roughly the half pixel resolution as the 1M camera, and in a gait lab the cameras has to cove a region at approximately 2.5x2.5 m. Due to this it is very unlikely that VGA resolution is sufficient to get a satisfactory result.

Comparing with the golden standard, presented in section 3.12, we did not even get close to such level of detail. The reason the significant better result by using the golden standard approach is first of all because the spatial resolution is much higher than the configuration provided in our test. Since the intension was to provide results from distances that would be similar to a setup suited for full body acquisition, it would be misleading to provide close-up photos on the face for this test. Secondly the light conditions and the projection of random patterns at least as important as the spatial resolution. The random pattern projection can be replaced by a proper choice of textiles.

3.11.5 Conclusion

Either the results from 2 m or the results from 4 m seemed to be acceptable even for BH-ratios less than 0.4. The standard deviation seems to be significantly larger than 1 mm as we saw for the flat surface with at the same BH-ratios. However the texture has shown to be of big importance. A considered choice of textured cloth and illumination is elementary to achieve accurate models and might even be sufficient together with synchronized cameras to achieve satisfactory results for the 2 m configuration. VGA cameras are not optional to perform the 3D models. The low pixel resolution will provide too large errors due to the low spatial resolution and will also smooth out unique information in the weak texture of the skin.

3.12 Golden standard models

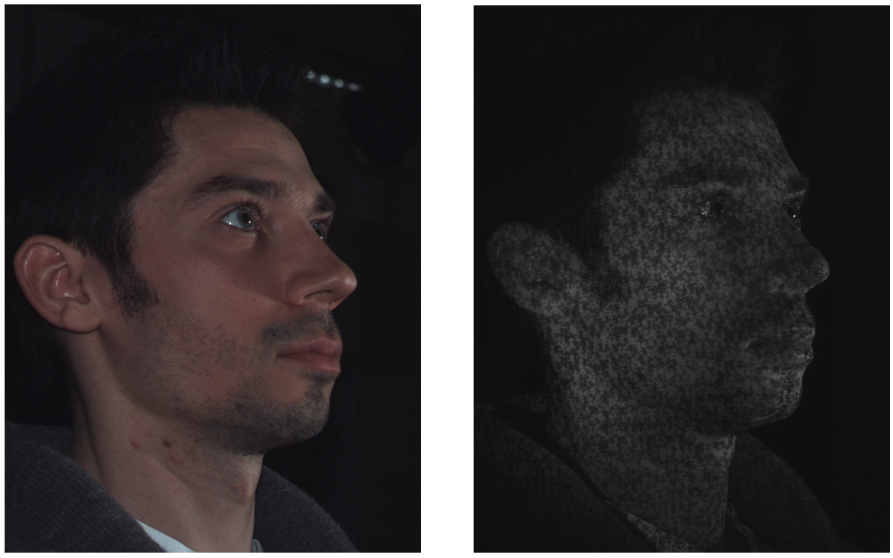


Figure 3.27: Left: Photo acquired by the 3DMD scanner system in 3D lab at the Panum institute; Photo with projection of random pattern used to obtain the dense surface.

Figure 3.28: Golden Standard model provided by 3DMD scanner

3.13 Testing skin as texture

3.13.1 Human textured with small random patterns

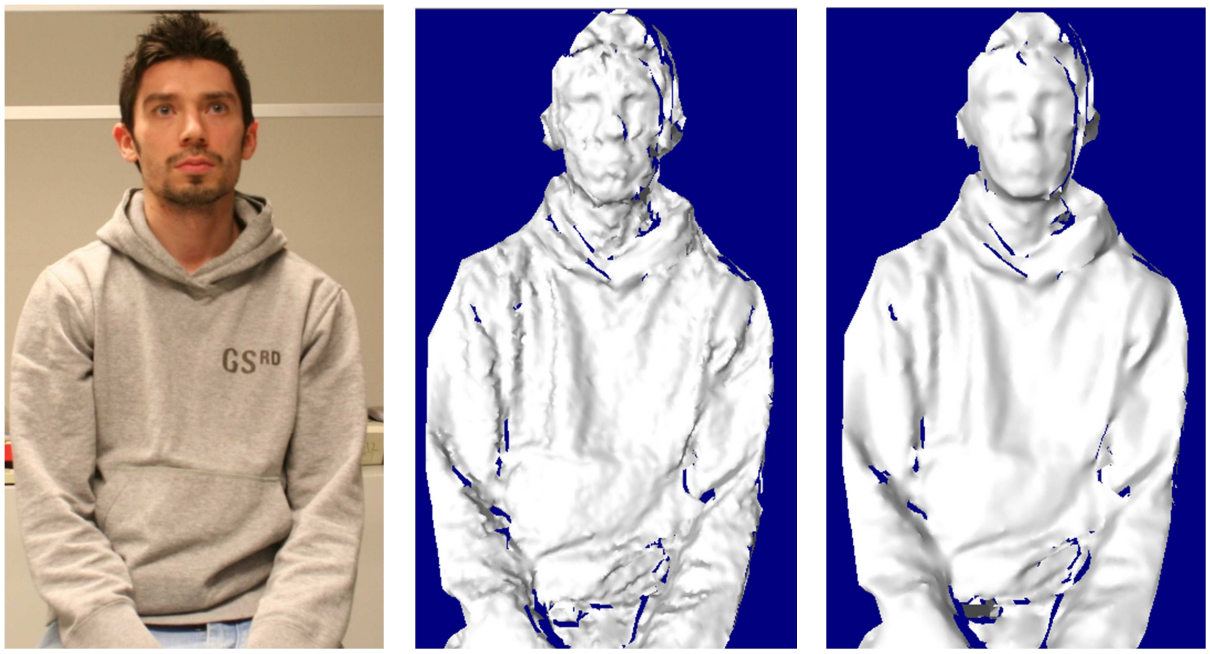


Figure 3.29: Left: Photo; Center: Triangulated mesh without processing[8]; Right: Processed triangulated mesh

[8] Processing constitutes of smoothing and removal of outliers

3.13.2 Naked upper body

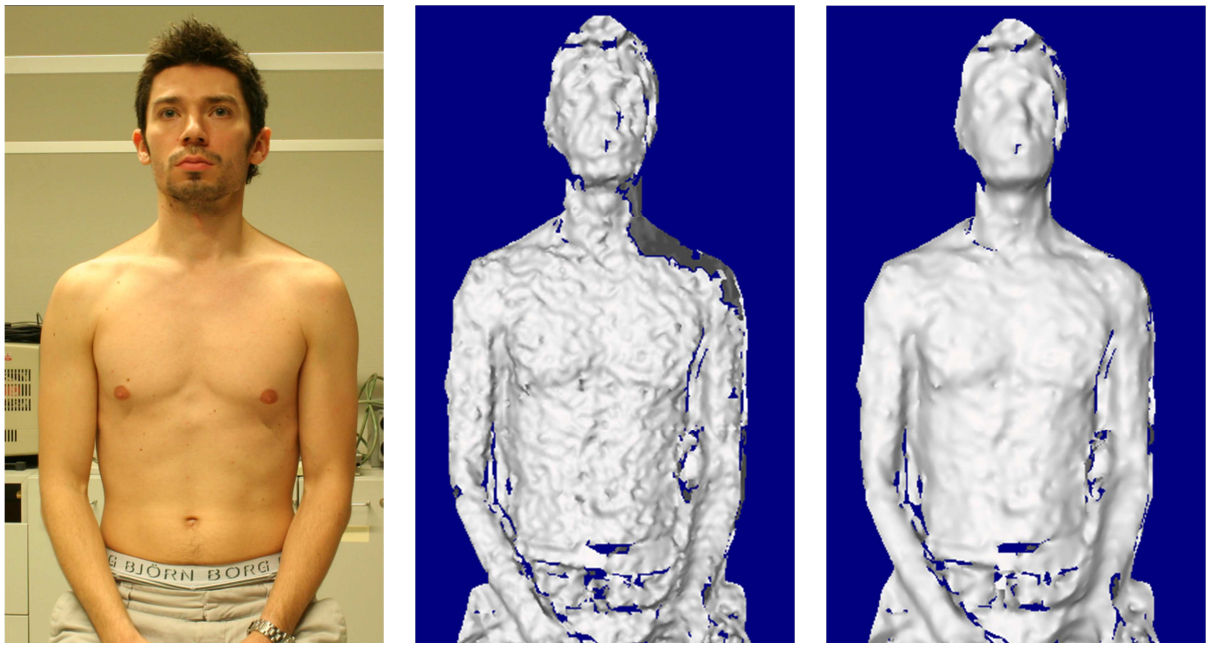


Figure 3.30: Photo; Center: Triangulated mesh without processing; Right: Processed triangulated mesh

3.14 Failure of shell fitting

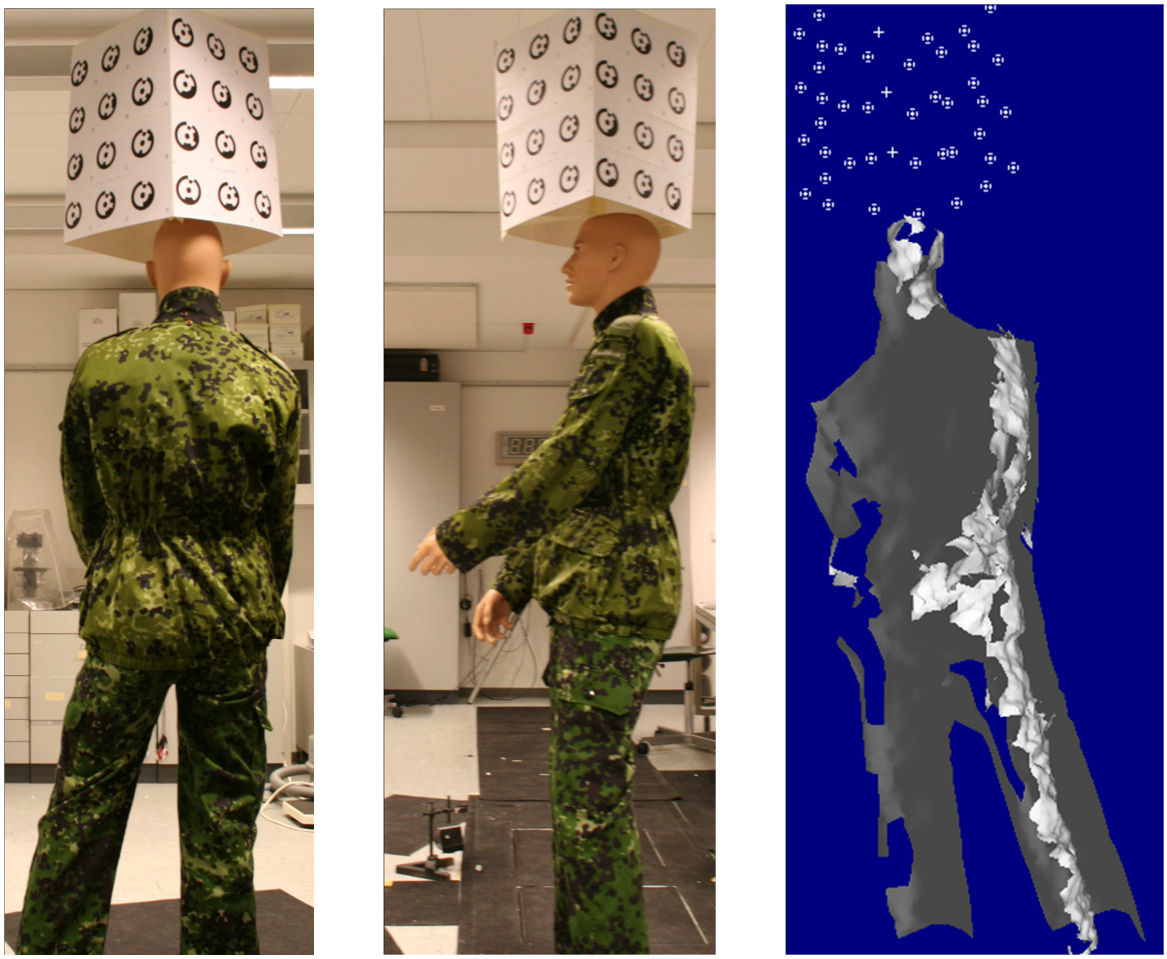


Figure 3.31: Left: Dummy from back; Center: Dummy from left; Right: Model composed of the mesh provided from the two stereoscopic viewpoints.

The fitting of the shells failed due to the narrow area of coded targets centered above the head. Figure 3.31 illustrates how the fitting fails. The Distance between the shells becomes larger and larger, the larger the distance of the shells becomes to the coded targets. The solution of this problem was to setup coded targets along the legs such they do not occlude the model for the cameras.

3.15 Full models

3.15.1 Dummy textured with large random pattern

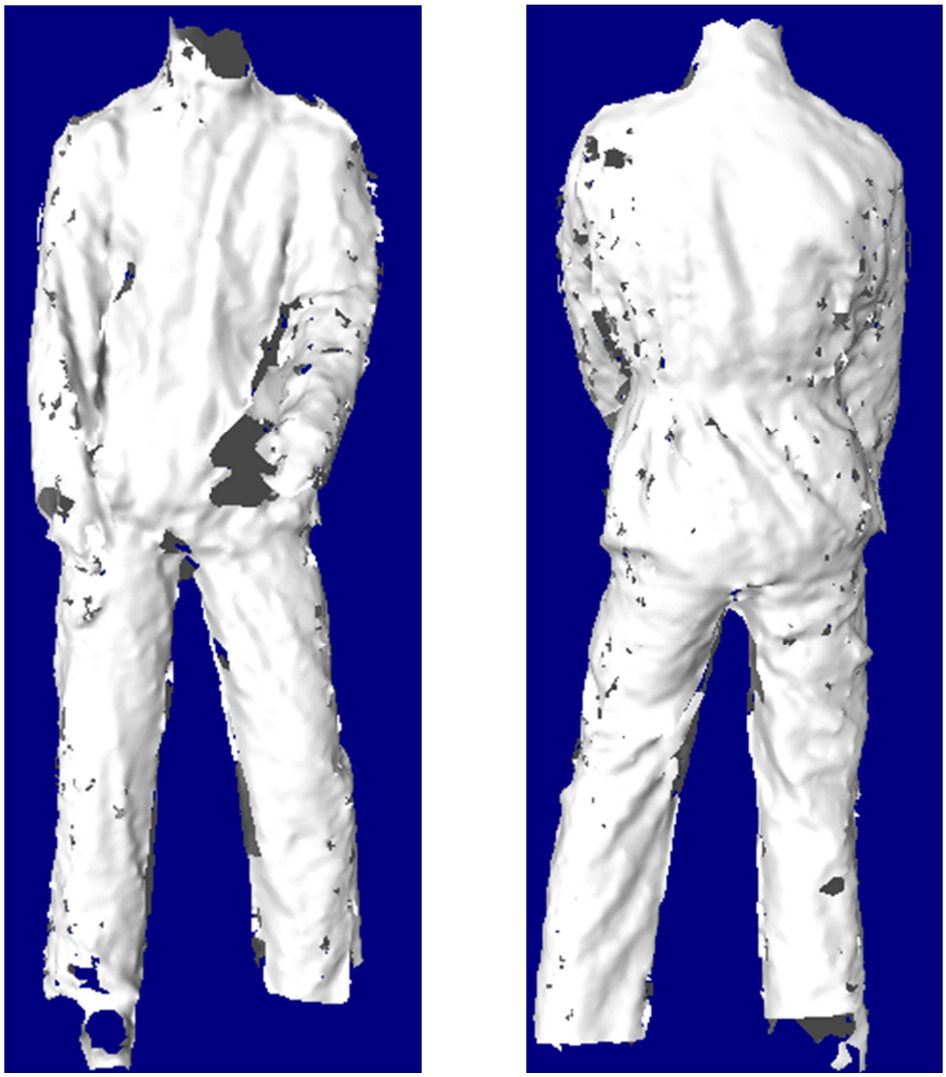


Figure 3.32: Left: Anterior view of the uniform; Right: Posterior view of the uniform

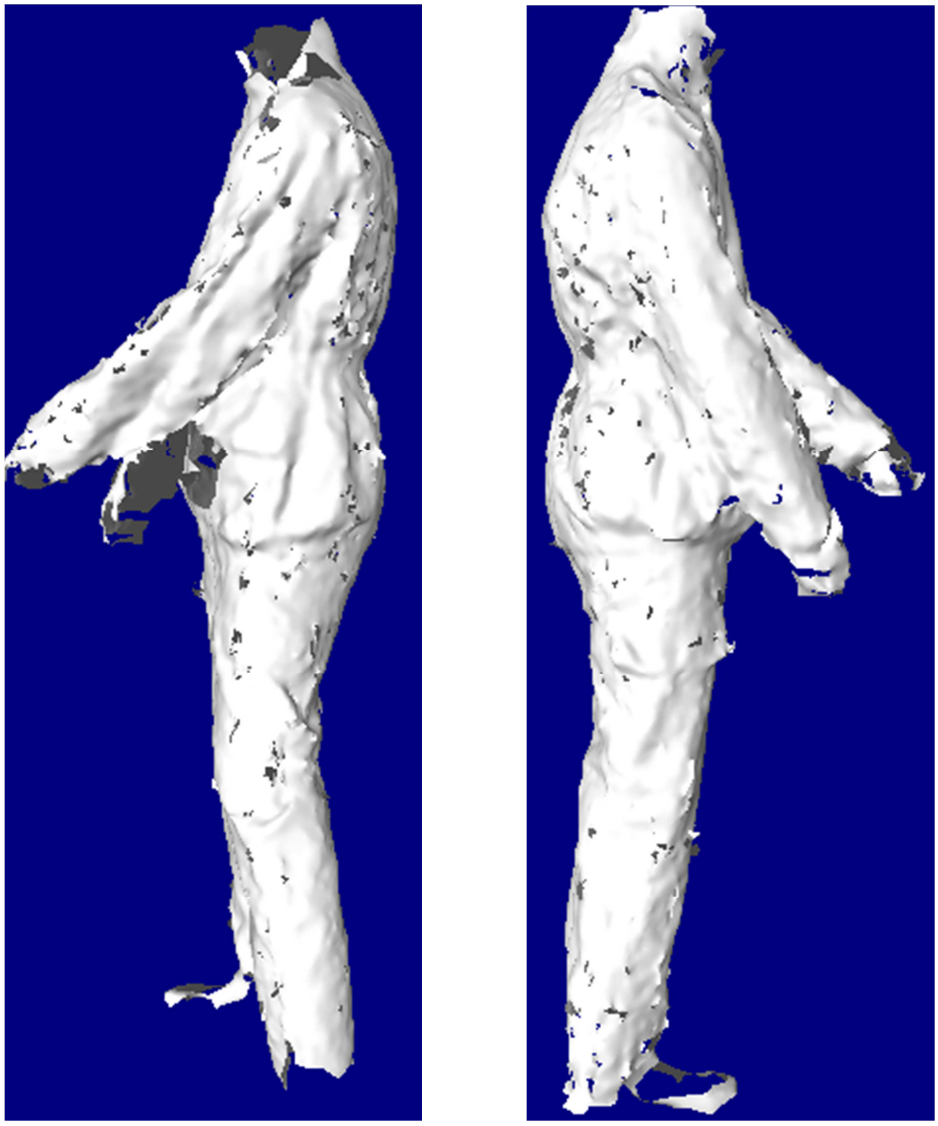


Figure 3.33: Left: Lateral view from right; Right: Lateral view from left

3.15.2 Dummy textured with structured pattern and weak random pattern

Figure 3.34: Anterior view of the dummy wearing jeans and checkered shirt; Right: Posterior view of the dummy

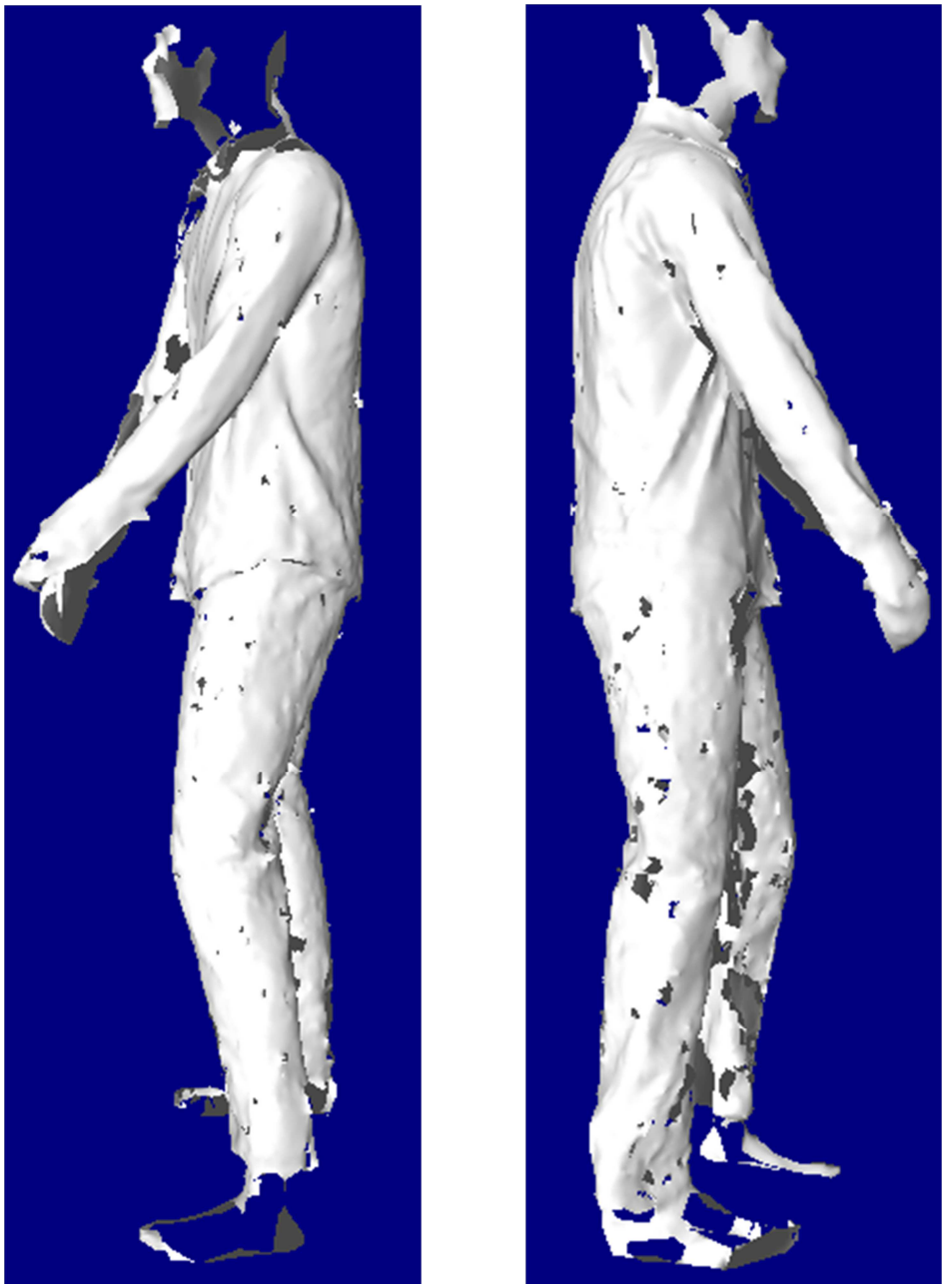


Figure 3.35: Left: Lateral view from right; Right: Lateral view from left

3.16 Anatomy

From the former section we saw that nearly all kinematic information can be extracted through the estimation of the joint centers. This section describes the anatomical location of the joint centers of the hip, knee and ankle since those joints are in main focus in gait analysis. Figure 3.36 shows a sketch of the skeleton system and the human body merged together.

Figure 3.36: Skeleton system. Left Ventral view; Right Lateral view. Source (Mc Graw Hill, 2011)

Figure 3.36 provides an insight in how the joints are located according to the superficial surface of the body. We see that the hip joint lies deep into the body, which makes it problematic to estimate the location from the superficial information, which will be provided through a reconstructed body model. The curvatures of the body might be useful to perform a body segmentation, since concavities are observed at the nearby the knees and ankles. More complex shapes are provided nearby the hips, but could most likely be segmented on basis of the concavities as well. We will return to this discussion later on in this section.

3.16.1 Joints of the lower limbs

Joints in general are divided into three classes: Fibrous, cartilage and synovial joints. All joints that are of interest in a biomechanical point of view are the synovial joints, characterized by the fluid filled joint cavities encapsulated by a synovial membrane. Synovial joints are further subdivided into six classes describing their functionality: Plane, saddle hinge pivot, 'ball and socket' and ellipsoid. The three joints that we like to extract for gait analysis are the Genu (knee), Talocrural (ankle) and Coxal (hip) joints. Those are described just below.

3.16.1.1 Hip joint

The hip joint are considered as a 'ball and socket' joint. It consists of Femur and Pelvis. The neck of femur extends though femur and forms the head of femur that functions as the ball of the joint. The pelvic bone forms a cavity called Acetabular labrum that functions as the socket.

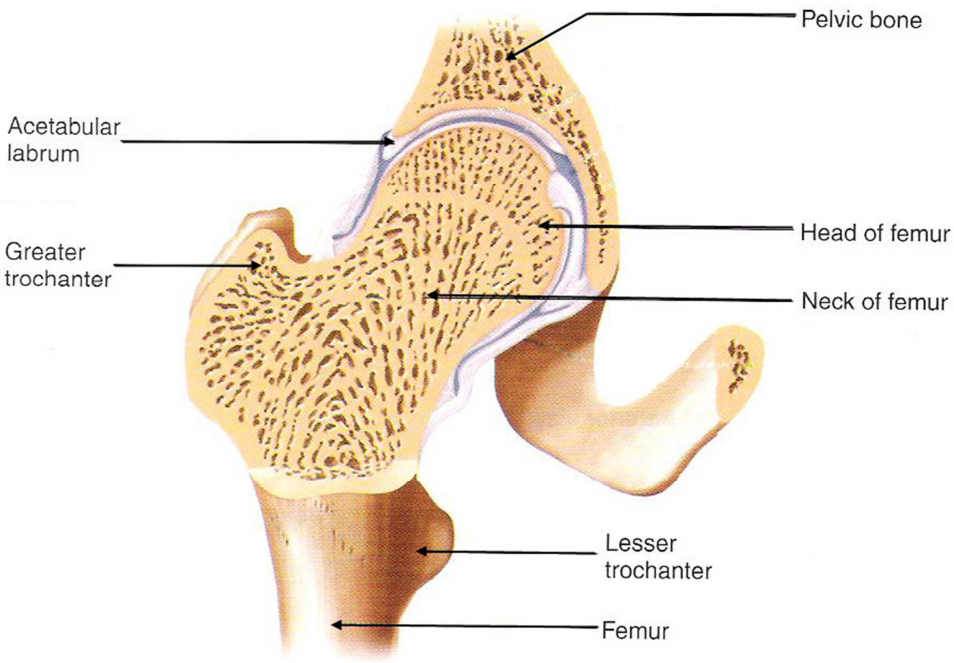


Figure 3.37: Right hip joint, anterior view Reference: Modified from (Seeley, et al., 2006)

The socket joint provides three degrees of freedom for the rotational movements. The hip movements are termed as follows:

1. **Lateral rotation** outward rotation of the thigh
2. **Medial rotation** inward rotation of the thigh
3. **Extension** moving the thigh posterior
4. **Flexion** moving the thigh anterior
5. **Abduction** moving the thigh in lateral direction
6. **Adduction** moving the thigh medial direction

3.16.1.2 Knee joint

The knee joint is considered as a hinge joint, providing one degree of freedom for rotational movement. The rotational axis goes transverse through the condyles of femur, which forms the joint with tibia and the meniscus as illustrated in Figure 3.38.

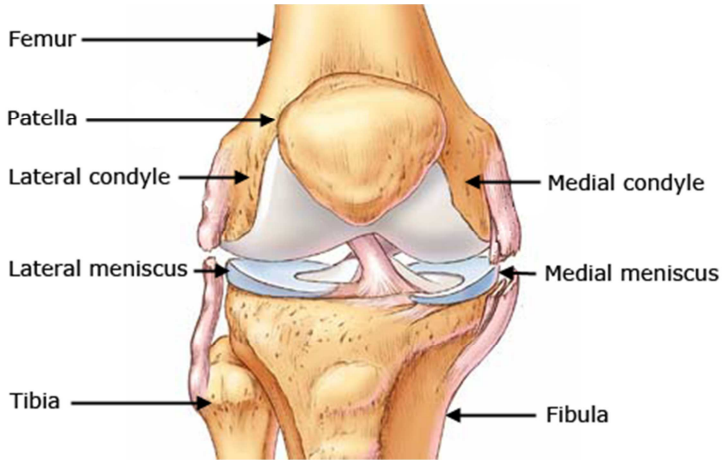


Figure 3.38: Right knee joint, anterior view Reference: Modified from (The Center for Orthopaedics & Sports Medicine, 2003)

The movements of the knees are termed as follows

1. **Extension,** providing an anterior rotation of the shank
2. **Flexion,** providing a posterior rotation of the shank

3.16.1.3 Ankle joint

The ankle is considered as a hinge joint as well. Providing one degree of freedom to provide following movements:

1. **Dorsiflexion**, pushing the toes upwards
2. **Plantar flexion**, pushing the tows downwards

However the ankle joint are also able to perform limited eversion and inversion, that means to perform rotations such the plantar surface faces lateral and medial respectively. This makes the ankle joint function more like a modified 'ball and socket' joint than a regular hinge joint.

The joint is formed by tibia, fibula from the shank and talus from the foot. Tibia and talus forms the majority of the joint, where fibula only plays a supporting role. A medial view of the ankle is shown in Figure 3.39.

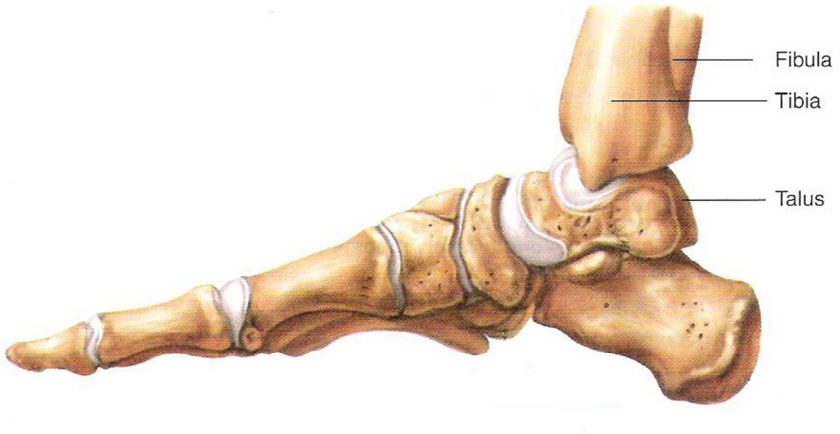


Figure 3.39: Right ankle, medial view. Reference: Modified from (Seeley, et al., 2006)

3.17 The Hungarian algorithm

The Hungarian algorithm is an optimization algorithm. The matrix interpretation gives the solution with the minimal costs of matching the row parameters with the column parameters in a matrix. The elements in the matrix have to represent the costs of matching the particular row parameter to the particular column parameter.

The Hungarian algorithm can be written as a pseudo code as follows:

1. Start with a n x n cost matrix
2. **For each** row {
3. Subtract the min. value in the row
4. **For each** column
5. Subtract the min. value in the column
6. With the minimum numbers of lines (columns or rows), extract all zeros
7. Suppose the number of extracted lines are k
8. **If** k<n
9. Let m be the minimum if the elements not extracted. Subtract m from every elements that are not extracted
10. Add m to every element covered by two lines
11. Return to 6.
12. **If** k==4
13. Perform a match between the columns and rows which common element is zero. If a row or column contains multiple zeros, then pick an arbitrary combination

The pseudo code is modified from (Castello, 2007).

3.18 Test of the shape context 3D function in MATLAB

Point matching using shape contexts has been applied in two ellipsoids with different parameters. The two ellipsoids are illustrated in Figure 3.40.

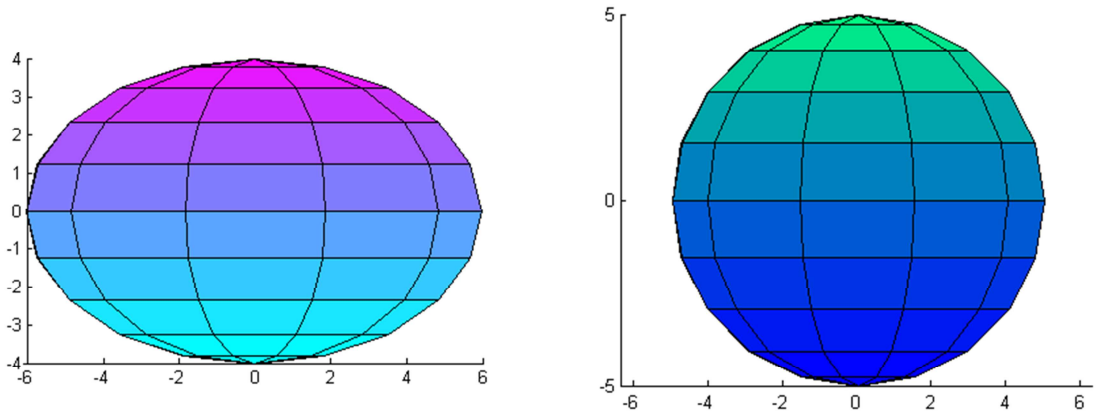


Figure 3.40: Left: Reference model; Right: Template model

Calculating the costs using the function presented in Equation 2.2 on page 82 and finding the solution with the minimal costs using the Hungarian algorithm we get the result presented in Figure 3.41. The max radius of the shape context histogram is set to 60% of the model size. This means that all points within a range of approximately 3.5 units are appended into the shape context histogram for a particular point.

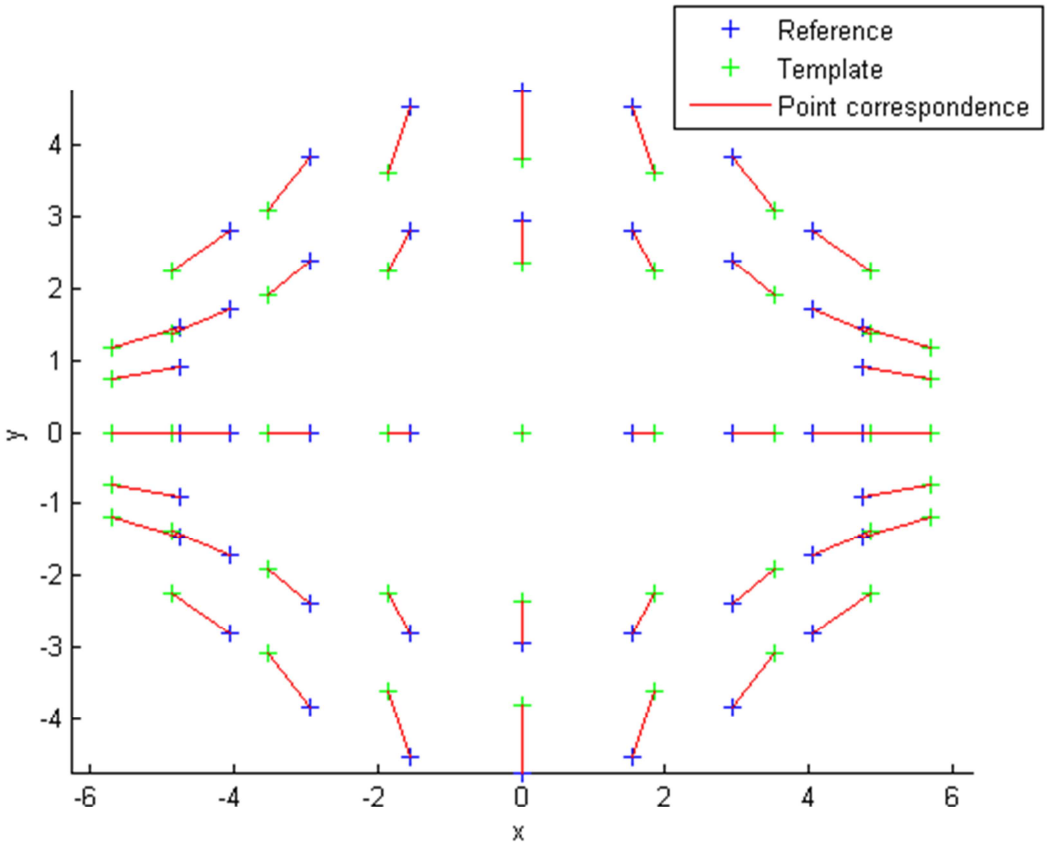


Figure 3.41: Diagram illustrating the point correspondences matched by 3D shape contexts and the Hungarian algorithm

As expected we see a correct estimation of the point correspondences. No suspicious crossovers has appears in the figure, which indicates the approach is working out fine for simple models.

3.19 Test of the curvature function in MATLAB

3.19.1 Introduction

The MATLAB-file 'Curvature.m' has been created with inspiration from the approach presented in (Meyer, et al., 2000) to calculate the mean curvature of a 3D point in a triangulated mesh. The mean curvature is defined by the mean of the principal curvatures. The estimation of the curvature is part of the expansion of the cost function presented in Equation 2.3 in section 2.6.1.2. The algorithm has been tested on a 3D mesh of a hand. The hand is chosen for the test, because a hand represents various curvatures, form the smallest in the wrist region to the highest by the fingertips. Using the hand model, which is created like the models we have used to the tests in part II, we would also be able see how the curvatures of the models actually look like. The model of the hand is acquired by the 3D scanner in 3D lab at the Panum institute.

3.19.2 Results

Figure 3.42 shows the results of the algorithm applied on the raw data and a smoothed version of the hand.

Figure 3.42: Left: hand model, plantar view; center smoothed hand model, plantar view; Right: Color bar with labeled values in [mm^{-1}].

According to the raw model, it was expected to see a much smoother surface. The dense triangulated mesh seems to provide many fluctuations in the curvature, which is considered as noise. It is hard to see whether the algorithm is performing right or not, based on the raw model. A smoothing has therefore been applied on the model, which is illustrated in the right model. Here we truly see higher curvature values at the fingertips and along the fingers as oppose to the curvature on the arm as expected. Whether the raw model is infected by noise or not, the small fluctuations might disturb the cost function, since the curvatures seems not to be significantly different from one part of the hand to another. To find out how the expansion of the cost function are performing, registrations with various α values has been tested. The results are presented below.

3.20 Testing the cost function for various alpha values

3.20.1 Introduction

Recalling from section 2.6.1.2, the expanded cost function was defined as follows:

$$C_{mn}^{e} = \alpha C_{mn} + (1 - \alpha) C_{mn}^{c}$$

Equation 3.19: Expansion of the cost function proposed by (Xiao, et al., 2009)

Through this section, the performance of the registration will be tested for various α-values in the cost function. The stretched hand models from the experimental results in section 2.7.1.3 are used for the test. As formerly explained the models are acquired in 3D lab in the Panum institute.

3.20.2 Results

3.20.2.1 α= 1.00

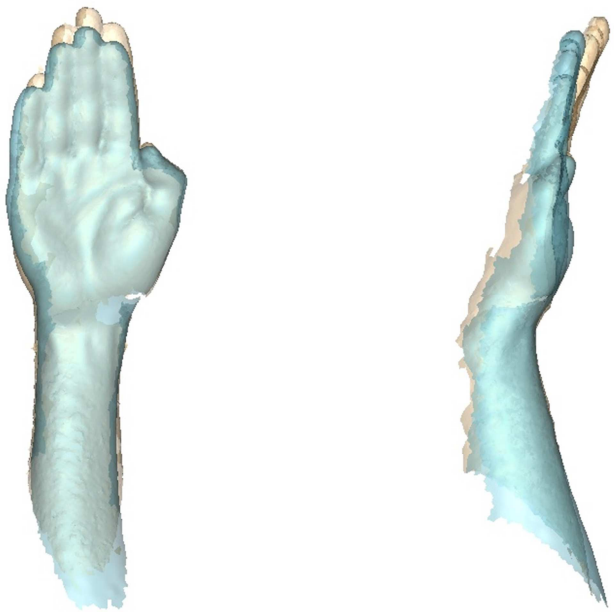


Figure 3.43: α = 1.00: Reference model (golden) and registered template model (cyan) plotted on top of each other from a plantar and medial view respectively

Approach for Error estimation	Value [mm]
Shortest distance RMS	4.6
Hungarian RMS	11.8

Table 3.4: Error estimation

3.20.2.2 $\alpha = 0.75$

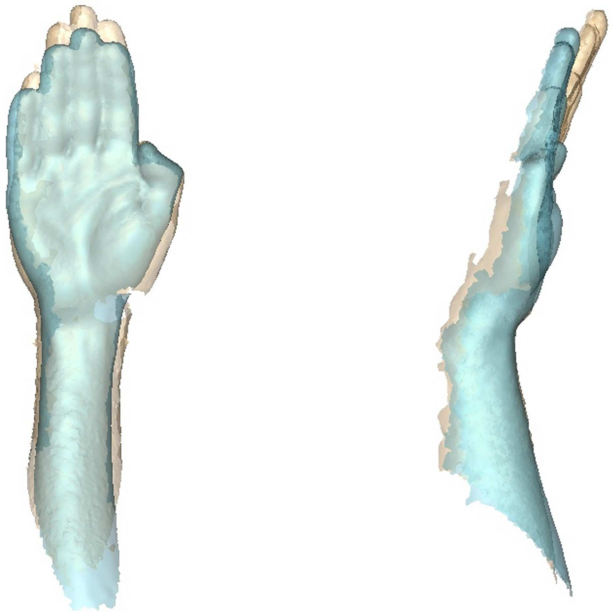


Figure 3.44: $\alpha = 0.75$: Reference model (golden) and registered template model (cyan) plotted on top of each other from a plantar and medial view respectively

Approach for Error estimation	Value [mm]
Shortest distance RMS	4.9
Hungarian RMS	12.2

Table 3.5: Error estimation

3.20.2.3 α = 0.50

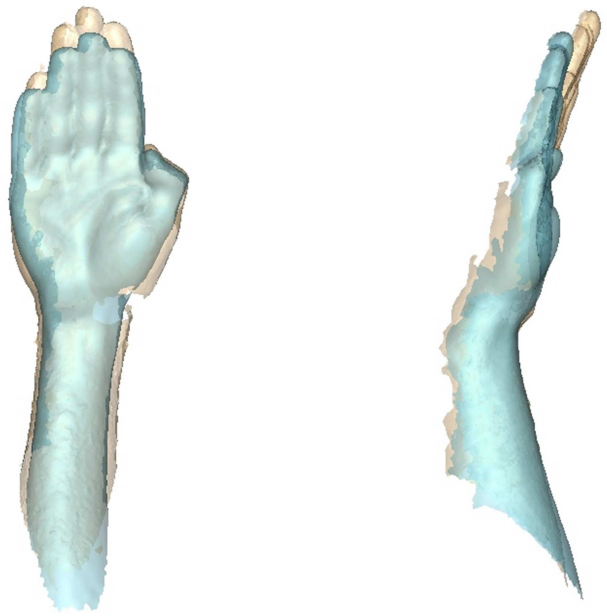


Figure 3.45: α = 0.50: Reference model (golden) and registered template model (cyan) plotted on top of each other from a plantar and medial view respectively

Approach for Error estimation	Value [mm]
Shortest distance RMS	4.9
Hungarian RMS	13.0

Table 3.6: Error estimation

Comparing the results for the various α-values, we do not see any significant differences. However if we look closer at the fingertips and the wrist we see a minor difference.

The shortening of the fingers varies for the different α values. α=1.00 is less shortened than the others.

In the area around the wrist we see the model has become narrower for α= 0.75 and α = 0.50. This artifact cannot be seen on the model for α = 1.00. From this it is concluded that the best result for this registration is achieved with α = 1.00.

3.21 Pose estimation, results of point tracking

3.21.1 Registration of flexing fingers

These results show the fatal consequences of large differences between the template and the reference models. The demonstration is based on a model flexing the proximal finger joints.

3.21.1.1 Template- and registration model

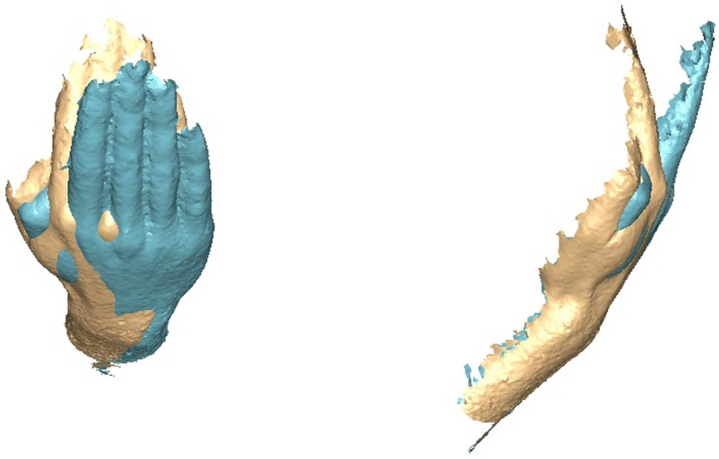


Figure 3.46: Left: Dorsal view of the template- (cyan) and the reference model (golden). Right: Lateral view of the models

In Figure 3.46 we see a large flex angle for the fingers that are going to be registered to one another.

3.21.1.2 Registration- and fitted template model

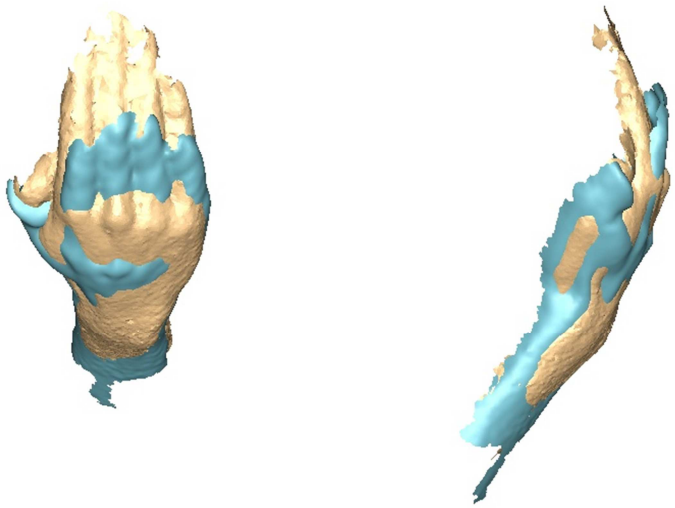


Figure 3.47: Left: Dorsal view of the registered template- (cyan) and the reference model (golden). Right: Lateral view of the models

Figure 3.47 shows a complete failure of the registration of all the fingers. Let's take a closer look on the point correspondences and the flow charts to see what's happening with the mesh.

3.21.1.3 Point correspondences between template- and reference model

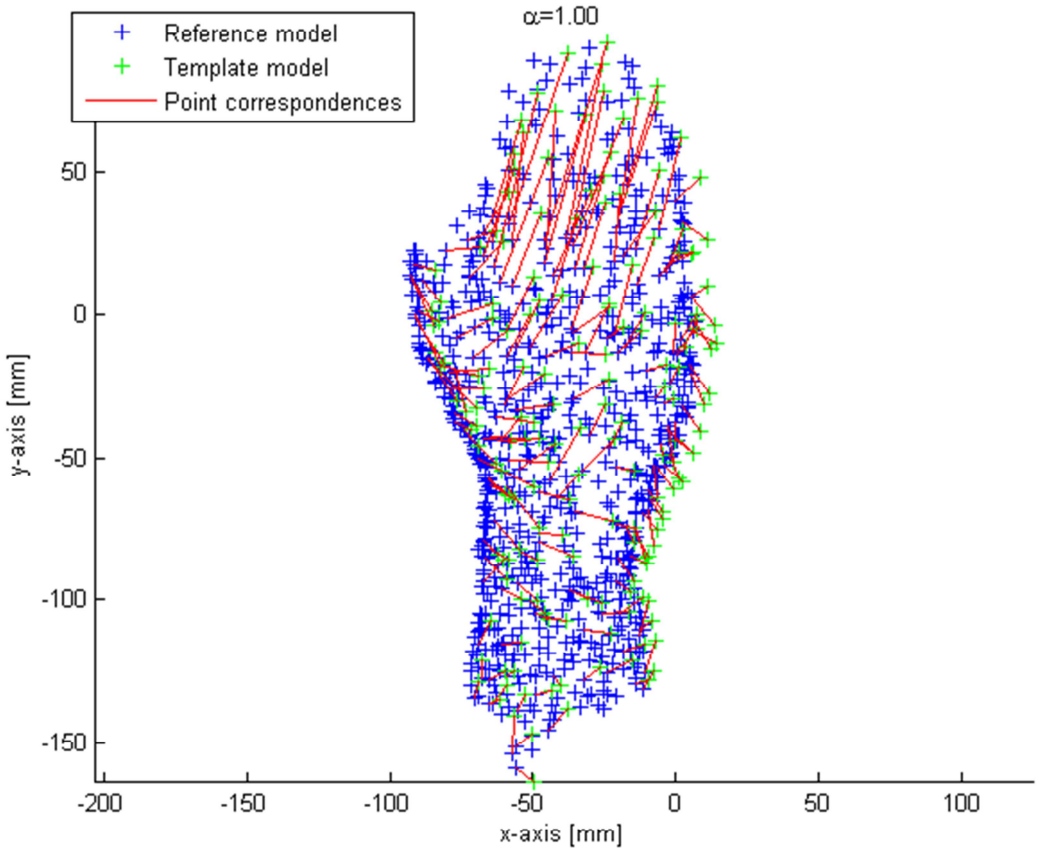


Figure 3.48: Plot of the point correspondences between template- and reference model from a dorsal view

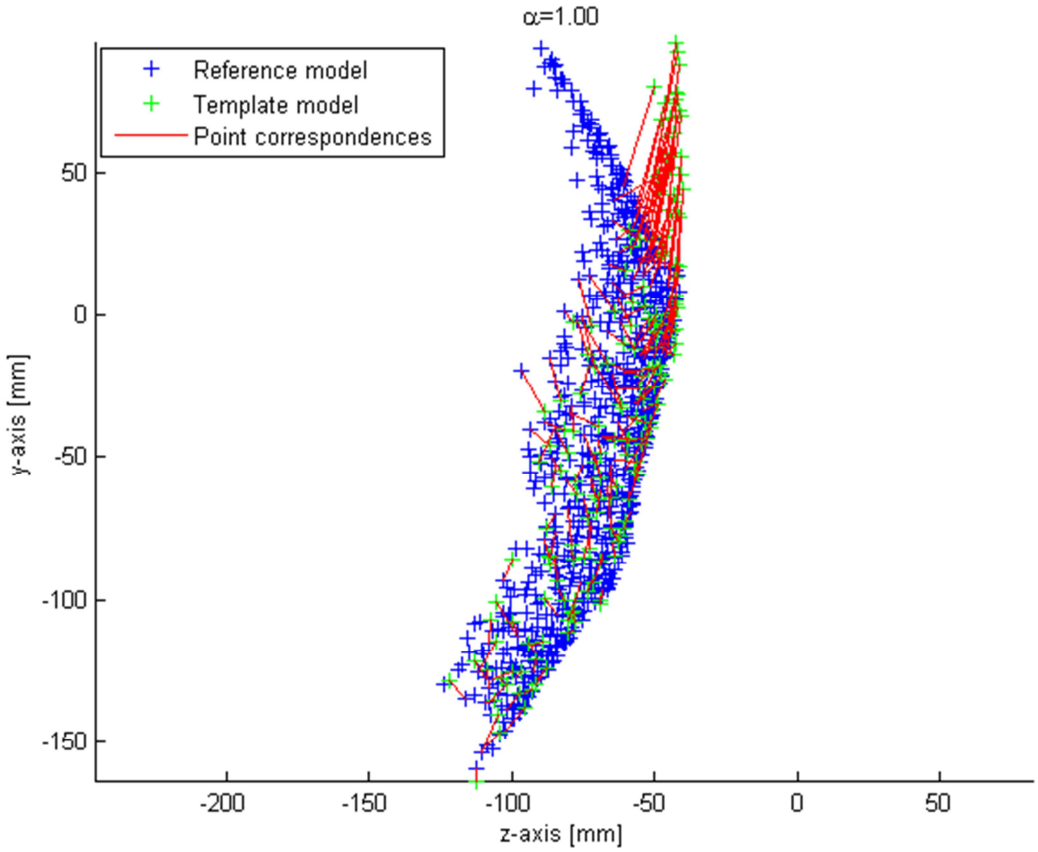


Figure 3.49: Plot of the point correspondences between template- and reference model from a lateral view

Looking at Figure 3.48 and Figure 3.49, we see no correspondences for the point in the distal end of the fingers in the reference model. The distal points in the template are matched to the proximal end of the fingers in the reference model. Recalling that the orientation of shape context histogram related to the orientation of the world coordinate system and not the normal of the surface, it seems quite reasonable that distal points in the template model are more similar to the proximal points of the fingers in the reference model.

The resulting flow of the mesh is illustrated on following figures.

3.21.1.4 Flow charts

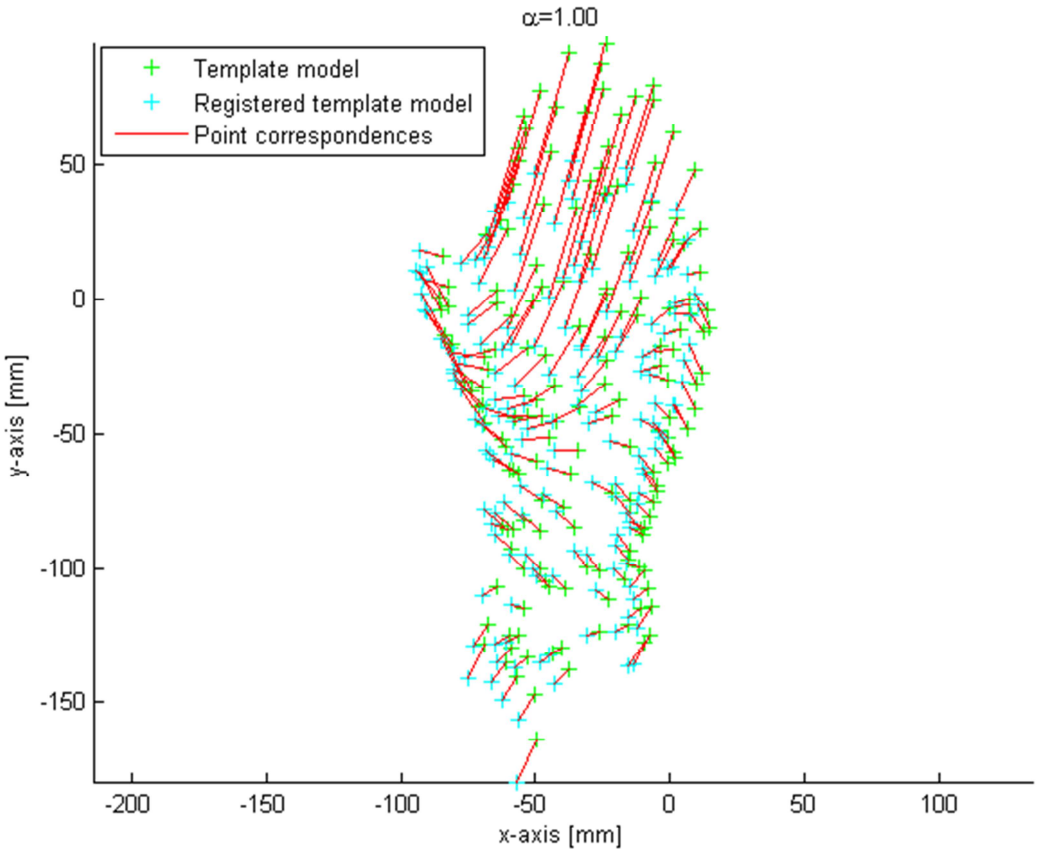


Figure 3.50: Flow chart of the template model from a dorsal view.

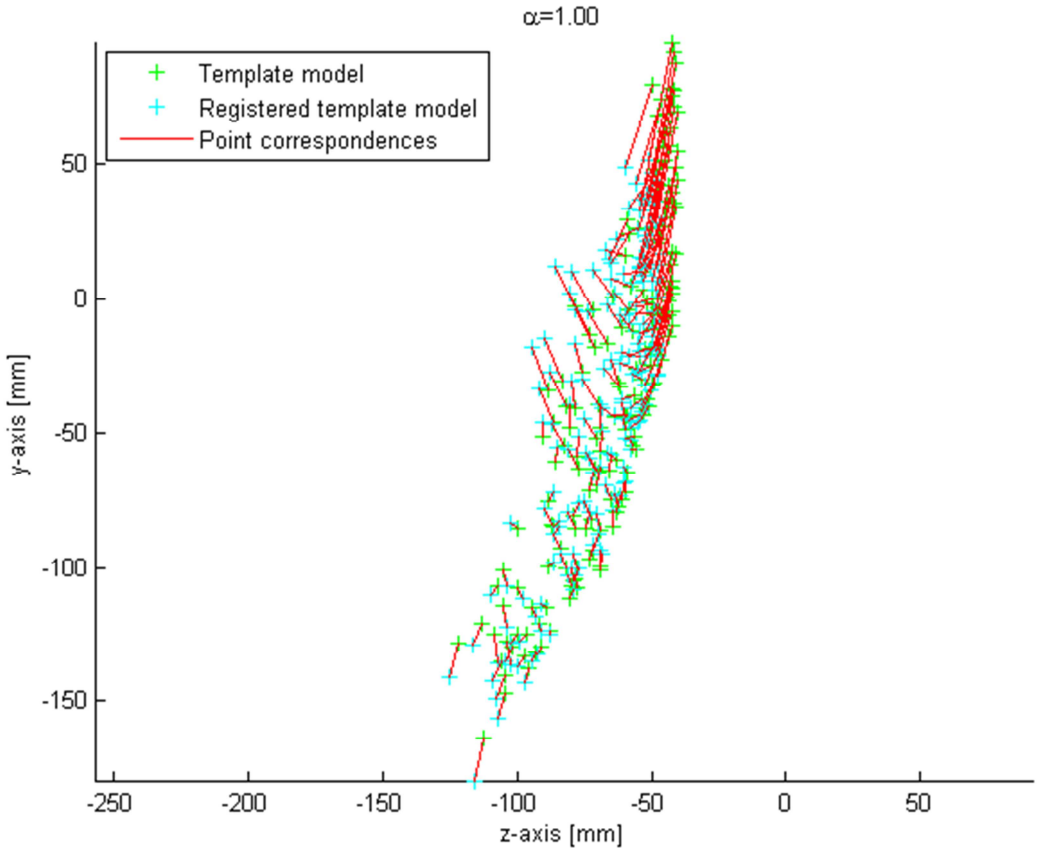


Figure 3.51: Flow chart of the template model from a lateral view.

In order to improve the registration it is necessary to make the shape context histogram invariant to the surface orientation in relation to the world coordinate system. This improvement could be part of a future work and be performed by orienting the shape context histogram in relation to the normal of the surface and another feature such as the Center of Mass.

3.21.1.5 Quantitative error estimation:

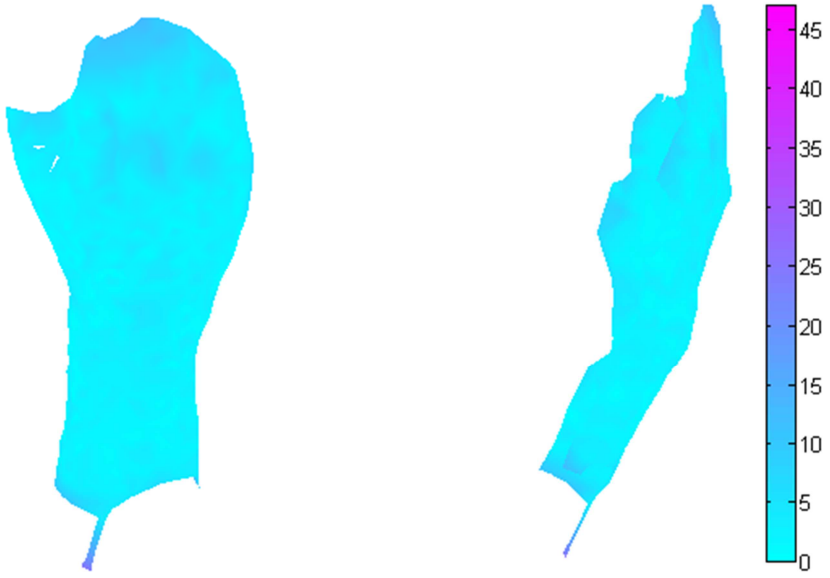


Figure 3.52: Deviation mapping of registered template model, based on the shortest distance. Cyan indicates small distance to the corresponding model, whereas magenta indicates large distance. Distance values on the color bar are labeled in millimeters.

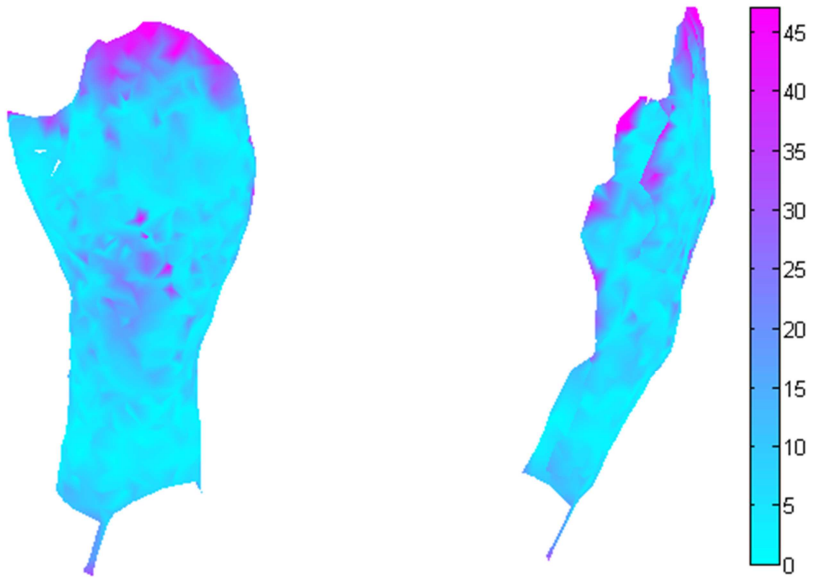


Figure 3.53: Deviation mapping of registered template model, based on the point catch using the Hungarian algorithm. Cyan indicates small distance to the corresponding model, whereas magenta indicates large distance. Distance values on the color bar are labeled in millimeters.

Approach for Error estimation	Value [mm]
Shortest distance RMS	4.4
Hungarian RMS	14.7

Table 3.7: Error estimation

The estimated RMS values are considered to be quite interesting. First of all the magnitude of the RMS provided by shortest distance approach is similar to the RMS values of the previous registrations presented in part II. On the other hand the Hungarian approach shows a significant higher value than we have seen from the previous registrations. It is clear that this is a case where the Hungarian approach is more reliable than the shortest distance approach, since the shortest distance fails because of the large difference between the two models.

3.21.2 Segmentation with 3 segments using multiple models

In this segmentation approach, the 3D points are clustered using the standard deviation of the point locations according to the local coordinate systems. The models that is used for this approach is illustrated below.

3.21.2.1 Hand models:

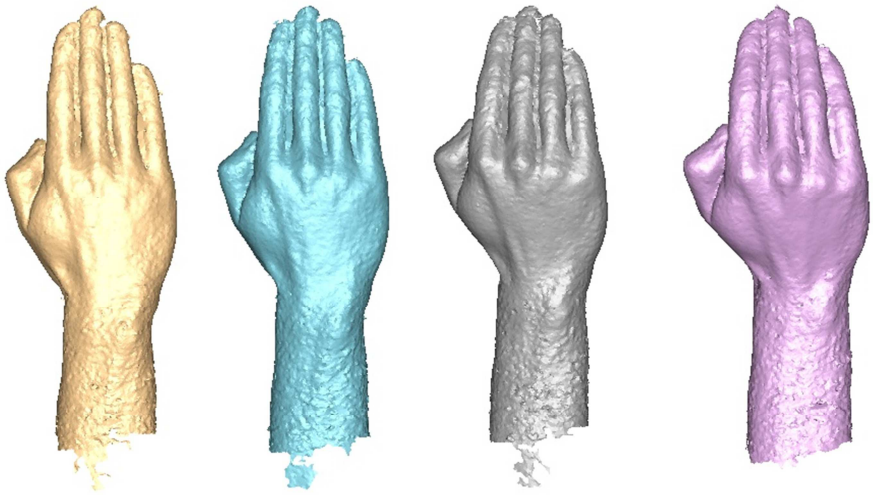


Figure 3.54: Four hand models with various flex angles in the wrist and proximal finger joint from a dorsal view.

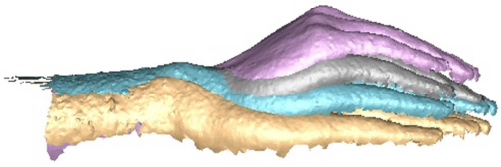


Figure 3.55: Four hand models with various flex angles in the wrist and proximal finger joint from a medial view.

The cyan colored model is used a template model, who is registered to the other models in 5 iterations. The results of the registration are listed below.

3.21.2.2 Results of registration

Figure 3.56: Template model registered to the reference models from a dorsal view.

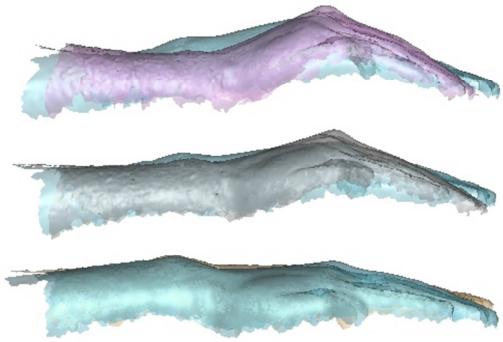


Figure 3.57: Template model registered to the reference models from a lateral view

The registration to the magenta model seems to fail. Looking at the registered models from the dorsal view, we see the registered template has become narrow when registered to the magenta model. From the lateral view we see poor registrations in both the wrist, proximal finger joint and at the finger tips.

From the lateral view the registration to the gray model does not seems to the successfully registered in the whole finger region. The flexion of the fingers is not sufficient to provide a satisfactory result.

Approach for Error estimation	Value [mm]
Golden model	
Shortest distance RMS	3.5
Hungarian RMS	6.5
Black model	
Shortest distance RMS	3.9
Hungarian RMS	6.6
Magenta model	
Shortest distance RMS	5.4
Hungarian RMS	12.7

Table 3.8: Error estimation

The RMS values seems to confirm the observations of the registrations

3.21.2.3 Multiple models segmented into two segments

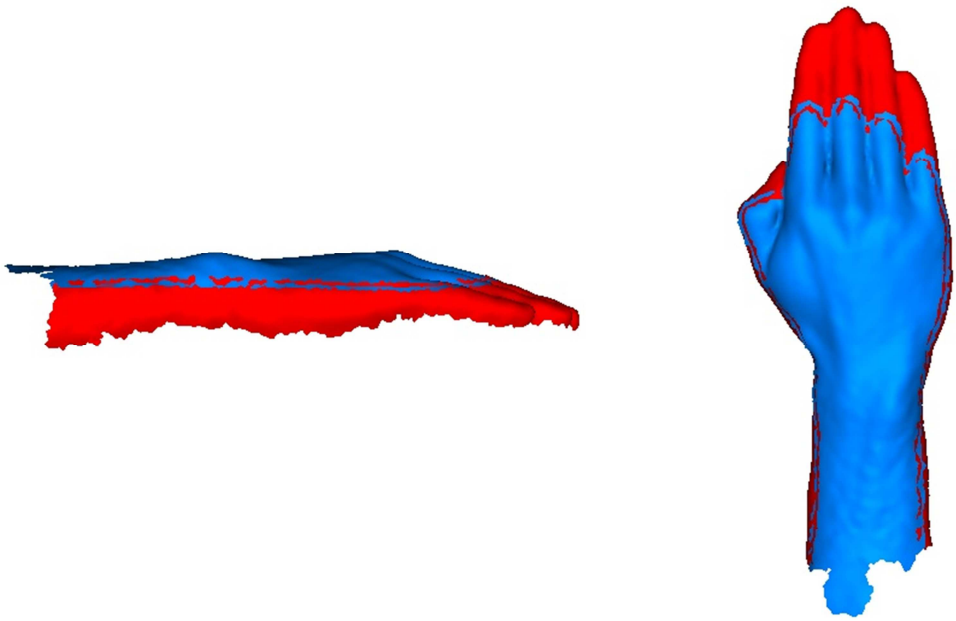


Figure 3.58: Results of segmentation with two clusters

Quite unexpected the results of the segmentation shows a segmentation of the hand along the coronal dimension instead of along the axial dimension. The segmentation is completely unusable to calculate any joint angles.

3.21.2.4 Multiple models segmented into three segments

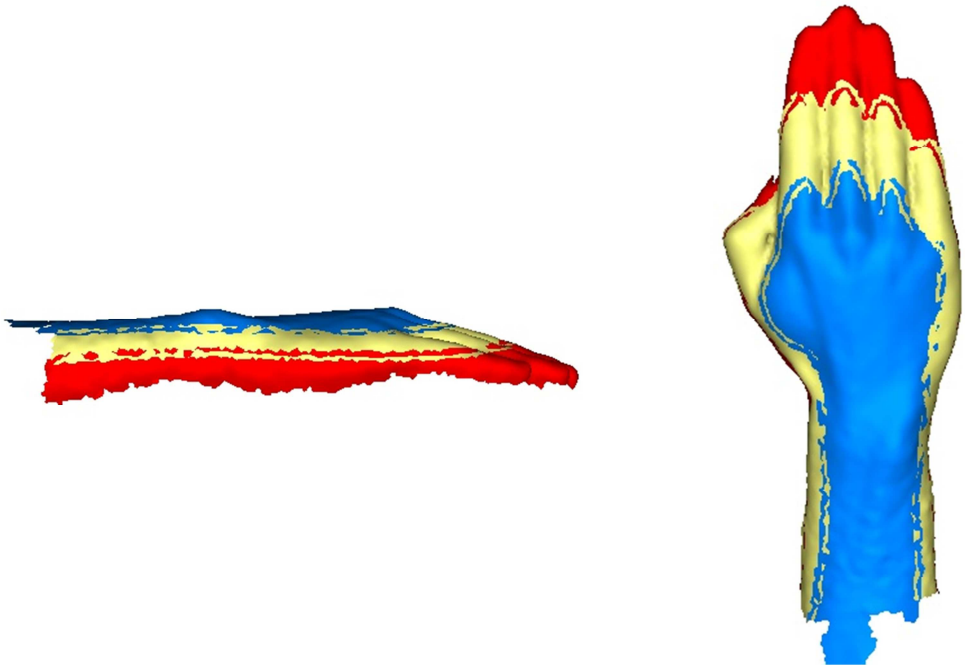


Figure 3.59: Results of segmentation with three clusters

For the segmentation with three segments we see the same tendencies as for the segmentation with two segments. The segmentations are cutting the model coronal instead of axial. The failure seems to be related to the unsuccessful registrations of the black and magenta model in particular. However using the golden model only has been tested as well and showed to be insufficient to segment the object properly in three segments but sufficient for a good segmentation into an arm and a hand segment.

Again we conclude that the segmentation fails due to errors in the registration. The registration approach is simply too primitive to provide satisfactory results for the segmentation.