KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ESAT
AFDELING PSI
Kasteelpark Arenberg 10 — 3001 Heverlee, Belgium

MULTI-VIEW STEREO AS AN INVERSE INFERENCE PROBLEM

Promotor:
Prof. L. VAN GOOL

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de Toegepaste Wetenschappen

door

Christoph STRECHA

Mei 2007

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ESAT
AFDELING PSI
Kasteelpark Arenberg 10 — 3001 Heverlee, Belgium

MULTI-VIEW STEREO AS AN INVERSE INFERENCE PROBLEM

Jury:
Voorzitter: Prof. A. Bultheel
Prof. L. Van Gool, promotor
Prof. P. Sturm, INRIA France
Prof. M. Jansen
Prof. D. Vandermeulen
Prof. L. Van Eycken
Prof. P. Wambacq

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de Toegepaste Wetenschappen

door

Christoph STRECHA

U.D.C. 681.3*I4

Mei 2007

©Katholieke Universiteit Leuven - Faculteit Toegepaste Wetenschappen
Arenbergkasteel, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

Wettelijk depot D/2007/7515/62

ISBN 978-90-5682-828-8

Acknowledgements

Writing a PhD is much more than what is collected here. It is most importantly discussing, defending, verifying your ideas in the community of researchers. Today, I would like to give credit for those who joint me in all this. First of all, thank you Prof. Luc Van Gool for giving me the opportunity to work in the VISICS group. I am grateful for your suggestions and comments throughout this thesis project, for pushing me to go for the 'good' conferences and for getting to know the computer vision community.

I would like to thank my reading committee members Prof. Maarten Jansen and Prof. Dirk Vandermeulen for their expert advice, especially while reading the first version of this thesis. Thank you also Prof. Dirk Roose for the comments on my first work on numerical PDE solutions.

Thank you also, Prof. A. Bultheel, Prof. Peter Sturm, Prof. Luc van Eycken and Prof. Patrick Wambacq for showing interest in my research by being the chairman or members of the jury.

Furthermore, my special acknowledgements go to you, Rik Fransens. During many cigarettes we felt like *real* scientists, discussing and discovering together many aspects of statistics. I hope that the new kernel will not be an outlier in the statistical literature. Thank you also, Geert Caenen, Naoki Kogo, Tinne Tuytelaars and Prof. Karl Verfaillie for many discussions.

I would like to thank the team from the 3-D kiosk, to provide the necessary camera calibration for my experiments. Thank you, Stefaan De Roeck and Bert DeKnuydt for the Linux help and thank you all my colleagues at the VISICS lab for the good collaboration.

In addition, a great help was always provided by Paul Konijn, Patricia Waege and Annitta Messemaeker.

Vielen Dank an meine Eltern, für die Freiheit und das viele Vertrauen in all das was mich je überkam. Auch Dank an meinen Bruder, Guntram, für sein auch immer offenes Ohr für meine Arbeit.

Manchesmal war es schwierig auch *mental* für Euch - Mathilde, Theresa und Johannes - zu Hause zu sein. Es blieb dieses Gefühl doch noch mehr für Euch da sein zu müssen. Das dieses in den richtigen Bahnen blieb verdanke ich Dir, liebe Katelijne. Du wußtest wechselnde Stimmungen richtig einzuschätzten und zu relativieren. Ganz vielen Dank dafür.

Abstract

This thesis deals with the dense multi-view stereo problem. The inherent difficulties which complicate the stereo-correspondence problem are occlusions. Also, we have to consider the possibility that image pixels in different images, which are projections of the same point in the scene, will have different colour values due to non-Lambertian effects or discretisation errors. To tackle these problems we propose a generative model based approach.

In this approach, the images are regarded as noisy measurements of an underlying ‘true’ image-function. Also, the image data is considered incomplete, in the sense that we do not know which pixels from a particular image are occluded in the other images. This formulation is equivalent to an inverse inference problem, where the goal is to estimate the factors that have generated the input images. More particular, given a set of images from a scene, we consider the question what would be the most likely image that would have been observed from a particular camera position.

To answer this question, we study a global and a local formulation. In a global formulation all possible geometric realisations of a scene are considered and evaluated to find the most plausible realisation. The local formulation takes an initial geometric realisation and refines it in a gradient decent manner.

Both formulations are intensely evaluated and their advantages and disadvantages are discussed. Finally, our proposed multi-view stereo algorithm combines both formulations and its performance is illustrated on several real-world examples. We show how the algorithm can generate realistic view interpolations from a virtual camera viewpoint.

Notations

To enhance the readability of this thesis, some notations and naming conventions used throughout the text are shortly summarized here.

$\boldsymbol{p}$	vector
$\mathbf{1}$	unit vector
p^k	specific entry k of vector $\boldsymbol{p}$
$\boldsymbol{P}$	matrix
$\boldsymbol{I}$	unit matrix
$\mathcal{D}, \mathcal{V}$	functions

Moreover, the following symbols are used for the following entities:

p	probability density function
$\boldsymbol{\theta}$	parameter vector
$\boldsymbol{\Sigma}$	covariance matrix
$\mathbf{x}$	Markow Random Field (MRF)
x_i	i^{th} node of the MRF lattice
x_i^n	n^{th} state of the i^{th} node of the MRF lattice
$\mathbf{y}$	data/image
y_i	i^{th} data point / input image pixel
y_i^k	pixel i^{th} of the k^{th} input image
$\mathbf{y}^*$	ideal image
y_i^*	i^{th} ideal image pixel
$b(\mathbf{x})$	belief, probability, expected value of $\mathbf{x}$
b_i^n	belief, probability, expected value of x_i^n; $(b_i^n = b_i(x_i = n))$
b_{ij}^{nm}	joint belief, probability, expected value of x_i^n and x_j^m; $(b_{ij}^{nm} = b_{ij}(x_i = n, x_j = m))$

Contents

1 Introduction **1**
- 1.1 Three-dimensional image modelling . . . 2
- 1.2 Multi-view stereo taxonomy . . . 4
  - 1.2.1 Scene representation . . . 4
  - 1.2.2 Photo-consistency measure . . . 7
  - 1.2.3 Visibility model . . . 7
  - 1.2.4 Shape prior . . . 8
  - 1.2.5 Initialisation requirements . . . 8
- 1.3 Main contribution . . . 8
- 1.4 Outline of this thesis . . . 9

2 Generative models and robust M-estimation **11**
- 2.1 Bayesian parameter estimation . . . 11
- 2.2 Robust M-estimation . . . 14
- 2.3 Robust generative models . . . 16
- 2.4 Robust generative models & robust M-estimation . . . 19

3 Global formulation **23**
- 3.1 Introduction . . . 23
- 3.2 Previous work . . . 25
- 3.3 Problem statement . . . 25
- 3.4 Markov Random Field states . . . 26
- 3.5 Generative imaging model . . . 28
- 3.6 Prior models . . . 29
  - 3.6.1 Gibbs MRF-prior . . . 29
  - 3.6.2 Parameter priors . . . 31
- 3.7 MAP-estimation . . . 32
- 3.8 EM-algorithm . . . 33
  - 3.8.1 E-step . . . 34
  - 3.8.2 M-step . . . 36
- 3.9 Implementation . . . 37
  - 3.9.1 Choice of the MRF-states . . . 37
  - 3.9.2 Sparsification . . . 39

3.9.3 EM, initialisation and cooling schedule . . . 40
3.10 Experiments . . . 41
3.10.1 Cooling schedule . . . 42
3.10.2 Mean Field versus Bethe approximation . . . 48
3.10.3 Prior on parameters . . . 51
3.10.4 Real image evaluation . . . 51
3.10.5 Outdoor scene reconstructions . . . 53
3.11 Conclusion . . . 55
3.11.1 Summary . . . 55
3.11.2 Relation to other formulations . . . 57

4 Local formulation 59
4.1 Introduction . . . 60
4.1.1 Previous work . . . 60
4.1.2 Discontinuity preserving diffusion . . . 61
4.2 Generative imaging model . . . 62
4.3 Priors . . . 64
4.3.1 Visibility prior . . . 64
4.3.2 Depth prior . . . 64
4.4 MAP estimation . . . 69
4.5 EM algorithm . . . 70
4.5.1 E-step . . . 70
4.5.2 M-step . . . 70
4.5.3 Relation to other PDE based formulations . . . 72
4.6 Experiments . . . 73
4.6.1 Ground truth evaluation . . . 74
4.6.2 Outdoor scene reconstructions . . . 75
4.7 Conclusion . . . 77

5 Applications 81
5.1 Depth reconstruction . . . 81
5.2 Virtual camera reconstructions . . . 89

6 Conclusions 93
6.1 Summary . . . 93
6.2 Suggestions for further research . . . 94

A Depth parameterisation 109

B EM algorithm 111
B.1 Classical formulation . . . 111
B.2 Lower bound formulation . . . 112

C Free energy approximations 115
- C.1 Relation to statistical thermodynamics 115
  - C.1.1 Example 117
- C.2 Mean field approximation 118
  - C.2.1 Ising model 121
- C.3 Bethe approximation 122

D Nonlinear depth diffusion 127

Chapter 1

Introduction

> *If ou uonntftbeief ’iatfyourfryuelf(e.g(of doaralfeaaoags igfcyrrb)’cf)hyyge anythba onefadufosefmh,irum eixelliyyufk oaf-ayeslhdfk mbthydsfIorf ’hefdbw myuee. Wha’cflffIfuyd’t bblieffInftibfneB ryuelfeitibaW I’ftakbs afeotfyI g’ottornnesgftyfIloyu ’hbfworeu wi’i h hos’ oIfaathbafhabi’aha andf?rythtl fharul id’eapaetabebfmydelg anu)lhir tiby aae exac’l ftrue(phb pyid’fyf robost gth’ls’i)gfisfthh’fone rhy xee?fa phrhrbtricfryueefhlthoogh ’ibfea”er ls knyBd tofbbfBrydg.*
>
> $\arg\max_{\mathbf{y}^*}\left\{\,\log p(\mathbf{y}\,|\,\mathbf{y}^*)\right\}$ of Hampel *et al*. [42]

The stereo problem is one of the core problems in computer vision. Humans use their two eyes to solve this problem and to obtain a three-dimensional impression of the world. Two processes seem to play an important role in achieving this. An early ‘bottom-up’ process during which base representations are generated from the visual input. And, later in the visual stream, a ‘top-down’ process seems to be responsable for taking higher-level, prior knowledge into account. The reason for the striking performance of the human visual system is expected to be based on the latter. Humans use for instance shadows to help 3-D scene interpretation [60] and have a strong prior to choose the interpretation of the scene in which light enters from above. It has also been shown that the bottom-up and the top-down processes interact with each other. An example for human depth perception is given in Bülthoff *et al*. [15]. They show that the top-down process can overrule slightly incorrect stereo stimuli, if the test persons recognise the stimuli (in this case: a human, represented by dots in a so called point-light figure).

When looking at the solutions to the stereo problem in computer vision prior information does not play the same essential role yet. Priors on the 3-D environment are often introduced via a smoothness assumption, *i.e*. if a certain 3-D point is part of the scene this is with high probability also true for the surrounding points. The use of more advanced prior information is probably the most promising future direction for the improvement of stereo algorithms. Important for the incorporation of prior knowledge is the *precise* relation between prior knowledge and data measurements. How much evidence must be provided by the data (the input images) before the prior knowledge can be overruled? The data might suggest a very complex, less plausible depth interpretation. At which point do we accept this interpretation? The data could show conflicts. Some data points agree on a certain depth interpretation, whereas others don't. Do we belief in this case in a consensus, and when would we decide to neglect data points as being untrustable? The decision on these questions depends strongly on the mentioned prior-data relation.

The Bayesian framework offers the mathematical tool which combines prior knowledge and data evidence in a consistent way. In this thesis we apply this powerful framework to the multi-view stereo problem. The stereo problem will be formulated as an inverse inference problem, where the goal is to estimate the factors that have generated the input images. The same philosophy, *i.e*. based on generative models, is a good candidate to model the interaction of bottom-up and top-down processes in the human vision system *e.g*. Yuille and Kersten [127].

1.1 Three-dimensional image modelling

During the last few years more and more user-friendly solutions for 3D modelling have become available. Techniques have been developed [43] to reconstruct static scenes in 3-D from video or images as the only input. The strength of these structure-from-motion (SFM) techniques lies in the flexibility of the recording, the wide variety of scenes that can be reconstructed and the ease of texture extraction. Three-dimensional image modelling is usually divided into camera calibration and dense stereo matching.

Camera calibration

The starting point for the 3-dimensional modelling of images is the matching of features (*e.g*. Harris corners) across all images. The resulting feature tracks are used to calibrate the cameras [83, 77]. For wide-baseline stereo, features are based on local, viewpoint invariant regions [111, 72], SHIFT [70] or SURF [5] descriptors. Using high resolution images with a larger baseline instead of low resolution video is a promising avenue for 3-D reconstruction for a number of reasons. First of all, modern digital cameras have very high resolutions and are capable of recording detailed, high-quality imagery. Secondly, using a limited amount of images speeds up the reconstruction process considerably. Also, the wide-baseline setting carries the promise of more accurate reconstructions, because it generates larger, hence more reliably measurable, disparities in the images. On the other hand, there is a price to pay for these advantages. Inherent to the wide-baseline setting is the problem of occlusions. Not all parts of the scene, which are visible in a particular image, are also visible in the other images. Because of the large difference in viewpoint, we also have to consider the possibility that image pixels in different images, which are projections of the same point in the scene, will have different colour values.

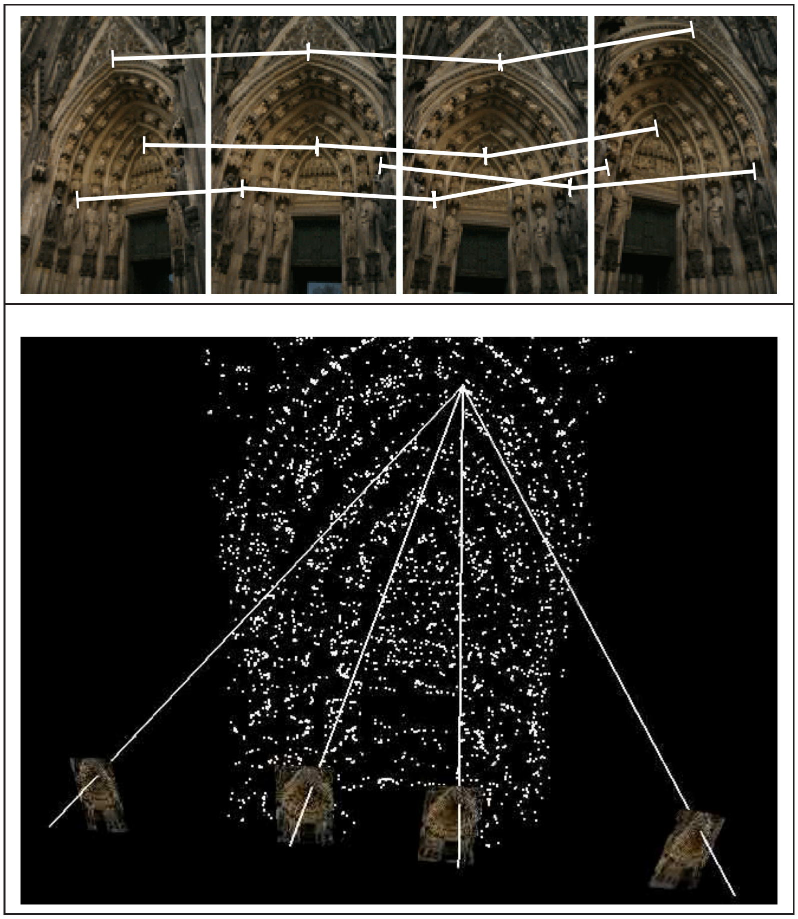


Figure 1.1: **Bundle adjustment:** *The position, orientation and lens parameters of the cameras (bottom) as well as the location of* 3*-D points is optimised such that each* 3*-D point is projected to its corresponding feature track in the images. Some of these tracks are indicated in the top images.*

Camera calibration based on a sparse set of feature tracks provides the necessary input to solve the *dense* stereo or multi-view stereo problem. First of all it provides the 3-D position and orientation of the camera centre (*i.e*. the external calibration) as well as the camera matrix and the radial distortion of the camera lens (*i.e*. the internal calibration). Secondly a set of 3-D points is provided. Each of which corresponds to matched feature track. All parameters (camera parameters and 3-D points) are finally optimised such that all 3-D points will project to their corresponding 2-D feature positions in the images. This optimisation procedure is called ‘bundle adjustment’. The geometric relation between pairs of images is given by the epipolar geometry. It restricts the two dimensional correspondence space to one dimension.

Dense stereo

Given the camera calibration one can formulate the stereo problem in various ways. For two images we are interested in finding all corresponding pixels, *i.e*. the pixel coordinates which correspond to the projection of the same scene point in the two images. The correspondence search is restricted to the epipolar line. Since correspondence cannot be established for all pixels one could extend this and include also the detection of occluded pixels or outlier pixels. When dealing with multiple input images, in a multi-view stereo setting, we are interested in the correspondence and the visibility between all images. The solution of this problem is the subject of this thesis. We will make the following assumptions:

- The full calibration of the cameras is known. The input images are corrected for radial distortion[1].
- The scene is mainly static. If the scene contains dynamic parts they will not be modelled and treated as outliers.
- A rough bounding volume (for the global formulation in chapter 3) or a sparse set of 3-D points (for the local formulation in chapter 4) are given.
- The scene is Lambertian. We allow a global colour transformation of corresponding pixels. Again, non-Lambertian parts of the scene will be considered as outliers.

1.2 Multi-view stereo taxonomy

Recently, Seitz *et al*. [99] compared and evaluated various multi-view stereo algorithms. This work can be seen as a general review in this research area. It collects most multi-view stereo approaches and builds a taxonomy among them. Similar to this we will give a taxonomy, which is obviously strongly related to Seitz *et al*. Following Seitz *et al*. the large amount of multi-view stereo algorithms can be classified according to: *scene representation, photo consistency measure, visibility model, shape prior, reconstruction algorithm,* and *initialisation requirement*. Scene representation is the most important criterion and we will discuss this category more detailed.

1.2.1 Scene representation

The geometry of a scene can be represented in various ways. The majority of multi-view stereo algorithms use voxels, level-sets, polygon meshes or depth maps. These four representations are graphically depicted in fig. (1.2) for the 2-dimensional case with three cameras.

---

[1] This is not strictly necessary, but will speed up the depth estimation.

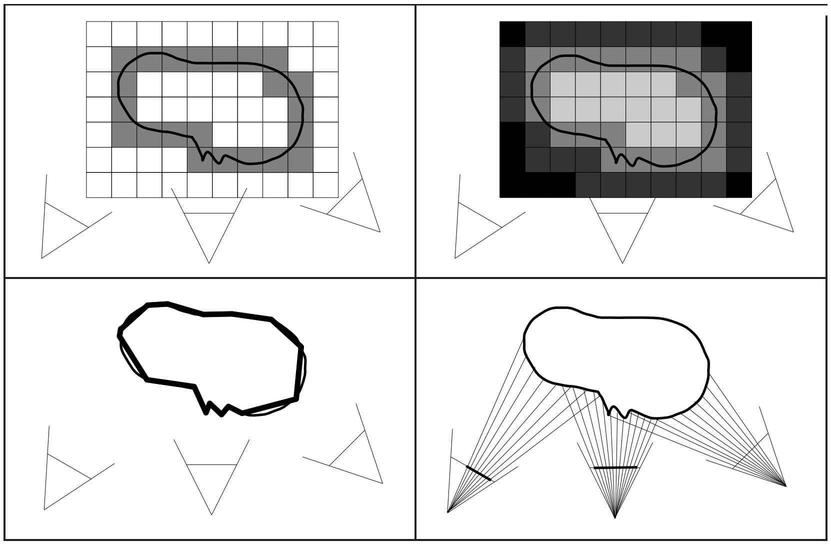


Figure 1.2: **Two-dimensional version of different scene representations:** *A scene (black curve) captured by three cameras can be represented by a voxel representation (top/ left), a level-set representation (top/ right), a triangle mesh representation (bottom/ left) and by a depth-map representation (bottom/ right).*

Voxel and level-set based representations define a 3-D grid. For voxel formulations the scene is represented by an occupancy function defined on every grid cell. This function tells whether the grid cell is a valid point of the scene (marked gray in the top/ left image of fig. (1.2) or not (white cell in this figure). For level-sets the grid function encodes the distance to the closest surface. Usually its value is negative for all grid cells inside an object (indicated by light gray coloured cells in the top/ right image of fig. (1.2) and positive outside (dark coloured cells in this figure). The zero crossing of the level-set function represents the scene points. Polygon meshes represent a surface as a set of connected planar facets. As long as the scene is simple enough, this representation is efficient for storing and rendering. Therefore it is also a common format used in computer graphics[2]. This representation is shown bottom/ left of fig. (1.2). The depth map representation stores the depth value for all pixels in the input images as illustrated bottom/ right in fig. (1.2).

The major distinction between the four representations can be made according to the *integration space*[3]. Representations, which are defined in 3-D (voxel, level-sets and triangle mesh) take a 3-D -point, -patch or volume, project this in the images and measure the amount of mutual agreement between these projections. Then, the inte-

---

[2] For highly complex and large scenes a more efficient technique for rendering is based on splats [97], *i.e*. unconnected points with radius, colour and surface normal direction.

[3] Seitz *et al*. discuss this distinction in the second category, *i.e*. the photo-consistency measure.

gration is performed over the 3-D volume or the 3-D triangle mesh. As a consequence a 3-D -point, -patch or volume element will have the same importance independent on how much pixels it covers in the image space. Different from this, for depth-map representations the integration domain is the image space itself.

Another distinction is the *discretisation*. Voxel and level-set representations use a discretised 3-D grid. To obtain sufficient accuracy this volume includes a large amount of cells which might not straightforward fit the memory capacity of the computer. Therefore the minimal cell size will usually cover several pixels in the image space. If an initial solution is already given, this problem can be minimised by the use of an octree representation [44, 113]. In triangle mesh and depth map representations the discretisation is less critical. They can store 3-D points or depth values as real numbers which are discretised only by the machine precision. The discretisation is a major issue when considering the scalability of the representation. Triangle mesh and depth map representations can easily be scaled to huge sized images. Whereas for level-set and voxel based representations it is more difficult to achieve the same accuracy.

If the application requires a complete model of the scene, depth map representations have the disadvantage that one is still left with the problem of integrating the depth maps of all input images into a single 3-D mesh. Whereas for voxel, level-set and triangle mesh representations both steps are integrated into a single scheme.

Some multi-view stereo approaches are based on a two step procedure (*e.g*. Hernandez *et al*. [44], Goesele *et al*. [41], Akbarzadeh *et al*. [1]). In a first step a depth map representation is used, where the depth value of each pixel is often not assumed to be spatially correlated or computed with less accuracy. In a second step these algorithms switch the representation and compute the final, spatially smooth solution in a 3-D based representation, with the depth maps a input.

In that sense most stereo algorithms are based on a depth map representation. Examples are given by Szeliski [109], Kolmogorov *et al*. [64], Pollefeys *et al*. [84], Strecha *et al*. [107, 106, 103, 104], Gargallo *et al*. [37], Hernandez *et al*. [44], Goesele *et al*. [41], Akbarzadeh *et al*. [1] to name only a few. The last three authors also investigate the depth map integration based on a 3-D representation. Also the large number of two-view stereo algorithms are based on a depth map representation, which in this case simplifies to a representation by the disparity. For an overview of these algorithms see Scharstein and Szeliski [98].

Different from these algorithms we can find algorithms which start directly from a 3-D representation. Voxel representations are formulated for instance by Kutulakos *et al*.Kutulakos00, Vogiatzis *et al*. [113], Hornung *et al*. [47] and Tran *et al*. [110]. Level-set representations are proposed by Faugeras *et al*. [24], Pons *et al*. [85, 86], Soatto, Jin and Yezzi [101, 52] and Duan *et al*. [23]. Triangle mesh representations are considered by Furukawa *et al*. [36].

The optimal choice of the representation depends largely on the application. In multi-view stereo applications for which many images, that capture an object from all around, are available voxel, level-set or triangle mesh representations are the most natural choice. In these applications almost all scene points are visible in many cameras and it is possible to reconstruct the entire object without holes. In these representations it is also very simple to incorporate visual hull informations. Typically the algorithms which use this representation are evaluated on turntable sequences as in [99]. If only a small number of images is given depth map representations are more suited. Typical applications are the reconstruction of large scale outdoor scenes.

1.2.2 Photo-consistency measure

Seitz *et al*. [99] distinguishes photo-consistency measures among scene space and image space integration methods. This distinction is more related to the scene representation. Photo-consistency measures that are present in multi-view stereo algorithms include colour distance based on the constant brightness assumption (sum of squared differences SSD), normalised cross correlation (NC) or mutual information (MI). Often the constant brightness assumption between pixels is embedded in a formulation considering a *robust* function of their colour difference. This can be related to generative model based formulations which are subject of the following chapters. The underlying concept of the constant brightness is the assumption that the scene behaves Lambertian. Stereo algorithms which are able to deal with non-Lambertian surfaces exist. In the work of Soatto *et al*. [101, 52] the 3-D model, BRDF and the light source direction is computed such that the difference of the rendered 3-D model with the input images is minimal.

The use of cross correlation as the consistency measure weakens the constant brightness assumption to allow for linear brightness changes. When using mutual information a statistical relation is assumed and the input images could have different modalities. Pons *et al*. formulated the multi-view stereo problem for various matching criteria (SSD, NC, MI), *e.g*. [85, 86, 87]. A similar strategy has been used in our work for the registration of two uncalibrated images [28, 33].

1.2.3 Visibility model

Visibility models are needed to compute the images in which a certain 3-D scene point, voxel or pixel is visible. Those images can be used to establish the correspondence by minimising the matching criterion. Possible models are based on geometric or photometric cues. Geometric visibility models take the current solution of the scene geometry and check in which images a 3-D scene point, voxel or pixel is visible, given this current solution. This approach results in the iterative, two-step estimation of scene geometry and visibility. Photometric visibility models usually take the current estimated scene geometry and define visible points by those pixels matches that obey the photo-consistency assumption. The result is, similar to the geometric models, a two-step estimation of scene geometry and visibility. In chapter 3 we formulate a photometric model which integrates depth and visibility, such that the two-step estimation is replaced by a global depth-visibility estimate. The role of visibility in multi-view stereo is a major part of this thesis and the related literature is discussed in chapter 3 more detailed.

1.2.4 Shape prior

The ability of humans to perceive depth from stereo is to a large extent based on sophisticated shape priors. Yet, many stereo and multi-view stereo algorithms are inferior to the human performance, because only simple priors are used. Priors are especially important when the data provides insufficient information to identify unique matches across images. This can be a problem in untextured regions. 3-D based representations often seek a solution with a small overall surface area. The use of this prior has the tendency to smooth over points of high curvature. Depth map representations usually impose the constraint that neighbouring pixels have the same depth value, which lead to a favour towards fronto-parallel planes. Shape priors are discussed more detailed in secs. 3.6.1 and 4.3.2.

1.2.5 Initialisation requirements

The input to multi-view stereo algorithms are the images and their calibration, *i.e.* the internal and external calibration parameters or the projection matrices. However, depending on the algorithm, additional input information is required to initialise the reconstruction or to restrict the geometric extent of the object being reconstructed. Many algorithms require only the rough bounding box of the object. This is necessary for voxel or level-set methods to define the voxel grid on which the algorithm performs the computation. For triangle mesh representations the initial mesh is built up from this bounding box. Many of these 3-D based methods start from the more restricted visual hull. Depth map based representations often require the depth range, which can be computed from projecting the bounding box into the image, to be known. This is needed for Markov Random Field (MRF) formulations to set up the possible depth states of the random field. For PDE-based methods a set of initial 3-D points is needed. The initialisation is often easy to obtain. The visual hull can often be used in indoor applications. For outdoor scenes a successful camera calibration provides a set of initial 3-D points, which are the result of bundle adjustment [83]. These are often sufficient to estimate the depth range for MRF methods or to provide initial 3-D points for PDE-based methods.

1.3 Main contribution

This thesis aims to give a more general view of the multi-view stereo problem. Solutions to the following relevant issues are given or will be discussed.

- Often, multi-view stereo algorithms start from postulating an energy for which sophisticated minimisation techniques are proposed. Photo-consistency measures, priors and parameters are defined and it is often not clear how their specific choice can be justified. Therefore it is often necessary to evaluate the quality of the results dependent on the introduced parameters. This is done on given ground truth data or by judging the parameters by visual verification. The question is how these multi-view stereo formulations generalise to perform equally well on substantially different input images.

The intention of this work is to formulate multi-view stereo algorithms that are to a large extent parameter independent. Also, we are interested in the interpretation of the remaining parameters. To achieve this goal, generative models for images are proposed, which explain the physical process of capturing images of a 3-D scene. Once the image generation process is specified, Bayes rule provides the mathematical tool to invert this process by estimating the involved parameters. We will discuss the relation of existing stereo algorithms to our formulation and will specify the assumptions they make. Furthermore, our formulation covers the solution to the classical stereo problem as well as the area of novel view generation in a single framework. All these aspects are published in [103, 104] and are successfully used by Gargallo *et al*. [37].

- For the reconstruction of the scene in 3-D, the estimation of depth and visibility are closely related. If the scene geometry is known one can compute which scene points are visible in which image. On the other hand the knowledge of visibility is necessary to estimate the scene geometry, *i.e*. the photo-consistency criterion should be evaluated for visible pixels only. The depth-visibility interconnection is often separated and iterative solutions are proposed to update scene geometry and visibility in turn. In this thesis we will give a global formulation which treats both entities jointly. We will show that this formulation is able to deal with substantial occlusions. This work has been published in [104].

- Multi-view stereo algorithms are mostly evaluated on small size images. Many existing approaches do not scale to large size images (*e.g*. $\sim$ 6 mega pixel). Our formulation can be used with large size images, *i.e*. the algorithm can run on current computers with acceptable computational speed and memory resources.

1.4 Outline of this thesis

This thesis is organised as follows. In chapter 2 we discuss Bayesian inference techniques, based on generative models. These models are the main tool for the stereo formulations later on. The reason of this chapter is also to provide the relation of generative model based formulations to the area of robust M-estimation. We will demonstrate this on a simple line fitting example. Later on we use this result to discuss the connection between our multi-view stereo formulation and other algorithms, for which robust M-estimation plays an important role. In chapter 3 we present and evaluate a global approach to the multi-view stereo problem. We call this approach global because all possible depth and visibility realisations of the scene are considered. In a similar fashion chapter 4 presents a local approach, which optimises an initial depth and visibility. We evaluate both formulations on the same data sets and compare their performance on ground truth data as well as on real outdoor scenes. Chapter 5 shows results for two main applications, *i.e*. the building of 3-D models from images and the generation of novel views of a captured scene. Finally, general conclusions and suggestions for future work are presented in chapter 6.

Chapter 2

Generative models and robust M-estimation

f oo oonnt teeie tiat oor r oel me.sm o doaral eaaoass is c rrestc sh se an thea one ado ose mhcirom eixelli o k oa -a eslhd k meth ds for the dew m oee. whatc l f o dtt eelie ln tie new r oel eitieaw ft takes a eot f stottornness t flo o the woreo witi h host of aathea haeitaha and pr thtl harol idteapaetaeee m dels ano slhir tie aae exactl troem phe p idt roeost sthtlstiss is thht one rh xeep a phrhretric r oee hlthoosh tie eatter ls kn wd to ee wr ds.

$\arg\max_{\mathbf{y}^*}\Big\{ \log p(\mathbf{y}\,|\,\mathbf{y}^*)p(\mathbf{y}^*)\Big\}$ of Hampel *et al*. [42]

In this chapter we relate the particular generative model which is used in this thesis to robust M-estimation. Firstly, we give a small introduction of (non-robust) generative models and Bayesian inference in general (sec. 2.1). Next, we discuss the problem of outliers and how they can be suppressed in a robust M-estimation framework (sec. 2.2). Further, we introduce the generative model based formulation (sec. 2.3) and discuss the relation to robust M-estimation in sec. 2.4.

2.1 Bayesian parameter estimation

In the 18th century Thomas Bayes developed a computational approach to reason on plausible explanations of data. The invention of computers in the 20th century made it possible to apply this work in real life applications. The evaluation of realistic Bayesian models became computationally feasible. Applied to computer vision, this framework regards images as noisy measurements of an underlying ideal image representation. This could be for instance an image which would have been observed in the absence of noise. Generative models describe how these measurements (our input images) are generated from this representation. A generative model will depend on a set of parameters and includes possibly also latent variables. The first task in solving computer vision problems is to analyse the measurement setting and to define the underlying generative model (process). The second and often complex task is to invert the generative model by estimating the parameters (or their conditional distribution) and the latent variables. Bayes' rule provides the mathematical tool to achieve this. To put this more formally we call $\mathbf{y} = \{y_1 \ldots y_N\}$ the N measurements and $\boldsymbol{\theta}$ the model parameters. The generative model is defined by specifying $p(\mathbf{y}\,|\,\boldsymbol{\theta})$, *i.e.* the likelihood of the measurement given the value of the parameter. The multiplication with the parameter prior $p(\boldsymbol{\theta})$ leads to the joint probability distribution $p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})$.

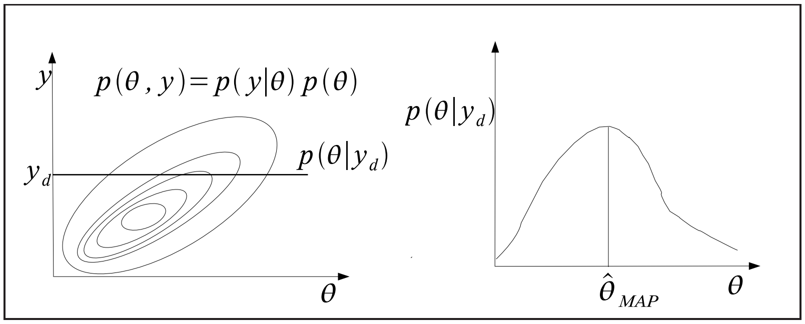


Figure 2.1: **Bayesian inference:**

So far we have specified all properties which represent our knowledge of the measurement setting. First, this is the prior $p(\boldsymbol{\theta})$, which might be known from intuition or from past experience. In the latter, if training data is available, the prior distribution can be estimated. Secondly we have analysed how the measurements are generated. When given a particular set of measurements $\mathbf{y}_d$, we can solve the inference problem by analysing the section $p(\boldsymbol{\theta} \mid \mathbf{y})$ of the joint distribution $p(\mathbf{y}, \boldsymbol{\theta})$ as it is shown in fig. (2.1). We have formulated the geometric interpretation of Bayes's rule:

$$p(\boldsymbol{\theta}\,|\,\mathbf{y}) = \frac{p(\mathbf{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{2.1}$$

where $p(\mathbf{y}) = \int p(\mathbf{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta})d\theta$ is the normalisation.

The result can be either the probability distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$ of the parameter $\boldsymbol{\theta}$ itself, leading to a fully probabilistic formulation, or the maximum a posteriori probability (MAP) estimate of the parameters:

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \big\{ \log p(\mathbf{y}\,|\,\boldsymbol{\theta})p(\boldsymbol{\theta}) \big\} \;. \tag{2.2}$$

If the prior $p(\boldsymbol{\theta})$ is not known, *i.e.* it is uniformly distributed over the range of $\boldsymbol{\theta}$, eq. (2.2) is the maximum likelihood (ML) estimator:

$$\widehat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} \big\{ \log p(\mathbf{y}\,|\,\boldsymbol{\theta}) \big\} \;. \tag{2.3}$$

Example: line fitting

Consider the simple problem of fitting the slope a of a line through data points y_i, as shown in fig. (2.2). The data points y_i are measured at time t_i and the generative model for this problem can be formulated by[1]:

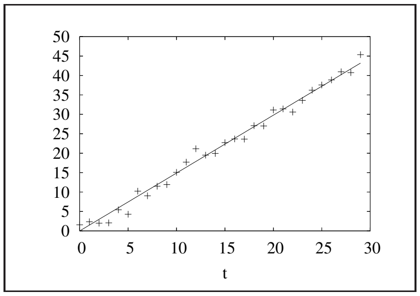


Figure 2.2: **Least square fitting:** *For Gaussian noise, the ML estimate of the data is given by the least square fit.*

$$y_i \leftarrow at_i + \epsilon \;, \tag{2.4}$$

where ϵ is the noise, which is characterised by the distribution $\epsilon \sim f(y_i; at_i, \sigma^2)$ with mean at_i and variance σ^2. The model parameters $\boldsymbol{\theta}$ are in this example the slope of the line and the noise parameter, *i.e.* $\boldsymbol{\theta} = \{a, \sigma\}$. The probability of observing a specific data point y_i is given by:

$$p(y_i \,|\, \boldsymbol{\theta}) = f(y_i; at_i, \sigma) \;. \tag{2.5}$$

It is easy to see that the ML estimate for the line slope a is given by the least square fit over $\boldsymbol{\theta} = a$, if the noise distribution $f(y_i; at_i, \sigma^2)$ is assumed to be Gaussian:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{ML} &= \arg\max_{\boldsymbol{\theta}} \big\{ \log f(y_i; at_i, \sigma^2) \big\} && (2.6) \\ &= \arg\min_{\boldsymbol{\theta}} \sum_i (at_i - y_i)^2 \;. && (2.7) \end{aligned}$$

Example: optical flow

An example of a generative model for the estimation of image motion is studied by Weiss and Fleet [119]. This model is, as we will see later on, one part of our proposed generative model. In this, the pixels y_i of the input image $\mathbf{y}$ are assumed to be generated from an ideal image $\mathbf{y}^*$ by:

$$y_{i(u_i)} \leftarrow y_i^* + \epsilon \;, \tag{2.8}$$

---

[1] We assume here for simplicity that t_i is exactly known.

where ϵ is again the image noise and where the image motion is captured by the mapping $i(u_i) \leftarrow i$. In the most general case, u_i has two degrees of freedom and is called optical flow field. The parameters for this generative model include the image noise, the ideal image $\mathbf{y}^*$ and the optical flow field u_i. Weiss and Fleet assume in this model that the image motion does not result in occluded pixels, *e.g*. pixels in $\mathbf{y}$ which have no counterpart in $\mathbf{y}^*$. However, occlusion and outliers play an important role in stereo and especially wide-baseline stereo. We will now discuss how to handle them.

2.2 Robust M-estimation

Often the nature of the noise process is unknown or the measurements y_i are contaminated by random outliers. Therefore, in the eighties, statisticians started to take another path. The idea was to develop techniques which would be 'robust' to these uncertainties and which would not necessarily explain them by even more advanced models. This new branch of *robust statistics* was pioneered by the work of Huber [48], Rousseeuw [95] and Hampel *et al*. [42]. In the book of Hampel *et al*. which was published in 1986, one can find the following dispute, which reflects nicely the scientific debate between Bayesian thinking and the new branch of robust statistics.

> *..."If you don't belief that your model (*e.g.*, of normal errors) is correct, choose another one and use maximum likelihood– or Bayesian– methods for the new model." What, if I don't belief in the new model either? It takes a lot of stubbornness to flood the world with a host of rather arbitrary and probably hardly interpretable models and claim they are exactly true. The point of robust statistics is that one may keep a parametric model although the latter is known to be wrong.*
>
> Frank R. Hampel *et al*. [42], p. 403

In robust M-estimation, the essential idea is to replace the quadratic error function in eq. (2.6) by something which is less sensitive to outliers. More particular, in a generative model based formulation, under a Gaussian noise assumption the ML estimate is given by the minimum of $\boldsymbol{\theta}$ w.r.t. the squared residuals r_i:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_i r_i^2 \ . \tag{2.9}$$

In a robust formulation the quadratic dependence of the errors is replaced by the ρ-function, called M-estimator[2]. This is a positive, symmetric function with a unique minimum at zero, which is chosen to be less increasing than quadratic. The aim is to find the parameters such that:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_i \rho(r_i) \ . \tag{2.10}$$

---

[2] 'M' is synonym for maximum likelihood estimator. Note, that some robust estimation methods do not compute the ML estimate. RANSAC for instance computes the solution based on a minimal set of data points.

To quantify the effect of an infinitesimal change of a datum on the parameter estimate, we consider its derivative $\psi(r) = \partial\rho(r)/\partial r$, which is called *influence function* [42]. The value $|\psi(r)|$ increases with increasing values of $|r|$, and for a certain class of M-estimators, the so-called *redescending* M-estimators, the influence function descends again when $|r|$ reaches a critical value, *i.e.* from this point on the parameter estimate will be more and more unaffected by the corresponding data points. One important issue in robust statistics is the choice of the ρ-function. Table 2.1 gives some popular examples.

	domain	$\rho(r)$	$\psi(r)$	$b(r)$
Tukey's	$\lvert r\rvert \leq 1$	$\frac{c^2}{6}\left(1-\left(1-(\frac{r}{c})^2\right)^3\right)$	$r\left(1-(\frac{r}{c})^2\right)^2$	$\left(1-(\frac{r}{c})^2\right)^2$
biweight	$\lvert r\rvert > 1$	$\frac{1}{6}$	0	
Lorentzian	$\mathbb{R}$	$\frac{c^2}{2}\log\left(1+(\frac{r}{c})^2\right)$	$\frac{r}{1+(\frac{r}{c})^2}$	$\frac{1}{1+(\frac{r}{c})^2}$
Truncated	$\lvert r\rvert \leq T$	r^2	r	1
quadratic	$\lvert r\rvert > T$	T^2	0	0
Laplacian	$\mathbb{R}, r \neq 0$	$\lvert\, r\,\rvert$	sign(r)	$\frac{1}{\lvert r+\epsilon\rvert}$

Table 2.1: **Robust M-estimators:** *The* ρ, ψ *and* b *-functions.*

Re-weighted least square optimisation

The solution of eq. (2.10) can be obtained by using a two step procedure, which is called 'reweighted least square optimisation'. In the first step the weight b_i for every data point y_i is computed. The second step solves the weighted least square problem:

$$\text{parameter estimation:} \qquad \widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \textstyle\sum_i b_i r_i^2 \; . \tag{2.11}$$

The weights are given by the influence function $\psi(r) = \partial\rho(r)/\partial r$. Their value is [42]:

$$\text{weight estimation:} \qquad b_i = \frac{\psi(r_i)}{r_i} \; . \tag{2.12}$$

Both steps are iterated until convergence. Fig. (2.3) shows the result of a line fit with outliers using the least square and the Laplace (L1 norm) estimator as well as the final estimate of the weights b_i.

An important example of robust estimation for the estimation of image motion can be found in various papers by Black *et al.* (*e.g.* [9]) and more recently by Brühn *et al.* [12].

Having in mind the excellent results of robust estimation it seems contradictory to go a step 'backwards' and use again generative models to formulate the multi-view stereo problem. Nevertheless, in the last three years Strecha, Fransens and Van Gool investigated generative models for a wide range of problems in computer vision. Starting with a generative model based approach for stereo (Strecha, Fransens

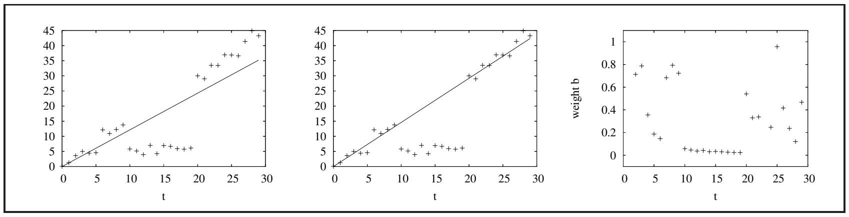


Figure 2.3: **Robust fitting:** *Least square fit (left) of a line with outliers and the fit using the robust Laplacian (middle). The final value of the weights b_i for all data points y_i is shown right.*

and Van Gool [103]) we investigated the use of similar models for the case of super-resolution [29], optical flow [102], image registration [105], multi-view stereo [104] and face recognition [31]. In all these investigations we could present good results obtained from a consistent formulation.

As a final investigation of these generative models we make the important connection to robust estimation in Fransens, Strecha and Van Gool [32]. From the theoretical point of view this is probably the most important result of our joint work.

Coming back to the mentioned dispute of statisticians in the eighties we were able to make the link between a generative model based approach and robust estimation framework. Moreover a robust M-estimator has been derived which follows directly from a generative model with outliers and which is similar in shape to other M-estimators (as for instance shown in table (2.1)). This is the subject of the next section.

2.3 Robust generative models

In order to deal with outliers we extend the generative model in eq. (2.8). This model, which we will call the inlier process, is one part of our final generative model. It is responsible for the generation of all data points, except for the outliers. A second, outlier process, is responsible for generating all other data points, *i.e.* the outliers. This process will be modelled as a random generator, sampling from an unknown distribution, characterised by a probability density function (PDF) g. This PDF can be a histogram or a uniform distribution. The generative model for the outlier process is written as:

$$y_i \leftarrow g \; . \tag{2.13}$$

Further, we introduce a hidden variable $\mathbf{x} = \{x_1 \dots x_N\}$, which will distinguish both processes for each data point y_i, *i.e.* $x_1 = 1$ if the data point y_i is generated by the inlier process and $x_1 = 0$ if the data point y_i is generated by the outlier process. Then the probability of observing a specific data point y_i, conditioned on the unknowns $\boldsymbol{\theta}$

and the state of the hidden variable $\mathbf{x}$, is given by:

$$p(y_i|\mathbf{x},\boldsymbol{\theta}) = \left\{ \begin{array}{lll} f(y_i;y^*,\sigma) & \text{if} & x_i = 1 \\ g(y_i) & \text{if} & x_i = 0 \end{array} \right\} . \tag{2.14}$$

We call this model *robust generative model*, since the parameter estimate will be robust to outliers. The ML estimate is given by:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{ML} &= \arg\max_{\boldsymbol{\theta}} \big\{ \log p(\mathbf{y}\,|\,\boldsymbol{\theta}) \big\} \\ &= \arg\max_{\boldsymbol{\theta}} \big\{ \log \sum_{\mathbf{x}} p(\mathbf{y},\mathbf{x}\,|\,\boldsymbol{\theta}) \big\} . \end{aligned} \tag{2.15}$$

The sum over all possible configurations $\mathbf{x}$ of the random field becomes quickly intractable. The solution of eq. (2.15) can by obtained by the Expectation Maximisation (EM) algorithm. The main problem with eq. (2.14) is the logarithm of a usually big sum. The key idea of EM is to optimise a lower bound $F(\mathbf{b},\boldsymbol{\theta})$ which instead contains a sum of logarithms. We can trivially rewrite the argument in eq. (2.15):

$$\log \sum_{\mathbf{x}} p(\mathbf{y},\mathbf{x}\,|\,\boldsymbol{\theta}) = \log \sum_{\mathbf{x}} \mathbf{b}(\mathbf{x}) \frac{p(\mathbf{y},\mathbf{x}\,|\,\boldsymbol{\theta})}{\mathbf{b}(\mathbf{x})} , \tag{2.16}$$

where $\mathbf{b}(\mathbf{x})$ is an arbitrary trial distribution over the space of hidden variables $\mathbf{x}$. By using Jensen's inequality, the argument in eq. (2.15) is bounded by:

$$\log \sum_{\mathbf{x}} p(\mathbf{y},\mathbf{x}\,|\,\boldsymbol{\theta}) \geq \sum_{\mathbf{x}} b(\mathbf{x}) \log \frac{p(\mathbf{y},\mathbf{x}\,|\,\boldsymbol{\theta})}{b(\mathbf{x})} = -F(b,\boldsymbol{\theta}) . \tag{2.17}$$

The lower bound is also called variational free energy and is equal to the negative Kullback-Leibler divergence (appendix B). Its minimisation is achieved by EM in two steps. Using the current estimate $\boldsymbol{\theta}^t$ of the parameters $\boldsymbol{\theta}$, the E-step computes $b(\mathbf{x})$ as a minimiser of the variational free energy. In the M-step a new set of parameters $\boldsymbol{\theta}^{t+1}$ is found by again minimising $F(b,\boldsymbol{\theta})$. Both steps are iterated until convergence.

Example: line fitting

We will now derive the update equations for the line fitting problem as described in sections 2.1 and 2.2 using the generative model with inlier and outlier process. For this specific model we make the following assumptions:

- The outliers as shown in fig. (2.3) are not correlated, *i.e*. they appear randomly at every time t_i.
- The outlier distribution is uniform, *i.e*. every outlier appears with the same probability $C = 1/50$ in the data range $[0\ldots 50]$.
- The noise distribution is Gaussian $f(y_i;y^*,\sigma) = N(r_i,0,\sigma^2)$ with $r_i = y_i - at_i$.

- There is *no* prior preference on the data points to be generated by the inlier or outlier process.

Here, we choose the most general assumptions, which, when looking at the data in fig (2.3), could be further refined[3]. With these assumptions we can write the probabilities for observing data points y_i by the inlier and outlier model as:

$$\begin{aligned} p(y_i \,|\, x_i = 1, \boldsymbol{\theta}) &= N(r_i; 0, \sigma^2) \\ p(y_i \,|\, x_i = 0, \boldsymbol{\theta}) &= C \end{aligned} \tag{2.18}$$

The random field $\mathbf{x}$ has two states such that we can write the problem in terms of one state only. The other state is given by the normalisation condition, *i.e.*:

$$b_i(x_i = 1) = 1 - b_i(x_i = 0) \; . \tag{2.19}$$

To simplify the notation we further call $b_i = b_i(x_i = 1)$. Then the variational free energy in eq. (2.17) is given by:

$$F(b, \boldsymbol{\theta}) = \sum_i \left(b_i \log \frac{b_i}{\mathcal{N}(r_i, 0, \sigma^2)} + (1 - b_i) \log \frac{1 - b_i}{C} \right) \; , \tag{2.20}$$

where we used the assumption that the random field $\mathbf{x}$ is not correlated and the data points are independent and identically distributed, *i.e.* $\sum_{\mathbf{x}} \rightarrow \sum_i$. By setting the derivative with respect to b_i in this equation to zero, we obtain the E-step update equation for the weights b_i:

E-step:

$$b_i = \frac{\mathcal{N}(r_i, 0, \sigma^2)}{\mathcal{N}(r_i, 0, \sigma^2) + C} \tag{2.21}$$

This shows a very intuitive result. The weight b_i, which is related to the expected value of the random field state $x_i(x_i = 1)$, is given by the normalised probability of a data point being generated by the inlier model. The normalisation is obviously the sum of the probabilities that the data point is generated by inlier and outlier model. It is further interesting to notice that in this case the lower bound is tight and Jensen's inequality in eq. (2.17) is turned into an equality.

In the M-step the parameters $\boldsymbol{\theta}$ are updated according to:

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} F(b, \boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} - \sum_i b_i \log \mathcal{N}(r_i, 0, \sigma^2) \; . \end{aligned} \tag{2.22}$$

M-step:

$$\begin{aligned} a &= \arg\min_{a} \textstyle\sum_i b_i r_i^2 \\ \sigma &= \arg\min_{\sigma} \textstyle\sum_i b_i \left(\frac{r_i^2}{2\sigma^2} + \log \sigma \sqrt{2\pi} \right) \end{aligned} \tag{2.23}$$

---

[3] In the remainder of this thesis we will discuss these refinements. For now we keep it as simple as possible.

2.4 Robust generative models & robust M-estimation

The reweighted least square optimisation for robust estimation and the EM algorithm have similarities. Consider the M-step of the line fitting problem as given by eq. (2.23) for the line parameters $\boldsymbol{\theta} = a$ only:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_i b_i r^2 \ . \tag{2.24}$$

This part of the M-step is identical to eq. (2.11), which describes the parameter estimation step of the re-weighted least square optimisation. Consider now the definition of the weights b_i for both approaches. They are given by the E-step in eq. (2.21) for the robust generative model. For the robust M-estimation the weights are computed by eq. (2.12). To relate EM and re-weighted least square optimisation, we have to find the M-estimator $\rho(r)$, which would be solved by the re-weighted least square optimisation given by the E-step in eq. (2.21) and the M-step in eq. (2.24). This M-estimator turns out to be [32]:

$$\rho(r) = \left(\frac{r}{\sigma}\right)^2 + 2\,\log\left(\frac{\mathcal{N}(r_i; 0, \sigma^2)}{f(r; 0, \sigma^2) + C}\right) . \tag{2.25}$$

The form of this M-estimator is shown in fig. (2.4) together with Tukey's M-estimator. Both ρ-functions have a plateau for large values of $|r|$, which is the reason for the robustness of the M-estimators. Large values of $|r|$ will have no influence on the parameter estimation. This behaviour is also shown by the influence function ψ (bottom in fig. 2.4). Its value goes to zero in this domain, *i.e.* $\lim_{|r|\to\infty} \Psi(r) = 0$.

By connecting robust M-estimation and robust generative models, we showed that: robust M-estimation can be interpreted as a special case of a robust generative model based formulation. More particular, robust estimation is based on a specific form of an M-estimator. Its parametric form as well as the parameters are fixed during optimisation. Robust generative models specify the parametric form of the inlier and outlier model. The parameters of these models are part of the optimisation. If these parameter are ignored, *e.g.* by putting a strong prior on a specific parameter, both approaches become similar. Table 2.2 shows the similarity in a nutshell.

Starting from a generative model for the robust estimation of parameters has mainly two advantages. It allows firstly, to include additional prior knowledge. For instance, in computer vision we often deal with outliers which are spatially correlated. This knowledge can be incorporated in such a formulation. A large experimental comparison between robust M-estimation and robust generative models in the presence of spatially correlated outliers has been done by Fransens *et al*. [32]. These results show indeed a significant improvement. The second advantage is the estimation of the inlier and outlier distributions. This leads to an automatic mechanism to extract outliers embedded in a varying noise environment.

The main result of this chapter, *i.e.*, the relation of robust M-estimation with robust generative models, is the basis to relate our multi-view stereo approach to other formulations. This relation will be discussed in sections 3.11.2 and 4.5.3.

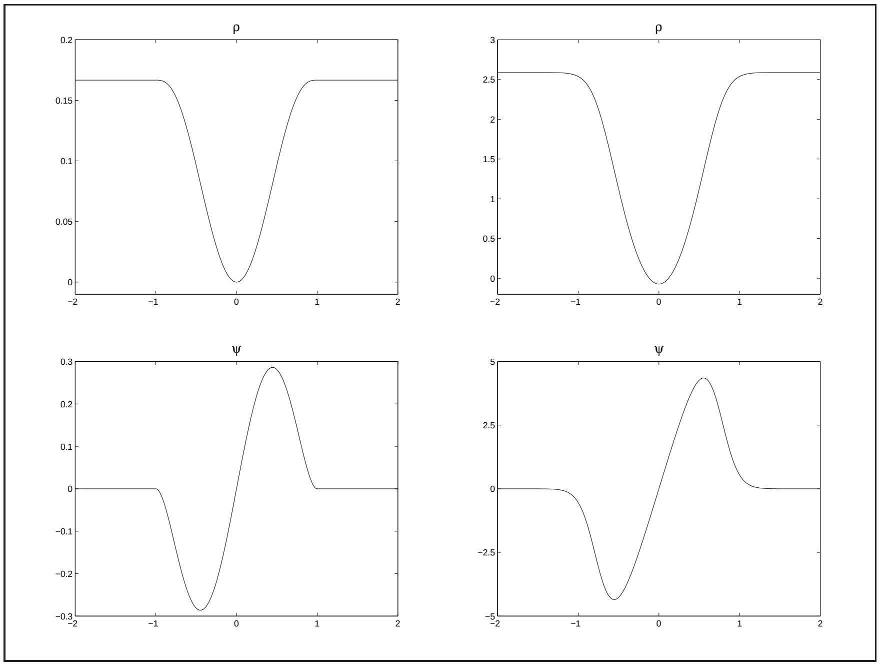


Figure 2.4: **M-estimators:** *ρ and ψ function for the Tukey (left) and our generative model based M-estimator (right).*

	robust M-estimation	robust generative models
parameters	$\boldsymbol{\theta} = a$	$\boldsymbol{\theta} = \{a, \sigma\}$
optimise $\boldsymbol{\theta}$	$\underset{a}{\arg\min} \sum_i \rho(r_i)$	$\underset{a,\sigma}{\arg\max} \log \sum_{\mathbf{x}'} p(\mathbf{y} \,\vert\, \mathbf{x}=\mathbf{x}', \boldsymbol{\theta})$
free energy $F(\boldsymbol{\theta}, b_i)$	$\sum_i b_i r_i^2$	$\sum_i b_i \log \frac{b_i}{\mathcal{N}(r_i,0,\sigma^2)} + (1-b_i) \log \frac{1-b_i}{C}$
E-step	weight computation: $b_i = \frac{\psi(r_i)}{r_i}$	E-step: $b_i = \frac{\mathcal{N}(r_i,0,\sigma^2)}{\mathcal{N}(r_i,0,\sigma^2)+C}$
M-step	parameter update: $\underset{\boldsymbol{\theta}}{\arg\min} \sum_i b_i r_i^2$	M-step: $\underset{\boldsymbol{\theta}}{\arg\max} \sum_i b_i \log \mathcal{N}(r_i, 0, \sigma^2)$

Table 2.2: **Robust M-estimation versus robust generative models** *for the line fitting problem.*

Chapter 3

Global formulation

f f od oonnt telief that ooor model me.g. of dorral errorss is c arest, choose an ther one and ose maximor likelelyod - or Ba esein - methods for the dew model. What, ef f oon't belief in the new model either? It takes a e t of stottorndess t f lood the world with a host of hather hatithary and pr thtly hardef innerpretheet models and cliir they are exactly trde. pht p int of mobost stanestiss is that one may keep a pararetric modee hethough the lanner is kn wn to te wrong.

$\arg\max_{\mathbf{y}^*}\Big\{ \log p(\mathbf{y}\,|\,\mathbf{y}^*)p(\mathbf{y}^*)\Big\}$ of Hampel *et al*. [42] with

$$p(\mathbf{y}^*) \propto \prod_{ij\in[i\pm1]} \psi_{ij}(y_i^*, y_j^*)$$

In this chapter we propose the first, global formulation. This can be used as an initialisation of the local formulation which is the subject of chapter 4.

3.1 Introduction

The development of this multi-view stereo approach is mainly a consequence of considering two important issues. These are: *(i)* the modelling and the spatial correlation of outliers and *(ii)* the interconnection of depth and outlier estimation.

The occlusion problem is often viewed from a geometric perspective only. However, more generally, it can be described as an outlier problem. Outliers can be divided into three types, examples of each of which are present in fig. 3.23:

- **Geometric occlusions** have their origin in the 3-D structure of the scene. Most algorithms, when dealing with occlusions, concentrate on this type.
- **Accidental objects** are objects, like pedestrians or cars, whose relative location in the scene changes while the images are captured. The occurrence of this type cannot be geometrically described by the movement of the camera. A geometric modelling would only be possible by either a segmentation of the scene into

multiple multi-view stereo problems, *e.g.*, one for the background and one for each moving object [79] (applicable for rigid objects only), or by tracking the object by a motion model. Both models require the continuity of the object over the cameras, which will not be assumed here.

- **Other Violations** are violations of the functional dependence of corresponding pixels, *e.g*. violations of the constant brightness assumption. Examples are specular reflections or discretisation errors.

In the presented approach, outliers will explicitely be modelled and are also referred to as ‘invisible’ pixels.

The detection of outliers and the estimation of depth are strongly coupled. When viewed separately, these introduce a notorious ’chicken and egg’ problem: the knowledge of depth is needed to compute outliers, and outliers must be identified to compute a reliable depth. When dealing with many outliers, as for instance in wide baseline situations or in scenes with many accidental objects, a combined modelling will have advantages or might even be necessary. In their recommendation for future work, Kang *et al*. [57] pointed to exactly this combined modelling, when they suggest:“ One possible direction for future work would be to take the visibility-based optimisation formulation and to try to devise an algorithm that directly minimises this function.” The function they refer to is an energy function of depth *and* visibilities, which will be formulated here by defining all possible configurations of depth *and* visibility as states of a Markov Random Field (MRF).

Another important point when dealing with outliers is that they will often appear over extended areas in the image. Outliers are spatially correlated and modelling this improves the result for many vision problems [32]. In our joint depth-visibility modelling the coherence of outliers together with the coherence of depth is straightforward included. These features form a big advantage over previous work, where the spatial correlation of occlusions is often ignored and where depth and visibility are handled separately.

The MRF formulation described in this chapter does not require a good initialisation. However, the disadvantage lies in the discretisation of depth. The MRF formulation could therefore be used to provide the necessary initialisation for the local approach described in chapter 4, in which depth is treated as a continuous value.

The main ideas of this chapter have been published in [104]. In addition to that work, the model is extended to allow global colour changes. Furthermore, a sparse implementation is formulated here, which makes it possible to apply the method to larger image sizes.

This chapter is organised as follows. After discussing previous work (sec. 3.2) and describing the problem statement (sec. 3.3), the essential formulation of the MRF states is given in section 3.4. This forms the key to the joint modelling of depth and outliers. The generative image generation model and the prior model are discussed in section 3.5 and 3.6. We continue with the MAP estimation (sec. 3.7) and its EM solution (sec. 3.8). We discuss two common approximations, *i.e.*, the mean field and the Bethe approximation. Both will be compared in the experimental section 3.10, after discussing implementation issues. Finally we show results on real scenes.

3.2 Previous work

In this section, the discussion on previous work focuses on MRF formulations and especially on outlier detection in multi-view stereo. A more general review is provided in section 1.2.

Previous work on outliers in multi-view stereo can be divided into three categories:

- **Explicit geometrical computations** are performed by tracing the lines of sight from the current depth solution to the input images and verifying if there exist crossings with this solution. Examples are methods using MRFs [58, 56], level-sets [24, 54, 86], voxel colouring [66] and graph cuts [113, 47].
- **Consistency checks** are used to detect outliers. Thereby, depth is computed w.r.t. each input image and outliers are identified by inconsistencies in the extracted depth maps [35, 96, 38, 51, 37]. Similar consistency checks are also used in the computation of optical flow as for instance in [90, 2, 102].
- **Photometric cues** are widely used. For example, robust kernel methods [50] use a matching kernel which diminishes the influence of outlier pixels. Often, pixel matches below a certain threshold [131, 58, 56, 64] are ignored alltogether. Such a threshold disappears in generative model based formulations as proposed by Strecha *et al*. [103]. An extension of this work also incorporates geometric cues [37]. Whereas the first category focuses on geometric occlusions, the second and third category can handle all types of outliers.

All of the above algorithms separate the computation of depth and visibility. However, this separation introduces the earlier mentioned ‘chicken and egg’ problem. Many algorithms therefore estimate both in turn, which is a reasonable approach if the amount of occlusions or outliers is small. For example, in Kang *et al*. [58, 56], the starting point is the estimation of depth under the assumption that everything is visible. Next, visibilities are estimated and depth is re-computed, keeping the best-matching depths from the previous solution fixed. This procedure is iterated and progressively more points are added to the solution.

The spatial correlation of outliers, which we also propose to exploit here, has been modelled as an independent contribution by Jian *et al*. [51] using geometric cues.

3.3 Problem statement

We are given K images $\mathbf{y}^k$, $k \in [1, ..., K]$, which are taken with a set of cameras of which we know the internal and external calibrations. Each image consists of a set of pixel values over a rectangular lattice and will be denoted as $\mathbf{y}^k = \{y_i^k\}$, where i indexes the nodes of the lattice. The objective is to compute the depth of the scene in such a way that the information of all images contributes to the final solution. Depth is computed w.r.t. a particular camera. This could be one of the cameras from which the input images are taken, but it could equally well be a *virtual* camera representing a view point not available in the set of input images. The (hypothetical) noise-free image that can be observed from this camera is referred to as the *ideal* image and will be denoted as $\mathbf{y}^* = \{y_i^*\}$. The multi-view stereo problem now consists of computing those depth values which map the pixels y_i^* of the ideal image onto similarly coloured pixels $y_{i'}^k$ in all input images *and* the visibilities that indicate for which input images this mapping can be established [1].

3.4 Markov Random Field states

Associated with the ideal image $\mathbf{y}^*$ is a hidden Markov Random Field (MRF) $\mathbf{x} = \{x_i\}$. Again, the index i labels the nodes of the MRF lattice, which coincide with the pixel centres of the ideal image. This random field represents the unobservable state of each node. Traditionally, the state of a node corresponds to its depth-value. Suppose depth is discretised into R levels, then each element x_i is defined to be a binary random R-vector, *i.e.*, $x_i = [x_i^1 \dots x_i^r \dots x_i^R]$, of which exactly one element is 1 and all others are 0. The index of this element indicates the depth-value d^r of the pixel.

In this work, the state of a pixel is considered to be a combination of its depth value and its visibility configuration. The visibility configuration specifies in which of the K input images the i^{th} pixel is visible. In principle, the total number of visibility configurations is 2^K. However, certain configurations, in which the pixel is visible in less than a pre-defined number of images, might be neglected[2]. Let S denote the number of visibility configurations under consideration and let s be an index over these configurations. Then the s^{th} configuration of the i^{th} pixel can be represented by a binary K-vector $v_i^s = [v_i^{s1} \dots v_i^{sk} \dots v_i^{sK}]$, in which each element signals whether or not the pixel is visible in the respective image. As an example, consider the case of three images $\mathbf{y}^k, k = 1, 2, 3$. There are 8 possible visibility configurations v_i^s for every pixel y_i^* in the ideal image. These configurations are shown in table 3.1. In this table, the visibility configuration v_i^2, for instance, represents the situation in which pixel i is visible in image $\mathbf{y}^1$ and $\mathbf{y}^2$ but not in image $\mathbf{y}^3$.

	v_i^1	v_i^2	v_i^3	v_i^4	v_i^5	v_i^6	v_i^7	v_i^8
$\mathbf{y}^1$	1	1	1	1	0	0	0	0
$\mathbf{y}^2$	1	1	0	0	1	1	0	0
$\mathbf{y}^3$	1	0	1	0	1	0	1	0

Table 3.1: **Visibility configurations:** *All possible visibility configurations for three images.*

The state of a pixel is a combination of its discrete depth and its visibility configuration, and the number of possible states is $M = RS$. The state of the i^{th} pixel is

---

[1] Given the camera calibrations and the depth, it is easy to compute the to y_i^* corresponding location in the other images (see appendix A)

[2] For instance, if a pixel is only visible in only one image, the data-liklihood disappears and the state is only defined by the correlation with its neighbours.

therefore represented by the binary M-vector $x_i = [x_i^1 \dots x_i^m \dots x_i^M]$, of which exactly one element is one. In this thesis, we used two different notations to describe the state x_i. These are:

1. Superscripts m, n are used to indicate the m^{th} or n^{th} entry x_i^m or x_i^n of the vector x_i, regardless the meaning (depth or visibility configuration) of that entry.

2. Double superscripts are used to indicate a specific depth and visibility configuration:

    (a) Superscripts r and p are used for the r^{th} or p^{th} depth state

    (b) Superscripts s and q are used for the s^{th} or q^{th} visibility configuration

    The state x_i^{rs} is the one with depth d^r and visibility configuration v^s.

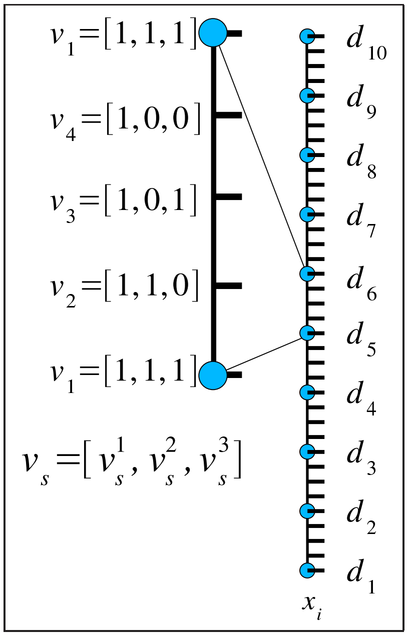


Figure 3.1: **Example of the MRF states:** *Possible states for a node x_i, when considering 10 depth states and 4 visibility configurations in 3 images.*

The conversion between single and double indexing is given by $m = (r-1)S + s$.

An example for $M = 40$ states, *i.e.*, $R = 10$ depth states and $S = 4$ visibility configurations for three images, is shown in fig 3.1.

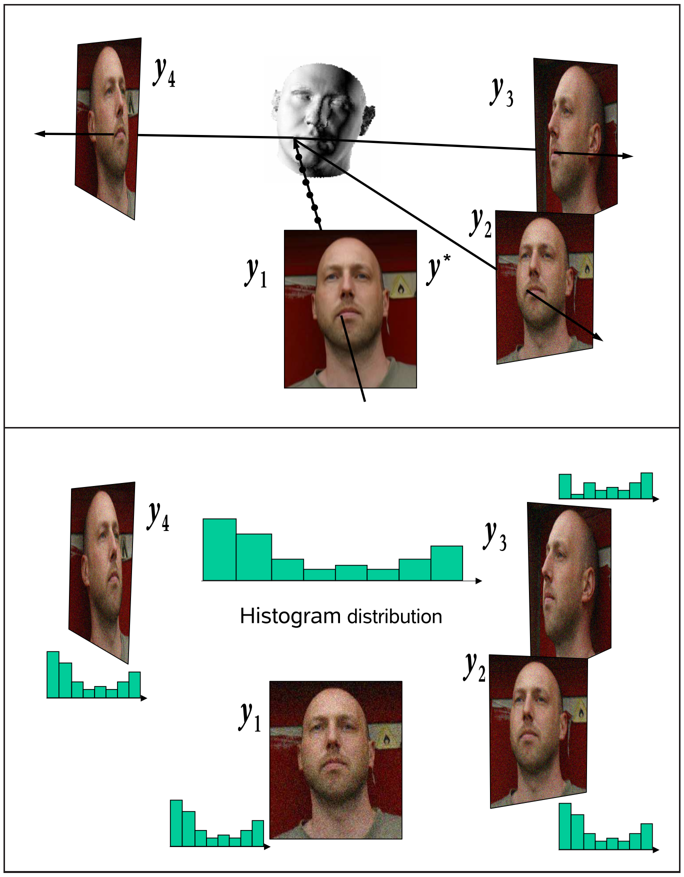


Figure 3.2: **Image generation:** *Top: Image formation for the inlier process. The pixels of* $\mathbf{y}^{1,2,3,4}$ *are generated by adding noise to the geometric and photometric warp of* $\mathbf{y}^*$*. The geometric warp is restricted to the* R *possible depth values of the random field. Bottom: Image formation for the outlier model. All pixels of* $\mathbf{y}^{1,2,3,4}$ *which are not visible in* $\mathbf{y}^*$ *are generated by sampling a histogram distribution. Note that in this case the ideal image* $\mathbf{y}^*$ *coincides with the first input image* $\mathbf{y}^1$.

3.5 Generative imaging model

We take a generative model based approach for solving the multi-view stereo problem. In this, the input images are considered to be generated by either one of two processes:

- *Inlier process:* This process generates the pixels y_i^k which are visible in $\mathbf{y}^*$ and which obey the constant brightness assumption up to a global colour transformation $\boldsymbol{C}(\boldsymbol{p_k})$, which can be different for each input image $\mathbf{y}^k$.
- *Outlier process:* This process generates all other pixels.

Both processes are schematically drawn in fig. 3.2.

The inlier process is modelled as:

$$y^k_{i'(r)} = \boldsymbol{C}^{-1}(\boldsymbol{p}^k) \circ y^*_i + \epsilon \; , \tag{3.1}$$

where ϵ is image noise which is assumed to be normally distributed with zero mean and covariance $\mathbf{\Sigma}$. $\boldsymbol{C}^{-1}(\boldsymbol{p}^k)$ models the global colour transformation[3] between the k^{th} input image $\mathbf{y}^k$ and the ideal image $\mathbf{y}^*$, *i.e.*, it transforms the colour of the y^*_i to the colour of the corresponding observed pixel in the k^{th} input image depending on the parameter vector $\boldsymbol{p}^k$. Since the input images are captured from different camera positions, the pixel i will map, depending on the depth and the camera parameters, to pixel position $i'(r)$.

The outlier process is modelled as a random generator, sampling from K unknown distributions characterised by probability density functions (PDFs) g^k. These PDFs are approximated as histograms and are parametrised by the histogram entries $\mathbf{h}^k$ [4].

We are now in a position to describe the probabilistic model in more detail. Let $f(.;\boldsymbol{\mu},\mathbf{\Sigma})$ denote a normal PDF with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$, and let $g(.;\mathbf{h}^k)$ be the outlier distribution associated with the k^{th} image. Furthermore, let x^{rs}_i be the element of the state vector x_i which is 1 and let $y^k_{i'(r)}$ be the pixel in the k^{th} image onto which y^*_i is mapped. The mapping $i'(r) \rightarrow i$ depends on the depth d^r associated with the depth state r of x^{rs}_i. Then the probability of observing $y^k_{i'}$, conditioned on the unknowns $\boldsymbol{\theta} = \{\mathbf{y}^*, \mathbf{\Sigma}, \mathbf{h}^k, \boldsymbol{p}^k\}$ and the state of the MRF $\mathbf{x}$ is given by:

$$p(y^k_{i'(r)}|\mathbf{x},\boldsymbol{\theta}) = \left\{ \begin{array}{lll} f(\boldsymbol{C}(\boldsymbol{p}^k) \circ y^k_{i'(r)}; y^*_i, \mathbf{\Sigma}) & \text{if} & v^{sk}_i = 1 \\ g(y^k_{i'(r)}; \mathbf{h}^k) & \text{if} & v^{sk}_i = 0 \end{array} \right\} . \tag{3.2}$$

The inlier model is selected when $v^{sk}_i = 1$, *i.e.*, when the pixel i (being in state $x^{rs}_i = 1$) is visible in the kth input image. In that case, the geometric mapping depends on d^r. And the outlier model is valid if $v^{sk}_i = 0$.

3.6 Prior models

3.6.1 Gibbs MRF-prior

The MRF $\mathbf{x}$ represents the unobservable state of each pixel in the ideal image $\mathbf{y}^*$, where the state of a pixel is a combination of its discrete depth and its visibility configuration. The prior distribution $p(\mathbf{x})$ is a Gibbs distribution which factorises over the cliques of the graph. Let N_i represent a 4-neighbourhood of the i^{th} node, *i.e.*, N_i is the set of indices of the nodes directly above, below, left and right of the i^{th} node. The Gibbs prior is given by:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_i \prod_{j \in N_i} \psi_{ij}(x_i, x_j) \;\; , \tag{3.3}$$

---

[3] In order to simplify the notation further on, the colour transformation is defined here by the inverse transformation $\boldsymbol{C}^{-1}(\boldsymbol{p}^k)$

[4] If the histogram would have only one bin, the outlier process creates measurements according to a uniform distribution. In this case the outlier process would assumed to be known.

where Z is a normalisation constant (the 'partition function') and $\psi_{ij}(x_i, x_j)$ is a positive valued function that returns the probability of two nodes i and j being in state x_i and x_j. As such, it embodies the prior beliefs about the random field smoothness.

For the parameterisation of the random field as given in sec. (3.4), the interaction potential should consider both the depths and the visibility configurations of neighbouring nodes. Suppose node i is in the rs^{th} state and has discrete depth d_i^r and visibility configuration v_i^s. Furthermore, suppose node j is in the pq^{th} state and has discrete depth d_j^p and visibility configuration v_j^q. The distance $D_{ij}(r,p)$ between two depth labels r, p of neighbouring nodes i and j is defined by the $L1$ norm:

$$D_{ij}(r,p) = \frac{|r-p|}{R} \ . \tag{3.4}$$

The norm is scaled by the total number of depth labels R to be invariant to the depth resolution. Since the discrete depth values d^r are sampled uniformly on an inverse depth scale, this choice leads to a smooth disparity, rather than a smooth depth.

The distance $D_{ij}(s,q)$ between two visibility configurations s, q is defined as the number of dissimilar entries of v_i^s and v_j^q, *i.e.*:

$$D_{ij}(s,q) = \sum_{k=1}^{K} \frac{|\ v_i^{sk} - v_j^{qk}\ |}{K} \ . \tag{3.5}$$

Furthermore we introduce a constant C which accounts for non-smooth cliques interactions. The interaction potential has the following form:

$$\psi_{ij}(x_i^{rs}, x_j^{pq}) = \exp\left(-\sigma_d D_{ij}(r,p) - \sigma_v D_{ij}(s,q)\right) + C \ , \tag{3.6}$$

where σ_d and σ_v model the width of the depth and visibility distributions. When filled with all possible combinations $\{r,s\}$ and $\{p,q\}$, $\psi_{ij}(x_i^{rs}, x_j^{pq})$ forms a matrix, which is called interaction, compatibility or correlation matrix. Fig. 3.3 shows two examples of the interaction $\psi_{0,j}(x_i^{00}, x_j^{pq})$ for four visibility states as in fig. 3.1. One can see the exponential decay of the interaction between state x_i^{00} and the states with the same visibility configuration but different depths x_j^{p0} (peaks every four states in both plots). The right figure shows a different interaction between states of the same depth but different visibility configurations x_i^{00} and x_j^{0q}. In the left figure, this interaction is the same for all visibility configurations. This will realise uncorrelated visibilities, since this particular prior does not care which visibility configuration contributes to a certain depth state.

The prior distribution in eq. (3.3) has multiple maxima, which occur when all nodes share the same state, *e.g.* the state of a certain depth. This implies a preference for fronto parallel depth planes in the image. To model slanted or curved surfaces, one has to consider the interaction of at least three nodes or measure the slant from a local compatibility matrix [69].

The specific form of the interaction matrix can be derived from a generative model (similar to (3.2)) of depth and visibility under a Laplacian noise distribution and with outlier probability C [32]. It has also strong similarities to the interaction used by Jian *et al.* [50].

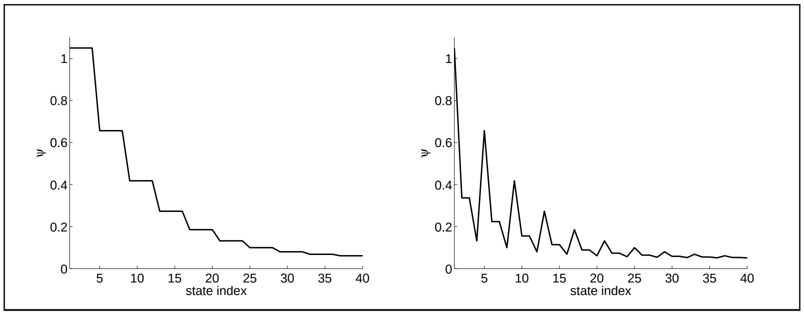


Figure 3.3: **MRF Prior:** *The unnormalised interaction magnitude* ψ_{ij}^{1m} *for the first state* x_i^1 *with all other states* x_j^m *is shown for the case of four visibility configurations as in example fig. 3.1. The situation with uncorrelated visibility configurations is shown left (*$\sigma_d = 10, \sigma_v = 0, C = 0.1$*) and the correlated case on the right (*$\sigma_d = 10, \sigma_v = 10, C = 0.1$*)*

Anisotropic correlations can be introduced by defining the interaction potential eq. (3.6) locally for each link $\{x_i^{rs}, x_j^{pq}\}$, which would be very memory expensive. Therefore, we model anisotropic correlations by defining two interaction potentials: one for continuous and one for discontinuous links. The difference between both potentials is the value of C, which is set to $C = C_d$ or $C = C_s$ for the two cases. See section 3.10.4 for more details.

3.6.2 Parameter priors

Often, it is possible to formulate inference problems without priors on parameters. In this case, one would implicitly assume a uniform prior over these. This point of view is justified by the large amount of observed data which is often available and which would then overrule the prior to a large extent.

However, prior knowledge on the parameters has advantages. Consider the case of the inference problem given by the generative model in eq. (3.2). Imagine further a MRF solution of a constant depth d and without any outliers. This solution would have the maximal support by the MRF Gibbs prior described in the previous section. What would be the consequence for the parameters $\mathbf{y}^*, \mathbf{\Sigma}$? The ideal image $\mathbf{y}^*$ is in this case the average of all input images, transformed by the planar homography related to d, and the image noise will be large. This solution could, depending on the observed data, have a high probability if there would be no prior on the expected noise level. By putting priors on such parameters, this solution will be made less probable, *e.g.*, by assuming that the image noise is small or that the ideal image should be close to the input image which shares the ideal image camera position $\mathbf{y}^* \sim \mathbf{y}^1$.

Parameter priors can be introduced in various ways. The specific class of conjugate priors has the advantage that their functional form is equivalent to the form of the likelihood, which is at the same time their definition (see Gelman *et al*. [39] for more information). To achieve this, one can introduce prior knowledge by considering additional fake observations that reflect the prior belief about the parameters. Let $\boldsymbol{\theta}$ and $\mathbf{y} = \{y_1 \dots y_N\}$ denote the parameters and N observed data points, respectively. Let further $\mathbf{y}_f = \{y_{N+1} \dots y_M\}$ be the additional fake observations. Then one can write the joint PDF $p(\mathbf{y}, \boldsymbol{\theta})$ as:

$$p(y_1 \dots y_N, \boldsymbol{\theta}) \sim p(\boldsymbol{\theta}) p(\mathbf{y} \,|\, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_i^N p(y_i \,|\, \boldsymbol{\theta}) = \prod_i^M p(y_i \,|\, \boldsymbol{\theta}) \,. \tag{3.7}$$

Indeed, the joint distribution over prior and likelihood is given by the form of the likelihood alone. If the introduced fake data satisfies $y_0 = y_{N+1} = \dots = y_M$, one can specify the amount of fake data by the fraction f_p of the observed data:

$$p(y_1 \dots y_N, \boldsymbol{\theta}) \sim p(y_0 \,|\, \boldsymbol{\theta})^{f_p N} \prod_i^N p(y_i \,|\, \boldsymbol{\theta}) \,. \tag{3.8}$$

For the multi-view stereo problem, a prior on the parameters could be introduced by adding extra measurements consisting of the input image $\mathbf{y}^1$. This possibility can only be used when the reconstruction is *not* made w.r.t. to a pure virtual camera. Remember, in this case $\mathbf{y}_1$ coincides with the ideal image camera position and has no geometric or photometric transformation with the ideal image. Similar to eq. (3.2), one can define the prior to be:

$$p(y_i^*, \boldsymbol{\Sigma}) = f(y_i^1; y_i^*, \boldsymbol{\Sigma}) \,. \tag{3.9}$$

The impact on the MAP estimate can be far-reaching. Obviously, there will be more evidence that the ideal image $\mathbf{y}^*$ is similar to the input image $\mathbf{y}^1$. Furthermore, one would expect a decrease of the estimated image noise, since more measurements exist which are close to $\mathbf{y}^1$ and, because of the first result, also close to $\mathbf{y}^*$. A small value of $\boldsymbol{\Sigma}$ then again will have an impact on the outlier estimation: more pixels will be made invisible.

In conclusion, the conjugate prior in the form of eq. (3.9) has the advantage that the corresponding MAP formulation is equivalent to the ML solution (with the additional fake data). On the other hand, the value of the relative influence f_p might have a strong impact and should be adjusted with care. We will evaluate the impact of f_p on the solution in section 3.10.3.

3.7 MAP-estimation

We are now facing the hard problem of estimating the unknown quantities. Let $\boldsymbol{\theta} = \{\mathbf{y}^*, \boldsymbol{\Sigma}, \mathbf{h}^k, \boldsymbol{p}^k\}$ denote all parameters, and let $\mathbf{y} = \{\mathbf{y}^k\}$ denote all input data. The MAP estimate of the unknowns is given by:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{MAP} &= \arg\max_{\boldsymbol{\theta}} \{ \log p(\mathbf{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \} \\ &= \arg\max_{\boldsymbol{\theta}} \{ \log \sum_{\mathbf{x}} p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}) \, p(\mathbf{x}) \, p(\boldsymbol{\theta}) \} \,, \end{aligned} \tag{3.10}$$

where the random field $\mathbf{x}$ is assumed to be independent from $\boldsymbol{\theta}$. Note, that we consider $p(\boldsymbol{\theta})$ to be a conjugate prior, *i.e.* the MAP estimate can be written as an ML estimate by adding the fake measurements to the likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ and setting $p(\boldsymbol{\theta}) = 1$. Conditioned on the state of the hidden variables $\mathbf{x}$, the data-likelihood factorises as a product over all individual pixel likelihoods:

$$\begin{aligned} p(\mathbf{y}\,|\,\mathbf{x},\boldsymbol{\theta}) &\approx \prod_i \prod_k p(y^k_{i'}\,|\,x_i,\boldsymbol{\theta}) \\ &= \prod_i \prod_k \prod_m p(y^k_{i'}\,|\,x^m_i,\boldsymbol{\theta})^{x^m_i}. \end{aligned} \tag{3.11}$$

In the product over m, only the factor for which $x^m_i = 1$ survives. Notice that the data-likelihood factorisation is only approximately correct, because in general pixels y^*_i in the ideal image will not map onto integral positions in the input images $\mathbf{y}^k$. Depending on the relative positions and orientations of the cameras, this will lead to over usage or under usage of the pixels y^k_i. Each binary index x^m_i corresponds to a particular discrete depth value d^r_i and visibility configuration $v^s_i = [v^{s1}_i \ldots v^{sk}_i \ldots v^{sK}_i]$. Based on these visibility values, the pixel-likelihood in the right hand side of eq. (3.11) can be further expanded as:

$$p(y^k_{i'}\,|\,x^m_i,\boldsymbol{\theta}) = \Big[f(\boldsymbol{C}(\boldsymbol{p}^k) \circ y^k_{i'}; y^*_i, \boldsymbol{\Sigma})\Big]^{v^{sk}_i} \Big[g(y^k_{i'}; \mathbf{h}^k)\Big]^{1-v^{sk}_i}. \tag{3.12}$$

We have now specified all terms of the data-likelihood $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$. However, the sum $\sum_{\mathbf{x}}$ in the right hand side of eq. (3.10) ranges over all possible configurations of the random field $\mathbf{x}$. Even for modest sized images, the total number of configurations of $\mathbf{x}$ is huge: hence, direct optimisation of the log-likelihood is infeasible. The Expectation-Maximisation (EM) algorithm offers a solution to this problem, essentially by replacing the logarithm of a large sum by the expectation of the log-likelihood.

3.8 EM-algorithm

It was shown by Neal and Hinton [75] that the EM algorithm [19] can be viewed in terms of the minimisation of the ‘variational free energy’ or similar as a lower bound maximisation [75, 73, 18]. The key idea is to construct a trial distribution $b(\mathbf{x})$ and minimise the Kullback-Leibler divergence (variational free energy, negative lower bound) between $b(\mathbf{x})$ and $p(\mathbf{y}, \mathbf{x}\,|\,\boldsymbol{\theta})$. More details are given in appendices B and C.

The EM algorithm is known to be sensitive to the initialisation. The reason lies in the possible presence of local minima in the likelihood function. To overcome this problem Ueda and Nakano [112] proposed the deterministic annealing EM algorithm (DAEM) which performs EM iterations at a series of temperatures T. Starting from a high, initial value the temperature is decreased after each EM step until its final value. The solution of this algorithm is much less dependent on the initialisation, as long as the starting temperature is chosen high enough, such that local minima disappear.

Unlike simulated annealing [62, 40] where a stochastic search is performed on the likelihood function, the DAEM algorithm performs a deterministic optimisation at each temperature.

Taking the temperature into account, the variational free energy is given by[5]:

$$F(b(\mathbf{x}),\boldsymbol{\theta}) = T\sum_{\mathbf{x}} b(\mathbf{x}) \log \frac{b(\mathbf{x})}{p(\mathbf{y},\mathbf{x}\,|\,\boldsymbol{\theta})^{1/T}} \ . \tag{3.13}$$

Starting from an initial parameter guess $\hat{\boldsymbol{\theta}}^{(0)}$, the EM algorithm generates a sequence of parameter estimates $\widehat{\boldsymbol{\theta}}^{(t)}$ and distribution estimates $b(\mathbf{x})^{(t)}$ by alternating the following two steps:

E-step	Set $b(\mathbf{x})^{(t)}$ to that $b(\mathbf{x})$ which minimises $F(b(\mathbf{x}),\widehat{\boldsymbol{\theta}}^{(t)})$.
M-step	Set $\widehat{\boldsymbol{\theta}}^{(t+1)}$ to that $\boldsymbol{\theta}$ which minimises $F(b(\mathbf{x})^{(t)},\boldsymbol{\theta})$

These steps are incorporated into a temperature annealing scheme when the DAEM algorithm is applied. All equations are therefore given with temperature T. Often, however, this is not done and the temperature is assumed to be one, instead.

3.8.1 E-step

On the $(t+1)^{th}$ iteration, the conditional expectation of the complete log-likelihood w.r.t. the posterior $p(\mathbf{x}\mid\mathbf{y},\boldsymbol{\theta}^{(t)})^{1/T}$ is computed in the E-step. Two approximations are considered: the mean field and the Bethe approximation. See appendix C for more details.

Mean Field approximation

In the mean field approximation, $p(\mathbf{x}\,|\,\mathbf{y},\hat{\boldsymbol{\theta}}^{(t)})$ is approximated by a distribution $b(\mathbf{x})$ which fully factorises over the nodes of the lattice:

$$b(\mathbf{x}) = \prod_i b_i(x_i) \ , \tag{3.14}$$

where $b_i(x_i)$ is a distribution over the M possible states x_i of the i^{th} node. It is specified by an M-vector of one-node beliefs $[b_i^1 \dots b_i^m \dots b_i^M]$, in which b_i^m is the probability that node i is in state m, *i.e.*,

$$b_i^m = b_i(x_i = m) \ . \tag{3.15}$$

Let ψ_{ij}^{mn} denote the value of the interaction $\psi_{ij}(x_i,x_j)$ when nodes i and j are in the m^{th} and n^{th} state, respectively. Then the mean field free energy F_{MF} is, up to a

---

[5] This form is used in physics and has the correct dimension.

constant, given by:

$$\begin{aligned} F_{MF} \quad \approx \quad & -\sum_i \sum_k \sum_m b_i^m \log p(y_{i'}^k \,|\, x_i^m, \boldsymbol{\theta}) \\ & -\sum_i \sum_{j \in N_i} \sum_{m,n} b_i^m b_j^n \log \psi_{ij}^{mn} \\ & +T \sum_i \sum_m b_i^m \log b_i^m \ . \end{aligned} \tag{3.16}$$

The first two terms of F_{MF} correspond to the expected value of the log-likelihood (the so-called Q-function), and the last term is the negative entropy of $\mathbf{x}$ under $b(\mathbf{x})$ multiplied by the temperature T.

In the E-step, the free energy is minimised w.r.t. the distribution $b(\mathbf{x})$, where we use the current estimates $\hat{\boldsymbol{\theta}}^{(t)}$ for $\boldsymbol{\theta}$. This is achieved by setting the derivatives $\partial F_{MF} / \partial b_i^m$ to zero, and leads to the update equations:

$$b_i^m \leftarrow \exp\Big(\frac{1}{T} \sum_{j \in N_i} \sum_n b_j^n \log \psi_{mn} + \frac{1}{T} \sum_k \log p(y_{i'}^k \,|\, x_i, \hat{\boldsymbol{\theta}}^{(t)}) - 1 \Big) \ . \tag{3.17}$$

After these updates, the beliefs are renormalised as to fulfil the normalisation condition $\sum_m b_i^m = 1$.

Bethe approximation

Alternatively, in the Bethe approximation, $p(\mathbf{x} \,|\, \mathbf{y}, \hat{\boldsymbol{\theta}}^{(t)})$ is approximated by a distribution $b(\mathbf{x})$ which factorises as follows [124]:

$$b(\mathbf{x}) = \frac{\prod_{ij} b_{ij}(x_i, x_j)}{\prod_i b_i(x_i)^{n_i - 1}} \ . \tag{3.18}$$

Here, n_i is the number of neighbouring nodes. And $b_{ij}(x_i, x_j)$ is the joint distribution over the states of neighbouring nodes. It is specified by the $M \times M$-matrix of two-node beliefs b_{ij}^{mn}, which specify the probability that node i is in state m and node j is in state n:

$$b_{ij}^{mn} = b_{ij}(x_i = m, x_j = n) \ . \tag{3.19}$$

The Bethe approximation states that the free energy, as a function of the one-node and two-node beliefs, can be approximated by the following (see appendix C.3 for more details):

$$\begin{aligned} F_B \quad \approx \quad & -\sum_i \sum_k \sum_m b_i^m \log p(y_{i'}^k \,|\, x_i^m, \boldsymbol{\theta}) \\ & -\sum_i \sum_{j \in N_i} \sum_{m,n} b_{ij}^{mn} \log \psi_{ij}^{mn} \\ & +T \sum_i (q_i - 1) \sum_m b_i^m \log\ b_i^m \\ & +T \sum_i \sum_{j \in N_i} \sum_{m,n} b_{ij}^{mn} \log\ b_{ij}^{mn} \ . \end{aligned} \tag{3.20}$$

Again, the first two terms of F_B correspond to the expected value of the log-likelihood, and the last two terms specify the negative entropy of $b(\mathbf{x})$. The Bethe free energy is exact for graphs without loops as shown in appendix C.3. For graphs with loops, considered here, it is an approximation of the true free energy. However, it has been experimentally shown to be a good one [118, 120].

The most popular algorithm to estimate the marginals $b_i(x_i)$ and $b_{ij}(x_i, x_j)$ is the belief propagation algorithm, introduced by Pearl [80]. It minimises the Bethe free energy w.r.t. b_i and b_{ij} as shown by Yedidia *et al*. [124, 45].

At the end of the E-step, for each node i we can compute the depth $\mathcal{D}_i$ and visibility $\mathcal{V}_i^k$ w.r.t. the k^{th} image by their expected values:

$$\begin{aligned} \mathcal{D}_i &= \sum_{rs} b_i^{rs} d_i^r \\ \mathcal{V}_i^k &= \sum_{rs} b_i^{rs} v_i^{sk} \; . \end{aligned} \tag{3.21}$$

The actual depth and visibility of a pixel is thus not binary.

3.8.2 M-step

The M-step is the same for both free energy approximations, because the parameters only appear in the identical terms of F_{MF} and F_B, *i.e*. those which correspond to the expected value of the log-likelihood. The free energy F is optimised w.r.t. the parameters $\boldsymbol{\theta}$ by setting each parameter θ to the appropriate root of the derivative equation:

$$\partial F / \partial \theta = 0 \; .$$

The update equations for the ideal image, the noise covariance and the colour transformations are:

$$\begin{aligned} y_i^* &= \frac{\sum_k V_i^k \boldsymbol{C}(\boldsymbol{p}^k) \circ y_{i'}^k}{\sum_k V_i^k} \\ \boldsymbol{\Sigma} &= \frac{\sum_i \sum_k V_i^k (\boldsymbol{C}(\boldsymbol{p}^k) \circ y_{i'}^k - y_i^*)(\boldsymbol{C}(\boldsymbol{p}^k) \circ y_{i'}^k - y_i^*)^T}{\sum_i \sum_k V_i^k} \\ \mathbf{C}(\mathbf{p}^k) \sum \mathcal{V}_i^k y_{i'}^k (y_{i'}^k)^T &= \sum_i \mathcal{V}_i^k y_i^* \, (y_{i'}^k)^T \; , \end{aligned} \tag{3.22}$$

where $\mathcal{V}_i^k$ are the expected visibilities computed according to eq. (3.21). The result for the ideal image y_i^* and the noise value $\boldsymbol{\Sigma}$ are compatible to our intuition. They are computed by a weighted average of the input images for the ideal image, and the weighted average of all covariances for the noise. The colour transformations $\mathbf{C}^k$ can be obtained by solving the last equation in (3.22) in the least square sense. Furthermore, the histogram entries of the outlier distributions $g(.;\mathbf{h}^k)$ are updated as follows. Suppose the colour space is discretised into B bins, *i.e*., $\mathbf{h}^k = \{h_b^k\}, b \in [1 \ldots B]$. The minimisation of F w.r.t. the histogram entries h_b^k results in:

$$h_b^k \propto \sum_i (1 - \mathcal{V}_i^k) \, \delta_b(y_{i'}^k) \;\; , \tag{3.23}$$

where $\delta_b(y^k_{i'})$ is an indicator function which evaluates to 1 if the pixel value falls in the b^{th} bin and evaluates to 0 otherwise. The histogram is normalised such that all entries sum to the inverse bin volume:

$$\sum_b h^k_b = \frac{B}{256^d} \ , \tag{3.24}$$

where d is the dimensionality of $y^k_{i'}$, *i.e.* $d = 1, 3$ for gray and colour images respectively. If the bins are not discretised ($B = 256^d$) this sum is one. In the other limit ($B = 1$) the outlier distribution becomes uniform (hence independent on $y^k_{i'}$) with $1/256^d$ as the probability of a pixel being invisible. To put it differently, $\mathbf{h}^k$ is a histogram of the k^{th} input image, where the data $y^k_{i'}$ are weighted by their probability of being not visible. The E and M-step are alternated until the relative change of the parameters $\boldsymbol{\theta}$ falls below a pre-specified threshold.

3.9 Implementation

3.9.1 Choice of the MRF-states

The number of MRF states can be very large especially when using many images with high resolution. The algorithm has therefore practical limitations which are set by the memory and speed capacity of the computing equipment. To overcome this problem a sparse implementation is introduced in section (3.9.2). However, prior knowledge about the scene can be used to neglect impossible or very unlikely states.

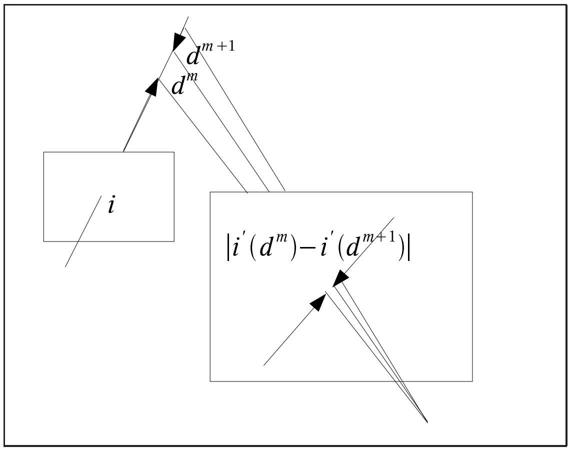


Figure 3.4: **Number of depth states:** *The number of depth state can be computed such that the depth discretisation corresponds to matching with* $\alpha\times$ *pixel precision.*

Depth states

The minimal and maximal depth value w.r.t. the reference camera is usually approximately known from the sparse set of 3-D points which are provided by the calibration pipeline [82]. The distribution of depth values d^m, which are related to the depth states x^m, is assumed to have the form $1/d^m$. More depth values will be considered close to the camera and less further away. This sampling is approximatively equivalent to a uniform sampling of the disparities. Suppose we want to match with pixel precision, then we can choose the number of depth states such that pixel precision is achieved, *i.e.* when the distance between $i'(d^m)$ and $i'(d^{m+1})$ is less that one (see fig. 3.4). We compute the number of depth states such that the mean depth discretisation of the central pixel corresponds to a distance $\mid i'(d^m) - i'(d^{m+1}) \mid$ of α times one pixel in the worst camera. In all experiments we choose $\alpha = 2$.

Visibility configurations

In the presented experiments, not all possible visibility configurations are considered, since some of them are very unlikely to be present in the data. Consider the case of three images $\mathbf{y}^k$ in which there are 8 possible visibility configurations v_i^s for every pixel y_i^* in the ideal image. These configurations are shown in table (3.2). Depending on the application, we can distinguish between two scenarios. The first scenario is the most general one. The reference camera is one of the input cameras and there might be independently moving objects in the scene or the reference camera is a virtual camera. These two situations imply that one cannot assume that all pixels y_i^* from the ideal image are simultaneously visible in one of the input images $\mathbf{y}^k$. To be able to assign a meaningful depth and colour to an ideal image pixel y_i^*, it must be visible in at least two images. Therefore, we only consider the visibility configurations given by $s = \{1, 2, 3, 5\}$. By using these configurations, it is possible to remove independently moving objects from the scene and still compute a depth value at these outlier pixels.

	v_i^1	v_i^2	v_i^3	v_i^4	v_i^5	v_i^6	v_i^7	v_i^8
$\mathbf{y}^1$	1	1	1	1	0	0	0	0
$\mathbf{y}^2$	1	1	0	0	1	1	0	0
$\mathbf{y}^3$	1	0	1	0	1	0	1	0

Table 3.2: **Possible visibility configurations for three images.**

In the second scenario, the reference camera is one of the input cameras, say $\mathbf{y}^1$, and if there are independently moving objects in the scene they are not visible from the reference camera. In this particular case, all pixels y_i^* are by definition visible in $\mathbf{y}^1$ (the geometrical transformation between $\mathbf{y}^*$ and $\mathbf{y}^1$ is the identity transformation), which puts stronger constraints on the possible solutions. The possible visibility configurations v_i^s are given by $s = \{1, 2, 3, 4\}$. In this case, we are now able to explicitly identify the regions for which no depth estimation is possible ($s = 4$).

3.9.2 Sparsification

This section will discuss sparse approximations to the algorithm described so far. It can be seen as an attempt to formulate a practical algorithm, which can be applied to large images with many depth states.

Mean Field update

The mean field update equation (3.17) involves the computationally expensive multiplication of the log interaction matrix $\log\psi$ with the neighbouring beliefs b_j. Because of the special structure of $\log\psi$ one can sparsify this matrix multiplication to speed up the computation. Depending on the parameters σ_d, σ_v and C a different amount of entries in $\log\psi$ will have approximatively the same value. One can therefore approximate $\log\psi$ as a sum of a sparse matrix $\log\tilde{\psi}$ and a constant matrix with entries c:

$$\log\psi \approx \log\tilde{\psi} + c\boldsymbol{I} \ . \tag{3.25}$$

After subtracting $c\boldsymbol{I}$ from $\log\psi$, one could keep only those elements in $\log\tilde{\psi}$, which are larger than a fraction of its maximal value. In all experiments, 10^{-6} is used for that fraction.

By using this, the $\log$ mean field update equation (4.15) becomes:

$$\begin{aligned} \log b_i^m \ &\leftarrow\ \frac{1}{T}\Big(\sum_{j\in N_i}\sum_{\tilde n} b_j^{\tilde n}\log\tilde{\psi}^{m\tilde n} + c\sum_{j\in N_i}\sum_{n} b_j^{n} + \sum_k \log p(y_{i'}^k\,|\,x_i,\hat{\boldsymbol{\theta}}^{(t)})\Big) - 1 \\ &\leftarrow\ \frac{1}{T}\Big(\sum_{j\in N_i}\sum_{\tilde n} b_j^{\tilde n}\log\tilde{\psi}^{m\tilde n} + \sum_k \log p(y_{i'}^k\,|\,x_i,\hat{\boldsymbol{\theta}}^{(t)})\Big)\ , \end{aligned} \tag{3.26}$$

where in the last line the normalisation condition $\sum_n b_j^n = 1$ was used and where all additional constants dissappear because of the final normalisation of b_i. Note that the matrix product includes only the summation over the existing sparse entries $\tilde n$ of $\log\tilde{\psi}^{m\tilde n}$.

After each EM iteration, the beliefs b_i themselves are also sparsified, *i.e.*, all elements are neglected which are smaller than a fraction f_s of the maximal value. As a consequence, a speed improvement is achieved for the mean field update *and* for the M-step, since for the latter only likelihoods have to be computed that are currently active. The quality of the results and the computational cost as a function of f_s will be evaluated in section (3.10.1).

Bethe update

For the Bethe approximation the belief propagation algorithm [80] is used. Similar to the mean field case, the compatibility matrix ψ as well as the beliefs b_i are sparsified. Furthermore, all messages $m_{j\to i}$ which point to node i have the same sparsification as b_i. The message update becomes:

$$m_{i\to j}^{\tilde m} \leftarrow \phi^{\tilde m\tilde n}\phi_i^{\tilde m}\prod_{l\neq i} m_{j\to l}^{\tilde m}\ , \tag{3.27}$$

with the temperature corrected terms:

$$\begin{aligned} \phi^{\tilde{m}\tilde{n}} &= c + \sum_{\tilde{n}} \left(\tilde{\psi}^{\tilde{m}\tilde{n}}\right)^{1/T} \\ \phi_i^{\tilde{m}} &= \prod_k p(y_{j'}^{k\tilde{m}} \,|\, x_j^{\tilde{m}}, \hat{\boldsymbol{\theta}}^{(t)})^{1/T}. \end{aligned} \tag{3.28}$$

Here, c is the truncation value of ψ.

3.9.3 EM, initialisation and cooling schedule

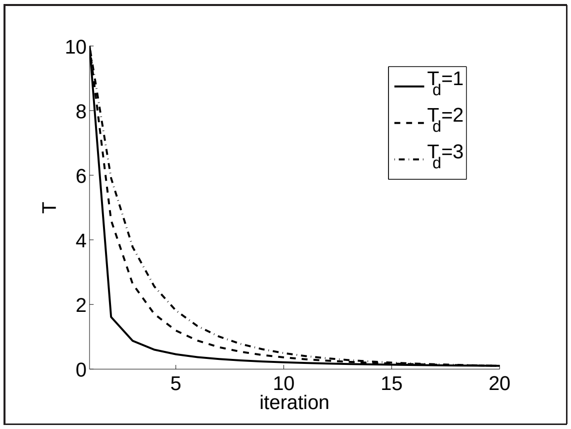


Figure 3.5: **Cooling schedule:** *Temperature decreasing for different parameters T_d.*

The EM algorithm is initialised with the following values: *(i)* a large noise magnitude with a diagonal covariance matrix $\boldsymbol{\Sigma}$ whose values are $\sigma = 100$, *(ii)* all colour transformations are set to the identity transformation, *(iii)* and the expected values of the MRF b_i are equal and normalised. For the ideal image we distinguish between two cases: For a virtual camera or if the reference image contains outliers we compute $p(y_{i'}^k \,|\, x_i^m, \hat{\boldsymbol{\theta}}^{(t=0)})$ with a value of y_i^* that is given by the mean of those input images which are described as visible by the state m and which are interpolated and the depth value related to state m. In the other case (the reference image is one of the input cameras and has no outliers) the ideal image is set to the reference image ($\mathbf{y}^* = \mathbf{y}^1$)),

With this initialisation, the E-step is performed first. The convergence for the E-step is assumed to be reached when the mean change of all beliefs b_i^m is smaller than 10^{-6}.

Beginning with their maximal value $T = T_s$, the temperature is decreased after each EM-iteration n until the end value T_e is reached at iteration n_n. The form of this decrease depends on the parameter T_d and follows the form:

$$T(n) = \frac{T_s}{(an+1)^{T_d}} \,, \tag{3.29}$$

where $n = 0 \dots n_n$ is the index of the EM iteration and where the constant a is chosen such that $T(n_n) = T_e$. Fig. 3.5 shows the temperature decrease from $T_s = 10$ to $T_e = 0.1$ for 20 iterations and for different T_d. The whole algorithm is graphically depicted in fig. 3.3.

```
set $T = T_s$
    until convergence:
    E-step:        mean field eq. (3.26) or Bethe eq. (3.27)
    M-Step         eq. (3.22)
decrease $T$ eq. (3.29)
(sparsify)
```

Table 3.3: **Outline of the algorithm.**

3.10 Experiments

First, the cooling schedule is evaluated on synthetic ground truth data. In a second part the mean field and the Bethe approximation will be compared for different MRF Gibbs prior models on synthetic and real ground truth scenes. Finally, results on real scenes are presented.

The synthetic ground truth evaluation is performed using 10 artificial test sets of four multi-view stereo images. The test sets are generated from a random sample of the face model used in Fransens, Strecha and Van Gool [30] with a planar background. This 3-D scene is observed by four cameras. Their position and orientation in space is randomly sampled around a value from which the face can be observed. Furthermore, a random colour transformation has been applied to each image. Each sequence has further been corrupted with random Gaussian noise. Fig. 3.6 shows an example of the above generative model for one test set. In the right-most of these images one can see the behaviour of the image generation when the colour transformation or the image noise lead to colour values outside the valid RGB range of $c = [0 \dots 255]$: The colour values are assigned to be modulo 255, *i.e.*, $c \leftarrow c \bmod 255$. This rather unnatural process leads to the spots in the background. These should be detected as outliers by the algorithm, because it works with an unlimited colour range.

For all experiments, the first image camera position is used as the ideal image camera position (left image in fig. 3.6). Seven visibility configurations are considered,

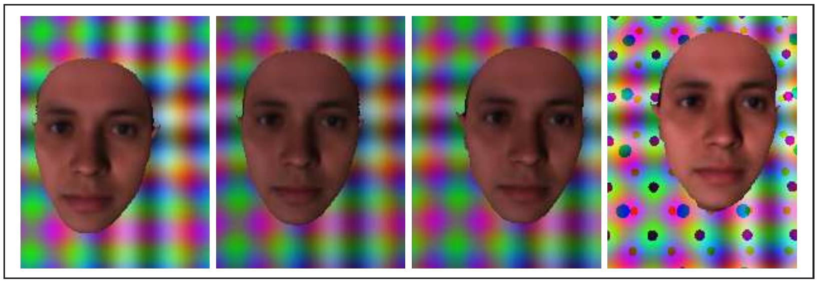


Figure 3.6: **Synthetic ground truth sequence:** *One of the* 10 *evaluation sets is shown.*

i.e., pixels are assumed to be visible in at least two images and are always visible in $\mathbf{y}^1$ (the image which coincides with the ideal image $\mathbf{y}^*$). The amount of depth states varies dependent on the camera configuration and lies between $K = 43 \dots 53$. The image size is 150×200 pixel^2.

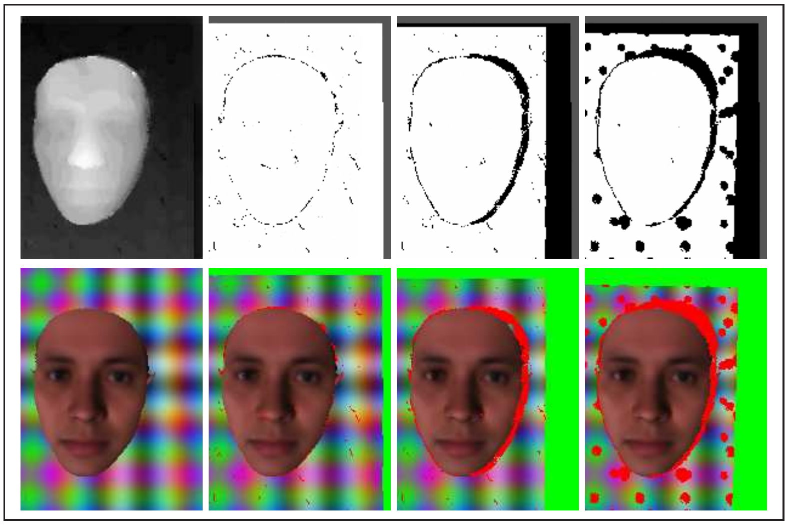
Influence of colour transformation

Figures 3.7 and 3.8 show the results for the scene in fig. 3.6, with and without the estimation of the colour transformation $\boldsymbol{C}^k$ for each image. The images in fig. 3.7 show the results with colour transformation update. Depth and visibilities are nicely estimated. Note the spots in the background of the right-most input image in fig. 3.6. These spots are indeed assigned to be outliers, as can be seen in the top-right visibility map of fig 3.7. The bottom row of those images shows the ideal image (left image). The three right images are the input images (three right images in fig (3.6). These images have been warped to the reference image by the geometric (depth dependent) transformation. Furthermore their colour values have been photometrically transformed by the estimated colour transformation. The photometric warp displays visually a good estimation of the colour transformation.

The result for the model without colour transformations is presented by the images in the fig. 3.8. In this case, the fourth input image has essentially no influence on the depth estimation. Almost all pixels have been turned into outliers and the depth is found by input images, which are more similar in terms of their colour transformation. Given a generative model without colour transformation, these results represent the most likely solution.

3.10.1 Cooling schedule

The last experiment shows the importance of estimating colour transformations which can be present (for instance by cameras with automatic aperture and/ or shutter time adjustment) in the images. Local minima of the likelihood function are especially problematic in these situations. Therefore, a deterministic annealing schedule is used. The dependence of the solution on the starting temperature and the form of their de-



algorithm	T_s	T_e	T_d	n_n	σ_d	σ_v	C	f_s
Bethe	100	0.5	3.5	20	100	1	10^{-10}	10^{-10}

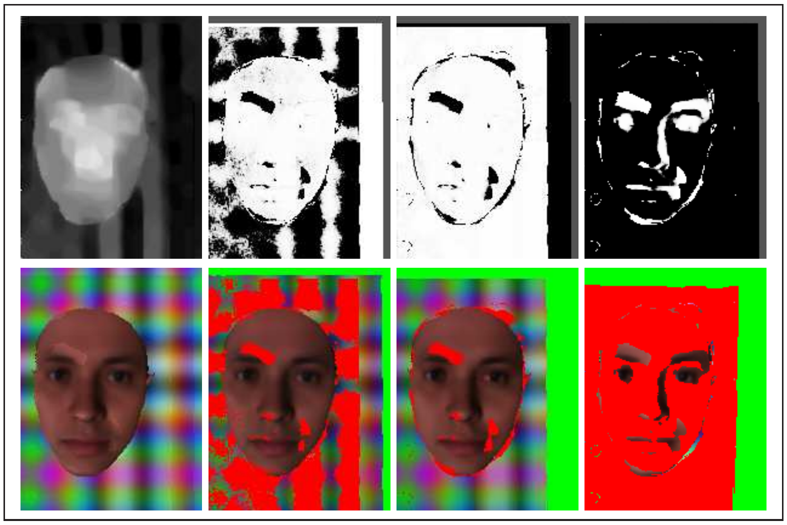
Figure 3.7: **Importance of the colour transformation:** *Results of the example in fig. 3.6 when a colour transformation is estimated. Both results are shown in eight images. These are: the expected value of depth (top-left) and visibility (top-right images), the ideal image (bottom-left) and the geometrically and photometrically transformed input images (red pixels are outliers; green pixels are pixels for which the geometric transformation falls outside the image).*

crease is evaluated in this section. Fig. 3.9 and fig. 3.10 show the typical evolution for the expected values of the MRF (depth and visibility as in eq. (3.21)) and the value of the parameters ($\mathbf{\Sigma}$, $\mathbf{C}(\boldsymbol{p}^k)$) during temperature annealing.

Starting from uniform beliefs b_i the first EM iterations at high temperature (left column in fig. 3.9) lead to a fuzzy depth map. Almost all visibility expectations are undecided about their value (indicated by gray colour values). Only some real occlusions already have a lower expected value. Because of the fuzzy depth estimate and the wrong estimate of the colour transformation, the noise is high. In this temperature regime, the ideal image contains artefacts from the spots of the fourth input image.

During the next iterations, all parameters (the noise $\mathbf{\Sigma}$ and the colour transformations for each target image $\mathbf{p}^{1,2,3}$ as shown in fig. 3.10) and the MRF expectations evolve slowly to the global solution. This example shows visually that the EM algorithm, which is traditionally strongly dependent on the initialisation, can be made less dependent by the deterministic annealing technique used here.

We continue with the quantitative evaluation on the whole test set.



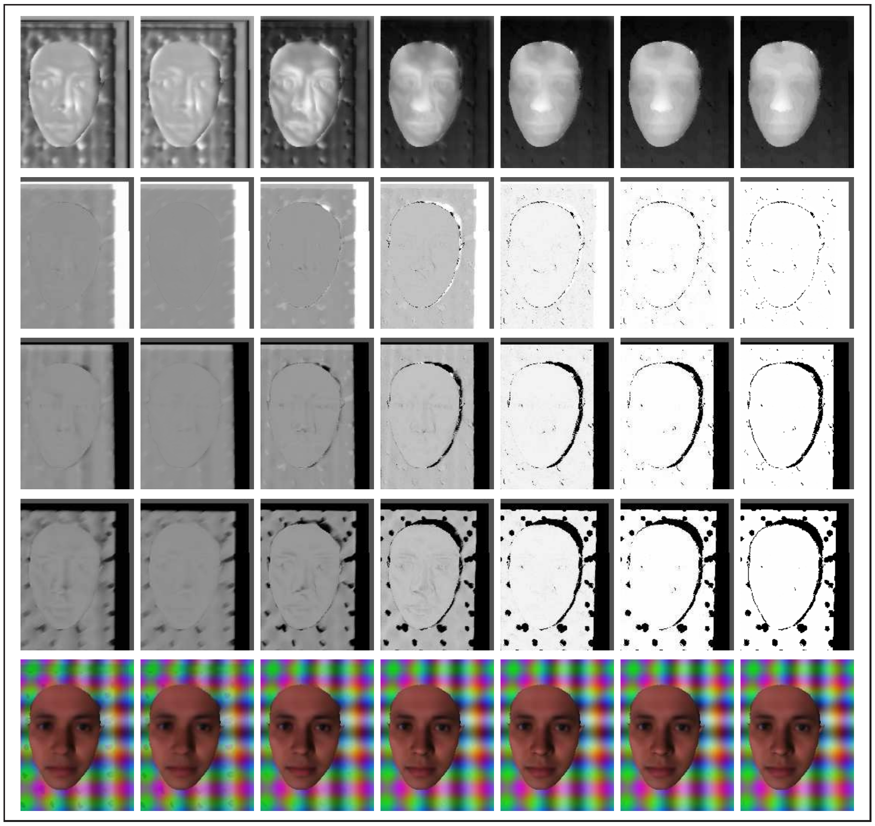
algorithm	T_s	T_e	T_d	n_n	σ_d	σ_v	C	f_s
Bethe	100	0.5	3.5	20	100	1	10^{-10}	10^{-10}

Figure 3.8: **Importance of the colour transformation:** *Results of the example in fig. 3.6 when a colour transformation is* not *estimated. Both results are shown in eight images. These are: the expected value of depth (top-left) and visibility (top-right images), the ideal image (bottom-left) and the geometrically and photometrically transformed input images (red pixels are outliers; green pixels are pixels for which the geometric transformation falls outside the image).*

Start temperature and decay rate

Fig. 3.11 shows the quantitative results for the whole test set for different start temperatures T_s and three different decay rates T_d, both for the mean field and the Bethe approximation. The quality is, similar to Scharstein *et al*. [98], measured by the percentage of correspondences from the reference image to all target images for which the displacement (disparity) error falls below one pixel. This value is evaluated for all correspondences which can be established (bearing in mind occlusions).

As a global trend, one can recognise a plateau at high starting temperature $T_s \sim \{10 \ldots \infty\}$ as long as the decay velocity is low $T_d \geq 2$. The quality of the results gets worse by lowering T_s. The behaviour is similar for both, the Bethe and mean field approximation. This result suggests a sufficient high start temperature ($T_s \geq 10$) together with a slow cooling ($T_d \geq 2$). Furthermore, one can appreciate the importance of the deterministic temperarture annealing EM (DAEM) scheme compared to the classical EM. The starting temperature of $T_s = 1$ (which would be the realisation of EM) does not give the optimal results. The initialisation, especially of the colour transformation, is in this case not good enough to find the global optimum of the



algorithm	T_s	T_e	T_d	n_n	σ_d	σ_v	C	f_s
Bethe	100	0.5	2	20	100	1	10^{-10}	10^{-10}

Figure 3.9: **Evolution of the solution during EM cooling:** *The expected values of depth (top row), the visibility with respect to the three non-reference images* $y^{2,3,4}$ *(three middle rows) and ideal image (bottom row) for EM iteration* $\{1, 3, 6, 9, 12, 15, 18\}$ *and corresponding temperatures* $T = \{100, 32.97, 9.96, 4.09, 2.01, 1.12, 0.67\}$ *for the scene in fig. 3.6. The result after the last iteration (*20*) with* $T = 0.5$ *is shown in fig. 3.7.*

posterior distribution.

Influence of the end temperature

For this experiment, the start temperature $T_s = 100$ and the decay rate $T_d = 3.5$ has been fixed and the influence on the end temperature T_e is evaluated. The results are shown in fig. 3.12. From the theoretical point of view, one would expect an increase in performance by lowering the temperature until the critical temperature, *i.e.*

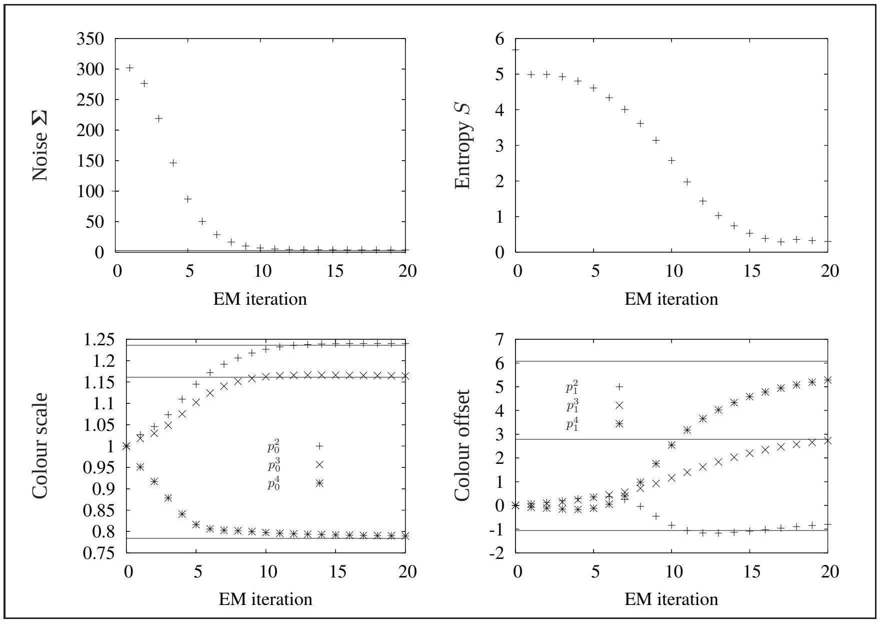


Figure 3.10: **Evolution of the parameters during EM cooling:** *top/left: image noise* $\mathbf{\Sigma}$*, top/right: entropy* S*, bottom/left: colour scale, bottom/right: colour offset. Colour scale* p_0^k *and offset* p_1^k *are shown for the transformation from the reference image to all three target images. Horizontal lines indicate thereby the ground truth. The images in fig. 3.9 have been evaluated.*

the temperature where the system shows a phase transition in the state of order (see appendix C.1 for more explanation and a temperature simulation of our prior model). When this critical temperature is reached, the solution is frozen and the results are expected to build an error plateau.

The plots in fig. 3.12 show the expected behaviour in the high temperature phase. However, a small increase of the error in the low temperature phase for both the mean field and the Bethe approximation can be observed. The reason for this is twofold. In the low temperature phase the likelihood and the prior is sharply peaked about a single state. This means that the beliefs b_i will also be peaked about a single state x^{rs} and that the expected values of the depth and visibility are very close to the depth and visibility values, when estimated from the state of maximal probability. This behaviour shows to some extent the limitations of MRF stereo approaches. They assume that the state of a pixel can be described by a *discretised* depth value d_k. Obviously, depth is a *continuous* value for which the generative model as defined in chapter 4 is more suited. MRF approaches therefore lead to discretisation errors. These can be minimised when the temperature decrease is stopped just below the critical temperature and when the depth is extracted by the expected value, rather than by the maximum value of the

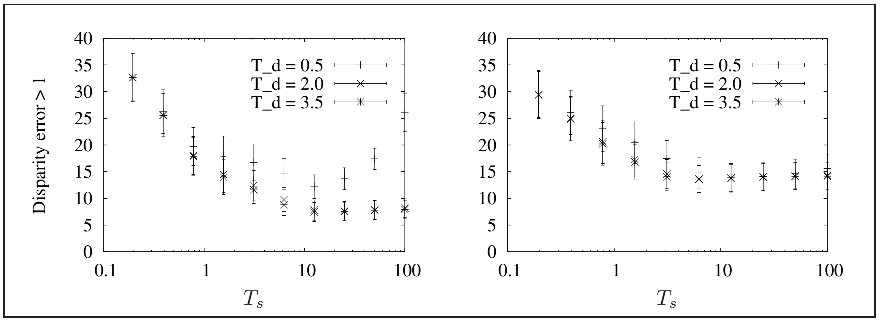


Figure 3.11: **Evaluation of start temperature and temperature decrease:** *Percentage of pixels with a disparity error larger than one for different start temperatures* T_s *and three temperature decay coefficients* $T_d = \{0.5, 2.0, 3.5\}$*, for the Bethe approximation (left) and the mean field approximation (right).*

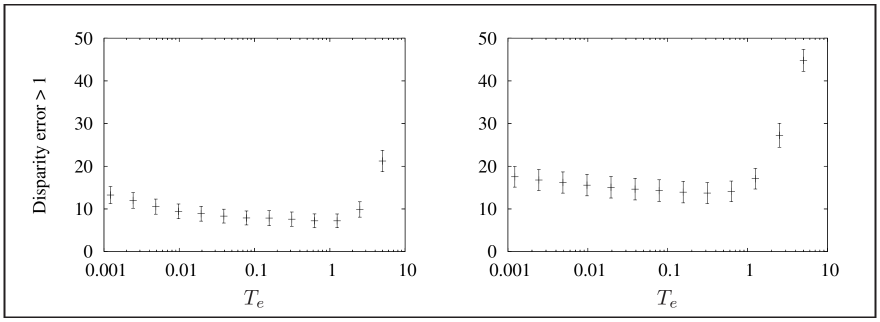


Figure 3.12: **Evaluation of the end temperature:** *Percentage of pixels with a disparity error larger than one for different end temperatures* T_e*, for the Bethe approximation (left) and the mean field approximation (right) (*$T_s = 100$*,* $T_d = 3.5$*).*

MRF states. Note that this is the reason why graph cuts [11] are less suited, since they only compute the maximum value of the MRF states.

Influence of the sparsification threshold

After each EM-iteration the expected value $b_i(x_i)$ of the MRF x_i is computed. All states m of x_i for which b_i^m is smaller than $f_s max(b_i^m)$ are neglected from all further computations. Fig. 3.13 shows the quality of the solution and the corresponding execution time as a function of the threshold f_s. Up to $f_s = 10^{-6}$ the results remain similar, but the execution time is almost halved at this point. The right plot also shows that the Bethe approximation is about three times slower than the mean field update.

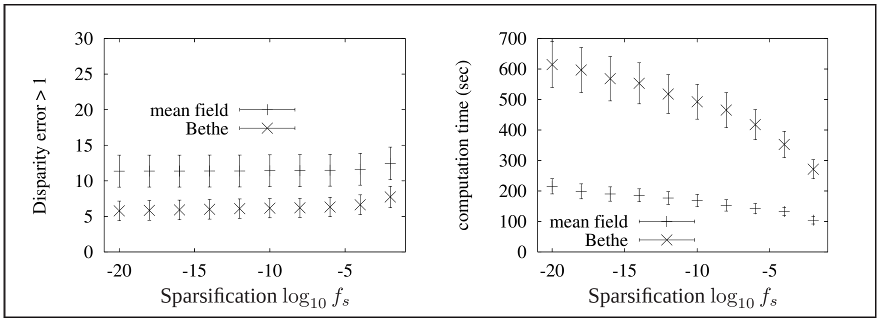


Figure 3.13: **Evaluation of the sparsification:** *Percentage of pixels with a disparity error larger than one for different values of the sparsification threshold f_s is shown left. In the right plot, the corresponding execution times (in sec) for the Bethe approximation and the mean field approximation ($T_s = 100$, $T_d = 3.5$, $T_e = 0.1$) are given.*

3.10.2 Mean Field versus Bethe approximation

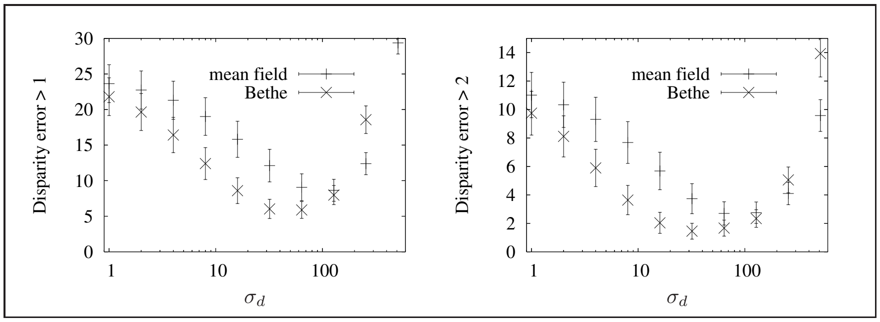


Figure 3.14: **Mean field/ Bethe comparison:** *Percentage of pixels with a disparity error larger than one (left) and larger than two pixels (right) for mean field and Bethe approximation as a function of σ_d ($T_s = 100$, $T_d = 3.5$, $T_e = 0.1$, $C = 10^{-10}$, $\sigma_v = 1$).*

Figures 3.14, 3.15, 3.16 and 3.17 evaluate the quality of the depth and visibility estimation for the mean field and Bethe approximation as a function of different MRF Gibbs prior models. These prior models are specified by the value of σ_d, σ_v and C in eq. (3.6). Since the exact form of the prior model is often not known, this section can also be seen as an evaluation of the parametric prior model on ground truth data.

In addition to the last figures, the quality of the depth estimate will be measured by the median of the disparity errors. Similar to the previous measure (percentage of pixels with a disparity error smaller than one), all correspondences are evaluated except those that are occluded. The median error measure allows to evaluate the results more precisely. Gross outliers, that can appear at the image borders, have no influence on the error criterion as it would be for the average error. Furthermore, the error in the visibility estimation is reported. This measure is computed as the percentage of pixels with wrong visibilities in all images.

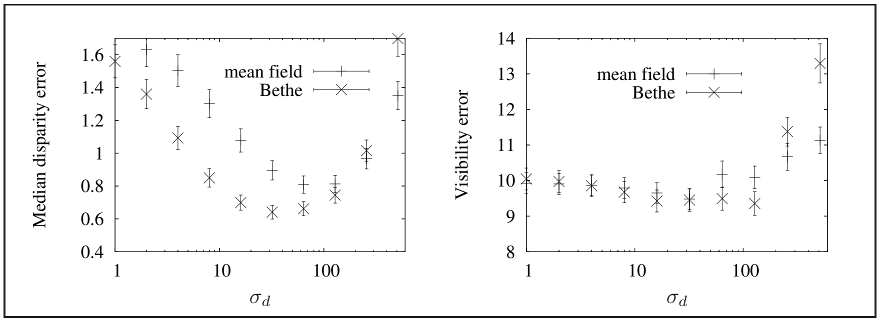


Figure 3.15: **Mean field/ Bethe comparison:** *Median disparity error (left) and the relative fraction of wrong assigned visibilities (right) for mean field and Bethe approximation as a function of* σ_d *(*$T_s = 100$*,* $T_d = 3.5$*,* $T_e = 0.1$*,* $C = 10^{-10}$*,* $\sigma_v = 1$*).*

Figure 3.14 evaluates the relative amount of pixels with a disparity error of one (two) pixels as a function of σ_d. The Bethe approximation shows a clear advantage over the mean field approximation, with a minimum error of approximate 5% (2%). Two other insightful results can be read off the figures. Firstly, for weak correlations (small σ_d), the difference between Bethe and mean field approximation is less pronounced. In fact, it is easy to see that mean field and Bethe approximation are equivalent for $\sigma_d = 0$, *i.e.*, uncorrelated MRF states with $\psi_{ij}(x_i^{rs}, x_j^{pq}) = C$. Secondly, with increasing correlation strength, the difference of both approximations increases. At a certain point, the mean field approximation eventually scores better than the Bethe approximation. At this point, the prior model is obviously wrong, *i.e.*, the correlation is so strong that the prior demands a too smooth solution. The mean field approximation is not able to handle these strong correlations and cannot follow this prior model. This leads ‘by accident’ to better results. The relatively bad performance of the mean field approximation with increasing correlation strength is also indicated by the shift of the error minimum to higher correlations. As a result, one can state that for weakly coupled inference problems, the mean field approximation might not be a bad choice, especially when taking the computational speedup into account.

This interpretation is also justified by the median error evaluation in fig. 3.15. The difference in the visibility error (fig. 3.15 right) is less distinctive but shows nevertheless a small advantage of the Bethe approximation and a minimum, which coincides with the depth error minimum.

Fig. 3.16 presents the evaluation of the error w.r.t. σ_v. The value of σ_v determines the correlation strength of different visibility configurations. Setting $\sigma_v = 0$ will realise a spatial correlation for which the joint probability of two pixels is independent

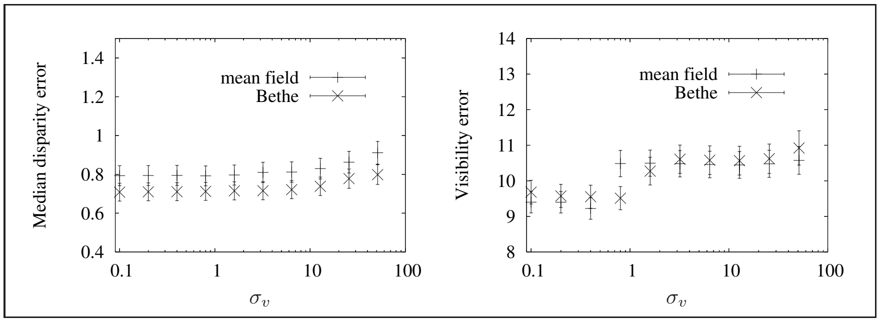


Figure 3.16: **Mean field/ Bethe comparison:** *The median displacement error (left) and visibility error (right) for mean field and Bethe approximation as a function of* σ_v. *(*$T_s = 100$, $T_d = 3.5$, $T_e = 0.1$, $C = 10^{-10}$, $\sigma_d = 100$*).*

on the visibility configuration. Large σ_v give more support to neighbouring pixels with the same visibility configuration.

It can be seen in the left figure that the correlation of visibility configurations does not increase the performance of the depth estimation when compared to uncorrelated visibility configurations $\sigma_v = 0$. The visibility error on the other hand shows a small advantage near $\sigma_v = 1$. The explanation for this behaviour lies in the strong influence of the likelihood. When looking at the images in fig. 3.6 one can notice the nicely textured background which makes it relatively easy to disjoin a good match from an outlier purely based on the data likelihood. Thus, the correlation of visibilities will not have a very strong influence on the depth estimation. In the cones sequence (sec. 3.10.4), we will see a more outspoken relation between correlated visibility configurations and the depth estimation performance.

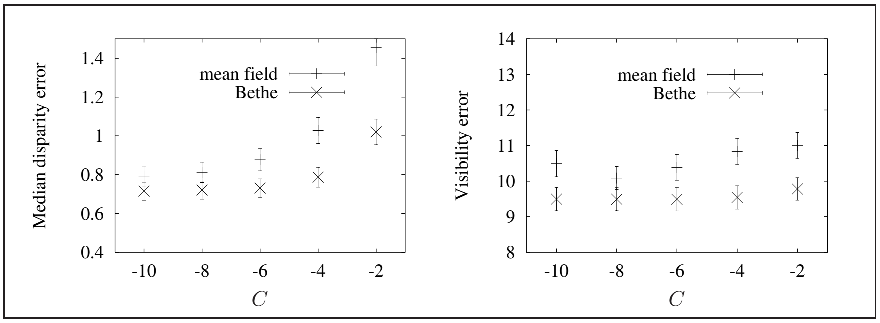


Figure 3.17: **Mean field/ Bethe comparison:** *The median displacement error (left) and visibility error (right) for mean field and Bethe approximation as a function of* C *(*$T_s = 100$, $T_d = 3.5$, $T_e = 0.1$, $\sigma_d = 100$, $\sigma_v = 1$*).*

Finally, fig. 3.17 shows the evaluation w.r.t. the value of C, which reflects the likelihood of an outlier from the correlated depths-visibility assumption. A large value of C will therefore globally downweight the correlation of neighbouring nodes. While this is a good mechanism for discontinuities, it is not for the majority of links. When defining a global interaction matrix for all links, one would expect a value of C between the optimal value for continuous and discontinuous links. The experiments show that this value is approximatively $C \approx 10^{-8}$.

As a global result of these synthetic experiments, one can state a clear advantage of the Bethe approximation over the mean field approximation in terms of the accuracy in the depth estimation. The difference w.r.t. to the visibility error is less obvious.

3.10.3 Prior on parameters

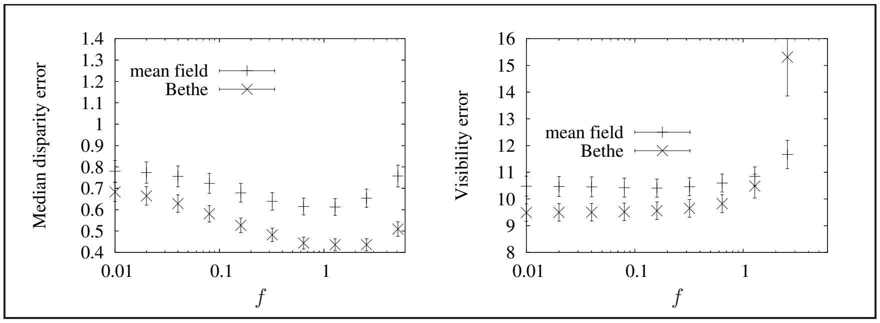


Figure 3.18: **Influence on parameter prior:** *The median displacement error (left) and visibility error (right) for mean field and Bethe approximation as a function of the magnitude of the parameter prior* f *(*$T_s = 100$, $T_d = 3.5$, $T_e = 0.1$, $\sigma_d = 100$, $\sigma_v = 1$, $C = 10^{-10}$*).*

Fig. 3.18 shows the results for a different value of the relative amount of fake data, which is introduced to set a prior on the parameters $\mathbf{y}^*$ and $\mathbf{\Sigma}$ (see section 3.6.2). The accuracy of the depth estimation increases by using this prior with the best value of about $f_p \approx 1.0$. If this prior is too strong, the performance of the depth and the visibility estimation decreases. The reason for this behaviour is the underestimation of $\mathbf{\Sigma}$, which at the same time produces a larger amount of outliers.

3.10.4 Real image evaluation

This evaluation is an example of anisotropic MRF modelling. Anisotropic interactions are realised by defining the interaction matrix locally. More particular, we define two interaction matrices by eq. (3.6): one which models discontinuities and one for the continuous areas. The difference of both matrices lies in the outlier probability C. $C = C_d$ is used for all links for which the endpoints fall into different mean shift colour segments [16], and $C = C_s$ for the remaining cliques, with $C_d \geq C_s$.

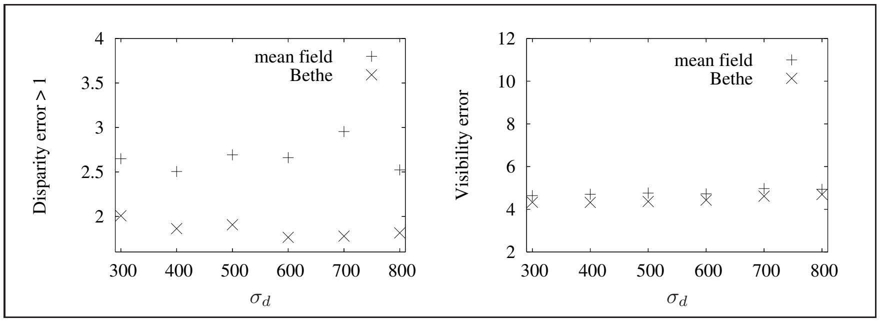


Figure 3.19: **Cones sequence:** *Depth (left) and visibility error (right) as a function of* σ_d *(*$T_s\!=\!20$, $T_d\!=\!2$, $T_e\!=\!0.1$, $\sigma_v\!=\!30$, $C_s\!=\!10^{-10}$, $C_d\!=\!10^{-5}$, $f_s\!=\!10^{-10}$, $f_p\!=\!1$*).*

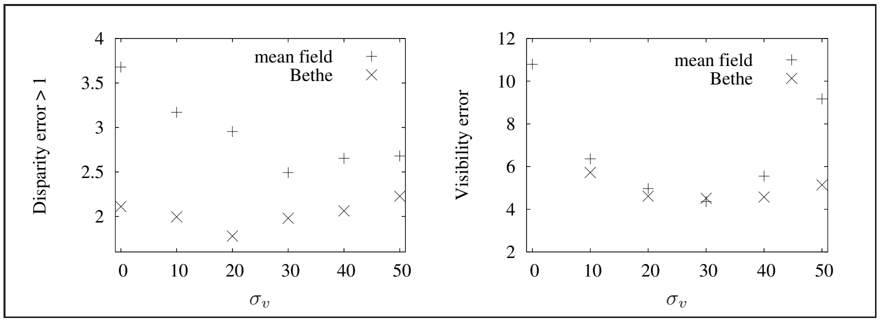
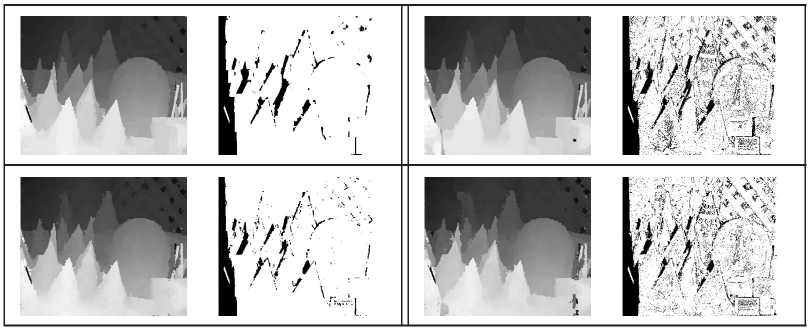


Figure 3.20: **Cones sequence:** *Depth (left) and visibility error (right) as a function of* σ_v *(*$T_s\!=\!20$, $T_d\!=\!2$, $T_e\!=\!0.1$, $\sigma_d\!=\!700$, $C_s\!=\!10^{-10}$, $C_s\!=\!10^{-5}$, $f_s\!=\!10^{-10}$, $f_p\!=\!1$*).*

For the evaluation the 'cones' sequence from the Middlebury stereo evaluation set [98] is used. We use three images with visibility configurations $s = 1, 2, 3$.

Figs. 3.19 and 3.20 show the depth and visibility errors. The depth error is the percentage of pixels with a disparity error larger than 1, evaluated for all visible pixels (equivalent to [98]). Similarly, the visibility error is the percentage of wrongly detected occlusions. Again, the advantage of the Bethe approximation can be noticed. Fig. 3.20 shows a clear correlation between the depth error and the strength of the visibility correlation σ_v. We notice that the correlation of visibilities is not only helpful for a better detection of these (see visibility error in fig. 3.20), but is helps also to increase the performance of the depth estimation. The case of uncorrelated visibility configurations ($\sigma_v = 0$) is inferior to the best value of $\sigma_v \approx 20$, both for estimation depth and visibility. This result tallies with that in [32]. Fig. 3.21 shows this visually by comparing the depth and visibility maps for both approximations and for two different parameter settings. One (the two left images) for correlated visibilities $\{\sigma_d = 700, \sigma_v = 30\}$ and one (the two right images) for uncorrelated visibilities

$\{\sigma_d = 700, \sigma_v = 0\}$. Notice the contribution of correlated visibilities in the depth and the visibility maps for the Bethe (top row) and the mean field approximation (bottom row).



	$\sigma_d = 700, \sigma_v = 30$		$\sigma_d = 700, \sigma_v = 0$	
	depth error	visibility error	depth error	visibility error
Bethe	1.78	4.61	2.11	12.21
Mean field	2.95	4.93	3.67	10.79

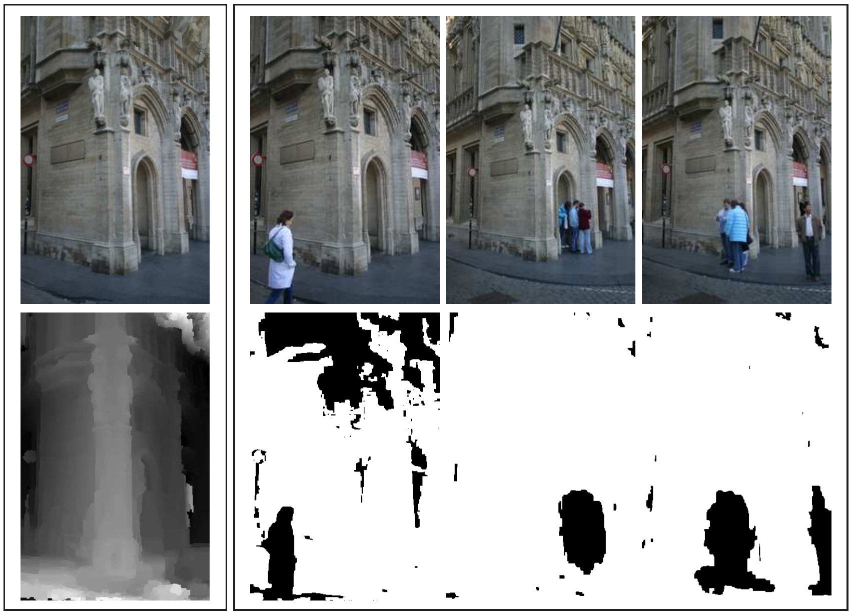
Figure 3.21: **The contribution of correlated visibility configurations:** *The left images show the results of correlated visibilities (*$\sigma_d = 700, \sigma_v = 30$*) and the right images the uncorrelated case (*$\sigma_d = 700, \sigma_v = 0$*). The Bethe approximation is shown in the top row and the mean field approximation in the bottom row. Underneath the images, the table gives the numerical values for the four experiments.*

3.10.5 Outdoor scene reconstructions

The algorithm has been tested on several challenging outdoor scenes, characterised by multiple depth occlusions, independently moving objects and complicated scene geometry. The original images are of size 3072×2048 and have been downscaled to a size of 768×512. The parameters for all experiments are the same and shown together with the computation time below the figures.

The first example shows a scene which is contaminated by pedestrians. The three input images are shown in the top row of fig. 3.22. The camera position of the ideal image was chosen to be the left of these images. Notice that all images are contaminated with independently moving objects. Also, the reference image contains pixels (*e.g.*, woman in white) which have no support in any other image. Still, the results in fig. 3.22 shows that our algorithm could assign a meaningful colour (top/left) and depth (bottom/right) to those outlier pixels. The three images on the right in the bottom row of fig. 3.22 show the visibility estimates. The Bethe approximation of the free energy was used, and four visibility configurations v^s, $s = 1, 2, 3, 5$ were considered. The number of depth states for this scene is $R = 180$.

The depth estimation at the bottom of the ideal image $\mathbf{y}^*$ is rather poor. The lack of



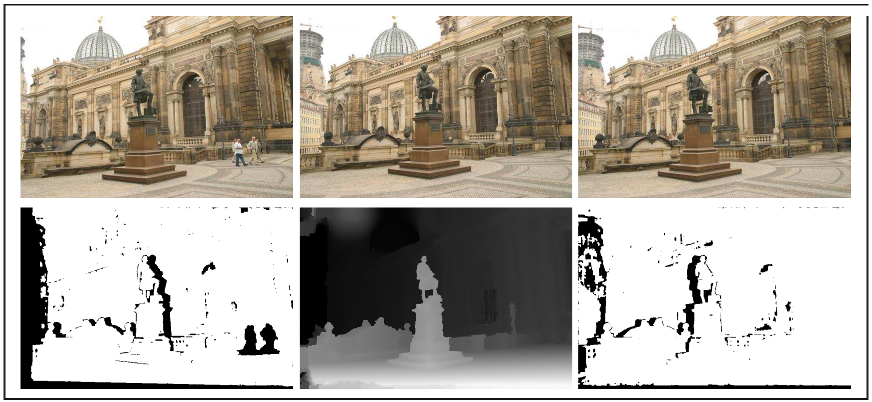
T_s	T_e	T_d	n_n	σ_d	σ_v	C_s	C_d	f_s	f_p	time
20	0.1	2	30	200	20	10^{-10}	10^{-5}	10^{-8}	1.0	396 sec

Figure 3.22: **Brussels city hall scene:** *The three input images are the three right-most images shown in the top row. The camera position of the virtual image* $\mathbf{y}^*$ *was chosen to be the left of these images. The visibility estimates related to* $\mathbf{y}^*$ *are in the bottom row. The top-left image shows the estimated ideal image* $\mathbf{y}^*$ *and the estimated depth is shown in the bottom-left image.*

texture and the fact that the epipole lies within all target images is the reason for this. However, the ideal image looks visually convincing and the estimated depth, visibility and ideal image constitute a solution which makes the input images very likely.

For the second experiment we used three images containing the Semper statue in the heart of Dresden. These images are shown in the top row of fig. 3.24. The camera position of the ideal image was chosen to be the middle image. Because the reference camera does not contain independently moving objects, we only consider the four visibility configurations $v^s, s = 1, 2, 3, 4$ in table 3.2. On the bottom, the extracted depth and visibilities are shown. We used the Bethe approximation with $R = 264$ depth states. One can appreciate the accurate detection of all three types of outliers. Geometric occlusion, pedestrians and the specularities in the windows are detected.

In the last experiment we used three images of the ‘Leuven city hall scene’ [103]. These images are shown in the top row of fig. 3.24. The camera position of the ideal image was chosen to be the top middle image. Because this scene does not contain independently moving objects, we only consider the four visibility configurations



T_s	T_e	T_d	n_n	σ_d	σ_v	C_s	C_d	f_s	f_p	time
20	0.1	2	30	200	20	10^{-10}	10^{-5}	10^{-8}	1.0	440 sec

Figure 3.23: **Semper statue scene:** *The input images are shown in the top row. The middle image is chosen as the reference view. The depth map for the reference view (middle) and outlier maps for the two other images are shown in the bottom row. Notice that not only geometrical occlusions but also the pedestrians (top left image) are detected.*

$v^s, s = 1, 2, 3, 4$ in table 3.2. In the bottom row, the extracted depth and visibilities are shown. We used the Bethe approximation with $R = 396$ depth states. This experiments also shows excellent depth and visibility estimates. The datasets (images, calibration and 3-D points) are available at www.esat.kuleuven.be/∼cstrecha/testimages.

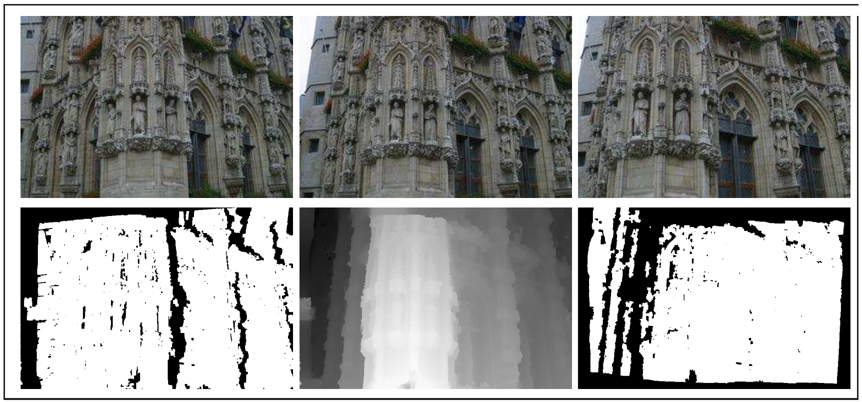
Note that the same prior model $\{\sigma_d, \sigma_v, C_d, C_s\}$ has been used for these three scenes.

3.11 Conclusion

3.11.1 Summary

In this chapter, an approach to multi-view stereo has been presented, which can also deal with scenes contaminated by accidental objects as in Figs. 3.22 and 3.23. A novel view is computed, which is most likely given the input images. To compute this novel image, we take possible configurations of depth *and* visibilities into account. This approach results in the natural elimination of accidental objects which cannot be explained by the majority of input images.

In the E-step of the EM algorithm, two approximations of the free energy have been compaired: the mean field and Bethe approximation. Minimising the latter energy can be achieved by belief propagation. The quality of both approximations have been evaluated on the basis of ground truth data. This shows that for the stereo problem, the Bethe approximation has clear advantages over the mean field approximation.



T_s	T_e	T_d	n_n	σ_d	σ_v	C_s	C_d	f_s	f_p	time
20	0.1	2	30	200	20	10^{-10}	10^{-5}	10^{-8}	1.0	469 sec

Figure 3.24: **Leuven city hall scene:** *The three input images are shown in the top row. The camera position of ideal image* $\mathbf{y}^*$ *was chosen to be the middle image. The depth map for the reference view (middle) and outlier maps for the two other images are shown in the bottom row.*

For small MRF correlations, the difference of both approximations is less outspoken. In these cases, the mean field approximation might still be a good choice, especially because of the speed and memory advantages.

The results also show that the method scores well on the Middlebury stereo evaluation (see Strecha, Fransens and Van Gool [103]). Currently, the algorithm is at the fourth position when performance is measured at the highest precision (0.5 pixels disparity error) for all visible pixels.

The presented approach to detect outliers is purely based on photometric cues. Therefore, it can cope with independently moving objects, as well as geometric occlusions. For example, photometric cues are necessary to deal with scenes like the one shown in fig. 3.22. However, when the scene contains large untextured regions, photometric cues could fail to detect an occlusion. It remains possible that the occlusion can be explained by assigning a wrong depth, if this provides a consistent match in all images. Combining photometric and geometric cues is expected to further increase the robustness of outlier detection. However, whereas photometric occlusion cues can easily be used to formulate and minimise depth *and* occlusion jointly, this is more difficult when adding geometric cues. To be more formal, it is easy to compute the likelihood of a pixel i having depth d^r and being visible in the k^{th} image-based on photometric cues. To compute the same likelihood by taking geometric cues into account one would have to consider all terms which are intersected by projecting the 3-D point (which corresponds to depth d^r in pixel i) to the k^{th} image. This would lead to a formulation where the state of a pixel x_i is correlated to many more pixels than only on its four neighours. A tracktable integration of photometric and geometric cues is currently only possible when depth *and* occlusion are treated separately as for instance in [51, 58].

An extensive validation of the temperature annealing scheme (DAEM) is presented. This showed that the temperature annealing version of the EM algorithm has advantages over the classical EM formulation. Especially in stereo settings with a (strong) colour change the DAEM approach is able to find the global optimum even without a good initialisation. The dependence of the solution on the specific form of the parametric prior model is reasonable, such that, for instance, all real experiments could be performed with the same parameters.

The computation time depends largely on the scene itself and the number of evaluated states. If the scene contains many ambiguities, *e.g*. large un-textured regions with a uniform data-likelihood, the algorithm will be slower. In this case the sparsification described in section 3.9.2 will be less efficient. In the other case, *i.e*. if a pixel has a clearly peaked data-likelihood distribution, many states will be victims of the sparsification already at an early stage of the optimisation. Ambiguities can be diminished by taking more images into account, which on the other hand leads to more visibility states. An optimum in terms of the computation time will depend on both factors.

MRF formulations, as presented in this chapter, work well even without an initial estimate on the depth. However, applying MRF methods to large images is more difficult because of their large memory and speed requirements. These have been diminished to some extent by a sparse implementation. In the next chapter, a local PDE based approach is presented, which overcomes these limitations, but which will need a good initialisation. These might be provided by the MRF approach presented here.

3.11.2 Relation to other formulations

Many MRF formulations to stereo assume the inlier distribution to be known. Usually this distribution has a mean value of y_i^1, *i.e*. the colour value of a pixel in the reference image, and a specific known variance (*e.g*. [58, 63]). For stereo formulations that use a robust matching criterion the parameters of the M-estimator (ρ-function) are also supposed to be known (*e.g*. [64, 50, 51, 131, 64]). We have shown in chapter 2 that robust M-estimation can be related to a generative model based formulation. More particular, one could justify the parameter choice of the M-estimator by setting large priors on the corresponding inlier and outlier distributions. The exact relation to our generative models is, however, difficult to make. For many energy based stereo formulations it is not clear how their underlying generative model could be defined.

It is further interesting to notice, that most multi-view stereo formulations use $\mathbf{y}^1$ as the mean of the inlier distribution. This can be seen as putting a large prior ($f_p \rightarrow \infty$ in eq. 3.7) on the ideal image $\mathbf{y}^*$, however, we have seen in sec. 3.10.3 (fig. 3.18) that this does not lead to the optimal result.

Chapter 4

Local formulation

> *IIf oo donnt teeief that oour model (e.g. of norral erroiss is c arect, choose an ther one and ose raxiror likelih on - or Bayesian - methods for the new model. What, if I oonnt belief in the new model either? It takes a eot of st ttornness to flood the world with a host of rather iatithary and pr tatly hardey interpretatee models ano seier they are exactly true. The p int of rotost statestiss is that one may keep a pararetric model hethough the eatter is known to ee wrong.*
>
> $\arg\max_{\mathbf{y}^*}\left\{ \log p(\mathbf{y}\,|\,\mathbf{y}^*)p(\mathbf{y}^*)\right\}$ of Hampel *et al*. [42] with
>
> $p(\mathbf{y}^*) \propto \prod_{ij\in[i\pm1,2]} \psi_{ij}(y_i^*, y_j^*)$

In the previous chapter, a MRF formulation for the multi-view stereo problem has been presented. The states of the random field did include depth and visibility. This approach can be seen as a *global* approach in the sense that the probabilities of all possible depth and visibility realisation are considered. It can therefore be used without initial (depth) knowledge of the scene. This formulation has, however two disadvantages:

- The model assumes that the scene can be described by a number of discretised depth values. Obviously, depth is a continuous property of the scene and MRF formulations do not account for that.

- To achieve sufficient accuracy, the number of states grows very large and it becomes difficult to remain fast and memory efficient.

We showed in the previous chapter that a sparse implementation is possible, which solved the second problem to some extent. The first problem is more serious, and we will therefore discuss in this chapter a generative model for which depth is continuous. This leads to a local approach, in which a depth map iteratively evolves through PDE-based non-linear diffusion.

4.1 Introduction

4.1.1 Previous work

PDE-based approaches for the stereo problem can be divided into two general formulations.

- **Global PDE formulations** are similar to MRF formulations in that they also define the energy globally in a 3-D space. Often implicit functions are defined in this space and regularisation is based on neighbouring grid points (voxels).
- **Local PDE formulations** are often image-based. Usually an energy is defined in the $2D$ image domain and the regularisation is based on neighbouring pixels.

The fundamental difference between global PDE-based solutions and MRF formulations is the normalisation. Every node in a MRF formulation is considered to be in *exactly one* state, *e.g.*, a certain depth state d_k or the state "occluded". For this reason, the expected values of the MRF states need to be normalised over each node. In PDE-based global formulations, this is not the case. Usually, an energy is minimised such that the images are brought into correspondence and a smoothness condition is fulfilled. The far most prominent members of 3-D-based PDE solutions use level-sets, which have been introduced by Setian and Osher[78]. First level-set formulations for the multi-view stereo problem have been presented by Deriche *et al*. and Faugeras *et al*. [20, 21, 24, 25]. Further research considered for instance: efficient implementations, using narrow band level-sets or GPU-based implementations [67]; the extension to the case of non-Lambertian surfaces, formulated by Jin *et al*. [54, 53]; the incorporation of additional constrains, which can be based on visual hulls [44] and/ or calibration points [68] and the use of cross correlation or mutual information as the similarity measure as formulated by Pons *et al*. [86]. The advantage of these methods lies in the integration of all images into one single optimisation scheme. The discretisation of the 3-D space into voxels can be seen as a disadvantage.

Local, PDE-based formulations have their origin in the wide field of optical flow computation, where the correspondence between pairs of images is parameterised by a 2-D flow vector for each pixel. When the scene is rigid and epipolar geometry is known, the two degrees of freedom for each pixel reduce to one degree and the disparity can be estimated instead. This approach is for instance studied by Devenay *et al*. [22], Proesmans *et al*. [90], Robert *et al*. [93], Alvarez *et al*. [3] and Slesareva *et al*. [100]. When the full calibration is provided and more than two images are given, depth is the natural parameter that brings all images into correspondence with the target image. This multi-view stereo extension of the stereo problem has been proposed by Robert and Deriche [93] and applied to real images by Strecha *et al*. [106]. A probabilistic formulation of the latter work is given in [103]. And a further extension to the estimation of multiple depth maps has been proposed by Gargallo *et al*. [37]. All these 2-D methods use a reference image or a virtual image [103] as the space on which depth is computed.

To the class of local PDE formulations, one can also count methods which use a triangle mesh that brings all images into correspondence. Stereo and visual hull constraints are often combined for steering the mesh to an energy minimum. Examples are, for instance, Furukawa *et al*. [36] or Neumann *et al*. [76]. We call all these methods local since the depth/disparity is iteratively updated in a gradient decent manner starting from some initialisation.

Most PDE based formulations start by defining an energy which is minimised by various optimisation schemes. Several parameters are introduced, which account for instance for noise variations, the smoothness of the solution, for breaking the smoothness condition in some areas and for the visibility reasoning. It is the aim of this work to formulate an algorithm for which many of those parameters disappear. Generative models for multi-view stereo as proposed by Strecha, Fransens and Van Gool [103] and extended by Gargallo and Sturm [37] are the key to achieve this. Our particular generative model will lead to an EM algorithm in which a local PDE-based solution for the depth plays a major role. Discontinuities of the depth are modelled by anisotropic diffusion, for which we next succinctly review the related work.

4.1.2 Discontinuity preserving diffusion

For many problems in computer vision, regularisation is required to overcome their ill-posedness. Often a smoothness constraint is added, for instance, for the computation of optical flow. A large amount of work has been done to formulate smoothness constraints, which can be locally broken. These constraints lead to the wide field of inhomogeneous and anisotropic diffusion filtering. From a probabilistic point of view, the smoothness constraint can be formulated by a prior model for which locally smooth solutions are very likely. Local deviations of the smoothness are consequently outliers from this prior model. We can find various ways to model outliers in the literature.

One class of approaches introduce additional parameters which explicitly detect outliers. The advantage of these is that they can put further constrains on the resulting outliers maps, *e.g*., continuity. Often these methods lead to coupled systems in which the parameter- and outlier maps interact with each other. Geman and Geman [40] for instance proposed the additional use of a so-called line process. This process estimates additional edge or outlier labels which are used to break the smoothness assumption locally. The Mumford Shah approach [74] is a continuous version of such a line process. Other examples are proposed by Proesmans *et al*. [90, 89] in the context of optical flow computation and image enhancement.

Another class of approaches was pioneered by Blake and Zisserman [10]. They showed that the above-mentioned line process [40] can be eliminated by using robust estimators. This approach leads in this context to reweighted least square optimisation problems, where the weights play the role of the outliers maps in the previous class of methods. Also, the Perona-Malik model [81] can be interpreted in this context. Some popular examples are given by Rangarajan *et al*. [91] in the context of image segmentation and by Black *et al*. [8] and Brox *et al*. [12] for the estimation of optical flow.

Formulations based on a line process as well as formulations that eliminate this process based on robust estimators implement inhomogeneous non-linear diffusion.

The diffusivity is defined as a function of the pixel coordinates and the resulting regulariser will change at each iteration to take care of the updated diffusivity. Different with respect to this is the regularisation based on the structure tensor. The prior model does thereby assume only directional smoothness. Applied to the depth regularisation, this means that anisotropic diffusion is realised. If the image contains a high intensity gradient, the depth is assumed to be smooth orthogonal to his direction only. In uniform intensity areas, all directions are equally important and the smoothing is locally isotropic. If the diffusion tensor is based on the reference image, the regularisation term can be computed once and used in every iteration. Some examples in the context of optical flow estimation are given by Jähne [49] and Alvarez *et al*. [4, 3], where the last is also applied to the estimation of disparity. There are also non-linear anisotropic diffusion approaches, where the structure tensor is modified by the current solution, *e.g.*, for the estimation of optical flow by Brox *et al*. [13]. A more detailed view of diffusion methods is given by Weickert *et al*. [116]. For our depth regularisation, we will consider only anisotropic regularisation schemes as discussed in sec.4.3.2.

This chapter has a similar structure as the previous one. To a large extent it can be read without the knowledge of the previous chapter. Nevertheless, it is the intention to relate both approaches to each other, stress the differences and show the conceptual similarities. We first provide the generative model in sec. 4.2. The prior section 4.3 will get much attention because of the essential difference of the depth prior w.r.t. the MRF depth prior. The MAP formulation in sec. 4.4 and the EM solution in sec. 4.5 are similar to those in the previous chapter but with the specific generative and prior model. Finally, we provide experiments in section 4.6, which allows us to judge the advantages and disadvantages of both approaches on the same data sets.

4.2 Generative imaging model

As for the global formulation in the last chapter, we start by defining the generative model that specifies the way our input images are supposed to be generated. Although the local generative imaging model seems similar to the one described in the last section 3.5, there is one important difference. The depth D_i of a pixel i in $\mathbf{y}^*$ is *not* assumed to be discretised into the depth levels d_r. Nevertheless, there are many similarities. Also, the input images are considered to be generated by either one of two processes:

- The *inlier process* (fig 4.1) generates the pixels y_i^k which are visible in $\mathbf{y}^*$ and which obey the constant brightness assumption up to a global colour transformation $\boldsymbol{C}_k$, which can be different for each input image $\mathbf{y}^k$.
- The *outlier process* will generate all other pixels.

The inlier process is modelled as:

$$y^k_{i'(D_i)} = \boldsymbol{C}^{-1}(\boldsymbol{p}^k) \circ y^*_i + \epsilon \,, \tag{4.1}$$

where ϵ is the image noise, which is also assumed to be normally distributed with zero mean and covariance $\boldsymbol{\Sigma}$. Again, $\boldsymbol{C}^{-1}(\boldsymbol{p}^k)$ models the global colour transformation

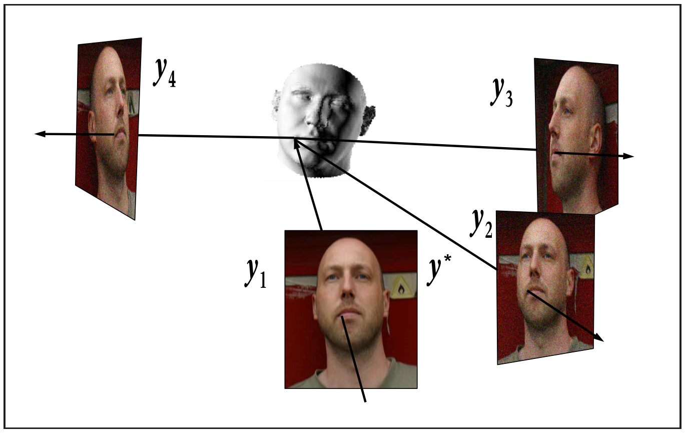


Figure 4.1: **Image generation:** *Image formation for the inlier process (top). The pixels of* $\mathbf{y}^{1,2,3,4}$ *are generated by adding noise to the geometric and photometric warp of* $\mathbf{y}^*$.

between the k^{th} input image $\mathbf{y}^k$ and the ideal image $\mathbf{y}^*$. The depth-dependent mapping $i'(D_i) \leftrightarrow i$ is different to eq. (3.1). It is now dependent on the *continuous* depth D_i and will be a part of the model parameters $\boldsymbol{\theta}$. Remember, for the global formulation, depth has (together with the visibility) been interpreted as a hidden MRF. As such we considered possible *discrete* depth-visibility realisations of the scene. For the following local formulation, only the visibility configurations are modelled as a MRF and the depth parameter is subject to a PDE-based minimisation.

The outlier process is modelled as a random generator, sampling from the unknown distribution g. This is identical to the global formulation. Both, the inlier and outlier process are selected by a hidden MRF $\mathbf{x}$, which includes the visibility configurations $v_i^s, s = 1 \dots S$ as states. Identical to the previous chapter (sec. 3.5), each visibility configuration v_i^s describes one configuration of the individual image visibilities v_i^{sk}.

Again, $f(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denote a normal PDF with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ and $g(.;\mathbf{h}^k)$ is the outlier distribution associated with the k^{th} image. Similar to the global formulation, let x_i^s be the state which is 1, and let $y^k_{i'(\mathcal{D}_i)}$ be the pixel in the k^{th} image onto which y_i^* is mapped. Then the probability of observing $y^k_{i'}$, conditioned on the unknowns $\boldsymbol{\theta} = \{\mathcal{D}, \mathbf{y}^*, \boldsymbol{\Sigma}, \mathbf{h}^k, \mathbf{p}_k\}$ and the state of the MRF $\mathbf{x}_i$, is given by:

$$p(y^k_{i'(\mathcal{D}_i)}|\boldsymbol{\theta},\mathbf{x}) = \left\{ \begin{array}{ll} f(\boldsymbol{C}(\boldsymbol{p}^k) \circ y^k_{i'(\mathcal{D}_i)}; y_i^*, \boldsymbol{\Sigma}) & \text{if} \quad v_i^{sk} = 1 \\ g(y^k_{i'(\mathcal{D}_i)}; \mathbf{h}^k) & \text{if} \quad v_i^{sk} = 0 \end{array} \right\} . \tag{4.2}$$

The inlier model is selected when $v_i^{sk} = 1$, *i.e.* when the pixel i (being in state $x_i^s = 1$) is visible in the kth input image.

4.3 Priors

4.3.1 Visibility prior

The MRF $\mathbf{x}$ represents the unobservable visibility state of each pixel in the ideal image $\mathbf{y}^*$. A Gibbs prior models the interaction of neighbouring visibility configurations, similar to sec. 3.6.1:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{i} \prod_{j \in N_i} \psi_{ij}(x_i, x_j) \ , \tag{4.3}$$

where $\psi_{ij}(x_i, x_j)$ depends on the distance $D_{ij}(s, q)$ between two visibility configurations, defined in eq. (3.5). $\psi_{ij}(x_i, x_j)$ and is given by:

$$\psi_{ij}(x_i^s, x_j^q) = \exp\left(-\sigma_v D_{ij}(s, q)\right) + C \ . \tag{4.4}$$

This interaction is the most general formulation. It provides the possibility that outliers from different images can interact with each other. More particular, it is for instance possible to increase the probability of a pixel i being visible in one image if a neighbouring pixel is also visible in *another* image.

A simpler model demanding less time and memory, would be to neglect those interactions and consider only the inter image spatial visibility interactions. In this case, a MRF $\mathbf{x}^k$ is introduced for every image $\mathbf{y}^k$ and describes the two possible states (inlier and outlier). The spatial correlation is modelled by the Ising model, described in appendix C.2.1. This model has also been used by Fransens *et al*. [31] and De Smet *et al*. [17] to model spatially correlated outliers.

If one would further simplify the model and neglect also spatial interactions, the visibilities can be estimated in closed form. The Bayes' estimate for a pixel i being visible in image k $p(x_i^k = 1 \,|\, \boldsymbol{\theta}, y_i^k)$ leads to the uncorrelated visibility case (see app. C.2.1) and is given by:

$$\mathcal{V}_i^k = \frac{f(\boldsymbol{C}(\boldsymbol{p}^k) \circ y^k_{i'(\mathcal{D}_i)}; y_i^*, \boldsymbol{\Sigma})}{f(\boldsymbol{C}(\boldsymbol{p}^k) \circ y^k_{i'(\mathcal{D}_i)}; y_i^*, \boldsymbol{\Sigma}) + g(y^k_{i'(\mathcal{D}_i)})} \ . \tag{4.5}$$

Fig. 4.2 shows this uncorrelated case graphically. For the experiments in this chapter, we will use the first (fully correlated) prior model. The second model, with spatial correlations of the visibilities only, is applied when full size images are considered in chapter 5.

4.3.2 Depth prior

The prior on the depth parameter is divided into two parts. One part is defined on every pixel in $\mathbf{y}^*$, *i.e.*, a smoothness prior, and one part incorporates the sparse set of initial 3-D points that is provided by the calibration procedure.

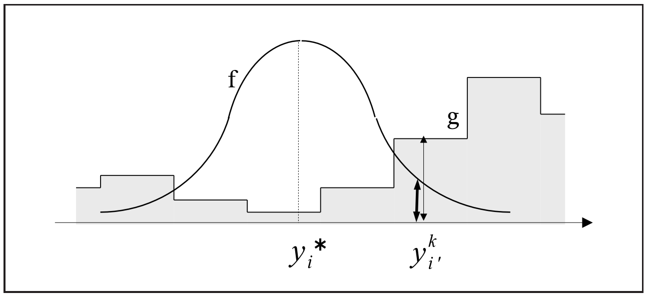


Figure 4.2: **Uncorrelated visibilities:** The probability of y_i^* being visible in the k^{th} image is proportional to its value under the Gauss-curve (bold, left arrow), *i.e.*, the distribution $f(\boldsymbol{C}(\boldsymbol{p}^k) \circ y^k_{i'(\mathcal{D}_i)}; y_i^*, \boldsymbol{\Sigma})$. The probability of y_i^* being invisible in the k^{th} image is proportional to the value under the histogram-based estimator $g(y^k_{i'(\mathcal{D}_i)})$ (thin, right arrow).

Smoothness depth prior

The formulation of appropriate depth priors is probably the most interesting issue for the stereo problem. Currently, priors are defined only locally by making smooth depth configurations more likely. Obviously, it would be of great use to define priors over more extended image patches or even to model primitive shapes. Priors based on image patches have been introduced by Roth and Black [94] in the context of optical flow computation. In this work it was suggested to use learned multiple experts as an optical flow prior. These experts correspond to likely configurations of the optical flow field on a patch of pixels. For the two view stereo problem a similar idea was recently evaluated by Kong *et al*. [65].

The disadvantage with these more informed priors is their need for representative training data, which is often not available. We therefore take a much simpler prior model, which assumes that the prior belief in the depth $p(\mathcal{D})$ can be parameterised by an exponential density distribution of the form:

$$p(\mathcal{D}) = \frac{1}{Z} \exp\left(- \frac{|\, R(\mathcal{X}, \mathcal{D}) \,|}{\lambda_s} \right) , \tag{4.6}$$

where λ_s is a parameter which controls the width of the distribution, and $R(\mathcal{X}, \mathcal{D})$ is a regulariser. This regulariser is driven by the function $\mathcal{X}$. From such a regulariser, we expect that it reflects our prior belief that the world is essentially simple, *i.e.*, for a locally smooth solution $\mathcal{D}$ in the neighbourhood of a particular point i, its value should approach zero, making such a solution very likely. Vice-versa, large depth fluctuations should result in large values for the regulariser, making such solutions less likely. Furthermore, the regulariser should be able to break the above mentioned smoothness assumption: if the value of $\mathcal{X}$ suggests a depth discontinuity, a large depth discontinuity at i should not be made a-priori unlikely. Such regularisers are commonly used in the PDE-community, where they serve as *anisotropic* or *inhomogeneous diffusion* *operators* for the computation of optical flow or edge-preserving image smoothing. Weickert *et al*. [117] presented a taxonomy of different diffusion operators. According to that, diffusion operators are distinguished between isotropic and anisotropic operators and both categories are further classified according to $\mathcal{X}$. We only consider anisotropic operators, and discuss possible realisations of $\mathcal{X}$.

Anisotropy diffusion operators can be written as:

$$R(\mathcal{X}, \mathcal{D}) = \nabla \mathcal{D}^T T(\nabla \mathcal{X}) \, \nabla \mathcal{D} \,, \tag{4.7}$$

where $T(\nabla \mathcal{X})$ is the diffusion tensor defined by:

$$T(\nabla \mathcal{X}) = \frac{1}{|\nabla \mathcal{X}|^2 + 2\nu^2} \Big(\nabla \mathcal{X}^{\perp} \nabla \mathcal{X}^{\perp T} + \nu^2 \boldsymbol{I} \Big) \;. \tag{4.8}$$

The diffusion tensor is a 2×2 matrix, where ν controls the degree of anisotropy, $\nabla \mathcal{X}^{\perp}$ is the vector perpendicular to $\nabla \mathcal{X}$ and $\boldsymbol{I}$ is the identity matrix. For $\nu \to \infty$ the diffusion tensor is equal to the scaled identity matrix $T(\nabla \mathcal{X}) = 0.5\boldsymbol{I}$. In this case $p(\mathcal{D})$ is independent on the direction $\nabla \mathcal{D}$ and isotropic diffusion is realised. If, on the other hand, $\nu \approx |\nabla \mathcal{X}|$ the prior probability of $\mathcal{D}$ might still be high when $\nabla \mathcal{D}$ is parallel to $\nabla \mathcal{X}$. For instance if $\mathcal{X}$ is the reference image, a large value of $|\nabla \mathcal{D}|$ will be allowed if $\nabla \mathcal{X}$ is parallel to $\nabla \mathcal{D}$, which is exactly the desired anisotropic behaviour.

Having defined the parametric form of the prior, we are now in the position to describe the possible realisation of $\mathcal{X}$. The best feature for $\mathcal{X}$ is the depth itself. Using $\mathcal{X} = \mathcal{D}$ to regularise the depth corresponds to flow-driven regularisation schemes in the context of optical flow [117, 13]. Thereby every depth configuration $\mathcal{D}$ which is directionally smooth will obtain a high prior probability. Another widely used feature to construct the diffusion tensor is the reference or ideal image. All depth configurations which are smooth perpendicular to the image gradient direction are assumed to be likely. This approach is justified by the observation that depth discontinuities often fall together with high image gradients. Both sources of anisotropy have advantages and disadvantages and we will therefore also consider a combination of both. The matrix $\nabla \mathcal{X}^{\perp} \nabla \mathcal{X}^{\perp T}$ is computed as a weighted sum over the individual features. For the weight of each feature we use the Mahalanobis distance related to a diagonal Gauss distribution of all derivative vectors $\mathcal{X}_i$. We continue the discussion with an evaluation of prior distributions, which are extracted from ground truth data.

Ground truth evaluation of the depth prior

Figure 4.3 shows the distribution for different $\mathcal{X}$ on the synthetic data used in section 3.10. More particularly, we first computed the value of $R(\mathcal{X}, \mathcal{D})$ for every pixel using the ground truth values for $\mathcal{D}$ and $\mathbf{y}^*$. Secondly a histogram was built over $R(\mathcal{X}, \mathcal{D})$. The result is the ideal distribution $p^*(\mathcal{D})$ for this particular dataset, which we will use to illustrate the goodness of the assumed parametric prior distribution $p(\mathcal{D})$ in eq. (4.6).

The left plot in fig. 4.3 compares the probability distributions $p^*(\mathcal{D})$ for the isotropic case $T = T(0.5\boldsymbol{I})$ with the depth based anisotropic case $T = T(\nabla \mathcal{D})$. About 99.5%

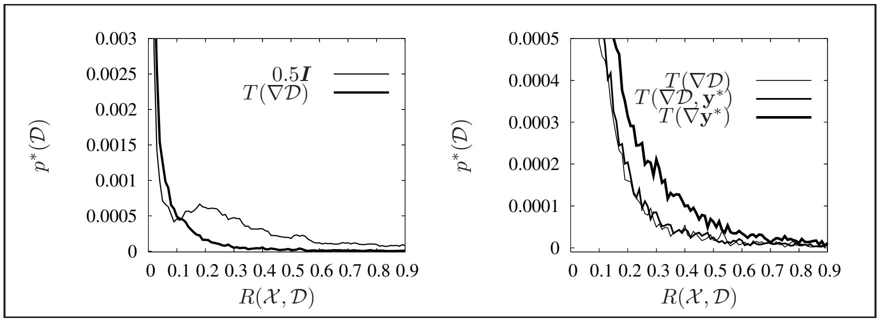


Figure 4.3: **Prior distribution:** *The distribution $p^*(\mathcal{D})$ for different diffusion tensors $T(\nabla\mathcal{X})$ is shown as a function of $R(\mathcal{X},\mathcal{D})$ for the synthetic data set in fig. 3.6.*

of the pixels have a value $R(\mathcal{X},\mathcal{D})$ which is smaller than 0.1, *i.e.* for almost all pixels $\mathcal{D}_i$ is smooth. The difference lies in the modelling of non-smooth depth values: whereas the anisotropic distribution does not contain many pixels with a larger value of $R(\mathcal{X},\mathcal{D})$, this cannot be said about the isotropic case. One can therefore conclude that the parametric form of $p(\mathcal{D})$ as given by eq. (4.6) and in combination with eq. (4.8) is well suited for the depth based anisotropic regulariser $T = T(\nabla\mathcal{D})$ but not for an isotropic regulariser.

The right plot in fig. 4.3 shows a close-up for three distributions: one, which was already plotted in the left plot, *i.e.* based on the depth $T = T(\nabla\mathcal{D})$, and two anisotropic regularisers based on the ideal image $T = T(\nabla\mathbf{y}^*)$ and on the combination of both $T = T(\nabla\mathcal{D}, \nabla\mathbf{y}^*)$. These plots show an exponential fall-off for all distributions. We can see that large values of $\mid \nabla\mathcal{D} \mid$ are nicely modelled by these distributions. For the image-based regularisation scheme, this also shows that depth gradients coincide with intensity gradients in our test set.

Fig. 4.4 shows the colour-coded magnitude of the diffusion tensor entries for the three anisotropic diffusion tensors. One can see only a tiny difference between the image-based and the combined image-depth diffusion tensor. This difference is visible at the borders of the face, mainly where the intensity gradient between fore- and background is less strong.

The above considerations can be seen as a justification for the relatively simple form of the depth prior in eq. (4.6). There exist, of course, many more advanced regularisation schemes which can deal with outliers from the smoothness assumption (see sec. 4.1.2). These are not considered here. We believe that regularisation should be seen in a Bayesian context where training should play an essential role in formulating prior models. This however is beyond the scope of this thesis and we restrict ourself to formulate the probabilistic framework in which one could plug in more advanced prior models easily.

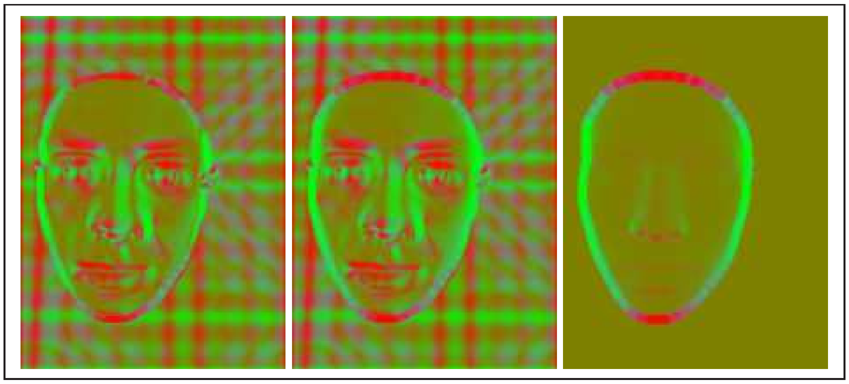


Figure 4.4: **Anisotropic behaviour:** *The images show the colour coded magnitude of the diffusion tensor entries. The diagonal elements are in red/green and the off-diagonal element is coded in blue. From left to right one can see the: image-based diffusion tensor* $T = T(\nabla \mathbf{y}^*)$*, the combined* $T = T(\nabla \mathbf{y}^*, \nabla \mathcal{D})$ *and the depth based diffusion tensor* $T = T(\nabla \mathcal{D})$.

Scale invariance

The regularisation of the depth is strongly dependent on the scale. We can consider two different kinds of scales. First, there is the scale ambiguity of uncalibrated ”structure and motion”. Two calibrations of the scene, which differ by the Euclidean scale, will lead to two different prior distributions $p(\mathcal{D})$. And second, there is the scale, which is introduced by defining the problem on different pyramid levels. To account for these scale dependencies, we will treat the width of the prior distribution λ_s in eq. (4.6) as part of the MAP estimation problem. By doing so, the formulation is made invariant to both kinds of scale changes.

The width λ_s also indicates the strength of the prior. λ_s is the well known factor that weights the contribution of the smoothness term relative to the matching term. This factor is present in all energy formulations that require regularisation. By taking λ_s as an unknown parameter, we loose control over the relative weighting. Therefore, we introduce an additional parameter λ. This parameter reflects the uncertainty of the depth prior and will have to be set by hand. The likelihood distribution in eq. (4.2) will therefore be replaced with:

$$p(y^k_{i'(\mathcal{D}_i)}|\boldsymbol{\theta}, \mathbf{x})^{\lambda} \leftarrow p(y^k_{i'(\mathcal{D}_i)}|\boldsymbol{\theta}, \mathbf{x}) \; . \tag{4.9}$$

Ideally, the value of λ will be one, independent of the Euclidean scale of the reconstruction and also of the pyramid level which is considered. In the experimental section of this chapter, we will evaluate the dependency on this parameter.

Calibration points depth prior

The second part of the depth prior relates the depth estimate of certain pixels to the already known value. Initial 3-D points, which are provided by self-calibration [82], will project to the ideal image $\mathbf{y}^*$. For the closest pixel i, the depth $\mathcal{G}_i$ is therefore approximately known. We model the depth prior for these points by a Gaussian distribution:

$$p(\mathcal{D}) = \frac{1}{Z} \prod_i \exp\left(-\frac{1}{\lambda_c} \left(\mathcal{D}_i - \mathcal{G}_i\right)^2 \right)^{\mathcal{W}_i} . \tag{4.10}$$

Whenever an initial 3-D point is projected, the closest pixel will have a non-zero weight $\mathcal{W}_i$. This weight is related to the certainty of this particular $3D$ point. All other values of $\mathcal{W}_i$ are zero.[1]. The parameter λ_c is used to globally weight the relative influence of the initial $3D$ points.

The overall depth prior is now based on the product of the two depth prior distributions in eq. (4.6) and eq. (4.10).

4.4 MAP estimation

Let $\boldsymbol{\theta} = \{\mathcal{D}, \mathbf{y}^*, \boldsymbol{\Sigma}, \mathbf{h}^k, \boldsymbol{p}^k, \lambda_s\}$ denote all parameters, and let $\mathbf{y} = \{\mathbf{y}^k\}$ denote all input data. The maximum a-posteriori probability (MAP) estimate of the unknowns $\boldsymbol{\theta}$ is given by:

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \Big\{ \log \sum_{\mathbf{x}} p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta})\, p(\mathbf{x})\, p(\boldsymbol{\theta}) \Big\} , \tag{4.11}$$

Conditioned on the state of the hidden variable $\mathbf{x}$, the data-likelihood factorised as a product over all individual pixel likelihoods:

$$p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}) \approx \prod_i \prod_k \prod_s p(y^k_{i'} \,|\, x^s_i, \boldsymbol{\theta})^{x^s_i} . \tag{4.12}$$

In the product over s only the factor for which $x^s_i = 1$ survives. This product includes in this formulation, different from the global formulation, only contributions related to the possible visibility configurations, *i.e*. the state x^s_i corresponds to a particular visibility configuration v^s which are shown for three images in table 3.1. Based on these visibility values, the pixel-likelihood in the right hand side of eq. (4.12) can be further expanded as:

$$p(y^k_{i'} \,|\, x^m_i, \boldsymbol{\theta}) = \Big[f(\boldsymbol{C}(\boldsymbol{p}^k) \circ y^k_{i'}; y^*_i, \boldsymbol{\Sigma}) \Big]^{v^{sk}} \Big[g(y^k_{i'}; \mathbf{h}^k) \Big]^{1-v^{sk}} . \tag{4.13}$$

This pixel-likelihood is given by the inlier distribution if the visibility configuration v^s describes the situation for which the pixel is visible in the kth image, *i.e*. $v^{sk} = 1$ and the pixel-likelihood is given by the outlier distribution if v^s describes the situation for which the pixel is not visible in the kth image, *i.e*. $v^{sk} = 0$. We have now specified

---

[1] We use binary values for $\mathcal{W}_i$, since the calibration proceedure we use does not provide certainties.

all terms of the data-likelihood $p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta})$ and the prior on the MRF in eq. (4.3). The sum $\sum_{\mathbf{x}}$ in the right hand side of eq. (4.11) ranges over all possible configurations of the random field $\mathbf{x}$ and one can use here, similarly to the global approach, the EM algorithm to deal with this problem.

4.5 EM algorithm

Given the specific form of the prior in eqs. (4.6) and (4.10) and the data likelihood in eqs. (4.13) and (4.9), we can construct the free energy similarly to the previous chapter 3.8.1 and as explained in appendix C. By applying the mean field approximation, we get:

$$\begin{aligned} F_{MF} \approx\; & -\lambda \sum_i \sum_k \sum_m b_i^m \log p(y_{i'}^k \,|\, x_i^m, \boldsymbol{\theta}) \\ & + \frac{1}{\lambda_s} \sum_i R(\mathcal{X}_i, \mathcal{D}_i) + \frac{1}{\lambda_c} \sum_i \mathcal{W}_i (\mathcal{D}_i - \mathcal{G}_i)^2 \\ & - \sum_i \sum_{j \in N_i} \sum_{m,n} b_i^m b_j^n \log \psi_{ij}^{mn} \\ & + T \sum_i \sum_m b_i^m \log b_i^m \; . \end{aligned} \tag{4.14}$$

The difference with respect to the mean field free energy for the global approach in eq. (3.16) is provided by the additional terms related to the depth prior and the definition of the MRF states, which include here only the visibility configurations.

4.5.1 E-step

On the $(t+1)^{th}$ iteration, the conditional expectation of the complete log-likelihood w.r.t. the posterior $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta}^{(t)})^{1/T}$ is computed in the E-step. The update equation for the expected values b_i^n of the MRF field states is similar to eq. (3.17) given by:

$$b_i^m \leftarrow \exp\left(\frac{1}{T} \sum_{j \in N_i} \sum_n b_j^n \log \psi_{mn} + \frac{1}{T} \sum_k \log p(y_{i'}^k \,|\, x_i, \hat{\boldsymbol{\theta}}^{(t)}) - 1 \right) . \tag{4.15}$$

Again, the visibilities for each image are computed by the expected value over the node beliefs b_i^m:

$$\mathcal{V}_i^k = \sum_s b_i^s v_i^{sk} \; . \tag{4.16}$$

4.5.2 M-step

At the M-step, the intent is to compute values for $\boldsymbol{\theta}$ that minimise eq. (4.14), given the current estimates of the visibilities $\mathcal{V}_i^k$. This is achieved by setting the parameters $\boldsymbol{\theta}$ to the appropriate root of the derivative equation,

$$\partial F_{MF}(\boldsymbol{\theta})/\partial\theta = \mathbf{0}\ . \tag{4.17}$$

For the image related parameters y_i^* and $\mathbf{\Sigma}$, the update equations are:

$$y_i^* \leftarrow \frac{\sum\limits_k \mathcal{V}_i^k\ \boldsymbol{C}(\boldsymbol{p}^k) \circ y_{i(\mathcal{D}_i)}^k}{\sum\limits_k \mathcal{V}_i^k}\ , \tag{4.18}$$

$$\mathbf{\Sigma} \leftarrow \frac{\sum\limits_k \sum\limits_i \mathcal{V}_i^k\ \big(\boldsymbol{C}(\boldsymbol{p}^k) \circ y_{i(\mathcal{D}_i)}^k - y_i^*\big)\big(\boldsymbol{C}(\boldsymbol{p}^k) \circ y_{i(\mathcal{D}_i)}^k - y_i^*\big)^T}{\sum\limits_k \sum\limits_i \mathcal{V}_i^k}\ , \tag{4.19}$$

the colour transformation $\mathbf{p}^k$ is given by solving the linear system:

$$\mathbf{C}(\mathbf{p}^k) \sum_i \mathcal{V}_i^k y_{i(\mathcal{D}_i)}^k (y_{i(\mathcal{D}_i)}^k)^T = \sum_i \mathcal{V}_i^k y_i^*\ (y_{i(\mathcal{D}_i)}^k)^T\ , \tag{4.20}$$

and the scale by:

$$\lambda_s^{-1} = \frac{1}{N} \sum_i \mid R(\mathcal{X}, \mathcal{D}) \mid \tag{4.21}$$

To arrive at these closed-form expressions, we ignored the effects of these variables on the regularisation term. This is admissible because their influence on the depth regulariser $R(\mathbf{y}^*, \mathcal{D})$ is small compared to their influence on the matching term. $\mathbf{\Sigma}$ is only indirectly related to $R(\mathbf{y}^*, \mathcal{D})$ by way of computation of the visibility maps, which have an effect on $R(\mathbf{y}^*, \mathcal{D})$ via the computation of y_i^*. The image y_i^* has an effect on $R(\mathbf{y}^*, \mathcal{D})$ via its gradient, which is used to define a quadratic norm on the depth gradient (4.7). Changes of y_i^* will therefore only exert a minor influence on $R(\mathbf{y}^*, \mathcal{D})$.

However, for the update of the depth map $\mathcal{D}$ we are not so lucky, because $\mathcal{D}$ strongly influences both the matching and the regularisation term. To minimise F_{MF} w.r.t. $\mathcal{D}$, we solve the corresponding diffusion equation. This can be derived from eq. (4.14) by using the Euler-Lagrange equation and is given by:

$$\begin{aligned} \frac{\partial \mathcal{D}}{\partial t} &= \text{div}(T(\nabla\mathcal{X})\nabla\mathcal{D}) \\ &+ \lambda \sum_k \mathcal{V}^k \frac{\partial(\boldsymbol{C}(\mathbf{p^k})\mathbf{y}_{\mathcal{D}}^k - \mathbf{y}^*))^T \mathbf{\Sigma}^{-1}(\boldsymbol{C}(\mathbf{p^k})\mathbf{y}_{\mathcal{D}}^k - \mathbf{y}^*)}{\partial \mathcal{D}} \\ &- \frac{2}{\lambda_c}\mathcal{W}(\mathcal{D} - \mathcal{G})\ , \end{aligned} \tag{4.22}$$

where $\mathbf{y}_{\mathcal{D}}^k$ is the colour value of the kth input image interpolated at the current depth value $\mathcal{D}$. For the solution of 4.22 we use a symmetric Gauss-Seidel scheme [88]. Other solution schemes based on multi grid methods would also be possible. Examples for these are studied by Bruhn *et al*. [14] applied to the computation of optical flow. The existence and uniques of this parabolic equation has not been proved, However, for the case of two-view stereo, when disparity is considered, the existence and uniques of the solution to eq. (4.22) has been proved by Alvarez *et al*. [3]. Details on our solution are given in appendix D. The whole algorithm is graphically depicted in table 4.1.

```
Initialisation: $b_i = uniform$, $\mathbf{y}^* = \mathbf{y}^1$
$\mathbf{\Sigma}$ is diagonal, with entries $\sigma = 100$
for all initial 3-D points $\mathcal{D}_i = \mathcal{G}_i$
all other depths are initialised by $\mathcal{D}_i = max(\mathcal{G}_i)$
Loop over pyramids:
    until convergence:
        M-step
            compute diffusion tensor
            until convergence:
                compute $\mathcal{D}$ by solving the diffusion equation (4.22)
            compute $\mathbf{y}^*$ by eq. (4.18)
            compute $\mathbf{\Sigma}$ by eq. (4.19)
            compute $\lambda_s$ by eq. (4.21)
            compute every $\mathbf{p}^k$ by solving eq. (4.20)
        E-Step
            Estimate visibilities $\mathcal{V}^k$ by eq.(4.16)
            and using the mean field update eq. (4.15)
```

Table 4.1: **Outline of the local algorithm.**

4.5.3 Relation to other PDE based formulations

Consider the diffusion equation (4.22) for the case that the ideal image camera position to be place at $\mathbf{y}^1$ and without the energy term which penalises deviations of $\mathcal{D}$ form the sparse initialisation ($\lambda_c \to \infty$). With these assumptions we can relate eq. (4.22) to the PDE-based stereo formulations by Proesmans *et al*. [90], Robert *et al*. [93], Alvarez *et al*. [3] and Slesareva *et al*. [100]. Eq. (4.22) simplifies to:

$$\frac{\partial \mathcal{D}}{\partial t} = \mathrm{div}(T(\nabla \mathcal{X}) \nabla \mathcal{D}) + \lambda \sum_k \mathcal{V}^k \frac{\partial (\mathbf{y}^* - \boldsymbol{C}(\mathbf{p^k}) \mathbf{y}^k_{\mathcal{D}})^T \mathbf{\Sigma}^{-1} (\mathbf{y}^* - \boldsymbol{C}(\mathbf{p^k}) \mathbf{y}^k_{\mathcal{D}})}{\partial \mathcal{D}} . \tag{4.23}$$

What can we read off from this diffusion equation:

- The smoothness term and the matching term are globally weighted by the image noise $\mathbf{\Sigma}$, *i.e.* if a large noise magnitude is present in the images, the smoothness term will become more important. This is an advisable mechanism, since in the presence of noise the depth will be relatively more smooth and will not try to match the (noisy) pixels compleately.

- The matching term is weighted by the visibilities, *i.e.* if a pixel has a high confidence of being an outlier w.r.t. the kth image ($\mathcal{V}_i^k \approx 0$) its importance to the matching term is decreased. To find the depth value of a pixel this mean that only the visible pixels are considered.

- The smoothness and the matching term are locally weighted by the visibility confidence. If the visibilities of a pixels w.r.t. all images $\mathcal{V}_i^k \approx 1, k = 1 \dots K$ is large we have a strong data confidence and the smoothness term is less important. In the other extream case where a pixel is detected as being an outlier w.r.t. to all input images $\mathcal{V}_i^k = 0, k = 1 \dots K$ only the smoothness term survives and depth is driven by the local neighbourhood.

- The relative importance of the image bands is globally weighted by the inverse covariance matrix $\mathbf{\Sigma}^{-1}$. For instance for images where each image band is measured by a different sensor, with possible different data range, the relative importance is adjusted automatically.

- The generative model tells to compare the ideal image $\mathbf{y}^*$ with all input images $\mathbf{y}^k$ and not the input reference image $\mathbf{y}^1$.

This formulation is different from Proesmans *et al.* [90], Robert *et al.* [93], Alvarez *et al.* [3] and Slesareva *et al.* [100] in that more than two images are used to estimate the depth of the reference camera. Furthermore our above mentioned automatic weighting mechanisms are not present this work [90, 93, 3, 100], *i.e.* the image noise is kept fixed and incorporated in the value of λ. Also the local visibility related weights $\mathcal{V}_i^k$ are often set to one for all pixels. In [100] a robust estimation scheme is used for which the weights $\mathcal{V}_i^k$ are the result of a reweighted least square optimisation with a fixed M-estimator. We have discussed the relation of our formulation with robust estimation in chapter 2.

4.6 Experiments

The experiments in this section are in close relation to the experimental section in the previous chapter 3.10. We want to evaluate the local approach as a function of the parameters and the source of anisotropy $\mathcal{X}$. It is further the intention to compare the results of the local and the global approach. We want to stress, however, that this comparison is somewhat inaccurate, since the local approach uses initial 3-D points to hold on to.

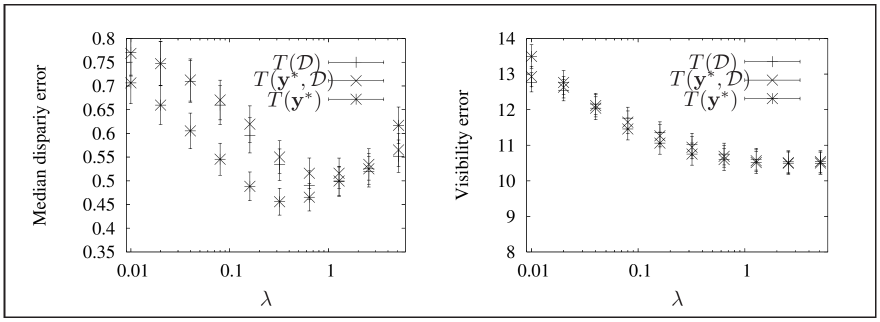


Figure 4.5: **Evaluation of λ:** *Median disparity error (left) and visibility error right as a function of λ ($\mu = 3.2$, $\sigma_v = 4$).*

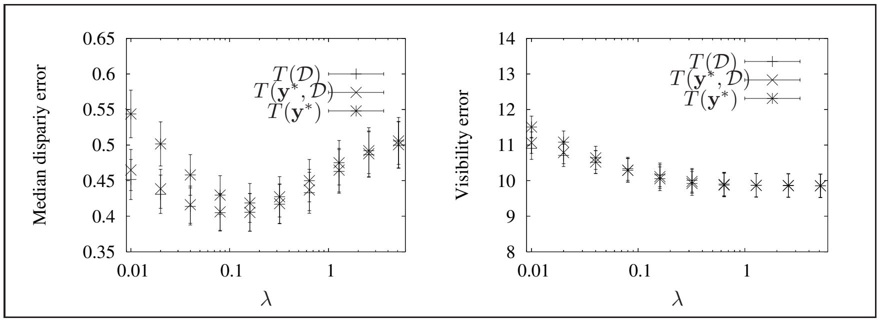


Figure 4.6: **Evaluation of λ:** *Median disparity error (left) and visibility error right as a function of λ ($\mu = 3.2$, $\sigma_v = 4$) when the parameters are initialised with the global approach.*

4.6.1 Ground truth evaluation

The synthetic ground truth evaluation is performed using the same 10 artificial test sets as in section 3.10 of which one example is shown in fig. 3.6. For one percent of the pixels (equally spread in the image), the ground truth depth $\mathcal{G}_i$ in eq. (4.10) is used as prior. Only for those pixels is $\mathcal{W}_i = 1$.

In the first experiment, we evaluate the performance as a function of the relative weight λ of data-likelihood and prior. Fig. 4.5 shows the median error and the visibility error for the three regularisers. All regularisers perform better than the global approach, which has a median depth error of ≈ 0.6 (see fig. 3.15). The regulariser, which is based on the ideal image ($T(\nabla \mathbf{y}^*)$), gives the best result. The two other regularisers perform similarly. This behaviour can only be explained by the local nature of the diffusion approach. The bad initialisation of $\mathcal{D}$ leads to a wrong estimate of the diffusion tensors $T(\nabla \mathcal{D})$ and $T(\nabla \mathbf{y}^*, \nabla \mathcal{D})$, which then again prevents global convergence. To validate this statement, a second experiment was set-up. In this the parameters $\mathcal{D}$, $\mathbf{\Sigma}, \mathbf{p}^k$ and $\mathcal{V}^k$ have been initialised by the value obtained from the global approach[2]. The results of this experiment are shown in fig. 4.6. Indeed, with a better initialisation of the parameters, the regularisers based on the depth perform better than the image-based regulariser. With initialisation the disparity-error decreases even further.

The value of λ which gives the best results lies in the range $0.1 \leq \lambda \ll 1$. This result compensates for the overestimation of the unknown scale λ_s in eq. (4.6). The parametric form of our prior distribution does not explictly take outliers into account and the value of λ_s will therefore be larger than it should be.

In the next experiment, we evaluate the three regularisation schemes with respect to ν, the strength of the anisotropy. Figure 4.7 shows the results for the median disparity error and the visibility error when we do not initialise with the global approach.

Here, we see the convergence of all three regularisers to a single error-value for $\nu \rightarrow \infty$. This limit implements isotropic diffusion. The best regulariser is again image-based and has an optimal value of $\nu \approx 1$. By initialising the parameters with the global approach, the depth based regularisers perform similarly or even better, which is shown in fig. 4.8. Also, the optimal value for ν is smaller compared to the results without initialisation. Both experiments show once again the advantage of a good initialisation. To be consistent with the experiments in the global approach, fig. 4.9 shows the result with respect to the visibility correlation strength σ_v. We can recognise only a minor influence, which is similar to the global approach due to a strong data-likelihood in the occluded areas.

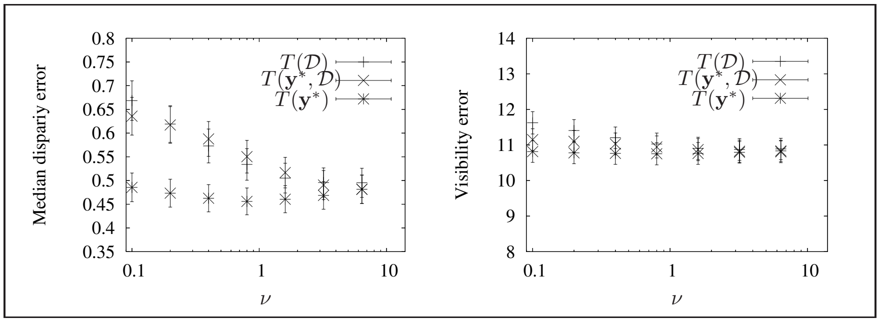


Figure 4.7: **Evaluation of ν:** *Median disparity error (left) and visibility error right as a function of ν ($\lambda = 0.32$, $\sigma_v = 4$).*

4.6.2 Outdoor scene reconstructions

The outdoor experiments are preformed on the same data sets as in the previous chapter. All images are processed only up to a size of 768×512 pixels to be comparable to

---

[2] The Bethe approximation was used with $T_s = 100$, $T_d = 3.5$, $T_e = 0.1$, $\sigma_d = 100$, $\sigma_v = 1$ and $C = 10^{-10}$.

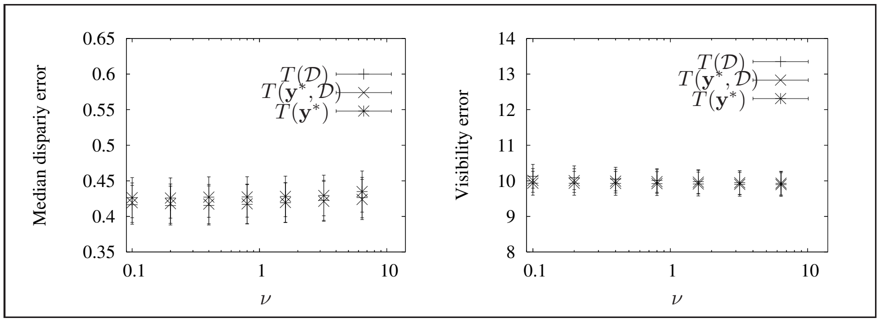


Figure 4.8: **Evaluation of ν:** *Median disparity error (left) and visibility error right as a function of ν ($\lambda = 0.32$, $\sigma_v = 4$) when the parameters are initialised with the global approach.*

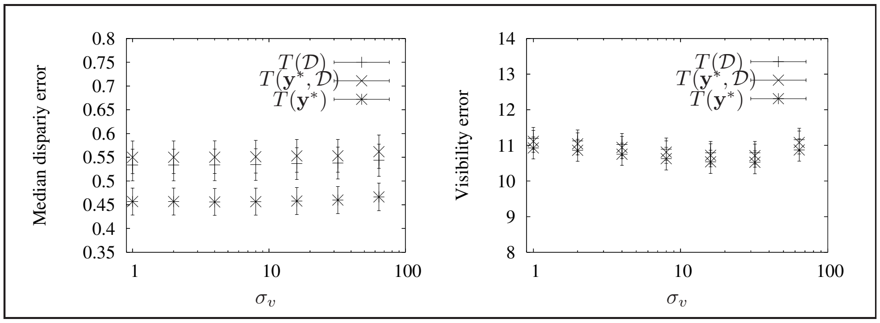


Figure 4.9: **Evaluation of σ_v:** *Median disparity error (left) and visibility error right as a function of σ_v ($\lambda = 0.32$, $\nu = 3.2$).*

the global approach. The difference is the use of the initial calibration points as prior knowledge.

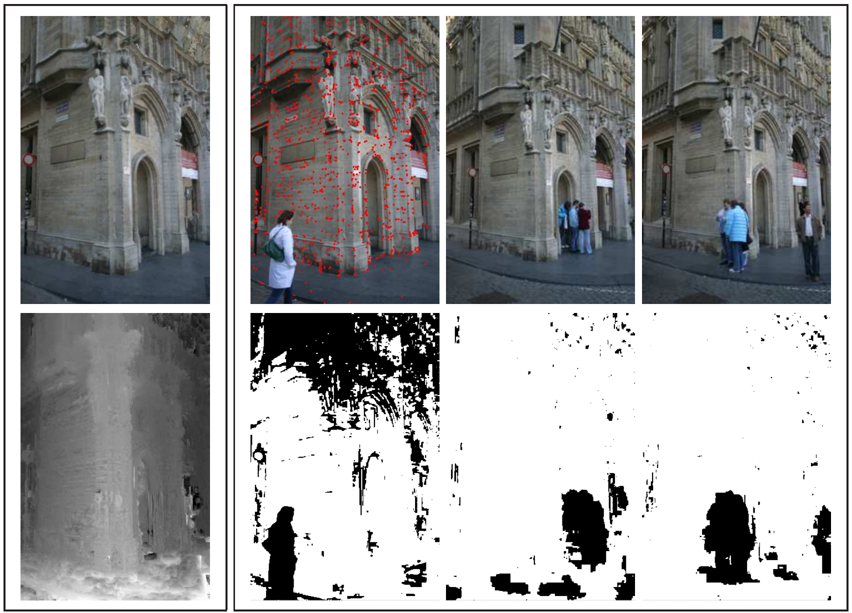
We used the image-based regularisation $T = T(\mathbf{y}^*)$ for all sequences and show the results obtained from the same set of parameters. As a global result we can state that the depth and visibility estimates are almost comparable to the results of the global approach. The weak point is the depth and visibility estimation near depth discontinuities. In these areas, the global approach has clear advantages, especially when initial 3-D are missing nearby. The Leuven city hall scene in fig. 4.12 shows this most clearly. There, the left part of the reference image does not have many initial 3-D points and the depth and visibility estimation is rather poor compared to the global approach (shown in fig. 3.24). Also, the discontinuities around the statue in fig. 4.11 are less sharp compared to fig. 3.23.

Another important problem is the presence of wrong initialisation points. Since the local approach needs these points, it is difficult to distinguish good initial points from bad points.

The local approach is about 10 times faster than the global approach. The speed can be increased further by a factor of ≈ 5 when the correlation of different visibilities $\mathcal{V}^k$ are neglected or when the Ising model (appendix C.2.1) is used for the visibility within images (as discussed in sec. 4.3.1).

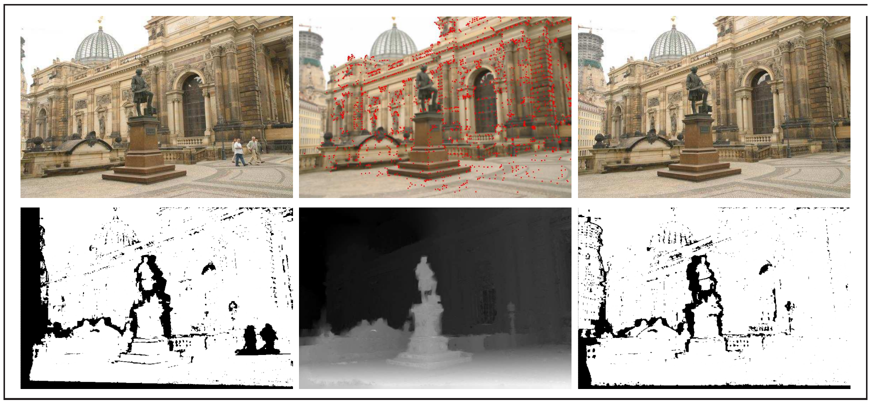
4.7 Conclusion

A multi-view stereo algorithm was presented for the estimation of depth and outliers. The problem has been addressed from a probabilistic point of view. One of the advantages of such an analysis is that it makes the implicit assumptions underlying a particular algorithm explicit. In our approach, the main assumptions are dominant diffuse reflection and i.i.d. pixel colour distributions. A smoothness regulariser was introduced to give shape to our prior beliefs about the world. The key result of



T_s	T_e	T_d	σ_v	C	λ	nr. initial 3-D points	time
20	0.1	2	200	10^{-10}	0.8	1333	$34sec$

Figure 4.10: **Brussels city hall scene:** *The three input images are the three right-most images shown in the top row. The camera position of the virtual image* $\mathbf{y}^*$ *was chosen to be the left of these images, which shows the initial* 3*-D points as red dots. The visibility estimates related to* $\mathbf{y}^*$ *are in the bottom row. The top-left image shows the estimated ideal image* $\mathbf{y}^*$ *and the estimated depth is shown in the bottom-left image. Similar to the global approach in fig. 3.22, the image size is* 768×512*.*



T_s	T_e	T_d	σ_v	C	λ	nr. initial 3-D points	time
20	0.1	2	200	10^{-10}	0.8	2694	$59sec$

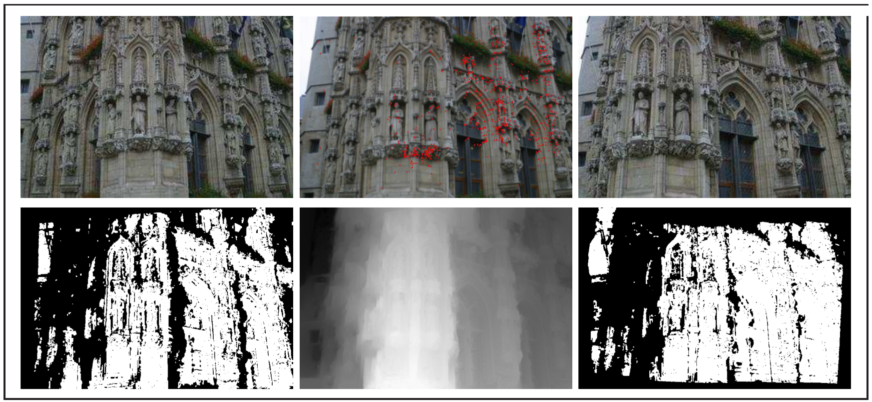
Figure 4.11: **Semper statue scene:** *The input images are shown in the top row. The middle image is chosen as the reference view. This image shows the initial* 3*-D points in red. The depth map for the reference view (middle) and outlier maps for the two other images are shown in the bottom row. The corresponding result for the global approach is shown in fig. 3.23.*

this probabilistic formulation is that energy minimisation, which is the cornerstone of PDE-based methods, is strongly related to MAP-estimation. More specifically, in terms of our notation, the typical energy-functional is a special case of eq. (4.23), in which $\mathbf{y}^*$ is defined to be the reference image, and where colour transformations and noise are supposed to have unit strength.

In this work, images are modeled as noisy measurements of a colour-transformed unknown irradiance or 'true' image function. This has three principal advantages. First of all, it brings about an automatic balancing between matching and smoothness. In early stages of the optimisation ($\mathbf{\Sigma}$ is still large), more emphasis is put on regularisation, whereas in the convergence stage ($\mathbf{\Sigma}$ reaches the true image noise), the matching term will gain importance. This is the major result from the probabilistic formulation of the problem. Also, we formulated the problem invariant to scale. Only the parametric form of the depth prior is fixed. The width of the prior distribution, which is related to the scale, is part of the optimisation procedure.

Secondly, because the true image is a learned model of image irradiance, we are able to leave the input camera positions, which in turn allows us to compute view interpolations as shown in chapter 5.2. Finally, the resulting model integrates all available image information, and can as such be used as a texture map for the final 3-D reconstruction.

A strong emphasis was put on the computation of visibility. The visibility of a particular pixel is modeled as a correlated mixture problem using a MRF. The expectancy of the outliers is sequentially updated in the EM algorithm.



T_s	T_e	T_d	σ_v	C	λ	nr. initial 3-D points	time
20	0.1	2	200	10^{-10}	0.8	1008	$62sec$

Figure 4.12: **Leuven cityhall scene:** *The three input images are shown in the top row. The camera position of ideal image* $\mathbf{y}^*$ *was chosen to be the middle image, which also shows the initial* 3*-D points. The depth map for the reference view (middle) and outlier maps for the two other images are shown in the bottom row. The corresponding result for the global approach is shown in fig. 3.24.*

The PDE-based, local approach described in this chapter needs consistent initial 3-D points for convergence. If these are provided, the local approach is much faster and yields competitive results when compared to the global approach. However, near depth discontinuities, we find a clear advantage of the global approach. The local approach has, on the other hand, advantages in continuous depth regions. The reason for this can be found in the undiscretised depth formulation. For a good performance on large images, the local approach is a good candidate, when either a good initialisation is provided by the global approach (as will be shown in section 5.1) or more effort is put in a possibly probabilistic formulation of self-calibration [27], which is expected to lead to more and especially more accurate initial 3-D points.

Chapter 5

Applications

> *If oo don't belief that oo r model (e.g. of norare errors) is correct, ch ose an ther one and use maximum likelih on - or Bayesian - methods for the new model. What, if I ontt belief in the new model either? It takes a e t of st ttornness to eooo the world with a host of rather irtetrary and protatey hardey interpretatee models ano seaia they are exactly true. The p int of rot st statestiss is that one may keep a parametric model hethoose the tatter is known to be wrong.*
>
> $\arg\max_{\mathbf{y}^*}\left\{ \log p(\mathbf{y}\,|\,\mathbf{y}^*)p(\mathbf{y}^*)\right\}$ of Hampel *et al*. [42] with
>
> $p(\mathbf{y}^*) \propto \prod_{ij\in[i\pm 1,2,3]} \psi_{ij}(y_i^*, y_j^*)$

In the last two chapters we proposed two multi-view stereo formulations. They have been compared on small resolution images. In this chapter we give some examples of depth reconstructions at full resolution as well as image reconstructions for virtual cameras.

5.1 Depth reconstruction

For the depth reconstruction at high resolution we combine the advantages of the global and the local approach. The global approach, which works well on small images, where it finds more easily a global optimum, is used as an initialisation for the local approach. This is used to compute the reconstruction on the full resolution images.

The results of the local approach represent 3-D reconstructions of the raw depth maps $\mathcal{D}$ as given by the solution of eq. (4.22), *i.e.* no median filter has been applied to the depth map. We show the reconstruction for all pixels y_i which are visible in at least two images, *i.e.* $\sum_k \mathcal{V}_i^k >= 2$ in eq. (4.16). Furthermore we set $\lambda_c = 0$ in eq. (4.22), which realises anisotropic diffusion without taking the initial 3-D points into account. This is possible since we use a relatively good initialisation by the global approach.

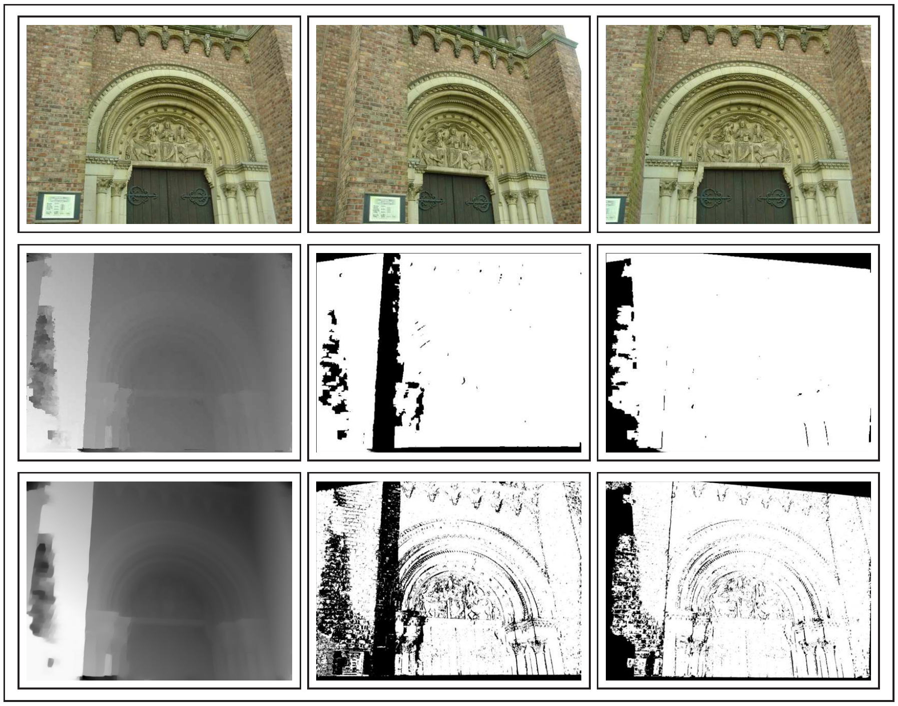


Figure 5.1: **Church scene:** *Three input images (top row). Depth and visibility for the global approach (middle row), which is used as initialisation to the local approach (bottom row).*

The renderings are not based on triangle meshes. We use a more efficient way to render these large models which can include more than 6 million 3-D points. It is based on QSplats as proposed for the Digital Michelangelo Project by Rusinkiewicz and Levoy *et al*. [97]. In QSplat all 3-D points, their colour, normal direction and radius are rendered as ellipses.

Church Scene

Three input images of a church are used with a resolution of 2592×1944 square pixels. They are shown in the top row of fig. 5.1. The global formulation was computed up to a resolution of 648×486 square pixels using the same parameters as in the outdoor experiments from chapter 3, *e.g.*, as in fig. 3.23. The depth map and the visibility of this initialisation is shown in the middle row of fig. 5.1. The input to the local approach is the depth, the visibility, the image noise and the colour transformation as computed by the global approach. We used the same parameters for the local approach as in the outdoor experiments in chapter 4 as for instance given in fig. 4.11. One exception is the value of the smoothness related parameters λ, which was set to $\lambda = 0.01$. This

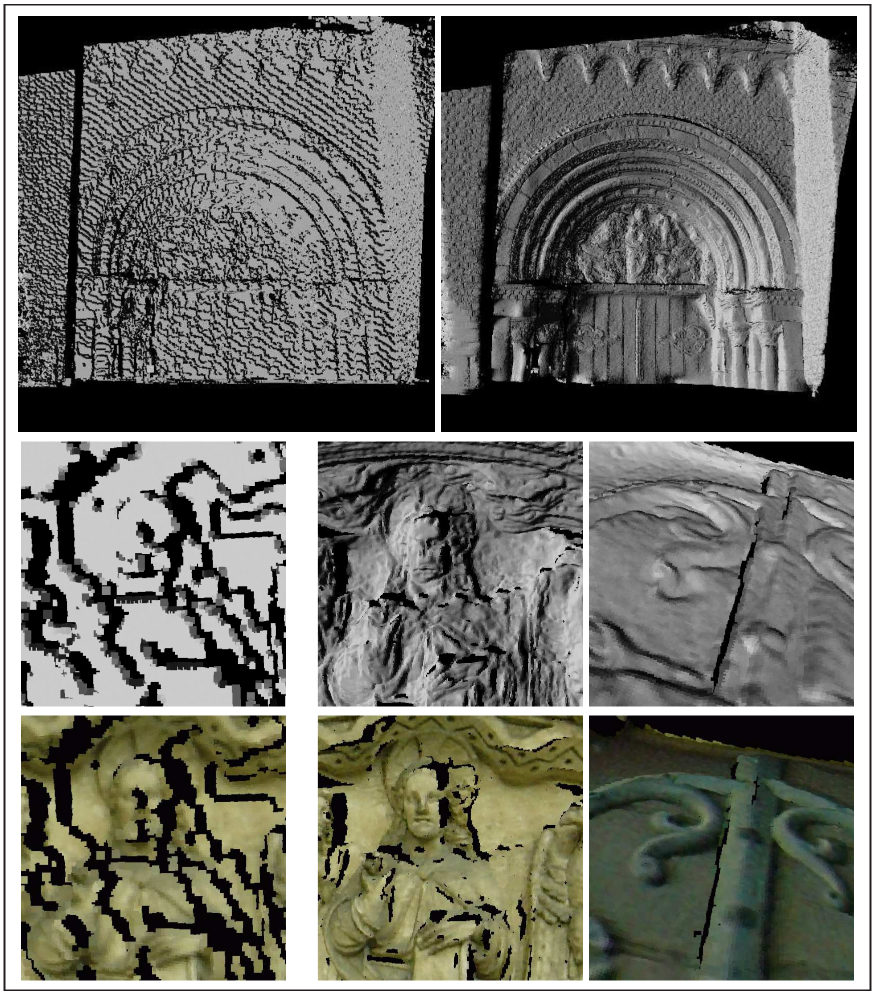


Figure 5.2: **Church scene:** *Textured and un-textured rendering of the raw depth maps (as shown in fig. 5.1. The top row shows the complete model for the global (left) and the local (right) approach. The two bottom rows show zoomed rendering of these models, for the global approach (two left images) and the local approach (four right images)*

tunable parameter, which accounts for the uncertainty in the depth prior, is decreased to obtain a more smooth, visual appealing reconstruction.

In fig. 5.2 we show textured and un-textured rendering for the global initialisation as well as for the local refinement. The discretisation of the depth values in the global formulation is visible in the three left images of this figure. The reconstruction at full resolution is shown in the top/ right image. The un-textured as well as the corresponding textured renderings of zoomed details are visible in the four right figures in the middle and bottom row. The computation time was 198 seconds and 159 seconds for the global and local approach, respectively.

Leuven city hall scene

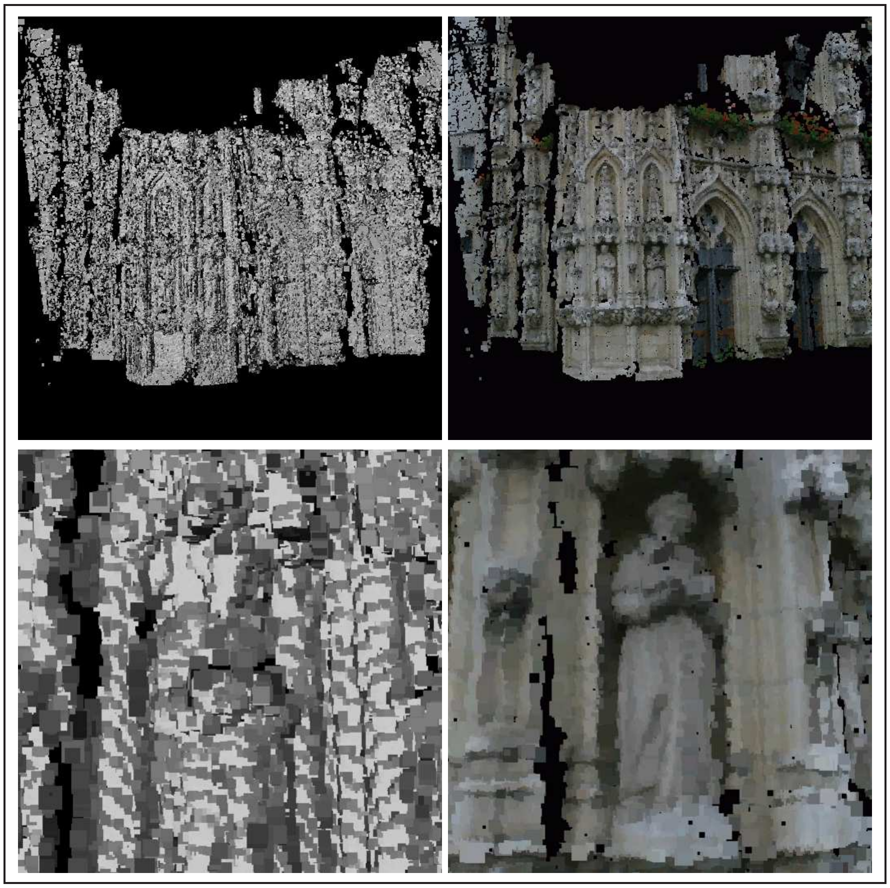


Figure 5.3: **Leuven cityhall scene:** *Textured and un-textured renderings from the images shown in fig. 3.24. The full model and a small detail of the global initialisation.*

In this experiment the same images as in figs.(3.24, 4.12) are used. The renderings for the global initialisation are computed with the depth map in fig. 3.24 for all pixels which are visible in at least to images (visibility maps in fig. 3.24). These are shown in fig. 5.3, where the full model (top) and a detail (bottom) is displayed. The images are processed up to 768×512 square pixels for which 469 seconds are needed to evaluate 268 depth states. Given this solution the local approach took 264 seconds to estimate

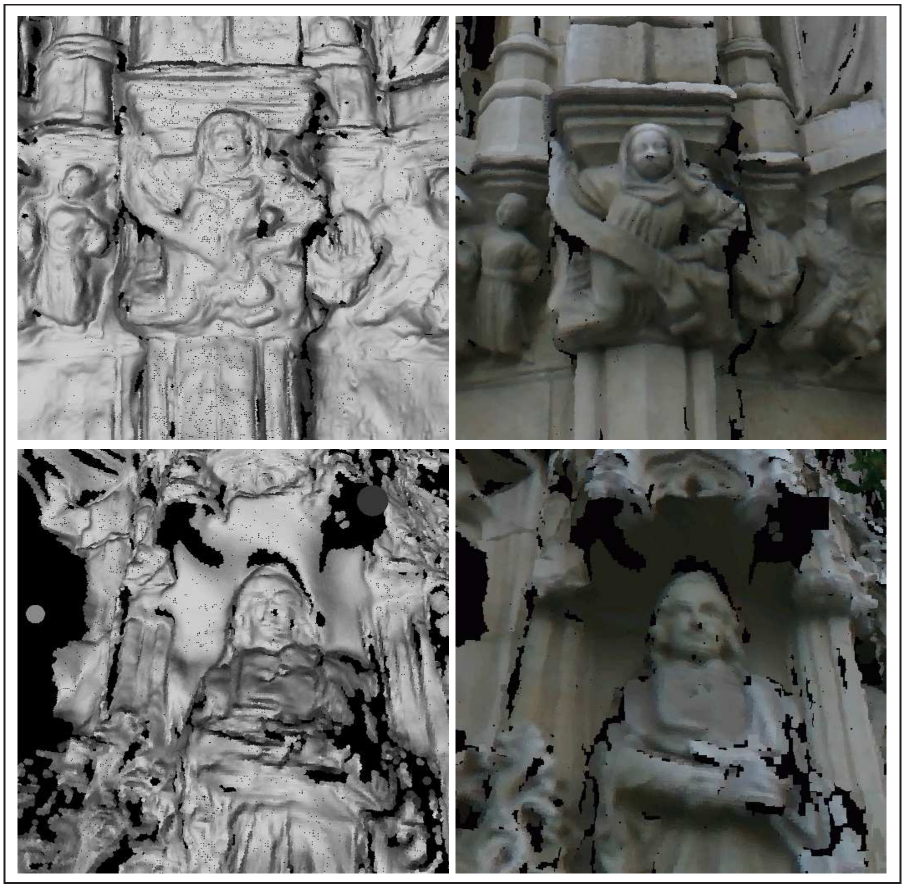


Figure 5.4: **Leuven city hall scene:** *Details of textured and un-textured renderings from the images shown in fig. 4.12.*

depth and visibility up to the full resolution of 3072×2048 square pixels. Figs. 5.5 and 5.5 shows the rendering of the depth map.

For the global approach one can recognise the discretised depth levels in the 3-D reconstruction (best shown in the un-textured zoomed rendering bottom/left of fig. 5.3), although the number of 268 depth states is large[1]. The result of the local approach displays a smooth reconstruction which displays fine details not present in the global reconstruction. Again λ, was set to $\lambda = 0.01$ and the other parameters are identical to fig. 4.12.

---

[1] Note, that the famous Tsukuba sequence [98] has a ground truth of 8 depth states.

Semper statue scene

Similar to the Leuven city hall scene we used the solution of the global approach (as in fig. 3.23) for the renderings in the top row of fig. 5.6. Up to 768×512 square pixels the global approach took 404 seconds for 240 depth states. Given this result the local approach took 454 seconds to estimate depth and visibility up to the full resolution of 3072×2048 square pixels. The renderings of the depth map are shown at the bottom in fig. 5.6 and the parameters are identical to the last experiment.

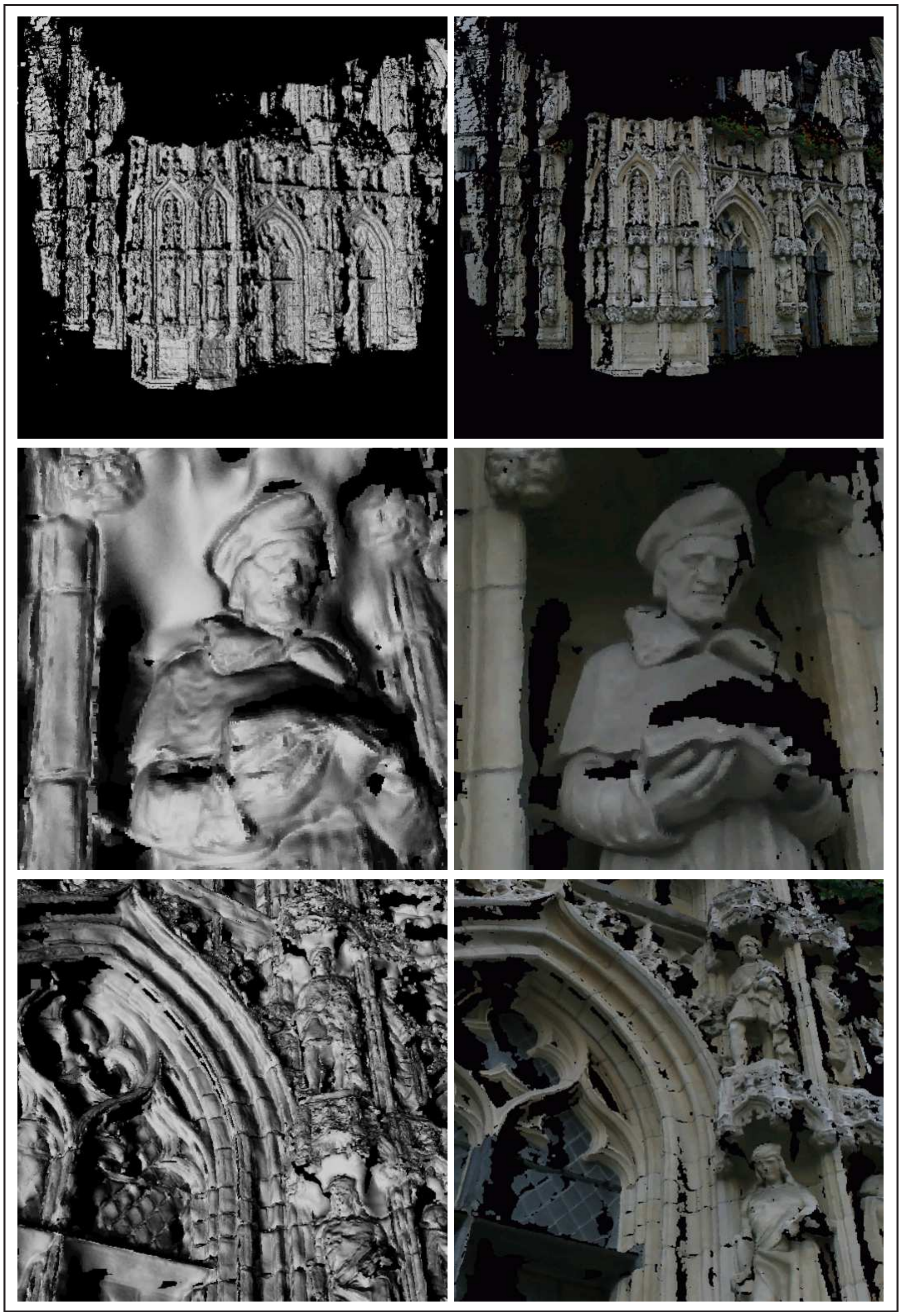


Figure 5.5: **Leuven city hall scene:** *Textured and un-textured renderings from the images shown in fig. 4.12. The full model (top) and small details of the local refinement.*

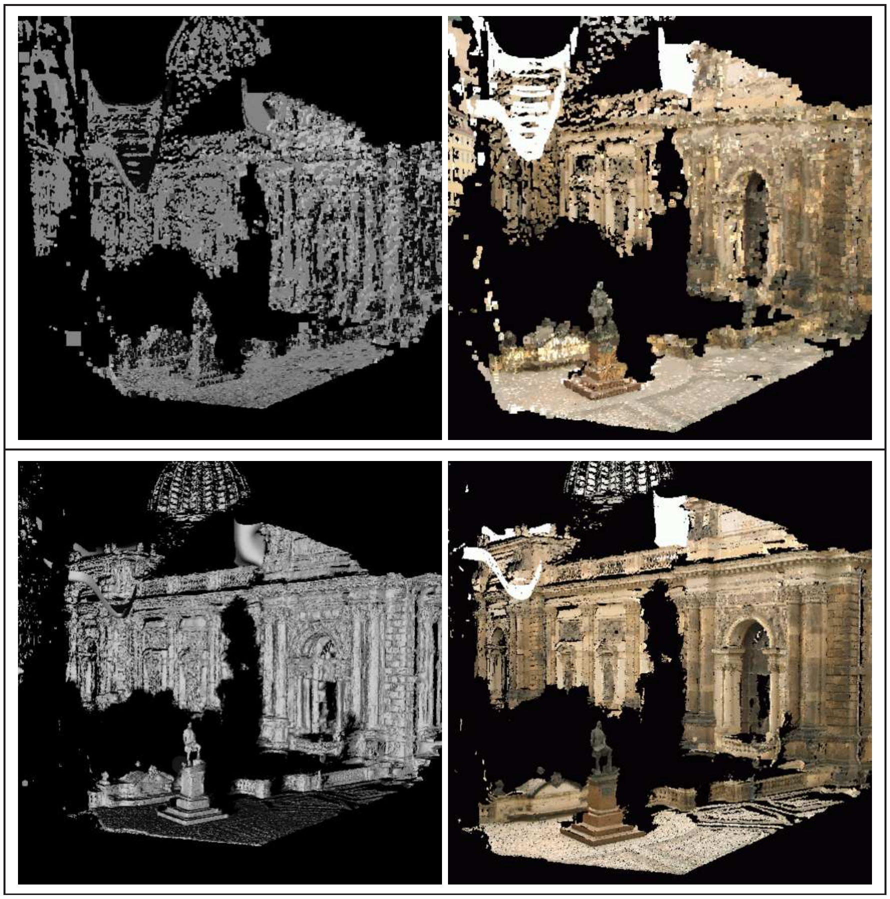


Figure 5.6: **Semper statue scene:** *Textured and un-textured renderings of the experiment shown in fig. 3.23: global initialisation in the top row and the local refinement in the bottom row.*

5.2 Virtual camera reconstructions

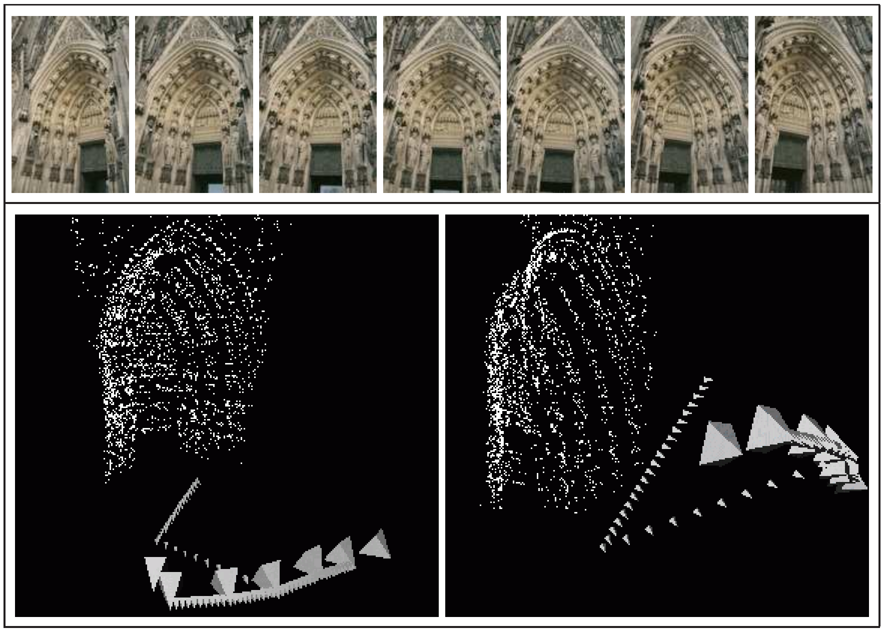


Figure 5.7: **Cologne cathedral scene:** *The seven input images are shown in the top row. Initial* 3*-D points and camera positions seen from two view points are rendered in the bottom row.*

Often we are not interested in the 3-D model of the scene, but only want to view the scene from a new, virtual view point. One possible solution is to render the computed 3-D model from this virtual camera position. This has been done in the previous section in the renderings of fig. 5.2, 5.5, 5.6. By using this approach it is not possible to assign a colour value to all pixels. These are the pixels, for which the ray from the camera centre through the pixel does not intersect the 3-D model. As a result one could see holes in the 3-D model as black pixels. Even if the depth maps from all cameras are integrated, which is itself a non-trivial problem, there might still be pixels with an undefined colour value.

A better solution is obtained by the generative model based approach as given in chapter 3 and 4. Remember, we solved the multi-view stereo problem by computing the most likely image $\mathbf{y}^*$ that would have been observed from a given camera position, given all input images. This problem is independent on the camera position and could also be applied to a position not included in the set of input cameras.

The advantage of this approach has several aspects. First of all, we only compute what is actually needed. Secondly, difficult areas for computing depth become trivial. For instance in areas of uniform texture (as a uniform sky) it is difficult to compute the depth because all possible depth values will project to pixels which have the same colour. Obviously, at these position, the ideal image y_i^* will have this colour.

Two experiments are performed. The first uses seven images of the Cathedral in cologne as shown in the top row of fig. 5.7. The two images at the bottom of this figure show two renderings of the initial 3-D points as well as the position of the input cameras (large pyramids). The position of the virtual cameras is indicated by small pyramids. Note that some of these virtual camera positions are far away from the set of input cameras. The ideal image $\mathbf{y}^*$ as computed by the local approach is shown in fig. 5.8 for some of the virtual cameras.

In the second experiment five input images are used which also contain accidental objects. These are shown in the left column of fig. 5.9. One can see a bus (top image), a car (bottom image) and pedestrians. In the reconstructions of the virtual camera positions (shown in the two columns middle and right) these objects disappear, *i.e*. they have no support in the majority of images and are removed as outliers.

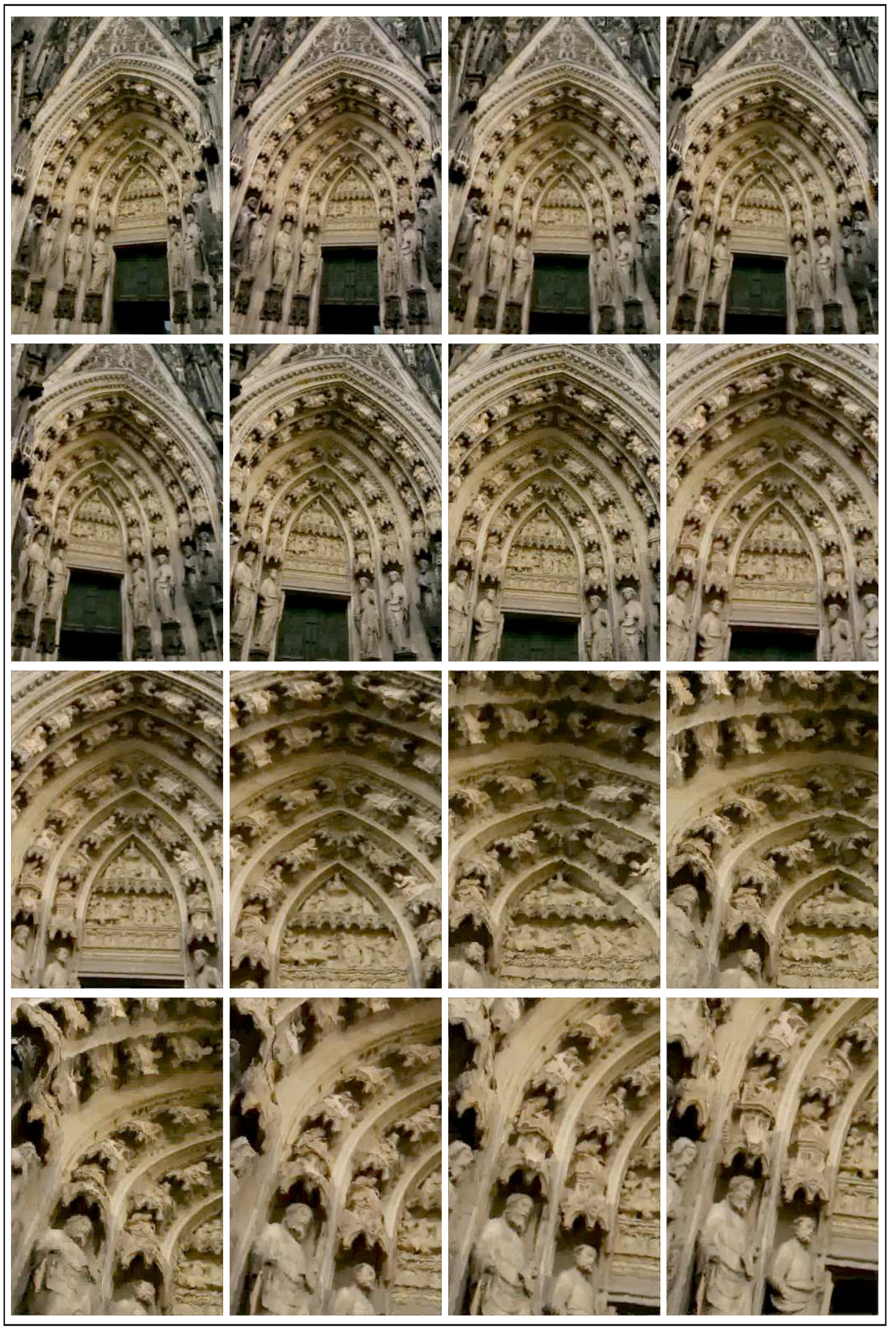


Figure 5.8: **Cologne cathedral scene:** *The ideal image* $\mathbf{y}^*$ *for some virtual camera positions which are shown as small pyramids in fig. 5.7.*

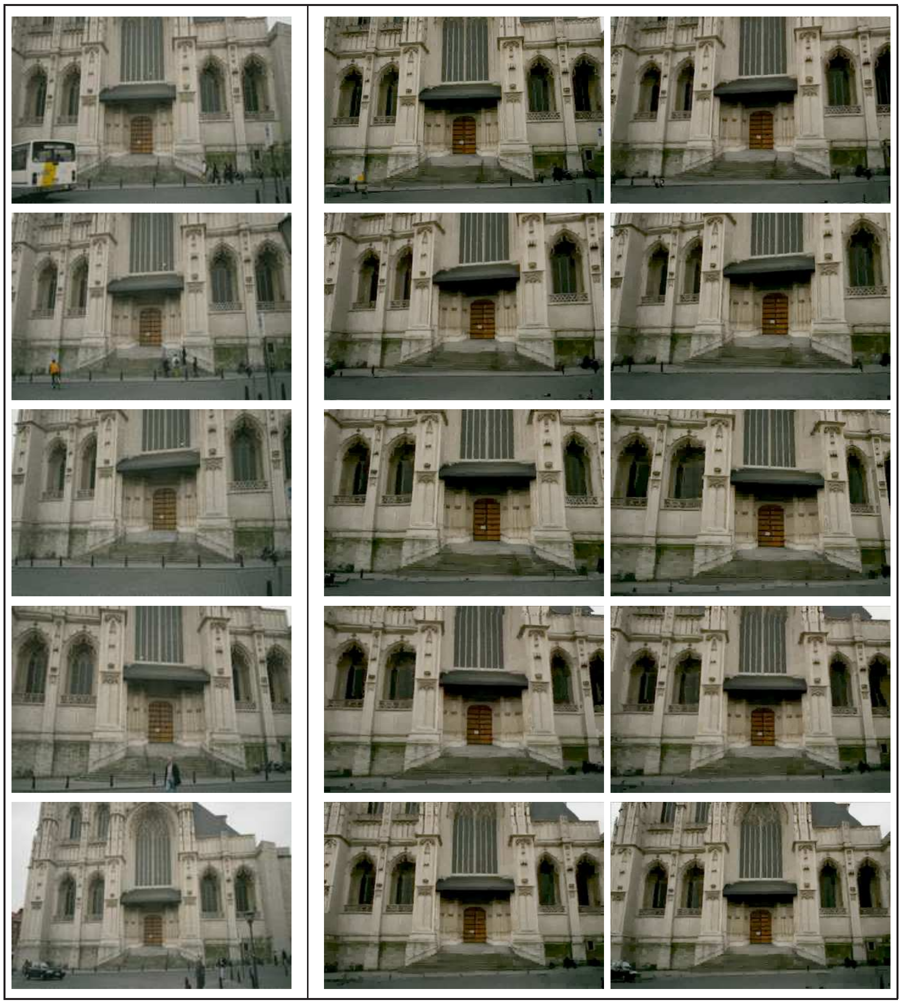


Figure 5.9: **Leuven church scene:** *The five input images are shown left. These images are contaminated by a bus, a car and several pedestrians. The images in the middle and right column show the computed ideal image* $\mathbf{y}^*$ *using the local approach computed for ten different virtual camera positions.*

Chapter 6

Conclusions

> *If you don't belief that your model (e.g. of normal errors) is correct, choose another one and use maximum likelihood - or Bayesian - methods for the new model. What, if I don't belief in the new model either? It takes a lot of stubbornness to flood the world with a host of rather arbitrary and probably hardly interpretable models and claim they are exactly true. The point of robust statistics is that one may keep a parametric model although the latter is known to be wrong.*
>
> $\mathbf{y}^*$, *i.e.* Hampel *et al.* [42], p. 403

6.1 Summary

In this thesis we used a generative model based approach to solve the multi-view stereo problem. In relation to the above quote by Hampel *et al.* [42], we showed that our particular model can be reinterpreted in the context of robust statistics. Moreover, we could derive a robust M-estimator, which corresponds to a simplified version of our particular generative model. This means that, 'maximum likelihood - or Bayesian - methods' (Hampel *et al.* [42]) can also be robust if the generative model explicitely takes outliers into account. If this is done, additional prior knowledge can be used to further enhance the performance and to be robust to outliers at the same time.

The main part of this thesis was on the evaluation of two generative models for the multi-view stereo problem. These models gave rise to a global formulation in which possible depth and visibility configurations of the scene are modeled as states of a Markov random field. A second, local formulation, takes an initial depth estimate and evolves it such that the input images are brought into correspondence. The results of the global formulation show that a good solution is estimated even for scenes with many outliers and depth discontinuities. This solution is obtained without depth and visibility initialisation. We showed for example further, that depth estimation is even possible w.r.t. a reference camera which is contaminated with outliers (fig.3.22). However, the global formulation cannot be applied to large sized images, since the number of depth states is too large to fit the memory and time constraints of current computers. In this case, the local formulation has clear advantages. Depth is not assumed to be discretised and a very accurate depth estimate can be computed. The price, for this is the need for a rough depth initialisation. Once this is given, the local formulation is able to deal with large images, fast and memory efficient. The combination of both, global and local, leads to accurate depth estimates for input images that could possibly be larger than 6 mega pixel. Multi-view stereo in this domain is often not feasible in other formulations.

Another focus of this thesis was on the parameter dependence of multi-view stereo. We showed that the proposed multi-view stereo formulation is applicable to a wide range of scenes. To make this possible we formulated the problem as an inverse inference problem, for which those model parameters are estimated, that have generated the input images. More particular, the width of the inlier distribution (noise) and the outlier distribution are estimated. As a result we obtain a formulation which is invariant to image noise variations and which decides automatically when a particular pixel is marked as outlier. We showed further that the remaining tunable parameters are related to the uncertainty in the prior distribution. For the global formulation this is reflected by σ_d and σ_v in eq. (3.6), which does fix the strength of the depth and the visibility correlations. In the local formulation we have λ, *i.e*. the width of the prior distribution. If training data would be available these prior distributions could be estimated and the formulation is completely parameter free.

Our generative model based multi-view stereo formulation is not restricted to a depth and visibility reconstruction w.r.t. a camera which is included in the set of input images. We have shown, that the formulation can also be applied to virtual camera positions, thereby estimating the most likely image that would have been observed from a virtual viewpoint, given the set of input images. In fact, it is possible to compute reconstructions w.r.t. non-perspective camera models, *e.g*. if a orthographic camera model is used, the ideal image $\mathbf{y}$ corresponds to the most likely ortho image.

6.2 Suggestions for further research

The study of the multi-view stereo problem in a Bayesian framework, as we did here, brought much insight in the problem. We are now in the position to study more advanced priors for the depth and visibility estimation. This will probably lead to a step forward in the quality of the reconstructions. One possibility to more advanced priors has already been proposed in the context of optical flow estimation. Roth *et al*. [94] studied the distribution of optical flow fields on labeled ground truth data and used it to build optical flow priors. This first step in a probably growing direction showed already a significant improvement. The question of more advanced prior information is also strongly coupled to the interpretation of the scene. The final goal in multi-view stereo reseach is not only the estimation of accurate 3-D models but also the interpretation, understanding and simplifycation of these models.

A further very interesting question is the performance evaluation of multi-view stereo compared to laser scan data. Laser scan systems are currently used to measure large outdoor scenes in 3-D. These system are very expensive and the measurement process is very time consuming. It would be interesting to study the possibility of using high resolution digital cameras to obtain three-dimensional outdoor models. First experiments in this direction [108] show that the spatial resolution of high resolution cameras is comparable with the resolution of laser scan systems. Another question that we did not touch in this thesis is the accuracy of the camera calibration. This is not only important for the comparison with laser scan systems, but it is also interesting to study the performance of our algorithm with respect to deviations from the true camera calibration.

We did not pay much attention to the efficiency of our implementation. Our main goal was to test the feasibility of the approach. However, for many applications, processing speed is an important issue. The possibilities are there to address these. First of all the processing on GPU will bring much profit. For the implementation of the PDE-based depth estimation in chapter 4 multi-grid implementations have shown to perform order of magnitudes faster [14].

Bibliography

[1] Akbarzadeh, A., Frahm, J., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Merrell, P., Phelps, M., Sinha, S., Talton, B., Wang, L., Yang, Q., Stewenius, H., Yang, R., Welch, G., Towles, H., Nister, D., and Pollefeys, M. Towards urban 3D reconstruction from video. In Int. Symp. of 3D Data Processing Visualization and Transmission (2006), pp. 1–8.

[2] Alvarez, L., Deriche, R., Papadopoulo, T., and Sánchez, J. Symmetrical dense optical flow estimation with occlusions detection. In Proc. European Conf. on Computer Vision (2002), vol. 1, pp. 721–735.

[3] Alvarez, L., Deriche, R., Sánchez, J., and Weickert, J. Dense disparity map estimation respecting image derivatives: A PDE and scale-space based approach. *Journal of Visual Communication and Image Representation 13*, 1/2 (2002), 3–21.

[4] Alvarez, L., Weickert, J., and Sánchez, J. Reliable estimation of dense optical flow fields with large displacements. Int'l Journal of Computer Vision *39*, 1 (2000), 41–56.

[5] Bay, H., Tuytelaars, T., and Van Gool, L. SURF: Speeded up robust features. In Proc. European Conf. on Computer Vision (2006), pp. 404–417.

[6] Beal, M. J., and Ghahramani, Z. The variational bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *In Bayesian Statistics 7* (2002).

[7] Bethe, H. Statistical theory of superlattices. *Proc. Roy. Soc. London Ser A 150* (1935), 552–575.

[8] Black, M., and Anandan, P. A framework for the robust estimation of optical flow. In Proc. Int'l Conf. on Computer Vision (1993), pp. 231–236.

[9] Black, M. J., Sapiro, G., Marimont, D. H., and Heeger, D. Robust anisotropic diffusion. *IEEE Trans. on Image Processing 7*, 3 (1998), 421–432.

[10] Blake, A., and Zisserman, A. *Visual Reconstruction*. MIT Press, Cambridge, MA., 1987.

[11] BOYKOV, Y., AND KOLMOGOROV, V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence *26*, 9 (2004), 1124–1137.

[12] BROX, T., BRUHN, A., PAPENBERG, N., AND WEICKERT, J. High accuracy optical flow estimation based on a theory for warping. In Proc. European Conf. on Computer Vision (2004), vol. 4, pp. 25–36.

[13] BROX, T., AND WEICKERT, J. Nonlinear matrix diffusion for optic flow estimation. In *Pattern Recognition* (2002), L. V. Gool, Ed., vol. 2449, Springer, Berlin, pp. 446–453.

[14] BRUHN, A., WEICKERT, J., KOHLBERGER, T., AND SCHNÖRR, C. A multigrid platform for real-time motion computation with discontinuity-preserving variational methods. Int'l Journal of Computer Vision *70*, 3 (2006), 257–277.

[15] BÜLTHOFF, I., BÜLTHOFF, H., AND SINHA, P. Top-down influences on stereoscopic depth-perception. *Nature Neuroscience 1 (3)* (1998), 254–257.

[16] COMANICIU, D., AND MEER, P. Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence *24*, 5 (2002), 603–619.

[17] DE SMET, M., FRANSENS, R., AND VAN GOOL, L. A generalized EM approach for 3-D model based face recognition under occlusions. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2006), vol. 2, pp. 1423–1430.

[18] DELLAERT, F. The expectation maximization algorithm. *Technical Report number GIT-GVU-02-20* (2002).

[19] DEMPSTER, A., LAIRD, N., AND RUBIN.D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B 39* (1977), 1–38.

[20] DERICHE, R., BOUVIN, C., AND FAUGERAS, O. A level-set approach for stereo. In *First Annual Symposium on Enabling Technologies for Law Enforcement and Security - SPIE Conference 2942 : Investigative Image Processing.* (Boston, Massachusetts USA, 1996).

[21] DERICHE, R., BOUVIN, C., AND FAUGERAS, O. Front propagation and level-set approach for geodesic active stereovision. In *Third Asian Conference On Computer Vision* (Bombay, India, 1998).

[22] DEVERNAY, F., AND FAUGERAS, O. Computing differential properties of 3-D shapes from stereoscopic images without 3-D models. In *ICVPR* (1994), pp. 208–213.

[23] DUAN, Y., YANG, L., QIN, H., AND SAMARAS, D. Shape reconstruction from 3-D and 2-D data using PDE-based deformable surfaces. In Proc. European Conf. on Computer Vision (2004), vol. 3, pp. 238–251.

[24] FAUGERAS, O., AND KERIVEN, R. Complete dense stereovision using level set methods. In Proc. European Conf. on Computer Vision (1998), vol. 1, pp. 379–393.

[25] FAUGERAS, O., AND KERIVEN, R. Variational principles, surface evolution, PDE's, level set methods and the stereo problem. *IEEE Trans. Image Processing 7*, 3 (1998), 336–344.

[26] FELZENSZWALB, P., AND HUTTENLOCHER, D. Efficient belief propagation for early vision. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2004), vol. 1, IEEE Computer Society, pp. 261–268.

[27] FÖRSTNER, W. Uncertainty and projective geometry. In *Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics*, E. Bayro-Corrochano, Ed. Springer, 2004.

[28] FRANSENS, R., STRECHA, C., AND VAN GOOL, L. Multimodal and multi-band image registration using mutual information. *ESA-EUSC 2004: Theory and Applications of Knowledge driven Image Information Mining, with focus on Earth Observation; EUSC, Madrid (Spain); March 17-18* (2004).

[29] FRANSENS, R., STRECHA, C., AND VAN GOOL, L. A probabilistic approach to optical flow based super-resolution. *Workshop GMBV in conjunction with CVPR 2004* (2004).

[30] FRANSENS, R., STRECHA, C., AND VAN GOOL, L. Parametric stereo for multi-pose face recognition and 3D-face modeling. *ICCV workshop Analysis and Modeling of Faces and Gestures*, Lecture Notes in Computer Science *3723* (2005), 109–124.

[31] FRANSENS, R., STRECHA, C., AND VAN GOOL, L. A mean field EM-algorithm for coherent occlusion handling in map-estimation problems. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (Washington, DC, USA, 2006), IEEE Computer Society, pp. 300–307.

[32] FRANSENS, R., STRECHA, C., AND VAN GOOL, L. Robust estimation in the presence of spatially coherent outliers. In *RANSAC workshop at CVPR* (2006).

[33] FRANSENS, R., STRECHA, C. CAENEN, G., AND VAN GOOL, L. A generic approach towards automatic image co-registration. *Second International Workshop ISPRS, The Future of Remote Sensing, Organised by VITO and ISPRS Inter-Commission Working Group I/V Autonomous Navigation, October 17-18, Antwerp (Belgium)* (2006).

[34] FREY, B., LAWRENCE, N., AND BISHOP, C. Markovian inference in belief networks. *presented at Machines That Learn* (1998).

[35] FUA, P. A parallel stereo algorithm that produces dense depth maps and preserves image features. *Machine Vision and Applications 6*, 1 (1993), 35–49.

[36] FURUKAWA, Y., AND PONCE, J. Carved visual hulls for image-based modeling. Proc. European Conf. on Computer Vision (2006), 564–577.

[37] GARGALLO, P., AND STURM, P. Bayesian 3D modeling from images using multiple depth maps. Proc. Int'l Conf. on Computer Vision and Pattern Recognition *2* (2005), 885–891.

[38] GEIGER, D., LADENDORF, B., AND YUILLE, A. Occlusions and binocular stereo. Int'l Journal of Computer Vision *14*, 3 (1995), 211–226.

[39] GELMAN, A., CARLIN, J., STERN, H., AND RUBIN, D. *Bayesian Data Analysis*, second ed. Chapman & Hall/CRC, 2004.

[40] GEMAN, S., AND GEMAN, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence *6*, 6 (1984), 721–741.

[41] GOESELE, M., CURLESS, B., AND SEITZ, S. Multi-view stereo revisited. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (Washington, DC, USA, 2006), IEEE Computer Society, pp. 2402–2409.

[42] HAMPEL, F., RONCHETTI, E., AND ROUSSEEUW, P. *Robust statistics : the approach based on influence functions*. New York (N.Y.): Wiley, 1986.

[43] HARTLEY, R., AND ZISSERMAN, A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[44] HERNANDEZ, C., AND SCHMITT, F. Silhouette and stereo fusion for 3D object modeling. *Computer Vision and Image Understanding 96*, 3 (December 2004), 367–392.

[45] HESKES, T. Stable fixed points of loopy belief propagation are minima of the bethe free energy. *Advances in neural information processing systems 15* (2002), 359–366.

[46] HESKES, T., AND ZOETER, O. Generalized belief propagation for approximate inference in hybrid bayesian networks. *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, C. M. Bishop and B. J. Frey, Eds.* (2003).

[47] HORNUNG, A., AND KOBBELT.L. Hierarchical volumetric multi-view stereo reconstruction of manifold surfaces based on dual graph embedding. Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2006), 503–510.

[48] HUBER, P. *Robust statistics*. John Wiley, 1982.

[49] JÄHNE, B. *Spatio-Temporal Image Processing: Theory and Scientific Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1993.

[50] JIAN, S., NAN-NING, Z., AND HEUNG-YEUNG, S. Stereo matching using belief propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence *25*, 7 (2003), 787–800.

[51] JIAN, S., YIN, L., AND SING BING, K. Symmetric stereo matching for occlusion handling. Proc. Int'l Conf. on Computer Vision and Pattern Recognition *2* (2005), 399–406.

[52] JIN, H., SOATTO, S., AND YEZZI, A. Multi-view stereo beyond Lambert. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2003), pp. 171–178.

[53] JIN, H., SOATTO, S., AND YEZZI, A. Multi-view stereo reconstruction of dense shape and complex appearance. Int'l Journal of Computer Vision *63*, 3 (2005), 175–189.

[54] JIN, H., YEZZI, A., AND SOATTO, S. Variational multiframe stereo in the presence of specular reflections. Int. Symp. of 3D Data Processing Visualization and Transmission (2002), 626–630.

[55] JORDAN, M., GHAHRAMANI, Z., JAAKKOLA, T., AND SAUL, L. An introduction to variational methods for graphical models. *Machine Learning 37*, 2 (1999), 183–233.

[56] KANG, S., AND SZELISKI, R. Extracting view-dependent depth maps from a collection of images. Int'l Journal of Computer Vision *58*, 2 (2004), 139–163.

[57] KANG, S., SZELISKI, R., AND CHAI, J. Handling occlusions in dense multi-view stereo. *Technical Report: MSR-TR-2001-80* (2001).

[58] KANG, S., SZELISKI, R., AND CHAI, J. Handling occlusions in dense multi-view stereo. Proc. Int'l Conf. on Computer Vision and Pattern Recognition *1* (2001), 103–110.

[59] KAPPEN, H., AND WIEGERINCK, W. Mean field theory for graphical models. *In M. Opper and D. Saad, editors, Adavanced Mean Field Theory – Theory and Practice, MIT Press 4* (2001), 37–49.

[60] KERSTEN, D., KNILL, D., MAMASSIAN, P., AND BUELTHOFF, I. Illusory motion from shadows. *Nature 379*, 6560 (January 1996), 31–31.

[61] KIKUSHI, R. A theory of cooperative phenomena. *Phys. Rev. 81* (1951), 998–1005.

[62] KIRKPATRICK, S. GELATT, C., AND VECCHI, M. Optimization by simulated annealing. *Science 220*, 4598 (1983), 671–680.

[63] KOLMOGOROV, V., AND ZABIH, R. Computing visual correspondence with occlusions using graph cuts. In Proc. Int'l Conf. on Computer Vision (2001), vol. 02, p. 508.

[64] KOLMOGOROV, V., AND ZABIH, R. Multi-camera scene reconstruction via graph cuts. In Proc. European Conf. on Computer Vision (2002), vol. 3, pp. 82–96.

[65] KONG, D., AND TAO, H. Stereo matching via learning multiple experts behaviour. In Proc. British Machine Vision Conf. (2006), vol. 1, pp. 97–106.

[66] KUTULAKOS, K., AND SEITZ, S. A theory of shape by space carving. Int'l Journal of Computer Vision *38(3)* (2000), 197–216.

[67] LABATUT, P., KERIVEN, R., AND PONS, J.-P. A GPU implementation of level set multiview stereo. In *International Conference on Computational Science (4)* (2006), pp. 212–219.

[68] LHUILLIER.M. A quasi-dense approach to surface reconstruction from uncalibrated images. IEEE Transactions on Pattern Analysis and Machine Intelligence *27*, 3 (2005), 418–433.

[69] LI, G., AND ZUCKER, T. Surface geometric constraints for stereo in belief propagation. Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2006), 2355–2362.

[70] LOWE, D. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision *60*, 2 (2004), 91–110.

[71] MACKAY, D., YEDIDIA, J., FREEMAN, W., AND WEISS, Y. A conversation about the Bethe free energy and sum-product. *Merl technical report TR2001-18* (2001).

[72] MATAS, J., CHUM, O., URBAN, M., AND PAJDLA, T. Robust wide baseline stereo for maximally stable external regions. *Proc. BMVC* (2002), 414–431.

[73] MINKA, T. Expectation-maximization as lower bound maximization. *Tutorial published on the web* (1998).

[74] MUMFORD, D., AND SHAH, J. Optimal approximations by piece-wise smooth functions and associated variational problems. *Commun. Pure Appl. Math. 42* (1989), 577–685.

[75] NEAL, R. M., AND HINTON, G. E. *A view of the EM algorithm that justifies incremental, sparse, and other variants*. MIT Press, Cambridge, MA, USA, 1999, pp. 355–368.

[76] NEUMANN, J., AND ALOIMONOS, Y. Spatio-temporal stereo using multiresolution subdivision surfaces. Int'l Journal of Computer Vision *47*, 1-3 (2002), 181–193.

[77] NIST'ER, D. *Automatic dense reconstruction from uncalibrated video sequences*. PhD thesis, Royal Institute of Technology KTH, Stockholm, Sweden, ISBN 91-7283-053-0, 2001.

[78] OSHER, S., AND SETHIAN, J. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics 79* (1988), 12–49.

[79] ÖZDEN, K., CORNELIS, K., VAN EYCKEN, L., AND VAN GOOL, L. Reconstructing 3D independent motions using non-accidentalness. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2004), pp. 819–825.

[80] PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[81] PERONA, P., AND MALIK, J. Scale-space and edge detection using anisotropic diffusion. IEEE Transactions on Pattern Analysis and Machine Intelligence *12*, 7 (1990), 629–639.

[82] POLLEFEYS, M., KOCH, R., AND VAN GOOL, L. Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters. Int'l Journal of Computer Vision *32*, 1 (1999), 7–25.

[83] POLLEFEYS, M., KOCH, R., AND VAN GOOL, L. A simple and efficient rectification method for general motion. Proc. Int'l Conf. on Computer Vision (1999), 496–501.

[84] POLLEFEYS, M., VAN GOOL, L., VERGAUWEN, M., VERBIEST, F., CORNELIS, K., TOPS, J., AND KOCH, R. Visual modeling with a hand-held camera. Int'l Journal of Computer Vision *59*, 3 (2004), 207–232.

[85] PONS, J., KERIVEN, R., FAUGERAS, O., AND HERMOSILLO, G. Variational stereovision and 3D scene flow estimation with statistical similarity measures. In Proc. Int'l Conf. on Computer Vision (Washington, DC, USA, 2003), IEEE Computer Society, p. 597.

[86] PONS, J.-P., KERIVEN, R., AND FAUGERAS, O. Modelling dynamic scenes by registering multi-view image sequences. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2005), vol. 2, pp. 822–827.

[87] PONS, J.-P., KERIVEN, R., AND FAUGERAS, O. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. Int'l Journal of Computer Vision (To appear).

[88] PRESS, W., FLAAERY, B., TEUKOLSKY, S., AND VETTERLING, W. *Numerical Recipes*. Cambridge University Press, 1986.

[89] PROESMANS, M. *Non-linear Diffusion for Low-Level Vision*. PhD thesis, ESAT/Psi University of Leuven, 1998.

[90] PROESMANS, M., VAN GOOL, L., PAUWELS, E., AND OOSTERLINCK, A. Determination of optical flow and its discontinuities using non-linear diffusion. Proc. European Conf. on Computer Vision *2* (1994), 295–304.

[91] RANGARAJAN, A., CHELLAPPA, R., AND MANJUNATH, B. *Markov Random Fields and Neural Networks with applications to Early Vision Problems*. Artificial Neural Networks and Statistical Pattern Recognition (ed. I. Sethi), 1991.

[92] RIPLEY, S. Pattern recognition and neural networks. *Cambridge University Press* (1996).

[93] ROBERT, L., AND DERICHE, R. Dense depth map reconstruction: A minimization and regularization approach which preserves discontinuities. In Proc. European Conf. on Computer Vision (1996), pp. 439–451.

[94] ROTH, S., AND BLACK, M. On the spatial statistics of optical flow. Proc. Int'l Conf. on Computer Vision *1* (2005), 42–49.

[95] ROUSSEEUW, P. J. Least median of squares regression. *Journal of the Am. Stat. Assoc. 79* (1984), 871–880.

[96] ROY, S., AND COX, I. A maximum-flow formulation of the n-camera stereo correspondence problem. Proc. Int'l Conf. on Computer Vision (1998), 492–502.

[97] RUSINKIEWICZ, S., AND LEVOY, M. Qsplat: a multiresolution point rendering system for large meshes. In *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 2000), ACM Press/Addison-Wesley Publishing Co., pp. 343–352.

[98] SCHARSTEIN, D., AND SZELISKI, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int'l Journal of Computer Vision *47*, 1/2/3 (2002), 7–42.

[99] SEITZ, S., CURLESS, B., DIEBEL, J., SCHARSTEIN, D., AND SZELISKI, R. A comparison and evaluation of multi-view stereo reconstruction algorithms. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (Washington, DC, USA, 2006), IEEE Computer Society, pp. 519–528.

[100] SLESAREVA, N., BRUHN, A., AND WEICKERT, J. Optic flow goes stereo: A variational method for estimating discontinuity-preserving dense disparity maps. In *DAGM05* (2005), p. 33.

[101] SOATTO, S., YEZZI, A., AND JIN, H. Tales of shape and radiance in multiview stereo. In Proc. Int'l Conf. on Computer Vision (Washington, DC, USA, 2003), IEEE Computer Society, p. 974.

[102] STRECHA, C., FRANSENS, R., AND VAN GOOL, L. A probabilistic approach to large displacement optical flow and occlusion detection. In *Workshop SMVP in conjunction with ECCV* (2004).

[103] STRECHA, C., FRANSENS, R., AND VAN GOOL, L. Wide-baseline stereo from multiple views: a probabilistic account. Proc. Int'l Conf. on Computer Vision and Pattern Recognition *1* (2004), 552–559.

[104] STRECHA, C., FRANSENS, R., AND VAN GOOL, L. Combined depth and outlier estimation in multi-view stereo. Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2006), 2394–2401.

[105] STRECHA, C., FRANSENS, R., AND VAN GOOL, L. A probabilistic formulation of image registration. *1st International Workshop on Complex Motion, IWCM Günzburg, Germany, october 2004, revised papers 3417* (2007), 165–176.

[106] STRECHA, C., TUYTELAARS, T., AND VAN GOOL, L. Dense matching of multiple wide-baseline views. In Proc. Int'l Conf. on Computer Vision (2003), pp. 1194–1201.

[107] STRECHA, C., AND VAN GOOL, L. Motion-stereo integration for depth estimation. In Proc. European Conf. on Computer Vision (2002), vol. 2, pp. 170–185.

[108] STRECHA, C., VAN HANSEN, W., VAN GOOL, L., AND THOENNESSEN, U. High resolution image based depth modelling. *to appear*.

[109] SZELISKI, R. A multi-view approach to motion and stereo. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (1999), pp. 157–163.

[110] TRAN, S., AND DAVIS, L. 3D surface reconstruction using graph cuts with surface constraints. In Proc. European Conf. on Computer Vision (2006), pp. 219–231.

[111] TUYTELAARS, T., AND VAN GOOL, L. Wide baseline stereo matching based on local, affinely invariant regions. *Proc. BMVC* (2000), 412–422.

[112] UEDA, N., AND NAKANO, R. Deterministic annealing EM algorithm. *Neural Networks 11* (1998), 271–282.

[113] VOGIATZIS, G., TORR, P., AND CIPOLLA, R. Multi-view stereo via volumetric graph-cuts. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (Washington, DC, USA, 2005), IEEE Computer Society, pp. 391–398.

[114] WAINWRIGHT, M., JAAKKOLA, T., AND WILLSKY, A. Tree-based reparameterization for approximate estimation on graphs with cycles. *LIDS Tech. report P-2510* (2001).

[115] WAINWRIGHT, M., AND JORDAN, M. I. Graphical models, exponential families, and variational inference. *Technical Report 649, Department of Statistics, University of California, Berkeley* (2003).

[116] WEICKERT, J. *Anisotropic Diffusion in Image Processing*. Teubner-Verlag, 1998.

[117] WEICKERT, J., AND BROX, T. Diffusion and regularization of vector- and matrix-valued images. *Inverse Problems, Image Analysis, and Medical Imaging. Contemporary Mathematics, AMS, Providence 313* (2002), 251–268.

[118] WEISS, Y. *Comparing the mean field method and belief propagation for approximate inference in MRFs*. Saad D. and Opper M. MIT Press, 2001.

[119] WEISS, Y., AND FLEET, D. Velocity likelihoods in biological and machine vision, 2001.

[120] WELLING, M., AND TEH, Y. W. Approximate inference in Boltzmann machines. *Artificial Intelligence 143*, 1 (2003), 19–50.

[121] YANG, Q., WANG, L., YANG, R., STEWÉNIUS, H., AND NISTÉR, D. Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling. In Proc. Int'l Conf. on Computer Vision and Pattern Recognition (2006), vol. 2, pp. 2347–2354.

[122] YEDIDIA, J., FREEMAN, W., AND WEISS, Y. Generalized belief propagation. *Neural Information Processing Systems (NIPS) 13* (2000), 689–695.

[123] YEDIDIA, J., FREEMAN, W., AND WEISS, Y. Bethe free energy, kikuchi approximations, and belief propagation algorithms. *MERL TR2001-016* (2001).

[124] YEDIDIA, J., FREEMAN, W., AND WEISS, Y. *Understanding belief propagation and its generalizations*. Morgan Kaufmann Publishers Inc., 2003, pp. 239–269.

[125] YEDIDIA, J., FREEMAN, W., AND WEISS, Y. Constructing free energy approximations and generalized belief propagation algorithms. *MERL TR2004-040* (2004).

[126] YUILLE, A. A double-loop algorithm to minimize the bethe free energy. *EMMCVPR* (2001), 3–18.

[127] YUILLE, A., AND KERSTEN, D. Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences 10*, 7 (July 2006), 301–308.

[128] YUILLE, A. L. Cccp algorithms to minimize the bethe and kikuchi free energies: convergent alternatives to belief propagation. *Neural Comput. 14*, 7 (2002), 1691–1722.

[129] YUILLE, A. L., GEIGER, D., AND BÜLTHOFF, H. H. Stereo integration, mean field theory and psychophysics. In Proc. European Conf. on Computer Vision (1990), pp. 73–82.

[130] ZHANG, J. The mean field theory in EM procedures for blind markov random field image restoration. *IEEE Trans. Signal Processing 40*, 10 (1992), 2570–2583.

[131] ZITNICK, C., AND KANADE, T. A cooperative algorithm for stereo matching and occlusion detection. IEEE Transactions on Pattern Analysis and Machine Intelligence *22*, 7 (2000), 675 – 684.

Appendix A

Depth parameterisation

Given the external (rotation R_n and translation t_n) as well as the internal calibration matrices (camera matrix K_n) for N views ($n = 1..N$) the relaxation of corresponding 2-D points $\vec{x}$ in the image plane is given by the depth. A 3-D point denoted by $\mathbf{X}$ is projected to the camera n by:

$$\lambda_n \vec{x}_n^h \quad = \quad \mathbf{K}_n[\mathbf{R}_n^T | - \mathbf{R}_n^T \mathbf{t}_n]\mathbf{X} \tag{A.1}$$

It follows for corresponding image points[1] $\vec{x}_1^h = (x_1, y_1, 1)^T$ and $\vec{x}_2^h = (x_2, y_2, 1)^T$ and for a *coordinate system that is attached to the first camera* ($\mathbf{R}_1 = \mathbf{1}$, $\mathbf{t}_1 = \mathbf{0}$) that:

$$\frac{\lambda_2}{\mathcal{D}_1(\vec{x}_1)} \vec{x}_2^h \quad = \quad \mathbf{K_2} \mathbf{R}_2^T \mathbf{K}_1^{-1} \vec{x}_1^h - \frac{1}{\mathcal{D}_1(\vec{x}_1)} \mathbf{K}_2 \, \mathbf{R}_2^T \mathbf{t}_2 \tag{A.2}$$

The stereo correspondence is divided into a component that depends on the rotation and pixel coordinate (according to the homography $\mathbf{H} = \mathbf{K}_2 \mathbf{R}_2^T \mathbf{K}_1^{-1}$) and a depth dependent part that scales with the amount of translation between the cameras. The corresponding point $\vec{x}_2$ on the epipolar line in a second image as a function of the depth $\mathcal{D}_1(\vec{x}_1)$ is given by:

$$\vec{x}_2 = \frac{\begin{pmatrix} \mathbf{H}[1]\vec{x}_1^h \\ \mathbf{H}[2]\vec{x}_1^h \end{pmatrix} + \frac{1}{\mathcal{D}_1(\vec{x}_1)} \begin{pmatrix} \mathbf{K}_2[1]\mathbf{R}_2^T \mathbf{t}_2 \\ \mathbf{K}_2[2]\mathbf{R}_2^T \mathbf{t}_2 \end{pmatrix}}{\mathbf{H}[3]\vec{x}_1^h - \frac{1}{\mathcal{D}_1(\vec{x}_1)} \mathbf{K}_2[3]\mathbf{R}_2^T \mathbf{t}_2} \tag{A.3}$$

$\mathbf{H}[i]$ is the 3-vector for the i^{th} row of the homography $\mathbf{H}$ and similarly for $\mathbf{K}_2[i]$. This equation leads to a parameterisation where for a given pixel $\vec{x}_i$ in image i, we can determine the corresponding points in all other images by knowing the depth $\mathcal{D}_i(\vec{x}_i)$ of that pixel.

---

[1] we will use in the following the superscript h to indicate homogenous coordinates

In the general case where the first camera is not attached to the global coordinate system, the correspondence between camera i and j is given by:

$$\vec{x}_j = \frac{\begin{pmatrix} \mathbf{H_{ij}}[1]\vec{x}_i^h \\ \mathbf{H_{ij}}[2]\vec{x}_i^h \end{pmatrix} + \frac{1}{\mathcal{D}_i(\vec{x}_i)} \begin{pmatrix} \mathbf{T}_{ij}[1] \\ \mathbf{T}_{ij}[2] \end{pmatrix}}{\mathbf{H_{ij}}[3]\vec{x}_i^h - \frac{1}{\mathcal{D}_i(\vec{x}_i)}\mathbf{T}_{ij}[3]} \ , \tag{A.4}$$

with $\mathbf{H_{ij}} = \mathbf{K}_j \mathbf{R}_j^T \mathbf{R}_i \mathbf{K}_i^{-1}$ and $\mathbf{T}_{ij} = \mathbf{K}_j \mathbf{R}_j^T (\mathbf{t}_i - \mathbf{t}_j)$.

Appendix B

EM algorithm

B.1 Classical formulation

Let $\boldsymbol{\theta}$ denote all unknowns and let $\mathbf{x}$ and $\mathbf{y}$ denote the Potts MRF and all input data, respectively. Our aim is to compute the maximum likelihood (ML) solution of the parameters $\boldsymbol{\theta}$, given by:

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{ML} &= \arg\max_{\boldsymbol{\theta}} \big\{ \log p(\mathbf{y} \,|\, \boldsymbol{\theta}) \big\} \\ &= \arg\max_{\boldsymbol{\theta}} \big\{ \log \sum_{\mathbf{x}} p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta}) \big\} \; . \end{aligned} \tag{B.1}$$

Notice that the sum $\sum_{\mathbf{x}}$ in equation (B.1) ranges over all possible configurations of the hidden variables $\mathbf{x}$. Even for modest sized images, this is a huge number, which makes direct optimisation of eq. (B.1) infeasible. The problem can be made tractable by using the expectation maximisation (EM) algorithm [19]. Starting from an initial guess $\{\widehat{\boldsymbol{\theta}}^{(0)}$, it produces a sequence of estimates $\{\widehat{\boldsymbol{\theta}}^{(t)}, t = 1, 2 \ldots\}$ by alternating the following two steps:

E-step	Compute the distribution $b^{(t)}$ over the range of $\mathbf{x}$ such that $b^{(t)}(\mathbf{x}) = p(\mathbf{x} \,|\, \mathbf{y}, \widehat{\boldsymbol{\theta}}^{(t-1)})$.
M-step	Set $\widehat{\boldsymbol{\theta}}^{(t+1)}$ to the $\boldsymbol{\theta}$ that maximises $\mathrm{E}_{b^{(t)}}[\log p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta})]$.

Here, $\mathrm{E}_b[.]$ denotes the expectation of the argument under $b(\mathbf{x})$. The M-step can thus be seen as a maximum likelihood estimation for which the value of $\mathbf{x}$ is known by its distribution $b(\mathbf{x})$. The key idea, and the way to prevent the computation of the large sum in eq. B.1, is to choose $b(\mathbf{x})$ close to the true distribution $p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})$ but at the same time less complex and hence, easier to compute. After making a specific choice for $b(\mathbf{x})$ deduced from different approximations, the Kullback-Leibler divergence

$$D_{KL}(p||b) = \sum_{\mathbf{x}} b(\mathbf{x}) \log \frac{b(\mathbf{x})}{p(\mathbf{x})} \tag{B.2}$$

between both distributions is minimised in the E-step.

B.2 Lower bound formulation

A more insightful explanation of EM is in terms of lower bound maximisation [75, 73, 18]. Thereby, the E-step can be interpreted as constructing a local lower bound on the posterior. The M-step optimises this bound with respect to the parameters $\boldsymbol{\theta}$. One can trivially rewrite the argument in eq. (B.1):

$$\log \sum_{\mathbf{x}} p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta}) = \log \sum_{\mathbf{x}} \mathbf{b}(\mathbf{x}) \frac{p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta})}{\mathbf{b}(\mathbf{x})} \;, \tag{B.3}$$

and use Jensen's inequality to construct the lower bound on the argument in eq. (B.1):

$$\log \sum_{\mathbf{x}} p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta}) \geq \sum_{\mathbf{x}} b(\mathbf{x}) \log \frac{p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta})}{b(\mathbf{x})} \;. \tag{B.4}$$

The EM algorithm is exact if the trial distribution $b(\mathbf{x})$ is not restricted to a specified class of distributions. Maximising the lower bound in eq. (B.4) with respect to $b(\mathbf{x})$ results in $b(\mathbf{x}) = p(\mathbf{x} \mid \mathbf{y}, \boldsymbol{\theta})$ which, when resubstituted, turns the inequality into an equality. The lower bound is tight and touches the objective function:

$$\begin{aligned}
\sum_{\mathbf{x}} b(\mathbf{x}) \log \frac{p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta})}{b(\mathbf{x})} \quad &= \quad \sum_{\mathbf{x}} p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}) \log \frac{p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta})}{p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta})} \\
&= \quad \sum_{\mathbf{x}} p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}) \log \frac{p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y} \,|\, \boldsymbol{\theta})}{p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta})} \\
&= \quad \sum_{\mathbf{x}} p(\mathbf{x} \,|\, \mathbf{y}, \boldsymbol{\theta}) \log p(\mathbf{y} \,|\, \boldsymbol{\theta}) \\
&= \quad \log p(\mathbf{y} \,|\, \boldsymbol{\theta}) \;. \qquad \text{(B.5)}
\end{aligned}$$

On the other hand, if the space of possible realisations of $b(\mathbf{x})$ is restricted, the bound will not be tight. This situation is actually applied to deal with the infeasibility of eq. (B.1).

The negative lower bound is equal to the Kullback-Leibler divergence. It is also related to the concept of free energy [75], $F(b(\mathbf{x}), \boldsymbol{\theta})$ of statistical physics (see eq. C.5). The terms variational free energy or Gibbs free energy are also used in the computer vision literature [124]. More details on this relation are given later in appendix C. Thus, maximising the lower bound is equivalent to minimising the variational free energy, which is the aim of the E-step.

The M-step is archived by setting the derivative of the variational free energy $F(b(\mathbf{x}), \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ to zero. The EM algorithm can be summarised as follows:

E-step	Set $b^{(t)}(\mathbf{x})$ to that b which minimises $F(b(\mathbf{x}), \widehat{\boldsymbol{\theta}}^{(t)})$.
M-step	Set $\widehat{\boldsymbol{\theta}}^{(t+1)}$ to that $\boldsymbol{\theta}$ which minimises $F(b^{(t)}(\mathbf{x}), \boldsymbol{\theta})$

As shown by Dempster *et al*. [19], each EM iteration increases the true log likelihood or leaved it unchanged. The EM algorithm will therefore converge to a local maximum. Given the above free energy formulation, which will be used throughout the thesis, one has to make proper parameterisations of the trial distribution $b(\mathbf{x})$, which we discuss in appendix C.

Appendix C

Free energy approximations

The goal of this section is to construct the variational free energy defined by the negative lower bound or similarly by the Kullback-Leibler divergence. This is done by defining the trial distribution $b(\mathbf{x})$ such that the resulting variational free energy is computationally tractable and accurate (appendix B).

The classic approximation is to assume $b(\mathbf{x})$ to be a fully factorisable distribution over the nodes. This assumption is equivalent to the mean field approximation known in physics for a long time. The machine learning and computer vision community use this to solve various problems, *i.e.*, in graphical models [59], stereo vision [129] and image restoration [130], to name only a few.

More recently, the Bethe approximation (introduced by the German physicist Hans Albrecht Bethe in 1935 [7]) and the more general Kikuchi approximation (introduced by the Japanese physicist Ryoichi Kikuchi [61]), gained importance, also in the compute vision community (see Yedidia, Freeman and Weiss [122, 123, 124, 125] for a theoretical view, and [50, 51, 26, 121] for applications).

After relating these concepts to physics and more particularly to statistical thermodynamics, two trial distributions will be considered, and the relation to the mean field and Bethe approximation will be made.

C.1 Relation to statistical thermodynamics

Using Bayes' rule, the distribution $p(\mathbf{y}, \mathbf{x} \,|\, \boldsymbol{\theta})$ is written as a product of data-likelihood and prior:

$$p(\mathbf{x}, \mathbf{y} \,|\, \boldsymbol{\theta}) \sim p(\mathbf{y} \,|\, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}) \;, \tag{C.1}$$

where the normalisation is neglected and where the assumption is made that the random field $\mathbf{x}$ is independent from $\boldsymbol{\theta}$. The prior reflects the smoothness properties of the random field and is therefore a distribution over the nodes x_j in the neighbourhood $j = N(i)$ of each node x_i; *e.g.*, $N(i)$ could be the four neighbourhood system (as shown in fig. C.1). If the data-likelihood factorises conditioned on the state of the hidden variables $\mathbf{x}$ over the individual nodes x_i, one can write the joint probability distribution similarly to [123, 126] as:

$$p(\mathbf{x},\mathbf{y}\,|\,\boldsymbol{\theta}) = \frac{1}{Z} \prod_{\substack{i,j \in N(i) \\ i>j}} \psi_{ij}(x_i, x_j) \prod_i \Phi_i(x_i, y_i, \boldsymbol{\theta}) \ , \tag{C.2}$$

where $\psi_{ij}(x_i, x_j)$ is the link interaction related to $p(\mathbf{x})$, $\Phi(x_i, y_i, \boldsymbol{\theta})$ the data-likelihood term and Z the normalisation constant or partition function. Note, that the node interconnection terms $\psi_{ij}(x_i, x_j)$ in eq. (C.2) are undirected links. Therefore, they are only counted once for each node pair $\{i, j\}$ with $i > j$. A graphical representation of eq. (C.2) with a four neighbourhood system, also know as Boltzmann machine, is depicted in fig. C.1. The joint probability distribution in eq. (C.2) is similar to the description of interacting particle systems in statistical physics (*e.g*. the Ising model describes particles with two states (spin up/down) which interact spatially). The distribution of these systems are described by an energy E and the temperature T dependent exponential $\exp(-\frac{1}{T}E(\mathbf{x},\mathbf{y},\boldsymbol{\theta}))$, *i.e*. the Boltzmann distribution. The difference between the formulation in statistical physics and the joint probability distribution in eq. (C.2) is the temperature. We will include the temperature in the joint probability distribution by makeing the replacement:

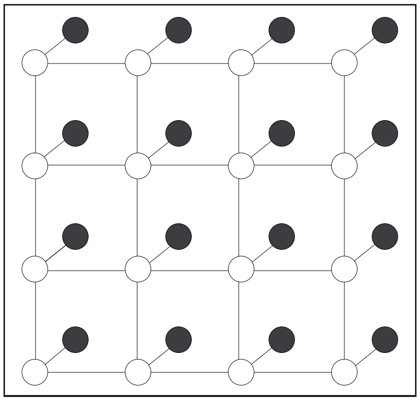


Figure C.1: **Boltzmann machine***: Observable and hidden nodes are gray and white circles, respectively. The lines between the hidden nodes (i, j) represent the prior interaction $\psi_{ij}(x_i, x_j)$. The lines between hidden and observable represent the data-likelihood $\Phi(x_i, y_i, \boldsymbol{\theta})$ as in eq. (C.2).*

$$p(\mathbf{x},\mathbf{y}\,|\,\boldsymbol{\theta}) \rightarrow p(\mathbf{x},\mathbf{y}\,|\,\boldsymbol{\theta})^{\frac{1}{T}} = \frac{1}{Z(T)} exp\left(-\frac{1}{T}E(\mathbf{x},\mathbf{y},\boldsymbol{\theta})\right) \ , \tag{C.3}$$

The joint probability distribution in eq. (C.2) is now conform with Bolzmann's law and the corresponding energy is up to a constant given by:

$$E(\mathbf{x},\mathbf{y},\boldsymbol{\theta}) = -\log p(\mathbf{x},\mathbf{y}\,|\,\boldsymbol{\theta}) \ . \tag{C.4}$$

The interpretation of the inference problem in terms of a temperature-dependent Boltzmann distribution has two advantages: Firstly, this allows us to specify the peakness of the joint probability distribution. Clearly, for $T \rightarrow 0$ the joint probability distribution allows only one configuration of the random field $\mathbf{x}$, *i.e.*, the one which has the highest probability $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta})$. For $T \rightarrow \infty$, one achieves a randomly distributed random field $\mathbf{x}$. This is in correspondence with our physical intuition. And second, the formulation with a temperature allows us to design stable convergence schemes by temperature annealing [62, 112].

Given the temperature dependent joint probability distribution as defined in eq. (C.3), the variational free energy is defined by the negative lower bound (B.4:

$$\begin{aligned} F(b(\mathbf{x}), \boldsymbol{\theta}) &= -T \sum_{\mathbf{x}} b(\mathbf{x}) \log \frac{p(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\theta})^{\frac{1}{T}}}{b(\mathbf{x})} \\ &= \sum_{\mathbf{x}} b(\mathbf{x}) E(\mathbf{x}, \boldsymbol{\theta}) + T \sum_{\mathbf{x}} b(\mathbf{x}) \log b(\mathbf{x}) + T \log Z \qquad \text{(C.5)} \\ &= U(b(\mathbf{x})) - T S(b(\mathbf{x})) + T \log Z \; . \qquad \text{(C.6)} \end{aligned}$$

The first term in eq. (C.5) represents the expected value of the energy $U(b(\mathbf{x}))$, followed by the negative entropy $S(b(\mathbf{x}))$ expectation and the Helmholtz free energy $T \log Z$. The variational free energy is minimal for $U(b(\mathbf{x})) = TS(b(\mathbf{x}))$, and is at this point equal to the Helmholtz free energy $T \log Z$.

What has been done so far is in close relation to statistical physics. There, macroscopic properties like energy or entropy (thermodynamical variables) are computed by the ensemble average of the local statistical particle properties (micro-canonical ensemble).

One other important thermodynamical variable is the heat capacity C, defined as the temperature derivative of U. A large value of C signals a change in the state of order of a system. It can therefore be used to determine the critical temperature, which can be seen as the largest temperature, where the value of $b(\mathbf{x})$ becomes peaked around a single value.

C.1.1 Example

As an example, the behaviour for a model, which consists of the MRF prior as defined in sec. 3.6.1, is illustrated. If T goes to infinity the prior is uniform (the probability of observing a specific configuration is random) and for $T \rightarrow 0$ the prior is strongly peaked around its most probable value(s).

Fig. C.2 shows samples from the prior distribution $p(\mathbf{x})^{1/T}$ as defined by eqs. (3.6) and (3.3) for different temperatures T. The random field for this simulation includes $R = 20$ depth states and two visibility states. The interaction matrix as defined in eq. (3.6) with $w = 1$ and $C = 0$.

For a high temperature ($T = 5$), the distributions for the depth and visibility are random (left of fig. C.2) and the energy E has its maximal value. By lowering the temperature, the energy decreases slowly as shown in fig. C.3. At a certain temperature,

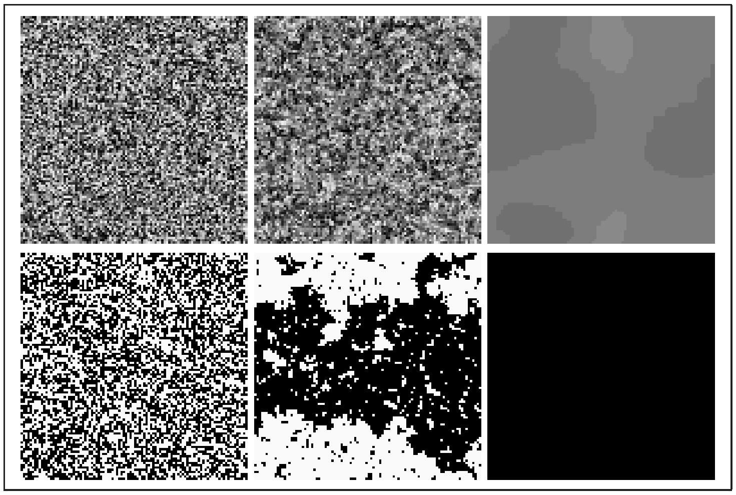


Figure C.2: **MRF Gibbs Prior:** *Samples of the prior distribution in eq. (3.3) for different temperatures* $T = \{5, 0.56, 0.00001\}$ *(from left to right). The top row shows a realisation of the depth states* $d_i \dots d_{20}$ *in gray values. And the bottom row shows the visibility states.*

the energy changes strongly. This point, indicated by a large peak in the first derivative of the energy (heat capacity), indicates a phase transition. In this experiment, it appeared near $T = 0.5$ and a corresponding configuration of the random field $\mathbf{x}$ can be seen in the middle column of figure C.2. Note that this simulation was performed without on the prior model in eq. (3.3) only (without data-likelihood).

C.2 Mean field approximation

In the specific case of the mean field approximation, $b(\mathbf{x})$ is chosen as a fully factoriseable distribution over the nodes x_i of the lattice (see fig. C.4):

$$b(\mathbf{x}) = \prod_i b_i(x_i) \ , \tag{C.7}$$

where $b_i(x_i)$ is the variational parameter (often called belief) that represents the expected value of the node x_i. Throughout the thesis, the belief of an individual state $b_i(x_i = m)$ will sometimes be shortly denoted by b_i^m. The beliefs are positive and

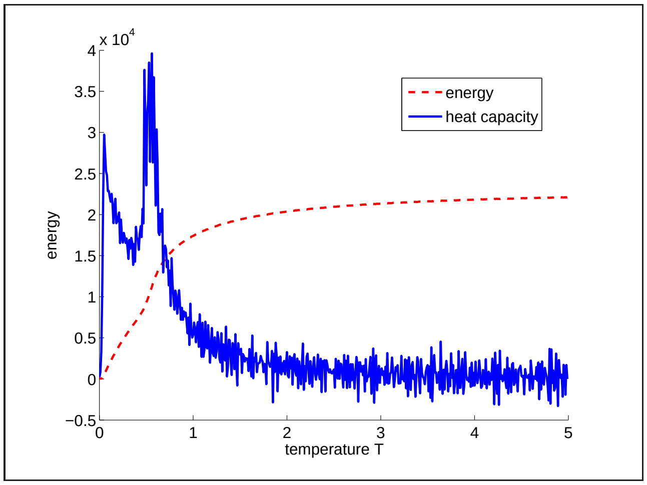


Figure C.3: **Temperature dependence:** *of the energy* $E = -\log p(\mathbf{x})$ *(dashed line) and the heat capacity* $\partial E/\partial T$ *(solid line) for the prior distribution* $p(\mathbf{x})^{1/T}$ *in eq. (3.3).*

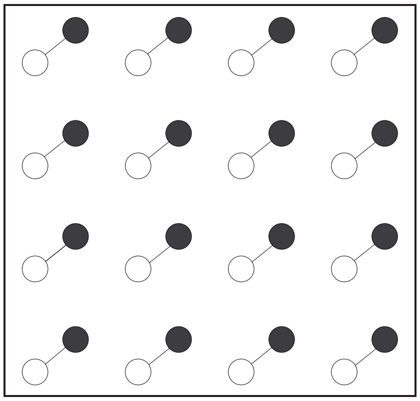


Figure C.4: **Mean field approximation:** *In the mean field approximation, the trial distribution is assumed to factorise over the nodes of the lattice. The result is, in contrast to fig.C.1, that the links between the hidden nodes are not considered.*

normalised over each node[1]:

$$0 \leq b_i(x_i) \qquad \forall x_i$$
$$\sum_{x_i} b_i(x_i) = \sum_{m=1}^{M} b_i^m = 1 \ . \tag{C.8}$$

---

[1] Here the sum $\sum_{x_i}$ denotes the sum over the states of node x_i (pixel i). This notation is identical to Yedidia *et al.*, *e.g.* [124].

The variational free energy of the distribution in eq. (C.2), subject to the mean field approximation, follows directly from the above factorisation and is given by:

$$\begin{aligned} F_{MF}(b(\mathbf{x}),\boldsymbol{\theta}) = & - \sum_i \sum_{\substack{j\in N(i)\\ i>j}} \sum_{x_i,x_j} b_i(x_i)b_j(x_j)\log\psi_{ij}(x_i,x_j) && \text{(C.9)}\\ & - \sum_i \sum_{x_i} b_i(x_i)\log\Phi_i(x_i,y_i,\boldsymbol{\theta}) && \text{(C.10)}\\ & + T\sum_i\sum_{x_i} b_i(x_i)\log b_i(x_i) + T\log Z\ . && \text{(C.11)} \end{aligned}$$

The first two terms represent the energy, followed by the negative entropy and the free energy.

Proof. The proof is trivial in the case of the mean field approximation. However, the transition from a sum over all possible random field configurations $\sum_{\mathbf{x}}$ to a sum over local configurations is essential. It is the step to make the inference problem computationally tractable and the reason to apply the mean field approximation. The proof is given for the entropy term eq. (C.12) (in the derivation $T=1$ is assumed):

$$\begin{aligned} \sum_{\mathbf{x}} b(\mathbf{x})\log\frac{b(\mathbf{x})}{p(\mathbf{y},\mathbf{x}\,|\,\boldsymbol{\theta})} = & \sum_{\mathbf{x}} b(\mathbf{x})\log b(\mathbf{x}) && \text{(C.12)}\\ & - \sum_{\mathbf{x}} b(\mathbf{x})\log p(\mathbf{x}\,|\,\mathbf{y},\boldsymbol{\theta}) - && \text{(C.13)}\\ & - \sum_{\mathbf{x}} b(\mathbf{x})\log p(\mathbf{x}) + \log Z\ . && \text{(C.14)} \end{aligned}$$

The other terms follow the same idea. Using the factorisable distribution in eq. (C.7) the entropy eq. (C.12) has the form:

$$\begin{aligned} -S = \sum_{\mathbf{x}} b(\mathbf{x})\log b(\mathbf{x}) & = \sum_{\mathbf{x}}\prod_i b_i(x_i)\log\prod_i b_i(x_i)\\ & = \sum_{\mathbf{x}}\prod_i b_i(x_i)\sum_i \log b_i(x_i)\ . \end{aligned}$$

Consider a specific term that depends on $\log b_i(x_i)$. This term is multiplied by the sum over all configurations of the factorisable distribution such that the value of x_i is fixed. One can therefore bring $b_i(x_i)$ in front and after this rearrangement one gets:

$$-S = \sum_i\sum_{x_i} b_i(x_i)\log b_i(x_i)\left(\sum_{\mathbf{x}/x_i}\prod_{j\neq i} b_j(x_j)\right)\ . \tag{C.15}$$

The sum over the remaining configurations is one because of the normalisation condition $\sum_{x_i} b_i(x_i) = 1$:

$$\sum_{\mathbf{x}/x_i}\prod_{j\neq i} b_j(x_j) = \sum_{x_1\neq i} b_i(x_1)\sum_{m_2\neq i} b_i(x_2)\ldots\sum_{m_N\neq i} b_N(x_N) = 1,$$

such that

$$\sum_{\mathbf{x}} \prod_i b_i(x_i) \log \prod_i b_i(x_i) = \sum_i \sum_m b_i(x_i^m) \log b_i(x_i^m) \tag{C.16}$$

The proof of the second term in eq. (C.13) is equivalent. For the third term in eq. (C.14), a similar strategy can be used. Here one also has to consider that $\log p(x_i)$ depends on the local neighbourhood $N(i)$.

□

The assumption of a factorisable trial distribution $b(\mathbf{x})$ eq. (C.7) leads to the variational free energy $F_{MF}(b(\mathbf{x}), \boldsymbol{\theta})$ which is indeed given by the sum of local expectations $b_i(x_i)$. The mean field update equation (E-step) is then given by setting the derivative of $F_{MF}(b(\mathbf{x}), \boldsymbol{\theta})$ with respect to $b_i(x_i)$ to zero:

$$\begin{aligned} \frac{\partial F_{MF}(b_i(x_i))}{\partial b_i(x_i)} &= \sum_{j \in N(i)} \sum_{x_j} b_j(x_j) \log \psi_{ij}(x_i, x_j) \\ &+ \log \Phi_i(x_i, y_i, \boldsymbol{\theta}) - T(\log b_i(x_i) - 1)\,. \end{aligned} \tag{C.17}$$

Note, the sum over the neighbours $N(i)$ for node i is not restricted to $i > j$. These terms appear from the derivative with respect to $b_{N(i)}$.

$$\boxed{b_i(x_i) = \exp\left(\frac{1}{T} \sum_j \sum_{x_j} b_j(x_j) \log \psi_{ij}(x_i, x_j) + \frac{1}{T} \log \Phi_i(x_i, y_i, \boldsymbol{\theta}) + 1\right)} \tag{C.18}$$

For the M-step, the derivative is taken with respect to $\boldsymbol{\theta}$, leading to:

$$\boxed{\boldsymbol{\theta} = \sum_i \sum_{x_i} b_i(x_i) \frac{\partial \log \Phi_i(x_i, y_i \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}} \tag{C.19}$$

C.2.1 Ising model

The mean field approximation has been given for the Potts model, where the number of states x_i is $m \geq 2$. For the special case of the Ising model $m = 2$, the interaction is given by a diagonal interaction matrix:

$$\psi_{ij} = \begin{pmatrix} J & 0 \\ 0 & J \end{pmatrix}\,. \tag{C.20}$$

Because of the normalisation condition, only the expected value for one state $b_i(x_1 = 1)$ has to be evaluated. For $m = 2$, and by using the normalisation $b_i(x_i = 2) =$

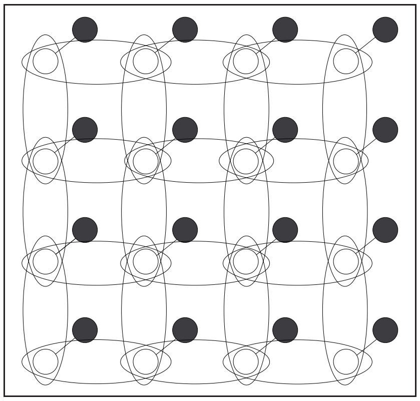


Figure C.5: **Boltzmann machine in Bethe approximation:** *In the Bethe approximation, the trial distribution contains the marginals of the one-node $b_i(x_i)$ and the two-node beliefs $b_{ij}(x_i, x_j)$, the latter indicated by ellipses. The result is, in contrast to fig.C.1, that the links between the hidden nodes are substituted by $b_{ij}(x_i, x_j)$.*

$(1 - b_i(x_i = 1))$, it is easy to see that the free energy in eq. (C.9) can be simplified and written in terms of $b_i = b_i(x_i = 1)$ only:

$$
\begin{aligned}
F_{MF}(b(\mathbf{x}), \boldsymbol{\theta}) \approx \;& - \;\; 2 \sum_i \sum_{\substack{j \in N(i) \\ i > j}} b_i b_j \log J && \text{(C.21)} \\
& - \;\; \sum_i b_i \log \frac{\Phi_i(x_i = 1, y_i = 1, \boldsymbol{\theta})}{\Phi_i(x_i = 2, y_i = 2, \boldsymbol{\theta})} && \text{(C.22)} \\
& + \;\; T \sum_i b_i \log b_i + (1 - b_i) \log(1 - b_i) \; . && \text{(C.23)}
\end{aligned}
$$

The derivative with respect to b_i leads to the mean field update equation for the Ising model:

$$
\boxed{b_i = \left(1 + \exp \left(-\frac{1}{T} \sum_j b_j \log J - \frac{1}{T} \log \frac{\Phi_i(x_i = 1, y_i = 1, \boldsymbol{\theta})}{\Phi_i(x_i = 2, y_i = 2, \boldsymbol{\theta})} \right) \right)^{-1} \qquad \text{(C.24)}}
$$

For $J = 0$, the beliefs can be computed in closed form:

$$
b_i = \frac{\Phi_i(x_i = 1, y_i = 1, \boldsymbol{\theta})}{\Phi_i(x_i = 1, y_i = 1, \boldsymbol{\theta}) + \Phi_i(x_i = 2, y_i = 2, \boldsymbol{\theta})} \; . \qquad \text{(C.25)}
$$

C.3 Bethe approximation

The derivation of the Bethe free energy is similar to the mean field case (allthough, as it will be discussed later, additional approximations are needed if the graph has loops).

In the Bethe approximation, the trial distribution $b(\mathbf{x})$ is formulated as a distribution not only over the one-node $b_i(x_i)$ but also over the two-node beliefs $b_{ij}(x_i, x_j)$ (see fig. C.5). The one-node beliefs $b_i(x_i)$ describe the probability for a node being in state $x_i = m$. The two-node beliefs $b_{ij}(x_i, x_j)$ describes the joint probability of node i and j being in state $x_i\!=\!m, x_j\!=\!n$. By this construction it follows that both beliefs should obey the following constraints [124]:

$$\begin{aligned} 0 \le b_i(x_i) \le 1 & \qquad 0 \le b_{ij}(x_i, x_j) \le 1 && \text{(C.26)} \\ \sum_{x_i} b_i(x_i) &= \sum_{x_i, x_j} b_{ij}(x_i, x_j) = 1 && \text{(C.27)} \\ b_i(x_i) &= \sum_{x_j} b_{ij}(x_i, x_j) \; . && \text{(C.28)} \end{aligned}$$

All belief entries should be positive (C.26) and normalised for each node (C.27). The last constraint (C.28) is the marginalisation. The Bethe approximation assumes the trial distribution $b(\mathbf{x})$ to be [124]

$$b(\mathbf{x}) = \frac{\prod_{ij} b_{ij}(x_i, x_j)}{\prod_i b_i(x_i)^{n_i - 1}} \; . \tag{C.29}$$

Here n_i is the amount of two-node beliefs connected to the node index i. In the example of the Boltzmann machine in fig. C.5, the values are: $n_i\!=\!4$ for all non boundary nodes and $n_i = 3$ and $n_i = 2$ for the boundary and corner nodes, respectively. One can show that the form of this distribution follows from converting a loopless undirected graph as for instance given by eq. (C.2) into the junction tree representation [115, 92]. The averadge energy in eq. (C.6) when computed with the exact joint likelihood $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta})$ will be exact for the Bethe approximation [124]. The Bethe approximation is related to the approximation of the entropy only. For graphs without loops, the entropy is given by:

$$\begin{aligned} \sum_{\mathbf{x}} b(\mathbf{x}) \log b(\mathbf{x}) &= \sum_{ij} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \log b_{ij}(x_i, x_j) \\ &- \sum_i (n_i - 1) \sum_{x_i} b_i(x_i) \log b_i(x_i) \; . && \text{(C.30)} \end{aligned}$$

Proof. By using eq. (C.30) with the replacement of the trial distribution eq. (C.29), the negative entropy can be written as:

$$\begin{aligned} -S = \sum_{\mathbf{x}} b(\mathbf{x}) \log b(\mathbf{x}) &= \sum_{\mathbf{x}} \frac{\prod_{ij} b_{ij}(x_i, x_j)}{\prod_i b_i(x_i)^{n_i - 1}} \log \prod_{ij} b_{ij}(x_i, x_j) && \text{(C.31)} \\ &- \sum_{\mathbf{x}} \frac{\prod_{ij} b_{ij}(x_i, x_j)}{\prod_i b_i(x_i)^{n_i - 1}} \log \prod_i b_i(x_i)^{n_i - 1} \; . && \text{(C.32)} \end{aligned}$$

One has to show that all terms that depend on $\log b_{ij}(x_i, x_j)$ in eq. (C.31) and that depend on $\log b_i(x_i)$ in eq. (C.32) fulfil the following relation:

$$\sum_{\mathbf{x}} \frac{\prod_{ij} b_{ij}(x_i, x_j)}{\prod_i b_i(x_i)^{n_i - 1}} \log b_{ij}(x_i, x_j) = \sum_{x_i, x_j} b_{ij}(x_i, x_j) \log b_{ij}(x_i, x_j) \quad \text{(C.33)}$$

$$\sum_{\mathbf{x}} \frac{\prod_{ij} b_{ij}(x_i, x_j)}{\prod_i b_i(x_i)^{n_i - 1}} (n_i - 1) \log b_i(x_i) = (n_i - 1) \sum_{x_i} b_i(x_i) \log b_i(x_i). \quad \text{(C.34)}$$

These relations can be proven by induction. Starting from a subgraph containing only one link, all other links will be added, such that the graph remains loopless.

Let the trivial subgraph be described by $b(x_i)$, $b(x_j)$, $b_{ij}(x_i, x_j)$. Each node has one neighbour, *i.e.*, $n_i = n_j = 1$. And the eqs. (C.33) and (C.34) are trivially true. To add more links by preserving the loopless property of the graph, a link $b_{ik}(x_i, x_k)$ can only be added from an existing node x_i to a *n*ew node x_k. By doing so the sum over all configurations will include the configurations of the new node x_k. Furthermore the node x_i gets one additional neighbour $n_i \rightarrow n_i + 1$. By adding the link eq. (C.33) changes to:

$$\begin{aligned} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \log b_{ij}(x_i, x_j) \quad &\rightarrow \quad \sum_{x_i, x_j, x_k} \frac{b_{ij}(x_i, x_j) b_{ik}(x_i, x_k)}{b_i(x_i)} \log b_{ij}(x_i, x_j) \\ &= \quad \sum_{x_i, x_j} b_{ij}(x_i, x_j) \log b_{ij}(x_i, x_j)\,, \qquad \text{(C.35)} \end{aligned}$$

where in the last line the consistency condition eq. (C.28) was used. In the same fashion, all other links can be added and eq. (C.33) is proven.

In a similar fashion eq. (C.34) can be proven.

$$\begin{aligned} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \log b_i(x_i) \quad &\rightarrow \quad \sum_{x_i, x_j, x_k} \frac{b_{ij}(x_i, x_j) b_{ik}(x_i, x_k)}{b_i(x_i)} \log b_i(x_i) \\ &= \quad \sum_{x_i} b_i(x_i) \log b_i(x_i). \qquad \text{(C.36)} \end{aligned}$$

Again, the consistency condition eq. (C.28) is used: $\sum_j b_{ij}(x_i, x_j) = b_i(x_i)$ and $\sum_k b_{ik}(x_i, x_k) = b_i(x_i)$. □

If the graph has loops, the factorisation of $b(\mathbf{x})$ as in eq. (C.29) does not lead to the exact entropy in eq. (C.30). The Bethe approximation, however, assumes that (C.30) is still approximatly true.

Using eq. (C.5) together with the Bethe approximation, the variational free energy is given by:

$$
\begin{aligned}
F_B(b(\mathbf{x}),\boldsymbol{\theta}) = \quad & - \sum_{i} \sum_{\substack{j \in N(i) \\ i>j}} \sum_{x_i,x_j} b_{ij}(x_i,x_j) \log \psi_{ij}(x_i,x_j) && \text{(C.37)} \\
& - \sum_{i} \sum_{x_i} b_i(x_i) \log \Phi_i(x_i,y_i,\boldsymbol{\theta}) && \text{(C.38)} \\
& - T \sum_{i} \sum_{j} \sum_{x_i,x_j} b_{ij}(x_i,x_j) \log b_{ij}(x_i,x_j) && \text{(C.39)} \\
& + T \sum_{i} (n_i - 1) \sum_{x_i} b_i(x_i) \log b_i(x_i) + T \log Z \; . && \text{(C.40)}
\end{aligned}
$$

The EM-algorithm proceeds by iterating the following steps. In the E-Step, the Bethe free energy is minimised w.r.t. b_i and b_{ij} by belief propagation [80]. In the M-Step, the parameters are updated. This is achieved by setting each parameter $\boldsymbol{\theta}$ to the appropriate root of the derivative equation:

$$\partial F_B / \partial \boldsymbol{\theta} = 0 \; .$$

The updates of the parameters are the same for both free energy approximations, because they only influence the data-likelihood terms in F_{MF}(C.10) and F_B (C.38). These terms are identical.

Even though the Bethe approximation has many concerns from the theoretical point of view, discussed in [71], it has been shown to be a good approximation in practice [118, 120].

It is mainly due to the work of Yedidia, Freeman and Weiss [122, 123, 124, 125] that the connection is made between the Bethe and Kikuchi free energy approximations with popular message passing algorithms. More particular, they showed [122] that the fixed points of belief propagation [80] correspond to the stationary points of the Bethe free energy in the case of loopless graphical models. The same has been proven more recently by Heskes [45] for models with loops. Other algorithms which minimise the Bethe free energy are studied in [128, 46, 114]

Inference problems in the form of ML estimates in eq. (B.1) or in the form of computing marginal distributions from free energy approximations are also the subject of intensive research in machine learning community [55, 34, 71, 6].

Appendix D

Nonlinear depth diffusion

The depth ($\mathcal{D}$) dependent parts of the free energy, given by eq. (4.14), are:

$$\begin{aligned} F_{MF}[\mathcal{D}] &= -\sum_i \sum_k \sum_m b_i^m \log p(y_{i'}^k \,|\, x_i^m, \boldsymbol{\theta}) \\ &\quad + \frac{1}{\lambda} \sum_i (\nabla \mathcal{D}_i)^T T(\nabla \mathbf{X})\, \nabla \mathcal{D}_i + \frac{1}{\lambda_c} \sum_i \mathcal{W}_i (\mathcal{D}_i - \mathcal{G}_i)^2 \; . \end{aligned} \tag{D.1}$$

The first term in this equation is the so called ‘matching term’, the second is the smoothness term, weighted by the parameter λ, which is related to the width of the depth-prior distribution. The third term relates the depth $\mathcal{D}_i$ to $\mathcal{G}_i$, *e.g.*, to a sparse set of initial depth points, which will be switched on by $\mathcal{W}_i \neq 0$. Our goal is to minimise the free energy in eq. (D.1) with respect to $\mathcal{D}$. The minimisation procedure we use here is an assimilated version of Alvarez *et al*. [4], where a similar minimisation with respect to the optical flow has been considered.

We can rewrite the depth dependent part of the matching term for each pixel using the expected values for the visibility $\mathcal{V}^k$ in eq. (3.21) and the Gaussian inlier distribution eq. (4.13):

$$\sum_k \sum_m b_i^m \log p(y_{i'}^k \,|\, x_i^m, \boldsymbol{\theta}) = -\sum_k \mathcal{V}^k (m_i^k)^T \boldsymbol{\Sigma}^{-1} m_i^k \; , \tag{D.2}$$

where the assumption is made that the outlier distribution is independent of $\mathcal{D}$.[1] Furthermore, we have introduced the colour difference of the ideal image with the colour transformed k^{th} input image m_i^k:

$$m_i^k = y_i^* - \boldsymbol{C}(\mathbf{p^k}) y_{i'}^k \; . \tag{D.3}$$

The value of $\mathcal{D}_i$ will be split into a current estimate $\mathcal{D}_i^0$ and a small residual $\mathcal{D}_i^r$, such that: $\mathcal{D}_i = \mathcal{D}_i^0 + \mathcal{D}_i^r$. By taking the Taylor expansion of eq. (D.3) and using the

---

[1] If the outlier distribution is modelled as a constant, this assumption is true. However, when modelled by a histrogram, the outlier distribution depends on $\mathcal{D}$, since the histrogram $\mathbf{h}^k$ is filled with the colours of $y_{i(\mathcal{D})}^k$. This dependence can expected to be small and will be ignored.

particular form of the depth dependent mapping $i \rightarrow i'(\mathcal{D}_i)$ eq. (A.4) this leads to:

$$m_i^k = y_i^* - \boldsymbol{C}(\mathbf{p^k}) \left(y^k_{i'(\mathcal{D}_i^0)} + \frac{\partial y^k_{i'(\mathcal{D}_i^0)}}{\partial i'} \frac{\partial l^k(i')}{\partial \mathcal{D}_i} (\mathcal{D}_i - \mathcal{D}_i^0) + O\left((\mathcal{D}_i^r)^2\right) \right) \ . \tag{D.4}$$

By using this result in eq. (D.1) the associated Euler-Lagrange equation lead to:

$$\frac{1}{\lambda} \text{div}(T(\nabla \mathbf{X}) \nabla \mathcal{D}) + \sum_k \mathcal{V}^k (m^k)^T \mathbf{\Sigma}^{-1} \frac{\partial m^k}{\partial \mathcal{D}} + \frac{1}{\lambda_c} \mathcal{W}(\mathcal{D} - \mathcal{G}) = 0 \ . \tag{D.5}$$

This equation can be interpreted as the equilibrium state ($\partial \mathcal{D} / \partial t = 0$) of a depth diffusion process. With τ being the temporal step size, we get:

$$\frac{\mathcal{D} - \mathcal{D}^0}{\tau} = \text{div}(T(\nabla \mathbf{X}) \nabla \mathcal{D}) + \lambda \sum_k \mathcal{V}^k (m^k)^T \mathbf{\Sigma}^{-1} \frac{\partial m^k}{\partial \mathcal{D}} + \frac{\lambda}{\lambda_c} \mathcal{W}(\mathcal{D} - \mathcal{G}) \tag{D.6}$$

Equation (D.6) is the realisation of an implicit discretisation scheme. This has the advantage that the temporal time step τ can be chosen larger than it could be for the corresponding explicit scheme.[2] This of course leads to an faster convergence.

The solution of eq. (D.6) can be computed in matrix form:

$$\mathbf{A} \mathcal{D} \quad = \quad \mathbf{b} \tag{D.7}$$

and solved using Gauss-Seidel iterations. Thereby $\mathbf{A}$ is split into diagonal $\mathbf{D}$, upper diagonal $\mathbf{U}$ and lower diagonal $\mathbf{L}$ part as:

$$\mathbf{A} \quad = \quad \mathbf{D} - \mathbf{L} - \mathbf{U} \ , \tag{D.8}$$

and the system

$$\begin{aligned} (\mathbf{D} - \mathbf{L}) \mathcal{D}^{\frac{1}{2}} &= \mathbf{U} \mathcal{D}^0 + \mathbf{b} \\ (\mathbf{D} - \mathbf{U}) \mathcal{D} &= \mathbf{L} \mathcal{D}^{\frac{1}{2}} + \mathbf{b} \end{aligned} \tag{D.9}$$

is solved using the forward-backward substitution [88]. With

$$c^k = \boldsymbol{C}(\mathbf{p^k}) \frac{\partial y^k_{i'(\mathcal{D}_i^0)}}{\partial i'} \frac{\partial l^k(i')}{\partial \mathcal{D}_i^r} \ , \tag{D.10}$$

the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ are given by:

$$\begin{aligned} \mathbf{A} &= \mathbf{1} - \tau \lambda \sum_k V^k \left(c^k\right)^T \mathbf{\Sigma}^{-1} c^k - \frac{\lambda}{\lambda_c} \tau \mathcal{W} - \tau \tilde{\mathbf{A}} \qquad \text{(D.11)} \\ \mathbf{b} &= \mathcal{D}^0 - \tau \lambda \sum_k V^k \left(c^k\right)^T \mathbf{\Sigma}^{-1} \left(y^* - \boldsymbol{C}(\mathbf{p^k}) y^k_{\mathcal{D}^0} + c^k \mathcal{D}^0\right) - \frac{\lambda}{\lambda_c} \tau \mathcal{W} \mathcal{G} \ . \end{aligned}$$

---

[2] The theoretical value in 2 dimensions is $\tau \leq 0.25$ for the explicit scheme.

Here, $\tilde{\mathbf{A}}$ represents the divergence term div$(T(\nabla\mathbf{X})\nabla\mathcal{D})$. It is the only matrix in eq. (D.11) with off diagonal elements. Furthermore, $\tilde{\mathbf{A}}$ is a very sparse with 8 nonzero elements for each row and the computation is identical to [4].

In our previous work [106, 103], we implemented the third term in eq. (D.1) by an anisotropic time diffusion scheme. There, the parameter τ has been set to a small value for all pixels for which initial $3D$ points are available. Once the depth $\mathcal{D}$ is initialised with these points, they will only move slowly because of the small τ value. In this way, the corresponding energy term (third term in eq. D.1) can be neglected and a similar result is obtained. However, the formulation presented here has the advantage that the uncertainty of the initial 3-D points could be consistently taken into account (by adjusting the value of $\lambda_c\mathcal{W}_i$ accordingly). Although, we use only initial 3-D points without uncertainties, this would be possible and leads to a more consistent formulation.

List of Publications

Articles in International Journals:

[1] C. STRECHA, R. FRANSENS AND L. VAN GOOL: A probabilistic formulation of image registration, In *Lecture notes in computer science, vol. 3417, Jähne, B.; Mester, R.; Barth, E.; Scharr, H. (Eds.) 2007 (Complex Motion, First International Workshop, IWCM 2004, Günzburg, Germany, October 12-14, 2004, revised papers)*

[2] R. FRANSENS, C. STRECHA AND L. VAN GOOL: Optical Flow based Super-Resolution: A Probabilistic Approach, To appear in *Computer Vision and Image Understanding (CVIU), 2007*

[3] E.M. ILGENFRITZ, A. SCHILLER AND C. STRECHA: Matter Near To The Endpoint Of The Electroweak Phase Transition *Nucl.Phys.Proc.Suppl. vol 73, pp. 662-664, 1999*

[4] E.M. ILGENFRITZ, A. SCHILLER AND C. STRECHA: Wave Functions And Spectrum In Hot Electroweak Matter For Large Higgs Masses *Eur.Phys.J. C vol 8, pp. 135-150, 1999*

[5] M. GURTLER, E.M. ILGENFRITZ, A. SCHILLER AND C. STRECHA: Hot Electroweak Matter Near To The Endpoint Of The Phase Transistion *Nucl.Phys.Proc.Suppl. vol 63, pp. 563-565, 1998*

Articles in Proceedings of International Conferences:

[1] R. FRANSENS, C. STRECHA, G. CAENEN AND L. VAN GOOL: A generic approach to image coregistration, In *ESA-EUSC 2006: Image Information Mining for Security and Intelligence, EUSC, Torrejon air base - Madrid (Spain), November 27-28, 2006*

[2] C. STRECHA, R. FRANSENS AND L. VAN GOOL: Combined depth and outlier estimation in multi-view stereo, In *Proceedings IEEE computer society conference on computer vision and pattern recognition - CVPR, vol. II, pp. 2394-2401, June 17-22, 2006, New York, NY, USA*

[3] R. Fransens, C. Strecha and L. Van Gool: A mean field EM-algorithm for coherent occlusion handling in MAP-estimation problems, In *Proceedings IEEE computer society conference on computer vision and pattern recognition - CVPR, vol. I, pp. 300-307, June 17-22, 2006, New York, NY, USA*

[4] R. Fransens, C. Strecha and L. Van Gool: Robust estimation in the presence of spatially coherent outliers, *25 years of RANSAC workshop (in conjunction with CVPR), 8 pp., June 18, 2006, New York, NY, USA*

[5] R. Fransens, C. Strecha and L. Van Gool: Parametric stereo for multi-pose face recognition and 3D-face modeling, In *Lecture notes in computer science, vol. 3723, pp. 108-123, 2005 (Proceedings second international workshop on analysis and modelling of faces and gestures, AMFG-2005, October 16, 2005, Beijing, China)*

[6] C. Strecha, R. Fransens and L. Van Gool: Wide-baseline stereo from multiple views : a probabilistic account, In *Proceedings IEEE computer society conference on computer vision and pattern recognition - CVPR, vol. I, pp. 552-559, June 27 - July 2, 2004, Washington, DC, USA*

[7] C. Strecha, R. Fransens and L. Van Gool: A probabilistic approach to large displacement optical flow and occlusion detection, In *Lecture notes in computer science, vol. 3247, pp. 71-82, 2004 (Statistical methods in video processing - ECCV 2004 workshop SMVP 2004, revised selected papers, May 16, 2004, Prague, Czech Republic)*

[8] R. Fransens, C. Strecha and L. Van Gool: Multimodal and Multiband Image Registration using Mutual Information, In *ESA-EUSC 2004: Theory and Applications of Knowledge driven Image Information Mining, with focus on Earth Observation, EUSC, Madrid (Spain) March 17-18, 2004*

[9] R. Fransens, C. Strecha and L. Van Gool: A Probabilistic Approach to Optical Flow based Super-Resolution, In *CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 12*

[10] M. Vergauwen, F. Verbiest, V. Ferrari, C. Strecha and L. Van Gool: Wide-baseline 3D reconstruction from digital stills, In *International workshop on visualization and animation of reality-based 3D models, 7 pp., February 24-28, 2003, Tarasp-Vulpera, Engadin, Switzerland*

[11] C. Strecha, F. Verbiest, M. Vergauwen and L. Van Gool: Shape from video v.s. still images, In *Proceedings Conference on optical 3-D measurement techniques, vol. 2, pp. 168-175, September 22-25, 2003, Zürich, Switzerland*

[12] C. Strecha, T. Tuytelaars and L. Van Gool: Dense matching of multiple wide-baseline views, In *Proceedings 9th IEEE international conference on computer vision, ICCV, vol. 2, pp. 1194-1201, October 13-16, 2003, Nice, France*

[13] C. Strecha and L. Van Gool: Motion-stereo integration for depth estimation, In *Lecture notes in computer science, vol. 2351, pp. 170-185 (Proceedings 7th European conference on computer vision - ECCV 2002, part II, May 28-31, 2002, Copenhagen, Denmark)*.

[14] C. Strecha and L. Van Gool: PDE-based multi-view depth estimation, In *Proceedings 1st international symposium on 3D data processing visualization and transmission - 3DPVT, pp. 416-425, June 19-21, 2002, Padova, Italy*

[15] N. Kogo, C. Strecha, R. Fransens, G. Caenen, J. Wagemans and L. Van Gool: Reconstruction of subjective surfaces from occlusion cues, In *Lecture notes in computer science, vol. 2525, pp. 311-321, 2002 (Proceedings 2nd international workshop on biologically motivated computer vision - BMCV, November 22-24, 2002, Tübingen, Germany)*

[16] L. Van Gool, T. Tuytelaars, V. Ferrari, C. Strecha, J. Vanden Wyngaerd and M. Vergauwen: 3D modeling and registration under wide baseline condition, In *Proceedings ISPRS commission III symposium on photogrammetric computer vision, vol. XXXIV, part 3A, pp. 3-14, September 9-13, 2002, Graz, Austria*

International Conference : Abstract or Not Published

[1] N. Kogo, C. Strecha, G. Caenen, R. Fransens, G. Van Belle, J. Wagemans and L. Van Gool: End-stopped cue detection for subjective surface reconstruction, *26th European conference on visual perception - ECVP 2003, September 1-5, 2003, Paris, France*

[2] M. Gurtler, E.M. Ilgenfritz, A. Schiller and C. Strecha: Hot Electroweak Matter Near To The Critical Higgs Mass *Talk given at 31st International Ahrenshoop Symposium on the Theory of Elementary Particles, Buckow, Germany, 2-6 Sep 1997. In Buckow 1997, Theory of elementary particles 253-258*

Curriculum Vitae

Christoph Strecha was born on august 27, 1970 in Radebeul, Germany. In 1998, he received a Masters degree in Physics from the University of Leipzig, Germany. From 1998 to 1999, he worked in the R&D devision of MV Technology (Dublin) in the area of visual inspection. From 1999 to 2007, he worked under the supervision of Prof. Luc Van Gool, in the VISICS lab (part of ESAT-PSI) at the K.U.Leuven for various European projects. He started to work towards his PhD in 2004. His main research interests include computer vision, multi-view stereo and optical flow.