Challenges in 3D scanning: Focusing on Ears and Multiple View Stereopsis

Jensen, Rasmus Ramsbøl

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Challenges in 3D Scanning: Focusing on Ears and Multiple View Stereopsis

Rasmus Ramsbøl Jensen

Kongens Lyngby 2013
Ph.D.-2013-301

Technical University of Denmark
DTU Compute
Department of Applied Mathematics and Computer Science
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253031, Fax +45 45882673
compute@compute.dtu.dk
www.compute.dtu.dk Ph.D.-2013-301

Summary

It is the goal of this thesis to address some of the challenges in 3D scanning. This has been done with focus on direct in-ear scanning and on Multiple View Stereopsis. Seven papers have been produced over the course of the Ph.D., out of which, six have been included.

Two papers concern volumetric segmentation based on Markov Random Fields. These have been formulated to address problems relating to noise filtering in direct in-ear scanning and Intracranial Volume estimation.

Another two papers have been produced on the topic of recovering surface data based on a strong statistical prior. This was done in particular on scans of ear canals, but the methods are general.

Finally, an experimental setup has been constructed, which has produced a large versatile data set. The data set has been used as the foundation for two papers on the evaluation of Multiple View Stereopsis. The data have a great potential to be used for advances in Multiple View Stereopsis, robust surface reconstruction and photorealistic modelling.

Resumé

Målet med denne afhandling at behandle nogle af udfordringerne i 3D scanning. Dette er gjort med fokus på direkte scanning af øregange og flerbilled stereopsis. Syv artikler er blevet produceret i løbet af ph.d. forløbet, hvoraf seks er inkluderet.

To af artiklerne vedrører volumetrisk segmentering baseret på Markov Random Fields. Disse omhandler løsningen på et støjfiltrerings-problem i forbindel med ørescanning samt estimering af intrakraniel volumen.

Yderligere to artikler er blevet skrevet omkring gendannelse af overflade-data baseret på en stærk statistisk prior. Dette er blevet udført på scanninger af øregange, men metoderne er generelle.

Endelig er en forsøgsopstilling blevet bygget, som har produceret et stort alsidigt datasæt. Dette datasæt er blevet brugt som grundlag for to artikler om evaluering af flerbilled stereopsis. Data har et stort potentiale for at blive anvendt til fremskridt i flerbilled stereopsis, robust overflade rekonstruktion samt fotorealistisk modellering.

Preface

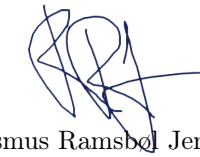
This thesis was prepared at the Image Analysis and Computer Graphics section at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark (DTU). It was done in fulfilment of the requirements for acquiring a doctor of philosophy degree (Ph.D.) within the topic of image analysis.

The presented work was financed by the *Danish National Advanced Technology Foundation* (project no. 019-2009-3) in a collaboration between 3Shape A/S, the Audiological Clinic at Bispebjerg Hospital and DTU.

One part of the research presented in the thesis deals with challenges in relation to direct in ear scanning, while the other part focuses on Multiple View Stereopsis. The thesis consists of an introductory part, an overview of the methods applied, a technical description of the experimental setup created and a concluding chapter. Following the conclusion are the 6 manuscripts prepared during the course of the Ph.D. study.

The project has been supervised by Associate Professor Rasmus R. Paulsen at DTU and Mike van der Poel, Group leader of optics and project manager at 3Shape. The research has mainly been carried out at DTU but also at 3Shape and Computer Science Division at the University College at Berkeley during an external stay under supervision of Professor Jonathan R. Shewchuk.

Lyngby, 1st of May-2013



Rasmus Ramsbøl Jensen

Acknowledgements

During my three years as a Ph.D. student I have had the honour to collaborate, work, discuss, research and also have fun with a lot of dedicated and inspiring people. I would like to thank the colleagues at the Image Analysis and Computer Graphics Section at the Department of Applied Mathematics and Computer Science at The Technical University of Denmark for creating an open door work environment, which is both engaging, motivating, fun and nerdy (in the most positive sense of the word).

Also, I thank my main supervisor Rasmus R. Paulsen from the Image Analysis and Computer Graphics Section, for his guidance and encouragement and my supervisor Mike van der Poel from 3Shape, for insights into cutting edge hardware developments in scanner technology. A special thanks goes to all my other collaborators, who helped me conduct my research and finish my dissertation. In no special order: Oline V. Olesen, Signe S. Thorup, Sujung Kim, Henrik Aanæs, Anders L. Dahl, George Vogiatzis, Anders N. Christensen, Jannik B. Nielsen and Rasmus Larsen. I would also like to thank Jonathan Shewchuk for his time during my external stay at UC Berkeley.

Finally, I feel lucky to have so much love and support from all my friends, bonus family and family. For that I am grateful.

Contributions

Papers included in this thesis

Chapter 6 Rasmus R. Jensen, Mike van der Poel, Rasmus Larsen, and Rasmus R. Paulsen, *Ultra Fast Optical Sectioning: Signal preserving filtering and surface reconstruction,* In proceedings for MeshMed, Toronto, 2011.

Chapter 7 Rasmus R. Jensen, Signe S. Thorup, Rasmus R. Paulsen, Tron A. Darvann, Nuno V. Hermann, Per Larsen, Sven Kreiborg, and Rasmus Larsen, *Genus Zero Graph Segmentation: Estimation of Intracranial Volume,* SCIA, 2013.

Chapter 8 Rasmus R. Jensen, Oline V. Olesen, Rasmus R. Paulsen, Mike van der Poel, and Rasmus Larsen, *Statistical Surface Recovery: A Study on Ear Canals,* In proceedings for MeshMed, Nice, 2012.

Chapter 9 Rasmus R. Jensen, Jannik B. Nielsen, Rasmus Larsen, and Rasmus R. Paulsen *Anatomically correct surface recovery: A statistical approach,* Submitted to Journal for Computer-Aided Design, 2013.

Chapter 10 Sujung Kim, Seong D. Kim, Anders L. Dahl, Knut Conradsen, Rasmus R. Jensen, and Henrik Aanæs, *Multiple View Stereo by Reflectance Modeling*, 3DIMVT, Zürich, 2012.

Chapter 11 Rasmus R. Jensen, George Vogiatzis, Anders L. Dahl, and Henrik Aanæs *On the Performance of Calibrated Multiple View Stereopsis,* Submitted to ICCV, Sydney, 2013.

Not included in thesis

- Oline V. Olesen, Rasmus R. Paulsen, Rasmus R. Jensen, Sune H. Keller, Merence Sibomana, Liselotte Hœjgaard, Bjarne Roed, and Rasmus Larsen, *3D Surface Realignment Tracking for Medical Imaging: A Phantom Study with PET Motion Correction*, Image-Based Geometric Modeling and Mesh Generation, Springer, 2013.

Contents

Chapter 1

Introduction

Development and integration of range scanners and imaging devices for 3D reconstruction make surface and volumetric scanning an increasingly common part of the world we live in. From the naval radar, developed in the 1930s (and named in the 1940s), to today's self parking car. Scanners are found in hospitals, where CT (Computed Tomography), MRI (Magnetic Resonance Imaging) and PET (Positron Emission Tomography) are used in diagnostics, treatment and research. In airports, similar methods are used in the security body scanners. In the entertainment industry motion capture is used to create realistic animations of characters in films and games. Motion capture is also used as a full body interaction in console games such as Microsoft's Xbox. Time-of-flight cameras and stereo camera setups are used for computer vision in industry to guide robots. Laser scanners are used to create 3D models of buildings and dental implants. Also, most people have a phone with a camera, which has the potential to become a 3D scanner through stereopsis.

Regardless of the scanning modality, most applications involve several processing steps. An immediate problem is noise, which is always present to some extent and should ideally be filtered. Another problem is lack of completeness, which can be solved by scanning from more angles or by somehow estimating missing parts. As a final step an analysis might be applied to the data. Depending on the application, examples of such analyses could be: finding a true surface, segmenting the data or recognising an object.

In special relation to this thesis, the industry of custom fit hearing aids currently use 3D scanners to create models of impressions of ear canals. This thesis is in part a collaboration with 3Shape [2], where a direct in-ear scanner is being developed. This is done in an effort to facilitate the fitting process. A direct scanning of a patients ear canal would be beneficial for a number of reasons. For one, it would remove some of the discomfort of having a silicon compound

injected into the ear canal and have it sit while it hardens. Also, by replacing the impression and scanning steps with a single direct scanning, production time would decrease. This could lead to a one time visit to the audiologists, whereas today getting a custom fit device, involves several visits. Shorter production time and fewer manual steps also leads to lower production costs, which could open up potential markets such as custom fit ear plugs and head phones. It is the aim to solve some of the problems relating to direct in-ear scanning.

Another focus of the thesis is Multiple View Stereopsis (MVS). MVS is used to reconstruct 3D data based on several images. It can be applied to uncalibrated images as well as calibrated. As the number of image sensors and images produced is increasing, good data for development and evaluation of new methods are needed. The body of available data sets is not only limited in the number of sets, it is also limited in the range of scenes [95, 87]. The scenes are also limited to mainly highly textured surfaces. To add completeness to the body of available data, an experimental setup has been constructed to provide a good reference data set with a wide range of scenes acquired under varying light. With more data available, the significance of future evaluations will be greatly improved.

1.1 Objectives

The objective of this thesis is to solve some of the challenges related to 3D scanning, and in particular with focus on direct in-ear scanning and Multiple View Stereopsis.

In relation to in-ear scanners, the thesis has been done in parallel with developments in the direct scanner prototypes from 3Shape. Consequently, it also involves foreseeing what the problems of the final scanner will be, such as:

- Noise filtering.
- Recovery of missing surface areas using statistical priors.

Due to the parallel developments of scanners and methods, the data will be from both laser scanned ear impressions as well as from direct scanner probe prototypes.

Another objective is to create an experimental platform for 3D data acquisition on a wide range of objects. Relating to Multiple View Stereopsis (MVS), the aim is to create a versatile data set that can be used for:

- Evaluation of (new) methods by providing a good reference.
- A real (not computer generated) data set for development of surface reconstruction algorithms.
- Testing the influence of light and texture.
- Creating photorealistic models.

Currently, the numbers and versatility of good reference data sets are very limited. Our work in this area will hopefully add to the completeness of the body of available data sets.

1.2 Thesis Overview

The thesis is structured with an introductory part on ear scanning and devices, an overview of the methods applied in the contributions, a description of the experimental work relating to Multiple View Stereopsis and finally a conclusion (Chapter 2-5). This is followed by the six papers prepared during the course of the Ph.D. (Chapter 6-11,

A brief introduction to the included contributions is included in the following. The papers are grouped as follows: two papers on volumetric segmentation, two papers on surface recovery using a strong statistical prior and finally two papers on the evaluation of Multiple View Stereopsis using a large data set created during the course of the thesis. The contributions should preferably be read before the conclusion in Chapter 5.

1.3 Contributions overview

Chapter 6 - Ultra Fast Optical Sectioning: Signal preserving filtering and surface reconstruction
We present a novel algorithm based on a Markov Random Field that uses a distance constraint to robustly classify a 3D scan volume. Through this classification a signal preserving filtering of the data set is done. The remaining data are used for a smooth surface reconstruction creating very plausible surfaces.

Chapter 7 - Genus Zero Graph Segmentation: Estimation of Intracranial Volume

In this paper, we present a fully automatic 3D graph-based method for segmentation of the intracranial volume (ICV) in non-contrast CT scans. We reformulate the ICV segmentation problem as an optimal genus 0 segmentation problem in a volumetric graph.

Chapter 8 - Statistical Surface Recovery: A Study on Ear Canals

We present a method for surface recovery in partial surface scans based on a statistical model. The framework is based on multivariate point prediction, where the distribution of the points are learned from an annotated data set.

Chapter 9 - Anatomically correct surface recovery: A statistical approach

We present a method for 3D surface recovery in partial surface scans. The method is based on an Active Shape Model, which is used to predict missing data. The model is constructed using a bootstrap framework, where an initially small collection of hand-annotated samples is used to fit to and register unknown samples, resulting in an extensive statistical model. The statistical recovery uses a multivariate point prediction, where the distribution of the points is given by the Active Shape Model.

Chapter 10 - Multiple View Stereo by Reflectance Modeling

In this paper, we propose to construct visual metrics of more than one dof using the DAISY methodology, which compares favorably to the state of the art in the experiments carried out. These experiments are based on a novel data set of eight scenes with diffuse and specular surfaces and accompanying ground truth. The performance of six different visual metrics based on the DAISY framework is investigated experimentally, addressing whether a visual metric should be aggregated from a set of minimal images, which dof is best, or whether a combination of one and two dof should be used.

Chapter 11 - On the Performance of Calibrated Multiple View Stereopsis

We test a simple image texture characterisation and demonstrate how it correlates to the performance. Our study is based on a large multiple view data set containing 80 scenes acquired from a setup based on a six-axis industrial robot. Accurate reference surface reconstructions are obtained from scans based on stereopsis with structured light support. In addition to predicting the reconstruction quality this study highlights the limitations for current state of the art surface reconstruction. Our data will be made available online.

Chapter 2

Ear devices and custom fitting

The hearing aid industry in Denmark is world leading with big market players such as Oticon, Widex and GN Resound. A custom fit hearing aid has a part of the device fitted to the subjects ear canal as seen in Figure 2.1. Hearing aids, which are fitted inside the ear canal are known as *in-the-ear* (ITE) or *completely-in-canal* (CIC). CIC is the same as ITE but sits deeper in the ear canal and is less visible. The *behind-the-ear* (BTE) hearing aid device consist of receiver part sitting behind the ear and a sound emitter in the ear canal. The sound emitter can be either a plastic tube or a wire connected to a loudspeaker in the ear canal. While the receiver part is generally stock the emitter part might be custom fitted to the ear canal.

(a) Custom fit hearing aid

(b) Hearing aid completely in canal (CIC)

Figure 2.1: A custom fit hearing aid made to fit completely in the ear canal (b) and a device fitted in the ear canal (a) (both pictures from [72])

2.1 Custom fit hearing aids

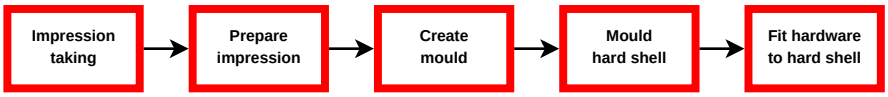
Fitting a hearing aid involves a lot of manual work and in the traditional way, also, handicraft of the audiologist and operator. The normal work flow involves the audiologist taking an impression of the ear canal. First a rubbery silicon compound is injected into the subjects ear (see Figure 2.2(a)). The compound hardens within a few minutes and the impression of the ear canal can be extracted (see Figure 2.2(b)). To protect the ear drum, a cotton ball is placed in the ear canal prior to the injection. The cotton ball is connected to a piece of string to ensure that it does not stay lodged inside the ear. Some find this procedure unpleasant and even claustrophobic and painful.

Traditionally, the impression would be handed to an operator, who would create a mould from the impression. The final casing of the device would then be created in this mould. The process of grinding the impression to the right size before creating the mould acquires a lot of experience. In the mould a hard shell is created, in which the hardware of the device is later fitted (see Figure 2.3(a)). The manual craftsmanship after the impression taking is in the process of being replaced by a 3D scanning of the impression and a digital prototyping. Based on the 3D scanning, the device is designed in a CAD-program and 3D printed to fit

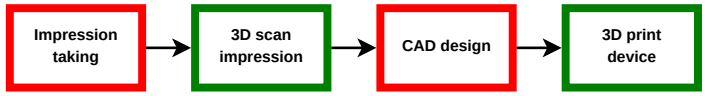
(a) Compound injected into ear canal (b) Impression of the ear canal

Figure 2.2: Manual impression taking of the ear canal (with Professor Rasmus Larsen as a stunt double). (a) shows a silicon rubber compound being injected into the subjects ear. (b) shows the resulting impression of the ear canal.

the hardware (see Figure 2.3(b)). The 3D printing is generally done outhouse. If a direct in-ear scanner is added to the workflow, this would shorten the fitting pipeline. More over, with advances in CAD, designing the final device will need less and less attention from the audiologist. As 3D printers become more widely available a realistic workflow would only require manual interaction from the audiologist during the actual in-ear scanning. This would allow for only one visit to the hearing aid clinic and a patient could go home with the device the same day. Regardless of the process, the device amplification has to be profiled to compliment the hearing impairment of the patient.



(a) Traditional fitting



(b) Fitting with impression scanning

(c) Future fitting with direct in-ear scanning

Figure 2.3: This figure shows the workflows of custom fitting done the traditional way (a), with impression scanning (b) and a probable future with direct in-ear scanning and automated CAD design of the device (c). Manual interaction is shown in **red** and **green** shows automated parts of the fitting process.

2.2 Head sets

While the market for a direct in-ear scanner in relations to custom fit hearing aids is apparent, there are also other markets. Since most people now have a mobile phone, which can be used as music player or radio, this could be a huge potential market. Normally, mobile phones are sold with an earplug headset. Frequently, these headsets come with a set of 3 different sized rubber ear canal fittings, which provides a reasonable fit for most people. Possibly, direct ear canal scanning and 3D printing would allow for dispensing of custom fit head sets at an attractive price.

2.3 Ear plugs

Normal ear plugs are made of foam, which is compressed and inserted into the ear canal, where it expands and stays in place. While this offers some protection, a better alternative is a custom fit ear plug. Other than providing a better fit and staying in place, a custom fit ear plug can offer much better protection and can come with different inserts that allow dampening of different magnitude and frequencies. However, these ear plugs are quite expensive, and

are therefore mainly used by professionals such as the Danish army, where they are provided for soldiers serving abroad. As with the fittings for head sets, a lower production cost as a result of direct ear scanning could make custom fit ear plugs more generally available.

2.4 Direct scanning of the ear canal

For surface scanning, different methods have already been proven such as laser scanning, stereo/multiview-vision, and time of flight. These scanning techniques are all 2.5D scanners, which means that 3D data are only captured from the perspective of the scanner. To create a full 3D scan several scans need to be merged into one. Scanning the ear canal involves several challenges. For one, the canal is very confined, which leads to restrictions on probe size and optical system. Also the scanner cannot be fixed relative to the ear canal. In laser scanning of impressions, the impression is placed on a rotating disk, which allows for the scanner to model the impression from all sides. This makes it a much more controllable task of merging the scan data into a full model. Environment control is not possible for the ear canal, and if the scanner is not able to cover the whole ear canal from one position, it has to rely on the data, when merging partial scans. Another challenge is the presence of noise coming from hair, ear wax, and the scanner itself. Finally, the s-shaped anatomy of the ear canal might also induce some occlusion. Figure 2.4 shows a direct in-ear scanner from 3Shape [2], the scanner uses Ultra Fast Optical Sectioning^{TM}. Optical sectioning is know from microscopy, where adjustments in focus are used to capture depth [68]. Since the scanner from 3Shape is hand held, the changes to the optical system and the image acquisition have to be very fast. The actual design implementation of the scanner is not public.

We had access to data made with a direct scanner probe. These data are used in the validation of the method presented in Chapter 9.

Figure 2.4: The direct in-ear scanner from 3Shape.

Chapter 3

Methods

A wide range of methods have been applied in the production of the included contributions presented in Chapter 6 through Chapter 11. This chapter provides a brief introduction to some of the methods used in this thesis. Hopefully this will facilitate the comprehension, when reading the contributions.

3.1 Segmentation in images and graphs

A common problem in image analysis is classification. Given some input data, regions are to be segmented into two or more classes. The simplest segmentation is thresholding, where some defined values divides the data into different classes. Relying on statistics, a common classifier is based on Bayes rule:

$$P(x|v) = \frac{P(v|x)P(x)}{P(v)} \tag{3.1}$$

Where x is a class variable taking on a label from the selection of classes L, while v is the observed data. As an example, in an image each pixel would get a labelling x and the data v would be the pixel values, which can either be scalar for grayscale or multidimensional for color and multispectral images. The classifier can be used to find the Maximum a Posterior (MAP) between a number of known classes. The MAP simply chooses the label that maximises this probability. Since the term $P(v)$ is constant for all classes it is normally omitted, when comparing the probability between classes. If the classes are normally distributed with mean μ_x and standard deviation σ_x corresponding to

the label of x, this becomes:

$$P(x|v) = P(x)\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(-\frac{(v-\mu_x)^2}{2\sigma_x^2}\right) \tag{3.2}$$

The distributions with μ_x and σ_x, and the prior probabilities for $P(x)$, need to be found. This can be done using manual segmentation in some representative data set, or it can be unsupervised as done in Expectation Maximisation (see Section 3.1.1).

A more advanced method, the Markov Random Fields theory, also takes a local neighbourhood (i.e. the surrounding pixels or voxels) into account in the classification. This will be addressed in Section 3.1.4.

3.1.1 Expectation Maximisation of a Gaussian Mixture Model

In classification, it is sometimes needed to estimate the distribution parameters. The Expectation Maximisation algorithm [45] (EM) can be used to estimate the distribution of a Mixture of Gaussians (superposition of a number of different Gaussian distributions). A class k from N classes is described with the distribution parameters μ_k and σ_k (σ would be Σ in the multivariate case) and the class prior π_k. The data set has index i and point values d_i. Given an estimate (this can be a bad estimate) of the class parameters $\tilde{\mu}_k$, $\tilde{\sigma}_k$ and $\tilde{\pi}_k$ the algorithm works by calculating the posterior probability in each point w_k^i for each class:

$$w_k^i = \frac{\mathcal{N}(d_i|\tilde{\mu}_k, \tilde{\sigma}_k)\tilde{\pi}_k}{\Sigma_{k'}\mathcal{N}(d_i|\tilde{\mu}_{k'}, \tilde{\sigma}_{l'})\tilde{\pi}_{k'}} \tag{3.3}$$

With an updated posterior the class parameters can be updated as follows:

$$\begin{aligned}
\tilde{\mu}_k &\leftarrow \frac{\Sigma_i w_k^i d_i}{\Sigma_i w_k^i} \\
\tilde{\sigma}_k^2 &\leftarrow \frac{\Sigma_i w_k^i (d_i - \tilde{\mu}_k)^2}{\Sigma_i w_k^i} \\
\tilde{\pi}_k &\leftarrow \frac{\Sigma_i w_k^i}{N}
\end{aligned} \tag{3.4}$$

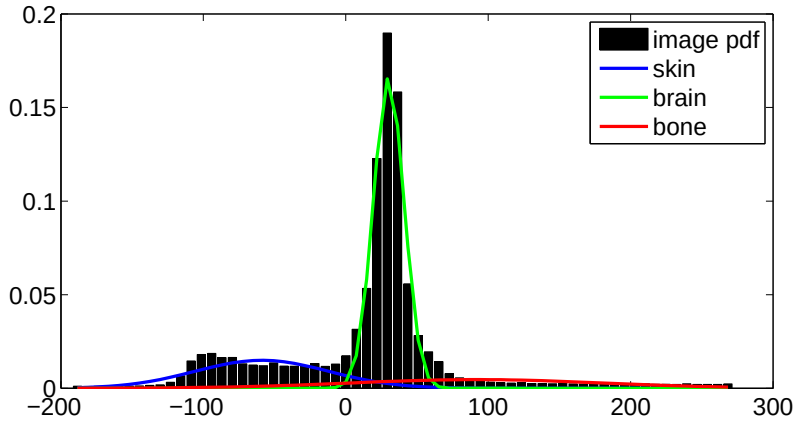
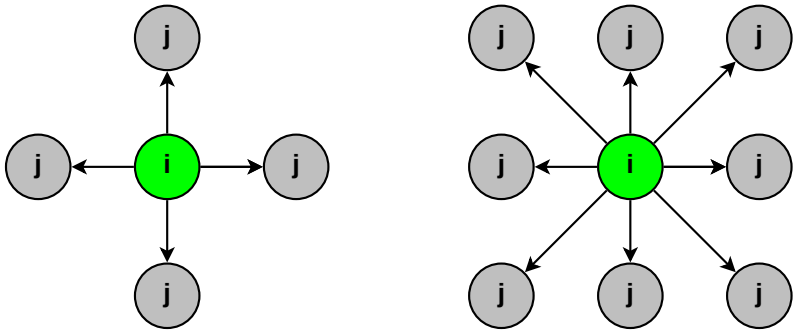


Figure 3.1: A histogram showing the probability density function of the response in Hounsfield units in a CT-scan of a head. Three distributions: brain matter, bone and skin in the scan have been estimated. It is nowhere near a perfect fit using only the normal distributions but the found three distributions serves perfectly well in classifying between Intracranial Volume (brain matter) and not Intracranial Volume (bone and skin).

Calculating the posterior probability and updating the class parameters will converge to a best fit to the data given a number of classes. This estimation of parameters is very useful for image data, as it removes the manual classification step. We used the EM algorithm to estimate the distributions in the CT scans in the paper on Intracranial Volume estimation (ICV) in Chapter 7. The three distributions were brain matter, bone and skin. An estimation of the three classes can be seen in Figure 3.1. It is not a perfect fit using only three normal distributions but for the purpose of classifying between ICV (brain matter) and non ICV (bone and skin) it is sufficient. To add consistency to the labelling the classification was done in an MRF framework.

3.1.2 Graphs

When considering an image segmentation problem it can be useful to create a graph using the data points as vertices and connect them with edges according to their spatial distribution. Using the spatial connectivity in the segmentation allows for a more robust segmentation as a neighbourhood dependency can be included. In image data, the data sets often provide a good structure for



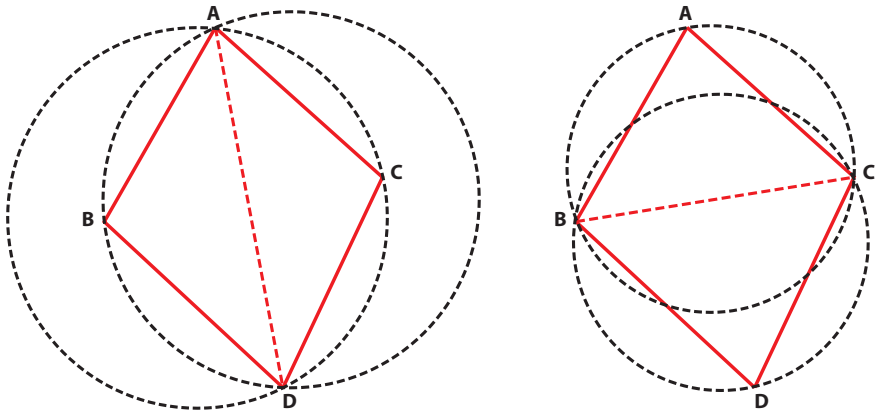
(a) 4-neighbourhood (b) 8-neighbourhood

Figure 3.2: Neighbourhood connectivity in 2D images, (a) shows the 4-way connectivity between a pixel i and its neighbourhood. (b) shows the 8-way connectivity.

the connectivity of the graph. If the data is a 2D image the neighbourhood connectivity of a pixel i could be the nearest 4 or 8 neighbours in the pixel grid (see Figure 3.2). If the image data is a voxel volume this neighbourhood connectivity becomes the 6-way or the 24-way neighbourhood. The implicit voxel connectivity was used for the Signal Preserving Filtering (see Chapter 6).

3.1.3 Delaunay tetrahedralization

For the Intracranial Volume (ICV) estimation (see Chapter 7) we used a different graph formulation than that provided in the data. We downsampled the volume using sample points quasi-equidistantly distributed in a sphere. To create a neighbourhood connectivity the points were connected using Delaunay tetrahedralization [90]. This is a well known method for mesh generation as it has the nice property that it maximises the angles of the triangles constituting the tetrahedra in the mesh. In Figure 3.3 a 2D example is shown with two adjoining triangles created by four points. The property of a Delaunay triangulation is that no circumcircle of a triangle can contain another triangle. If it does, the joining edge between the triangle and the contained triangle must be flipped [89]. In Figure 3.3(a) two triangles are shown for which the circumcircle contains the other triangle. After an edge flip the circumcircles do no longer contain other triangles and the triangulation is Delaunay (se Figure 3.3(b)). While Delaunay triangulation builds on this simple property, the implications are not so simple. When an edge is flipped the edges of the joining triangles most be checked to see if they are still Delaunay. If not their edges need to



(a) Bad triangles (b) Good triangles

Figure 3.3: Four points creating two triangles shown in red. Each triangle's circumcircle is shown as a dashed black line and the connecting edge shown as a dashed red line. Triangulation before edge flip, (a), and after edge flip, (b). After the edge flip the triangles are Delaunay.

be flipped and so on and so forth. The Delaunay property expands to any dimension. In 3D, the circumspheres of two joining tetrahedra cannot contain the other tetrahedra, if they do, the joining surface must be flipped. As with with the 2D example efficient flipping is not trivial. However, the end result should in any case be Delaunay regardless of implementation. For our work we simply used the implementation in Matlab.

3.1.4 Markov Random Fields

In both the paper on Signal Preserving filtering (Chapter 6) and Genus Zero Graph Segmentation (Chapter 7) Markov Random Fields (MRF) were used. This is a classifier that, apart from a statistical probability, also includes a dependency on a local neighbourhood [10, 63]. We define a random field with the input data defined with spatial data v_i at the vertices (pixels, voxels or graph nodes) with the index set I. Each vertex classification variable x_i takes on a label from the set of class labels L. The configuration of the full set of vertices is $\mathbf{x}$. A neighbourhood system to v_i is defined as $N = \{N_i | i \in I\}$ for which it holds that $i \notin N_i$ and $i \in N_j \Leftrightarrow j \in N_i$. A random field is said to be a Markov field, if the probability $P(\mathbf{x})$ of any configuration of $\mathbf{x}$ satisfies the

positivity property:

$$P(\mathbf{x}) > 0 \qquad \forall \mathbf{x} \in L \tag{3.5}$$

And the Markovian property:

$$P(x_i|\{x_j : j \in I \backslash \{i\}\}) = P(x_i|\{x_j : j \in N_i\}) \tag{3.6}$$

Or in other words the probability of x_i given the index set $I\backslash\{i\}$ is the same as the probability given the neighbourhood of i. Combining the probabilities given the data and the neighbourhood, we get:

$$P(x_i|v_i, \{x_j : j \in N_i\}) = P(x_i|v)P(x_i|\{x_j : j \in N_i\}) \tag{3.7}$$

The joint neighbourhood probability to be maximised can be expressed as:

$$\begin{aligned} P(x_i|\{x_j : j \in N_i\}) &= \frac{1}{Z} \prod_{j \in N_i} P(x_i|x_j) \\ &= \frac{1}{Z} \prod_{j \in N_i} \exp\left(-\lambda(x_i, x_j)\right) \\ &= \frac{1}{Z} \exp\left(-\sum_{j \in N_i} \lambda(x_i, x_j)\right) \end{aligned} \tag{3.8}$$

Where Z is a normalising constant, which can be omitted. If neighbours that have the same label should produce a higher probability than ones with different label, $\lambda(x_i, x_j)$ can be defined as:

$$\lambda(x_i, x_j) = \begin{cases} K & x_i \neq x_j \\ 0 & x_i = x_j \end{cases} \qquad \text{since} \qquad \frac{1}{Z}\exp\left(-K\right) < \frac{1}{Z}\exp\left(-0\right) \tag{3.9}$$

Where K is a non negative real number. This is known as the general smoothness prior but other neighbourhood dependencies can be formulated. Instead of the probability maximisation in Equation (3.7), the problem can be formulated by using the negative log of the combined probability. This results in an energy minimisation problem:

$$E(\mathbf{x}) = \sum_{i \in I} \left(-\log(P(x_i|v_i)) + \sum_{j \in N_i} \lambda(x_i, x_j) \right) \tag{3.10}$$

The first part of this equation is referred to as the observation term. Other probabilities can be used instead of the Bayesian Maximum a Posterior. The second term is the neighbourhood prior term, which local consistency to the classification. The simplest neighbourhood term is the general smoothness prior in Equation (3.9). The general smoothness term penalises neighbouring vertices with different labels with the constant K. For $K = 0$ the solution to the MRF becomes the Maximum a Posterior classification. A data driven neighbourhood term can also be used, where the penalty is less if the data in v_i and v_j are very different. This would favour shifts in the labelling, where there are changes in the data. As with the probability term the neighbourhood term can be formulated as suited for a given problem. While the formulation of the MRF is simple, the mutual dependency between a given point and its neighbour makes it a complex problem to solve. If the labelling is binary an optimal solution can be found using Graph Cuts [57]. If the problem has multiple labels there is no optimal solution using Graph Cuts but a good solution can usually be found using this method in the form of the alpha expansion.

In the paper on noise filtering (see Chapter 6), the scanner frustum (the volume covered by the scanner) is considered an MRF. The problem was to find a true skin surface and remove noise from hair and scanner. We approached this by labelling the frustum voxels as being either under or over the skin surface rather than under or over the scan surface, which is what the scanner provides. For the main part the two definitions will coincide except in the presence of noise. A simple probability term was used, giving high probability for voxels over scan surface also being over skin surface and under scan surface also being under skin surface. As neighbourhood priors, we used the general smoothness prior along with a vicinity prior. The vicinity prior was added to make changes in labelling happen close to the surface. For every voxel a distance was estimated to the nearest scan point using a distance transform [49]. Neighbours with a different label would get a penalty based on the distance to the nearest scan point. Based on this, changes in label close to the scanned surface would get a very little penalty, while points far from the scan surface would get a big penalty. This was exactly the quality sought for in the MRF formulation.

When estimating intracranial volumes (ICV) (see Chapter 7) we also formulated an MRF. Using Expectation Maximisation (see Section 3.1.1) we found estimates of the distributions for the three dominant tissues: brain matter, bone and skin. These distributions were used in the probability term, where we used the two classes: ICV and non ICV. The ICV probality was set equal to that of brain matter, while the non ICV was the combined probability of the remaining two classes. A general smoothness prior was added. Also, a data driven neighbourhood dependency was included. Two neighbouring voxels with different labels were penalised proportionally to the exponential of the negative absolute gradient. Since this function goes towards zero as the gradient gets big, this

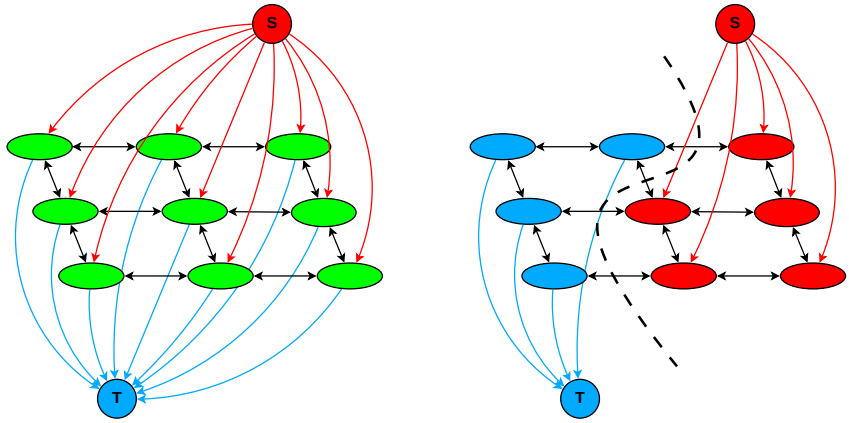
neighbourhood term favours difference in labelling along high gradients in the data. To create a smaller problem the volume was subsampled. A set of sample points outside the head of the subject had the labelling clamped to non ICV, while a selection of nodes inside the ICV had their labelling clamped to being ICV. Clamping in MRF is done by setting the probability term to infinity for one class and to zero for the other. The clamping combined with the formulation of neighbourhood dependencies made the algorithm consistently produce good segmentations of genus zero.

These were just two examples of different implementations of MRF. While using a probability and a neighbourhood dependency is a simple idea, MRFs can be used to solve a wide range of problems, if formulated right.

3.1.5 Graph Cuts

As mentioned in the previous section, Graph Cuts [57] can find an optimum solution to a binary MRF. Graph Cuts uses an older min-cut/max-flow algorithm originally designed to find both the bottle necks and maximum throughput in a transport network. Figure 3.4 shows are very simple segmentation problem of labelling nine vertices to one of two classes. The vertices are connected to the immediate neighbours. The graph is constructed by adding two extra terminal vertices such that all vertices are connected to these. The two terminal vertices are named source (s) and sink (t). The capacities of the edges from the vertices to the terminals are found using the observation term, while capacity of the neighbour edges is defined according to the neighbourhood prior term. The source edges were found using the probability term of one class and all the sink edges using the probability term of the other. The capacities of the edges between vertices are found using the neighbourhood prior term (as described in Equation (3.9)). The shown example in Figure 3.4 is a two dimensional image but the method expands to higher dimensions.

On the graph, the min-cut/max-flow algorithm finds the minimal cut. Three stages of the min-cut/max-flow algorithm are shown in Figure 3.5. This is not the typical layout for a Graph Cut as all nodes are normally connected to both source and sink but this is an easier way to get the grasp of how the min-cut/max-flow works. Consider Figure 3.5(a) a transport network with edges with a flow and a maximum capacity. In this figure there is no flow going from source to sink. The first step of the algorithm is to find an augmenting path from source to sink. An augmenting path allows for an increase in flow. This is shown in 3.5(b), where an initial augmenting path is found. The algorithm repeats to find augmenting paths until no more flow can be pushed through the network. When the network is exhausted, the graph can be separated by the

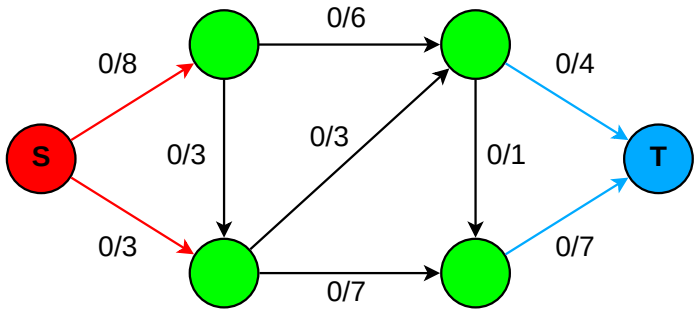


(a) Full graph (b) Segmented graph

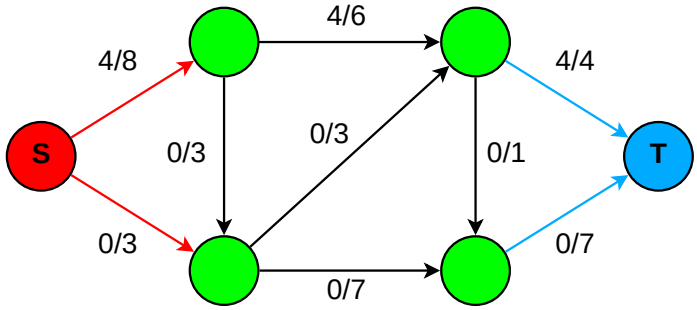
Figure 3.4: A graph as constructed for a Graph Cut solution to an MRF in (a) and after the cut has been found in (b).

cut for which all edges going from source towards sink are exhausted. This is the bottle neck of the network and the minimal cut. Figure 3.5(c) shows the minimal cut and the segmentation based on this cut. Note that one edge in the cut is not exhausted; this edge is however going in the opposite direction of the flow and from sink to source (the only edge going from blue to red in the final segmentation).

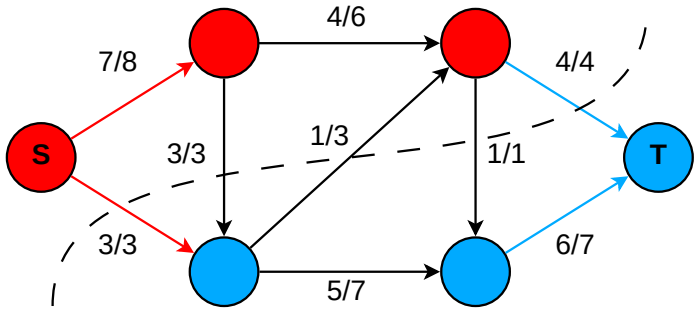
Graph Cuts were used to find the solution to the MRFs formulated in Chapter 6 and 7. This method is much faster than the original iterative method [10] for solving MRFs. Also, the iterative method did not guarantee an optimal solution. Since the Graph Cut method is a very good solution to the widely used MRF problem formulation, a lot of advances have been done to speed it up. For lange data sets or realtime applications, improvements on the Graph Cut solutions are desirable. For video, the classification can be improved using Dynamic Graph Cuts [54], where the graph and solution from one frame are used for the next. Generally changes will be small and a great speed up can be achieved. Another approach is to find the solution using parallel computation [64], where smaller subproblems are solved and merged. While this speed up will be relatively small running on the CPU (4-8 cores) the performance gain of a parallel implementation is very substantial if implemented on a graphics card with 100+ cores, or on a cluster.



(a) No flow



(b) First augmented path



(c) Exhausted network

Figure 3.5: The min-cut/max-flow algorithm working on a simple graph. (a) shows the graph with no flow and the maximum edge capacities. In an MRF the edges connecting the start and end notes would have capacities according to the observation term, while all other edge capacities are found according to the neighbourhood prior term. (b) shows the first augmenting path. (c) shows the minimal cut and the resulting segmentation.

3.2 Alignment of 3D data

A common problem relating to 3D surface scanners is that they only cover a scene from a single point of view. This is the reason why surface scanners are sometimes referred to as 2.5D scanners as they do not provide a full 3D model. To create a 3D model from such a scanner, the first step is to align and merge several 2.5D scans in the same reference system. In Chapter 4 and 11, the calibrated positions of the robot are used to put data in a common reference. Also, in this work several scan sweeps, with a 90 degree rotation between each sweep, are put into the same reference using a data driven method. The method is explained in the following.

3.2.1 Iterative Closest Point

To align scans in the same reference without knowledge of orientation, a common method is the Iterative Closest Points (ICP) algorithm [110, 82]. The algorithm seeks to find the optimal rigid transformation T of the source point set (x) such that it aligns with the target point set (y) already in the reference coordinate system:

$$\arg\min_{T(\mathbf{x})} \|T(\mathbf{x}) - \mathbf{y}\|_2 \tag{3.11}$$

Generally, the transformation is rigid and has 6 degrees of freedom in 3D consisting of a rotation R and a translation t:

$$T(\mathbf{x}) = R\mathbf{x} + t = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \tag{3.12}$$

Where the rotation matrix is a rotation around the coordinate system axes $R = R_x R_y R_z$. The ICP algorithm works as shown in Algorithm 3.1.

While the ICP algorithm is quite simple in theory, there are a lot of things that need attention and the algorithm comes with a wide range of extensions [82]. For one, the algorithm converges to a local minimum. Therefore, it needs an initialisation, which is reasonably close to the global minimum. If the starting point is not close to the solution, the algorithm might not converge to the desired minimum. When finding the correspondence between source and target, one has to consider how this is done. In the simplest form, correspondence is found simply as the shortest Euclidian distance from a point in the source set to a point in the target set. However, there might be better alternatives such as looking for correspondence in the surface normal direction (if a normal can be estimated) or use some descriptor in the data. Also, if the data are

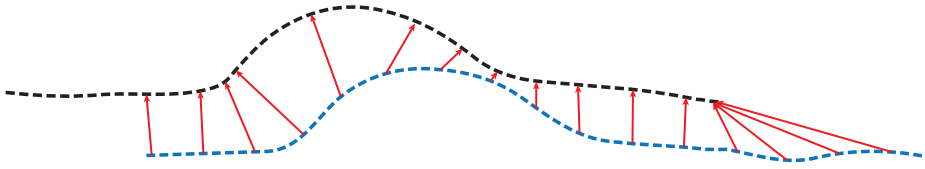
Algorithm 3.1 Iterative Closest Points

procedure ICP($\mathbf{x}, \mathbf{y}$) ▷
  repeat
    Find correspondence points $\mathbf{x_c}$ and $\mathbf{y_c}$ between $\mathbf{x}$ and $\mathbf{y}$
    Find global transformation between correspondence points, $\arg\min_{T(\mathbf{x_c})} \|T(\mathbf{x_c}) - \mathbf{y_c}\|_2$
    Update source point set $\mathbf{x} \leftarrow T(\mathbf{x})$
  until convergence
  return x ▷
end procedure

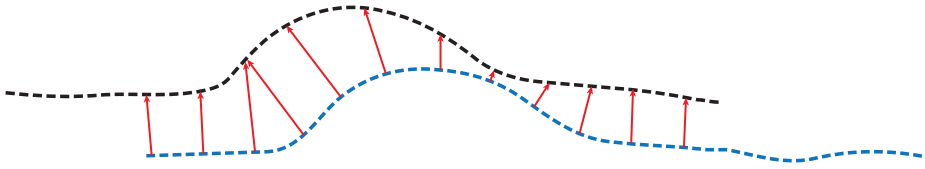
not fully overlapping the non-overlapping parts should not be included in the correspondence when finding the transformation [82]. Figure 3.6(a) shows an example of correspondence found as the shortest Euclidian point distance. A more consistent correspondence can be found using the normal direction as shown in Figure 3.6(b).

In Chapter 9 we used what can be described as an ICP algorithm with an embedded Active Shape Modelling (ASM) (see Section 3.3). This approach was used when building the statistical model that constituted the foundation for the surface recovery of missing parts in partial scans.

The ICP algorithm was also used to align data from sweeps with 90 degree rotations in the work presented in Chapter 4 and 11. Figure 4.14 on page 51 shows different views of a statuette of two doves. The data has been acquired from 4 different sides and aligned using ICP.



(a) Euclidian point correspondence



(b) Normal direction correspondence

Figure 3.6: Two different approaches to finding correspondence between patches. (a) shows correspondence using closest points. (b) shows correspondence found in the normal direction. Source point set is shown in blue and the target in black.

3.3 Statistical Shape modelling

In Chapter 8 and 9 a strong statistical prior was used to recover missing data in partial scans. This prior comes from a statistical shape model built by finding corresponding points throughout a set of shapes and consequently aligning these accordingly.

3.3.1 Procrustes analysis

To create a statistical shape model using a population of shapes with corresponding points, the shapes have to be aligned. The Generalized Procrustes Analysis (GPA) [93] uses rigid transformation to align the shapes to a common mean. This is done in an effort to remove bias as opposed to aligning to just one sample shape. A shape is denoted by $\mathbf{x_i}$ and the total set of shapes is contained in $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$. The estimated mean of the aligned shapes $\bar{\mathbf{x}}$ is used as a reference for the alignment of the shapes. The algorithm works as shown in Algorithm 3.2.

Sometimes it is desirable to remove scale from the data, in which case a scale factor is added to the rigid transformation. This is known as the Euclidian

Algorithm 3.2 Generalized Procrustes Analysis

procedure GPA(X) ▷
  $\bar{\mathbf{x}} \leftarrow \mathbf{x_1}$ ▷ assign a random shape as reference shape
  repeat
    Align all shapes $\mathbf{x_i}$ to $\bar{\mathbf{x}}$ ▷ rigid transformation (and sometimes scale)
    Update reference shape $\bar{\mathbf{x}}$ to be the mean of the aligned shapes $X_{aligned}$
  until convergence
  return $\bar{\mathbf{x}}$ and $X_{aligned}$ ▷
end procedure

Similarity transform. With scale added to the transformation, the algorithm will shrink during iterations. This problem is solved by translating to the origo and then normalising all shapes. When the algorithm terminates the shapes can be scaled back.

3.3.2 Principal Component Analysis

With a high dimensional data set, Principal Component Analysis (PCA) [107] is a strong tool to reduce the dimensionality and remove correlation from the analysis. Figure 3.7 shows a synthetic example of a population of men with weight as a function of height. These are obviously highly correlated. PCA is a projection into a new coordinate system, where the first axis direction maximises the variance in the data set, the second axis is orthogonal to the first pointing in the direction with the second most variance and so on and so forth. In this example the new axes might be translated into physical features such as *size*, as the first component, and *build*, as the second. Since these are now uncorrelated they form a better basis for analysis. It might not always be possible or sensible to try to get an interpretation of the components. Especially for higher dimensions it can be hard to give a meaningful interpretation of what variance a given component describes. A PCA can also help remove dimensions that only contain noise.

In our work on statistical modelling and surface recovery of ears, an ear shape consists of about 3000 correspondence points. Since each point has coordinates in 3D, a shape can be formulated as a single point in a hyper dimensional space with the dimensionality 9000 (3x3000). It seems obvious that for an ear shape in this high dimensional space the 9000 variables are not independent, i.e. points next to each other on the surface are highly correlated. If $\mathbf{x} =$

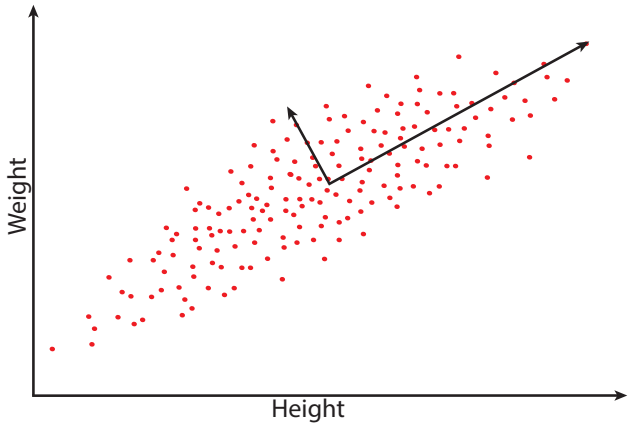


Figure 3.7: Plot of a population of men with weight as a function of height (data made up for illustrative purposes). These measures are obviously highly correlated, which means that if one measure changes it is very likely the other measure will too. The first two principal component axes are shown. The new axes removes correlation in the data and could be labelled *size* and *build*.

$(x_1, y_2, z_3, \ldots, x_n, y_n, z_n)$ is a vector with the spatial coordinates of a shape and the observation matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^T \in \mathbb{R}^{n \times m}$, then the covariance matrix, $\mathbf{\Sigma}$, of $\mathbf{X}$ is:

$$\mathbf{\Sigma} = \frac{1}{m}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n \times n}. \tag{3.13}$$

This covariance matrix can be represented as its Eigenvalue decomposition:

$$\mathbf{\Sigma} = \mathbf{P}\Lambda\mathbf{P}^T, \tag{3.14}$$

where $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_m)$ is a matrix consisting of columns of Eigenvectors and $\Lambda = diag(\lambda_1, \ldots, \lambda_m)$ is a diagonal matrix holding the Eigenvalues (the variance in the direction of the corresponding Eigenvector). Summing up the Eigenvalues account for the total variance in the data. If the Eigenvectors are ordered according to their Eigenvalues, a cumulative summation of the first to the last Eigenvalues will show how many are needed to explain a given percentage of the total variance. Figure 3.7 shows this cumulative summation of variance as a function of the number of Eigenvectors included. This is also referred to as the number of modes. The data set is the final Active Shape Model (ASM) used for the statistical recovery in Chapter 9. The final model consists of 180 aligned shapes with around 3000 correspondence points making the shape space 9000 dimensional (3x3000). As can be seen in Figure 3.7 more than 90% of the variance of the shape model can be explained using the 40 first modes. This is a drastic reduction of dimensionality. When doing the missing surface recovery,

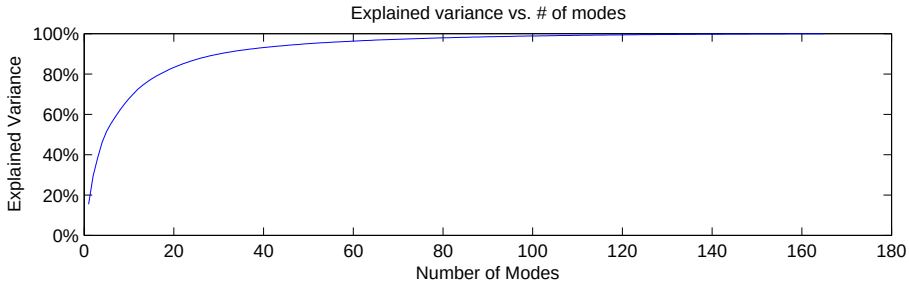


Figure 3.8: Plot of modes vs. variance for the final ASM.

a partial scan was fitted to the statistical model, then fully recovered. This was done in an iterative scheme, where more and more Eigenvectors where included for each iteration. While the collection of Eigenvectors explain all the variance in a data set, it might be reasonable to only use a subset. The vectors with the smaller Eigenvalues might explain variance in the measuring rather than actually explain variation in the population.

3.3.3 Active Shape Modelling

The statistical model used in Chapter 9 was created using a bootstrapped Active Shape Modelling (ASM). A small subset of samples were manually annotated and registered using the approach described by Paulsen et al. in [73, 71] (this work formed the statistical model used in Chapter 8). An Active Shape Model (ASM) was constructed as described in [22, 73]. The statistical model is aligned and fitted to each unknown sample. The fitting is done by aligning an changing the dominant modes according to a PCA model of the registered data. This is done iteratively, allowing co-registration to and inclusion in the ASM, thereby expanding the model sample by sample. The ASM grows in size as unknown samples are registered and included. This allows the model to explain more and more shape variation. Intuitively this leads to the expectation that the algorithm will become increasingly better at fitting to unknown shapes and that latter samples are better registered than former. As described in the paper on Anatomical Surface Recovery (Chapter 9) the ability to fit new shapes improved as more shapes were included.

3.4 Stereo Reconstruction

In Chapter 4 and 11 a stereo rig mounted on a robotic arm was used to create 3D data of a scene. Stereo vision works much like the human vision, where seeing a point from two views can be used to triangulate the points position in space. Figure 3.9 shows a stereo setup seing the same point from two cameras. The cameras are displayed as backprojection, which will be explained in Section 3.4.1. The terms focal point and epipolar line used in the figure, will be explained in Section 3.4.1 and 3.4.2). The triangulation is done with the focal points and a 3D point as the triangle corners. The triangulation relies on the ability to match points in both images. When more than two images are used for a reconstruction it is known as Multiple View Stereopsis (MVS). Adding more images when finding correspondence has the benefit that it makes the correspondence more robust as it is much less prone to produce false positives in the matching. For an image based 3D reconstruction to work, the intrinsic (internal camera) and the extrinsic (orientation) parameters need to be estimated. This can be done using the images alone[92]. However, estimating the parameters during a calibration can produce even better results.

3.4.1 Camera model

The basic camera model is the pinhole camera model, where every ray of light passes through the infinitely small pinhole, the focal point, and is projected onto the image sensor. When light passes the focal point left/right and up/down are flipped. To remove the confusion from this image flipping, the model is portrayed with the sensor in front of the focal point. This is known as backprojection and makes up, down, left, and right correspond in both the image and the real world. The pinhole model does not include lens distortion. Lens distortion will be addressed later.

The pinhole camera model [40] can be defined as follows:

$$\mathbf{x} = K\,[R\mathbf{t}]\,\mathbf{X} \tag{3.15}$$

Where $\mathbf{x} = [x, y, 1]^{\mathrm{T}}$ is the image point and the world point is $\mathbf{X} = [X, Y, Z, 1]^{\mathrm{T}}$, both in homogeneous coordinates. The transformation from world coordinates to camera coordinates is done by R and $\mathbf{t}$, that are a rotation and a translation

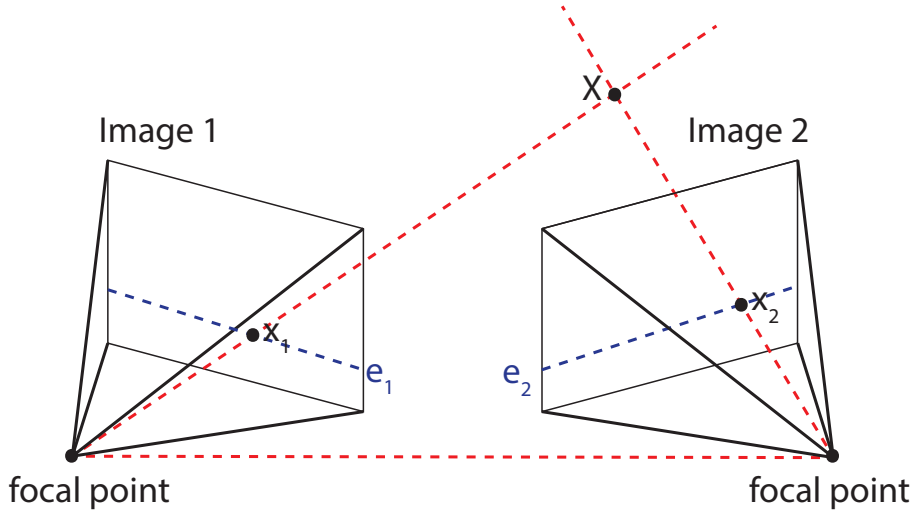


Figure 3.9: Two back projection cameras seing a spatial point $\mathbf{X}$ from different views. The figure also shows the focal points, the image points $\mathbf{x_1}, \mathbf{x_2}$ and the epipolar lines $\mathbf{e_1}, \mathbf{e_2}$.

matrix respectively. The camera matrix is:

$$K = \begin{bmatrix} f * k_x & s & x_0 \\ 0 & f * k_y & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.16}$$

Where $[x_0, y_0]^{\mathrm{T}}$ is the principal point where the optical axis intersects the sensor and f is the focal length. The camera parameters k_x and k_y accommodate for scaling and s is the skewness parameter.

3.4.2 Correspondence between images

To perform stereo reconstruction a robust matching of correspondence points must be implemented. This normally relies on finding features to match in each image [100]. To improve both robustness and speed in matching, the epipolar geometry between images is used. This is given by knowing the camera geometry, both intrinsic and extrinsic. What is seen as a point in one image, is essentially a ray of light. The projection of this ray of light can be expressed as a line going across the image in the other camera as shown in Figure 3.9. Therefore, a point in one image has a line of possible matches in the other. This is an epipolar line, shown as e_1 and e_2 in 3.9. When matching features from one

image to the other one only has to look for matches along the epipolar line. This line eliminates the risk of false positives from matchings that cannot exist (the intersection of two rays that intersect in reality). To speed things up further, both camera images can be rectified such that the rows in each corrected image constitute the epipolar lines. This way the lines does not have to be recomputed for each matching.

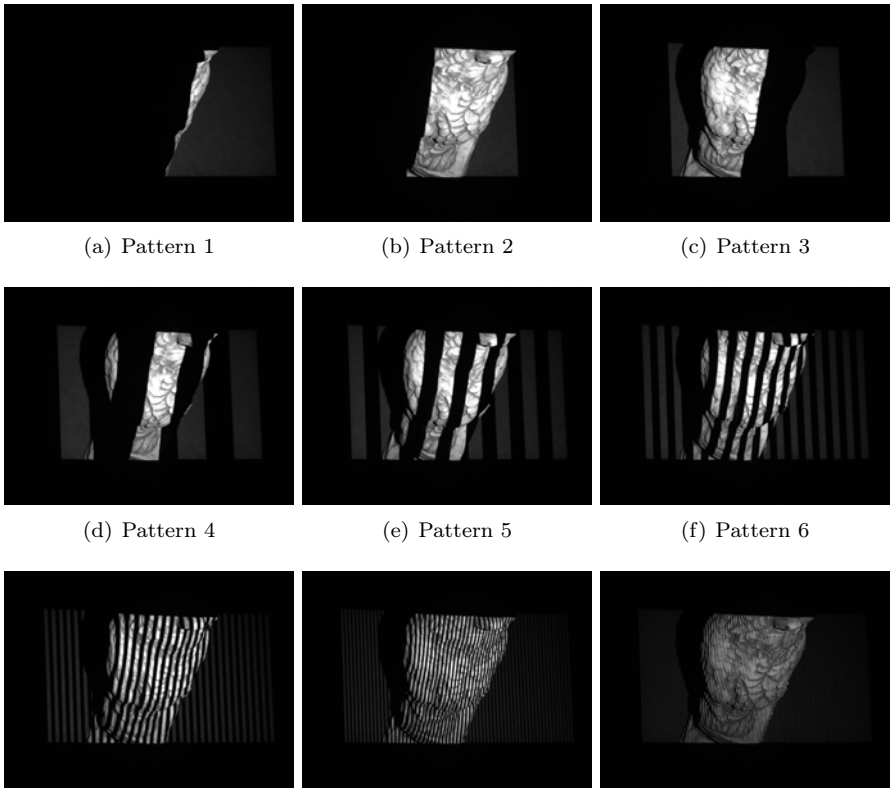
3.4.3 Gray encoding

To do the stereo matching in our setup we chose not to rely on normal feature matching. Instead we opted for an active method of projecting structured light to support the matching [8]. Stereo vision with structured light can be done with a projector replacing one of the cameras. We did however add a projector to the stereo rig such that we could rely on good dependable optics and only use the projection of patterns as an aid in the correspondence matching.

The patterns used are known as a Gray coding, where a stack of projected patterns create a linewise bit encoding [8]. Figure 3.10 shows nine encoding patterns. The nine patterns allows for a total of $2^9 - 1 = 511$ different line codes across the width of the projector. As the native resolution of the projector is 854x480, nine is the maximum number of patterns that can be used. There exist more sophisticated ways of projecting patterns and the color channels can be used to project several patterns simultaneously. However, the Gray encoding is very robust. Because the patterns consist of only black and white, they have a good signal to noise ratio. To improve even further on the signal to noise ratio, we also projected each pattern in its negative version and used the difference between the two images as the final response. In Chapter 10 and 11 the projection scheme allowed for a 3D surface reconstruction of very difficult objects. Both with high complexity and specularities that are normally hard to reconstruct.

3.4.4 Lens distortion

While being a good model, the pinhole camera with an infinitely small focal point would let in very little light, which defies the purpose of capturing light on a sensor. A camera lens allows for more light through a bigger aperture as a trade-off by limiting the depth of field. Since no optical system is flawless, lens distortion has to be taken into account when doing the 3D reconstruction. The



(a) Pattern 1 (b) Pattern 2 (c) Pattern 3

(d) Pattern 4 (e) Pattern 5 (f) Pattern 6

(g) Pattern 7 (h) Pattern 8 (i) Pattern 9

Figure 3.10: The nine patterns of a Gray Encoding implemented projected onto a statuette of an owl. To get the maximum signal to noise ratio each pattern is projected in both a positive and negative version. The effective pattern is the difference between the two. Stacking the patterns on top of each other will create a unique encoding going from left to right.

lens distortion can be modelled as [41]:

$$\begin{bmatrix} x_d \\ y_d \end{bmatrix} = \left(1 + k_1 r^2 + k_2 r^4 + k_3 r^6\right) \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 2k_4 xy + k_5(r^2 + x^2) \\ k_4(r^2 + y^2) + 2k_5 xy \end{bmatrix} \quad (3.17)$$

Where $r^2 = x^2 + y^2$ is the radius, and k_1, k_2, k_3 are parameters for the radial distortion (the higher order k_3 can generally be omitted). The parameters k_4, k_5 controls the tangential distortion. In the calibration of our setup k_3 was in fact omitted and the remaining four parameters where sufficient to model the lens distortion. Generally, for practical use, the inverse solution is wanted. This has to be numerically estimated.

Chapter 4

Data set for stereo and multiple view 3D reconstruction and validation

Stereo and Multiple View reconstruction of 3D data is a well studied problem [40]. However, availability of data sets is very limited [95, 87]. The variety is also somewhat limited to surface textures, which are easy for correspondence matching between frames. Therefore, developments has to be done either on the limited data sets or acquired for the purpose, which might be in a less controlled fashion. Development and validation on a limited range of data might result in overfitting or as expressed by Everingham et al. [28]:

> A question often overlooked by the computer vision community when comparing results on a given dataset is whether the difference in performance of two methods is statistically significant.

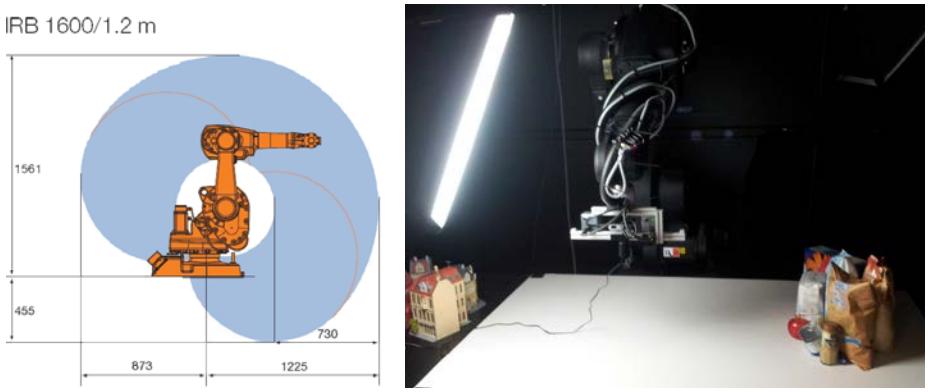
Our objective was to create a more complete data set, acquired using a controlled setup and with a wider range of scenes. In an existing data set created at The Technical University of Denmark [4], data was acquired with a camera mounted on a six axis industrial robot arm. The robot would move to several positions on arcs in a plane, and images were acquired from each position. A stationary structured light projector was used to support correspondence between image pairs from different positions in the sweep. Therefore, the stereo reconstruction was limited to what is illuminated by the projector and viewable from the different camera positions.

In an effort to get a better coverage and a more versatile acquisition, we mounted two cameras and a mini projector on the actual industrial robot arm as seen in Figure 4.1(b). This allows for a better coverage as the structured light projector moves with the cameras. The robot has been programmed to move to positions distributed on concentric spheres. The robot arm has very high repeatability and provides full control over the camera positions in the experimental design. This is a very agile solution compared to a rigid cameras setup on a frame, which would require a camera at each position. Also, being able to take images from positions placed on concentric spheres is only possible as the robot arm does not obstruct the view of itself. The ability to freely design the camera positions in the setup makes the combination of the robot arm and a stereo setup quite unique. At each position a series of structured light images are acquired followed by a number of images with the projector turned of and a variable illumination provided by an array of LEDs (Light Emitting Diodes). The resulting data set can be used for evaluation of stereo and multiple view algorithms along with a lot of other uses. In addition, the produced point clouds are very dense and form a good basis for advances in surface reconstruction algorithms. Since texture information is available for all points in the set, it can also be used to create some nice photo realistic models. The data set includes:

- Images: raw, cleaned and rectified. These can be used both for multiple view reconstructions and as texture for the 3D point sets.
- Light variation: images acquired under different light settings.
- Calibration: intrinsic and extrinsic parameters.
- Accurate, dense and unbiased 3D point sets.

4.1 Industrial robot arm

The core of the setup for the data acquisition is an industrial robotic arm, which allows for free camera positioning. The ABB IRB1600 robotic arm has a very high repeatability. Once the arm is programmed to go to a number of positions, this movement or sweep can be repeated with no stochastic error. The repeatability allows for a calibration sweep prior to the data acquisition. The robot arm allows for an experiment design, only restricted by the reach of the robot. Figure 4.1(a) shows a schematic of the robot arm and its range of motion. The setup of the robot arm inside the black robot cage is shown in Figure 4.1(b). The robot is controlled by dedicated robot controller, which is connected via ethernet to a PC.



(a) Schematic of robot (b) Robot setup in cage

Figure 4.1: (a) a schematic of the robot arm and its range of motion (ABB Group, *www.abb.com*). (b) the robot arm inside the robot cage. The arm itself and the walls of the cage are painted black to remove reflecting light (it also makes it hard to take good photos of the setup).

4.2 Acquisition setup

On the end of the arm, the acquisition setup is mounted. This consists of two Point Grey Scorpion 2.0M pixel cameras and an Acer C20 mini projector. The cameras are mounted with 12mm Schneider optics with the aperture fixed to a narrow setting to provide a good depth of field. The mount is shown in Figure 4.2. The cameras are connected via Firewire and the projector via VGA. While the cameras are obviously used to record images, the projector is used to project a structured light pattern used for stereo matching. At each position of the robot sweep, the cameras acquire images during the projection of the structured light. Then the projector turns off and a new series of images are recorded during which, the light is varied using an array of LEDs mounted on the ceiling of the robot cage. An image is also taken with no illumination. This image is used to remove *dark* noise from the other images.

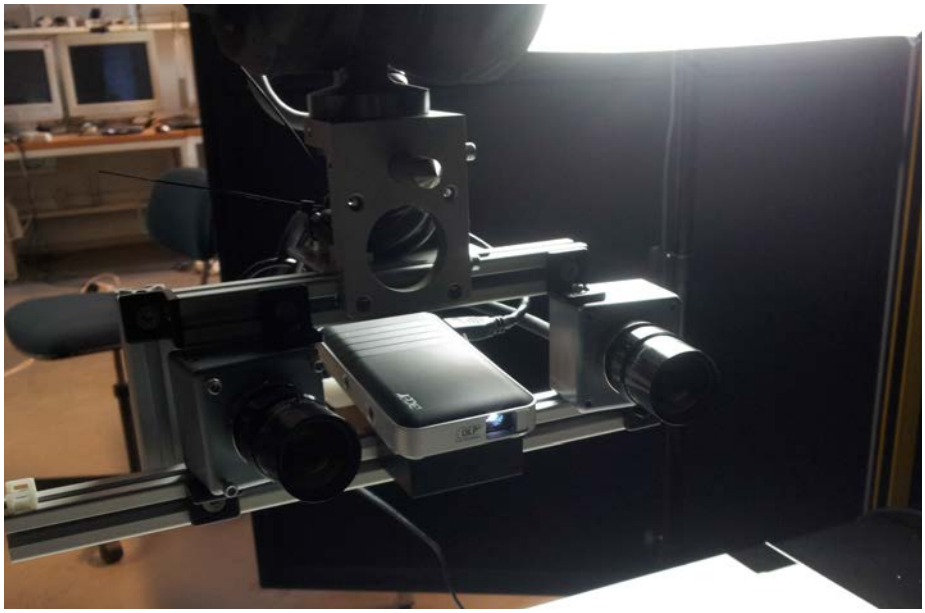


Figure 4.2: Two Point Grey Scorpion cameras and an Acer C20 mini projector mounted on the robot arm.

4.3 Scene illumination

Mounted on the ceiling of the robot cage is an array of LEDs. During acquisition the LEDs are turned on row-wise from left to right until all LEDs are lit, then they are turned of row-wise from left to right. This illumination scheme somewhat simulates morning to evening. Figure 4.3 shows simulated morning, noon and evening. The LEDs are turned on and off by a light controller connected to a PC via a serial port. The variation in illumination allows for analysis of the influence of light in image based 3D reconstruction algorithm.

(a) *Morning* (b) *Noon* (c) *Evening*

Figure 4.3: A cute little bunny with simulated morning, noon and evening illumination (every university should have a detailed scan of a bunny).

4.4 Controlling the setup

At the center of it all is a computer running a program coded in C++, interfacing with the camera, robot, and light setup. The projection of patterns is handled by OpenGL[1]. The robot controller runs a script communicating through an ethernet connection with the PC. The cameras are connected via Firewire, the projector via VGA and finally the LEDs are controlled by a controller connected via a serial port. A schematic of the setup is shown in Figure 4.4. The flowchart routine for the data acquisition sweep is shown in Figure 4.5.

[1] OpenGL, www.opengl.org

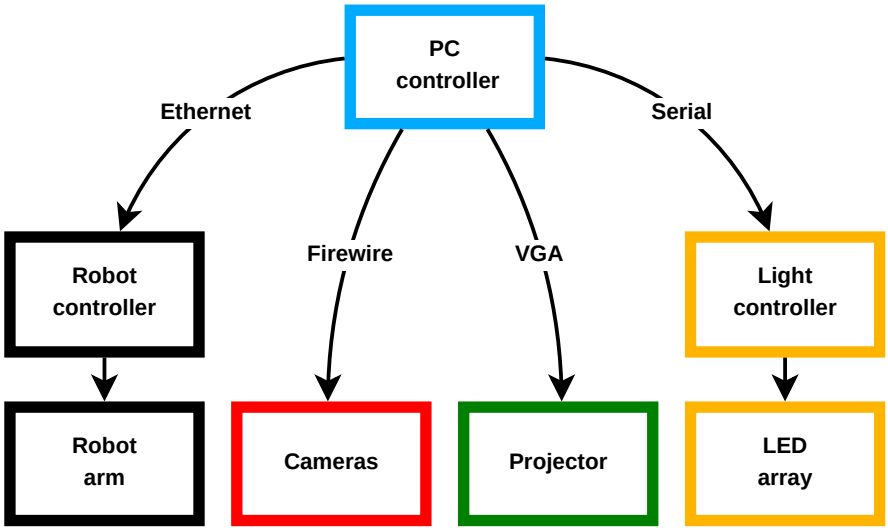


Figure 4.4: A schematic showing the control of the experimental setup.

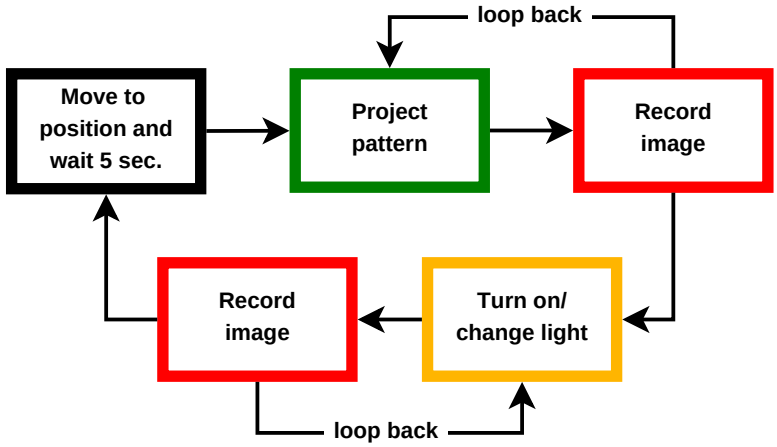


Figure 4.5: Flowchart showing the actions of the system. First thing is the robot moving to a new position. The process has loop backs during the structured light image acquisition and during the varying light acquisition. In the end the process loops back to the robot moving to a new position.

4.5 Calibration

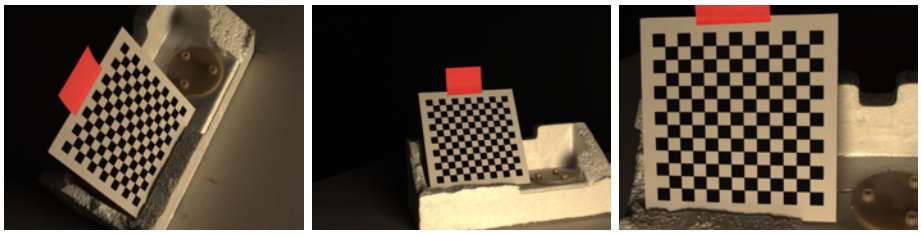
Once everything is connected the whole setup is calibrated, estimating both the extrinsic and intrinsic parameters. We used a publicly available toolbox for Matlab [47]. The toolbox provides a full calibration of both extrinsic and intrinsic parameters given a series of images of a planar checkerboard.

During the calibration the extrinsic parameters, being position and rotation, are found. The relative transformation between each camera is estimated. This remains constant. The spatial position and orientation, in each acquisition point in the sweep, are also found. The intrinsic camera parameters are also estimated. The calibration of the robot is done such that both the calibration of camera parameters and positions are found simultaneously. To do this the cameras should be able to see the calibration object from all positions. As the calibration relies on a quality and precise calibration object, we used a checkerboard in a high quality print, glued onto a glass plate to make it perfectly planar. The robot is programmed to move to image acquisition positions distributed on a sphere with the camera pointing towards the center of the sphere. To get an even sample distance in the positions, the robot moves along lines of latitude with constant spacing between positions. The same spacing is used between the lines of latitude making the spacing almost equidistant. This distribution removes the oversampling near spherical poles, that a point distribution using both lines of longitude and latitude would inflict.

In the final data set, the robot was programmed to move to 49 spherical positions with radius of 500mm to the center. In addition, on some sweeps an extra 15 positions were added, placed on a concentric sphere with a radius of 700mm. These 49 or 64 positions were used during data acquisition. To get a good estimate of the intrinsic parameters it is also important to get images with good coverage of the calibration object. On top of the 64 positions used in the data acquisition another 30 positions were added, where the robot moves much closer to the object to get almost full coverage of the checkerboard. The closeup images are very important in robustly estimating the intrinsic parameters, which directly influence the quality of the estimated extrinsic parameters. Figure 4.6 shows three calibration images with radius 500mm and 700mm to the center and a closeup with good coverage.

4.5.1 Validation of extrinsic parameters

The system depends on the robots ability to accurately go to the programmed positions with no stochastic error. Also, it is very important that over time the



(a) 500mm (b) 700mm (c) close up with good coverage

Figure 4.6: Three of the 94 calibration images used to find both intrinsic and extrinsic parameters.

robot has little to no drift. Ten calibration sweeps where carried out over a two month period during the time the data were collected. This allowed for a check of repeatability and drift as the estimated camera positions can be compared. The evaluation of the repeatability is done in the following subsection. Based on a calibration Figure 4.7 shows the calculated 49 positions that constitute the inner shell in the acquisition sweep. The numbers are assigned according to order of the positions. In some scans an additional 15 positions are added with a greater radius to the center. These are not shown in the figure.

4.5.2 Aligning the calculated positions

For each calibration, the checkerboard was placed manually on markers in the center of the scene. The orientation of the calibration object dictates the orientation of the resulting camera position estimates. Therefore, offset in the positioning of the checkerboard shows as offset in the estimated camera positions. This is not an error, it is just a rigid transformation. For the positions to be compared they have to be aligned. This is done using a rigid Procrustes Analysis, which produces the best available alignment (see Section 3.3.1). Figure 4.8 shows the resulting camera position estimates from 10 calibrations before and after alignment. The close up of the graph presents a better view of the difference in positions, which is why not all points are showing.

The aligned shapes are compared to the resulting mean from the Procrustes Analysis. The Euclidian deviation from the mean is shown in Figure 4.9(a). The deviation is shown as a function of the position for each of the 10 calibrations. The result shows that all but one deviations are well below 0.2mm. Figure 4.9(b) shows the standard deviation for the 10 sweeps as a function of position, while

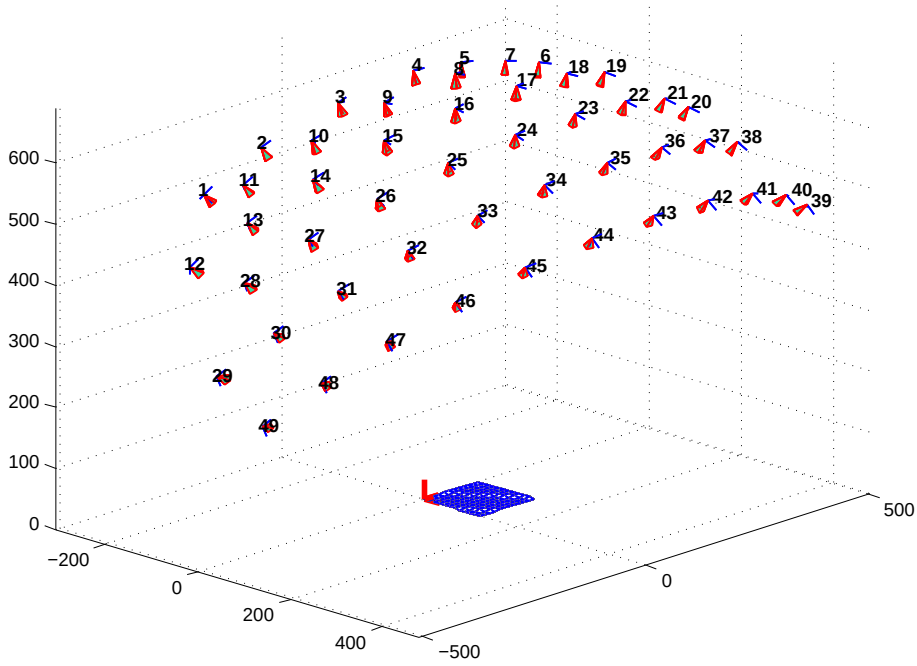
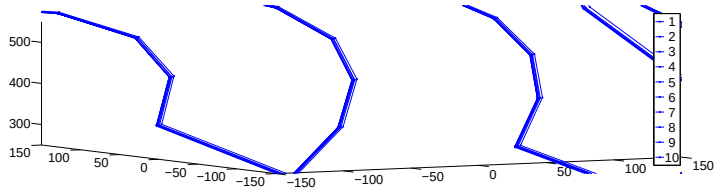
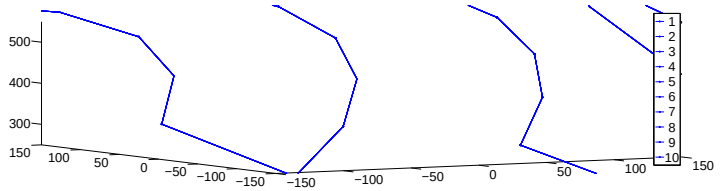


Figure 4.7: The 49 camera positions as found during calibration. The origo is in the corner of the calibration checkerboard and the distances are measured in millimeters.



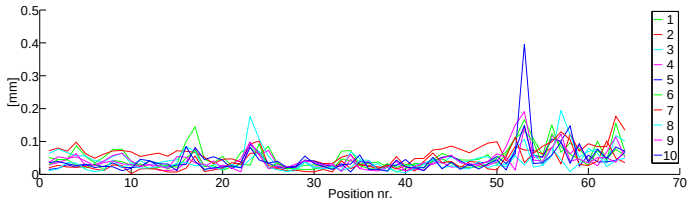
(a) Unaligned positions.



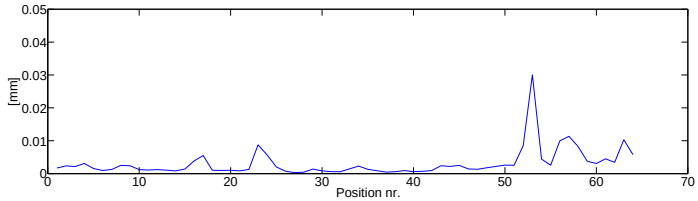
(b) Aligned positions.

Figure 4.8: The estimated positions from the calibration before alignment (a) and after rigid alignment using Procrustes Analysis (b). After alignment the positions and connecting edges between positions are perfectly overlapping.

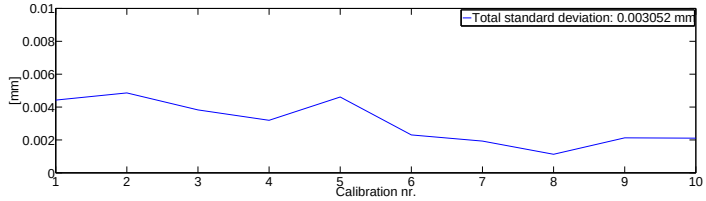
Figure 4.9(c) shows the standard deviation as a function of the calibration sweep. As can be seen on all three figures the deviations are very small. The overall standard deviation for all positions in the 10 sweeps is a negligible 0.0031mm. Since calibration 8 is the one with the smallest standard deviation this was used for the reconstruction.



(a) Deviation from mean as a function of position for all calibration sweeps.



(b) Standard deviation as a function of position.



(c) Standard deviation as a function of calibration.

Figure 4.9: The deviation of estimated positions from a mean found using Procrustes Analysis (a). The standard deviation is shown as a function of position and calibration in (b) and (c).

4.6 Point reconstruction

With a full calibration of the camera setup and the positions during a scan sweep, points can now be reconstructed. The structured light is used for correspondence and the extrinsic parameters of the relative position between the two stereo cameras as well as their individual intrinsic parameters allows for reconstruction of 3D patches from each position. Knowing the global position of the cameras during a sweep allows for a transformation into a common coordinate system. The images recorded during the variable lighting can be used to colour the reconstructed points by back projection. To create an image with minimal shadows, the images recorded can be stacked and a maximum value for each pixel can be found. This creates a *shadowless* maximum image. In Figure 11.8 a scan of a collection of scale model houses are shown along with an example of a *maximum* image.

(a) Model house with textured points

(b) Texture *without* shadows

Figure 4.10: A scan of scale model houses is shown in (a). A example of a *shadowless* image used to texture 3D points is shown in (b). The *shadowless* image is created by stacking the images taken during varying lighting and taking the pixelwise maximum.

4.7 Point validation

Checking the quality of the reconstructed data is not a simple task, though a few quality checks can be made manually. Figure 4.11 shows four different views of a reconstructed owl. The overall impression of the scan is good, there seem to be next to no noise, close ups reveal good detail and no discontinuities from the individual scan alignments. When seen from the side the surface, on which the owls stands, is rendered perfectly flat and paper thin. Other surfaces areas are also paper thin. Bad alignment of scans and bad estimates of the intrinsic parameters would show up as discontinuities in the resulting point reconstruction and finally bad correspondences in the stereo matching would produce noisy surfaces. The reconstruction has none of the above.

4.7.1 Validation from a bowling ball

To test the data quality we have scanned a bowling ball because its spherical shape makes it a useful calibration object. Good bowling balls have very little deviation from their spherical shape, which is why they are so expensive. Figure 4.12 shows the bowling ball, where 49 partial scans have been placed in a common coordinate system. The points from each position are coloured with a different colour. From the reconstructed surface points, we can estimate the centre position and radius. Given the center and radius estimates, how each point deviates from the sphere can be used to recover the reconstruction variance.

Let the center point be $\mathbf{P}_c$, a reconstructed point on the sphere be $\mathbf{P}_i$, and the radius be r. For any point on the sphere we have:

$$||\mathbf{P}_c - \mathbf{P}_i||_2 = r^2 \ . \tag{4.1}$$

In coordinates this can be written as:

$$(x_c - x_{p,i})^2 + (y_c - y_{p,i})^2 + (z_c - z_{p,i})^2 - r^2 = 0 \ . \tag{4.2}$$

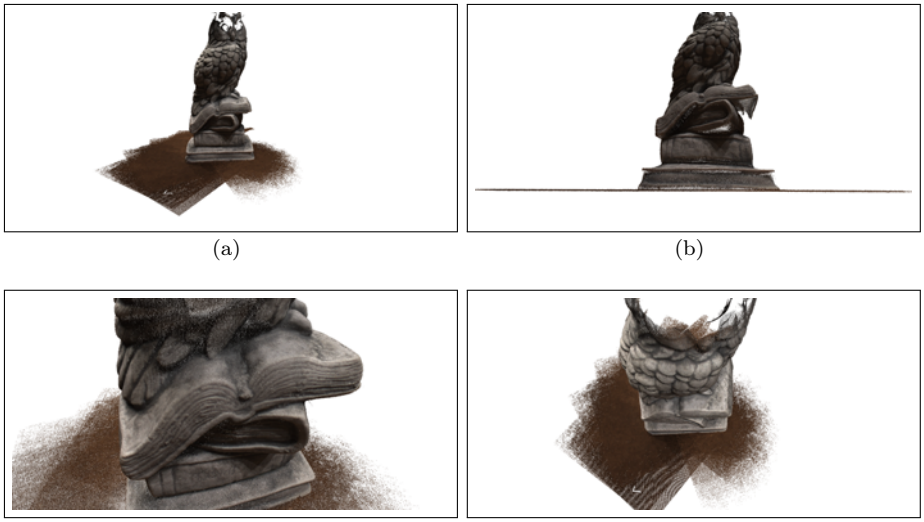


Figure 4.11: A reconstruction of an owl from 4 different views. There has been no surface reconstruction, it is just a very dense point cloud (the owl is not what it seems).

This can be written as a matrix multiplication:

$$[x_{p,i}^2 + y_{p,i}^2 + z_{p,i}^2, -2x_{p,i}, -2y_{p,i}, -2z_{p,i}, 1] \begin{bmatrix} 1 \\ x_c \\ y_c \\ z_c \\ x_c^2 + y_c^2 + z_c^2 - r^2 \end{bmatrix} = 0 \ . \quad (4.3)$$

The left side of Equation 4.3 is stacked in a matrix $\mathbf{A}$. From the inner product matrix $\mathbf{Z} = \mathbf{A}^T\mathbf{A}$ we obtain the eigenvalue decomposition $\lambda\mathbf{Z} = \mathbf{L}\mathbf{Z}$, where λ are the eigenvalues and $\mathbf{L}$ the corresponding eigenvectors. $\mathbf{l}$ is the eigenvector corresponding to the smallest eigenvalue and contains the least squares solution of the centre and radius. From this we estimate the center by:

$$\overline{\mathbf{P}}_c = \begin{bmatrix} \bar{x}_c \\ \bar{y}_c \\ \bar{z}_c \end{bmatrix} = \begin{bmatrix} \mathbf{l}[2]/\mathbf{l}[1] \\ \mathbf{l}[3]/\mathbf{l}[1] \\ \mathbf{l}[4]/\mathbf{l}[1] \end{bmatrix} \ . \quad (4.4)$$

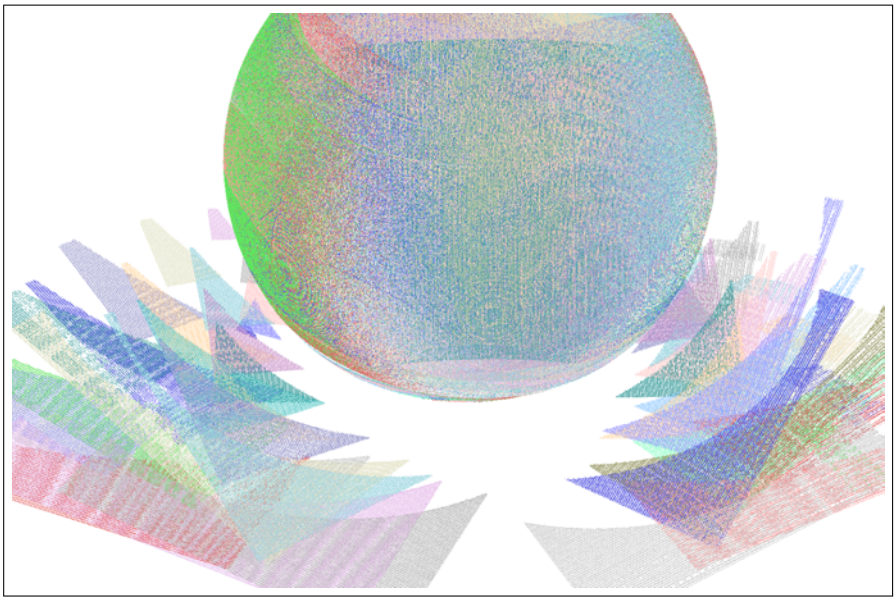


Figure 4.12: The globally aligned 3D points of a bowling ball acquired from 49 positions. The individual scans from each position has a different colour.

$\mathbf{l}[n]$ is the n'th element of $\mathbf{l}$. The radius is estimated as:

$$\bar{r} = \sqrt{\bar{x}_c^2 + \bar{y}_c^2 + \bar{z}_c^2 - \mathbf{l}[5]/\mathbf{l}[1]} \ . \tag{4.5}$$

4.7.2 Results of point evaluation

Table 4.1 shows the parameter estimates for all points in one set. Figure 4.13 shows the estimated parameters for the individual scans, Table A.1 and A.2 in Appendix A.1 list the estimates.

$\bar{x}_c$	$\bar{y}_c$	$\bar{z}_c$	$\bar{r}$	σ
-11.39	-25.41	630.16	133.27	0.1449

Table 4.1: Estimated center coordinates $(\bar{x}_c, \bar{y}_c, \bar{z}_c)$, radius $(\bar{x}_r)$ and standard deviation of individual points (σ). All numbers are in mm.

The bowling ball data set has a standard deviation of 0.14 mm for all scans

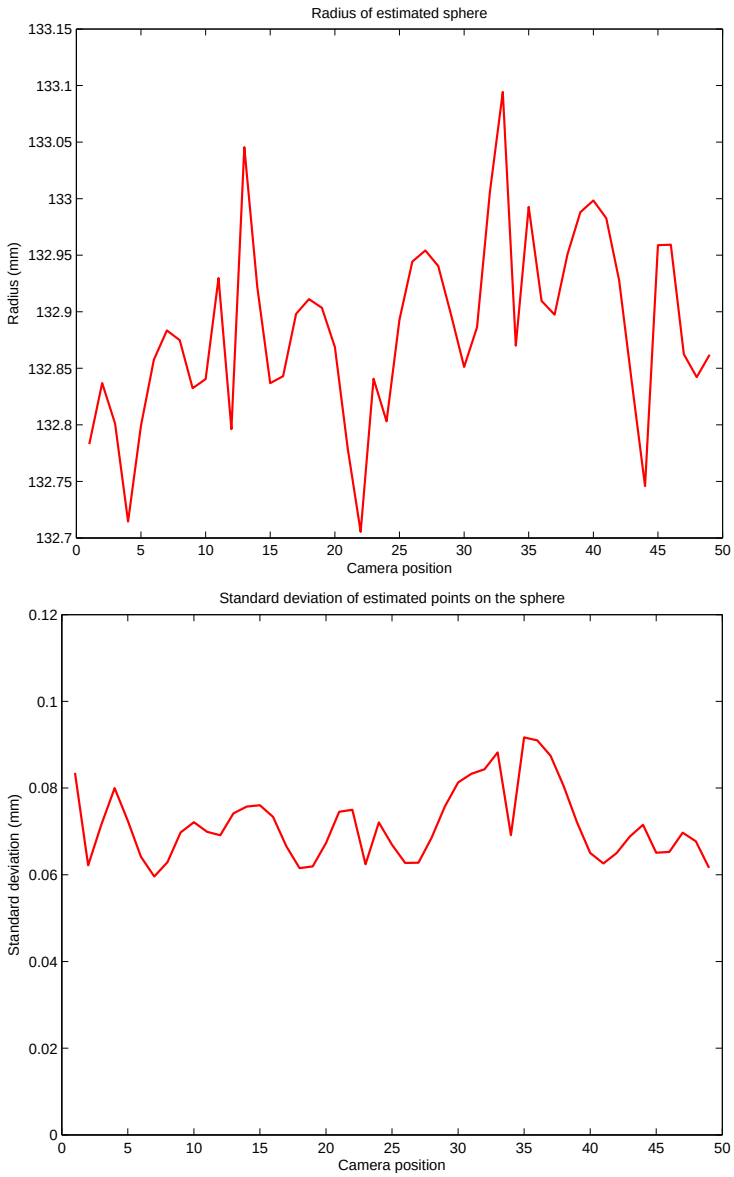


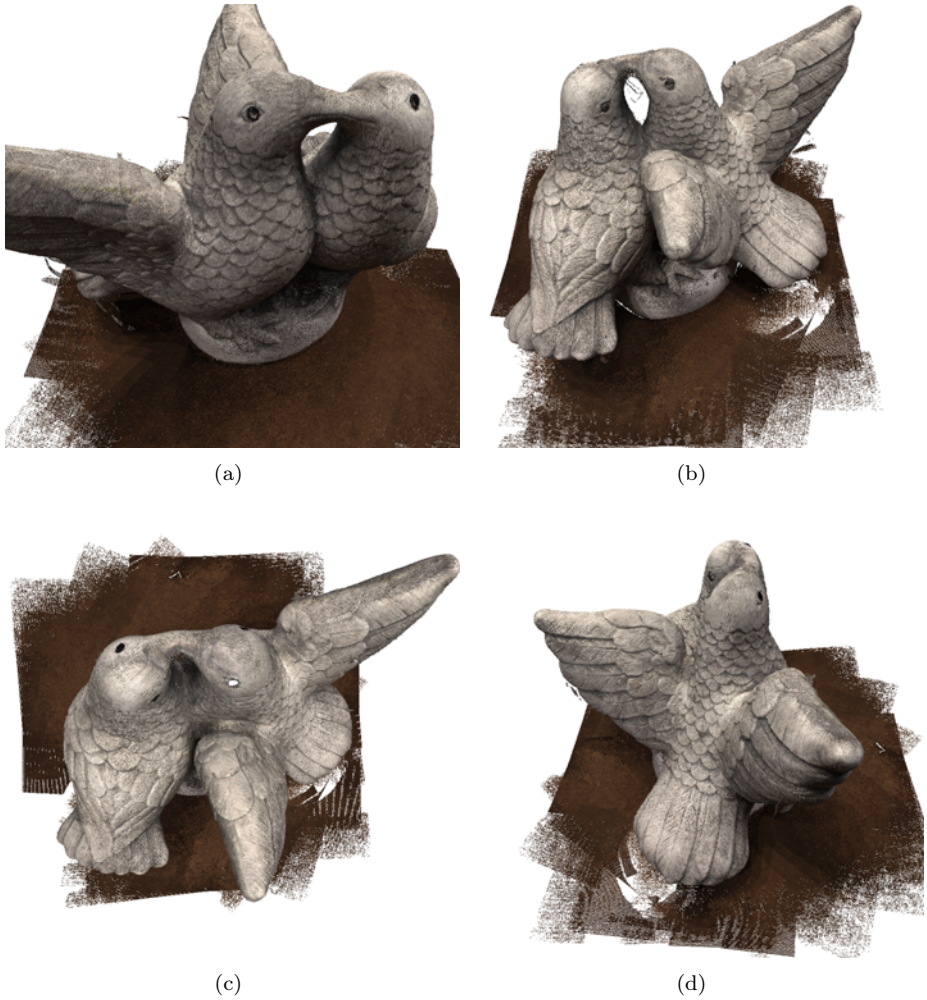
Figure 4.13: Estimated radii of the bowling ball (left) and point standard deviation from the radius for the individual scans (right).

collected and around 0.08 mm for the individual scans. With a pixel size of around 0.22 mm we have a standard deviation of around 0.6 pixels. Given the estimated standard deviation on the bowling ball and the standard deviation of the positions in the 10 calibration sweep a conservative guarantee of the point reconstruction would be $<$0.25mm error.

4.8 The final data set

The data set consists of over 100 scenes, out of which some are 360 degree scans. In the 360 degree scans, the scenes were scanned four times with a 90 degree rotation between each scan sweep. This was done with the scene placed on a wooden board that pivots around a simple screw. While this is indeed simple, the 90 degree rotations are almost identical between scans and the final alignment can be done using ICP (see 3.2.1). Figure 4.14 shows a full 360 degree reconstruction of a statuette of a couple of doves.

With about 15 million points in each scan, and over 100 scenes, the total data set contains over a 1 billion points. The vast number of points makes it ideal for statistical analysis. In Chaper 11 we used both the solid statistical foundation along with the versatile selection of scenes to do analysis on how MVS algorithms work in relation to texture.



(a) (b)

(c) (d)

Figure 4.14: A 360 degree scan created by combining four 90 degree rotations of the scene. The final model consists of 47 million points.

Chapter 5

Discussion and conclusion

One of the objectives of this thesis was to find a robust method that could filter noise and hair in surface scans from a direct in-ear scanner. The work on noise filtering in Chapter 6 was done at a time, when the scanner probe was in its very early stages. Therefore data was used from a more developed prototype for dental impression taking. The presence of noise from hair in the scans was therefore not really known for the in-ear scanner. The method proposed was based on an expectation that noise would not be Gaussian. Scanner noise was expected to be random (salt and pepper) and noise from hair would be fine structures far from the surface. The method was developed such that actual skin surface data would not be corrupted. It was developed to handle the worst case scenario in scans with a lot noise from hair. The algorithmic work on the direct surface scan data has been done in parallel with hardware developments. This is complicated as it is hard to predict the final outcome and the resulting challenges. It might be that the filtering method proposed in Chapter 6 will the not be part of the final product. The method proposed in Chapter 8 and 9 might solve the final noise problem.

The contribution on Intracranial Volume estimation in Chapter 7 is a little bit the odd one out, as it does not relate directly to either ear scanning or stereopsis. The method presented is however very related to the method proposed on noise filtering, where a specifically formulated MRF is used to find the correct volumetric segmentation. The method is an example of the flexibility of an MRF formulated to suit a given problem.

Another issue relating to direct in-ear scanning was data recovery in partial scans due to occlusion, noise and hair. This problem was addressed in both Chapter 8 and 9, where statistical models were used as a strong prior to estimate missing surface areas. In the first contribution the statistical model was created on a set of 29 manually annotated ear impressions. In the second, impressions were

also used but this time a small annotated set was used to create an ASM, which was automatically and iteratively. In both cases my contribution was mainly in how to use the statistical model to do the recovery rather than achieving correspondence across the population, be that manually or automatically. The results are promising for this approach. However, it is a very specific approach as it needs a statistical model of the shape to be reconstructed. This is not always straight forward to create. In the second contribution the method proved to handle noise in scans very well. This approach might be a solution to the problem of noise from both scanner and hair altogether in direct ear scanning. There is good reason to think that given a statistical model this approach can be useful in other areas. An example could be facial scans, where facial features can support model creation, and the method can recover missing parts. The method has been successfully tested on 10 test persons, where direct in-ear scans have resulted in a good custom ear plugs for all 20 scans. It seems likely that the surface recovery method in Chapter 9 will be part of a final framework for the direct scanning.

Great care and effort had to be exercised in the creation of the experimental setup described in Chapter 4. The design and implementation of the setup had to be refined during several iterations. Each iteration served the purpose of getting a improvements in the calibrations and maximising the coverage in each sweep, while still keeping the movement within the reach of the robot. The resulting *near* non-existing error in the position estimates along with the low standard deviation in the points estimates strongly supports the high quality of the data. In the final setup each scan takes nearly 1h45 of acquisition time. Therefore, it was of outmost importance that the system functioned consistently, as errors could potentially ruin many days, if not months, of work. The final data set amounts to more than 1TB of data and the processing time including the stereo and multiview reconstructions for comparison was several weeks. Producing data might not be the most glorious work, and while being a cumbersome task, it is however important. Only with a large set of data can methods and metrics be significantly evaluated. This solid foundation was used in Chapter 10 to test metrics with respect to specularities. Also, the data set was used to analyse performance of state of the art Multiple View Stereopsis algorithms in the contribution in Chapter 11 in respect to texture. Since the data will be available online, they have a great potential of being used for research in Multiple View Stereopsis, surface reconstruction and photo realistic modelling. Whatever they are used for, evaluations, analyses, and observations based on more than a billion points can safely be called significant.

As a concluding remark I think my combined work comes close to fulfilling the objectives of the thesis. In some small way, it might even contribute to making the world a better place.

Chapter 6

Ultra Fast Optical Sectioning: Signal preserving filtering and surface reconstruction

Rasmus R. Jensen, Mike van der Poel, Rasmus Larsen, and Rasmus R. Paulsen

Abstract

In 3D surface scanning it is desirable to filter away *bad* data without altering the quality of the remaining *good* data. Filtering of raw scanner data before surface reconstruction can minimize the induced error and improve on the probability of reconstructing the true surface. If outliers consist of actual data such as hair, and not just evenly distributed noise, these outliers tend to err smoothing algorithms away from the wanted result. We present a novel algorithm based on a Markov Random Field that uses a distance constraint to robustly classify a 3D scan volume. Through this classification a signal preserving filtering of the data set is done. The remaining data are used for a smooth surface reconstruction creating very plausible surfaces. The data used in our work comes from a newly developed hand held 3D scanner. The scanner is an Ultra Fast Optical Sectioning scanner, which is able to extract high quality 3D surface points from 2D images recorded at over 3000 fps. The scanner has been developed for digital impression taking in the dental area. Our work relates to future in-ear scanning for fitting custom hearing aids without impression taking.

Keywords: *3D scanning, Markov Random Field, computer vision, surface reconstruction, noise filtering*

6.1 Introduction

3D surface acquisition is an established and active research and development area. Novel applications and devices continues to emerge, where the data scale ranges from minuscule in microscopy optical sectioning [68] to large scale aerial surface laser scanning of the earth [104]. A variety of scanners and cameras exist; each with their own strengths and weaknesses. While some scanners produce 3D surface data, scanning an object from one direction is known as 2.5D scanning as it only portrays the object from one side and does not provide a full 3D model. To construct a 3D model several 2.5D scans need to be patched creating a full reconstruction [75].

We have worked with a new scanner, the Ultra Fast Optical Sectioning TRIOS scanner from 3Shape[3], which has been developed to facilitate 3D impression taking in the dental area. Our work is a study on how to reconstruct surfaces in the presence of structured noise. The study is a preliminary study, which relates to fitting custom hearing aids, where the construction of an in-the-ear scanner would make the ear canal impression step obsolete.

Anticipating the problem of structured noise from hair in the ear canal we want an algorithm that filters the data and leaves only valid surface data. As no scanner has yet been produced that will actually go into the ear, we have used data of hairy arms and bearded chins recorded by the TRIOS scanner. The scanner produces high quality data with some very sparse *salt-and-pepper* type noise and also good scanning of actual hair strands. As the scanner is a dental scanner with both high precision and accuracy, we were faced with a specific problem of removing only the hair without degrading the remaining data.

Simple mean or median filtering can make any surface fair (if one smoothens enough) but these filters also distort the data set and are useable only when the noise is Gaussian (mean filtering) or when the number of outliers are few (median filtering). An adaptive application of such filters used only on outliers can remove noise without too much degrading of the data, but such an approach would break down in areas with a lot of outliers. Generally, local filtering does only preserve local structure; for areas with a lot of hair a global method is needed.

Implicit functions err towards outliers and noise. Splines [80] are a well known tool for creating both smooth curves and surfaces, but if noise is not Gaussian the smoothing will be skewed. This is the case in our data, where outliers are mainly found above the surface. Splines are also continuous and therefore do not handle discontinuity well.

A way of removing outliers in a data set is to use random sampling such as RANSAC [29] and fit to the random sample until a fit matches the data well. Even though this removes the influence of outliers it does not guarantee an optimal match. A RANSAC approach is described in [86], where an algorithm is created that finds basic shapes and structures in noisy data sets.

There exist a large body of literature covering noise properties and handling in direct surface scanners [24]. However, the used scanning device is not directly comparable with devices previously investigated.

To maintain as much of the good data as possible in our scans, we have solved our problem based a *Markov Random Field* (MRF) formulation on a 3D voxel grid. An early description of MRFs in 2D image analysis is on the noise removal in dirty pictures addressed in [10]. The novelty of our work is the use of a 3D MRF with a distance based smoothness prior that classifies the data set into surface data and not surface data. The classification allows for a signal preserving filtering of the data set before any actual surface reconstruction.

6.2 Data

The data come from an Ultra Fast Optical Sectioning TRIOS scanner, which records a stream of 2D images. Approximately 130 images are collected into a set that constitutes a voxel volume. From known changes in the scanner during the volume acquisition a scan surface is constructed which can be converted into a real world coordinate system through a calibration step. The scanner computes the voxels that defines the interface between air and solid material as seen from the scanners viewpoint. Since the scanner is known to be above the surface it is possible to label the voxels along the depth axis as either *above* or *below* the sought surface. When the scanner firmware has determined which voxels belong to the interface, the row, column and depth volume is transformed into real world 3D points using the calibrating parameters. The resulting 2.5D point cloud represents the interface between air and solid material as seen from the scanners viewpoint. To perform a full 3D surface scan, the scanner is moved around the object and the partial scans are merged together using a proprietary algorithm.

The scanner is normally used for digital impression taking in dental work, which requires a very high accuracy. Therefore, the quality of scans are very high and the noise levels due to the scanner hardware is minimal. However, real physical objects as for example hair will also be captured by the scanner. In the current application (direct ear-scanning) we are interested in the true surface of the ear

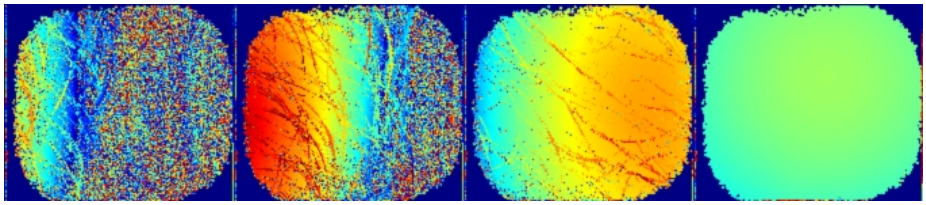


Figure 6.1: Four depthmaps of scan surfaces. From left to right: Little surface coverage with hair, half coverage with hair, full coverage with hair, full coverage without hair.

and therefore hair is considered noise and should be removed from the scan. In the following we consider hair as being structured noise.

A 2.5D scan can also be considered a depth map, where the pixel value reflect the distance to the object. Figure 6.1 shows four depth maps of scans from different surfaces. The scans constitute a range from very bad to perfect, and the algorithm should be able to handle all examples. When there is a lot of structured noise in the scan, we need an algorithm that does not break down but does a good filtering leaving only the actual surface data (even if it is very little).

6.3 Markov Random Field volume classification

The raw output from the scanner (and the scanner firmware) is a voxel set where the scanner is virtually placed above the voxel volume looking down the depth direction. Each voxel is labeled as being either *above* or *under* the *scan surface*. The scan surface, S_{scan}, is the initial surface that can be extracted as the interface between the *above* and *under* voxel sets. However, S_{scan} is noisy (in the sense that hair is present in the scan) and does not represent the true underlying skin surface. We aim to produce a consistent and locally smooth skin surface, S_{skin}, from the data set. In order to re-label the voxel set and thereby implicitly producing, S_{skin}, a Markov Random Field (MRF) classification/regularisation approach is chosen. In the following, a short introduction to MRFs and a description of the chosen models are given.

6.3.1 The volume random field

We define a random field with spatial voxel positions $\{v_1, v_2, \ldots, v_n\}$ in the volume V with the index set I. In this set each voxel v_i takes a value x_i from the binary label set L=$\{$*under*,*above*$\}$, where *under* is under and *above* is over the skin surface. Notice that we make a distinction between scan surfaces S_{scan} and skin surfaces S_{skin}; the classification relates to the latter. All values of x_i are represented by the vector $\mathbf{x}$, which is the configuration of the random field.

A neighbourhood system to v_i is defined as $N = \{N_i | i \in I\}$ for which it holds that $i \notin N_i$ and $i \in N_j \Leftrightarrow j \in N_i$. A random field is said to be a Markov field, if the probability P of any configuration of $\mathbf{x}$ satisfies the positivity property:

$$P(\mathbf{x}) > 0 \qquad \forall \mathbf{x} \in L \tag{6.1}$$

And the Markovian property:

$$P(x_i | \{x_j : j \in I \backslash \{i\}\}) = P(x_i | \{x_j : j \in N_i\}) \tag{6.2}$$

Or in other words the probability of x_i given the index set $I\backslash\{i\}$ is the same as the probability given the neighbourhood of i. Our use of neighbourhood is limited to the direct 6-neighbours in the volume. The goal is to compute the configuration of the field that maximizes the probability.

6.3.2 Defining the Markov Random Field

We aim to produce a labelling $\mathbf{x}$ of the voxel volume V, such that the boundary of the labelling coincides with the skin surface. A likelihood term and two priors are defined on the following:

- Under scan surface is likely to be under skin surface (*under*), while above scan surface is likely to be above skin surface (*above*).
- Skin surface points are in the vicinity of scan points.
- Skin surface is locally smooth.

Here *surface points* and *scan points* are defined as the local interface between *above* and *under* labelled voxels. Using these priors a MRF is created for which a minimal energy problem is defined using the following:

Likelihood term: This term is based on the relative position of a voxel and the scan surface:

$$\Phi(v_i|x_i) = -\log P(v_i|x_i) \tag{6.3}$$

Generally, we expect the scan and skin surfaces to coincide and therefore make a simplification of the likelihood function, such that voxels below the scan surface S_{scan} have low energy if labelled *under* and high energy if labelled *above* and vice versa for voxels above the scan surface. The energy then becomes:

$$\begin{aligned} \Phi(v_i < S_{\text{scan}}|x_i) &= \begin{cases} 1 & x_i = under \\ 0 & x_i = above \end{cases} \\ \Phi(v_i > S_{\text{scan}}|x_i) &= \begin{cases} 0 & x_i = under \\ 1 & x_i = above \end{cases} \end{aligned} \tag{6.4}$$

This somewhat loosely defined term alone would just produce the scan surface.

Vicinity prior: The skin surface should be close to scan surface points. This is induced by adding an energy penalty to changes in label, which relates to the distance from the voxel v_i to the nearest scan point S_{scan}:

$$\lambda(x_i, x_j) = \begin{cases} K_{\text{dist}} \cdot \text{dist}(v_i, S_{\text{scan}}) & x_i \neq x_j \\ 0 & x_i = x_j \end{cases} \tag{6.5}$$

The distance is approximated using an Euclidean distance transform (EDT) [49]. This is a fast linear time algorithm that approximates distance in a number of sweeps. This penalty should only affect voxels that are not in the immediate neighbourhood of scan surface points, which is why 1 is subtracted from the distances, such that both an actual surface point and its direct neighbours have distance 0. The vicinity constraint mainly effects areas with high discontinuity but it also forces the resulting surface to be true to the data in areas with continuity.

Smoothness prior: Neighbouring voxels are expected to have the same label with higher probability than having different labels, therefore an energy penalty is given to adjacent voxels with different label:

$$\psi(x_i, x_j) = \begin{cases} K_{ij} & x_i \neq x_j \\ 0 & x_i = x_j \end{cases} \tag{6.6}$$

This is a general smoothness constraint.

Combining the likelihood term with the vicinity and smoothness prior, we get the following energy minimization problem for the scan volume:

$$E(\mathbf{x}) = \sum_{i \in I} \left(\Phi(v_i|x_i) + \sum_{j \in N_i} (\lambda(x_i, x_j) + \psi(x_i, x_j)) \right) \tag{6.7}$$

Where $\mathbf{x}$ is the classification of the whole volume. The vicinity constant K_{dist} and the smoothness constant K_{ij} relates to each other and the likelihood term, which we defined to be 0 or 1. The solution to the *MRF* ensures maximum probability with the constraint that the labelling is both highly consistent with the data and smooth. To solve the minimization problem the *Graph Cut* algorithm [57] is used. This algorithm is an efficient way to find the optimal solution for such binary problems. From the resulting MRF classification a new surface S_{MRF} is extracted as the interface between the *under* and *above* labelled voxels.

6.4 Filtering based on the MRF solution

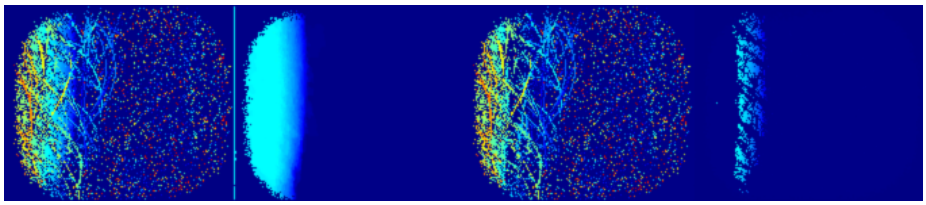
With a solution to the MRF and a new surface estimate S_{MRF} the changes compared to the original surface S_{scan} can be analyzed. As the MRF is set up so create a surface that coincides with the skin surface S_{skin}, we set up a filtering based on the following:

$$S_{\text{skin}} = S_{\text{scan}} \cap S_{\text{MRF}} \quad \wedge \quad S_{\text{hair}} = S_{\text{scan}} \setminus S_{\text{MRF}} \tag{6.8}$$

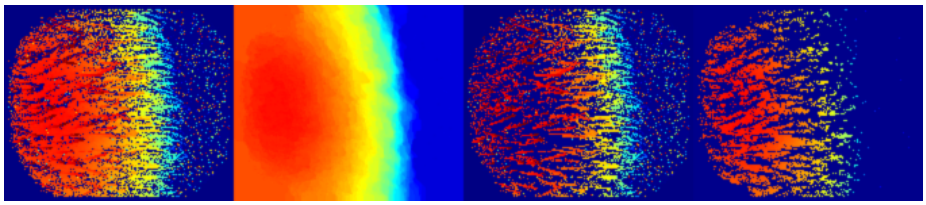
This results in a pure skin surface estimate and an estimate of the structured noise (hair) removed from the original scan to create the skin surface. We have used the strict definition but to give some flexibility one might add a threshold on how much is considered a change when comparing the original data with the MRF surface. Figure 6.2 shows examples of the resulting depthmaps of the *MRF* based filtering. Even though it is difficult to quantify the result, clearly both hair and noise are removed leaving only the smooth skin surface data.

Setting the relations between the likelihood term, the vicinity and smoothness prior is not trivial and based on *trial-and-error*. However, a reasonable approach is (at least for this type of data) to set the smoothness prior K_{ij} to 0, while incrementing the vicinity prior K_{dist} until a good result is achieved (Fig. 6.3(a)). This parameter effects areas with discontinuity such as skin to hair, while it actually forces the algorithm to be true to the input data in areas with continuity such as skin to skin. Setting the parameter removes most of the hair strands leaving only a little stubble, which can then be removed by adding the smoothness constraint. Figure 6.3(b) shows the difference between surface data found using only the vicinity prior and using both vicinity and smoothness, the difference is seen as green, and the blue is the filtered surface data.

Even though the *MRF* solution actually creates a surface without holes, this surface tends to have bumps where the surface is closed below the hair strands.

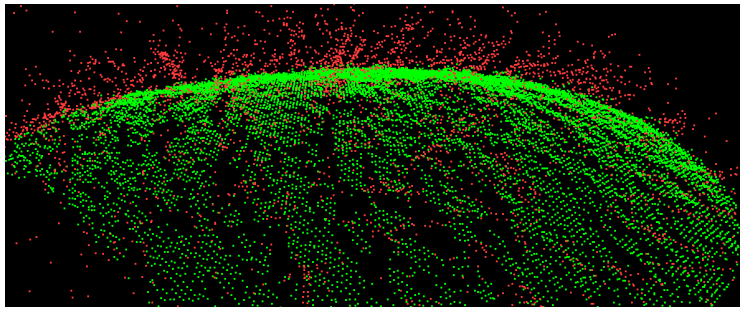


(a) Very little skin with hair

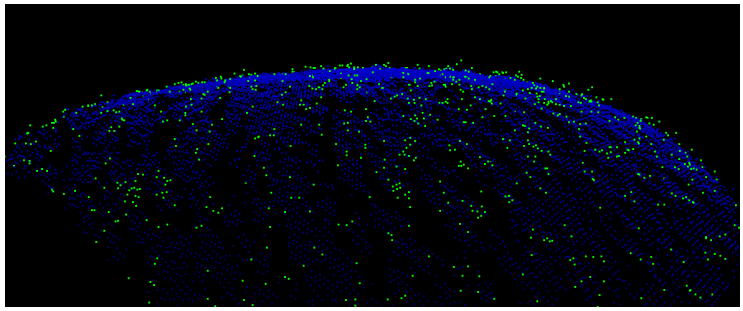


(b) Half skin coverage with hair

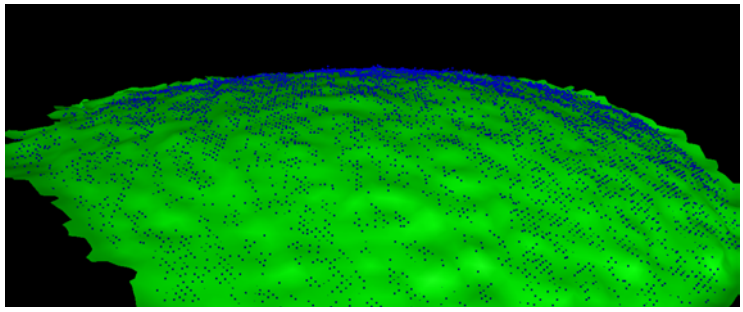
Figure 6.2: From left to right: depth map of original scan, surface based on volume classification, changes made in new surface, unchanged data in new surface. It is especially worth noting that the first scan is of very poor quality. The scanner is only focusing on very little skin surface and on top of that there is a lot of hair. In spite of this the algorithm returns the small amount of actual skin surface present in the scan.



(a) Filtering done using only the vicinity prior



(b) Filtering after adding the smoothness prior



(c) Filtered data and surface reconstruction

Figure 6.3: The figure shows a side view of a hairy surface. In the first image the effect of adding the vicinity prior alone is shown. Hair strands are coloured red and the surface green. The second image shows the difference between the surface using only vicinity prior and the surface, where both vicinity and smoothness prior has been applied. The difference is shown in green, while the resulting surface points of both vicinity and smoothness prior filtering are shown in blue. The third image shows a surface reconstruction using the filtered point set.

This is why only the unchanged part of the MRF surface should be kept. These remaining data are a much better starting point for a smooth surface reconstruction, as these data belong to the actual surface. Figure 6.3(c) shows a surface reconstruction based on the filtered point set. The surface has been reconstructed with the Markov Random Field Surface Reconstruction [74]. This algorithm uses a Markov Random Field to regularize a point distance field and it creates plausible hole filling, where data is missing. If computation time is a factor a simple 2D Delaunay triangulation [89] (considering the data as a 2D height map) would produce a reasonably fair surface.

6.5 Conclusion

In this paper data from a novel surface scanner has been used in an approach to remove structured noise from raw scanner data. The approach is based on relabelling a voxel set using a Markov Random Field classification and extracting the sought surface as the interface between two voxel labels.

We have achieved good results on filtering noise and hair from our surface scans. Even though it is hard to quantify the ability to filter both noise and hair, qualitative visual inspections of the results are very promising. Because of the high quality expectance of the scans it is a strong point that the data is filtered such that the remaining data is unchanged.

Our algorithm has been implemented in Matlab with crucial parts in C++ `.mex`-files and it filters in a matter of a few seconds on a 2.8 GHz Intel processor laptop. There is good reason to believe that the surface filtering could be done in realtime with a full implementation in a precompiled programming language.

One could argue that a full volume classification as part of the filtering is not necessary. This is meant in the sense that large parts of the scan volume will be *far* from the actual surface and therefore a smarter way of classification close to the surface could speed up the process. This would however complicate the approach and is a topic for future research.

Chapter 7

Genus Zero Graph Segmentation: Estimation of Intracranial Volume

Rasmus R. Jensen, Signe S. Thorup, Rasmus R. Paulsen, Tron A. Darvann, Nuno V. Hermann, Per Larsen, Sven Kreiborg, and Rasmus Larsen

Abstract

The intracranial volume (ICV) in children with premature fusion of one or more sutures in the calvaria is of interest due to the risk of increased intracranial pressure. Challenges for automatic estimation of ICV include holes in the skull e.g. the foramen magnum and fontanelles. In this paper, we present a fully automatic 3D graph-based method for segmentation of the ICV in non-contrast CT scans. We reformulate the ICV segmentation problem as an optimal genus 0 segmentation problem in a volumetric graph. The graph is the result of a volumetric spherical subsample from the data connected using Delaunay tetrahedralisation. A Markov Random Field is constructed on the graph with probabilities learned from an Expectation Maximisation algorithm matching a Mixture of Gaussians to the data. Results are compared to manual segmentations performed by an expert. We have achieved very high Dice scores ranging from 98.14% to 99.00%, while volume deviation from the manual segmentation ranges from 0.7%-3.7%. The proposed method is expected to perform well for other volumetric segmentations.

Keywords: *I*ntracranial volume, CT, craniosynostosis, graph cut, segmentation

7.1 Introduction

Unicoronal synostosis (UCS) is a congenital craniofacial malformation characterized by the premature fusion of one of the coronal sutures, potentially leading to asymmetric head shape, craniofacial growth disturbances, increased intracranial pressure and developmental delays. Computed Tomography (CT) scanning is usually performed to confirm the diagnosis and to facilitate surgical treatment planning. The intracranial volume (ICV) in children with premature fusion of one or more sutures in the calvaria may become reduced, leading to risk of increased intracranial pressure [106]. Challenges for automatic estimation of ICV include holes in the skull in newborns (the fontanelles), but also holes in the cranial base (e.g. the foramen magnum and other foramina, fissures and synchondroses). The main contribution of our work is a fast and fully automatic method for segmentation and estimation of the ICV in CT scans of children with craniosynostosis. The method is based on the construction of a volumetric graph description of the skull volume using tetrahedralization followed by a graph cut forced to robustly perform a genus 0 segmentation. Validation is carried out by comparing the automatic segmentation model to a semi-automated model.

7.2 Brief Review of the Previous Research

Current work on automatic ICV[1] estimation has focused on Magnetic Resonance Imaging (MRI) volumes [67, 76, 91]. However, these methods are not well suited for ICV estimation in craniosynostotic cases due to the limited bone-tissue contrast in MRI. In the case of craniosynostosis, the best contrast of the cranial bones, e.g. for diagnosis and surgery planning, is obtained from CT scans. Furthermore, standard methods often use atlases based on a normal population, which may lead to a bias in the estimation of the ICV in craniosynostotic cases. The current standard for ICV estimation from CT is a manual method based on thresholding followed by a seed-growing algorithm. The challenge of this method is the need for manual editing in the various foramina in the skull base as well as in regions where craniosynostosis or lacking suture fusion have caused gaps between the cranial bones [88, 5, 6].

Anatomical segmentation such as the segmentation of the ICV in medical images is addressed in the literature by a series of approaches. In [61], deformable template matching is applied in a Bayesian setting; in [62], deformable surface models are proposed using a graph cut approach; and in [109], a multiclass Markov Random Field (MRF) is used for voxel classification. In the latter case

[1] In MRI they often estimate the total intracranial volume (TIV).

it is interesting that, for two-class models, global optimal segmentation can be obtained using a graph-cut-based approach [57]. In this work we propose a two class segmentation of the ICV, where the classes (inside and outside) are modeled as mixtures of Gaussians. In addition to a label prior, we use a gradient-dependent interaction term. Moreover, we employ a tetrahedralization of a spherical equidistant sample distribution leading to a graph. The graph has dedicated outside and inside nodes, which robustly forces the graph segmentation to be of genus 0.

7.3 Approach

The data consist of pre-surgical CT head scans of 15 children diagnosed with UCS (either left- or right-sided). Age ranged from 6 to 18 months. All scans were acquired at Copenhagen University Hospital, Rigshospitalet, except for one which was acquired at Helsinki University Central Hospital. Because of the UCS and the different age the data set is not homogeneous. All scans were obtained at 512 x 512 pixels in-plane size and a complete volume consists of between 167 and 350 slices.

The aim of the method is to create a volumetric segmentation that follows the transition between brain matter and bone, while also closing holes in the bone structure. As the intensities of the CT scans vary, we fit a mixture of Gaussians to each individual scan. The mixture of Gaussians is carried out using expectation maximization and results in three normal distributions describing: skin, brain matter and bone (see Fig. 7.1). Skin and bone have higher variance compared to brain tissue, which is used to classify the distributions unsupervised. Brain matter is by far the dominant, but also that with the least variance. Using the probability density function, where v is a sample value, we define the following two probabilities: $\mathrm{p}(v|x = \mathrm{ICV}) = \mathrm{pdf}_{brain}$ and $\mathrm{p}(v|x \neq \mathrm{ICV}) = \mathrm{pdf}_{skin} + \mathrm{pdf}_{bone}$. Generally, the brain matter distribution fits well to the data, while the other two tissues just stay below and above both with a wider standard deviation. Using only the mixture of Gaussians to classify brain-tissue and non-brain tissue would lead to misclassification as the distributions are crude, while the proposed method is insensitive to this.

Before the segmentation, the volumes were interpolated in the slice-wise direction to create isotropic voxels and ensure a regular sampling. A graph is created on sampling points in the volumes. The sampling points are found using a spherical volume of quasi-equidistantly distributed nodes. The nodes are distributed on the surfaces of concentric spheres, where the differences between their radii are equal to the spacing between their longitudes. Similarly, points

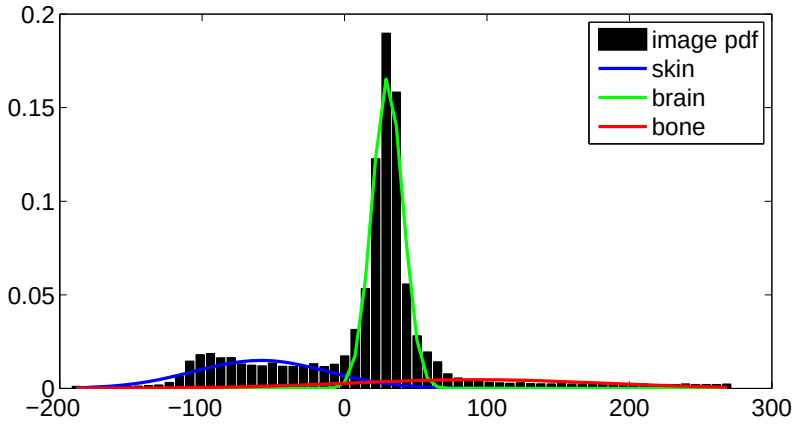


Figure 7.1: A mixture of Gaussians shown on the intensity histogram of a CT scan.

on the longitudes are distributed with this spacing. This suggested sampling approach has two benefits: the sampling density can easily be changed for different resolutions, and it also removes the over/under-sampling problem of a spoke-like graph directed from the center and out. The sample volume is centered in the middle of the calvaria and created such that it covers the entire skull. In the center of the graph, we leave a small empty sphere, which will be used to clamp the inside of the graph to the ICV (see Fig. 7.2). The spherical graph is centered automatically by summing the voxel-wise probability of brain matter given the distribution prior. We find the coordinates of voxels with higher probability than $\frac{2}{3}$ of the maximum summation of the sagittal, coronal and transverse planes, respectively. The center is found as the median of these coordinates, while the radius of the sphere is found as 2.5 times the maximum interquartile range in the sagittal and coronal planes only. Fig. 7.2 illustrates the sample point distribution in the transverse and sagittal planes. For the actual ICV estimation we have used a much higher sample density, using an even voxel distance of two between sample points. A robust way of connecting each of the sample points to the immediate spatial neighborhood in a highly connected graph can be achieved by Delaunay tetrahedralization [90]. As this approach produces doublets of edges represented by several adjacent tetrahedra the connectivity has to be cleaned up such that edges are represented only once. On the graph with index set $\mathcal{I}$, we define the following MRF, which is solved

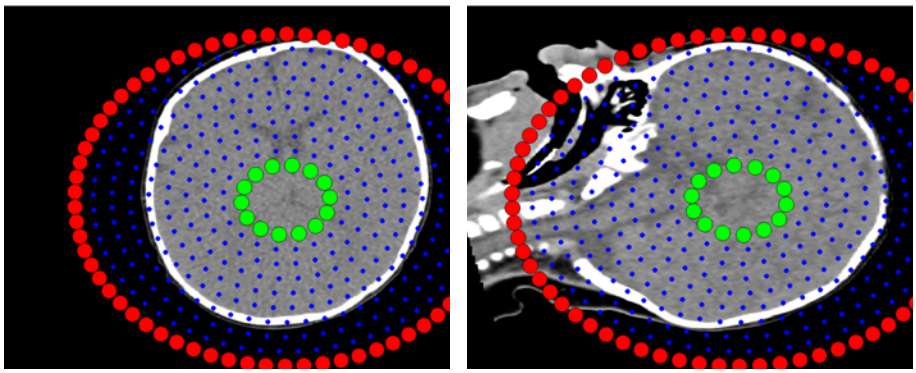


Figure 7.2: Examples of distribution of sample points in the transverse and sagittal planes. The sample density is much higher in the actual application with a voxel distance of two between sample points. The green nodes show the inner sphere which is forced to be part of the ICV, while the red nodes shows the outer sphere which is forced to be outside the ICV. The slices are contrast enhanced based on the mixtures of Gaussians.

using graph cuts [57]:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{I}} \left(\Phi(v_i|x_i) + \sum_{j \in N_i} (\lambda(x_i, x_j) + \psi(x_i, x_j)) \right) \tag{7.1}$$

$\Phi(v_i|x_i) = -\log \mathrm{p}(v_i|x_i)$ defines a log-likelihood function. The function returns high values for low probabilities and vice versa. The outer and inner sphere log-likelihood values are clamped as follows:

$$\begin{aligned} \Phi(i \in \text{outer}|x_i) &= \begin{cases} \infty & x_i \neq \text{ICV} \\ 0 & x_i = \text{ICV} \end{cases} \\ \Phi(i \in \text{inner}|x_i) &= \begin{cases} 0 & x_i \neq \text{ICV} \\ \infty & x_i = \text{ICV} \end{cases} \end{aligned} \tag{7.2}$$

N_i denotes the neighborhood of the i'th voxel and has terms defined as:

$$\lambda(x_i, x_j) + \psi(x_i, x_j) = \begin{cases} K_\nabla e^{-|\nabla \mathrm{f}(x_i, x_j)|} + K_{ij} & x_i \neq x_j \\ 0 & x_i = x_j \end{cases} \tag{7.3}$$

Where $|\nabla \mathrm{f}(x_i, x_j)|$ is the absolute gradient; K_∇ controls the power of the gradient term, while K_{ij} is a general smoothness prior. With the outside sphere clamped as *outside ICV* and the inside sphere clamped as *ICV*, a setting of K_∇ and K_{ij} exists for which the resulting graph cut will only be on the inner edge

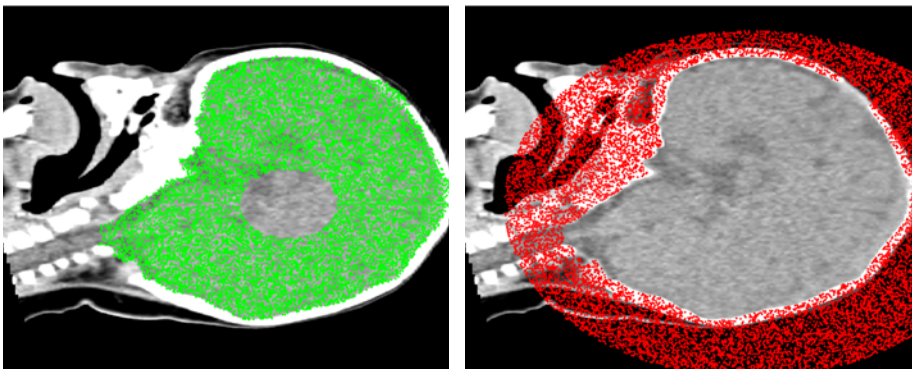


Figure 7.3: The result of the segmentation between *ICV* and *outside ICV* shown in the sagittal plane. The slices are contrast enhanced based on the mixtures of Gaussians.

of the skull, close the holes (e.g. fontanelles, optic canals, foramen magnum, and other foramina), and it will produce a segmentation of genus 0. We achieved our results with K_∇ being of the same magnitude as $\max \Phi(v_i|x_i = \text{ICV}), i \in I$ and K_{ij} $\frac{1}{100}$ of that. We found the results to be rather insensitive to fine tuning of K_∇ and K_{ij}. The result of a segmentation is shown in Fig. 7.3. With a segmentation of the graph, the volume can be estimated using the tetrahedralization of the sample points contained in the cut.

7.4 Results and Discussion

Ground truth was made by expert manual segmentations using a semi-automatic, slice-wise method based on a seed-growing algorithm incorporated in Analyze$^{\text{TM}}$ (BIR Research Lab, Mayo Clinic, Rochester, MN, USA). This method requires a user-specified intensity threshold and manual editing.

As gaps and small fractures are present in the data, the semi-automatic segmentation algorithm often breaks down and manual editing is necessary. Easy cases for manual editing are when the natural curvature of the skull is present and the gaps are small. Unfortunately, severe cases with large gaps and no a priori information potentially lead to large errors. Average processing time for the manual method is two hours, including threshold estimation. Average runtime including all steps of the process is 12 minutes on a fast consumer desktop computer (Intel i7 3.6@4.2 GHz processor with 16 GB ram) running Matlab. Figure 7.4

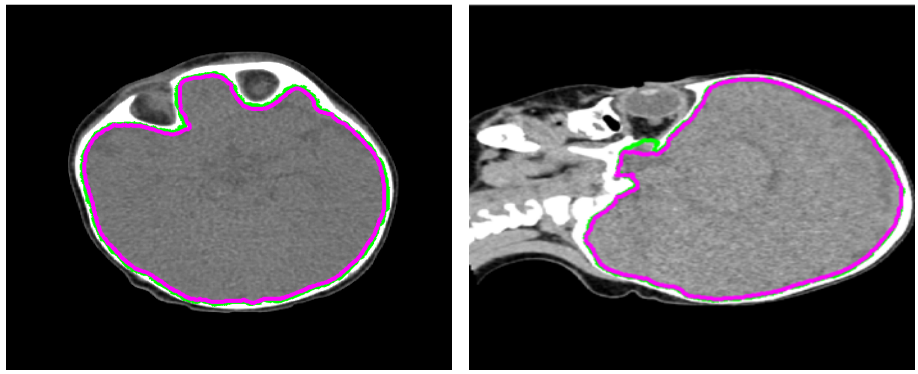


Figure 7.4: Examples of the manual (magenta) and automatic segmentation (green) shown on a transversal (Patient #8) and sagittal slice (Patient #11), respectively. The slices are contrast enhanced based on the mixtures of Gaussians.

illustrates the two methods on example slices. Looking at the transversal slice, the methods are consistent, while on the sagittal slice the automatic method includes part of the optic canal. Fig. 7.5 shows the automatically estimated volumes as a function of the manual volumes. The linear regression of the two lines has been forced through origo. For the volume measured entirely inside the cut, we get R^2-value of 99.54%, with a slope of $\alpha = 0.9768$ being a 2.32% underestimate. Including the tetrahedra partially inside the cut, we get and R^2-value of 99.50%, with a slope of $\alpha = 1.0246$ being a 2.46% overestimate. We have used the *worst* of the volume estimates (i.e. including the partially cut tetrahedra) to calculate volume deviation, Dice score [27] and Hausdorff distance [46].

Table 7.1 shows the comparison between the two segmentation models. For each patient the deviation from the manual volume, Dice coefficient and Hausdorff distance were calculated to evaluate the proposed method. While the Dice coefficient measures the volume overlap, the Hausdorff distance measures the maximum error between the two segmentations. Since the Hausdorff distance measure is sensitive to single error the 95 % Hausdorff distance is also included.

Deviations in volume are very small and lie between 0% and 3.7%. The Dice coefficients also show a high consistency in overlap between the methods. The lowest Dice score is 98.14, while the highest is 99.00. An explanation for the differences might relate to the graph cut lying slightly outside the manual border (see Fig. 7.4). This behavior might relate to the chosen threshold, and it would be interesting to assess the consequence of using various thresholds as well as the manual error in future work.

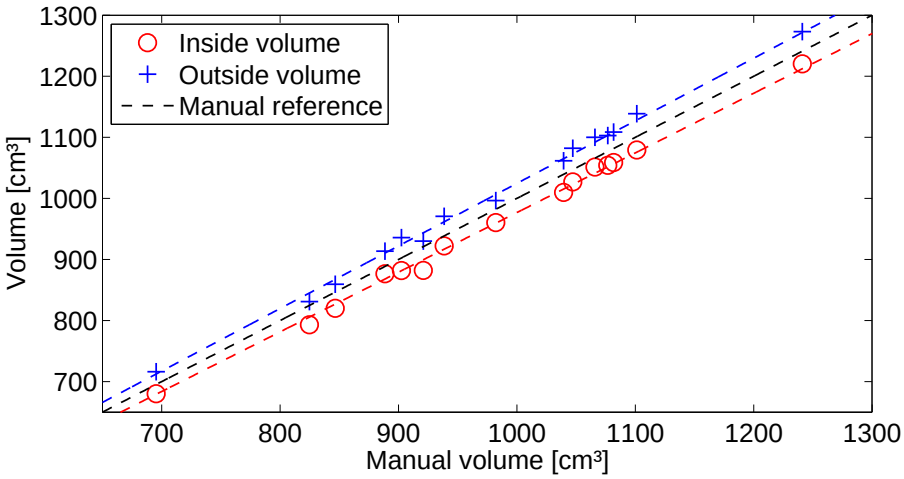


Figure 7.5: Automatic volumes as a function of the corresponding manual volume including linear regression lines. Red denotes the volume with tetrahedra fully inside the segmentation, while blue also includes tetrahedra partially inside. As a reference the manual volume is also shown.

The Hausdorff distance, which is a conservative measure, showing the maximum error in overlaps, shows differences of up to 9.81 mm. Examining the images, large differences between the two methods occur where the foramen magnum and optic canals are closed. Both regions are hard to segment consistently, since the spinal cord and the optic nerves are similar in intensity to that of brain matter.

Patient #	Man. volume $[cm^3]$	Auto. volume $[cm^3]$	Volume deviation [%]	Dice score [%]	Hausdorff distance $[mm]$	95% Haus. distance $[mm]$
1	846.6	859.5	1.5	98.87	7.62	2.22
2	824.8	831.0	0.7	98.88	5.09	2.16
3	888.6	913.5	2.8	98.55	4.94	1.16
4	1076.9	1102.9	2.4	98.67	5.30	1.10
5	1039.8	1061.7	2.1	98.71	6.88	1.72
6	1241.1	1273.2	2.6	98.66	4.74	1.10
7	982.2	996.5	1.5	99.00	9.81	3.65
8	902.6	935.8	3.7	98.14	4.88	1.38
9	1081.4	1108.8	2.5	98.67	4.73	1.12
10	1066.1	1100.0	3.2	98.36	7.97	1.75
11	1101.2	1138.8	3.4	98.28	6.05	1.34
12	921.1	930.2	1.0	98.87	8.11	2.98
13	695.2	716.2	3.0	98.39	6.09	1.89
14	1047.5	1082.1	3.3	98.28	6.48	1.27
15	938.6	970.8	3.4	98.29	5.16	1.22

Table 7.1: Result overview comparing the manual segmentation to the proposed automatic method. *Volume deviation* denotes the relative percentage-wise difference between the numeric volume estimates, *Dice score* measures the overlap between the volumes and *Hausdorff distance* assesses the maximum error. The *95 % Hausdorff distance* is shown in the last column, this measure is much less sensitive to single error.

7.5 Concluding Remarks

In conclusion, we have implemented an automatic, fast method for accurate estimation of the ICV in children with UCS. The method is fairly insensitive to fine tunning of parameters. There is good reason to believe that the proposed method can be used for other applications in volumetric image segmentation.

Chapter 8

Statistical Surface Recovery: A Study on Ear Canals

Rasmus R. Jensen, Oline V. Olesen, Rasmus R. Paulsen, Mike van der Poel, and Rasmus Larsen

Abstract

We present a method for surface recovery in partial surface scans based on a statistical model. The framework is based on multivariate point prediction, where the distribution of the points are learned from an annotated data set. The training set consist of surfaces with dense correspondence that are Procrustes aligned. The average shape and point covariances can be estimated from this set. It is shown how missing data in a new given shape can be predicted using the learned statistics. The method is evaluated on a data set of 29 scans of ear canal impressions. By using a leave-one-out approach we reconstruct every scan and compute the point-wise prediction error. The evaluation is done for every point on the surface and for varying hole sizes. Compared to state-of-the art surface reconstruction algorithm, the presented methods gives very good prediction results.

Keywords: *surface recovery, hole closing, multivariate statistics, shape modeling, in ear scanning*

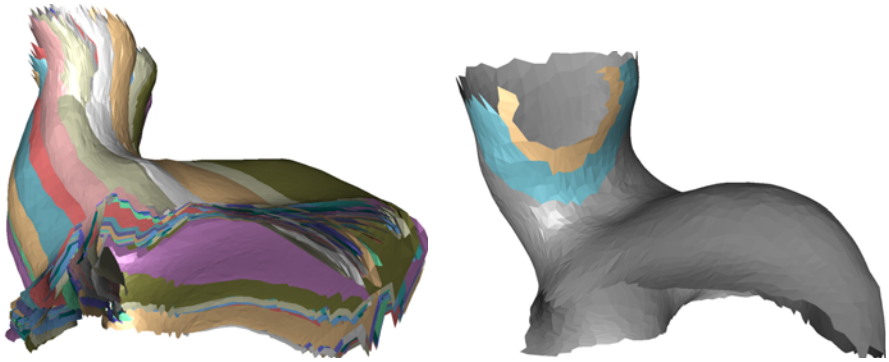
8.1 Introduction

Direct surface scanning of humans is an increasingly used modality where the applications range from model creation in the entertainment industry, plastic

surgery planning and evaluation, craniofacial syndrome evaluation [60, 39], and in particular hearing aid production [72]. In this paper, we are concerned with a particular surface shape namely that of the ear canal. Ear canal surface scans are used in custom hearing aid fitting. This is a very large industry that probably makes the ear the most scanned part of the human anatomy. A standard hearing aid producer generates more than a thousand scans per week. When producing custom in-the-ear devices like hearing aids and monitors, the standard routine is to inject silicone rubber in the patients ear and then laser scan this impression, thereby creating a model of the ear canal. While this technique normally creates complete surfaces, direct ear scanners are emerging and it is expected that scans with these devices will require handling of missing data due to the complex anatomy of the human ear and the limited space for the scanner probe.

In this paper we are presenting a method for predicting missing data based on the information in the partial scan. Hole filling and missing data recovery is a well studied problem, in particular for 2D images. In 3D, data recovery is sometimes considered a by-product of the surface reconstruction algorithm. The algorithms used to generate triangulated surfaces from a point clouds will usually try to cover missing areas using some mathematical or physical assumptions. One series of approaches uses Delaunay triangulation of border points [56]. Such methods are obviously susceptible to noise in the border points and will typically require some form of smoothing. An alternative strategy is to interpolate implicit (signed distance) functions locally or globally under various forms of regularization [51, 74]. Other methods, inspired from 2D inpainting approaches have also been investigated [101, 21, 18]. These are typically based on a variational definition of the behavior of the surface where the holes are. In [18] it is the mean curvature of the surface that is being regularized and in [21] it is the Willmore energy over the surface. The reported results of these techniques seem similar to the results from the Markov Random Field surface reconstruction algorithm [74] that we are using as reference.

In our method, we aim at predicting the missing points based on the existing points in the scan. Instead of using a variational formulations or physical assumptions on the behaviour of the surface, we utilize a population statistics of the given class of surfaces learnt from an annotated and co-registered training set. In the chosen example, we base our population statistics of the ear canal on a statistical shape model of the ear canal originally presented in [73]. However, the method is general and can be applied to all types of surface scans.



(a) Aligned scans (b) Shape with holes

Figure 8.1: a) shows the Procrustes aligned scans. Each scan has a different color. b) shows a surface scan with three holes with identical center but different radii, $r = [2\text{mm}, 3\text{mm}, 5\text{mm}]$.

8.2 Data

The data consists of 29 scanned ear impressions. For further processing, point correspondence over the set was created using the method initially described in [73]. Here a template mesh is thin plate spline warped to fit each shape using a sparse set of landmarks. This is followed by a propagation of template mesh vertices to each shape based on a nearest neighbor search. Since the ears are topologically equivalent to open cylinders special care must be taken in the non-overlap regions and the template is pruned accordingly. Furthermore, the Markov Random Field regularization of the correspondence field described in [71] was used to further optimize the dense correspondence. The 29 scans with dense correspondence was then aligned using the Generalized Procrustes Alignment [93] and the shapes scaled to the scale of the mean shape. The result of this is a set of aligned shapes (triangulated surfaces), with dense point correspondence (3000 vertices per scan) as seen in Fig. 8.1(a). This type of data is normally used to build statistical shape models as for example described in [22, 73].

In Fig. 8.1(b) an example of a partial scan is seen. It is synthetically created by cutting a hole in a complete ear canal surface scan. In the example, holes with radii of 2, 3, and 5 mm are shown.

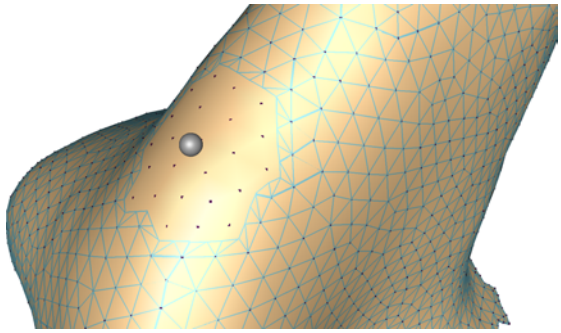


Figure 8.2: A hole in a mesh. To predict the unknown vertex shown $\mathbf{x}_1$ as a gray sphere, the known vertices $\mathbf{x}_3$ shown vertices on the blue mesh are used. The remaining unknown vertices $\mathbf{x}_2$ are shown as black dots in the hole.

8.3 Statistical Surface Recovery

The aim of the method is to predict the unknown data from the known data. The unknown data is in this case the vertices placed in the missing parts of the scanned surface. The known data is the vertices in the partial scan. Since the method predicts one vertex at a time, we can define three sets of vertices:

$$\begin{aligned}
\mathbf{x}_1^{\mathrm{T}} &= (u_{11}, v_{11}, w_{11}) : \text{the unknown vertex we want to predict.} \\
\mathbf{x}_2^{\mathrm{T}} &= (u_{21}, v_{21}, w_{21}, u_{22}, v_{22}, w_{22}, \ldots) : \text{remaining unknown vertices} \\
\mathbf{x}_3^{\mathrm{T}} &= (u_{31}, v_{31}, w_{31}, u_{32}, v_{32}, w_{32}, \ldots) : \text{known vertices.}
\end{aligned}$$

In Fig. 8.2 the situation is exemplified, with $\mathbf{x}_1$ shown as a gray sphere and the known vertices, $\mathbf{x}_3$, shown as vertices in the blue mesh. The remaining unknown vertices, $\mathbf{x}_2$, are obviously not used to predict $\mathbf{x}_1$. However, any point in $\mathbf{x}_2$ can be estimated by setting it to be $\mathbf{x}_1$. In the following we assume that a point correspondence has been established between the partial scan and the training set described in Sec. 8.2. The correspondence allows for differentiation between known vertices and missing vertices in the partial scan. We will elaborate on this later.

In conclusion, we will determine how the unknown vertex $\mathbf{x}_1$ is predicted from known vertices in $\mathbf{x}_3$. Without any prior knowledge of the distribution of data, we consider $\mathbf{x}_1$ and $\mathbf{x}_3$ as belonging to the normal distribution:

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_3 \end{bmatrix} \in N\left(\begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{13} \\ \mathbf{\Sigma}_{31} & \mathbf{\Sigma}_{33} \end{bmatrix} \right), \ \mathbf{\Sigma}_{13}^{\mathrm{T}} = \mathbf{\Sigma}_{31}$$

The expected value of $\mathbf{x}_1$ given $\mathbf{x}_3$ is:

$$\mathrm{E}\{\mathbf{x}_1|\mathbf{x}_3\} = \mu_1 + \mathbf{\Sigma}_{13}\mathbf{\Sigma}_{33}^{-1}(\mathbf{x}_3 - \mu_3)$$

With the variance:

$$\mathrm{V}\{\mathbf{x}_1|\mathbf{x}_3\} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{13}\mathbf{\Sigma}_{33}^{-1}\mathbf{\Sigma}_{31}$$

The vertex distributions are determined using the Procrustes aligned training shapes with point correspondence. So for example $\mathbf{\Sigma}_{11}$ is the covariance of the vertex $\mathbf{x}_1$ as learned from the training set. The covariance $\mathbf{\Sigma}_{13}$ between $\mathbf{x}_1$ and $\mathbf{x}_2$ is also estimated from the training set along with (μ_1, μ_3) that are the average vertex positions from the Procrustes average estimation. As there are far less shapes (28 leaving one out) than points (3000), $\mathbf{\Sigma}_{33}$ will be singular. Let $\mathbf{\Sigma}_{33} = \mathbf{P}\Lambda\mathbf{P}^{\mathrm{T}}$ be the Eigenvalue decomposition. We restrict $\mathbf{\Sigma}_{33}$ to its affine support, i.e. the dimensions spanned by the Eigenvectors corresponding to the k positive Eigenvalues, such that:

$$\mathbf{\Lambda}^* = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k) \text{ and } \mathbf{P}^* = [\mathbf{p}_1\ \mathbf{p}_2\ \ldots\ \mathbf{p}_k]$$

The projection of $\mathbf{x}_3$ using the k selected Eigenvectors $\mathbf{P}^*$: $\mathbf{y}_3 = \mathbf{P}^{*\mathrm{T}}\mathbf{x}_3$ has affine support for $\mathbf{x}_3$ and the variance:

$$\mathrm{V}\{\mathbf{y}_3\} = \mathrm{V}\{\mathbf{P}^*\mathbf{x}_3\} = \mathbf{P}^{*\mathrm{T}}\mathbf{\Sigma}_{33}\mathbf{P}^* = \mathbf{\Lambda}^*$$

The covariance of $\mathbf{x}_1$ and $\mathbf{y}_3$ is:

$$\mathrm{C}\{\mathbf{x}_1, \mathbf{y}_3\} = \mathrm{C}\{\mathbf{x}_1, \mathbf{P}^{*\mathrm{T}}\mathbf{x}_3\} = \mathrm{C}\{\mathbf{x}_1, \mathbf{x}_3\}\mathbf{P}^* = \mathbf{\Sigma}_{13}\mathbf{P}^*$$

Finally, the prediction of the unknown vertex $\mathbf{x}_1$ can be done using the projection $\mathbf{y}_3$:

$$\mathrm{E}\{\mathbf{x}_1|\mathbf{y}_3\} = \mu_1 + \mathbf{\Sigma}_{13}\mathbf{P}^*\mathbf{\Lambda}^{*-1}\mathbf{P}^{*\mathrm{T}}(\mathbf{x}_3 - \mu_3)$$

This expectancy can be used for any unknown vertex $\mathbf{x}_1$ given a known set of vertices $\mathbf{x}_3$. While the Eigenvalue decomposition only has to be done once for the known vertices $\mathbf{x}_3$, the $\mathbf{\Sigma}_{13}$ has to be calculated for each vertex to be predicted.

If every unknown vertex is predicted according to the described method the known triangulation from the training set can be propagated to the predicted point and will then constitute a full surface reconstruction.

The described point prediction method requires point correspondence to the training set. This means that given a new surface scan of an ear canal with missing data, the first step is to create point correspondence. This is a non-trivial matter with a variety of existing approaches [50] and not the main topic of this paper. A usable approach could be to use a variant of the iterative closest point algorithm that handles non-overlapping regions [82].

8.4 A Standardized Test

To compare the performance of the statistical prediction to a state-of-the art surface reconstruction algorithm, we do the following standardized test. We define:

$\mathbf{x}_1$: the vertex in the center of a hole.
$\mathbf{x}_2$: vertices within radius r of $\mathbf{x}_1$ defines a hole.
$\mathbf{x}_3$: known vertices.

We predict every vertex in a shape with the vertex being the center of a hole. We repeat this experiment for holes with different radii. It seems intuitive that in the prediction of a given hole, the center is generally the hardest to accurately reconstruct as it is furthest from any known vertices. This is also seen in Fig. 8.2, where the predicted vertex is the unknown vertex furthest away from known data. Therefore, the prediction of the center points is considered a soft upper bound on how well a given hole can be closed. Figure 8.1(b) shows an ear scan where holes with the same center but varying radii have been cut out.

The MRFSurface reconstruction described in the following will be used for comparison against our proposed method.

8.5 Markov Random Field surface reconstruction

Markov Random Field surface reconstruction (MRFSurface) as described in [74] is based on an implicit reconstruction and regularization of the true surface represented by the available, noise filled, and potentially hole infested data. A signed distance field on a regular voxel grid is used to maintain the implicit surface as the zero-level iso-surface. Obviously, the distance field is not properly defined in regions with missing data or holes and the novelty in [74] is the Markov Random Field based regularization of the distance field leading to good hole filling and noise handling properties. In the Bayesian setting, a prior voxel energy is used that implicitly defines how the surface behaves in areas with missing data. While several models are possible, we focus on two. The first is based on penalizing nearest neighbour voxel value differences leading to membrane like behavior of the surface. The second is using the difference of the Laplacian of the neighbouring voxel values leading to a higher-order, spline like behavior of the surface. In order to maintain distance field values in high-confidence areas a quadratic observation term is used that punishes distance values different from the original estimates. The balance between the prior and the observation term is governed by a confidence factor based on local point cloud sampling densities.

The goal is now to optimize the values of the entire distance field and thereby creating an implicit zero-level that behaves well in both dense sampled areas and areas with holes. Since the problem can basically be cast into a massive set of linear equations several approaches can be used. In [74] sparse Cholesky factorization, conjugate gradients and multi-scale, banded iterated conditional modes (MS-ICM) were evaluated. In this work, we use the MS-ICM solver. The result is a well regulated zero-level iso-surface that is then polygonised using a standard iso-surface extractor [14]. Furthermore, an iterative triangle-optimiser is used in the post-processing to create near-equilateral triangles [74].

8.6 Experiments and Results

To test the performance of our method, we use the Procrustes aligned data set of 29 scans with point correspondence and use 28 to build a statistical model and reconstruct the last one. To simulate that the scan is a new and previously unseen scan we start by removing the point correspondence to the training set. This is done by re-triangulating the surface, such that it contains two thirds the number of vertices being 2000 relative to 3000 in the original mesh. The re-triangulation is done using the method described in [15] and results in a mesh with vertices evenly distributed and with near equilateral triangles.

As described in Sec. 8.3 point correspondence to the training set is needed. For the sake of testing the point prediction algorithm, we have adopted a rather simple approach. We find the scan of the 28 that produces the best point correspondence. A point x_{original} is said to have correspondence if the nearest point x_{remeshed} in the re-meshed surface also has the also has x_{original} as its nearest point. This produces a unique point correspondence. We use the correspondence from the scan that creates the minimum average distance between correspondence points. This approach produces around 1500 unique correspondences.

One note about the alignment and point correspondence: We use the original Procrustes alignment for our re-meshed scan, this alignment is fair but not optimal as it is an alignment to the mean shape. Our point correspondence is also very crude, though speaking of correspondence is somewhat vague as good landmarks are hard to accurately determine on a shape such as an ear.

Using our statistical approach we reconstruct every shape. The reconstruction is done vertex by vertex, each time the vertex defines the center of a hole. This is repeated for holes with radii $r = [2\text{mm}, 3\text{mm}, 5\text{mm}]$. Figure 8.3 shows two statistically reconstructed scans with reconstruction error as color. The measure of the error is the distance from the reconstructed hole center to the original

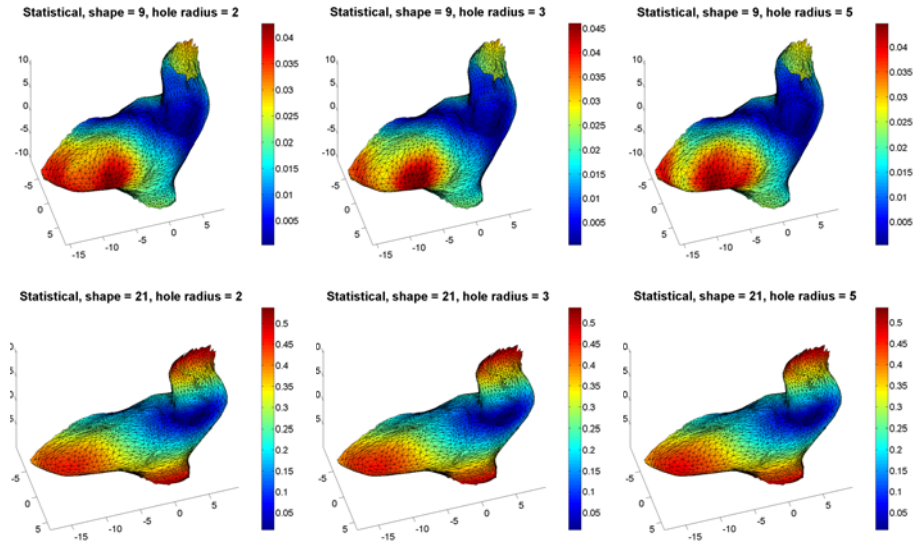


Figure 8.3: The reconstruction of two ears ranging from very good to reasonable, the reconstruction is done for holes with radius $r = [2\text{mm}, 3\text{mm}, 5\text{mm}]$. The reconstruction error is shown in color. Notice the difference in color scale for the to reconstructions.

hole center, this error constitutes a upper bound (meaning worst) for the reconstruction error, whereas a point to surface distance should produce the lower bound (most optimistic). It is worth noticing that we get good reconstructions even though the holes are big relative to the size of the scans. The reconstruction error is very similar for different hole sizes which leads to assume that there is a strong correlation between points across the shape of the ear canal. To compare the proposed reconstruction with a recent surface reconstruction algorithm, the experiments are repeated using MRFSurface to again reconstruct all points in all shapes for varying hole sizes. It has been used with two different hole closing priors: Laplacian and Membrane. While the reconstruction using these two priors is fair for small holes the reconstruction collapses as the holes get bigger. Figure 8.4 shows the standard deviation from zero of the reconstruction of the 29 scans using all three methods for holes with varying radius, $r = [2\text{mm}, 3\text{mm}, 5\text{mm}]$. As with the statistical method the upper bound error is used, again being the distance from the original hole center to the closest point in the reconstructed surface. Comparing the results of the three different reconstructions it can be seen that all reconstructions are good for hole sizes with $r = 2\text{mm}$, with the statistical reconstruction outperforming the other to. As the hole sizes increases to $r = 3\text{mm}$ the Membrane prior reconstruction begins to

Table 8.1: Total median error and interquartile range for the three reconstructions (in mm)

	$r = 2$mm		$r = 3$mm		$r = 5$mm	
Reconstruction	med	iqr	med	iqr	med	iqr
Statistical	0.1111	0.2003	0.1115	0.2004	0.1124	0.2011
Laplacian	0.3387	0.1573	0.3798	0.1713	0.5666	0.4947
Membrane	0.3745	0.1758	0.5515	0.4130	1.4683	1.3314

fail while the Laplacian prior reconstruction is still good. The statistical method stays almost unchanged. Finally the holes are increases to $r = 5$mm and the Laplacian prior reconstruction also begins to fail. The statistical reconstruction stays very constant (the three top figures are very close to but not identical). This again suggest a very strong correlation between the surface points in ear scans, even though the shapes of the ear impressions vary a lot, as can be seen in Fig. 8.1(a). Using a statistical model also allows for surface reconstruction of large areas near borders. The distribution of reconstruction of error is right skewed. A less skew biased measure of the reconstruction is the median and the interquartile range for the three reconstructions. Table 8.1 shows the median and interquartile range over all shapes and vertices. The table clearly shows that our proposed method performs better in the reconstruction. The interquartile range is also quite narrow making the overall reconstruction error consistently small.

It is well known that the ear canal deforms when people are chewing or doing other facial movements. This is partly due to the movement of the mandibular bone [26]. Furthermore, the impression taking and scanning is also influenced by several factors, including viscosity [78] of the impression material and the subsurface scattering encountered during laser scanning. The use of wax alone also adds to the error during today's impression taking [77]. The article suggest a tolerance of 0.3mm, which puts our surface recovery results within tolerance.

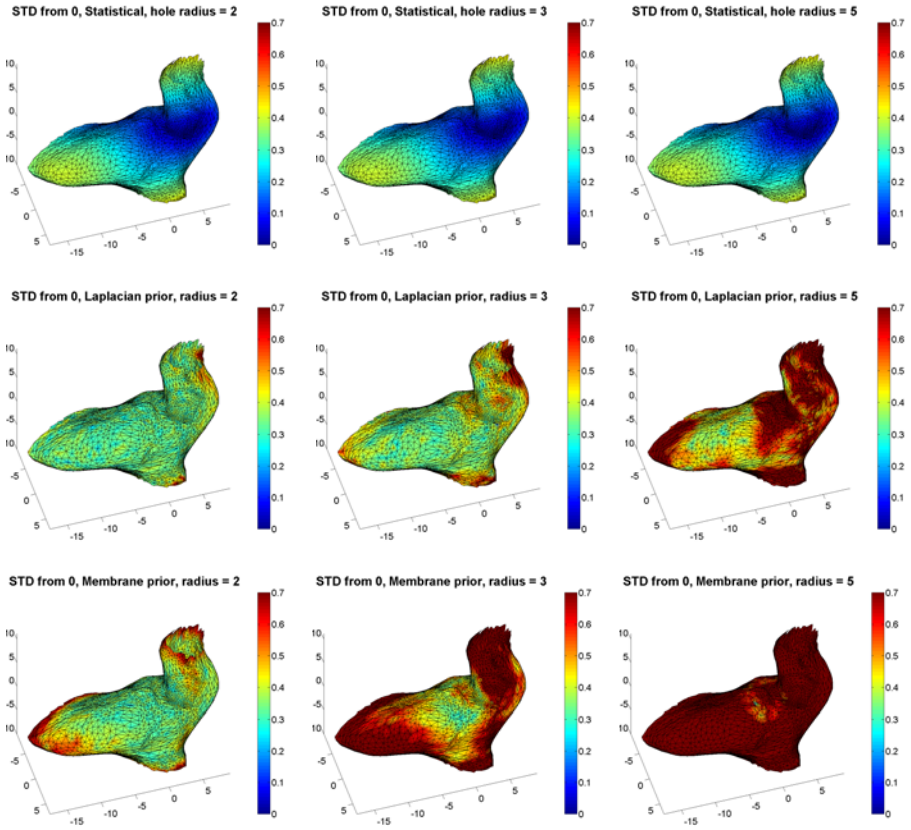


Figure 8.4: Standard deviation from zero of the reconstruction error using our proposed Statistical reconstruction and MRFSurface reconstruction with both Laplacian and Membrane prior energy for holes with radius r = [2mm, 3mm, 5mm]. Notice that the three top figures are not identical, the reconstruction results are just very close to being identical.

8.7 Conclusions and discussion

We have shown that we can predict the missing parts of partial scans using a statistical model. Especially when the holes become big relative to the size of the shape, our method outperforms surface reconstructions that only use the immediate vicinity in the reconstruction.

We have focussed mainly on the statistical method given an alignment and then creating some point correspondences. Future work should focus on getting a better alignment and point correspondence. Different variations of the ICP algorithm [82] are know to produce good alignment. Also, it is possible that non-rigid alignment will improve on the point correspondence and thus the final result. We have also made the assumption that a partial scan can be aligned as well as a full scan, this assumption holds true for holes up to a certain size. As the holes grow and scans become more partial, at some point alignment will no longer be functional and the method can not be applied or it might need to be manually assisted.

The performance of the proposed method is very good, the only really time consuming operation (a matter of seconds for our data set) of the statistical hole closing method is the singular value decomposition used for the variance matrix of *known* points in the statistical model. For a real world application once an alignment and a point correspondence has been established, the singular value decomposition is only done once and can be used to reconstruction all the missing parts. The proposed method is not restricted to the used data, but can be applied to data, where point correspondence throughout the data set can be obtained.

Chapter 9

Anatomically correct surface recovery: A statistical approach

Rasmus R. Jensen, Jannik B. Nielsen, Rasmus Larsen, and Rasmus R. Paulsen

Abstract

We present a method for 3D surface recovery in partial surface scans. The method is based on an Active Shape Model, which is used to predict missing data. The model is constructed using a bootstrap framework, where an initially small collection of hand-annotated samples is used to fit to and register unknown samples, resulting in an extensive statistical model. The statistical recovery uses a multivariate point prediction, where the distribution of the points is given by the Active Shape Model. We show how missing data in a partial scan, once point correspondence is achieved, can be predicted using the learned statistics. A quantitative evaluation is performed on a data set of 10 laser scans of ear canal impressions with minimal noise and artificial holes. The holes are cut in a region known to be hard to cover with a direct scanner. We also present a qualitative evaluation on authentic partial scans from an actual direct in ear scanner prototype. Compared to a state-of-the-art surface reconstruction algorithm, the presented method gives matching prediction results for the synthetic evaluation samples and superior results for the direct scanner data.
Keywords: *surface recovery, hole closing, multivariate statistics, shape modeling, in ear scanning, active shape model*

9.1 Introduction

Direct surface scanning of humans is an increasingly used modality where the applications range from model creation in the entertainment industry, plastic surgery planning and evaluation, craniofacial syndrome evaluation [60, 39], and in particular hearing aid production [72]. In this paper, we are concerned with a particular surface shape namely that of the ear canal. Ear canal surface scans are used in custom hearing aid fitting. This is a very large industry that probably makes the ear the most scanned part of the human anatomy. A standard hearing aid producer generates more than a thousand scans per week. When producing custom in-the-ear devices like hearing aids and monitors, the standard routine is to inject silicone rubber in the patients ear and then laser scan this impression [72]. While this technique normally creates complete surfaces, direct ear scanners are emerging and it is expected that probe scans with these devices will require handling of missing data due to occlusion in the complex anatomy of the human ear and the limited space for the scanner probe.

In this paper we are presenting a method for predicting missing data based on the information in the partial scan. Hole filling and missing data recovery is a well studied problem, in particular for 2D images. In 3D, data recovery is sometimes considered a by-product of the surface reconstruction algorithm. The algorithms used to generate triangulated surfaces from point clouds will usually try to cover missing areas using some mathematical or physical assumptions. One series of approaches uses Delaunay triangulation of border points [56]. Such methods are obviously susceptible to noise in the border points and will typically require some form of smoothing. An alternative strategy is to interpolate implicit (signed distance) functions locally or globally under various forms of regularisation [51, 74]. Other methods, inspired from 2D inpainting approaches have also been investigated [101, 21, 18]. These are typically based on a variational definition of the behavior of the surface where the holes are. In [18] it is the mean curvature of the surface that is being regularised and in [21] it is the Willmore energy over the surface. The reported results of these techniques seem similar to the results from the Markov Random Field surface reconstruction algorithm [74] that we are using as reference.

In our method, we predict the missing points based on the existing points in the scan. Instead of using variational formulations or physical assumptions on the behaviour of the surface, we utilise a population statistics of the given class of surfaces learnt from an annotated and co-registered training set. In the chosen example, we base our population statistics of the ear canal on an extensive statistical shape model of the ear canal constructed in a bootstrap framework based on what was originally presented in [73]. The method is general and is applicable to all surface scans, where a statistical shape distribution can be

estimated.

The 3D morphable models introduced for the analysis and synthesis of 3D faces [12] can also be used to recover missing data in surface scans [13]. In [12] a 3D statistical shape and texture model is built based on a set of registered training samples and from this a principal component analysis is performed giving a set of eigenvectors and values. To recover missing data the set of known points are found in a pre-processing step and the missing data points are found by computing the optimal linear combination of eigenvectors fitting the known data. This is combined with a ridge regression regularisation [30] to avoid non-plausible shapes. The approach described in [12] is similar to our prediction step, but in contrast we also include the steps needed to identify the missing points in the described framework. Furthermore, we also weight the geodesic distance from the missing points to the known points in the prediction.

9.1.1 Data and Preprocessing

The data consists of 310 scanned left-ear impressions, as the one shown in figure 9.1. The scans have been acquired from a traditional 3D scanner, resulting in meshes of arbitrary triangulation. From this collection, 12 representatives are chosen and from these point correspondence over the selected impressions is created using the method initially described in [73]. Here a template mesh is thin plate spline warped to fit each shape using a sparse set of landmarks. The template mesh is furthermore optimized to have an equilateral triangulation. This is followed by a propagation of template mesh vertices to each shape based on a nearest neighbour search. Since the ears are topologically equivalent to open cylinders special care must be taken in the non-overlap regions and the template is pruned accordingly. Furthermore, the Markov Random Field regularization of the correspondence field described in [71] was used to further optimize the dense correspondence. This small subset of impressions form the basis for the bootstrapping framework used to encompass the entire collection of ear impressions, with the goal of constructing a statistical shape model as for example described in [22, 73].
In addition to the traditionally scanned impressions presented above, a small collection of so-called partial scans also exists. The scans have been acquired by a prototype in-ear 3D scanner[2]. They are partial in the sense that some areas of the surface are missing due to noise and/or occlusion. Finally, a small set of scanned ear-impressions, not part of the original 310 samples, have had holes cut in them to mimic the nature of the partial scans. We denote these manually created partial scans as synthesized partial scans. This set is used for controlled evaluation of our method. Figure 9.2 shows examples of authentic and synthesized partial scans.

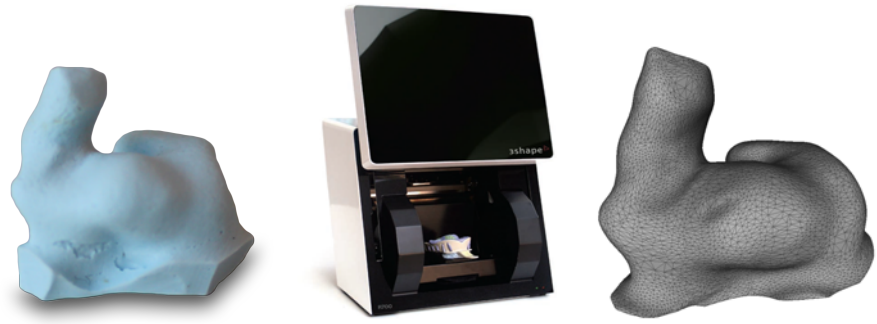


Figure 9.1: An ear impression (left) and the corresponding point cloud (right) obtained using a traditional 3D scanner [2] (middle). For clarity, only the points on the visible part of the surface are shown.

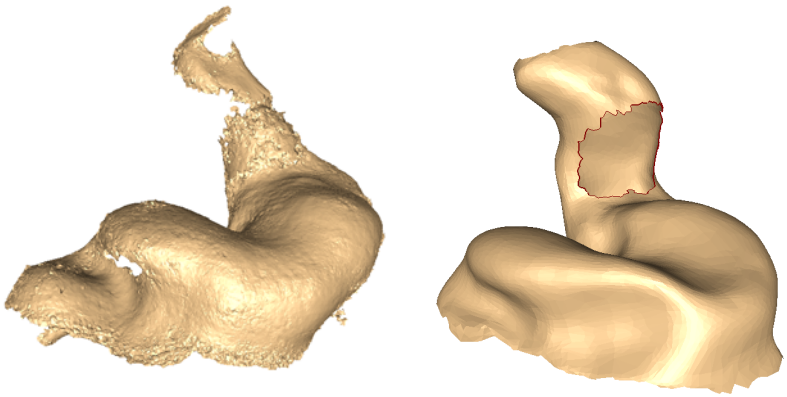


Figure 9.2: Left: Authentic partial scan from prototype in-ear scanner. Right: Synthesized partial scan, used for controlled evaluation.

9.2 Bootstrapped Active Shape Model

In order to accurately recover missing information in a partial scan, a significant statistical basis must be constructed. Acquiring such a basis is not trivial although, as here, a large dataset is available. This is due to the fact that generally no point-correspondence exists between scanned samples in a dataset, leading to a need for co-registration of samples. Often, the task of co-registering meshes is done manually, as is seen in e.g. [72]. However, annotating larger collections of samples becomes an extensive and time-consuming task. We address this problem by proposing a bootstrapped registration procedure.

Initially a small subset of samples is manually annotated and registered using the approach described by Paulsen et al. in [73, 71]. Using this small subset, an Active Shape Model (ASM) is constructed as described in [22, 73]. The statistical model is aligned and fitted to each unknown sample. This is done iteratively, allowing co-registration to and inclusion in the ASM, thereby expanding the model sample by sample. The ASM thus grows in size as the bootstrapping procedure processes unknown samples, allowing it to explain more and more shape variation from the dataset. Intuitively this leads to the expectation that the algorithm will become increasingly better at fitting to unknown shapes and that latter samples are better registered than former, wherefore a revisit of early registrations may be chosen as a finalising step. An overview of our bootstrap-framework is outlined in Algorithm 9.1.

9.2.1 Initial Active Shape Model

Initially we introduce the concept of the Active Shape Model (ASM), that forms the basis for the bootstrapping and later the data recovery process. Initially introduced by Cootes et al. [22], the ASM is a statistical model that seeks to describe the shape variation of a collection of samples.
Assuming a collection of m aligned shapes, each consisting of p 3D points

$$\mathbf{v}_i = (x_1, y_1, z_1, \ldots, x_p, y_p, z_p)^T \in \mathbb{R}^n, \tag{9.1}$$

these shapes can be interpreted as being points in an $n = 3p$ -dimensional space. The average shape is thus

$$\bar{\mathbf{v}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{v}_i \tag{9.2}$$

and the shape deviation from mean

$$\mathbf{x}_i = \mathbf{v}_i - \bar{\mathbf{v}}. \tag{9.3}$$

Algorithm 9.1 Bootstrapping Procedure

Given a collection of samples with no point-correspondence:

1. Manually annotate and co-register a small subset of samples.
2. Build an Active Shape Model (ASM) based on the manually annotated samples.
3. For each unknown sample in the dataset:
   (a) Roughly align sample to the mean of the ASM using Shape Context features.
   (b) Adjust alignment of sample using the Iterative Closest Point (ICP) algorithm.
   (c) Fit ASM to the adjusted alignment.
   (d) Repeat (b) and (c) until convergence.
   (e) Project the fitted ASM mesh onto the sample.
   (f) Check quality of projection and add result to ASM if valid.
4. *(Optional)* Repeat 3, updating samples in the ASM where projection quality is better than the previous run.

In order to investigate the variation of the data, an observation matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathbb{R}^{n \times m}$ can be constructed. The covariance matrix, $\mathbf{\Sigma}$, of $\mathbf{X}$ is found by

$$\mathbf{\Sigma} = \frac{1}{m}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{n \times n}. \tag{9.4}$$

Performing an Eigenvalue decomposition of this covariance matrix, thus provides insight in the primary modes of variation within the dataset

$$\mathbf{\Sigma} = \mathbf{P}\Lambda\mathbf{P}^T, \tag{9.5}$$

where $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_m)$ is a matrix consisting of columns of Eigenvectors and $\Lambda = diag(\lambda_1, \ldots, \lambda_m)$ is a diagonal matrix holding the Eigenvalues. These Eigenvalues corresponds to the variation expressed in the respective Eigenvector directions, i.e. $\lambda_i = \sigma_i^2$. In scenarios where $m < n$, only a subset of the Eigenvalues will be non-zero, the size of this subset will be denoted m'.

Given the collection of non-zero Eigenvalues and corresponding Eigenvectors described above, these can be used as a basis. Any shape $\tilde{\mathbf{v}}$ can then be synthesised by a linear combination of the Eigenvectors, weighted by their Eigenvalues:

$$\mathcal{M}(\mathbf{c}) = \tilde{\mathbf{v}} = \sum_{i}^{m'} c_i \lambda_i \mathbf{p}_i = \mathbf{P}\Lambda\mathbf{c}, \tag{9.6}$$

where $\mathbf{c} = (c_1, \ldots, c_{m'})$ is a vector of weights determining how much the individual Eigenvectors contributes in the synthesis. This constitutes the Active Shape Model and hereby the ASM can be interpreted as a function of the weights in $\mathbf{c}$, i.e. $\mathcal{M}(\mathbf{c})$.

9.2.2 Automatic Pre-Alignment

The raw samples to be included in the ASM may not be positioned or oriented correctly relative to each other. This may not always be the case, but in scenarios where it is so, pre-alignment prior to the final, fine, alignment is needed in order to ensure a proper, global as well as local, alignment, i.e. ensuring that the fine alignment procedure does not terminate in a local minimum.

Multiple approaches to automatic alignment of shapes exists, we have chosen to use 3D Shape Context Descriptors [31] in this implementation [9, 58, 38]. The descriptors describes a point on a 3D surface by a histogram of its local neighbourhood, indicating the local geometric distribution of points. Given a point $\mathbf{q}$ on a surface, any neighbouring point's relative position to $\mathbf{q}$ can be

expressed in spherical coordinates (r, θ, φ). The radial distance r between $\mathbf{q}$ and a neighbour $\mathbf{q}_n$, can be computed by

$$r = \|\mathbf{q} - \mathbf{q}_n\|. \tag{9.7}$$

The inclination angle θ and the azimuthal angle φ requires choice of reference-frame in order to be intercomparable between differently aligned samples. In this experiment, molds were acquired from a laser scanner using a rotating platform. The 3D representations of the moulds thus have a consistent vertical axis. This consistency can be utilised to construct a common frame of reference. In this frame of reference the third basis element is aligned with the normal of the point $\mathbf{q}$. This is formulated as $\mathbf{b}_3 = \mathbf{n}_\mathbf{q} = (n_x, n_y, n_z)^T$. The first basis element is aligned with the vertical axis, with the restraint of being orthogonal to $\mathbf{b}_3$. Denoting a vector pointing along the fixed vertical axis $\mathbf{v} = (0, 1, 0)^T$, this is found by $\mathbf{b}_1 = \mathbf{v} - (\mathbf{v} \cdot \hat{\mathbf{b}}_3)\hat{\mathbf{b}}_3$, i.e. a vector rejection of $\mathbf{v}$ on $\mathbf{n}_\mathbf{q}$, where $\hat{\mathbf{b}}$ denotes the normalised value of $\mathbf{b}$. As a result of orthogonal basis vectors in a right-handed coordinate-system, the second basis element is thus restrained to being $\mathbf{b}_2 = \mathbf{b}_3 \times \mathbf{b}_1$. From this basis, a rotation matrix $\mathbf{R}$, rotating to the local frame of reference can be constructed:

$$\mathbf{R} = \left[\hat{\mathbf{b}}_1\ \hat{\mathbf{b}}_2\ \hat{\mathbf{b}}_3\right]. \tag{9.8}$$

Any neighbouring point, $\mathbf{q}_n$, can thus be described in $\mathbf{q}$'s local frame of reference by:

$$\tilde{\mathbf{q}}_n = \mathbf{R}(\mathbf{q}_n - \mathbf{q}) \tag{9.9}$$

Within this frame of reference, the inclination angle and the azimuthal angle of the point is given by:

$$\theta = \arccos\left(\frac{\tilde{q}_{n,z}}{r}\right) \qquad \varphi = \arctan\left(\frac{\tilde{q}_{n,y}}{\tilde{q}_{n,x}}\right). \tag{9.10}$$

Based on the coordinates (r, θ, φ), points in the proximity of $\mathbf{q}$ can be grouped in a discrete set of bins. Hereby a histogram over the 3-dimensional distribution of points surrounding $\mathbf{q}$ can be constructed and used as a feature vector. In our experiment, (r, θ, φ) of points within a radius of $10mm$. were divided into $(8, 13, 4)$ bins respectively, yielding a 416-dimensional feature vector or Shape Context Descriptor.

The choice of utilising the vertical axis to construct a common frame of reference poses a constraint on the geometry as points having normals parallel to the vertical axis cannot be used. In practice this means that perfectly horizontal surfaces cannot be evaluated. Is this constraint not tolerable, or is there no common axis, one may use other methods for constructing a local frame of reference. [99] suggests obtaining the basis by doing an Eigenvalue Decomposition of a weighted Covariance Matrix of the local neighbourhood.

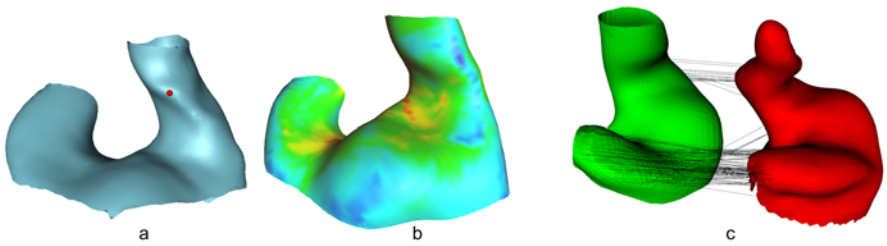


Figure 9.3: The descriptor in the point marked by the red marker in a) is compared to the descriptors in all points of b). Colouring in b) corresponds to the χ^2 distance between the descriptors. The minimum distance, and thus the most similarity, is correctly recognised on the ridge of the right shape. c) Lines indicating the 100 most significant Shape Context matches between the two samples. Most lines indicate a correct matching, there are, however, some mismatches.

Through the Hungarian method [59], point-descriptors are matched and based on this matching a corresponding transformation can be computed in a least squares sense. The cost function used in the matching is given by

$$C(\mathbf{q}_{i,j}) = \frac{1}{2}\sum_{k=1}^{K}\frac{|h_i(k) - h_j(k)|^2}{h_i(k) + h_j(k)}, \tag{9.11}$$

describing the cost C of matching points i and j. K is the number of bins used in the local neighbourhood histograms and $h(k)$ denotes the k^{th} bin in the respective histogram. This is also known as the χ^2 histogram distance. An illustration is shown in figure 9.3. Given the result of this matching with $\mathbf{q}_i$ and $\mathbf{q}_i'$ being a matched pair of points, translation and rotation is obtained by solving the least squares problem

$$\min_{\mathbf{R},\mathbf{t}} \sum_i \|(\mathbf{R}\mathbf{q}_i + \mathbf{t}) - \mathbf{q}_i'\|^2. \tag{9.12}$$

Using this method, unknown samples are aligned to the mean of the ASM, $\bar{\mathbf{v}}$, thus supplying a plausible pre-alignment. Failed pre-alignments are easily detected by evaluating the average Euclidean point to point distance between the mean shape, $\bar{\mathbf{v}}$ and aligned shape $\mathbf{v}_a$. In our dataset, alignments with average point-to-point distances above 5 mm are rejected.

9.2.3 Iterative Fitting and Re-Alignment

Given a roughly aligned unknown sample, $\mathbf{v}_a$, the alignment is refined and the ASM, $\mathcal{M}(\mathbf{c})$, is fitted. This is done in an iterative manner, where an Iterative Closest Point (ICP) alignment of the sample is followed by an ASM-fitting of the model, and repeated upon until convergence is met.

The ICP alignment is an iterative process where two surfaces are aligned using a rigid-body or similarity transform. ICP was initially described in [110, 11] and has later been extended in a variety of ways [82]. Modifications to the ICP includes point-to-surface matching, handling of non-overlapping regions, curvature weighting and statistical point correspondence rejection. In our ICP implementation the points are matched to their nearest neighbours, with the constraint that points connected to a border should be ignored. Using ICP, $\mathbf{q}'_i$ in equation 9.12 therefore denotes the nearest neighbour.

For the fitting, we seek to find a deformation of the Active Shape Model that minimises the error between the model and the unknown sample. An ASM constructed from the shape analysis of m samples, each consisting of n points, will be parametrised in an m'-dimensional space and thus have m' modes of variation.

Let $\mathbf{q}_i \in \mathbb{R}^3$ be a point belonging to the ASM, $\mathcal{M}(\mathbf{c})$, and let $\mathbf{q}'_i \in \mathbb{R}^3$ be the closest point on the target sample surface $\mathbf{v}_a$, we seek to find the set of weights $\mathbf{c}^*$ that minimises the sum of distances between $\mathbf{q}$ and $\mathbf{q}'$:

$$\arg\min_{\mathbf{c}} \|\mathcal{M}(\mathbf{c}) - \mathbf{v}_a\| = \arg\min_{\mathbf{c}} \frac{1}{m'} \sum_{i=1}^{m'} \|\mathbf{q}_i - \mathbf{q}'_i\| \tag{9.13}$$

We solve this optimisation problem by utilising an implementation of the Nelder-Mead method [69], however any multidimensional optimiser may be used. One may compress the parametrisation of the model by reducing the number of parameters used to a number $k \leq m'$, often based on a pre-determined requirement for fraction of explained variance. In our implementation, we reduced the number of parameters used to a number corresponding to 99% explained variance. For datasets with few samples this reduction is insignificant, but as the ASM grows, the model is compressed considerably.

As the model-fitting is basically a synthesisation from a k-dimensional (assumed) normal-distribution, a confidence level for an obtained set of parameters $\mathbf{c}^*$ can thus be computed by utilising the Mahalanobis distance M between the parameter set and the ASM distribution since $M^2 \sim \chi^2_n$. This allows validation of fittings by setting a reasonable confidence limit. In our implementation, a confidence level of 99.9% was used.

9.2.4 Registration of Fitting

Having determined $\mathbf{c}^*$ for $\mathcal{M}$ and $\mathbf{v}_a$, the model mesh $\mathcal{M}(\mathbf{c}^*) = \tilde{\mathbf{v}}$ is propagated to the sample shape $\mathbf{v}_a$ in order to perform a point-wise registration. The procedure of co-registration is described in [71], and utilises a Markov Random Field regularisation of the correspondence field between the two shapes. A successful registration will thus result in a mesh of p points, following the shape of $\mathbf{v}_a$, all with correspondence to the model, $\mathcal{M}$. In order to verify that registrations are valid, two mesh features are inspected. First the triangulation of the resulting mesh is inspected. Ideally the angles of all triangles in the mesh should be 60°, however, as $\tilde{\mathbf{v}}$ is projected onto $\mathbf{v}_a$, mesh-distortions may arise. In failed registrations one may thus observe significant changes in the resulting triangles. By inspecting the minimum angle in each triangle α_{min}, the value of the projection's 20^{th} percentile of α_{min} can be found and used as a measure for triangulation quality. Due to its nature this method allows registrations to have some degeneracies without being rejected. We choose to reject registrations having a 20^{th} percentile value below 15°, recalling, as stated in 9.1.1, that the template mesh has near equilateral triangulation.

Secondly the normals of the projection are compared to the normals of $\tilde{\mathbf{v}}$. In this case the normals would ideally be the same for the two, but also here the projection may distort the resulting mesh and for erroneous registrations, the normals may change significantly. The dot-product between the normals of $\tilde{\mathbf{v}}$ and the registration yields insight to this distortion. Here, the average dot-product is evaluated. In our implementation, registrations having an average dot-product below 0.75 are rejected.

9.2.5 Finalisation

When an alignment and a registration is obtained using the iterative scheme above, they are both refined iteratively. During each iteration the registration-mesh is smoothed using simple meaning of the nearest neighbours. The surface normals of the smoothed mesh are found and regularised using local averaging of directions. A new set of correspondence points are found in the sample scan in the direction of the regularised normals and the alignment, $\mathbf{v}_a$, is adjusted accordingly. This process is repeated until convergence. As the sample input scan is expected to be more densely sampled than $\mathcal{M}(\mathbf{c}^*)$ the iterative update ensures a regular mesh with evenly distributed vertices.

9.2.6 Inclusion

For each sample where the ASM is able to successfully fit with a valid set of parameters $\mathbf{c}^*$, and where the registration produces a mesh of good quality with respect to the quality-measures described above, the ASM, $\mathcal{M}$, can be expanded to also include this registration. The number of samples m, forming the basis for the statistical model, is thus incremented by 1. Assuming a correct registration, the knowledge of the ASM is hereby improved to cover additional shape variation, making it better prepared for future samples. This results in the statistical model becoming increasingly better as it is presented to more samples. As a result of this behaviour, latter samples are expected to be registered more easily than former, thus suggesting that former samples should be revisited by the final model. An optional final step is therefore to recompute the ASM fittings and registrations on all samples a chosen number of times. This may be of interest if the ASM has difficulties fitting to samples placed early in the dataset.

9.3 Co-registration of Partial Scans

As described, a crucial, and not easily solved, part of recovering missing data is to co-register an unknown mesh with the ASM. This is required in order to obtain point-correspondence between model and surface, creating a partial scan with a mesh structure identical to that of the model.

The process of co-registering an unknown scan to the ASM is basically addressed in section 9.2. In the case of reconstructing partial scans, however, the exact same approach may not suffice. This is mainly due to the fact that an automatic alignment between a partial scan and model may prove to be difficult for Shape Context features. The difficulties arise in scenarios where the key shape features of the model are not present on the scan or vice-versa. The general problem of creating point-correspondence between shapes is a non-trivial matter with a variety of existing approaches [50] and not the main topic of this paper. In our implementation we limited ourselves to the already existing Shape Context alignment approach, and where this failed, manual alignment was used. The result of registering and fitting the ASM to a partial scan is that the ASM template mesh is deformed and propagated to the partial scans in areas where there are valid data. The template mesh vertices are marked as missing when the corresponding point or area in the partial scan is not present or valid. If an area is missing in the partial scan, the point projection will often result in that the project point is placed on a boundary in the partial scans, thus enabling

detection of missing point correspondences.

9.4 Surface Recovery

Given a registration, we aim to recover missing surface data in a partial scan such that the recovered data are anatomically correct. We approach this by using a statistical model and define the set of known and unknown data in a partial scan as follows:

$$\begin{aligned} \text{missing vertices: } \mathbf{s}_1^{\mathrm{T}} &= (x_{11}, y_{11}, z_{11}, x_{12}, y_{12}, z_{12}, \ldots) \\ \text{known vertices: } \mathbf{s}_2^{\mathrm{T}} &= (x_{21}, y_{21}, z_{21}, x_{22}, y_{22}, z_{22}, \ldots) \end{aligned}$$

The correspondence allows for differentiation between known vertices and missing vertices in the partial scan. We will determine how the unknown data $\mathbf{s}_1$ are predicted from known vertices in $\mathbf{s}_2$.

Without any prior knowledge of the distribution of data, we consider a shape $\mathbf{s}$ consisting of $\mathbf{s}_1$ and $\mathbf{s}_2$ as belonging to the normal distribution:

$$\mathbf{s} = \begin{bmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \end{bmatrix} \in N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \right), \ \mathbf{\Sigma}_{12}^{\mathrm{T}} = \mathbf{\Sigma}_{21} \tag{9.14}$$

The expected value of $\mathbf{s}_1$ given $\mathbf{s}_2$ is:

$$\mathrm{E}\{\mathbf{s}_1|\mathbf{s}_2\} = \mu_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{s}_2 - \mu_2) \tag{9.15}$$

With the variance:

$$\mathrm{V}\{\mathbf{s}_1|\mathbf{s}_2\} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21} \tag{9.16}$$

From the ASM we get an aligned set of shapes. This training set is denoted $X_{aligned}$. From the training set the covariances $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{12}$ as well as the means (μ_1, μ_2) are learned. As there are far less shapes than points, $\mathbf{\Sigma}_{22}$ will be singular. Let $\mathbf{\Sigma}_{22} = \mathbf{P}\Lambda\mathbf{P}^{\mathrm{T}}$ be the Eigenvalue decomposition. We restrict $\mathbf{\Sigma}_{22}$ to its affine support, i.e. the dimensions spanned by the Eigenvectors corresponding to the k positive Eigenvalues, such that:

$$\mathbf{\Lambda}^* = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k) \text{ and } \mathbf{P}^* = [\mathbf{p}_1 \ \mathbf{p}_2 \ \ldots \ \mathbf{p}_k] \tag{9.17}$$

The projection of $\mathbf{s}_2$ using the k selected Eigenvectors $\mathbf{P}^*$: $\mathbf{y}_2 = \mathbf{P}^{*\mathrm{T}}\mathbf{s}_2$ has affine support for $\mathbf{s}_2$ and the variance:

$$\mathrm{V}\{\mathbf{y}_2\} = \mathrm{V}\{\mathbf{P}^*\mathbf{s}_2\} = \mathbf{P}^{*\mathrm{T}}\mathbf{\Sigma}_{22}\mathbf{P}^* = \mathbf{\Lambda}^* \tag{9.18}$$

The covariance of $\mathbf{s}_1$ and $\mathbf{y}_2$ is:

$$\mathrm{C}\{\mathbf{s}_1, \mathbf{y}_2\} = \mathrm{C}\{\mathbf{s}_1, {\mathbf{P}^*}^{\mathrm{T}}\mathbf{s}_2\} = \mathrm{C}\{\mathbf{s}_1, \mathbf{s}_2\}\mathbf{P}^* = \mathbf{\Sigma}_{12}\mathbf{P}^* \tag{9.19}$$

Finally, the prediction of the unknown data $\mathbf{s}_1$ can be done using the projection $\mathbf{y}_2$:

$$\mathrm{E}\{\mathbf{s}_1|\mathbf{y}_2\} = \mu_1 + \mathbf{\Sigma}_{12}\mathbf{P}^*{\mathbf{\Lambda}^*}^{-1}{\mathbf{P}^*}^{\mathrm{T}}(\mathbf{s}_2 - \mu_2) \tag{9.20}$$

This expectancy can be used for any unknown set of vertices $\mathbf{s}_1$ given a partial scan $\mathbf{s}_2$, be that a single missing vertex or all the missing data. If every unknown vertex is predicted according to the described method the known triangulation from the training set can be propagated to the predicted data set and will then constitute a full surface reconstruction. The method can also be used to filter data for noise if the known scan data are also recovered. By varying the number k values of Eigenvalues and vectors used in the projection the fraction of described variance can be controlled.

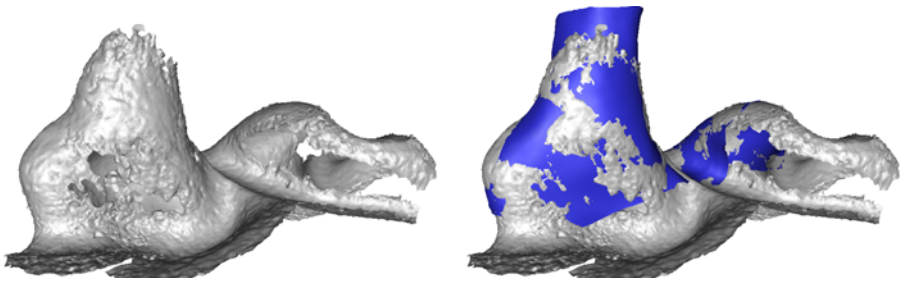
9.4.1 Algorithm for Surface Recovery

The known vertices according to the ASM found during the registration of the partial scan is $\mathbf{s}_2$ and the full scan as provided by the scanner is $\mathbf{s}_{scan}$. Let $\mathbf{s}_1^*$ be the predicted missing data, $\mathbf{s}_2^*$ the prediction of the partial scan and $\mathbf{s}^*$ be the full reconstructed shape. The full average shape is denoted μ. With an initial registration the algorithm works as follows:

Algorithm 9.2 Anatomical surface recovery

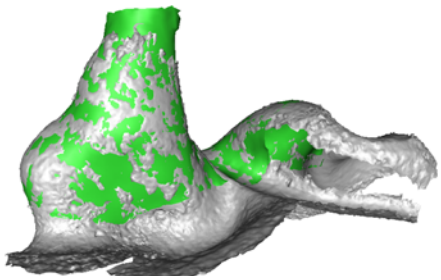
procedure AnatomicalSurfaceRecovery($X_{aligned}, s_2, s_{scan}$) ▷
  $\mathbf{s}^* \leftarrow \begin{bmatrix} \mathbf{0} \\ \mathbf{s}_2 \end{bmatrix}$
  repeat
    Procrustes align $\mathbf{s}^*$ to μ and apply same transformation to $\mathbf{s}_{scan}$
    Predict $\mathbf{s}_1^*$ using the described method ▷ the described variance is increased in each iteration
    Predict $\mathbf{s}_2^*$ using the described method ▷ the described variance is increased in each iteration
    $\mathbf{s}^* \leftarrow \begin{bmatrix} \mathbf{s}_1^* \\ \mathbf{s}_2^* \end{bmatrix}$
    Find vertex correspondence between $\mathbf{s}^*$ and $\mathbf{s}_{scan}$
    Update $\mathbf{s}_2$ and $\mathbf{s}^*$ with the correspondence vertices from $\mathbf{s}_{scan}$
  until convergence
  return $\mathbf{s}_1^*$ and $\mathbf{s}_2^*$ ▷
end procedure

We repeat the loop body with two different recovery approaches. First $\mathbf{s}_1^*$ and $\mathbf{s}_2^*$ are predicted all at once. In the last few loops the data are predicted vertex by vertex using only the nearest vertices in the prediction. The vertex distances are found as the geodesic distances on the mean shape, so these only have to be calculated once. The geodesic distances are used to ensure topological consistency when selecting a neighbourhood. Our shape model has 3096 vertices and in the local recovery we only use the 10 nearest of these. In the local prediction the recovered data is locally very true to the original scan. We restrict the Eigenvalues in the recovery to the ones describing 30% of the variance and then gradually raise this to 99.9%. Gradually raising the percentage of described variance helps the algorithm produce anatomically correct shapes and prevents the influence of bad correspondences in the initial iterations. Figure 9.4(a) shows a scan with parts missing from the ear canal, tragus and concha. Figure 9.4(b) and Figure 9.4(c) shows the surfaces created using the full recovery followed by one where the final recovery has been done locally. The latter is much more true to the data.

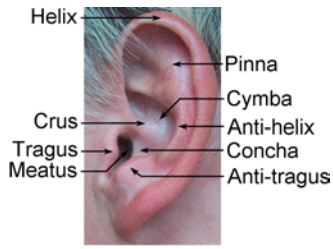


(a) Scan

(b) Recovered full



(c) Local recover



(d) Anatomical reference

Figure 9.4: 9.4(a) shows a scan with parts of the ear canal, tragus and concha missing. 9.4(b) shows the recovered surface when the shape is recovered in full. 9.4(c) shows the recovered surface where the surface recovery is done locally only. This is much more true to the actual data. 9.4(d) shows an anatomical reference of the ear.

9.5 Results

9.5.1 Bootstrapped Active Shape Model

Based on the method described in section 9.2, we were able to construct an extensive Active Shape Model of the left ear based on the available dataset. A total of 310 samples were processed and from these, 241 passed the automatic quality verification. In Figure 9.5 a random selection of aligned scans are presented. From this figure it is clearly seen that the 3D Shape Context Descriptors are able to correctly match and align unknown meshes to the ASM, providing a good pre-alignment for the ICP and ASM fitting procedures. As the Active Shape Model processes new samples, the complexity of the model increases and thus the fraction of variation explained per principal component must be expected to drop. This is seen in Figure 9.6 where the fraction of variation explained by the 10 first principal components were plotted as the ASM grew in size. In addition, it can be seen that the explained variance seems to stabilise somewhat as the model grows in size, indicating that the shape model eventually captures the class variability. As stated, the bootstrapping resulted in a shape model covering a total of 241 samples, and thus containing 241 modes of variation. The variation explained by the individual modes has been plotted in Figure 9.7. 90% of all variance is contained within the first 37 modes of variation and the curve indicates a somewhat compact model. We do, however, expect that the automatic registration procedure has induced a significant amount of false variation in form of vertex drifting along the sample surfaces. Such variation of course directly affects the compactness of the ASM in the form of low-variance principal components, assuming that the drifts are uncorrelated. The actual shape variation from the ear is therefore expected to be found within the former principal components. After having run the bootstrap a finalising step was introduced. This step consisted of having all mesh registrations briefly inspected by an operator in order to verify that no meshes contained significant distortions. The operation is easily done in any 3D mesh browser. The inspection resulted in an additional 80 registrations being removed from the ASM. Effectively this resulted in the final ASM consisting of 161 shapes.

9.5.2 Surface Recovery in Synthesized Partial Scans

In order to compare our approach with existing methods for reconstruction, a controlled experiment was set up. A collection of 10 3D ear-scans, not included in the training data, was chosen and all scans had a reasonable sized hole cut in them. The holes were cut between first and second bend of the ear canal, in an

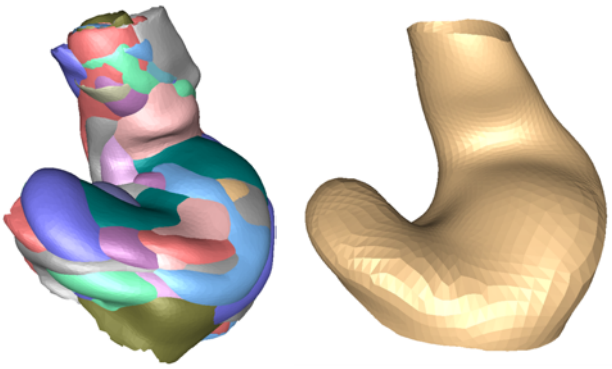


Figure 9.5: Left: A selection of 10 scanned ear-molds that have been automatically pre-aligned using Shape Context Descriptors. Right: Mean shape of aligned samples.

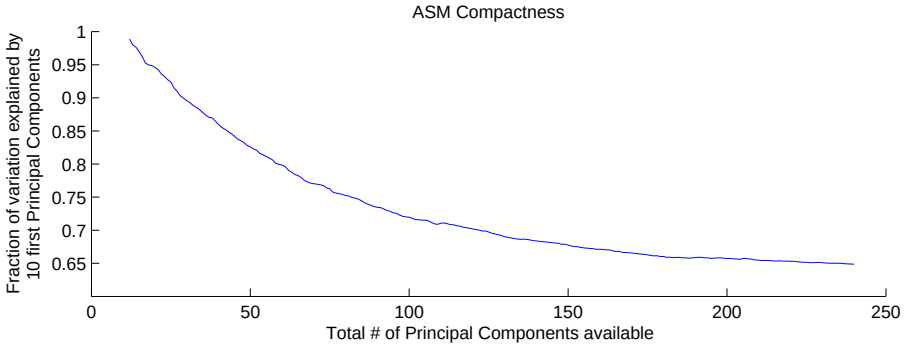


Figure 9.6: Plot of the explained variance by the 10 first principal components in the Active Shape Model as the it grows in size. The explained variance seems to stabilise somewhat as a sufficient number of samples has been included.

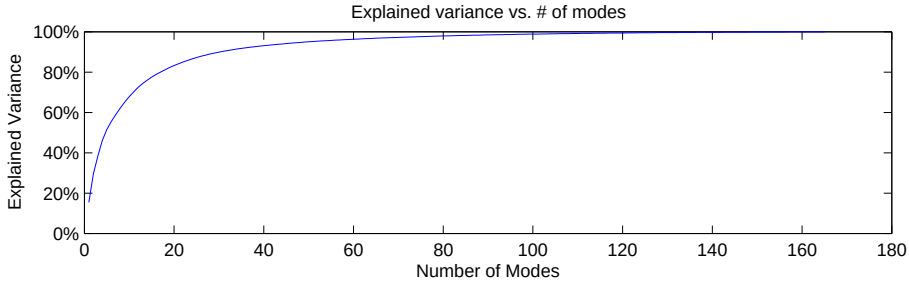


Figure 9.7: Plot of modes vs. variance for the final ASM.

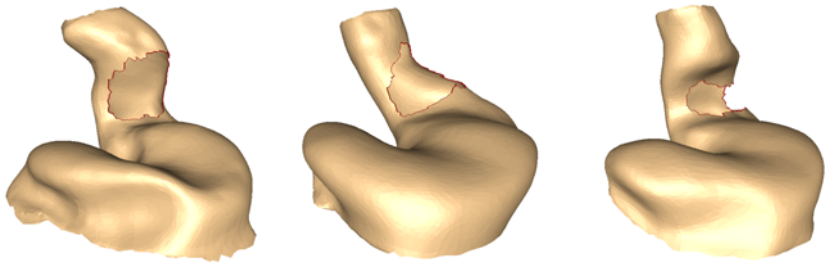


Figure 9.8: 3 meshes where a hole has been cut between first and second bend. This is an area that is often found occluded with experimental optical in-ear scanners. A total of 10 meshes had holes cut like the ones shown here.

area that is known to often be occluded when using experimental optical in-ear scanners. 3 of these samples are shown in Figure 9.8. Hereby any reconstruction of these partial scans can be compared to the ground-truth, allowing for a quantitative comparison of methods. For each mesh in the collection of *synthesised* partial scans the missing data was recovered. This was done using our method, both with and without smoothing, and the Markov Random Field (MRF) surface reconstruction approach, described in e.g. [74]. The choice of comparing to the MRF approach is based on the fact that it is known to produce top-quality reconstructions of smooth surfaces. All reconstructions were then compared to the ground truth, by computing a signed distance (based on surface normals) between all reconstructed points and the original surface. In Figure 9.9 the reconstructions of the previously shown synthetic partial scans are shown, where the surface values denotes the signed distance between reconstruction and truth.

Statistics on all reconstructions are shown in figures 9.10, 9.11, and 9.12. They clearly indicate that the performance of our method is indeed comparable to the

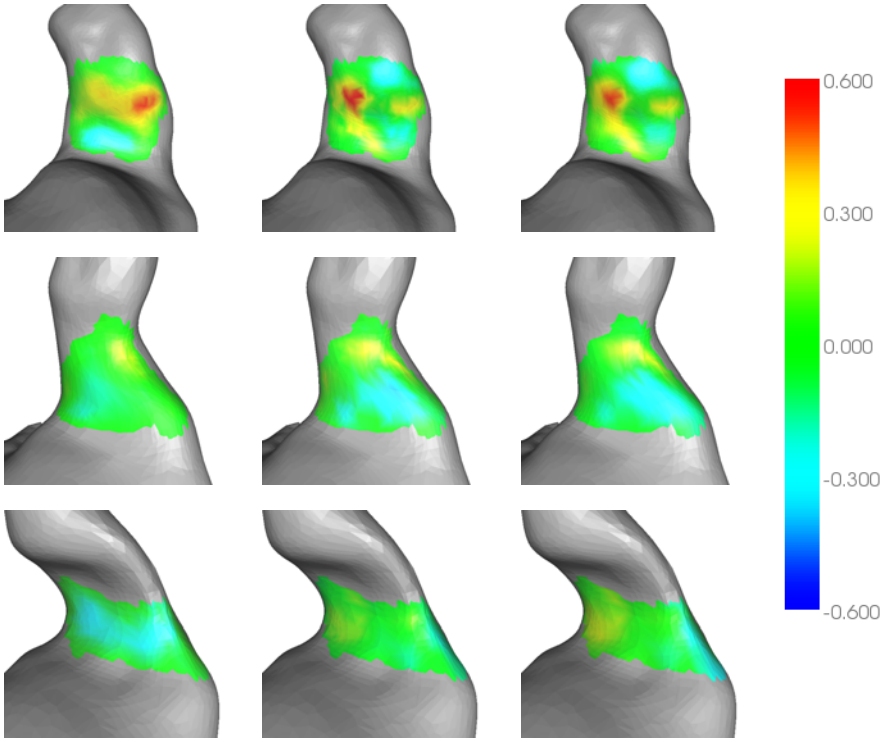


Figure 9.9: Reconstruction of missing data for 3 different scans (rows), using Markov Random Field (MRF) reconstruction (column 1), our method (column 2) and the smoothed variant of our method (column 3). Surface values corresponds to the signed distance between reconstruction and ground truth.

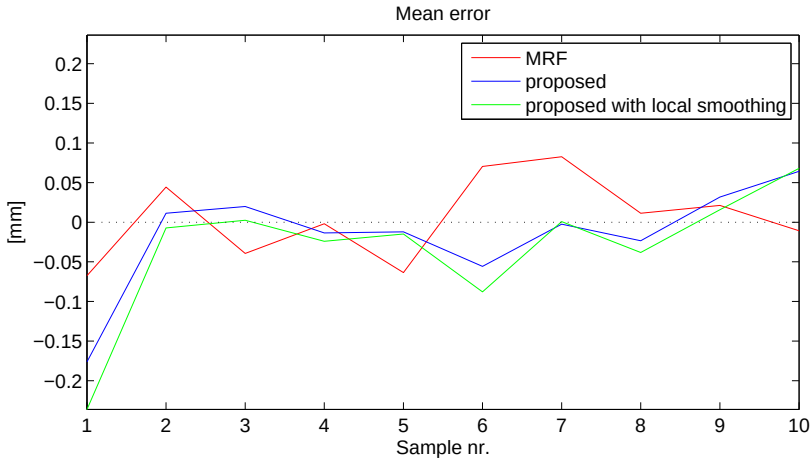


Figure 9.10: Mean of signed point-distances between surface reconstructions and ground truth, for 10 different samples. Our method does not outperform the MRF method, but performs almost equally. The large errors seen in sample 1 are caused by a significant physical deformity in the patients ear.

MRF approach. A significant jump in both the mean and standard deviation of the reconstruction error is observed in sample #1. After inspection, this sample revealed an abnormal cavity in the skin of the ear-canal, explaining the higher error. It should be noted that no prior, neither statistical or physical, would be able to predict such errors. Although this comparison proves high performance of our method, it does not fully illustrate the strength of having a statistically based prior. The MRF approach predicts missing points based on the existing curvature of data in contrast to our method that predicts missing points based on knowledge of the shape variation of an ear population. Effectively this means that where either noisy edges exists or data is sparse, the MRF approach has little chance of estimating the true surface. In this test each hole is surrounded by smooth noiseless surface areas providing an optimum setting for the MRF reconstruction. In the following, we will present a qualitative comparison based on authentic optical 3D scans of the ear, suffering from high noise and sparse point support.

9.5.3 Surface Recovery in Direct Ear Scan Data

We have shown that our proposed method produces good results on a somewhat artificial data set. Artificial in the sense that molds have been laser scanned

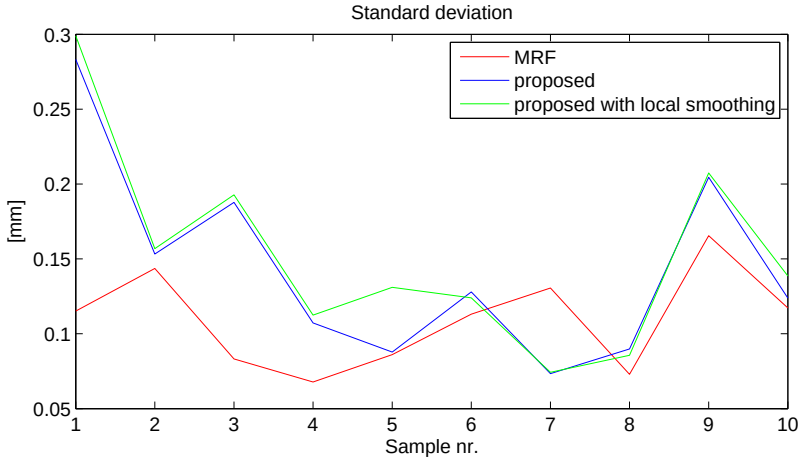


Figure 9.11: Standard deviation of signed point-distances between surface reconstructions and ground truth, for 10 different samples.

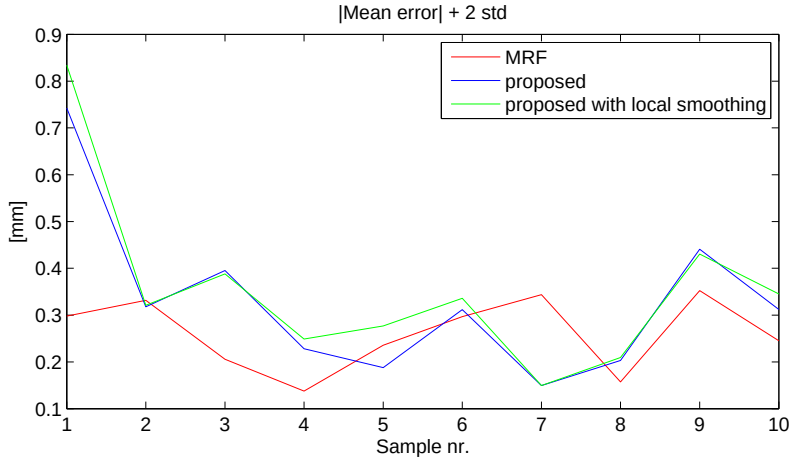
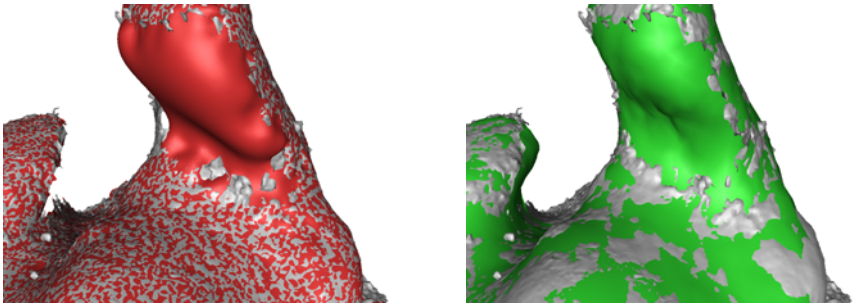


Figure 9.12: Mean plus 2 standard deviations of the absolute values of point-distances between surface reconstructions and ground truth, for 10 different samples. This indicates how large an error fluctuations one should expect.



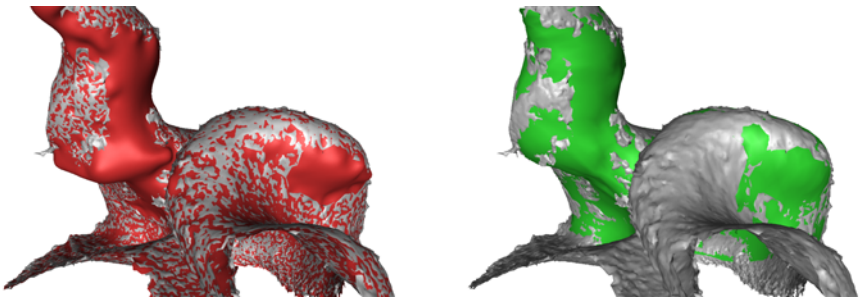
(a) MRF (b) Proposed

Figure 9.13: This figure shows a close up of the original scan with missing data and noise and the MRF reconstruction and our proposed method. Using only the surrounding data the holes closed with MRF are much more bulky as a result of the noise, while our proposed method creates a more probable surface.

before we have manually created a hole. The remaining surfaces are therefore near perfect and far from what an actual in ear scanner will produce. In direct ear scans different kinds of noise will be present coming from hair and ear wax but also possible scanner noise as well as occlusion.

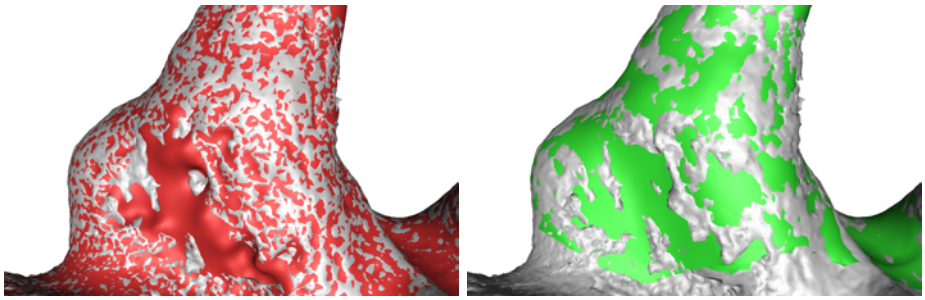
We have tested our algorithm on 12 scans from a prototype direct in ear scanner [2]. In cases with a lot of noise having a strong prior, which is what our statistical model provides, proves to be very useful. In the presence of a lot of noise our method still produces anatomically correct meshes, which are locally true to the scan data in the covered areas. Qualitative inspection shows very good hole closing in the 12 scans. In addition all 12 scans were 3D printed as earplugs and tested by the respective test subjects with positive feedback.

Figure 9.13 shows a scan with a big part of the ear canal missing. The missing part has been recovered with both the MRF method and the proposed method. As can be seen in the scan our proposed method produces what seems to be a much more plausible surface in the missing part. Figure 9.15 shows an ear with even more data missing of the ear canal as well as a missing part of concha. Again the noise in the data is handled well in the statistical surface recovery.



(a) MRF (b) Proposed

Figure 9.14: An ear scan with large parts of the ear canal missing and a hole in concha, the missing data has been recovered with both MRF and our proposed method.



(a) MRF (b) Proposed

Figure 9.15: An ear scan with a large part of the tragus missing, the missing data has been recovered with both MRF and our proposed method. The inside of the ear canal has large fragments, which could come from hair, wax or general scanner noise.

9.6 Discussion

We have shown that we can predict the missing parts of partial scans using a statistical model. The ability to predict missing data is comparable to state-of-the-art algorithms, when holes are relatively small and the data is fair without too much noise. This has been quantified in an experiment with synthesised holes in samples of laser scanned molds. The holes have been placed in an area known to be difficult to scan.

On scans from a real in-ear scanner probe prototype, the qualitative results produced with the proposed method are much more plausible when visually inspected. The more extensive prior knowledge about the shape to be reconstructed makes the recovery much more robust, when recovering larger holes. The results also seem invariant to the presence of noise, and as such the method can also function as a noise filter. Surface reconstruction algorithms that only use the immediate vicinity in the reconstruction are very sensible to noise on the edges of the area to be recovered.

After using the proposed data recovery method on 12 scans they were 3D printed on a stereolithograpy (SLA) machine [1] and worn by the test subjects for a substantial time. They all proved to be well fitting in the subjects ears even though the hard material from the SLA machine makes the ear plugs very susceptible for non-accurate fitting. We have therefore demonstrated a complete pipeline from direct ear scanning to production of well fitting hearing devices.

The performance of the proposed method is from a usability viewpoint fair. While the preprocessing steps of creating the ASM is time consuming the data recovery step of partial scans produces the results in a few minutes in our Matlab implementation. The preprocessing step should only be performed once or potentially repeated when even more ear canal data has been acquired.

For further work a few points should be addressed. When building the ASM some scans get rejected automatically due to poor alignment. We have not investigated if the rejects are from a subpopulation and therefore our ASM is biased in not including these. During the manual quality check, the rejected shapes are mainly having a registration problem on the very edge of the part of concha included in our model. These are much less likely to constitute a subpopulation of the total set and therefore less likely to bias the final model in their absence.

Regarding the fitting process, we experienced that the Nelder-Mead optimiser in some scenarios produced a somewhat sparse model-synthesisation. An alternative approach could be a fusion with the ridge regression method proposed

by [13] and is thus a subject of future research.

As such the method is in no way restricted to the used data. It should be applicable to data, where a consistent training set can be acquired, where the shape variation can be meaningfully modelled, and where point correspondence throughout the data set can be obtained. Furthermore, the method can also be extended to predict the missing colour values in textured surface as for example facial scans.

Chapter 10

Multiple View Stereo by Reflectance Modeling

Sujung Kim, Seong D. Kim, Anders L. Dahl, Knut Conradsen, Rasmus R. Jensen, and Henrik Aanæs

Abstract

Multiple view stereo is typically formulated as an optimization problem over a data term and a prior term. The data term is based on the consistency of images projected on a hypothesized surface. This consistency is based on a measure denoted a *visual metric*, e.g. normalized cross correlation. Here we argue that a visual metric based on a surface reflectance model should be founded on more observations than the degrees of freedom (dof) of the reflectance model. If (partly) specular surfaces are to be handled, this implies a model with at least two dof. In this paper, we propose to construct visual metrics of more than one dof using the DAISY methodology, which compares favorably to the state of the art in the experiments carried out. These experiments are based on a novel data set of eight scenes with diffuse and specular surfaces and accompanying ground truth. The performance of six different visual metrics based on the DAISY framework is investigated experimentally, addressing whether a visual metric should be aggregated from a set of minimal images, which dof is best, or whether a combination of one and two dof should be used. Which metric performs best is dependent on the viewed scene, although there are clear tendencies for the two dof minimal metric to be the preferred one.

10.1 Introduction

Multiple view stereo or the dense 3D reconstruction of the surface of an object from multiple calibrated images is one of the persistent central challenges of computer vision. This paper addresses this challenge by investigating image similarity measures – *the visual metrics* for surfaces with light reflectance properties that contain both specular and diffuse components.

A massive effort has recently been put in multiple view stereo, and advances have been achieved by recent benchmark datasets like the Middlebury multi-view stereo sets [87], the dense multi-view stereo of buildings from Strecha et al. [95], as well as works on large scale urban reconstruction of Furukawa et al. [32] and Gallup et al. [34]. Many recent landmark achievements [23, 33, 34, 36, 79, 102, 103, 108] have been obtained. These recent efforts have mainly focused on methods for optimization and regularization. The visual metrics used have been sums of squared differences (SSD) or normalized cross correlation (NCC) between image pairs. These visual metrics are well suited for diffuse reflecting surfaces, where the surface appearance is independent of the viewing direction, but not for more complex reflecting surfaces with specularities. Both the Middlebury datasets [87] and the buildings from [95] consist of diffuse objects, and therefore fit well with the simple visual metrics such as SSD and NCC.

Many real world objects are not well modeled as diffuse reflecting. Multiple view stereo algorithms can, however, handle a lot of these objects using NCC or SSD by robust statistics and an abundance of images. Such an abundance is, however, often not possible or practical, and in these cases, the SSD and NCC based frameworks brake down, and a more elaborate visual metric is needed.

In [48, 94], it is shown that visual metrics dealing with objects with more complex reflectance properties, e.g. specular, cannot be based on comparing image pairs. In this paper, we further this work and propose novel visual metrics based on modeling the reflectance. To do this the number of images should exceed the dof of the reflectance model, which is one in the diffuse case. Based on this realization, we investigate how to construct visual metrics dealing with diffuse *and* specular objects, and thus reflectance models with more than one dof. This results in a visual metric with better properties than the radiance tensor of Jin et al. [48].

Considerable evidence exists to support that the SIFT framework is superior to NCC when dealing with salient feature point matching [25, 66]. This has been exploited by Tola et al. [97] in the stereo case by changing the binning of the descriptors to the output of Gaussian filters, whereby the computations could be performed more efficiently resulting in the DAISY descriptor. To deal with

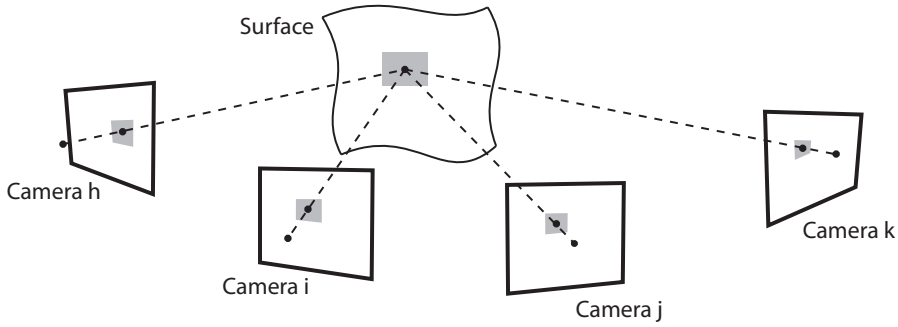


Figure 10.1: Illustration of the relationship between a surface patch and corresponding image patches. Visual metrics typically evaluate the support for the presence of a given surface patch in the data by comparing these image patches

more than two images, the DAISY framework combines the scores of image pair matchings as commonly done with NCC. As part of the investigation we also propose how to construct a visual metric with more than one dof by constructing a tensor of DAISY descriptors. We experimentally show that the DAISY tensor is superior.

The investigation of the proposed visual metric is based on a new data set of eight different scenes with diffuse and specular objects. This data is accompanied by ground truth obtained by a structured light scanner[1]. Firstly we demonstrate that the DAISY tensor is to be preferred to raw pixels because it is more robust and approximates the ground truth better. Following this, we investigate the difference between using one or two dof and a combination of the two. As for the latter, if the part of a scene is diffuse, then extra dof could lead to overfitting so that it might cause possible performance degradation. Lastly, we investigate if a visual metric should be aggregated from a minimal set of images, i.e. two in the diffuse case and three for the proposed visual metric, or directly based on all relevant images. The performed experiments are done in a 2.5D manner via the alpha expansion of [16]. This is a relatively simple reconstruction algorithm, which we deliberately have chosen over the state of the art algorithms because our focus is on the visual metric. If we chose an algorithm with stronger modeling capabilities, this could clutter the effect of the visual metric. Choosing an algorithm that does not use contextual information might, however, not reveal the potential of the visual metric in a realistic setup. We found the choice of the graph cut algorithm [16] a good tradeoff.

[1]This data set is available at *http://roboimagedata.imm.dtu.dk/reflectance/*.

In this paper, we investigate the similarity metric similar to the work of [43], but we focus on multiple views opposed to stereo in their work. An in depth review of the multiple view stereo literature and introduction of this field can be found in [87, 20].

10.2 Visual Metrics

Multiple view stereo deals with estimating the 3D surface of an object or a scene from multiple images with the known camera calibrations. These known calibrations allow us to compute where a given 3D point is projected in the images, cf. Figure 10.1. Multiple view stereo is typically handled as an optimization problem, where we want to find the surface $\mathcal{S}$, which is most consistent with the images. Normally a prior is added. This prior is often formulated as a smoothing. The image consistency is formulated as a visual metric, $\mathrm{V}(\mathbf{x}, \mathbf{n})$, evaluated at each point $\mathbf{x}$ on the surface with normal $\mathbf{n}$. The optimization problem thus becomes[2]

$$\min_{\mathcal{S}} \sum_{\mathbf{x} \in \mathcal{S}} \mathrm{V}(\mathbf{x}, \mathbf{n}(\mathbf{x})) \ + \mathrm{Prior}(\mathcal{S}) \ , \tag{10.1}$$

where $\mathbf{n}(\mathbf{x})$ is the surface normal at $\mathbf{x}$. The visual metric is based on a planar patch at $\mathbf{x}$ with normal $\mathbf{n}$, whereupon the relevant images are projected as illustrated in Figure 10.1. Different visual metrics then employ different measures to quantify the consistency. A typical example is the use of NCC between pairs of projected patches, e.g. projections from cameras i and j in Figure 10.1. It is the construction of these consistency measures we investigate further in this paper.

10.2.1 The Radiance Tensor

In Jin et al. [48], a visual metric is constructed via a radiance tensor. For a given surface patch described by $\mathbf{x}$ and $\mathbf{n}$, this radiance tensor is constructed by firstly enumerating the relevant, visible, images by $i \in \{1, \ldots, n\}$. Denote the m pixel intensities of the associated projected patches as $\mathbf{r}_i(\mathbf{x}, \mathbf{n})$, where $\mathbf{r}_i(\mathbf{x}, \mathbf{n})$ is an m dimensional vector. These vectors are then combined into the $m \times n$ radiance tensor

$$\mathbf{R}(\mathbf{x}, \mathbf{n}) = \begin{bmatrix} \mathbf{r}_1(\mathbf{x}, \mathbf{n}) & \mathbf{r}_2(\mathbf{x}, \mathbf{n}) & \cdots & \mathbf{r}_n(\mathbf{x}, \mathbf{n}) \end{bmatrix} \ . \tag{10.2}$$

In the ideal case where the patch coincides with the surface and no other noise is present either, a patch should look the same from all directions, up to scale,

[2] This is typically formulated as an integral over $\mathcal{S}$, which is then later discretized.

in the diffuse case. In this ideal case, all $\mathbf{r}_i(\mathbf{x}, \mathbf{n})$ should thus be scaled versions of each other, and the rank of $\mathbf{R}(\mathbf{x}, \mathbf{n})$ becomes one. A main result of [48] is that if the reflectance model of a surface is described by the diffuse plus specular Phong model, then the rank of $\mathbf{R}(\mathbf{x}, \mathbf{n})$ should be two in the ideal case.

In the rank two case of [48], the singular values[3] of $\mathbf{R}(\mathbf{x}, \mathbf{n})$, $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$, form the basis of a visual metric. Given a patch on the true surface in the ideal case, only the first two singular values σ_1 and σ_2 should be non-zero. This corresponds to $\mathbf{R}(\mathbf{x}, \mathbf{n})$ having rank two. The visual metric, $J(\mathbf{x}, \mathbf{n})$, from [48] is thus

$$J(\mathbf{x}, \mathbf{n}) = \sum_{i=3}^{n} \sigma_i^2 \ , \tag{10.3}$$

which is equal to the total variation of the noise for the patch being on the true surface. A similar visual metric corresponding to a diffuse model would similarly be [4]

$$\sum_{i=2}^{n} \sigma_i^2 \ . \tag{10.4}$$

10.2.2 Visual Metric as Model Fitting Residual

An interpretation of the visual metric in (10.3) is that a linear subspace is fitted to the data, i.e. the $\mathbf{r}_i(\mathbf{x}, \mathbf{n})$, and that the visual metric is put equal to the squared residual error. This linear subspace has dimension two, corresponding the the models dof. The same interpretation can be made of (10.4) except that a 1D subspace is fitted. Similarly, the cross-correlation, ρ_{ij}, between $\mathbf{r}_i(\mathbf{x}, \mathbf{n})$ and $\mathbf{r}_j(\mathbf{x}, \mathbf{n})$ is the best fit of the model

$$\left\| \frac{\mathbf{r}_i(\mathbf{x}, \mathbf{n}) - \mu_i}{\|\mathbf{r}_i(\mathbf{x}, \mathbf{n}) - \mu_i\|} - \alpha \frac{\mathbf{r}_j(\mathbf{x}, \mathbf{n}) - \mu_j}{\|\mathbf{r}_j(\mathbf{x}, \mathbf{n}) - \mu_j\|} \right\|_2^2 \ , \tag{10.5}$$

where μ_i is the mean of $\mathbf{r}_i(\mathbf{x}, \mathbf{n})$ – i.e. $\alpha^* = \rho_{ij}$. The residual error is $1 - \rho_{ij}^2$. The NCC can thus also be interpreted as residual error after fitting a one parameter model.

An implication of viewing a visual metric as a model fitting residual is that we need more observations, i.e. $|n|$, than the dof of the underlying reflectance model. If not, the residual, and thus the visual metric, will always be zero. Thus, the diffuse model works well with only two observations, $(n = 2)$, since it has one dof.

[3] In general $n < m$, and there is thus n singular values of $\mathbf{R}(\mathbf{x}, \mathbf{n})$.

[4] Note the starting index of the summation.

The model fitting residual interpretation does not need to be possible for all conceivable visual metrics. However, given a reflectance model, then its dof is equal to the dimension of the possible ways in which a surface patch can change appearance between image views in general. Thus, at least one more image observation, $\mathbf{r}_i(\mathbf{x}, \mathbf{n})$, than the dof is needed. This again implies that if a visual metric is based on a reflectance model, then it needs to be based on at least one plus the dof observations.

The conclusions reached here are generalizations of [94], which is based on more formal arguments. Note also that visual metrics are often made invariant under different actions, e.g. rotation in the SIFT descriptor [66]. Such an invariance removes an effect instead of modeling it, and as such it does not increase the dof.

10.3 Visual Metrics for Specular Surfaces

Specular surfaces are best described by a two or larger dof reflectance model. So based on the above reasoning, we wish to investigate how we may best construct visual metrics of more than the usual one dof. First of all, we propose an extension of the SIFT methodology to the two or larger dof case via a DAISY tensor.

10.3.1 DAISY Tensor

A DAISY descriptor [97] of a gray scale image is computed from orientated image derivatives. These derivatives are convolved by Gaussian kernels and the filter output form the entries of a DAISY descriptor vector $\mathbf{d}_i(\mathbf{x}, \mathbf{n})$. We propose forming a tensor of the relevant DAISY descriptors, described by 3D point $\mathbf{x}$ and normal $\mathbf{n}$ as in line with (10.2)

$$\mathbf{D}(\mathbf{x}, \mathbf{n}) = \begin{bmatrix} \mathbf{d}_1(\mathbf{x}, \mathbf{n}) & \mathbf{d}_2(\mathbf{x}, \mathbf{n}) & \cdots & \mathbf{d}_n(\mathbf{x}, \mathbf{n}) \end{bmatrix} . \tag{10.6}$$

Let the singular values of $\mathbf{D}(\mathbf{x}, \mathbf{n})$ be given by $\{\varsigma_1, \varsigma_2, \ldots, \varsigma_n\}$. Then we can form visual metrics as[5]

$$D_1(\mathbf{x}, \mathbf{n}) = \sum_{i=2}^{n} \varsigma_i^2 \tag{10.7}$$

$$D_2(\mathbf{x}, \mathbf{n}) = \sum_{i=3}^{n} \varsigma_i^2 . \tag{10.8}$$

[5]Note the starting indices of the summations

10.3.2 Further Lines of Investigation

In line with findings for two view stereo [96] and salient features [25], our experiments show that the DAISY tensor outperforms the radiance tensor as a basis for a visual metric. Likewise, we only consider linear subspaces of a given degree as representatives of models of a given dof.

10.3.2.1 Minimal vs. All

For salient features, matching performance is increased for smaller differences in viewing angle between images. It is partly because the approximation of the planar patch assumption becomes less profound. It is thus relevant to ponder whether visual metrics should be based on aggregations of *minimal* sets of images, as done with NCC in [102], or if *all* relevant images should be used at once as done in [48], cf. (10.3). Using all relevant images at once increases the redundancy in the data giving bigger noise reduction. Also the larger difference in viewing angle will generally give a better baseline to depth ratio, and thus better depth estimation, cf. [35]. To shed light on this matter, we compare the two alternatives experimentally.

The visual metrics directly using all relevant images are given by (10.7) and (10.8). The size of the minimal sets is one plus the dof of the model since there needs to be a residual. In the two dof case, we denote these sets $\{i, j, k\} \in \mathcal{C}_3$. The visual metric is then aggregated from the squared third singular value of

$$\begin{bmatrix} \mathbf{d}_i(\mathbf{x}, \mathbf{n}) & \mathbf{d}_j(\mathbf{x}, \mathbf{n}) & \mathbf{d}_k(\mathbf{x}, \mathbf{n}) \end{bmatrix} \ , \tag{10.9}$$

which we denote $\Gamma^3_{ijk}(\mathbf{x}, \mathbf{n})$, i.e.

$$\begin{aligned} &\Gamma^3_{ijk}(\mathbf{x}, \mathbf{n}) = \varsigma_3^2 = \\ &\min_{\mathbf{v}_1, \mathbf{v}_2} \sum_{m \in \{i,j,k\}} \left\| \mathbf{d}_m(\mathbf{x}, \mathbf{n}) - [\mathbf{v}_1 \mathbf{v}_2][\mathbf{v}_1 \mathbf{v}_2]^T \mathbf{d}_m(\mathbf{x}, \mathbf{n}) \right\|_2^2 \ , \end{aligned} \tag{10.10}$$

where $\mathbf{v}_1, \mathbf{v}_2$ is an orthonormal basis of a 2D linear subspace. The two dof minimal visual metric considered here is then given by

$$M_2(\mathbf{x}, \mathbf{n}) = \sum_{\{i,j,k\} \in \mathcal{C}_3} \Gamma^3_{ijk}(\mathbf{x}, \mathbf{n}) \ , \tag{10.11}$$

which is the sum of ς_3^2 for all relevant image triplets. In an analog fashion, the one dof minimal visual metric is given by

$$M_1(\mathbf{x}, \mathbf{n}) = \sum_{\{i,j\} \in \mathcal{C}_2} \Gamma^2_{ij}(\mathbf{x}, \mathbf{n}) \ . \tag{10.12}$$

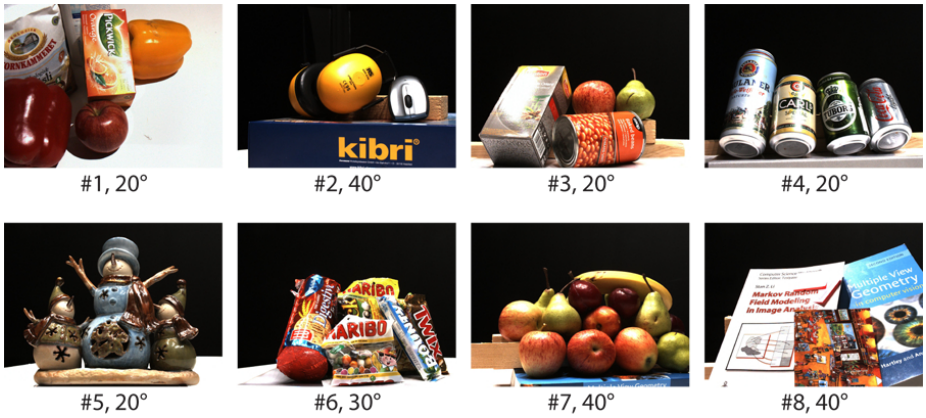


Figure 10.2: The scenes of our investigation numbered #1 - #8. The numbers after the comma indicate the baseline in degrees.

10.3.2.2 Model Averaging

Although two dof visual metrics are superior when dealing with specular surfaces, one dof visual metrics suffice for diffuse surfaces. In the latter case, a two dof visual metric would possibly overfit leading to performance loss. A visual metric averaging the one and two dof models is also investigated. We propose an additional pair of visual metrics

$$\begin{aligned} D_{1.5}(\mathbf{x},\mathbf{n}) &= \frac{1}{2}D_2(\mathbf{x},\mathbf{n}) + \frac{1}{2}D_1(\mathbf{x},\mathbf{n}) \\ &= \frac{1}{2}\varsigma_2^2 + \sum_{i=3}^{n}\varsigma_i^2 \qquad (10.13) \\ M_{1.5}(\mathbf{x},\mathbf{n}) &= \sum_{\{i,j,k\}\in\mathcal{C}_3}\Gamma_{ijk}^{2.5}(\mathbf{x},\mathbf{n}) \qquad (10.14) \\ \text{where} \quad & \Gamma_{ijk}^{2.5}(\mathbf{x},\mathbf{n}) = \frac{1}{2}\varsigma_2^2 + \varsigma_3^2 \ . \end{aligned}$$

10.3.2.3 Investigated Visual Metrics

In summary, our investigation is based on eight visual metrics. Two are based on the raw pixel intensities J, (10.3), with two different patch sizes. Six are based on the DAISY tensor, i.e. D_1, $D_{1.5}$, D_2, M_1, $M_{1.5}$, and M_2, investigating the combined possibilities of

- If one dof, two dof or an averaged alternative should be used.
- If the visual metric should be based directly on all relevant images or on a combination of minimal subsets.

10.4 Experimental Results

To perform multiple view stereo experiments on objects with specular and diffuse surface reflectance models, we compiled a new data set consisting of eight different scenes as shown in Figure 10.2. The scenes show specular reflectances and have planar to non-planar surfaces. We chose to vary the baseline of the different data sets to ensure significant specularities to challenge the visual metric. This is done by visual inspection. The number of images was kept constant at five and the maximum angles between images of the eight scenes were: #1 – 20°, #2 – 40°, #3 – 20°, #4 – 20°, #5 – 20°, #6 – 30°, #7 – 40°, #8 – 40°. These angles are an indication of the baselines used, and are listed in Figure 10.2.

The recorded images have a spatial resolution of 1200×1600 pixels recorded as 8 bit RGB converted to gray scale. The data set was recorded with an industrial robot arm using a setup similar to [4, 87]. We have, however, mounted the structured light scanner on the robot arm holding the camera. Hereby the ground truth 3D point-set was perfectly aligned with the camera position and provides a good coverage of the scenes. This enabled us to evaluate multiple view stereo algorithms by measuring the distance from the ground truth points of the structured light scan to the multi view reconstruction.

The average reconstruction errors and standard deviations are shown in Table 10.1 and the graph of averge reconstruction errors is illustrated in Figure 10.3. One reconstruction example is shown in Figure 10.4. The reconstruction errors were computed by taking the absolute difference between the estimated depths and ground truth for each pixel, but only where there were ground truth measurements.

To get a clearer picture of the performance of the visual metrics, we have solved the multiple view stereo reconstruction optimization problem (10.1) via the alpha-expansion algorithm of Boykov et al. [16], which is a very well understood optimization algorithm. For the same reason we have also avoided iterating over a visibility mask as done in [97]. In this way, we avoid complicating factors that impair the evaluation of the visual metrics.

The algorithm of [16] works by finding an optimal depth for each pixel in a reference image, where the depth is taken from a discrete set of ordered depth

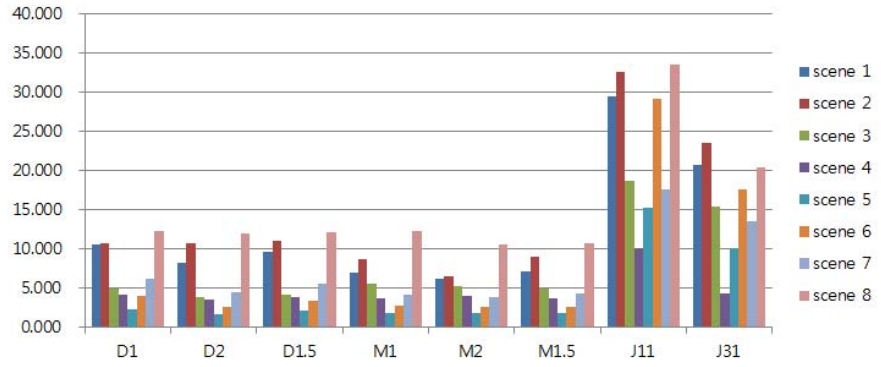


Figure 10.3: Graph of mean errors in Table 10.1. It shows the average reconstruction errors (in mm) on the vertical line and eight metrics on the horizontal line with eight scenes with different color. Note that DAISY based visual metrics are superior to the raw based visual metrics by a large margin, and M_2 is slightly better than other DAISY based metrics. The effect of subtle differences among DAISY based visual metrics can be seen in Figure 10.4.

values. The depth resolutions used for the different scenes are determined by range of the ground truth data points and divided into equal sized steps of approximately 1 mm. This resulted in between 110 and 180 discrete steps in the different scenes.

We evaluate six different DAISY based visual metrics. The $\mathbf{d}_i(\mathbf{x}, \mathbf{n})$ is computed similarly to the DAISY descriptor in [97]. We compute the descriptor on a 31×31 pixels[6] image patch with three spatial sampling rings of six positions resulting in 19 spatial sampling positions. At each position the eight smoothed signed derivatives are sampled resulting in a 152 dimensional descriptor. The smoothing factor of the center point and first ring is $\sigma = 3$, for the second ring $\sigma = 5.5$, and for the third ring $\sigma = 8$. The raw based visual metrics evaluated are $J(\mathbf{x}, \mathbf{n})$ from (10.3) with a patch size of 11×11 and 31×31. The first is chosen because it is the recommendation by Jin et al. [48], the second is chosen in order to have the same terms as the DAISY based visual metrics. In the following we denote these two visual metrics as $J^{11}(\mathbf{x}, \mathbf{n})$ and $J^{31}(\mathbf{x}, \mathbf{n})$ respectively.

A summary of experimental results is shown in Table 10.1, Figure 10.3 and Figure 10.4. From the quantified errors in Table 10.1 several things can be

[6] In this case $m = 31 \times 31 = 961$.

concluded. Firstly, the DAISY based visual metrics outperform the raw based visual metrics, $J^{11}(\mathbf{x}, \mathbf{n})$ and $J^{31}(\mathbf{x}, \mathbf{n})$, by a large margin. This is clear evidence that a DAISY based visual metric should be preferred supporting the findings of [97]. Also $J^{31}(\mathbf{x}, \mathbf{n})$ consistently outperforms $J^{11}(\mathbf{x}, \mathbf{n})$.

We also note that the best performing descriptor varies between the two 2-dof DAISY descriptors, D_2 and M_2, and the D_2 favors the data sets with the small baselines. This indicates that the minimal cases are better at dealing with perspective distortion, and this is more important than a good depth to baseline ratio.

$V(\mathbf{x},\mathbf{n})$	Scene #1		Scene #2		Scene #3		Scene #4	
	mean	std.	mean	std.	mean	std.	mean	std.
$D_1(\mathbf{x},\mathbf{n})$	10.67	22.55	10.79	17.43	4.99	13.59	4.18	4.69
$D_2(\mathbf{x},\mathbf{n})$	8.34	19.28	10.77	19.14	**3.82**	10.64	**3.59**	3.90
$D_{1.5}(\mathbf{x},\mathbf{n})$	9.70	21.68	11.07	18.83	4.25	12.75	3.91	4.23
$M_1(\mathbf{x},\mathbf{n})$	7.00	15.93	8.69	15.36	5.56	13.86	3.80	4.31
$M_2(\mathbf{x},\mathbf{n})$	**6.24**	14.08	**6.46**	12.47	5.35	13.52	4.01	4.78
$M_{1.5}(\mathbf{x},\mathbf{n})$	7.12	15.84	9.01	15.98	5.00	12.41	3.74	3.79
$J^{11}(\mathbf{x},\mathbf{n})$	29.45	42.91	32.67	33.93	18.79	31.11	10.16	21.78
$J^{31}(\mathbf{x},\mathbf{n})$	20.84	38.05	23.57	30.80	15.39	29.19	4.40	12.40

$V(\mathbf{x},\mathbf{n})$	Scene #5		Scene #6		Scene #7		Scene #8	
	mean	std.	mean	std.	mean	std.	mean	std.
$D_1(\mathbf{x},\mathbf{n})$	2.32	4.51	4.06	12.56	6.22	13.31	12.36	29.94
$D_2(\mathbf{x},\mathbf{n})$	**1.77**	3.28	2.68	6.45	4.53	10.99	11.94	29.87
$D_{1.5}(\mathbf{x},\mathbf{n})$	2.16	4.44	3.44	10.42	5.62	12.56	12.23	30.02
$M_1(\mathbf{x},\mathbf{n})$	1.90	2.61	2.76	6.15	4.22	9.16	12.30	29.87
$M_2(\mathbf{x},\mathbf{n})$	1.83	2.38	**2.65**	5.78	**3.90**	9.24	**10.59**	27.77
$M_{1.5}(\mathbf{x},\mathbf{n})$	1.91	2.90	2.70	6.36	4.41	9.44	10.76	27.77
$J^{11}(\mathbf{x},\mathbf{n})$	15.31	26.34	29.22	44.75	17.62	31.70	33.51	46.00
$J^{31}(\mathbf{x},\mathbf{n})$	10.10	22.09	17.65	36.90	13.62	26.66	20.51	38.79

Table 10.1: Average reconstruction errors and standard deviation (in mm) for the eight visual metrics and eight scenes. Note that the reported standard deviation is for the errors and not for the mean. If we assume a few hundred *independent* observations, the main differences between the means are significant. The ground truth consists of about 300.000 correlated observations, so a few hundred independent observations seems a reasonable assumption. The fact that the stdandard deviation is larger than the mean is a consequence of the reconstruction errors following a very skew distribution with a very fat tail in the direction of large errors. For each scene, the best mean value is denoted by **bold face**.

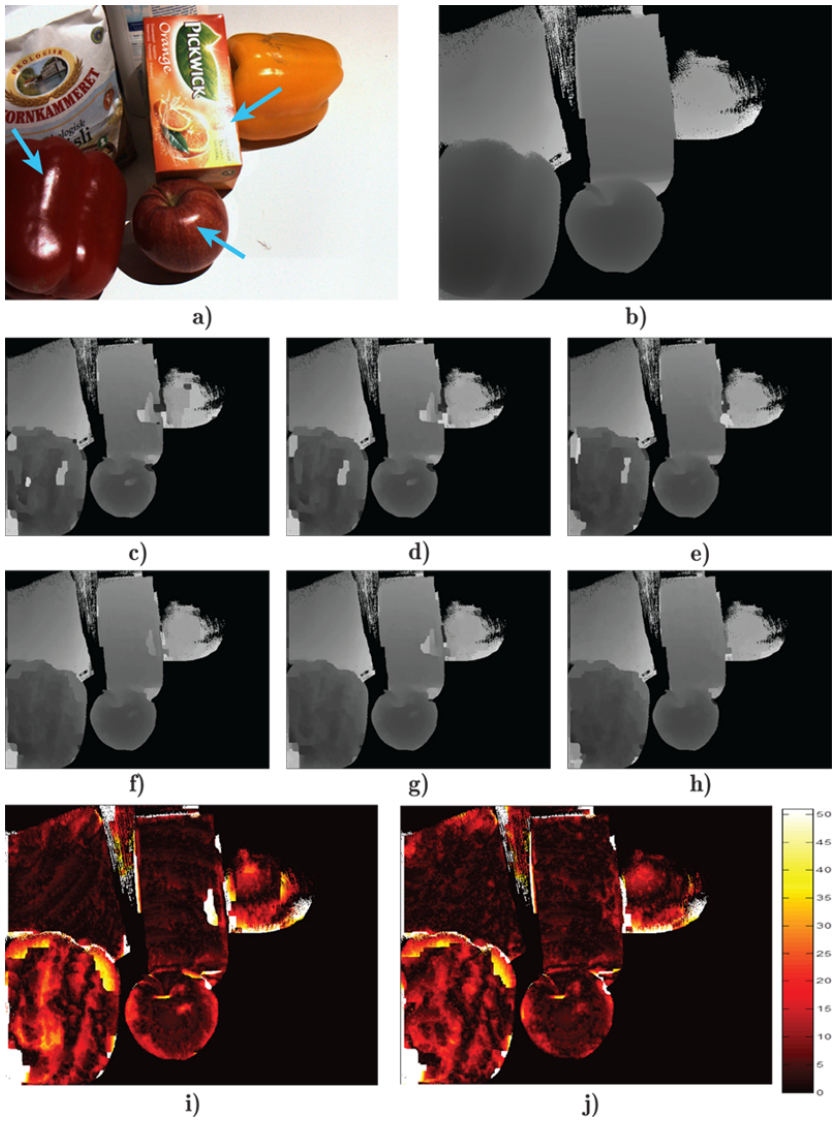


Figure 10.4: The reconstruction results of Scene #1, wrt. the DAISY based visual metrics. The figures illustrate **a)** The sample input image with blue arrows marking the distinct specularities to notice in the results. **b)** The ground truth. **c) - h)** reconstructed depth maps by the following visual metrics D_1, $D_{1.5}$, D_2, M_1, $M_{1.5}$, and M_2 respectively. **i)** and **j)** the reconstruction errors of M_1 and M_2 respectively, i.e. f) and h) minus b). Note the differences around the specularities.

10.5 Perspective and Conclusion

In this paper, we have linked surface reflectance models with the visual metrics used for multiple view stereo. We conclude that we need more observations for a visual metric than the dof of an underlying reflectance model. Thus, more than two observations are needed to handle (partly) specular objects. We proceeded by proposing a method for including more than two images or observations into a visual metric, while incorporating the DAISY framework. This proved superior to directly using raw pixels regarding the ability to approximate the ground truth of our data. This is consistent with findings for salient feature matching [25] and two view stereo [97].

We have also put forth a new multiple view data set with ground truth, which spans different reflectance models better than any available data set we are aware of. This data set is the basis of our experimental evaluations. The evaluations, first of all, consider the dof of the underlying reflectance model. Our experiments also address whether the visual metric should be aggregated from a minimal set of images, as done with NCC in [17], or if all relevant images should be used directly as in [48]. Our experimental results show that the use of two dof is favorable. The choice between all or the minimal case seems to depend on the baseline – with a small baseline favoring using all images. As argued in the introduction, elaborate visual metrics are mostly needed for limited image budget, and thus large baselines, favoring M_2.

Since the state of the art in visual metrics [97, 48] is also represented in the visual metrics we investigated, the M_2 proposed here looks like a strong choice for a visual metric in relation to multiple view stereo. To further argue the matter in relation to robustness, e.g. occlusions, some of the current good choices of addressing this [17, 102] use minimal cases, and thus our M_2 visual metric should be usable in these robust frameworks.

Our findings favor basing visual metrics on underlying surface reflectance models. This opens the new interesting question of how these models should be formulated. In this work, we have limited these reflectance models to be linear subspaces to avoid a combinatorial explosion since we already compared eight visual metrics. It is, however, likely that other models, e.g. more physical based models comprising nonlinear manifolds, would perform better. In this regard the work of [19] is inspirational. Also it is likely that probabilistic models of the reflectance should be formulated, but this would require much more than eight scenes.

Chapter 11

On the Performance of Calibrated Multiple View Stereopsis

Rasmus R. Jensen, George Vogiatzis, Anders L. Dahl, and Henrik Aanæs

Abstract

Calibrated multiple view stereopsis has become very accurate for a range of complicated objects, making this approach an alternative technique for surface recovery, compared to active light methods. The quality of the obtained surface reconstructions is, however, dependent on the reflectance properties of the objects to be reconstructed. Especially surface texture can aid the reconstruction whereas specularities and repeated textures have been reported to corrupt it. Assessing the quality of a reconstructed surface based on calibrated images can be hard, but it can be vital to know if a reconstruction is reliable, e.g. when using it for metrology. Our investigation shows, that performance is strongly related to the presence of texture. Specularities or repeated texture patterns severely corrupts two-view reconstructions, but it does not pose problems for the multiple views. We test a simple image texture characterisation and demonstrate how it correlates to the performance. Our study is based on a large multiple view dataset containing 80 scenes acquired from a setup based on a six-axis industrial robot. Accurate reference surface reconstructions are obtained from scans based on stereopsis with structured light support. In addition to predicting the reconstruction quality this study highlights the limitations for current state of the art surface reconstruction. Our data will be made available online.

Figure 11.1: Example images from eight scenes illustrating the variability in geometry, reflectance, and texture.

11.1 Introduction

Current multiple view stereo (MVS) reconstruction methods perform impressively for a range of highly complicated and detailed scenes [33, 102, 98]. Images originating from such scenes are typically rich in texture, which is central for successful reconstruction. Simple shaped objects can, however, be very hard to reconstruct, because they lack the necessary detail to find correspondence between images. Especially lack of texture makes it close to impossible to obtain precise 3D representations even for the simplest shapes like spheres or cubes. This counter-intuitive relation between object complexity and performance of MVS methods makes it difficult to predict, if a reconstruction is reliable or not.

Applications of 3D models range from film and video game industry [53] to quantitative metrology used in e.g. material and geo science [81]. In these and other applications high quality is essential, but the question is, if a given model is a precise reconstruction of an object or a scene. Quality cues are given by the texture and surface reflectance properties of the object or scene, because these influence the ability of solving the correspondence problem. In this paper we show that texture can be quantified from the scene images to predict the performance of state-of-the-art reconstruction methods. In addition to providing a measure of reliability, our investigation also reveals current challenges within MVS surface reconstruction.

Relating image texture cues to surface reconstruction requires data that span the expected complexity with regard to geometry, specularity, and texture. Popular datasets for evaluating multiple view stereopsis include the Middlebury Multi-View dataset [87] and the Multi-View Stereo dataset of [95]. Current state-of-

the-art reconstruction algorithms are able to reconstruct 3D models of these objects and scenes very accurately[1], with most reconstructed points within 0.4 - 0.5 mm from the reference surface. These datasets are chosen such that they contain both sharp and smooth features, complex topology, strong concavities, and both strongly and weakly textured surfaces – all aspects that challenge MVS algorithms [87]. Their reflectance properties are, however, limited to Lambertian or close to Lambertian reflectance. In our investigation we want to include higher degrees of specularity as well as a much larger range of materials with various shapes, textures, and reflectance properties. We have therefore collected a dataset to fulfil this purpose – example images are shown in Figure 11.1.

Our MVS dataset contains 80 scenes with 49 or 64 camera positions from each scene. The data has been collected using a six-axis industrial robot. The 3D surface references are obtained from a structured light scanner mounted on the robot arm. Hereby we obtain high density and completeness, because the light scanner provides a depth map from each camera position. Based on this data we have reconstructed 3D surfaces using the MVS stereo algorithms of Campbell et al. [17] and Furukawa and Ponce [33]. In addition, we use a simple two view version of [17]. We omitted the last surface reconstruction steps, including outlier rejection, regularisation, surface smoothing, etc., of these algorithms, such that the evaluated output was point clouds.This was done to isolate the raw data term from these other effects. The obtained point clouds were evaluated based on our structured light reference. The found results were related to image texture measures to assess performance in relation to texture. Our dataset will be available online[2].

11.2 Related work

The first work that attempted to benchmark MVS algorithms was [87] where the performance of six algorithms was measured across six different scenes. The authors subsequently invited submissions of reconstruction results from dozens of different algorithms that were publicly ranked against each other. The somewhat artificial, low-resolution setup of [87] was subsequently improved in the evaluation effort of [95] that consisted of high-resolution images of outdoor scenes. Both [87] and [95] made an invaluable contribution to the advancement of MVS technologies by providing a solid platform on which improvement to existing state-of-the-art could be measured and recorded.

Our work contributes to the evaluation of MVS, albeit with a different focus. In

[1] http://vision.middlebury.edu/mview/eval/
[2] http://www.homepage.to.come

[87, 95] the evaluators' basic question was "which MVS algorithm works best for *this scene*?" In our work we ask the question "what types of scenes work best for *this MVS algorithm*?" Even though superficially the two questions seem related, in practice they require two very different experimental setups and give rise to different types of analysis. The evaluations of [87, 95] consider a small number of 3D scenes that are thought to be representative of real-world application domains for MVS. In practice they choose well-textured diffuse-reflectance 3D objects on which MVS algorithms tend to perform quite well. They then apply several algorithms in order to create a performance ranking for each scene. Our approach is to consider the widest possible range of 3D scenes one might encounter in real applications, and then consider how particular types of MVS algorithm perform on each type of scene. This approach sheds light on the performance of MVS technology as a whole and its overall suitability for particular applications.

Most successful MVS algorithms fall under two main categories: point-cloud based (e.g. [98, 37, 103, 33, 102]) and volume-based methods (e.g. [65, 55, 42]). Volume-based methods aggregate photo-consistency data in a 3D volume and compute a 3D surface within that volume using surface optimisation. On the other hand point-cloud based methods convert photo-consistency data into a 3D point-cloud, which is then converted into a 3D surface using standard meshing techniques such as Poisson Reconstruction [51], Graph-cuts [102] or signed distance functions [70]. In this work we focus on point-cloud based methods because we can easily isolate the point-cloud stage from the surface extraction stage and all the filtering and regularisation this entails.

Within point-cloud based methods we can distinguish two different paradigms. Feature expansion [33] and depth-map fusion [98, 17, 37, 103, 102]. Under the feature expansion paradigm the algorithm starts from a set of 3D features in the scene, which then expand into nearby 3D points while outliers are filtered using occlusion reasoning. Depth-map fusion works by computing independent depth-maps for each image using neighbouring images. These depth-maps are then merged into a single point-cloud. We chose [33] and [17] as representative algorithms from the feature expansion and depth-map fusion families. It must be stressed again that our aim is not to directly compare the two methods or the two families of algorithms. Rather, by running these methods on a large selection of datasets we highlight the effect on performance of different types of 3D scene.

Perhaps closer in spirit to the present work are some previous attempts at investigating in detail, aspects of MVS performance. In [52] there is a theoretical analysis of the impact of scene geometry on feature-expansion MVS methods. A serious evaluation of MVS algorithms based on depth map fusion is presented in [44]. Our work can be seen as an empirical analysis of both families of MVS

algorithms.

A recent trend in MVS research has been to automate all aspects of the MVS pipeline, including viewpoint selection and image capture. For example, in [32, 7] MVS is applied to photographs of famous landmarks, harvested from online photo-collections. Similarly, the authors of [105] propose using MVS with sequences of images obtained by a remote controlled model helicopter for the purposes of automatic 3D mapping. These examples highlight the importance of automating the decision of when to use which MVS algorithm, as well as which images of a possibly huge data sequence to use. One of the aims of this work is to extract simple predictors for the performance of MVS algorithms that can be used to select the appropriate algorithm for each scene, or select appropriate subsets of images to use.

11.3 Data

The setup for data acquisition is illustrated in Figure 11.2 and 11.3. We have mounted two cameras and a projector on a six-axis industrial robot arm enabling us to acquire images from a range of positions. At each position we obtain a surface point cloud using structured light. Controlled illumination is obtained from a set of light emitting diodes (LEDs) mounted above the scene.

The robot provides very precise camera positioning, because it has very high positioning repeatability. However, the encoding precision is not as precise, so we obtain actual positioning by a set of predefined path positions, and this path is calibrated using a fixed checkerboard pattern. The encoded path has subsequently been used for acquiring images of the 80 scenes in our dataset.

The 80 scenes contain different number of images. 59 scenes contain 49 camera positions and 21 scenes contain 64 camera positions. Example data is shown in Figure 11.1 and 11.4. The camera positions of the smaller sets are placed spherically at a distance of 50 cm from the scene centres, i.e. around 35 cm from the scene surfaces. The larger sets contain an additional 15 positions placed spherically at a distance of 65 cm from the scene centres as shown in Figure 11.5. The scenes' content is chosen with varying reflectance, texture, and geometric properties, and include fabric, print, groceries, fruit, a bunny sculpture, and more. Scene examples are shown in Figure 11.1.

The scenes are illuminated with a set of 18 light emitting diodes (LEDs) placed above the scene. These are strobed in groups to generate different directional lighting variations. For this experiment only the uniform illumination images

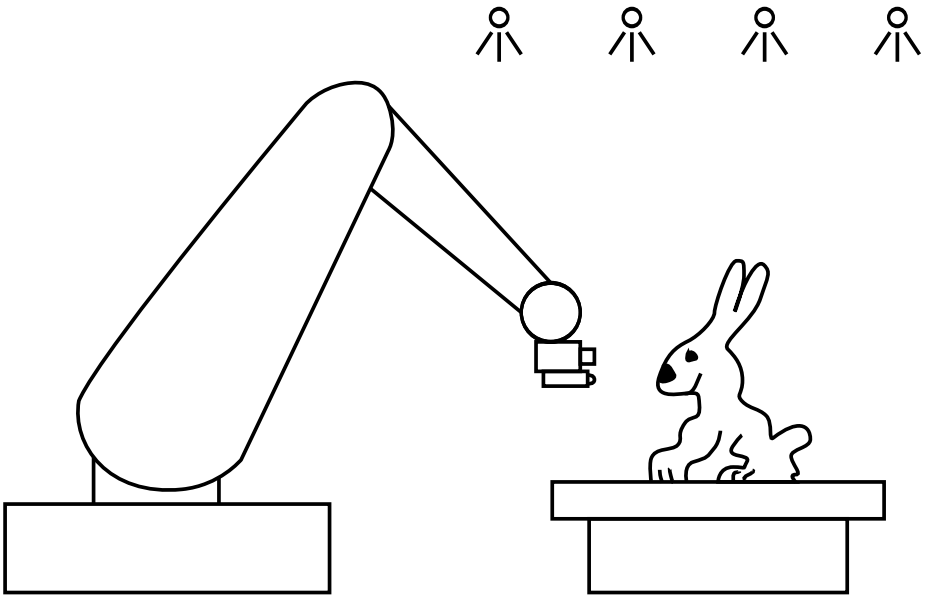


Figure 11.2: Image acquisition setup, consisting of industrial robot arm with two cameras and a projector mounted. LEDs are placed in ceiling and the scene placed on a table.

Figure 11.3: Left: Six-axis industrial robot. Right: Stereo cameras mounted on the robot arm together with a projector for structured light scanning.

are used, generated by using all 18 LEDs at once. The image resolution is 1200×1600 pixels in 8 bit RGB colour.

Reference surfaces are obtained from structured light scans using binary stripe encoding, which is recommended as being one of the most precise structured

Figure 11.4: Top row: Data example showing three of the 49 views. Bottom row: Surface reconstruction showing details of the reference data.

light methods [85, 83, 84]. We use a calibrated stereo setup mounted on the robot arm, as illustrated in Figure 11.3.

Our experiments are dependent on the accuracy of structured light scans, and we have therefore measured the accuracy using an object with known geometry. We chose a bowling ball, because it is a spherical object of suitable size with simple and known geometry. A reference scan was obtained from each camera position, and all these scans were combined to make up the total reference data for each scene. For each scan we estimated the centre position and the radius of the sphere form the surface points using linear least squares. This also enabled us to estimate the deviation of the individual points from the sphere's surface. We obtained a standard deviation of 0.17 mm on the centre position estimates, and an average standard deviation on the surface points of 0.14 mm corresponding roughly to 0.6 pixels. Positioning repeatability of the robot turned out to be very high. Over the two months data acquisition period we performed 10 calibrations, and the total standard deviation of the camera positions was 0.0031 mm.

The reference scans are not complete. The main cause is that we only cover the front of the objects, but still there are areas seen by the cameras that have not been covered. This occurs because the projector only covers the images partly, but there are also other small holes where the structured light images have been

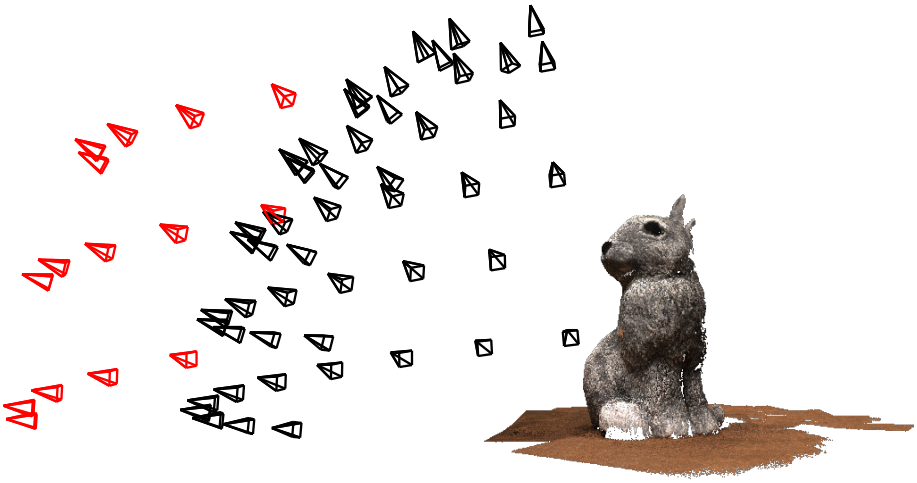


Figure 11.5: Camera positions at 50 cm distance (black) and 65 cm distance (red).

severely underexposed. Despite these minor incompletenesses the scans are very dense, each containing 13.4 million points on average. To ensure a more uniform distribution of points, we reduced the point sets such that no point was closer than 0.2 mm to another point. Hereby we avoid an evaluation bias in areas with high point density. Points were reduced by first making all points as belonging to the set. Then we randomly visited each point and checked if it was still part of the set. If it was, we marked all points within 0.2 mm from this point as outside the set. The final reduced point set contains all kept points.

Only the scene objects are used in the evaluation. This is done by removing the part of the reconstruction containing the table, simply by discarding points below a manually placed plane.

11.4 Method

The evaluation is based on measuring the difference between the multi-view stereo reconstructions and the reference structured light scans and vice versa. This measure is obtained by comparing closest points in the two sets. Since the reference scans are not complete we might get multi-view reconstructions in areas that are not covered by the reference scans. Just looking at the nearest points, these areas would be considered erroneous, because they are far from the reference, even though they might have reconstructed the surface perfectly. We

overcome this problem by computing a volumetric visibility mask for each scene that explains the areas where the reference scans can be seen. All multi-view points outside the mask are not considered in the evaluation.

The mask is initialised as a volume with voxels of 1 mm. Each voxel is a boolean telling if that voxel is a visible part of the structured light reconstruction. Points in the structured light scans are reconstructed from a specific camera position, and we use the ray from the point to the camera centre to generate the visibility volume. To allow for MVS reconstructions behind the structured light points, we extended the camera to point rays with 10 mm in depth. All voxels traversed by the rays was set as visible. 10 mm was chosen as a trade off between including outliers from the MVS and avoiding to include well reconstructed surfaces that was not covered by the structured light scans. To make our implementation efficient we employed z-buffering and used the full structured light points sets.

High density in the multi-view reconstructions can be a problem for measuring the error. If an algorithm e.g. produces many points in regions with good data support and less in areas with bad support, this algorithm would perform better than an algorithm of the same precision, but with uniform distribution of points. To avoid this bias we reduce the multi-view reconstructions to have a density that is similar or lower than the reference scans. We use the same point reduction procedure as for the reference scans.

With the volume mask, we can now compare the multi-view reconstructions with the reference scans. Let us denote the multi-view reconstructions within the visibility mask M and the reference scans R. Our performance measure is done two ways by finding the nearest point in M for each point in R and vice versa. The Euclidian distance between these neighbouring points is the error. Since only the visible parts of M are used, the error measure from all points in M to the nearest points in R will measure the precision of M. Measuring the error for each point in R will also include points in areas that might not be reconstructed in M. The measure from R to M will therefore be larger if M is incomplete. The match criteria are illustrated in Figure 11.6.

Based on this error measure the performance is correlated to image statistics. We do a point wise comparison of the measured error to a local texture measure. We chose a simple variance measure of a 15×15 pixels image patch around each surface point in M estimated as $\sigma_t^2 = \sigma_I^2/(\mu_I + \psi)$ and we set $\psi = 5$. This gives a texture and reconstruction error for each point that can be directly compared.

We have endeavoured to preserve the raw data term of MVS, isolating as far as possible the effects of outlier rejection, regularisation, surface smoothing etc. The reason for this is that the paper aims to shed light on the feasibility of the photo-consistency cue for surface reconstruction.

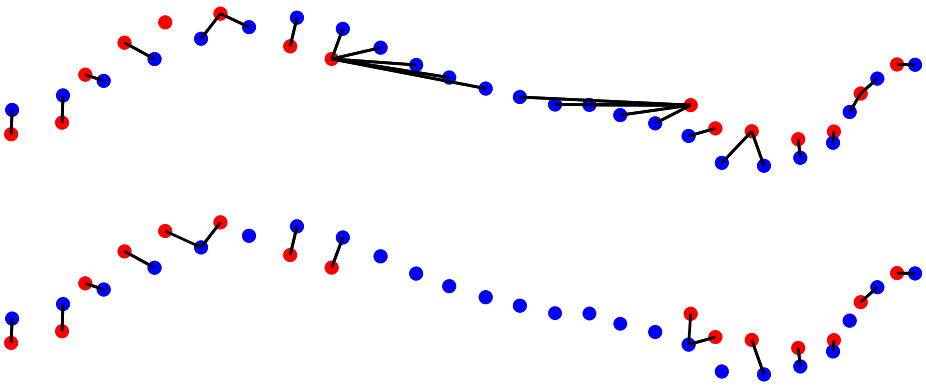


Figure 11.6: Match criteria – reference scans R are shown as blue points and MVS reconstructions M as red points. Top is matching from R to M and bottom is matching from M to R. Note that the incompleteness in the red points gives error in the R to M match.

11.5 Results

Performance of the three approaches are shown in Table 11.1. Both MVS algorithms perform superior to two-view stereo, and many complicated scenes are reconstructed with high accuracy. Especially the comparison from the reference scans to the MSV (R to M) is generally low. This shows that the MSV algorithms reconstruct surface points according to the reference. Some scenes have, however, larger errors primarily caused by holes in the MVS reconstructions. The comparison from the MSV to the reference (M to R) reveals the precision of the algorithms. Here the two-view stereo is clearly inferior to the multi-view counterparts. Comparing scenes with high error with low error scenes reveals, that especially scenes with repeated texture and specularities corrupt the reconstruction, but only for the two-view case. In contrast, the two algorithms based on multiple images can handle specular surfaces and repeated texture very well, as the examples in Figure 11.7 and 11.8 illustrate.

We have found that the primary factor corrupting surface reconstruction is lack of texture, with the result of holes in the MVS reconstructions in areas without texture. So, our focus has been on how texture can be characterised to predict performance of the MVS reconstructions. Holes in scans are primarily seen in the reference to MVS (R to M) error measure, and Figure 11.9 shows the performance error compared to the texture measure.

In each reference point we have a reconstruction error measure and a texture

Algorithm	Avg. err.	Median err.
Furukawa and Ponce [33]		
(Multi-view stereo)		
R to M	0.413	0.841
M to R	0.315	0.658
Campbell et al. [17]		
(Multi-view stereo)		
R to M	0.168	0.595
M to R	0.481	0.790
Campbell et al. [17]		
(Two-view stereo)		
R to M	0.194	0.612
M to R	0.598	1.999

Table 11.1: Error measured as distance in mm for the three surface reconstruction methods. Results are shown as the distance from the reference to the reconstruction method (R to M) and the reconstruction to the reference (M to R)

.

measure. For each scene we can look at the distribution of both the error and the texture measure, and we found distribution quantiles to be the feature most precisely explaining the MVS reconstruction. To select the most relevant percentile in the two distributions, we computed the correlation coefficient for a large number of percentiles for both the error and texture measure. We found that especially high percentiles in the error measure and low percentiles in the texture measure show highest correlation.

To estimate how much of the MVS can be predicted by texture we use the 90 % percentile of the error measure and the 10 % percentile of the texture measure. These features were used in a correlation analysis as illustrated in Figure 11.10. Correlation coefficient of -0.52 was obtained for [33] and -0.50 for [17]. This shows that our simple texture characterisation measured per image does not fully explain the MVS performance. But it should be noted that all scenes with high textured images have good performance. This is seen even clearer in Figure 11.11 that shows the total patch distribution of texture versus performance. The upper right corner is all zeros saying that no patch with good texture is poorly reconstructed.

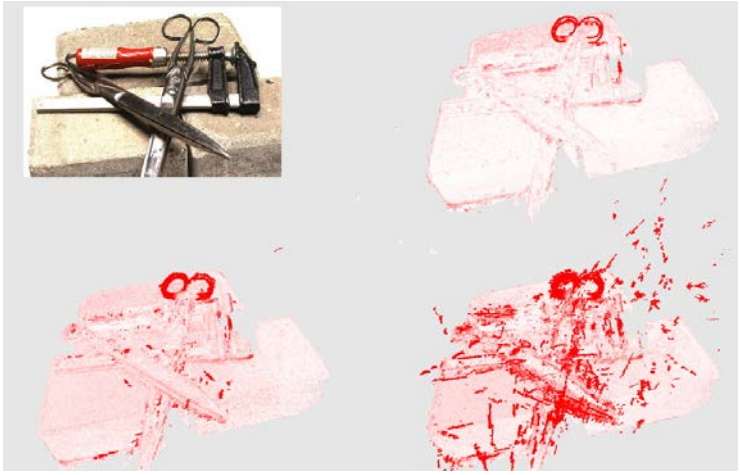


Figure 11.7: Example of a specular scene compared from MVS reconstruction to reference (M to R). Top left is an example image. Top right is [33]. Bottom is [17] – left is multi-view and right is two-view. The colour white to red shows the error, with red having error larger than 1 cm and white points having no error.

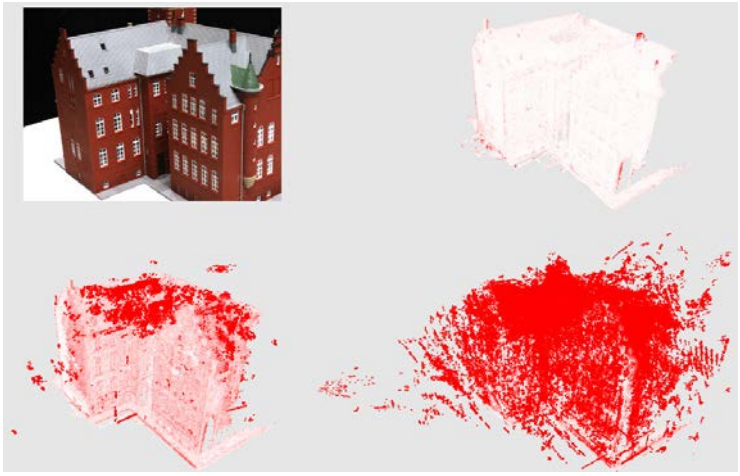


Figure 11.8: Example of a scene with repeated texture compared from MVS reconstruction to reference (M to R). Top left is an example image. Top right is [33]. Bottom is [17] – left is multi-view and right is two-view. The colour white to red shows the error, with red having error larger than 1 cm and white points having no error.

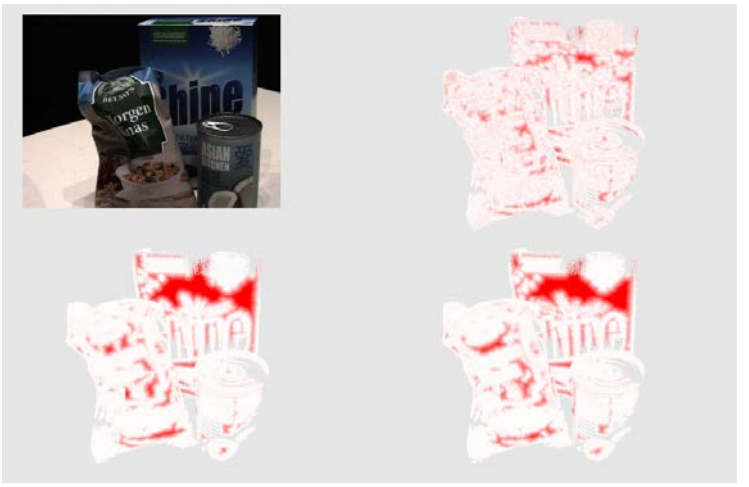


Figure 11.9: Example of a scene with missing texture compared from reference to MVS reconstruction (R to M). Top left is an example image. Top right is error of [33]. Bottom is reconstructions from [17] – left is multi-view and right is two-view. The colour white to red shows the error, with red having error larger than 1 cm and white points having no error.

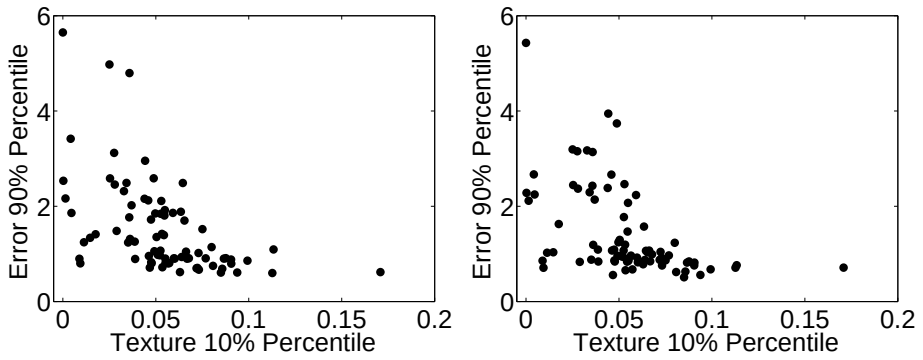


Figure 11.10: Performance error compared to texture characterisation. Left is [33] with a correlation coefficient of -0.52 and right [17] with a correlation coefficient of -0.50. The average error between the reference and the multi-view reconstructions is used (mean of the M to R and R to M measure).

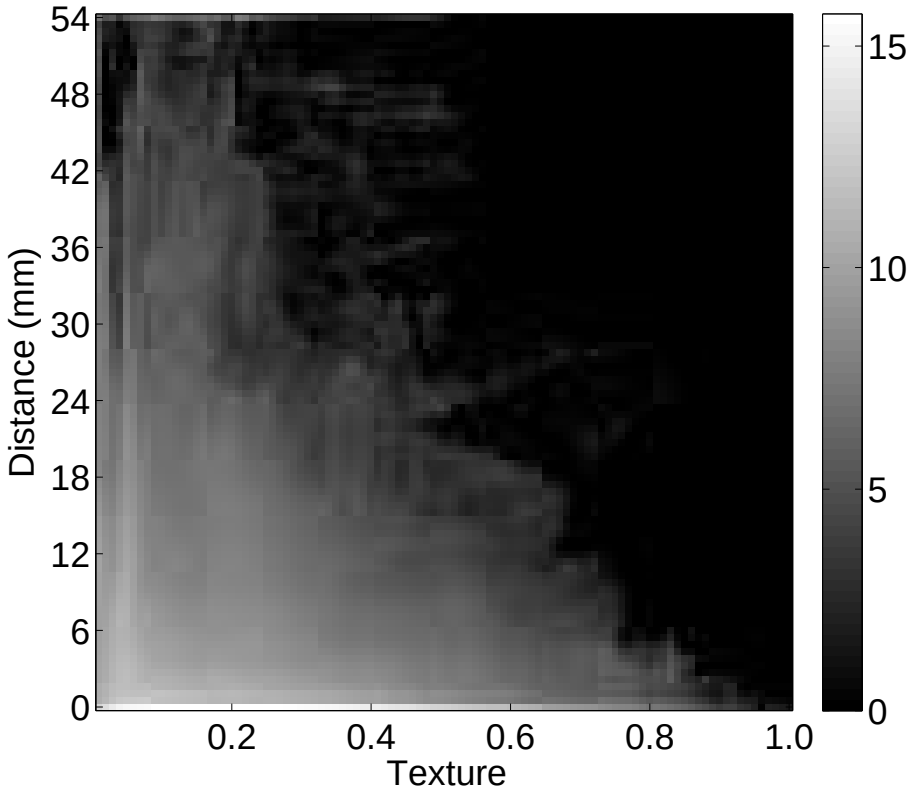


Figure 11.11: 2D histogram of texture and error measure (logarithm of number of patches). Note how all patches with high texture have low matching error.

11.6 Discussion

In this paper we posed the question of "what types of scenes work best for *this MVS algorithm*?" This is a very important question for judging the quality of a MVS reconstruction, because this can be very hard to tell just from a set of images and a 3D point cloud. Our answer is that scenes rich in texture will be reconstructed well for the two MVS algorithms by Furukawa and Ponce [33] and Campbell et al. [17] that we investigated. Specular surfaces, repeated texture or complicated geometry did not corrupt the reconstruction. The degree to which the MVS algorithms were able to precisely reconstruct 3D points despite highly specular surfaces came as a surprise. This is a major advantage of using multiple views compared to two-views. Experiments with two-views were severely influenced by all these factors, which indicates that situations with two or even just a few images is still a very hard and unsolved problem.

Our study is based on a ground truth augmented dataset of 80 scenes with at total of 4235 images, which is many times larger than anything reported – as far as we are aware. This size highly improves the statistical significants of our findings. This data will be made publicly available to e.g. investigate the challenges and research paths for other types of MVS algorithms, than the ones tested in our study.

The aim of the paper was to shed light on the feasibility of the photo-consistency cue for surface reconstruction. The raw data term was therefore preserved to isolate the effect of outlier rejection, regularisation, surface smoothing etc. The photo-consistency cues of MVS methods turned out to be stronger than expected. Our dataset was chosen to reflect aspects assumed to corrupt reconstruction, but a vast majority of the MVS reconstructed surface points were very close to the reference. Especially the specularities represented in our data turned out not to be a problem. We did not take this to the extreme of mirroring or transparent surfaces – also because this would require alternative methods for obtaining reference data.

The performance obtained based on using the raw data term, i.e. avoiding outlier rejection, regularisation, surface smoothing, etc., implies two things a) that results are likely to be better than presented here with the full fledged algorithms, and b) that issues with forming a dense surface from the points, i.e. meshing, has not been evaluated.

We found the presence of texture to result in good performance, and all poor performing scenes had areas with missing texture. However, some scenes with low texture were still reconstructed well, which indicates that even few texture cues can aid reconstruction. Our findings regarding texture were validated vi-

sually and confirmed by a quantitative analysis. So, a MVS reconstruction will be good if the employed images were rich in texture.

The fact that a lack of texture does not necessarily imply poor performance implies that there are more factors to determining poor performance than just a lack of texture. In our future work we aim at finding other factors determining performance of MVS surface reconstruction. To accomplish this, we plan to exploit the full potential of our dataset by investigating the effect of the number of images used for the reconstruction, the local geometry and lighting, as well as texture.

Appendix A

Stuff

A.1 Spherical estimates of individual scan

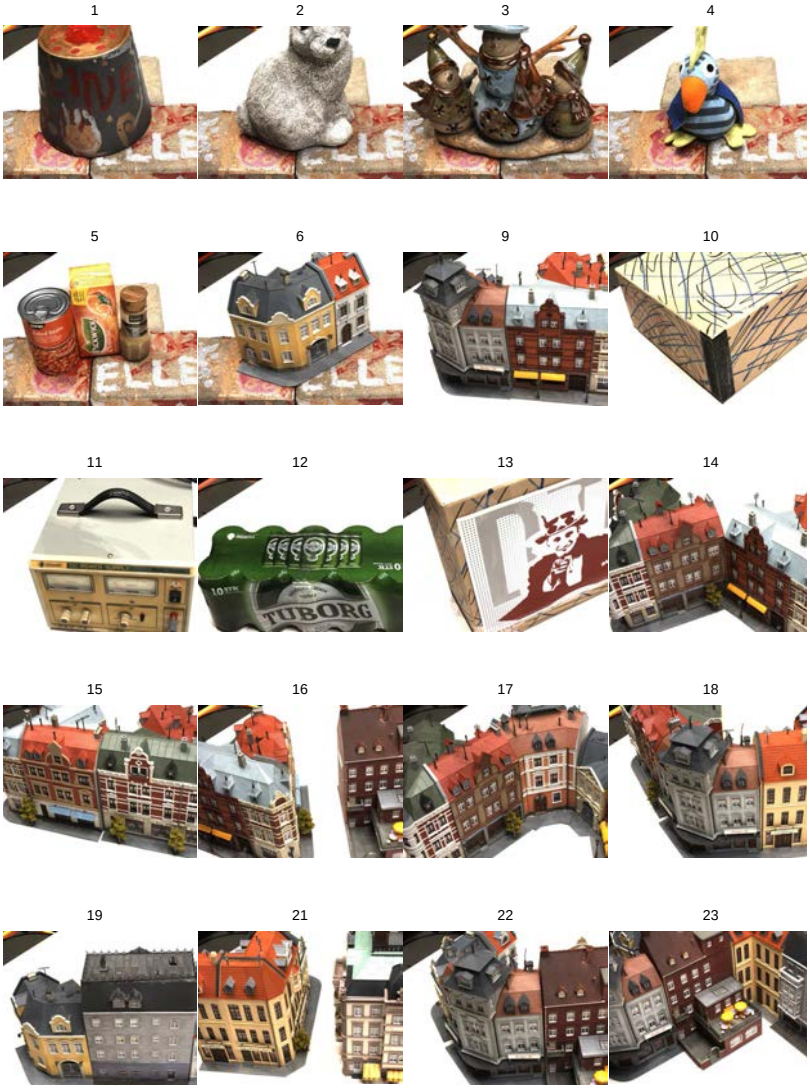
Observation	$\bar{x}_c$	$\bar{y}_c$	$\bar{z}_c$	$\bar{r}$	σ
1	-11.15	-25.18	629.51	132.66	0.0958
2	-10.88	-25.09	629.59	132.68	0.0767
3	-10.73	-25.42	629.73	132.74	0.0687
4	-10.92	-25.70	629.95	132.82	0.0724
5	-11.27	-25.83	630.15	132.95	0.0708
6	-11.35	-25.78	630.17	132.99	0.0725
7	-11.05	-25.68	630.11	132.88	0.0728
8	-10.80	-25.52	629.96	132.79	0.0754
9	-10.76	-25.28	629.80	132.75	0.0731
10	-10.84	-25.17	629.73	132.73	0.0723
11	-11.05	-25.01	629.66	132.69	0.0783
12	-11.17	-24.91	629.54	132.61	0.0891
13	-11.04	-25.03	629.70	132.70	0.0793
14	-10.93	-25.12	629.81	132.75	0.0758
15	-10.87	-25.23	629.86	132.77	0.0697

Table A.1: Estimated center coordinates $(\bar{x}_c, \bar{y}_c, \bar{z}_c)$, radius $(\bar{x}_r)$ and standard deviation of individual points (σ) for observations 1 - 25. All numbers are in mm.

Observation	$\bar{x}_c$	$\bar{y}_c$	$\bar{z}_c$	$\bar{r}$	σ
16	-10.80	-25.39	629.93	132.78	0.0735
17	-10.90	-25.56	630.06	132.81	0.0767
18	-11.08	-25.66	630.12	132.87	0.0788
19	-11.31	-25.77	630.20	132.99	0.0765
20	-11.44	-25.72	630.20	133.10	0.0832
21	-11.25	-25.67	630.15	133.00	0.0851
22	-11.03	-25.58	630.11	132.88	0.0846
23	-10.93	-25.47	630.06	132.83	0.0828
24	-10.87	-25.37	630.01	132.82	0.0795
25	-10.87	-25.24	629.94	132.82	0.0739
26	-10.93	-25.10	629.88	132.80	0.0745
27	-11.01	-25.00	629.79	132.77	0.0768
28	-11.09	-24.89	629.67	132.69	0.0833
29	-11.22	-24.71	629.64	132.62	0.0866
30	-11.07	-24.87	629.75	132.72	0.0818
31	-11.04	-24.98	629.85	132.80	0.0813
32	-10.99	-25.08	629.91	132.84	0.0755
33	-10.93	-25.19	630.00	132.86	0.0772
34	-10.86	-25.37	630.05	132.86	0.0818
35	-10.88	-25.48	630.11	132.86	0.0875
36	-10.95	-25.61	630.17	132.88	0.0932
37	-11.12	-25.69	630.24	132.97	0.0944
38	-11.31	-25.73	630.23	133.06	0.0945
39	-11.41	-25.71	630.35	133.12	0.0987
40	-11.25	-25.69	630.30	133.04	0.1018
41	-11.10	-25.61	630.27	132.98	0.1017
42	-10.97	-25.52	630.22	132.91	0.0980
43	-10.94	-25.45	630.18	132.90	0.0929
44	-10.97	-25.33	630.12	132.91	0.0865
45	-10.98	-25.21	630.06	132.90	0.0804
46	-11.02	-25.11	629.99	132.88	0.0772
47	-11.06	-25.01	629.92	132.84	0.0782
48	-11.16	-24.83	629.84	132.76	0.0805
49	-11.21	-24.71	629.75	132.67	0.0870

Table A.2: Estimated center coordinates $(\bar{x}_c, \bar{y}_c, \bar{z}_c)$, radius $(\bar{x}_r)$ and standard deviation of individual points (σ) for observations 26 - 49. All numbers are in mm.

A.2 Catalogue of scenes in the data set



24


25


28


29


30


31


32


33

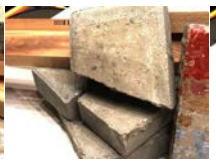

34


35


36


37


38


39


40

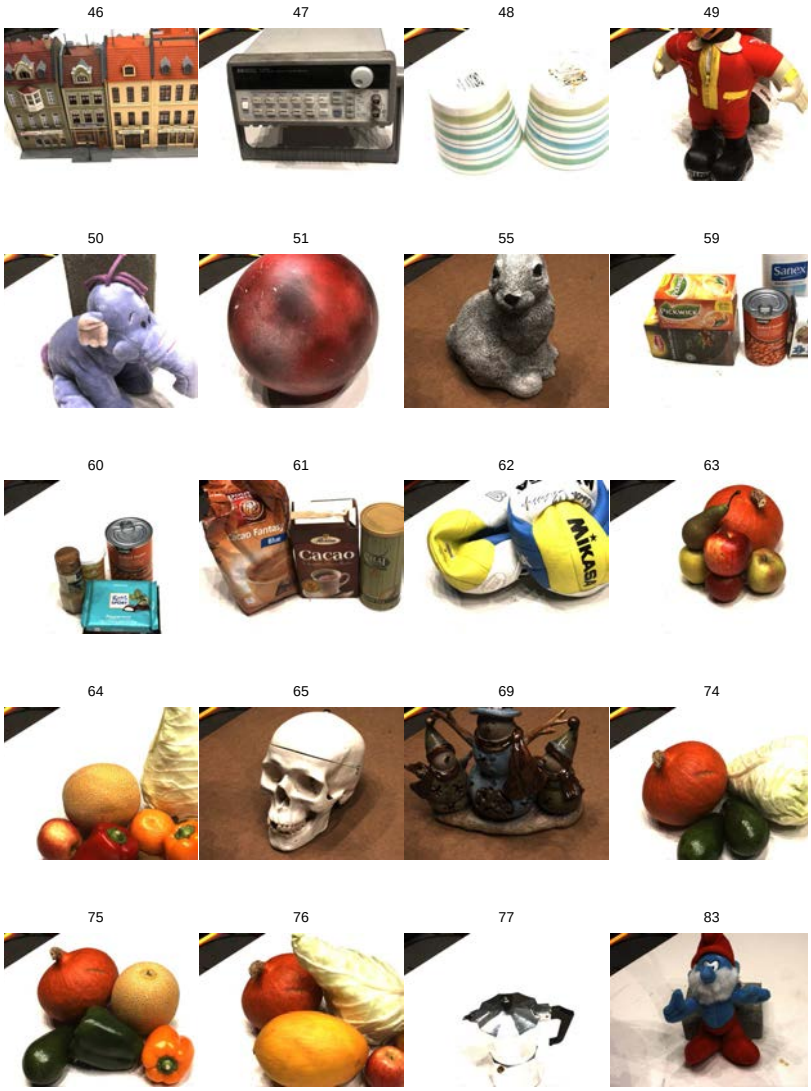

41


42


43


44


45



46
47
48
49
50
51
55
59
60
61
62
63
64
65
69
74
75
76
77
83
Cacao
MIKASA
Sanex

84 93 94 95

96 97 98 99

100 102 103 105

106 110 114 118

122 126 127 128

Bibliography

[1] 3Dsystems. Projet MP3000. URL http://www.3dsystems.com/.

[2] 3Shape A/S. Website. URL http://www.3shape.com.

[3] 3Shape A/S. TRIOS scanner. Website, 2011. URL www.3shapedental.com/restoration/dentist/digital-impression-taking.aspx.

[4] Henrik Aanæs, Anders Lindbjerg Dahl, and Kim Steenstrup Pedersen. Interesting Interest Points. *International Journal of Computer Vision*, 97 (1):1–18, June 2011. ISSN 0920-5691. doi: 10.1007/s11263-011-0473-8.

[5] P J Anderson, D J Netherway, A Abbott, and D J David. Intracranial Volume Measurement of Metopic Craniosynostosis. *J. Craniofacial Surg.*, 15(6):1014–1016, 2004.

[6] P J Anderson, D J Netherway, K McGlaughlin, and D J David. Intracranial Volume Measurement of Sagittal Craniosynostosis. *J. Clinical Neuroscience*, 14(5):455–458, 2007.

[7] Christian Bailer, Manuel Finckh, and Hendrik P A Lensch. Scale robust multi view stereo. In *Proceedings of the 12th European conference on Computer Vision - Volume Part III*, ECCV'12, pages 398–411, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-33711-6. doi: 10.1007/978-3-642-33712-3_29.

[8] J Batlle, E Mouaddib, and J Salvi. Recent progress in coded structured light as a technique to solve the correspondence problem: a survey. *Pattern recognition*, 31(7), 1998.

[9] S Belongie, J Malik, and J Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, April 2002.

[10] Julian Besag, Statistical Society, and Series B Methodological. On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302, 1986. ISSN 00359246.

[11] Paul J Besl and Neil D McKay. A method for registration of 3D shapes. *IEEE Transactions on pattern analysis and machine intelligence*, 14(2): 239–256, 1992.

[12] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999.

[13] Volker Blanz, Albert Mehl, Thomas Vetter, and H-P Seidel. A statistical method for robust 3D surface reconstruction from sparse data. In *3D Data Processing, Visualization and Transmission. Proceedings. 2nd International Symposium on*, pages 293–300. IEEE, 2004.

[14] J Bloomenthal. An implicit surface polygonizer. *Graphics Gems IV*, pages 324–349, 1994.

[15] M Botsch and L Kobbelt. A remeshing approach to multiresolution modeling. *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 185–192, 2004.

[16] Y Boykov, O Veksler, and R Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[17] Neill D Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using Multiple Hypotheses to Improve Depth-Maps for Multi-View Stereo. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 766–779, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-88681-5. doi: 10.1007/978-3-540-88682-2_58.

[18] V Caselles, G Haro, G Sapiro, and J Verdera. On geometric variational models for inpainting surface holes. *Computer Vision and Image Understanding*, 111(3):351–373, 2008.

[19] M Chandraker and R Ramamoorthi. What An Image Reveals About Material Reflectance. In *IEEE International Conference on Computer Vision*, 2011.

[20] R Cipolla, S Battiato, and G M Farinella. *Computer Vision: Detection, Recognition and Reconstruction*. Studies in Computational Intelligence. Springer, 2010.

[21] U Clarenz, U Diewald, G Dziuk, M Rumpf, and R Rusu. A finite element method for surface restoration with smooth boundary conditions. *Computer Aided Geometric Design*, 21(5):427–446, 2004.

[22] T F Cootes, C J Taylor, D H Cooper, and J Graham. Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[23] D Cremers and K Kolev. Multiview Stereo and Silhouette Consistency via Convex Functionals over Convex Domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1161–1174, 2011.

[24] B Curless and M Levoy. A volumetric method for building complex models from range images. *Proceedings of ACM SIGGRAPH*, pages 303–312, 1996.

[25] A L Dahl, H Aanæs, and K S Pedersen. Finding the Best Feature Detector-Descriptor Combination. In *International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 318–325, 2011.

[26] S Darkner, R Larsen, and R Paulsen. Analysis of deformation of the human ear and canal caused by mandibular movement. *Proc. Medical Image Computing and Computer-Assisted Intervention*, pages 801–808, 2007.

[27] Lee R Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):pp. 297–302, 1945.

[28] Mark Everingham, Luc Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, September 2009. ISSN 0920-5691. doi: 10.1007/s11263-009-0275-4.

[29] M A Fischler and R C Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. ISSN 0001-0782.

[30] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning.* Springer Series in Statistics, 2001.

[31] Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bulow, and Jitendra Malik. Recognizing Objects in Range Data Using Regional Point Descriptors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2004.

[32] Y Furukawa, B Curless, S M Seitz, and R Szeliski. Towards Internet-scale multi-view stereo. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1434–1441, 2010. doi: 10.1109/CVPR.2010.5539802.

[33] Yasutaka Furukawa and Jean Ponce. Accurate, Dense, and Robust Multi-View Stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, August 2010. ISSN 1939-3539. doi: 10.1109/TPAMI.2009.161.

[34] D Gallup, J M Frahm, P Mordohai, Q Yang, and M Pollefeys. Real-Time Plane-Sweeping Stereo with Multiple Sweeping Directions. In *IEEE Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[35] D Gallup, J M Frahm, P Mordohai, and M Pollefeys. Variable baseline/resolution stereo. In *IEEE Computer Vision and Pattern Recognition*, 2008.

[36] M Goesele, N Snavely, B Curless, H Hoppe, and S M Seitz. Multi-View Stereo for Community Photo Collections. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[37] Michael Goesele, Brian Curless, and Steven M Seitz. Multi-View Stereo Revisited. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2402–2409, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2597-0. doi: 10.1109/CVPR.2006.199.

[38] Matthias Grundmann, Franziska Meier, and Irfan Essa. 3D shape context and distance transform for action recognition. In *19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[39] P Hammond, T J Hutton, J E Allanson, L E Campbell, R Hennekam, S Holden, M A Patton, A Shaw, I K Temple, M Trotter, and Others. 3D analysis of facial morphology. *American Journal of Medical Genetics Part A*, 126(4):339–348, 2004.

[40] R Hartley and A Zisserman. *Multiple view geometry in computer vision*. 2000.

[41] J Heikkila and O Silven. A four-step camera calibration procedure with implicit image correction. *Computer Vision and Pattern Recognition, 1997. Proceedings*, pages 1106–1112, 1997.

[42] C Hernandez, G Vogiatzis, and R Cipolla. Probabilistic visibility for multi-view stereo. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007. doi: 10.1109/CVPR.2007.383193.

[43] H Hirschmuller and D Scharstein. Evaluation of Stereo Matching Costs on Images with Radiometric Differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, 2009.

[44] Xiaoyan Hu and P Mordohai. Evaluation of stereo confidence indoors and outdoors. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1466–1473, 2010. doi: 10.1109/CVPR.2010.5539798.

[45] DR Hunter and Kenneth Lange. A tutorial on MM algorithms. *The American Statistician*, 58:30–37, 2004.

[46] D P Huttenlocher, G A Klanderman, and W J Rucklidge. Comparing images using the Hausdorff distance. *PAMI, IEEE Trans. on*, 15(9):850–863, 1993.

[47] J.-Y. Bouguet. Camera calibration toolbox for Matlab.

[48] H Jin, S Soatto, and A Yezzi. Multi-view stereo reconstruction of dense shape and complex appearance. *International Journal of Computer Vision*, 63(3):175–189, 2005.

[49] M W Jones and Others. 3D distance fields: A survey of techniques and applications. *IEEE Transactions on Visualization and Computer Graphics*, pages 581–599, 2006. ISSN 1077-2626.

[50] O V Kaick, H Zhang, G Hamarneh, and D Cohen-Or. A survey on shape correspondence. In *Computer Graphics Forum*, volume 20, pages 1–23, 2010.

[51] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, SGP '06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association, Eurographics Association. ISBN 3-905673-36-3.

[52] R Klowsky, A Kuijper, and M Goesele. Modulation transfer function of patch-based stereo systems. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1386–1393, 2012. doi: 10.1109/CVPR.2012.6247825.

[53] Leif Kobbelt and Mario Botsch. A survey of point-based techniques in computer graphics. *Computers & Graphics*, 28(6):801–814, 2004.

[54] Pushmeet Kohli and Philip H S Torr. Dynamic graph cuts for efficient inference in Markov Random Fields. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2079–88, December 2007. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1128.

[55] K Kolev, T Brox, and D Cremers. Fast Joint Estimation of Silhouettes and Dense 3D Geometry from Multiple Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):493–505, 2012.

[56] R Kolluri, J R Shewchuk, and J F O'Brien. Spectral surface reconstruction from noisy point clouds. In *Proceedings of the Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 11–21. ACM, 2004.

[57] V Kolmogorov and R Zabin. What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):147–159, 2004. ISSN 01628828.

[58] M Körtgen and GJ Park. 3D shape matching with 3D shape contexts. In *In The 7th Central European Seminar on Computer Graphics*, 2003.

[59] H W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, 2:83–97, 1955.

[60] S Lanche, T Darvann, H Ólafsdóttir, N Hermann, A Van Pelt, D Govier, M Tenenbaum, S Naidoo, P Larsen, S Kreiborg, and Others. A statistical model of head asymmetry in infants with deformational plagiocephaly. *Image Analysis*, pages 898–907, 2007.

[61] K Van Leemput, F Maes, D Vandermeulen, and P Suetens. Automated Model-based Tissue Classification of MR Images of the Brain. *Med. Imag., IEEE Trans. on*, 18(10):897–908, 1999.

[62] K Li, X Wu, D Z Chen, and M Sonka. Optimal surface segmentation in volumetric images-a graph-theoretic approach. *PAMI, IEEE Trans. on*, 28(1):119–134, 2006.

[63] S Z Li. Markov random field models in computer vision. *Computer Vision - ECCV '94. Third European Conference on Computer Vision. Proceedings. Vol.II*, pages 361–370, 1994.

[64] Jiangyu Liu. Parallel graph-cuts by adaptive bottom-up merging. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2181–2188, June 2010. doi: 10.1109/CVPR.2010.5539898.

[65] Shubao Liu and D B Cooper. A complete statistical inverse ray tracing approach to multi-view stereo. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 913–920, 2011. doi: 10.1109/CVPR.2011.5995334.

[66] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[67] K Y Manjunath. Estimation of Cranial Volume - an Overview of Methodologies. *J. Anat. Soc. India*, 51(1):85–91, 2002.

[68] M A A Neil, R Juskaitis, and T Wilson. Method of obtaining optical sectioning by using structured light in a conventional microscope. *Optics Letters*, 22(24):1905–1907, 1997. ISSN 1539-4794.

[69] J A Nelder and R Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965.

[70] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. DTAM: Dense tracking and mapping in real-time. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2320–2327, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-1-4577-1101-5. doi: 10.1109/ICCV.2011.6126513.

[71] R Paulsen and K Hilger. Shape modelling using markov random field restoration of point correspondences. In *Information Processing in Medical Imaging*, pages 1–12. Springer, 2003.

[72] R R Paulsen. *Statistical shape analysis of the human ear canal with application to in-the-ear hearing aid design.* PhD thesis, 2004.

[73] R R Paulsen, R Larsen, S Laugesen, C Nielsen, and B K Ersbø ll. Building and Testing a Statistical Shape Model of the Human Ear Canal. In *Proc. Medical Image Computing and Computer-Assisted Intervention*, pages 373–380, 2002.

[74] R R Paulsen, J A Bæ rentzen, and R Larsen. Markov random field surface reconstruction. *Visualization and Computer Graphics, IEEE Transactions on*, 16(4):636–646, 2010.

[75] Rasmus R Paulsen and Rasmus Larsen. Anatomically plausible surface alignment and reconstruction. In *Proceedings of Theory and Practice of Computer Graphics*, 2010.

[76] G Pengas, J M S Pereira, G B Williams, and P J Nestor. Comparative Reliability of Total Intracranial Volume Estimation Methods and the Influence of Atrophy in a Longitudinal Semantic Dementia Cohort. *J. Neuroimaging*, 19(1):37–46, 2009.

[77] C Pirzanski. Despite new digital technologies, shell modelers shoot in the dark. *The Hearing Journal*, 59(10):28, 2006.

[78] C Pirzanski and B Berge. Earmold impressions: Does it matter how they are taken. *Hear Rev*, 10(4):18–20, 2003.

[79] J.-P. Pons, R Keriven, and O Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *International Journal of Computer Vision*, 72(2):179–193, 2007.

[80] CH H Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967. ISSN 0029-599X.

[81] Fabio Remondino and Sabry El-Hakim. Image-based 3D Modelling: A Review. *The Photogrammetric Record*, 21(115):269–291, 2006.

[82] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the ICP algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001.

[83] Joaquim Salvi, Jordi Pages, and Joan Batlle. Pattern codification strategies in structured light systems. *Pattern Recognition*, 37(4):827–849, 2004.

[84] Joaquim Salvi, Sergio Fernandez, Tomislav Pribanic, and Xavier Llado. A state of the art in structured light patterns for surface profilometry. *Pattern recognition*, 43(8):2666–2680, 2010.

[85] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I—-195. IEEE, 2003.

[86] R Schnabel, R Wahl, and R Klein. Efficient RANSAC for Point-Cloud Shape Detection. In *Computer Graphics Forum*, volume 26, pages 214–226. Wiley Online Library, 2007.

[87] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:519–528, 2006.

[88] S Sgouros, A D Hockley, J H Goldin, M J C Wake, and K Natarajan. Intracranial Volume Change on Craniosynostosis. *J. Neurosurg.*, 91:617–625, 1999.

[89] J Shewchuk. Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator. *Applied Computational Geometry Towards Geometric Engineering*, pages 203–222, 1996.

[90] JR R Shewchuk. Tetrahedral mesh generation by Delaunay refinement. *Proceedings of the 14th annual symposium on Computational geometry*, 4: 86–95, 1998.

[91] S M Smith. Fast Robust Automated Brain Extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

[92] N Snavely, SM Seitz, and R Szeliski. Photo tourism: exploring photo collections in 3D. *ACM transactions on graphics (TOG)*, 2006.

[93] Rothamsted Experimental Station and J C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[94] S Stefano, A J Anthony, and J Hailin. Tales of Shape and Radiance in Multi-view Stereo. In *IEEE International Conference on Computer Vision*, 2003.

[95] Christoph Strecha, Wolfgang von Hansen, Luc Van Gool, Pascal Fua, and Ulrich Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

[96] E Tola, V Lepetit, and P Fua. A Fast Local Descriptor for Dense Matching. In *IEEE Computer Vision and Pattern Recognition*, 2008.

[97] E Tola, V Lepetit, and P Fua. DAISY: an Efficient Dense Descriptor Applied to Wide-Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, 2010.

[98] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23(5):903–920, 2012.

[99] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *Computer Vision–ECCV 2010*, pages 356–369. Springer, 2010.

[100] Tinne Tuytelaars and Krystian Mikolajczyk. Local Invariant Feature Detectors: A Survey. *Foundations and TrendsÂ® in Computer Graphics and Vision*, 3(3):177–280, 2007. ISSN 1572-2740. doi: 10.1561/0600000017.

[101] J Verdera, V Caselles, M Bertalmio, and G Sapiro. Inpainting surface holes. In *Proceedings of International Conference on Image Processing*, volume 2, 2003.

[102] George Vogiatzis, Carlos Hernández, Philip H S Torr, and Roberto Cipolla. Multiview stereo via volumetric graph-cuts and occlusion robust photo-consistency. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2241–2246, 2007.

[103] H-H. Vu, R Keriven, P Labatut, and J.-P Pons. Towards high-resolution large-scale multi-view stereo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, June 2009.

[104] A Wehr and U Lohr. Airborne laser scanning–an introduction and overview. *ISPRS Journal of Photogrammetry and Remote Sensing*, 54 (2-3):68–82, 1999. ISSN 0924-2716.

[105] A Wendel, M Maurer, G Graber, T Pock, and H Bischof. Dense reconstruction on-the-fly. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1450–1457, 2012. doi: 10.1109/CVPR.2012.6247833.

[106] J L Whitwell, W R Crum, H C Watt, and N C Fox. Normalization of Cerebral Volumes by Use of Intracranial Volume: Implications for Longitudinal Quantitative MR Imaging. *AJNR*, 22:1483–1489, 2001.

[107] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, August 1987. ISSN 01697439. doi: 10.1016/0169-7439(87)80084-9.

[108] C Zach, T Pock, and H Bischof. A globally optimal algorithm for robust TV-L1 range image integration. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[109] Y Zhang, M Brady, and S Smith. Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm. *Med. Imag., IEEE Trans. on*, 1(20):45–57, 2001.

[110] Zhengyou Zhang. Iterative point matching for registration of free-form curves and surfaces. *International journal of computer vision*, 13(2):119–152, 1994.