Technical Report

Pattern Recognition and Image Processing Group
Institute of Automation
Vienna University of Technology
Treitlstr. 3/1832
A-1040 Vienna AUSTRIA
Phone: +43 (1) 58801-8161
Fax: +43 (1) 5054668
E-mail: men@prip.tuwien.ac.at

PRIP-TR-45 January 17, 1997

Robust Stereo and Adaptive Matching in Correlation Scale-Space

Christian Menard

Abstract

The stereo analysis method is similar to the human visual system. Due to the way our eyes are positioned and controlled, our brains usually receive similar images of a scene taken from nearby points of the same horizontal level. Therefore the relative position of the images of an object will differ in the two eyes. Our brains are capable of measuring this difference and thus estimating the depth. Stereo analysis tries to imitate this principle.
This work contains two complementary and original contributions, one combines stereo techniques with robust statistics and the other solves the correspondence problem in a multi-scale approach using correlation scale-space. Most standard algorithms make unrealistic assumptions about noise distributions, which leads to erroneous results that cannot be corrected in subsequent processing stages. In this work the standard area-based correlation approach is modified so that it can tolerate a significant number of outliers. The approach exhibits a robust behavior not only in the presence of mismatches but also in the case of depth discontinuities.
Another central problem in stereo matching using correlation techniques lies in selecting the size of the search window. Small windows contain only a small number of data points, and thus are very sensitive to noise and therefore result in false matches. Whereas large search windows contain data from two or more different objects or surfaces, thus the estimated disparity is not accurate due to different projective distortions in the left and the right image. In this work a new method is proposed providing a continuous scale for the matching process, so that for each region in the stereo pair depending on the local information an optimal scale can be estimated.
Results are given on synthetic images for the robust correlation technique. The adaptive matching method using correlation scale-space is tested on both synthetic and real images.

Acknowledgments

This thesis would not have been possible without the guidance, advice and support of numerous people in our PRIP (Pattern Recognition and Image Processing) laboratory at the Vienna University of Technology, particularly I want to thank:

- *Prof.Kropatsch*, my thesis supervisor, who demonstrated to me what it means to work scientifically. His valuable suggestions and criticism have influenced this thesis to a large extend and made the completion of this work possible.
- *Aleš Leonardis*, for reviewing parts of this thesis and for many thorough discussions in the field of robust statistics.
- *Horst Bischof*, for reviewing drafts of this thesis, for his advice in everyday office life, and for helping me solving various LaTeX problems.
- *Robert Sablatnig*, with whom I share the office for years, for his friendship and endless discussions on the dissertation topic.
- *Amy Krois-Lindner*, for carefully proofreading this thesis.
- *Alexius Korzinek*, for providing excellent resources, especially for the installation of Linux and Khoros.

Furthermore I would like to thank my students and colleagues, who have contributed to this work: *Ingeborg Tastl, Bergita Göbel, Souheil Ben-Yacoub, Norbert Brändle, Helmut Kofler, Martin Kampel, Hannes Föttinger and Michael Reiter.*

Last, but not least, I would like to express my gratitude to my parents, who have made my studies possible and who always encouraged and supported me in my goals.

Contents

1 Introduction and Overview **7**

1.1 Introduction 7
1.2 Goals of this Thesis 9
1.3 Overview 10

2 Stereo-Vision **11**

2.1 The Human Visual System 11
2.2 Stereo Image Geometry 12
2.3 Epipolar Geometry 14
2.4 Calibration of a Stereo System 15
2.5 The Correspondence Problem 19
2.6 Occlusion 22
2.7 Constraints in Stereo-Vision 23
2.7.1 Geometric Constraints 23
2.7.2 Object-Based Constraints 24
2.8 Area-Based vs. Feature-Based Stereo 24
2.9 Correlation between two Signals 25
2.10 Standard Area-Based Stereo 30
2.11 Discussion 32

3 Robust Correlation **35**

3.1 Introduction 35
3.2 Outliers in Stereo-Vision 36
3.3 Robust Computation of Correlation 38
3.3.1 M-estimator 38
3.3.2 Iterative weighted Least Squares Method 44
3.4 Experimental Results and Comparisons 45
3.4.1 Salt & Pepper Noise 49
3.4.2 Additive Gaussian Noise 54
3.5 Chapter Summary 56

4 Multi-Scale Stereo **59**
4.1 Introduction 59
4.1.1 Pyramid Representation 60
4.1.2 Scale-Space 62
4.1.3 Multi-Scale vs. Pyramid Representation 68
4.2 Hierarchical Matching Using Image Pyramids 69
4.3 Coarse to Fine Matching 72
4.4 Adaptive Matching in Correlation Scale-Space 77
4.4.1 Correlation with a Weighted Function 77
4.4.2 Scale Selection 83
4.5 Experimental Results 87
4.5.1 Results on Synthetic Images 87
4.5.2 Results on Real Images 88
4.6 Chapter Summary 93

5 Conclusion and Outlook **95**

Bibliography **97**

List of Figures

1.1 Stereo-vision paradigm by Barnard and Fischler. 9

2.1 Binocular vision. 12
2.2 Stereo image geometry. 13
2.3 Triangulation 14
2.4 Epipolar geometry 15
2.5 Camera geometry with perspective projection and radial lens distortion. 16
2.6 Detection of the calibration points in the image plane 18
2.7 Ambiguous correspondence 19
2.8 Correspondence problem 20
2.9 Disparity 21
2.10 Occlusion 22
2.11 Disparity $D(p')$ between two signals $I_L(x)$ and $I_R(x)$ at position p'. 26
2.12 1D rect function 27
2.13 $\delta_{1/w}(x)$ for $f(x + x_L)$. 28
2.14 2D rect function 29
2.15 Correlation with function $\delta_{1/w^2}(x, y)$ 30
2.16 Correspondence establishment 31
2.17 Synthetic stereo pair of a box on a flat ground with natural texture added on the surface. 32
2.18 Disparity map computed with the standard stereo method 32

3.1 Test performed on a stereo pair corrupted with one outlier 37
3.2 Different correlation functions for a clean and for a window corrupted with one outlier 38
3.3 Different objective functions ρ. 40
3.4 An objective function for three classes of doubts 41
3.5 Hard, soft-redescenders and monotone functions 43
3.6 Synthetic stereo pair: Box 45
3.7 *Outlier test:* Robust vs. Standard technique 47
3.8 Comparison of the results computed with both standard and robust stereo technique 48
3.9 *Residual test* under Salt & Pepper noise condition 50

3.10 Disparity maps computed with the standard and with the robust stereo method under Salt & Pepper noise condition 51
3.11 *MSE test* under Salt & Pepper noise condition 53
3.12 *Residual test* under Gaussian noise condition 54
3.13 Disparity maps computed with the standard and the robust stereo method under Gaussian noise condition 57
3.14 *MSE test* under Gaussian noise condition 58

4.1 Pyramid representation. 61
4.2 Six levels of a 2 $\times 2/4$ pyramid. 61
4.3 Multi-scale representation of a signal at different scales. 62
4.4 Signal $I(x)$ is successively smoothed using Gaussian kernels of increasing width t. 64
4.5 Scale-space representation of a one-dimensional signal $I(x)$. 65
4.6 Gray-level illustrations of sampled Gaussian kernels at different scale levels t. 66
4.7 Gray-level images of an archaeological sherd at different scales . . 67
4.8 Hierarchical matching algorithm 70
4.9 Disparity maps computed with different window sizes w. 71
4.10 Synthetic stereo pair: Pyramid on a flat ground with natural texture added on the surface. 72
4.11 Two-dimensional correlation function for a complete scanline in the left stereo image 74
4.12 Zoomed regions of the two-dimensional correlation function 75
4.13 The correlation function gets smoother with larger window sizes w 76
4.14 Correlation at different scales 77
4.15 Two-dimensional scale-space kernels for different values $t = 4, 3, 2, 1$ 79
4.16 Correlation Scale-Space 80
4.17 3D plots of the Correlation Scale-Space 81
4.18 Ambiguous correspondence 82
4.19 *CSS* for one point and the corresponding zero crossings of the first derivative 83
4.20 Correlation as a function of scale for different placements 84
4.21 *CSS* maxima 86
4.22 Test on a synthetic stereo pair: Pyramid 89
4.23 Test on a synthetic stereo pair: Sphere 90
4.24 Test on a real stereo pair: Archaeological sherd 1 91
4.25 Test on a real stereo pair: Archaeological sherd 2 92

Chapter 1

Introduction and Overview

1.1 Introduction

The task of perceiving three-dimensional information in a scene can be carried out by the human visual system already in the first days of life. Vision tasks try to imitate the capability to see objects, to estimate their three-dimensional position in the world, and to identify the objects for grasping them. But the tasks of perceiving depth, identifying objects and grasping are performed by humans in real time for everyday purposes, without being aware of this complex process. Until now computer vision has tried to imitate these processes, which are performed by the human brain. But so far it is not clear how these processes work and interact with each other. A good introduction to the human visual system, especially considering the wealth of literature in neurophysiology, psychology and psychophysics is given by Gregory [Gre78]. A more detailed description is given by Levine [LS91]. The eyes are only instruments to acquire images of the real world. These images are interpreted in our brain and for the same scene, seen by different people, there exist different interpretations. So the question arises, what do we really see? Do we see a scene or only what our brain wants us to see? It is clear why researchers in the field of computer vision do not simply build systems that emulate the human visual system, because what is known about the human visual system beyond the human eye has only a speculative background. The fallibility of the human visual system is demonstrated by the existence of visual illusions, such as the Necker cube, Frasers spiral, etc. (See Gregory and Frisby for more examples [Gre78, Fri80].) These illusions show us that the human visual system is not infallible and the all-important question arises whether human vision is just controlled by hallucination. Helmholtz expresses the view that "Every image is the image of a thing merely for him who knows how to read it, and who is enabled by the aid of the image to form an idea of the thing" [Hel24]. How can vision be defined? Marr stated that "vision is the process of discovering from images what is present in the world and where it is" [Mar82].

An important fact is the **robustness** of the human visual system. It can cope with various noise and lighting conditions. For example a person driving a car can guide it perfectly although the visibility is very poor because of snow or fog. These two situations represent different noise conditions. For snow some regions in the field of vision are **replaced** by snow-flakes, whereas for the situation with fog the scene is interfered with **additive** wrong information. Moreover humans are able to recognize objects even if they are partly **occluded** or have changed their appearance over time. For example humans are able to recognize a person by his face even if it has changed over years.

Another important fact is that of **scale**. If the scale is too low for a certain problem it can be refined by humans by foveating interesting structures, or if necessary, moving closer to the interesting object. This movement process makes it possible to acquire additional information about the three-dimensional structure of an object. So the main tasks to be solved by vision algorithms are what kind of information should be extracted at the earliest stages, and what kind of operators should be performed on the data that reach the visual sensors.

If an object in the real world is observed by a single camera the three-dimensional information is projected into a two-dimensional image plane. In this case it is photogrammetrically impossible to reconstruct the three-dimensional information for this object. In contrast to computer vision, in computer graphics first three dimensional models are created, which are then transformed in two dimensions by using shading and ray tracing techniques. But this process is not reversible, since the information of depth is lost during the projection process. In order to reconstruct the **three-dimensional information** of a scene, computer vision tries to imitate the human visual system to use two different views of the same scene. This technique is known under the term **stereo-vision** or **binocular vision**.

It is very important that a vision system incorporates the properties of robustness and that interesting structures for certain scales are found to the stereo-vision process. According to Barnard and Fischler [BF82a] a stereo-vision system contains the steps that are shown in Figure 1.1. All of these steps play important roles in the design of a stereo system and it is of interest to know how the different tasks depend on each other. Only small mistakes that are made during this processes can result in wrong depth informations for the object which is currently under consideration. Starting from the image acquisition, noise can be introduced for certain reasons into the images, which complicates the modeling of the used cameras and of course the search for the corresponding points in a stereo pair. But the success of the approach mostly depends on its ability to solve the problem of stereo matching, a topic upon which this thesis concentrates.

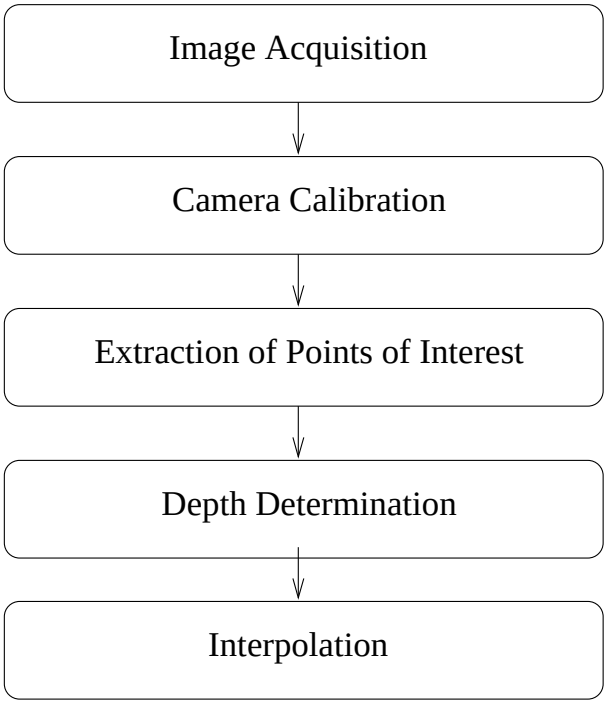


Figure 1.1: Stereo-vision paradigm by Barnard and Fischler.

1.2 Goals of this Thesis

This work contains two complementary and original contributions; one tries to combine stereo techniques with robust statistics to form

- a **robust** version of the **correlation** and the other solves the correspondence problem adaptively in a
- multi-scale framework using a **correlation scale-space**.

Standard stereo techniques are very sensitive to noise. Noise can be introduced to a stereo pair in many ways. It can be sensor noise or points belonging to different surfaces or objects. One main goal of this work is to provide a **robust** version of a standard **stereo matching technique**, which can tolerate a significant amount of noise.

Another central problem in stereo matching using correlation techniques lies in selecting the size for the operator used. The size of the operator must be large enough to contain enough data points to have a large gray-level variation, but small enough to avoid the effect of perspective distortion. If the size of the operator is too small and does not cover a sufficient gray-level variation the signal to noise ratio is low, thus giving only a poor estimate for the disparity value. If

the size of the operator is too large it contains data points belonging to different objects or surfaces, thus the estimated depth value is not accurate due to different projective distortions in the left and the right image.

In this work a new adaptive method is proposed in which the size of the operator is changed continuously depending on the local variations of intensity and three-dimensional information, thus making it possible to estimate an optimal size for the operator used for certain regions in a stereo pair.

1.3 Overview

The chapters in this work are written to be as self-contained as possible, so that they can be read independently of each other. Each chapter starts with a brief introduction and is concluded with a short summary. In chapter 2 basics to stereo-vision are given. Readers familiar with the field of stereo can skip this chapter and can direct their interest to the two main chapters 3 and 4 of this work. In chapter 3 the disadvantages of using a standard stereo technique are pointed out and a robust stereo approach is introduced. Tests are made under different noise conditions. Chapter 4 discusses the problem of resolution and scale. A new adaptive matching method is proposed using a correlation scale-space. Experimental results are given on a variety of synthetic and real images at the end of this chapter. Finally chapter 5 concludes this work and gives a brief outlook towards future work.

Chapter 2

Stereo-Vision

2.1 The Human Visual System

The attempt to visualize three-dimensional information began early, when artists tried to visualize 3D objects in paintings by combining perspective and shading. Leonardo da Vinci (1452-1519) was the first to describe binocular parallaxes [GR76]. He noted that the human visual system is able to see behind a small object by fixating this object, thus making it seem opaque. In 1670 the philosopher Malebranche described the geometric relation between the left and the right eye and an object in the scene [GGC85]. In 1838 the first stereoscope was invented by Charles Wheatstone [GR76]. In 1960 Julesz's experiments with random dot stereograms show that three-dimensional information can be seen although the objects cannot be seen in monocular images [Jul71]. So it is important in recent research work to incorporate the knowledge about the biological background of the human visual system in the field of stereo-vision.

The human visual system is one of the most complex senses and is used for everyday purposes, but the process of seeing and understanding objects is not simple. Photons, which reflect from a scene into our eyes, form our visual information. The light is focused by a lens and is projected as image on the retina of the eye. The retina transforms the information in form of electric impulses to the brain. By fixating an object with both eyes a **binocular fusion** of both monocular images takes place. The nearer an object is to the human eye the higher is the relative distance of the projections on the retinas of both eyes. Such a fusion is only possible if the projected images of the object fall on the corresponding areas of the retinas. As an approximation for these corresponding areas a cyclopean eye can be used, which is centered between the left and right eye. Both retinas overlap in the cyclopean eye that the centers of both fovea overlap. All the areas in the retina that overlap in the cyclopean eye are corresponding areas. These areas form a curved surface which is called **horopter** and goes through the optic system of both eyes and the fixation point. The horizontal

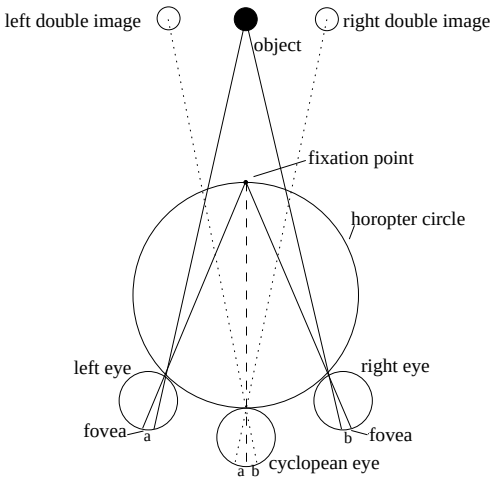


Figure 2.1: Binocular vision.

cross-section is called horopter circle. All scene points which are not projected on the horopter fall on non-corresponding areas on the retina, producing a double image [GGC85]. In Figure 2.1 this principle is shown schematically. Despite intensive research work in this field, the way the binocular integration of the visual information works is not completely known [CF75].

The term stereo used in computer vision and human vision is the recovery of a three-dimensional scene from multiple views. However stereo-vision tries to imitate the human visual system thus using only two views of the same scene. In order to determine three-dimensional information from stereo images, assumptions have to be made concerning geometric relations, which is described in the next section.

2.2 Stereo Image Geometry

The simplest image device is the so-called **pinhole camera**, which has a pinhole through which light enters the camera and forms a two-dimensional image. In a more common model the imaging geometry can be represented by a camera model in which a lens is used for the optical system. The image plane lies behind the lens, thus the image is reversed. A more convenient model rearranges the geometry in such a way that the projected image is not reversed. The world coordinate system X_w, Y_w, Z_w is chosen such that the X_w, Y_w plane coincides with the image plane, with its origin in its center. The center O of the lens lies on the Z_w axis at $Z_w = f$, where f is the focal length of the lens. In order to formulate the relation between a point in the world coordinate system and the corresponding point in the two-dimensional image coordinate system a point $P = (X_w, Y_w, Z_w)$ in the scene is mapped to a point $p' = (x, y, 0)$ in the image

plane. The coordinates for x, y are derived by the following relations:

$$x/f = \frac{X_w}{f - Z_w} \qquad (2.1)$$
$$y/f = \frac{Y_w}{f - Z_w}$$

This image geometry can easily be extended to stereo-vision. If two cameras with the same focal length f are used, the distance between the centers of projection O' and O'' is called the basis b. The angle between the optic axes is 2ϕ, and the two camera coordinate systems are defined by X_L, Y_L, Z_L and X_R, Y_R, Z_R. The Y_L and Y_R axes are parallel to each other. A coordinate system X, Y, Z is defined so that the Z axis bisects the angle between the Z_L and Z_R. The relation between these coordinate systems is illustrated in Figure 2.2. The relations can be derived as follows:

$$\begin{pmatrix} X_L \\ Y_L \\ Z_L \end{pmatrix} = \begin{pmatrix} cos\phi & 0 & sin\phi \\ 0 & 1 & 0 \\ -sin\phi & 0 & cos\phi \end{pmatrix} \begin{pmatrix} X + b/2 - f sin\phi \\ Y \\ Z \end{pmatrix} \qquad (2.2)$$

$$\begin{pmatrix} X_R \\ Y_R \\ Z_R \end{pmatrix} = \begin{pmatrix} cos\phi & 0 & -sin\phi \\ 0 & 1 & 0 \\ sin\phi & 0 & cos\phi \end{pmatrix} \begin{pmatrix} X - b/2 + f sin\phi \\ Y \\ Z \end{pmatrix} \qquad (2.3)$$

If the point $P = (X, Y, Z)$ is observed by two cameras and the corresponding

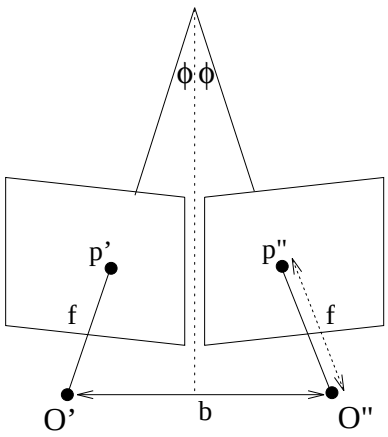


Figure 2.2: Stereo image geometry.

points are $p' = (x_L, y_L)$ and $p'' = (x_R, y_R)$ their relationship is given by

$$x_k = \frac{f X_k}{(f - Z_k)} \qquad (2.4)$$
$$y_k = \frac{f Y_k}{(f - Z_k)} \qquad for \quad k = L, R$$

If two corresponding points $p' = (x_L, y_L)$ and $p'' = (x_R, y_R)$ are given in the two image planes the three-dimensional position of the point $P = (X, Y, Z)$ can be derived from equations (2.2), (2.3) and (2.4).

2.3 Epipolar Geometry

In the field of stereo-vision the principle of **triangulation** is used. Under the assumption of perspective projection, each point p' in the image is the projection of some point in the real world along the ray to the center of projection O'. If the corresponding point in the second image is known, the object point P must also lie on the corresponding ray through that point. This principle is shown schemat-

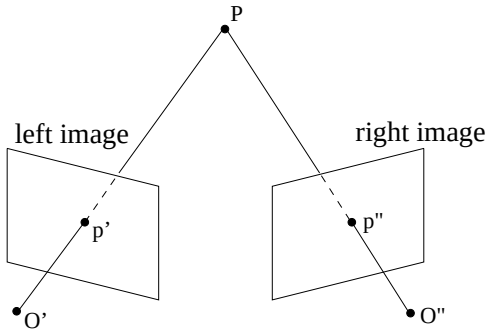


Figure 2.3: Given two images of one object point P from different views, the three-dimensional position can be computed if the point is visible in both images.

ically in Figure 2.3. Having two images from a scene from two different views the three-dimensional position of any object point that is visible in both images can be computed. This object point must lie at the intersection of both rays. The determination of the three-dimensional information of an object point P is called triangulation. The recovery of a point P in the real world using triangulation requires that the image location of the object point in the left image p' is matched to the location of the object point p'' in the right image. It is assumed that every image point is the perspective projection of an object point P, and that each image point in the left image has at most a single unique matching point in the right stereo image, thus excluding transparent objects. The plane going through the two centers of projection O',O" and the object point P is called the **epipolar plane**. The intersection of the epipolar plane and the two image planes results in two straight lines, which are called **epipolar lines** [Wen92]. These relations are depicted in Figure 2.4 schematically. In many applications the two image planes are often chosen to be coplanar and parallel to their baseline (line between the two centers of projection). In order to guarantee such an arrangement, the cameras used are positioned exactly so that the two image planes lie in one plane. Given the case that two stereo images are coplanar and parallel to their baseline

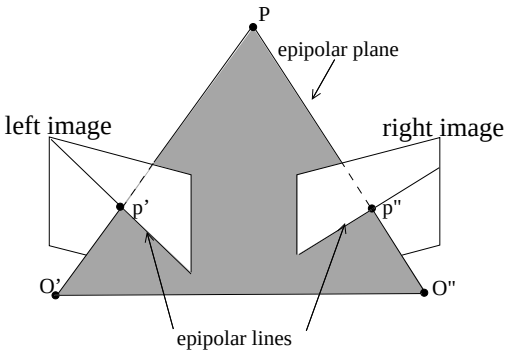


Figure 2.4: The plane built up by the two centers of projection O',O'' and the point P is called the epipolar plane. The intersection of this plane with the two image planes are called epipolar lines.

they are **rectified** [HS93, Bra97]. In rectified images the epipolar lines are the scanlines in both images. The main problem in using the epipolar geometry is that epipolar lines have to be computed with high accuracy using calibration techniques. There are several works dealing with the computation of epipolar geometry. Some methods compute the relative relations between the stereo cameras to determine the epipolar geometry [BWS91, VB91], whereas other techniques calibrate the complete stereo configuration [Tsa85, Fau93]. In the next section basic notations to the calibration process of a stereo system are given.

2.4 Calibration of a Stereo System

The calibration of cameras is considered an important issue in computer vision. **Camera calibration** is the process of determining the internal camera geometry and its optical characteristics and the three-dimensional position and orientation of the camera frame relative to a certain world coordinate system. The parameters of the transformation from the 3D object coordinate system to the computer image coordinates are called **intrinsic parameters**.

These parameters are

- f: effective focal length
- κ_1, κ_2: lens distortion coefficients
- C_x, C_y: computer image coordinates at the origin in the image plane

These parameters determine how the image coordinates of a point are derived, given the spatial position of the point with respect to the camera. The estimation of the geometrical relation between the camera and the scene, between different cameras, is also an important aspect of calibration. The corresponding

parameters that characterize such a geometrical relation are called **extrinsic parameters**. The six extrinsic camera parameters are:

- Ψ, Φ, Θ: rotation parameters
- T_x, T_y, T_z: translation parameters

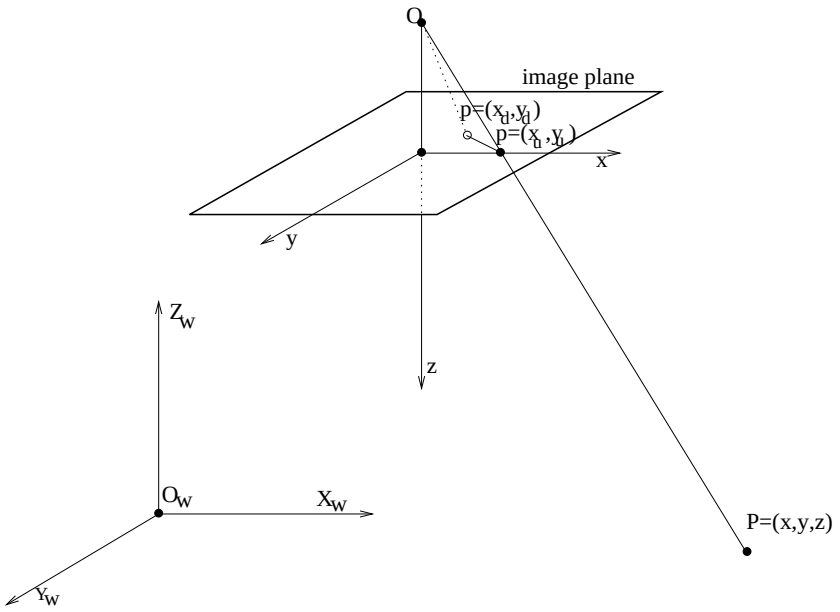


Figure 2.5: Camera geometry with perspective projection and radial lens distortion.

The relation between the two-dimensional image coordinate system and the three-dimensional world coordinate system can be written as:

$$\begin{bmatrix} x - C_x \\ y - C_y \\ f \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} X - T_x \\ Y - T_y \\ Z - T_z \end{bmatrix} \quad , \tag{2.5}$$

where r_{ij} are the coefficients of the rotation matrix $R = R_\Psi R_\Phi R_\Theta$. The relation between the two coordinate systems is visualized in Figure 2.5. If the lens distortion is modeled by the calibration technique it can be distinguished between distorted $p = (x_d, x_d)$ and undistorted image coordinates $p = (x_u, y_u)$. The relation is given by

$$\begin{pmatrix} x_u \\ y_u \end{pmatrix} = \begin{pmatrix} x_d \\ y_d \end{pmatrix} + \begin{pmatrix} x_d(\kappa_1 r^2 + \kappa_2 r^4) \\ y_d(\kappa_1 r^2 + \kappa_2 r^4) \end{pmatrix} \quad , \tag{2.6}$$

where

$$r = \sqrt{{x_d}^2 + {y_d}^2} \ . \tag{2.7}$$

The problem of camera calibration is to compute the camera intrinsic and extrinsic parameters based on a number of points whose object coordinates in the scene are known, and whose image coordinates are measured.

Several calibration techniques are proposed in literature and can be categorized into four major techniques:

- **Techniques involving full scale nonlinear optimization**
  Although the accuracy obtained by these methods is very high the process is very time consuming. An example for this technique is proposed by Faig [Fai75]. He uses 17 unknowns for each image. However, because of the large number of unknowns the accuracy is very good. A very similar approach is made by Sobel, who described a method which calibrates a camera by solving nonlinear equations, where 18 parameters need to be optimized [Sob73]. The problems using this techniques are that a good initial estimate is needed for the nonlinear search. An efficient and accurate camera calibration technique was proposed by Tsai [Tsa85]. In this work off-the-shelf TV cameras and lenses are used. This two-stage technique is aimed at efficient computation of camera extrinsic parameters as well as the effective focal length and radial lens distortion.

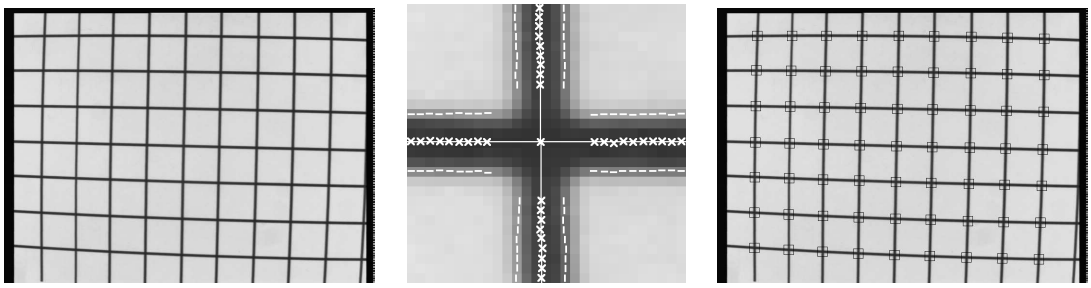
- **Techniques computing perspective transformation matrix first using linear equation solving**
  The main advantage of these methods is that no nonlinear optimization is needed. Although the transformation from the three-dimensional world coordinate system to the two-dimensional image coordinate system is nonlinear, it is linear if lens distortion is ignored. If for a number of points the three-dimensional world coordinates and the corresponding two-dimensional image coordinates are given, the coefficients in the perspective transformation matrix can be computed by least squares solution of an over-determined system of linear equations. For the definition of the perspective transformation matrix see [DH73]. Given the perspective transformation matrix, the camera model parameters can then be computed if needed. One example for this technique is the *DLT* (Direct Linear Transformation), which nowadays is used very often because of its simplicity. The *DLT* was developed by Abdel-Aziz and Karara [AAK71]. The problem using this method is that no lens distortion can be modeled and that the number of unknowns in linear equations is much larger than the actual degrees of freedom and that they are not linearly independent.

- **Two plane method**
  This method was proposed by Martins, Birk and Kelley [MBK81]. Also for

this method only linear equations have to be solved. The relative orientation between camera coordinate system and the world coordinate system is assumed to be known. The number of unknowns is at least 24, 12 for each plane, much larger than the degrees of freedom. A more general technique was proposed by Isaguirre and Pu, where a full nonlinear optimization is needed [IPS85].

- **Geometric techniques**
  This method was developed by Fischler and Bolles [FB86]. They used geometric construction to derive a direct solution for the position and orientation of the camera. Although only linear equations are to be solved the disadvantage of this method is that neither the focal length nor the lens distortion parameters can be computed.

In order to calibrate the cameras, one possibility is to use multiple planar rectilinear grids. To define such a three-dimensional grid, one grid on a plane is moved through the measurement area while taking different images at known positions. One image of a grid can be seen in Figure 2.6 (a). To compute the orientation parameters a calibration plate is placed into the measurement area,



(a) image of the calibration plate (b) detected point (c) marked points

Figure 2.6: Detection of the calibration points in the image.

which consists of black horizontal and vertical lines on a white background. The image coordinates of the calibration points are computed automatically at subpixel accuracy. In order to detect the center of a line in the image, the left and the right border of the line is determined with the help of a profile. A window of the size 30×30 is used to determine one intersection of two lines. Within this window the centers of the horizontal and vertical lines are computed. Since it is not possible to distinguish between a horizontal and vertical line near the intersection of two lines, these values are not taken into account. The location of one calibration point is determined by intersecting the two lines, which are computed

by linear regression [MB95]. The result is the position of one calibration point in sub-pixel accuracy. In Figure 2.6 (b) this principle is shown for one point and in Figure 2.6 (c) all detected points are marked in the image. This procedure is performed for each position of the calibration plate during the transportation through the measurement area. With the image coordinates of these calibration points and the known positions of the scene points, the intrinsic and extrinsic camera parameters can be computed using for instance the method proposed by Tsai, which is described in more detail in [Tsa85, LT86]. With the help of the intrinsic and extrinsic camera parameters the absolute depth information can be computed for two corresponding points in a stereo pair, for which the calibration procedure was performed. The problems which may occur in finding corresponding points in stereo images are described in the next section.

2.5 The Correspondence Problem

The search for the correct match of a point is called the **correspondence problem** and is one of the central and most difficult parts of the stereo problem [MP79, TWK87, BF82b]. Several algorithms have been published for finding correspondences between images like the correlation method [MP76, LM90, SHC90], the correspondence method [Gri85, HJS90], or the phase difference method [JJT91]. The difficulty in stereo-vision is to find corresponding points in both stereo images, so that each point in a pair of points is the image of the same point in space. Ambiguous correspondence between points in the two stereo images may lead to

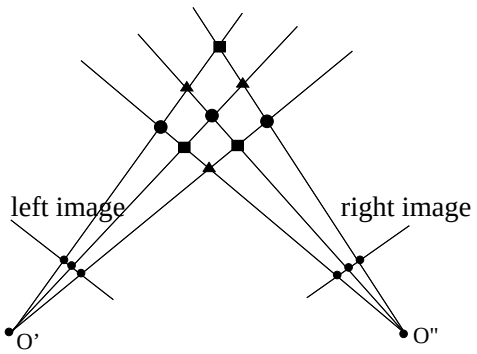


Figure 2.7: Ambiguous correspondence: If the correspondence between multiple points in two images is ambiguous, triangulation may lead to several different interpretations of the scene.

several different interpretations of the scene [Jul71]. The determination of the three-dimensional locations of the scene points using triangulation is performed such that each point in a pair of matched points is the image of the same object point. If the correspondence between multiple points in two images is ambiguous, triangulation may lead to several different interpretations of the scene. The

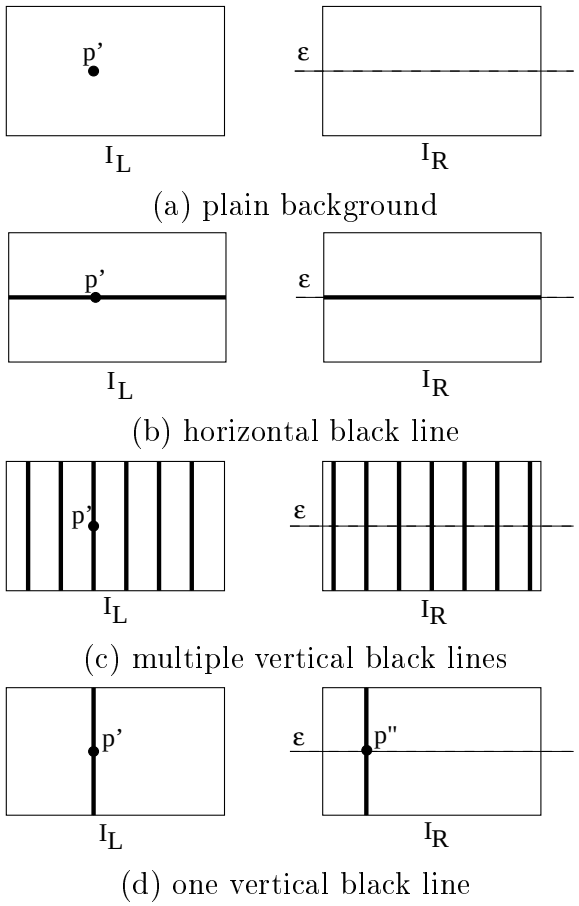


Figure 2.8: Correspondence problem: For one point p' in the left I_L image the corresponding point p" has to be found along the epipolar line ε in the right image I_R. It is not possible for (a)-(c) to find a unique solution, whereas for (d) the corresponding partner p'' could be detected.

problem of finding correspondences in a stereo pair is an ill-posed problem[1] even in quite simple scenes as depicted in Figure 2.7. Three object points in a row (circles) are observed from two different positions. Each of the three points in the left view can match any of the three projections of the right view. There are 9 possible matches between the two pictures, but 6 of them are not correct. Possible wrong matches are shown as squares and triangles.

So one of the most difficult problems in stereo-vision is to determine the correct corresponding points for a given stereo pair. It is not always easy for

[1]Hadamard stated three criteria for a well-posed mathematical problem: (i) a solution should exist; (ii) the solution should be unique; and (iii) the solution should depend continuously on the input data.

the human visual system to detect corresponding points in stereo images, if they contain insufficient information. As an example Figure 2.8 depicts four different stereo pairs, where the epipolar lines are the scanlines in both images. The stereo pair visualized in Figure 2.8 (a) contains no information, thus for a given point p' no corresponding point p'' can be found. In Figure 2.8 (b) one horizontal line is added to the images. Although information is present in both images, no correspondence can be established. In example (c) the images consist of multiple vertical lines. No unique solution can be found for this example either, since

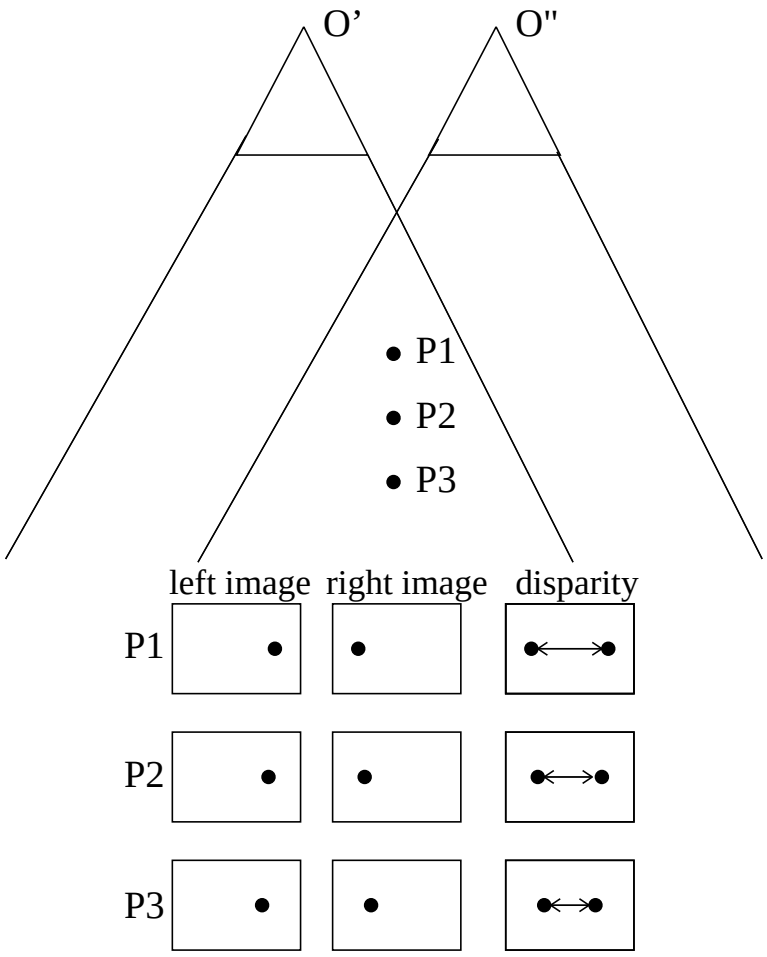


Figure 2.9: Disparity: Three different object points are observed by two cameras. The distance between the two different projections of one object point P is called disparity.

it is not clear to which vertical line p' corresponds, whereas for example (d) a unique solution can be found. This example shows that it depends on the information in the scene whether the correspondence problem can be solved or not. The three-dimensional position of the scene points can be determined only for a corresponding pair of points.

The distance between two corresponding points is related directly to the depth information. In Figure 2.9 this principle is shown schematically. Two parallel aligned cameras observe three different object points one after the other, which have different depth positions. For each exposure the projected image points in the left and in the right image plane are shown in the first two columns. In the third column the superimposed image planes are depicted. It can be seen that there is a distance between these two projections. This distance is called the **disparity** between the corresponding points. The nearer one object point is to the camera the bigger is the distance between its two projections. In order to solve the correspondence problem for each point in the left image corresponding partners have to be found. Each disparity value for the left image is stored, which results in a **disparity map**. A disparity map is an image where the disparity is represented by the gray-level. If for each pixel in the left stereo image disparity information is available, the disparity map is called a **dense disparity map**. Due to occlusions, highlights, and depth discontinuities many problems occur which cannot be solved in the common way.

2.6 Occlusion

Occlusion detection is an important problem in stereo-vision. The presence of occlusion complicates the matching process in the generation of 3D data. The three-dimensional information can only be computed for points which are visible in both stereo images. Thus a matching process taking into account occlusions is necessary to accurately recover the three-dimensional structure from two-dimensional stereo images. There are several works which propose some new computational frameworks for stereo matching incorporating occlusion information [CN91, JM92]. The problem with occlusion plays an important role in stereo matching, because some basic assumptions and the fundamental trian-

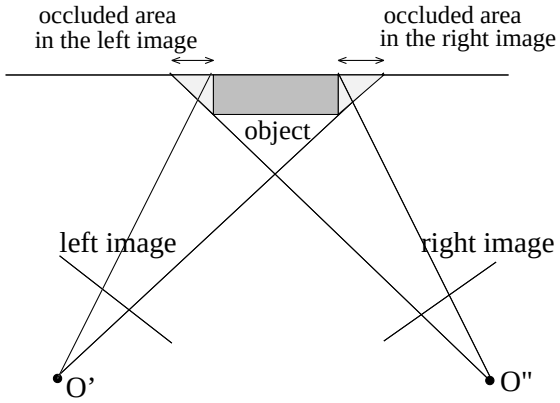


Figure 2.10: Occlusion: Points which are visible from one view may not be visible from another viewpoint.

gulation geometry are invalid in occluded regions. One cause for the absence of corresponding points is the fact that objects in the scene may occlude differently in the two stereo images.

In Figure 2.10 the principle is visualized schematically. One box on a flat ground is observed by two cameras. It can be seen that there are occlusion areas for the left and the right image. For these points no depth information can be computed. In order to avoid ambiguous correspondences and to detect occlusion areas in stereo images the information about the geometric relations or apriori knowledge about objects are used. Some important constraints in stereo-vision, which are also used in this work, are described in the next section.

2.7 Constraints in Stereo-Vision

In general there are no global rules for finding corresponding points in a stereo pair. Each stereo algorithm uses several assumptions about the image geometry or about the objects in the scene. By establishing correspondences in stereo images it happens that the results are ambiguous. Thus for one point in the left image more than one candidate point exists in the right image. One possibility to reduce ambiguity is to have apriori knowledge about the scene geometry or about the objects themselves. In the following subsections an overview of the constraints used in this work are given.

2.7.1 Geometric Constraints

Epipolar constraint:
Generally for a given image point in the left image the search for a corresponding point in the right stereo image has to be made over the complete image. Such a two-dimensional search is not necessary when using the **epipolar constraint**. For a given image point, its corresponding point in the right image has to be searched along the line that is the projection of the line through the given image point and its center of projection, the epipolar line.

Photometric constraint:
Corresponding points in a stereo pair are considered to have the same intensity values. This constraint holds for objects with nearly Lambertian surfaces and using a parallel aligned camera arrangement providing only small changes in the surface orientation. Moreover the ratio between baseline and depth should be kept small [IB92]. Using this constraint it is evident that there must be ambient illumination for the acquisition otherwise reflections and highlights will occur on the surfaces of the objects. This constraint can also be used for color images where two corresponding pixels in a stereo pair have the same color values.

Similarity constraint:
If a line or an edge is observed from two different views the two projections should have similar characteristics like orientation and length, thus a horizontal line can never be matched with a vertical one. Problems occur if one line in the left image is split into two lines in the right image or disappears because of occlusions.

Uniqueness constraint:
One pixel in the left image can only be matched to one pixel in the right image [Mar82]. This constraint does not hold in the case of ambiguous stereo where one point in the left image can be matched to two points in the right image. An example of ambiguous stereo is given in section 2.5.

2.7.2 Object-Based Constraints

Disparity limit:
In general for one point in the left stereo image the corresponding point is searched along the epipolar line in the right image, if the assumption is made that the disparity value for two corresponding points lying on one object is less than a given threshold [MP76]. This threshold defines the minimum distance between camera and objects in the scene.

Continuity constraint:
The disparity values computed for a complete image vary only smoothly over the image. This constraint is used for instance for determining the elevation models of terrains out of satellite data. This constraint does not hold for object boundaries and strong depth edges in the scene.

Occlusion constraint:
If an object is observed by two cameras there occur occluded areas which can not be seen by the left or by the right camera (see section 2.6). A vertical depth discontinuity observed with the left camera results in an occlusion in the right camera [GLA92].

2.8 Area-Based vs. Feature-Based Stereo

The methods for stereo matching can be grouped into two major categories:

- **Feature-Based Stereo**
  Feature-based stereo techniques match features in the left image to those in the right image. Features are selected in the image, such as, for example, edge points or edge segments. Feature-based techniques have the advantage of being less sensitive to photometric variations and being faster than

area-based techniques, because there are fewer candidates for matching corresponding points. Most approaches still use edges as features. Others, for example, use regions [CVSG89, LCK93] or topological structures [Fle92]. Although feature-based stereo techniques solve the correspondence problem in a fast and accurate way, the number of corresponding structures is low because of the small number of features.

- **Area-Based Stereo**
  Area-based stereo techniques find corresponding points on the basis of the similarity of the corresponding areas in a stereo pair. For one point in a stereo image the corresponding point is sought on the basis of the similarity of the neighboring regions. This neighboring region is called **window**. A similarity measure is then applied to search for a corresponding point with a matching neighborhood in the other image. Area-based techniques have the disadvantage of being sensitive to photometric variations during the image acquisition process and of being sensitive to distortions as a result of changing the viewing position. This sensitivity is due to direct comparison of intensity values in the stereo images.

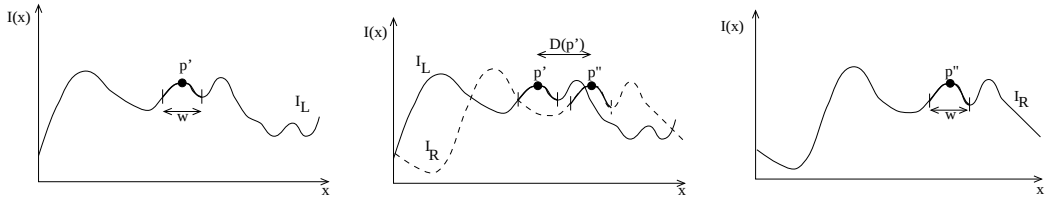
The area-based methods, however, have some advantages over feature-based methods. Using area-based methods a dense disparity map can be estimated directly and the performance of the algorithm does not rely on more or less reliable features. Extracting robust features in natural scenes can sometimes be very difficult and time-consuming. Moreover the algorithms can use the whole information of the images without loss through image preprocessing. The main causes for different intensity images that lead to difficult matching in area-based methods include photometric effects, occlusions, sensor and discretization noise. Thus the area-based methods are efficient and useful compared with other methods. In this work the area-based stereo technique is used as basis for the matching process between stereo images. In the next section some basic notations regarding similarity measures used for area-based techniques are described. At the end of the chapter the problems are noted which occur by using simple area-based stereo techniques.

2.9 Correlation between two Signals

The main goal in stereo-vision is to solve the correspondence problem between two given signals. For two one-dimensional signals $I_L(x)$ and $I_R(x)$, which are different projections of one signal in the scene, the corresponding points have to be found. A common characteristic among stereo matching methods is that two matching processes run separately (from the left image to the right and from the right to the left) and benefit a little from each other. This matching process is

known under the term bidirectional matching method. In this work the stereo matching process concentrates on the matching from the left to the right image.

If for a point p' of the left signal the corresponding point p'' in the right signal is searched a certain region w around p' is taken into account. The region which



(a) left signal (b) both signals (c) right signal

Figure 2.11: Disparity $D(p')$ between two signals $I_L(x)$ and $I_R(x)$ at position p'.

is most similar to the region in $I_L(x)$ is supposed to have the corresponding point in its center.

In order to compute the disparity value for each point in the left image the correlation between light intensities is used as similarity measure. For I_L (left stereo image) and I_R (right stereo image) the discrete definition of the correlation C between two regions of the size w in the one-dimensional case can be written as:

$$C(x_L, x_R, n) = \frac{\sigma^2_{LR}(x_L, x_R, n)}{\sqrt{\sigma^2_L(x_L, n)\sigma^2_R(x_R, n)}}, \tag{2.8}$$

where

$$\sigma^2_k(x, n) = \sum_{i=-n}^{n} \frac{[I_k(x+i) - \mu_k(x, n)]^2}{(2n+1)} \quad k = L, R \tag{2.9}$$

are the standard deviations in the two images and

$$\mu_k(x, n) = \sum_{i=-n}^{n} \frac{I_k(x+i)}{(2n+1)} \quad k = L, R \tag{2.10}$$

are the local mean values and

$$\sigma^2_{LR}(x_L, x_R, n) = \sum_{i=-n}^{n} \frac{[I_L(x_L+i) - \mu_L(x_L, n)][I_R(x_R+i) - \mu_R(x_R, n)]}{(2n+1)} \tag{2.11}$$

is the covariance between the two regions. The correlation $C(x_L, x_R, n)$ between two regions for a given n can be written as:

$$C(x_L, x_R, n) = \frac{\sum_{i=-n}^{n} [I_L(x_L+i) - \mu_L(x_L, n)][I_R(x_R+i) - \mu_R(x_R, n)]}{\sqrt{\sum_{i=-n}^{n} [I_L(x_L+i) - \mu_L(x_L, n)]^2 \sum_{i=-n}^{n} [I_R(x_R+i) - \mu_R(x_R, n)]^2}}. \tag{2.12}$$

Now the correlation can be formulated in a continuous way for later use. The local mean values and the standard deviations from equation (2.8) in the one-dimensional case are defined for a given region w as

$$\sigma_k^2(x,w) \;=\; \frac{1}{w}\int_{\xi=-w/2}^{w/2} [I_k(x+\xi)-\mu(x,w)]^2\; d\xi \quad k=L,R \;\;, \tag{2.13}$$

with

$$\mu_k(x,w) \;=\; \frac{1}{w}\int_{\xi=-w/2}^{w/2} I_k(x+\xi)\; d\xi \quad k=L,R \;\;. \tag{2.14}$$

The covariance between two regions of size w is

$$\sigma_{LR}^2(x_L,x_R,w) \;=\; \frac{1}{w}\int_{\xi=-w/2}^{w/2} [I_L(x_L+\xi)-\mu_L(x_L,w)][I_R(x_R+\xi)-\mu_R(x_R,w)]\; d\xi. \tag{2.15}$$

The correlation $C(x_L,x_R,w)$ between two regions of size w in the one-dimensional case can be written in a continuous way as:

$$C(x_L,x_R,w) \;=\; \frac{\int_{\xi=-w/2}^{w/2} [I_L(x_L+\xi)-\mu_L(x_L,w)][I_R(x_R+\xi)-\mu_R(x_R,w)]\; d\xi}{\sqrt{\int_{\xi=-w/2}^{w/2} [I_L(x_L+\xi)-\mu_L(x_L,w)]^2\; d\xi \int_{\xi=-w/2}^{w/2} [I_R(x_R+\xi)-\mu_R(x_R,w)]^2\; d\xi}}. \tag{2.16}$$

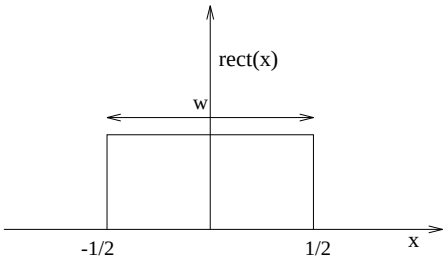


Figure 2.12: $rect(x)$

The correlation for infinite regions is defined by using a binary function $rect(x)$ (Figure 2.12) which is defined as

$$rect(x) \;=\; \begin{cases} 1 & |x| \le \frac{1}{2} \\ 0 & elsewhere \end{cases} \;\;, \tag{2.17}$$

and let

$$\delta_{1/w}(x) \; = 1/w \; rect(x/w), \qquad w > 0. \tag{2.18}$$

So $\delta_{1/w}$ can be equivalently rewritten as

$$\delta_{1/w}(x) \; = \begin{cases} \frac{1}{w} & |x| \leq \frac{w}{2} \\ 0 & elsewhere \end{cases}, \tag{2.19}$$

thus $\delta_{1/w}(x)$ is zero outside the region $|x| \leq w/2$ and has constant value $1/w$ inside that region and it follows that

$$\int_{\xi=-\infty}^{\infty} \delta_{1/w}(\xi) d\xi \; = \; 1, \tag{2.20}$$

so that

$$\int_{\xi=-\infty}^{\infty} f(x+\xi)\delta_{1/w}(\xi) \; d\xi \; = \; \frac{1}{w} \int_{\xi=-w/2}^{w/2} f(x+\xi) \; dx \tag{2.21}$$

is just the average of $f(x)$ over a region w centered at the origin for any w, which is depicted in Figure 2.13. For a more detailed explanation see [Ros84]. Using (2.21) the correlation C can be rewritten as an infinite function as:

$$C(x_L, x_R, w) \; = \; \frac{\int_{\xi=-\infty}^{\infty} \delta_{1/w}(\xi)[I_L(x_L+\xi) - \mu_L(x_L,w)][I_R(x_R+\xi) - \mu_R(x_R,w)]d\xi}{\sqrt{\int_{\xi=-\infty}^{\infty} \delta_{1/w}(\xi)[I_L(x_L+\xi) - \mu_L(x_L,w)]^2 d\xi \int_{\xi=-\infty}^{\infty} \delta_{1/w}(\xi)[I_R(x_R+\xi) - \mu_R(x_R,w)]^2 d\xi}}. \tag{2.22}$$

For a given size of $w > 0$ it can be shown that (2.16) is equal to (2.22).

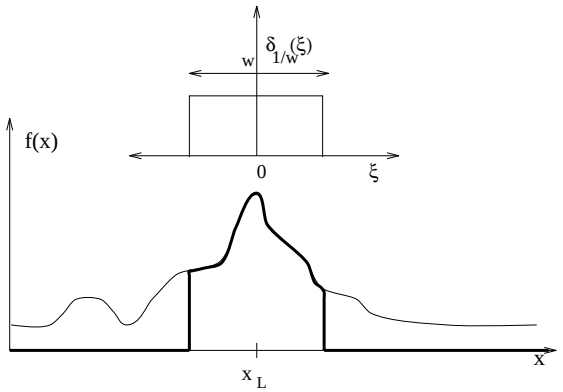


Figure 2.13: $\delta_{1/w}(x)$ for $f(x + x_L)$.

For an image $I : \mathbb{R} \mapsto [0,1]$ and $\mu : \mathbb{R} \times \mathbb{R} \mapsto [0,1]$, the mean $\mu(x,w)$ at location x and width w is defined as

$$\mu(x,w) \ = \frac{1}{w} \int_{\xi=-w/2}^{w/2} I(x+\xi)\ d\xi \tag{2.23}$$

Using the function $\delta_{1/w}$ the local mean can be rewritten so that

$$\mu(x,w) \ = \ \frac{1}{w} \int_{\xi=-w/2}^{w/2} I(x+\xi)\ d\xi \ = \int_{\xi=-\infty}^{\infty} I(x+\xi)\delta_{1/w}(\xi)\ d\xi \tag{2.24}$$

A convolution with the operator * is defined as

$$f(x) * g(x) \ = \int_{-\infty}^{\infty} g(x-\xi)f(\xi)\ d\xi \tag{2.25}$$

With (2.23) and (2.25) the mean $\mu(x,w)$ can be written as a convolution:

$$\mu(x,w) \ = \int_{-\infty}^{\infty} I(x-\xi)\delta_{1/w}(\xi)\ d\xi \ = I(x) * \delta_{1/w}(x)\ , \tag{2.26}$$

where $\delta_{1/w}$ is the convolution kernel. For $\sigma^2(x,w)$ the convolution with the function $\delta_{1/w}$ can be written as

$$\sigma^2(x,w) \ = \ [I(x) - \mu(x,w)]^2 * \delta_{1/w} = I^2(x) * \delta_{1/w} - \mu^2(x,w) \tag{2.27}$$

and the covariance is defined as

$$\begin{aligned} \sigma^2_{LR}(x_L,x_R,w) \ = \ & [I_L(x_L) - \mu_L(x_L,w)][I_R(x_R) - \mu_R(x_R,w)] * \delta_{1/w} \ = \\ & [I_L(x_L)I_R(x_R)] * \delta_{1/w} - \mu_L(x_L,w)\mu_R(x_R,w) \end{aligned} \tag{2.28}$$

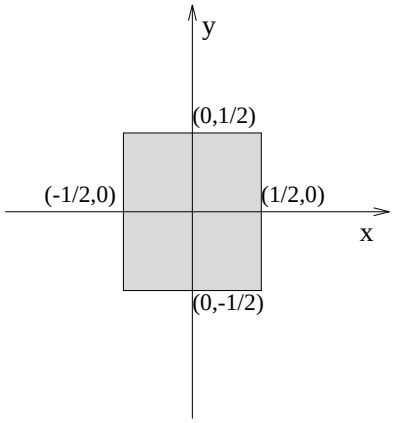


Figure 2.14: rect(x,y)

The correlation can be written as convolution with the kernel $\delta_{1/w}$ as

$$C(x_L, x_R, w) = \frac{[I_L(x_L)I_R(x_R)] * \delta_{1/w} - \mu_L(x_L, w)\mu_R(x_R, w)}{\sqrt{[I_L^2(x_L) * \delta_{1/w} - \mu_L^2(x_L, w)][I_R^2(x_R) * \delta_{1/w} - \mu_R^2(x_R, w)]}}. \tag{2.29}$$

In order to define the correlation for two-dimensional regions equation (2.18) is extended to

$$\delta_{1/w^2}(x, y) = 1/w^2 \; rect(x/w, y/w), \qquad w > 0 \tag{2.30}$$

with

$$rect(x, y) = \begin{cases} 1 & |x| \leq \frac{1}{2} \; , \; |y| \leq \frac{1}{2} \\ 0 & elsewhere \end{cases} \quad . \tag{2.31}$$

The two-dimensional function $rect(x, y)$ is shown in Figure 2.14. The correlation for two-dimensional regions with the kernel δ_{1/w^2} can be written as:

$$C(x_L, y_L, x_R, y_R, w) = \frac{[I_L(x_L, y_L)I_R(x_R, y_R)] * \delta_{1/w^2} - \mu_L(x_L, y_L, w)\mu_R(x_R, y_R, w)}{\sqrt{[I_L^2(x_L, y_L) * \delta_{1/w^2} - \mu_L^2(x_L, y_L, w)][I_R^2(x_R, y_R) * \delta_{1/w^2} - \mu_R^2(x_R, y_R, w)]}}, \tag{2.32}$$

where the local mean values are

$$\mu_k(x, y, w) = I_k(x, y) * \delta_{1/w^2} \qquad k = L, R \; . \tag{2.33}$$

In Figure 2.15 this principle is depicted. Within a given region of the size w the

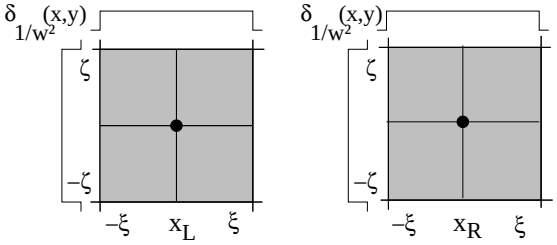


Figure 2.15: Correlation with function $\delta_{1/w^2}(x, y)$

intensity values are taken into account, whereas the other values are weighted to zero. In the next section a simple area-based stereo algorithm is explained on a synthetic stereo pair.

2.10 Standard Area-Based Stereo

In order to determine corresponding points in parallel aligned stereo images equation (2.32) is used as similarity function.

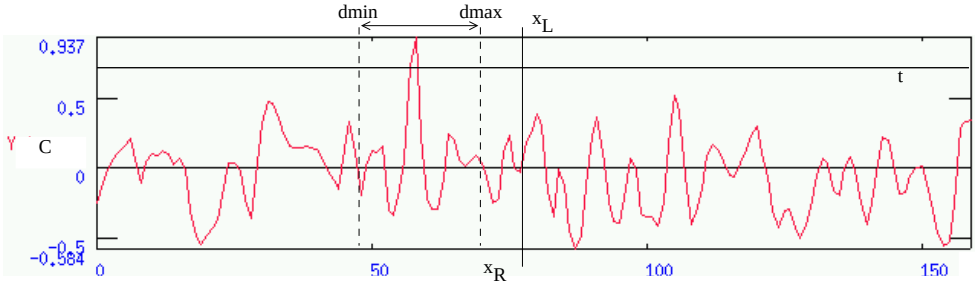


Figure 2.16: For one point in the left image $I_L(77, 40)$ the correlation function for the complete scanline in the right image I_R is computed using a fixed window size $w = 7$. The maximum of function C is determined at position x_R=55.

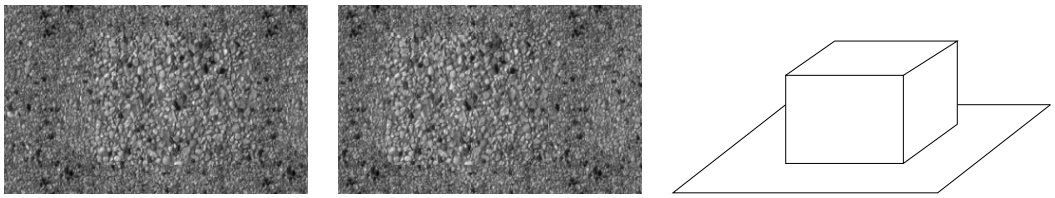
In Figure 2.16 this principle is depicted for one point in the left stereo image $I_L(x_L, y_L)$. For this point the corresponding point is determined in the right stereo image. For a given point in the left image the correlation function $C(x_L, y_L, x_R, y_R, w)$ is computed over the complete scanline $x_R \in [0, 155]$ in the right image $I_R(x_R, y_R)$, where $y_R = y_L$. Using parallel aligned cameras the corresponding partner is supposed to be along the scanline in the right image. Since it is impossible to have negative disparity values using this constraint, the correlation function C has only to be computed to the position of x_L in the right image. If the disparity value for two corresponding points is zero ($x_R = x_L$) the distance to the point in the real world is infinite. For each pair of points the correlation C is computed using square windows. The result is a correlation function the maximum of which is supposed to be the corresponding point, which can be seen in Figure 2.16. The disparity between x_L and x_R is computed by

$$D(x_L, y_L) = \begin{cases} |x_L - x_R| & x_R = argmax\{C(x_L, y_L, x_R, y_R, w)\} > T \\ -1 & else \end{cases}, \qquad (2.34)$$

where T defines the threshold accepting the corresponding point. The fact that the maximum of C is below this threshold T can come from occlusion, highlights or depth discontinuities. In the case of occlusion an occlusion map can be generated separately by reprojecting the computed disparity values from the left to the right image and determining the error. If the error is above a certain threshold an occlusion is detected at this point. The maximum of the correlation function is accepted if it is above the threshold T. In this case the position of the unique maximum defines the corresponding point in $I_R(x_R, y_R)$. Furthermore algorithms can be speeded up by using the disparity limit defined by the disparity minimum *dmin* and maximum *dmax*. In this case the correlation function only has to be computed between $x_L - dmax$ and $x_L - dmin$.

2.11 Discussion

Although stereo techniques have achieved a great progress, some problems have not yet been solved. One of the most important reasons is that depth discontinuities and occlusions are often not explicitly treated in many matching algorithms.



(a) left image (b) right image (c) box in the scene

Figure 2.17: Synthetic stereo pair of a box on a flat ground with natural texture added on the surface.

Often stereo algorithms utilize the constraints of uniqueness, smoothness and ordering to simplify the matching process. However these constraints are invalid assumptions in occluded regions.

The standard stereo technique described in the previous section is applied on a synthetic stereo pair by using a fixed window size w. The stereo pair consists of a box on a plane ground with natural texture added on the surface, which is shown in Figure 2.17. Figure 2.18 depicts the result of the standard area based method. In addition to the gray-level of the left image I_L the disparity value $D(x, y)$ is also available. It can be seen that for regions where the correlation values are below the threshold T, no disparity value is available. These mismatches can occur for the following reasons:

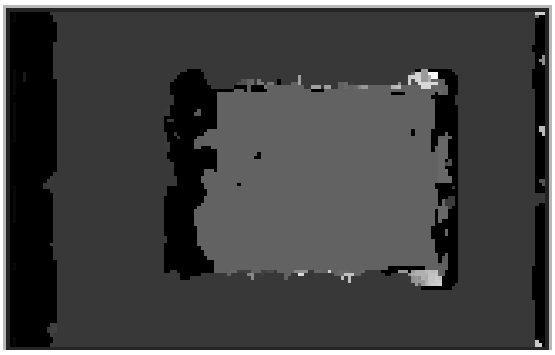


Figure 2.18: Disparity Map: Black areas define points where no correspondence can be established.

- **homogeneous regions**
  If the stereo pair contains too little gray-level variations it is impossible to correlate regions. In order to solve the correspondence problem it is important to have enough information in the regions which are currently under consideration.

- **specular reflections**
  Highlights on the surface of objects can complicate the correspondence establishment. A highlight on the surface changes its position when observed from two different views, thus resulting in a false match.

- **occlusion**
  The correspondence problem can only be solved if one object point in the scene can be seen by two cameras. If there are occluded regions in the stereo pair it depends on the size of the search window whether a corresponding partner can be found. In this case the depth value is smoothed. The problem with occlusions plays an important role in stereo matching, because the uniqueness and smoothness constraint are invalid in occluded regions. This simple area-based technique usually has troubles in the neighborhood of occluded regions, which leads to wrong matches [LB95].

- **different objects**
  Vision algorithms must be able to combine data while simultaneously discriminating between data that should be kept distinct, such as outliers and data from other regions. All points in a window used in standard correlation should have the same disparity, otherwise the estimate will be incorrect.

- **outliers**
  Outliers can either be large measurement errors (incorrect matches) or data points belonging to different objects [ML96].

- **windowing problem**
  Most of the problems mentioned above depend on the size of the search window. The problem with homogeneous regions can be tackled by enlarging the window, whereas in regions of occluded areas a large search window gives wrong estimates of depth. Also the problem with outliers cannot be solved satisfactory by changing the window size. In general, the size of the search window plays an important role in stereo-vision since it influences the accuracy of the solution and the computation time. There is always a trade-off in the selection of a window size. If the window is too small, the estimates are unreliable since they are based on a small number of data points in a window. If the size of the window is large, the risk of encompassing data that belong to several different objects or surfaces is enlarged.

The listed items are the main problems to be solved by stereo algorithms.

In this chapter basic notations to stereo-vision were given and an area based stereo algorithm was described using the correlation as similarity measure. The main goal to be solved by stereo algorithms is to establish correspondences in a stereo pair, which is called the correspondence problem. First the correlation was defined for discrete and then for continuous signals, which when reformulated can be rewritten as convolution. This provides an efficient computation and is needed for later use for a scale space version of the correlation.

Chapter 3

Robust Correlation

3.1 Introduction

Robust statistics is concerned with the fact that assumptions made in statistics are at most approximations to reality. One reason is the occurrence of **gross errors**, such as Salt & Pepper noise, which do not follow any distributions. These errors are usually **outliers**, which are far away from the bulk of the data, and are dangerous for many classical statistical procedures. While the problem of outliers is well known and is as old as statistical procedures and while a number of statisticians (such as S. Newcomb [New86], K. Pearson [PS36] et al.) were clearly aware of it, it has only been in the last years that attempts have been made to formalize the problem beyond limited and ad hoc procedures towards a theory of robustness. Now there is a great variety of approaches towards the robustness problem which can be found in [Hub81]. The theory of robustness plays an important role in organizing and reducing information about the behavior of statistical procedures to a manageable form. Robust window operators [BBW88, KKM$^+$89, MMRK91, SS92] have been brought into computer vision as an answer to the problems encountered by standard least squares methods in windows containing outliers or more than one statistical population.

The maximum permitted percentage of the outliers, such that they have only limited or no influence, will be measured by the concept of the **breakdown point** [HRRS86]. The breakdown point of an estimator is determined by the smallest portion of outliers in the data set at which the estimation procedure can produce an arbitrarily wrong estimate. It is especially the aspect of the breakdown point which is usually emphasized in the design of robust estimators for vision algorithms. This indicates that the use of robust estimators in computer vision is often more for rejecting outliers than for optimally estimating parameters of the models in the case of non-Gaussian data distributions that may arise from the nature of the physical data. Some robust local operators [BBW88] can theoretically achieve the maximum breakdown point of 50% which means that the estimate

remains unchanged if less than half of the data are outliers. Its oldest definition was restricted to one-dimensional estimation of location by Hodges [J.L67], whereas Hampel gave a more general formulation [Ham71]. Donoho and Huber introduced a single finite-sample version of the breakdown point [DH85].

Stereo computation is just one of the vision problems where the presence of outliers cannot be neglected. Most standard algorithms make unrealistic assumptions about the noise distributions leading to erroneous results which cannot be corrected in the subsequent processing stages. As an example W. Luo and H. Maitre tried to use a surface model to correct and fit disparity data in stereo-vision as a post-processing step after non-robust stereo computation [LM90]. By that time the result may already contain errors that cannot be removed. The major difference to the approach described in this chapter is that the principle of the robust estimators is integrated **directly** in the computation of stereo.

3.2 Outliers in Stereo-Vision

When depth maps are obtained by stereo techniques, any false correspondences produced by the matching algorithm will induce outliers in the disparity space. This poses a challenging problem to any surface reconstruction algorithm. For example, all the points in a window used in standard correlation should have the same disparity, otherwise the estimate will be incorrect. Hence a window containing data points from different surfaces or objects produces virtual outliers.

well-behaved noise	
• sensor noise	additive noise is introduced to the left and right stereo image
• depth discontinuities	disparity values are smoothed
• aliasing effects	disparity values are not computed in sub-pixel accuracy
outliers	
• specular reflections	highlights are matched instead of the object
• occluded areas	occluded regions can only be seen in one view $\rightarrow$ no correspondence possible
• repetition of texture	wrong region is matched, because of multiple maxima in the correlation function

Table 3.1: Different noise conditions.

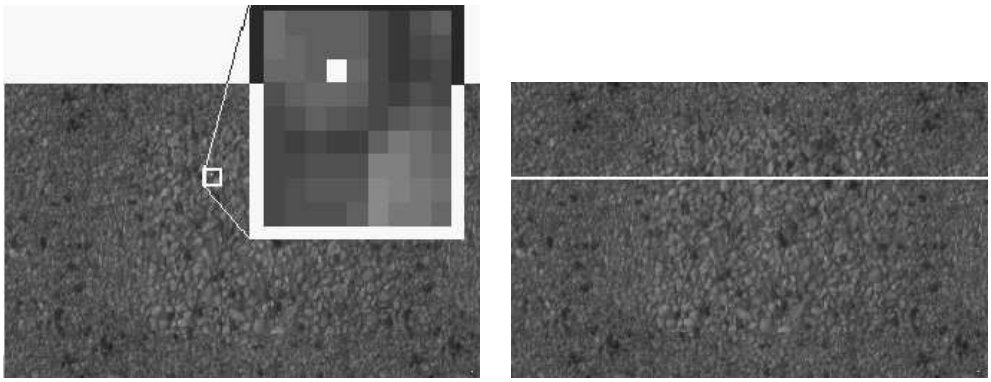
The majority of the points in the observed region represent proper data, whereas the rest are classified as outliers belonging to a different object or surface. The

sources of outliers are not confined to sensor defects or mismatches in binocular stereo. There are several reasons for wrong disparity values in stereo data.

It must be distinguished between

- **noise** that is **well-behaved** in a distributional sense and
- **outliers**, which are wrong matches.

In Table 3.1 examples of these two categories are shown. The major problem with the standard stereo computation based on correlation is that due to occlusions, highlights, discontinuities, and noise, one cannot avoid erroneous matches leading to incorrect estimation of disparities and consequently shape. The reason is that the standard correlation follows the least squared argument which is very sensitive to outliers in the data set. A single outlier which is sufficiently far away can ruin a least squared analysis completely. In order to show how sensitive the standard correlation technique is to outliers, a simple experiment is performed. For a given



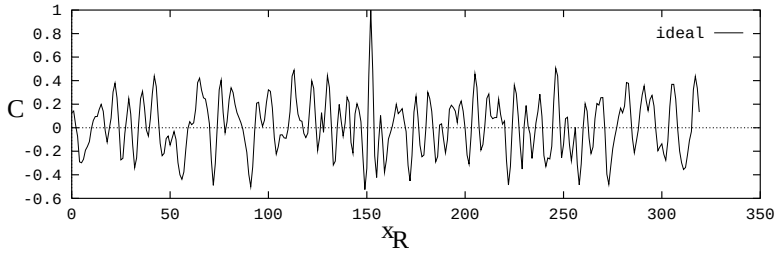
(a) with one outlier corrupted copy of the left stereo image I_L

(b) original left image I_L

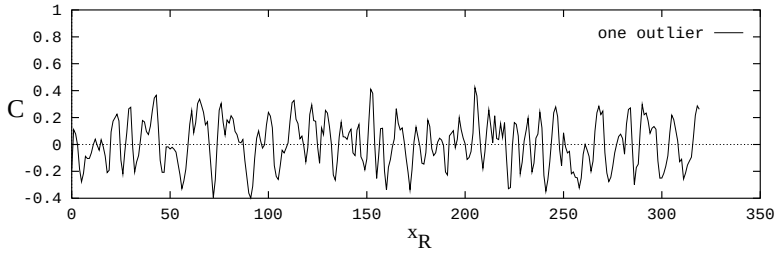
Figure 3.1: Test performed on a corrupted copy of the left stereo image: (a) zoomed region of a 9×9 window with one outlier; (b) the correlation values are computed along the complete scanline.

window of the size of 9×9 the correlation function C (Eq. (2.32)) is computed for one point in the left stereo image along the complete scanline of the clean and with one outlier corrupted left image. Figure 3.1 (a) depicts the corrupted image with the zoomed region including the wrong data point and in (b) the original left stereo image is visualized. The selected 9×9 window of the left image is correlated with 9×9 windows along the epipolar scanline of both the original left and the left image corrupted with one outlier. In Figure 3.2 (a) and (b) it is illustrated that only one outlier can corrupt the correlation function dramatically,

which results in a wrong match of the corresponding point. In the next section a robust version of calculating the correlation is proposed.



(a) correlation function for clean data



(b) correlation function with one outlier in the data set

Figure 3.2: Different correlation functions C computed at the same position in the left image for the clean and with one outlier corrupted window.

3.3 Robust Computation of Correlation

It is becoming clear that the algorithms should be designed such that they can cope with various types of noise. Thus, the idea is to incorporate the principles of robustness directly into the computation, in this case—the computation of correlation. The approach being developed not only tolerates a significant number of outliers but also gives robust results in the presence of depth discontinuities.

3.3.1 M-estimator

In Schunck [Sch90] it is argued that visual perception is a fundamental problem in discrimination: data must be combined with similar data (having the same property) and outliers must be rejected. In other words, vision algorithms must be able to combine data of the same class while simultaneously discriminating between data of different classes. The estimators that remain stable in the presence of various types of noise and can tolerate a certain portion of outliers are known

under the generic name of **robust estimators** [Hub81, RL87]. **M-estimators** (maximum likelihood type estimate), for example, tackle the problems by either rejecting outliers from the calculation or down-weighting their influence on the final result. This is achieved by first computing an initial estimate and then refining it by repeated re-weighting of the data points.

In M-estimator [HJL$^+$89, Hub81, MB91], the solution of a vector θ is given by a minimization problem for a given region $|\xi| \leq w/2$ of the following form

$$\min_{\theta} \int_{\xi=-w/2}^{w/2} \rho(x(\xi) - \theta) d\xi \tag{3.1}$$

or by an implicit equation

$$\int_{\xi=-w/2}^{w/2} \psi(x(\xi) - \theta) d\xi = 0 \tag{3.2}$$

where $x(\xi)$ is the residual error of what is actually observed and what is estimated. ρ is an arbitrary non-negative monotonically increasing function, which is called the **objective function**; $\psi(x(\xi) - \theta)$ is a derivative of $\rho(x(\xi) - \theta)$ with respect to θ and is called an M-estimator

$$\psi(x(\xi) - \theta) = \frac{\partial}{\partial \theta} \rho(x(\xi) - \theta) \ . \tag{3.3}$$

Equation (3.2) can equivalently be written as

$$\int_{\xi=-w/2}^{w/2} \tau(\xi)(x(\xi) - \theta) d\xi \ = \ 0 \ , \tag{3.4}$$

with

$$\tau(\xi) \ = \ \frac{\psi(x(\xi) - \theta)}{x(\xi) - \theta} \qquad |\xi| \leq w/2 \ . \tag{3.5}$$

This gives a formal representation of θ as a weighted mean

$$\theta \ = \ \frac{\int_{\xi=-w/2}^{w/2} \tau(\xi) x(\xi) d\xi}{\int_{\xi=-w/2}^{w/2} \tau(\xi) d\xi} \tag{3.6}$$

with weights depending on the data. It is known that M-estimators minimize objective functions more generally than for instance the standard sum of squared residuals. The estimators can be categorized into three major classes.

- **Hard redescenders**
  These estimators have $\psi(x) = 0$ for large $|x|$. Examples for this class of estimators are *Tukey's biweight* [Tuk81], *Andrew's Wave* [ABH+72] and *Talwar* [HT75].

- **Soft redescenders**
  These estimators have ψ functions that are asymptotic to zero for large $|x|$. One example is Welsch [DW76].

- **Monotone functions**
  This family has a monotone $\psi(x)$ function and includes Huber [Hub81], Logistic and Fair [Fai74].

These estimators are characterized by the concepts of efficiency and breakdown point. Efficiency refers to the relative ability of an estimator to yield optimal estimates of the assumed noise distribution.

For the objective function ρ different models can be used which are visualized in Figure 3.3. The first column in this figure defines objective functions for the

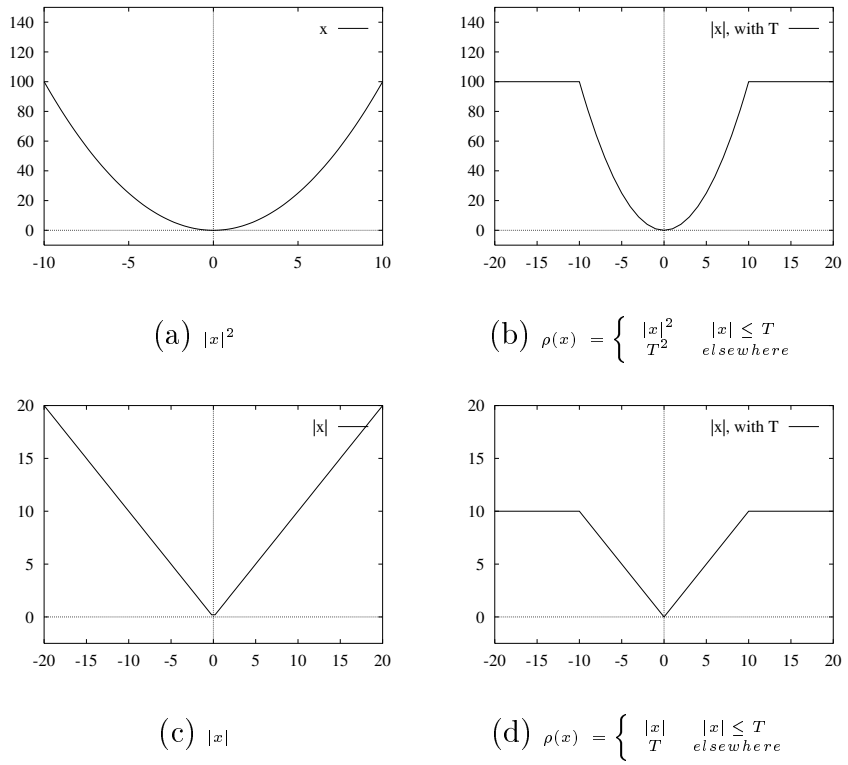


Figure 3.3: Different objective functions ρ.

least squares method (top) and the absolute difference (bottom). The second column depicts objective functions, where the residual error x is tested whether it is inside the given region defined by a certain threshold T. If the residual is above the threshold T the point is rejected by setting the residual value to T. The boundaries between rejection and non-rejection which yield good compromises between safety and efficiency, will have to be found by methods other than outlier tests. These boundaries may not, and, in general, will not coincide with the boundaries between "uninteresting" and "interesting" observations to be examined separately. The latter should include also all "doubtful" outliers which are correct, but which, in a broader context, may also turn out to be rather special. This suggests at least three categories, which are depicted in Figure 3.4:

- **clear outliers**, rejected by an outlier rule,
- **doubtful outliers**, where no clear apriori decision is possible and
- **proper data**.

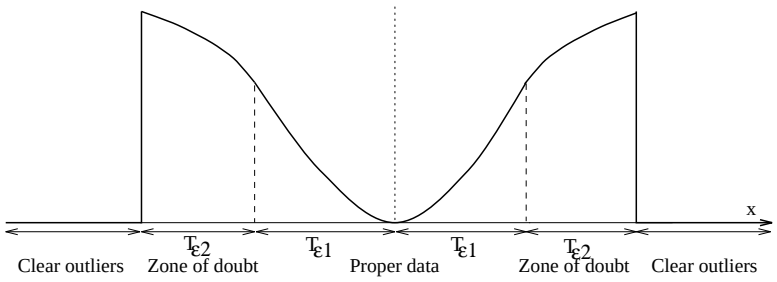


Figure 3.4: Three classes of doubts.

Some specific forms of functions ρ and ψ are proposed in literature, such as Huber's and Tukey's [Tuk81]. Huber derived the following robust ρ and ψ:

$$\rho(x) = \begin{cases} 0.5x^2 & |x| \leq aS \\ aS|x| - 0.5(aS)^2 & otherwise \end{cases} \tag{3.7}$$

$$\psi(x) = \begin{cases} -aS & x < -aS \\ x & |x| \leq aS \\ aS & x > aS \end{cases} .$$

Tukey's ρ and ψ function can be expressed as

$$\rho(x) = \begin{cases} \frac{1}{aS}[1 - (1 - (\frac{x}{aS})^2)^3] & |x| \leq aS \\ \frac{1}{aS} & |x| > aS \end{cases} \tag{3.8}$$

$$\psi(x) = \begin{cases} x[1 - (\frac{x}{aS})^2]^2 & |x| \leq aS \\ 0 & |x| > aS \end{cases} ,$$

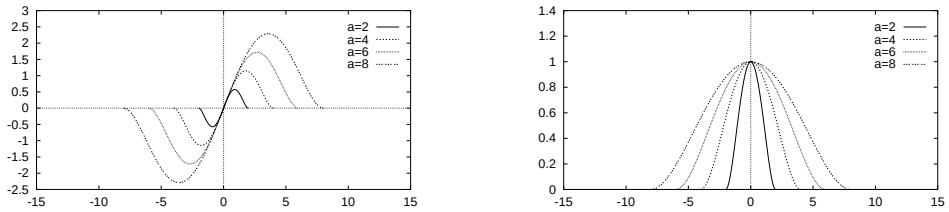
where a is a tuning constant, 1.5 for Huber's and 6 for Tukey's; S is a scale estimator which is usually *MAD* (median of absolute deviation). The weight function of Tukey's biweight is

$$\tau(x) = \begin{cases} [1-(\frac{x}{aS})^2]^2 & |x| \leq aS \\ 0 & |x| > aS \end{cases} . \tag{3.9}$$

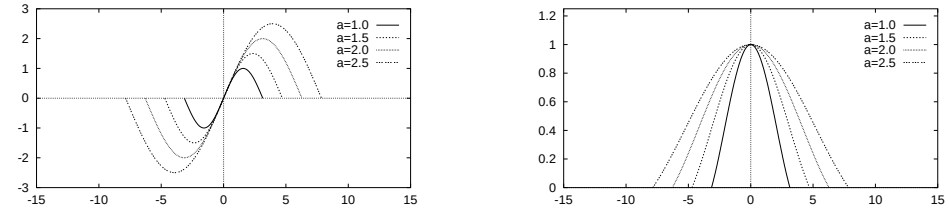
In Table 3.2 representatives of the three classes of estimators are given.

Estimator	$\rho_{Estimator}$	$\psi_{Estimator}$	$\tau_{Estimator}$	range	a
Tukey's biweight	$\frac{(aS)^2}{2}[1-(1-(\frac{x}{aS})^2)]^3$ $\frac{(aS)^2}{2}$	$x[1-(\frac{x}{aS})^2]^2$ 0	$[1-(\frac{x}{aS})^2]^2$ 0	$\|x\| \leq aS$ $\|x\| > aS$	$1.0 \leq a \leq 6.0$
Andrews wave	$(aS)^2[1-cos(\frac{x}{aS})]$ $2(aS)^2$	$aS[sin(\frac{x}{aS})]$ 0	$\frac{aS}{x}sin(\frac{x}{aS})$ 0	$\|x\| \leq aS$ $\|x\| > aS$	$1.0 \leq a \leq 2.5$
Talwar	$\frac{x^2}{2}$ $\frac{(aS)^2}{2}$	x 0	1 0	$\|x\| \leq aS$ $\|x\| > aS$	$1.0 \leq a \leq 5.0$
Welsch	$\frac{(aS)^2}{2}[1-exp[-(\frac{x}{aS})^2]]$	$xexp[-(\frac{x}{aS})^2]$	$exp[-(\frac{x}{aS})^2]$		$1.0 \leq a \leq 5.0$
Huber	$\frac{x^2}{2}$ $aS\|x\| - \frac{aS^2}{2}$	x $aS\ sgn(x)$	1 $\frac{aS}{\|x\|}$	$\|x\| \leq aS$ $\|x\| > aS$	$1.0 \leq a \leq 2.0$
Fair	$(aS)^2[\frac{\|x\|}{aS} - log(1+\frac{\|x\|}{aS})]$	$x(1+\frac{\|x\|}{aS})$	$\frac{1}{(1+\frac{\|x\|}{aS})}$		$1.0 \leq a \leq 1.5$
Logistic	$(aS)^2 log[cosh(\frac{x}{aS})]$	$aS\ tanh(\frac{x}{aS})$	$\frac{aS}{x}tanh(\frac{x}{aS})$		$1.0 \leq a \leq 2.0$

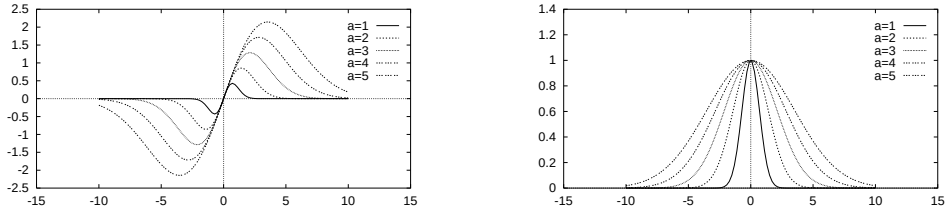
Table 3.2: Different classes of estimators: hard redescenders (Tukey's, Andrews and Talwar), soft redescenders (Welsch) and monotone functions (Huber, Fair and Logistic). For each function, ρ, ψ and the corresponding weighting function are depicted. The parameter a defines the tuning constant.



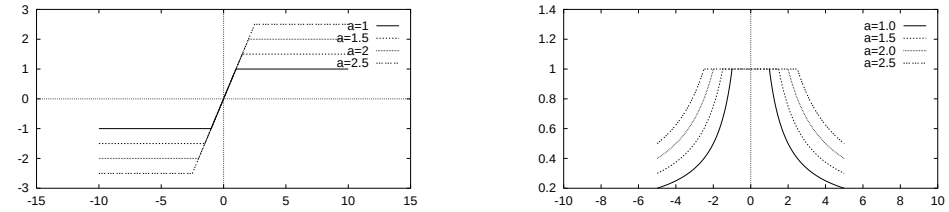
(a) Tukey's biweight: ψ_{tukey}, τ_{tukey}



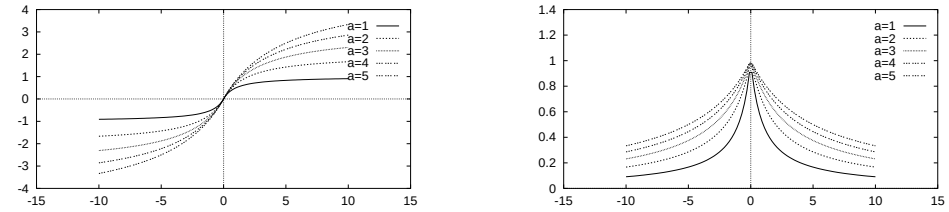
(b) Andrew's Wave: ψ_{sin}, τ_{sin}



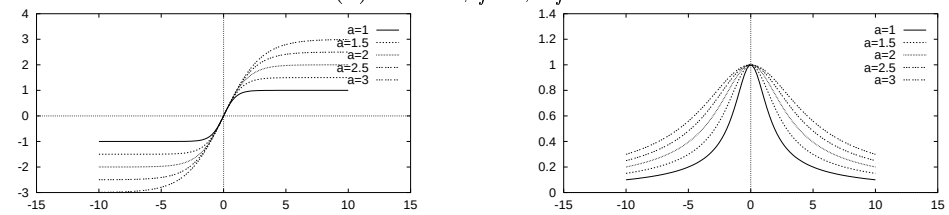
(c) Welsch: ψ_{welsch}, τ_{welsch}



(d) Huber: ψ_{huber}, τ_{huber}



(e) Fair: ψ_{fair}, τ_{fair}



(f) Logistic: ψ_{log}, τ_{log}

Figure 3.5: ψ and τ functions for M-estimators; hard (Tukey's, Andrew's and Talwar), soft-redescenders (Welsch) and monotone functions (Huber, Fair and Logistic). Multiple curves display the effect of increasing the tuning constant a.

The plots for the ψ and τ function are depicted in Figure 3.5. It can be seen that the tuning constant a determines the shape and the cutoff points of the weighting functions, thus influencing the estimated parameters of the estimator used.

3.3.2 Iterative weighted Least Squares Method

Since the weighting functions τ are used for this type of estimator it is difficult to find a closed form solution for a robust version of the correlation, thus an iterative method is used.

In order to define the correlation as minimization problem in the one-dimensional case, equations (2.19) and (2.29) are used to modify the correlation such that

$$C = \frac{2 - C_{LS}}{2} , \tag{3.10}$$

where

$$C_{LS}(x_L, x_R, w) = \int_{\xi=-\infty}^{\infty} \delta_{1/w}(\xi) \ \varepsilon^2(x_L + \xi, x_R + \xi, w) d\xi . \tag{3.11}$$

The residual error ε is defined as

$$\varepsilon^2(x_L, x_R, w) = [\{\frac{I_L(x_L) - \mu_L(x_L, w)}{\sigma_L(x_L, w)}\} - \{\frac{I_R(x_R) - \mu_R(x_R, w)}{\sigma_R(x_R, w)}\}]^2 . \tag{3.12}$$

Minimizing C_{LS} is equivalent to maximizing the correlation C. Using the re-weighted least squares (RLS) technique, equation (3.11) can be written in an iterative way as:

$$C_{\Omega_n}(x_L, x_R, w) = \int_{\xi=-\infty}^{\infty} \delta_{1/w}(\xi) \ \Omega_n(x_L, x_R, w) \ {\varepsilon_{\Omega_n}}^2(x_L + \xi, x_R + \xi, w) d\xi , \tag{3.13}$$

with the residual error

$${\varepsilon_{\Omega_n}}^2(x_L, x_R, w) = [\{\frac{I_L(x_L) - \mu_{L,\Omega_n}(x_L, w)}{\sigma_{L,\Omega_n}(x_L, w)}\} - \{\frac{I_R(x_R) - \mu_{R,\Omega_n}(x_R, w)}{\sigma_{R,\Omega_n}(x_R, w)}\}]^2 . \tag{3.14}$$

The weights from $\Omega_n \rightarrow \Omega_{n+1}$ are defined by

$$\Omega_{n+1}(x_L, x_R, w) = \tau({\varepsilon_{\Omega_n}}^2(x_L, x_R, w)) , \tag{3.15}$$

where τ is the weighting function (Table 3.2). The standard deviation can be written as

$$\sigma_{k,\Omega_n}(x, w) = \frac{1}{w} \int_{\xi=-\infty}^{\infty} \delta_{1/w}(\xi) \ \Omega_n(x_L + \xi, x_R + \xi, w)[I_k(x + \xi) - \mu_{k,\Omega_n}(x, w)] d\xi \quad k = L, R \tag{3.16}$$

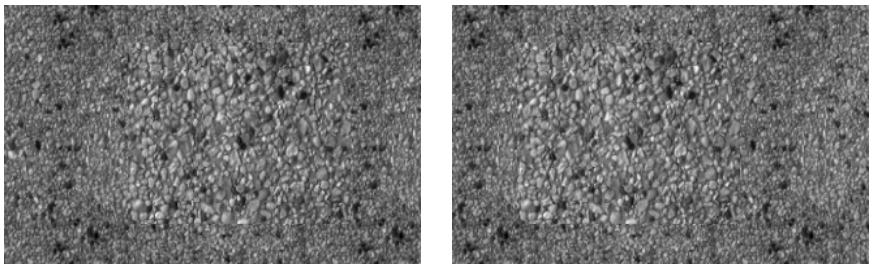
with

$$\mu_{k,\Omega_n}(x,w) \;=\; \frac{1}{w}\int_{\xi=-\infty}^{\infty} \delta_{1/w}(\xi)\;\Omega_n(x_L+\xi,x_R+\xi,w)I_k(x+\xi)d\xi \quad k=L,R\;. \tag{3.17}$$

The iteration starts with the intial condition $\Omega_0 \;= [1,\ldots,1]$. The correlation function can easily be extended from the one- to the two-dimensional case. The function C_Ω (Eq. 3.13) is used for the robust correlation technique. Now the complete algorithm for robust computation of the correlation can be outlined:

1. Compute the initial disparity value using the standard correlation C.
2. Based on the disparity $D = |x_L - x_R|$, compute weights Ω_n. For the computation different weighting functions τ can be used, which are depicted in Table 3.2.
3. Calculate a new correlation function C_{Ω_n} using the weights Ω_n and determine new disparities D.
4. Update the weights with $\Omega_{n+1} \;=\; \tau({\varepsilon_{\Omega_n}}^2)$
5. Iterate steps 2, 3 and 4 until convergence, or until a maximal number of iterations is reached[1].

3.4 Experimental Results and Comparisons



(a) box left (b) box right

Figure 3.6: Synthetic stereo pair: Box on a flat ground with natural texture added on the surface.

The robust approach is tested on a synthetic box image shown in Figure 3.6. In order to compare the accuracy of the stereo matching the disparity values for the

[1]Experiments have shown that usually a few iterations (2–4) are needed for convergence.

stereo pair are known. Different tests are performed on the box image.

Test for one and all points:

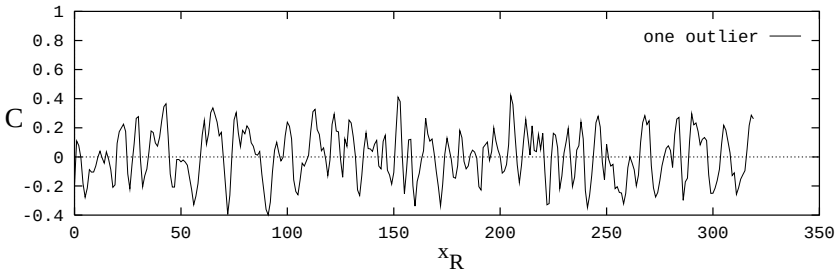
- *Outlier test:* The outlier test from section 3.2 is performed using the robust approach.
- *Robust Stereo Algorithm:* The disparity map is determined for the box image by using both the standard and the robust approach.

Test under different noise conditions[1] (Salt & Pepper and Gaussian noise):

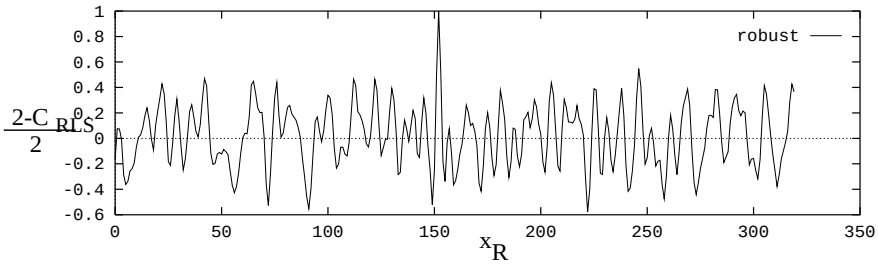
- *Residual test:* In order to compare the residual errors of the detected maximum the test is performed for only one position in the left box image using both the standard and the robust technique.
- *MSE test:* The disparity maps are computed for both the standard and the robust stereo method. The disparity maps are compared to each other by using the *MSE* (Mean Square Error). The *MSE* is determined between the ideal disparity map (Figure 3.8 (a)) and the computed results for both methods under different noise conditions. The following notation is used for this comparison $MSE(< method > (< noisecondition >) - ideal)$, where $< method >$ defines the used stereo method and $< noisecondition >$ the noise, which is introduced to the stereo pair.

[1]The robust correlation technique is tested for different weighting functions τ, which are depicted in Table 3.2

Outlier test: The same test from section 3.2 is performed on the corrupted left stereo image by using the robust approach. The correlation functions of one epipolar line for both standard and robust method are visualized in Figure 3.7. The outlier in the window could be detected and eliminated by using the robust approach, thus $C_\Omega(outlier) \approx C(ideal)$.



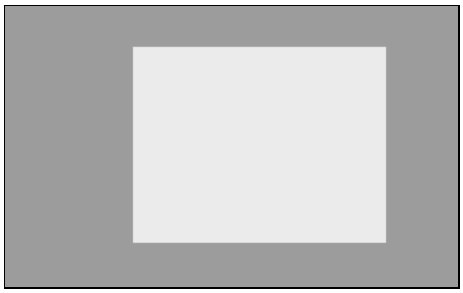
(a) correlation function C



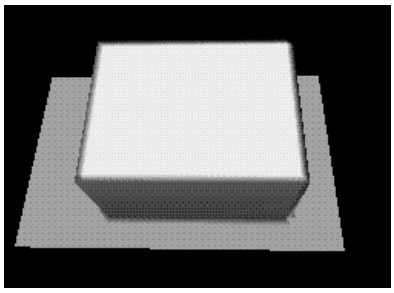
(b) correlation function C_Ω

Figure 3.7: Robust correlation: (a) the outlier corrupts the correlation value computed by C and is thus not accepted as a corresponding point. (b) The outlier could be detected and eliminated by using the proposed robust method.

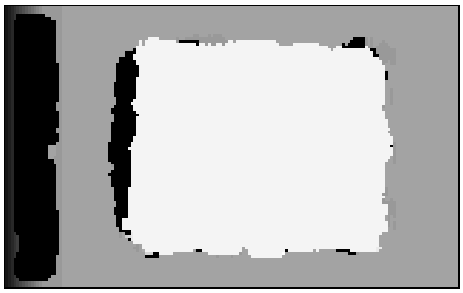
Robust Stereo Algorithm: A dense disparity map is computed for the box image for both the standard and the robust technique using a fixed window size w and a threshold T for accepting a corresponding point. For the robust technique Tukey's weighting function is used only to give one example. Figure 3.8 shows the ideal disparity map, the result of the standard stereo matching algorithm, the result with the robust correlation method and the disparity error distributions for both methods. It can be seen that in the regions near the depth discontinuity, edge points can be detected more exactly with the robust stereo technique with the exception of the occlusion area on the left side of the box.



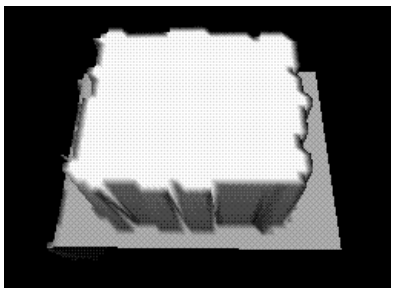
(a) ideal disparity



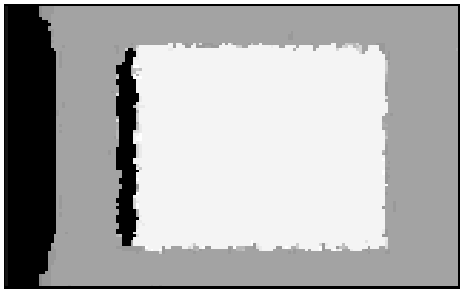
(b) 3D plot of (a)



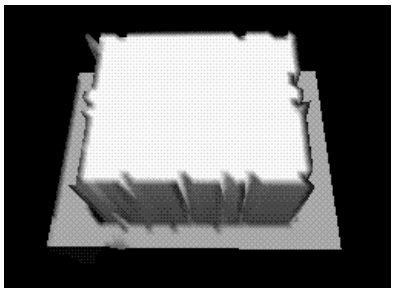
(c) C, w=7, T=0.9



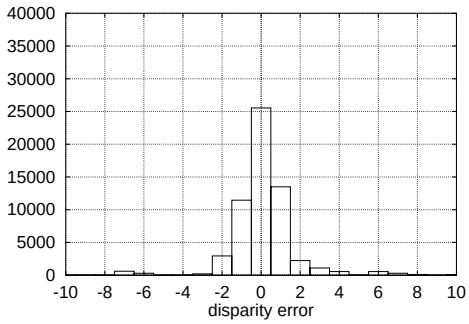
(d) 3D plot of (c)



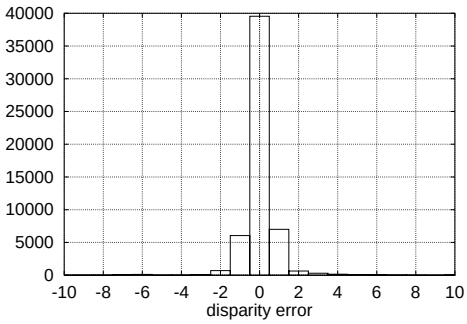
(e) C_Ω, w=7, T=0.9



(f) 3D plot of (e)



(g) error distribution for C



(h) error distribution for C_Ω

Figure 3.8: Comparison of the results: (a) ideal disparity map, (c) disparity maps computed with the standard, and (e) with the robust stereo method using Tukey's biweight τ_{tukey}. (b,d,f) show 3D plots of the computed disparity maps. (g),(h) depict the disparity error distributions for both methods.

3.4.1 Salt & Pepper Noise

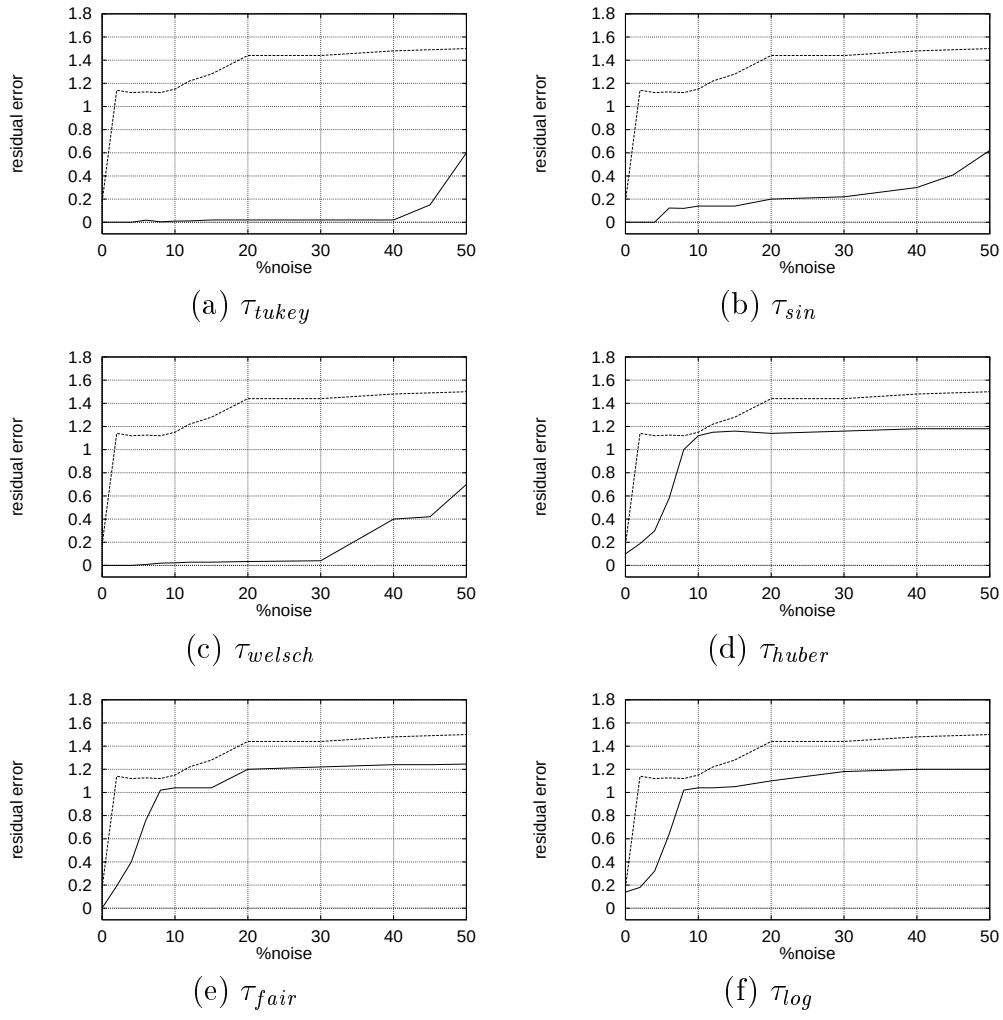
Images are sometimes corrupted by gross errors. In such a case, a random value replaces the value of the pixel. A commonly used model for this behavior is Salt & Pepper noise, in which some percentage of the pixels of the observed area is randomly replaced by white or black pixels.

Residual test: Up to 50% outliers are added to a 7×7 window. The corresponding point is determined using the standard and the robust technique with different weighting functions. The results of this test are visualized in Figure 3.9, where the plots are ordered in the same way as they appear in Table 3.2. The dotted curve defines the residual errors for the standard correlation method. It can be seen that for all weighting functions the residual error stays below the error computed with the standard method. A threshold of $T = 0.8$ for the correlation C is equivalent to the value of $T = 0.4$ for the function C_Ω. The relation is given in equation (3.10). If the correlation functions exceeds the threshold T the corresponding point is not accepted. In Table 3.3 the percentage of noise is given at which the residual error exceeds the threshold $T = 0.4$ for different weighting functions. The best results for this test are obtained by using Tukey's biweight.

weighting functions	τ	$T > 0.4$
Tukey's	(τ_{tukey})	48%
Andrew's Wave	(τ_{sin})	45%
Logistic	(τ_{log})	45%
Welsch	(τ_{welsch})	42%
Huber	(τ_{huber})	5%
Fair	(τ_{fair})	4%
.	.	
.	.	
Standard method		at 1%

Table 3.3: Results for different weighting functions.

The residual errors stay below the threshold for up to nearly 50% of noise. In the case of Salt & Pepper noise, the outliers have to be eliminated by the outlier rule and not only down-weighted.

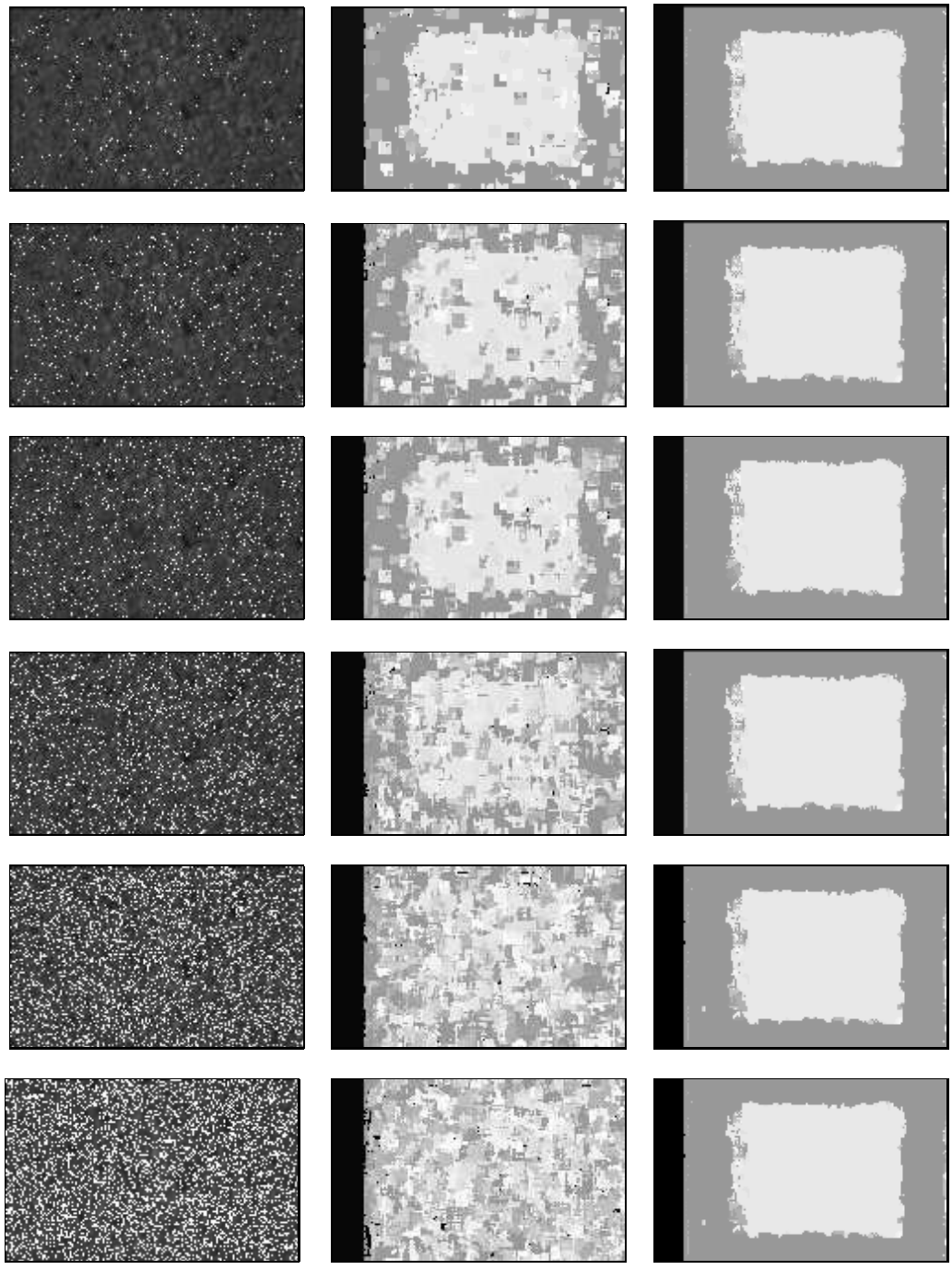


(a) τ_{tukey} (b) τ_{sin}

(c) τ_{welsch} (d) τ_{huber}

(e) τ_{fair} (f) τ_{log}

Figure 3.9: *Residual test* under Salt & Pepper noise condition: For one point in the left box image the residual error is computed for both methods. The residual error for the standard method is plotted as dotted curve and for the robust technique as solid curve using different weighting functions τ. The x-axis defines the percentage of Salt & Pepper noise added to the window.



(a) additive noise (b) C (c) C_Ω

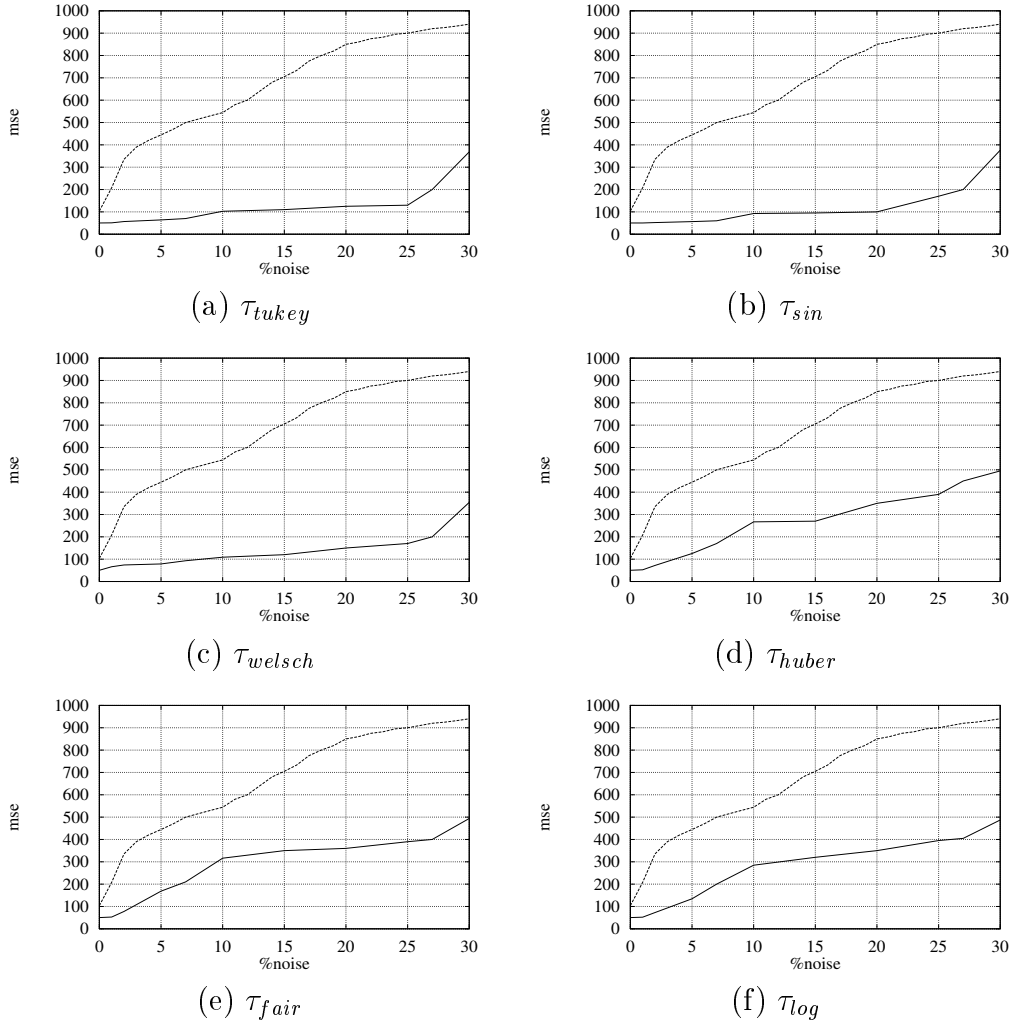
Figure 3.10: (a) Additive Salt & Pepper noise added to the box image with 1%, 2%, 3%, 5%, 20% and 30%. (b) Disparity maps computed with C. (c) Disparity maps computed with C_Ω using τ_{tukey}. The size of the search window is constant for both methods with $w = 7$.

MSE test: 30% of Salt & Pepper noise is added to the left box image. For this test the threshold T is set to zero to accept also wrong matches. The disparity maps for the standard and the robust technique are determined and compared to each other. For evaluation the Mean Square Error *MSE* is computed between the ideal depicted in Figure 3.8 (a) and computed disparity maps for both methods. The results are visualized in Figure 3.10, where the first column depicts the left image introduced with different percentages of Salt & Pepper noise, the second column depicts the disparity maps computed with the standard correlation method and the last column with the robust method using Tukey's biweight. Figure 3.11 shows the $MSE(C[S\&P] - ideal)$ and the $MSE(C_\Omega[S\&P] - ideal)$ for different weighting functions. The dotted curve represents the error for the standard correlation method. Also for this test the robust approach shows for all weighting functions a stable behavior against the standard method. Table 3.4 depicts the *MSE* for 25% noise for different weighting functions and for the standard method.

weighting functions	τ	**MSE** at 25%
Tukey's	(τ_{tukey})	120
Andrew's Wave	(τ_{sin})	170
Welsch	(τ_{welsch})	170
Huber	(τ_{huber})	395
Fair	(τ_{fair})	399
Logistic	(τ_{log})	400
.	.	
.	.	
Standard method		940

Table 3.4: $MSE(C_\Omega[S\&P] - ideal)$ for different weighting functions for 25% noise.

Especially for hard redescenders the robust technique delivers better results than using monotonic functions. The hard redescenders give better results, because gross errors like Salt & Pepper noise should be rejected and not down-weighted.



(a) τ_{tukey} (b) τ_{sin}

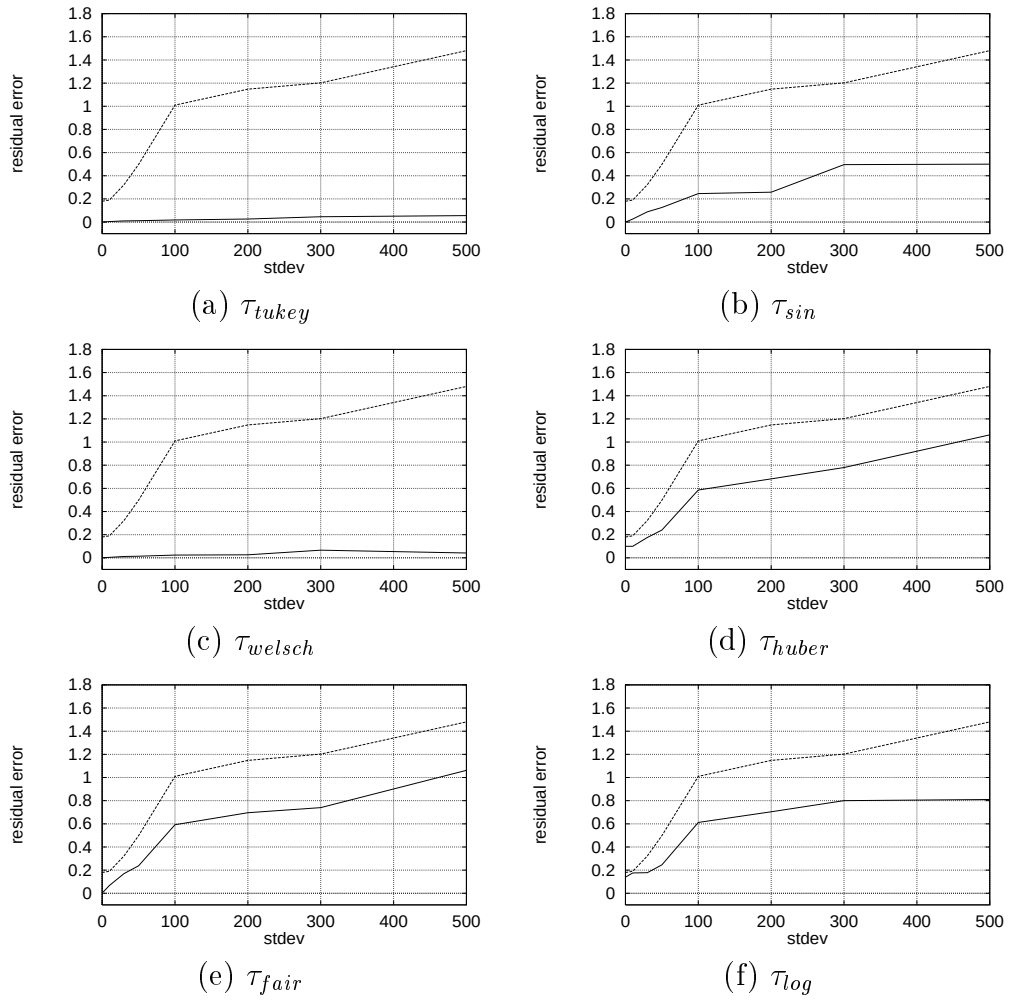
(c) τ_{welsch} (d) τ_{huber}

(e) τ_{fair} (f) τ_{log}

Figure 3.11: *MSE test* under Salt & Pepper noise condition: The $MSE(C[S\&P] - ideal)$ is plotted as dotted curve and the $MSE(C_\Omega[S\&P] - ideal)$ as solid curve for different weighting functions.

3.4.2 Additive Gaussian Noise

Another experiment tests the sensitivity of the standard and robust technique to additive Gaussian noise. Also for this noise condition the *residual test* and the *MSE test* are performed for the box image. The stereo pair is corrupted with additive Gaussian noise with a standard deviation varying from $\sigma = [0..500]$ with pixels clipped for > 255 and < 0.



(a) τ_{tukey} (b) τ_{sin}

(c) τ_{welsch} (d) τ_{huber}

(e) τ_{fair} (f) τ_{log}

Figure 3.12: *Residual test* under Gaussian noise condition: For one point in the left box image the residual error is computed for both methods as a function of σ. The residual error for the standard method is plotted as dotted curve and for the robust technique as solid curve using different weighting functions τ.

Residual test: The results of the computed residual errors for one position in the left box image are depicted in Figure 3.12. Also for this experiment a threshold of $T = 0.4$ is taken for accepting corresponding points. The test is performed for different weighting functions. It can be seen that also for this test the residual errors of the robust method are always smaller than the errors resulting from the standard method. In Table 3.5 the value of σ for different weighting functions is depicted, at which the residual error exceeds the threshold $T = 0.4$. The best

weighting functions	τ	σ:$T > 0.4$
Tukey's	(τ_{tukey})	500 error ≤ 0.4
Welsch	(τ_{welsch})	500 error ≤ 0.4
Andrew's Wave	(τ_{sin})	260
Logistic	(τ_{log})	65
Huber	(τ_{huber})	65
Fair	(τ_{fair})	69
.	.	
.	.	
Standard method		40

Table 3.5: Results for different weighting functions.

results are obtained for Welsch and Tukey's biweight. It can also be seen from Table 3.5 that the standard method accepts a certain percentage of Gaussian noise. The reason is that the standard correlation follows the least squared argument and is based on Gaussian data distribution.

MSE test: Also for Gaussian noise condition the disparity maps computed with both standard and robust method are compared to each other, by using the *MSE*. The stereo pair is introduced with additive Gaussian noise. The results are visualized in Figure 3.13, where the first column depicts the left box image introduced with Gaussian noise with increasing standard deviation, the second column depicts the disparity maps computed with the standard correlation method and the last column with the robust method using Tukey's biweight. The $MSE(C(Gauss) - ideal)$ and the $MSE(C_\Omega(Gauss) - ideal)$ for different weighting functions is visualized in Figure 3.14. Table 3.6 depicts the $MSE(C(\sigma = 500) - ideal)$ and $MSE(C_\Omega(\sigma = 500) - ideal)$ for different weighting functions.

Even under Gaussian noise the error of the robust stereo method is always smaller than the error obtained with the standard method. For this noise condition it is also obvious from Figure 3.14 that hard redescenders obtain better results than monotonic functions.

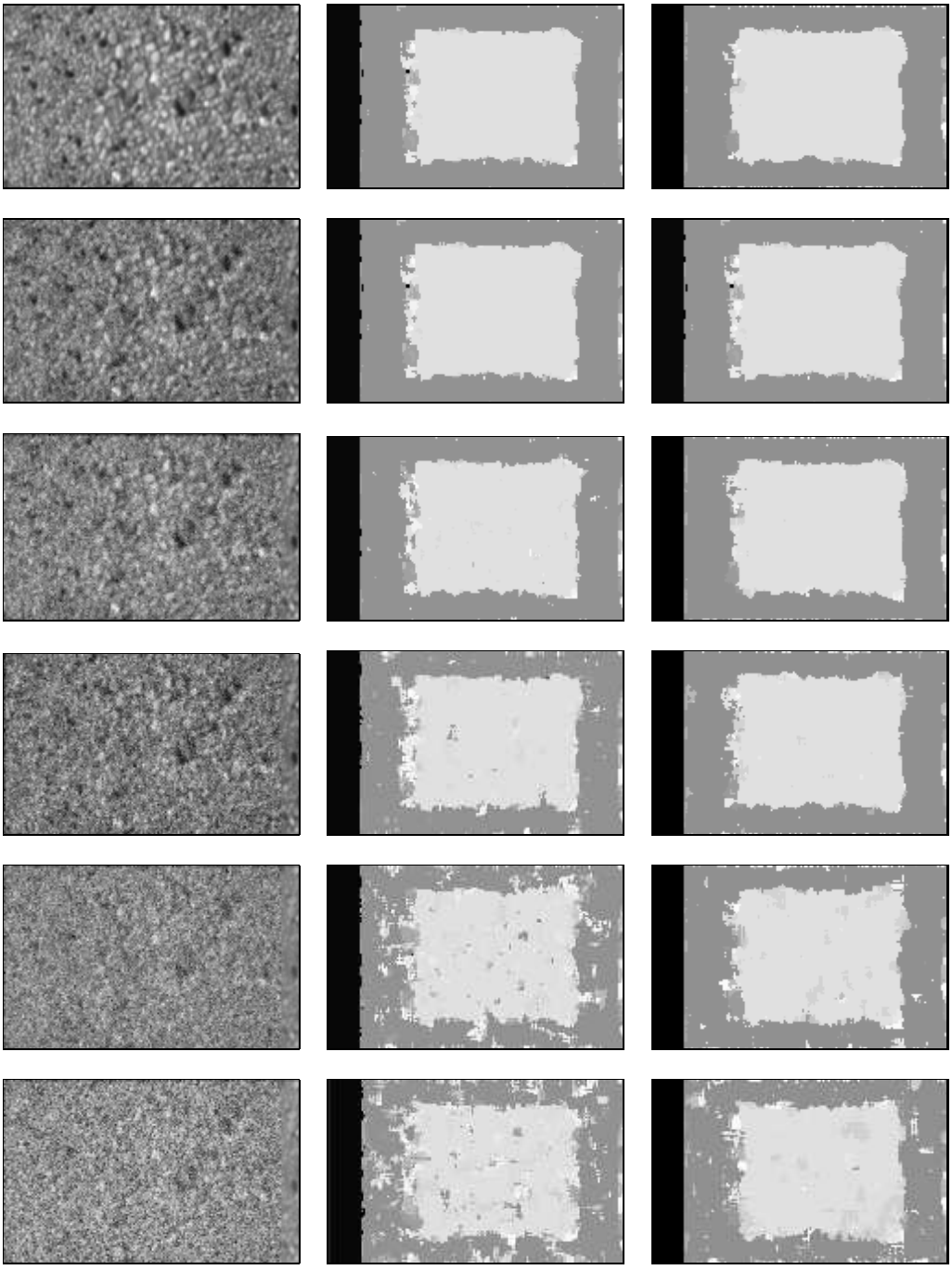
weighting functions	τ	**MSE**, σ=500
Welsch	(τ_{welsch})	190
Tukey's	(τ_{tukey})	210
Andrew's Wave	(τ_{sin})	210
Fair	(τ_{fair})	240
Huber	(τ_{huber})	250
Logistic	(τ_{log})	260
.	.	
.	.	
Standard method		450

Table 3.6: $MSE(C(\sigma = 500) - ideal)$ and $MSE(C_\Omega(\sigma = 500) - ideal)$ for different weighting functions.

3.5 Chapter Summary

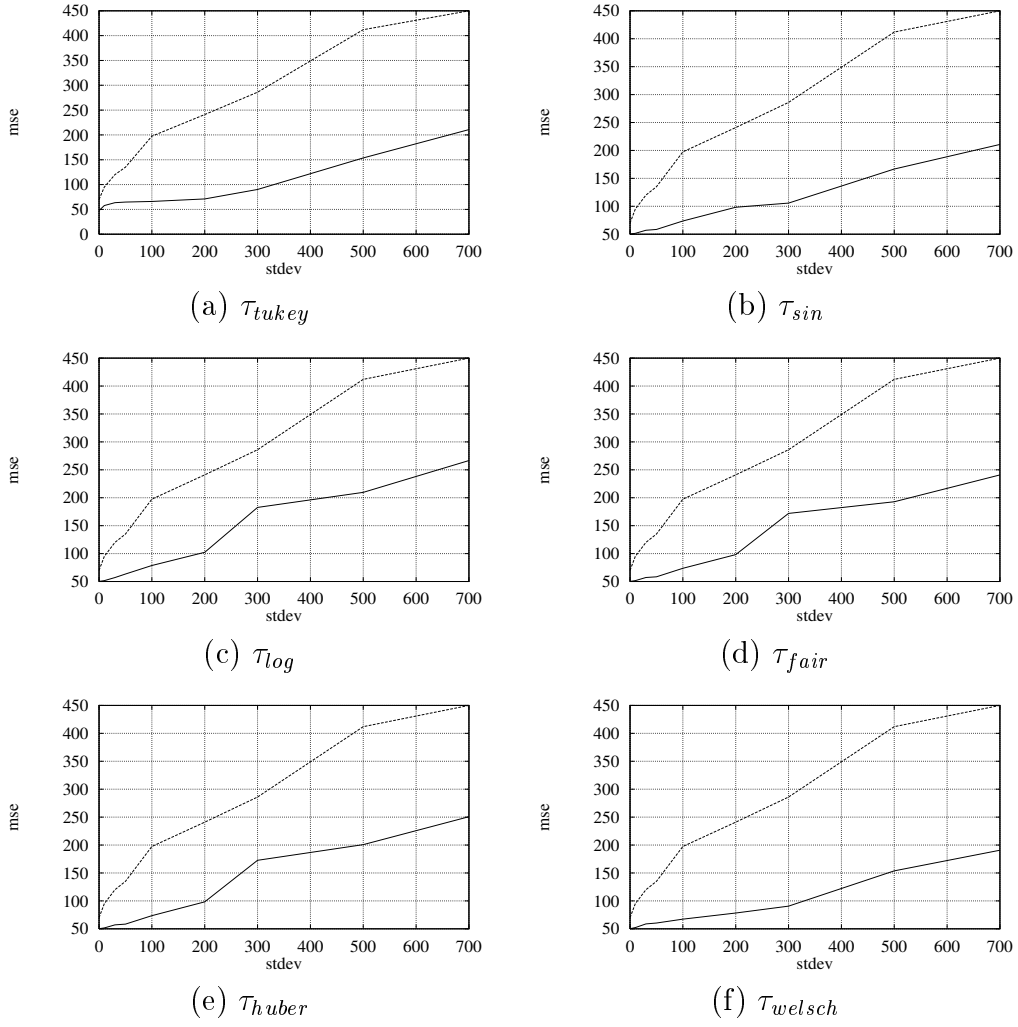
This chapter presented a modification of the standard area-based correlation approach. It can tolerate a significant number of outliers. The approach exhibits a robust behavior not only in the presence of mismatches but also in the case of depth discontinuities. Various tests were performed using different weighting functions under different noise conditions. It was shown, for example, that the robust approach can nearly tolerate 50% of noise. Outliers could be detected and eliminated. Tests were made for one (*Residual test*) and all points (*MSE test*) of the box image. It can be observed that the robust correlation method obtains better results than the standard method even under different noise conditions.

But there are still problems with window operators (local information and the fact that the information obtained in one window is usually *not* shared among the neighboring windows; also another consequence of such a local approach is that the number of outliers that can be tolerated is limited successfully). The problem of selecting the size of the search window is not sufficiently solved by using this robust approach, since it may contain too much or too little information.



(a) additive noise (b) C (c) C_Ω

Figure 3.13: (a) Additive Gaussian noise added to the left stereo image with $\sigma = 30$, 50, 100, 200, 300 and 500. (b) Disparity maps computed with C. (c) Disparity maps computed with C_Ω using τ_{tukey}. The size of the search window is constant with $w = 7$.



(a) τ_{tukey} (b) τ_{sin}

(c) τ_{log} (d) τ_{fair}

(e) τ_{huber} (f) τ_{welsch}

Figure 3.14: *MSE test* under Gaussian noise condition: The $MSE(C[Gauss] - ideal)$ is plotted as dotted curve and the $MSE(C_\Omega[Gauss] - ideal)$ as solid curve for different weighting functions.

Chapter 4

Multi-Scale Stereo

4.1 Introduction

Fine scale properties are available for objects in the real world. They are suppressed when used for everyday purposes. For instance when looking at moving objects like people at a distance (coarse scale) it is important to identify them as a woman or man or to determine the direction of movement. At a fine scale it is more appropriate to talk about small features of the individual person, like the hair style or the face. By refining the scale, the face of the person can be classified in several parts like the eyes, nose, mouth, etc. This fact is well-known in the experimental sciences. In physics for example, the world is described at several levels of scales, from partical physics at fine scales to astronomy at large scales. In order to model the structure of the real world the concept of scale plays an important role in computer vision.

The terms scale and resolution are sometimes used interchangeably in the literature, and their precise meaning is not always clear. In this work the following convention for scale and resolution is used:

- **resolution:** Resolution is the reduction of features into its constituent parts. Thus for images resolution is the spatial density of the grid points [HS93].
- **scale:** Scale is used for the characteristic length of operators used for image processing. In order to determine the scales at which a feature is present, image filters are applied with kernels of varying sizes (operator size). Thus the scale range of a feature is the range of operator size for which the feature can be detected after filtering.

According to resolution in numerical analysis, the accuracy can often be increased by refining the grid sampling. The selection of a large grid size is mainly motivated by efficiency, since exact equations are simulated. In computer vision the size of the grid for resolving structures is sometimes very small, thus making the solution even more difficult. A more serious problem is that of scale. In most

standard numerical problems the refinement of the grid used will improve the accuracy of the results. In easy problems the solutions contain variations which take place in one single space, whereas problems having solutions with variations on different scales are more complicated to handle. These fine-scale phenomena cannot always be resolved by discrete approximations. Moreover the occurrence of discontinuities in image data are known to complicate the situation further. In order to determine objects from a digital image it is necessary to extract information from it by using some operators. In this situation it is important what kind of operators should be used and how large they should be.

The general idea of representing a signal on multiple scales is not new. Early work was performed by Rosenfeld in 1971 [RT71], who observed the advantage of using operators of different sizes in edge detection. Several authors used different levels of resolution which is described in [Kli71, Uhr72]. These approaches have been developed further by Burt [BA83], Crowley [CS87], Kropatsch [Kro91], et al. to the **image pyramid**. In the next two subsections two major scale representations are discussed, namely the hierarchical approach using image pyramids and the scale-space approach.

In the further sections various stereo matching methods are presented using multi-scale representations. In the first method a hierarchical structure is used for the matching process, where both scale and resolution change. A coarse to fine matching method is described next where the resolution of the input images is not changed.

Finally, a new approach (multi-scale representation) is proposed where the size of the search window can be changed in a continuous way, thus making it possible to be more exact in the determination of interesting scales for certain regions in a stereo pair. Experimental results are given at the end of this chapter on synthetic and real images.

4.1.1 Pyramid Representation

A pyramid representation of a signal is a set of successively smoothed and subsampled representations of the original signal, which is visualized schematically in Figure 4.1. Tanimoto [Tan86] describes a pyramid as a collection of images of a single scene at different resolutions. The images of this collection can be ordered according to their resolution and are numbered as **levels**. The number of pixels from level to level decreases with a constant factor. Two terms describe the structure of a pyramid: the **reduction factor** and the **reduction window** [Kro91]. The reduction factor determines the rate at which the number of pixels decreases from one level to the next. The reduction window to every pixel in the entire pyramid associates a set of pixels in the level below. The reduction windows are mostly rectangularly shaped and are described by (**number of columns**) $\times$ (**number of rows**). In the classical pyramid every 2×2 block of pixels is merged recursively into one pixel of the coarser level. The structure

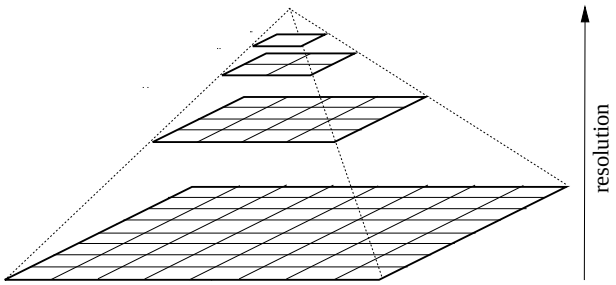
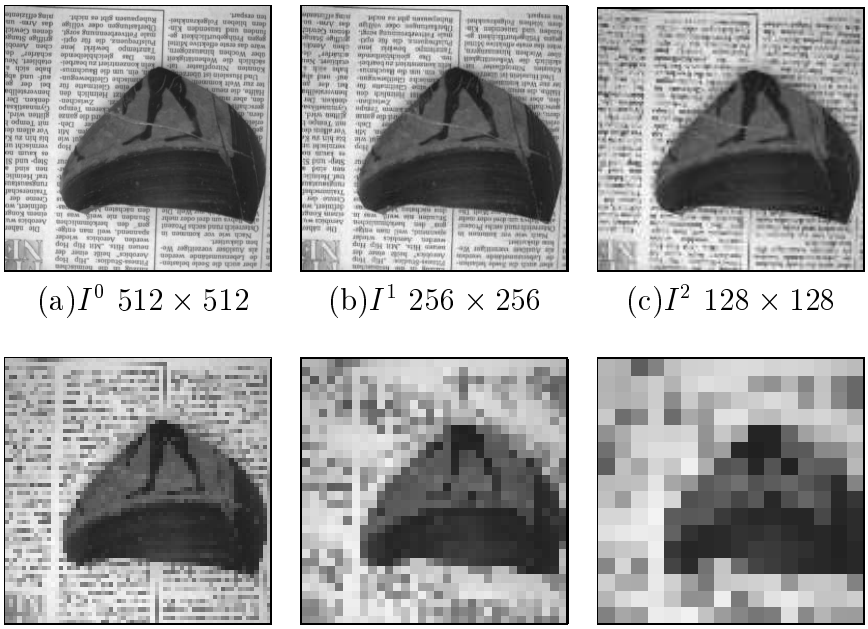


Figure 4.1: Pyramid representation.

can be described by $2 \times 2/4$ which specifies the 2×2 reduction window and the reduction factor of 4.

In order to create a pyramid structure this operation is performed recursively from level n to a coarser level $n + 1$. An example can be seen in Figure 4.2, where a pyramid representation of an image containing an archaeological sherd is constructed. The size of the original image is 512 ×512. The resolution at the top level of the pyramid in this example is 16×16.

The main advantage of the pyramid representation is the rapidly decreasing



(a)I^0 512×512 (b)I^1 256×256 (c)I^2 128×128

(d)I^3 64×64 (e)I^4 32×32 (f)I^5 16×16

Figure 4.2: Six levels of a 2 ×2/4 pyramid.

image size, thus reducing the computational work in preprocessing steps. The memory requirements are small and there are commercially available implementations of pyramids in hardware [BASv86, vGSJ85]. There is a large literature on different aspects of pyramid representation. Early works of pyramid structures were proposed by Burt [Bur81], Burt and Adelson [BA83], Crowley [CS87] and Meer [MBR87]. In general, pyramids are not translation invariant, which implies that the representation changes when the image is shifted. Some new works providing a translation invariant segmentation using irregular pyramid structures have been proposed by Nacken [Nac95] and Kropatsch [KY96]. An overview of the state of the art in the field of computer vision using new pyramid structures and curve representation schemes can be found in [Kro91] and [JR94].

4.1.2 Scale-Space

The main idea of creating a **scale-space representation** of a signal is to successively suppress fine-scale information. A mechanism is required that simplifies the data and removes high frequency information. This operation is known by the term **scale-space smoothing** [Lin94]. One of the major reasons for a multi-scale representation is to remove fine details for preprocessing steps in order to have low noise conditions and to eliminate unnecessary scene details. One method for a multi-scale representation of measured signals was proposed by Witkin and Koenderinck [Wit83, Koe84]. The main idea of a scale-space representation of a

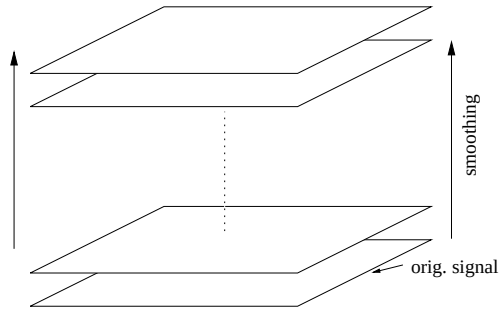


Figure 4.3: Multi-scale representation of a signal at different scales.

signal is to generate a one-parameter family of derived signals in which the fine scale information is successively suppressed, preserving the same resolution. A scale-space representation is visualized in Figure 4.3, where the original signal is at the bottom.

For a one-dimensional signal $f : \mathbb{R} \mapsto \mathbb{R}$ the scale-space representation $I : \mathbb{R} \times \mathbb{R}_+ \mapsto \mathbb{R}$ is defined in such a way that the representation at lowest scale is equal to the original signal

$$I(x;0) \;=\; f(x) \;, \tag{4.1}$$

and the representations at coarser scales are given by convolving the signal f with smoothing kernel $g(x,t)$

$$I(x;t) \ = g(x,t) \ * \ f(x) \ , \tag{4.2}$$

where the scale parameter $t \in \mathbb{R}_+$ is increased successively. Equation (4.2) can be written explicitly as

$$I(x;t) \ = \int_{\xi=-\infty}^{\infty} g(\xi,t) \ f(x-\xi) \ d\xi. \tag{4.3}$$

The one-dimensional Gaussian kernel is defined by

$$g(x,t) \ = \frac{1}{t\sqrt{2\pi}} \ e^{-\frac{x^2}{2t^2}}. \tag{4.4}$$

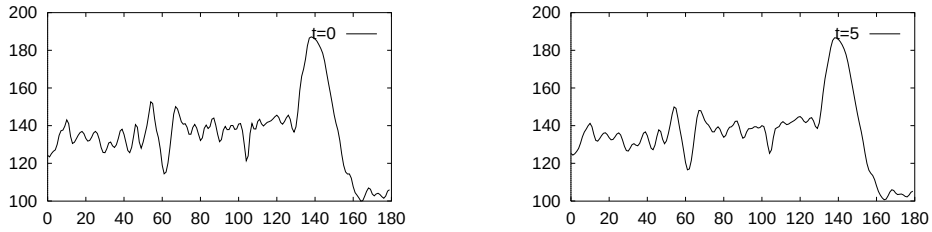
In general, instead of using the Gaussian kernel, any other symmetric function can be used. The scale-space theory has been developed for continuous signals and images. In order to approximate the convolution integral of equation (4.2) numerically the rectangular rule of integration is used without truncation of the infinite integral interval. This leads to the approximation

$$I(x;t) \ = \sum_{n=-\infty}^{\infty} g(n,t) \ f(x-n) \ , \tag{4.5}$$

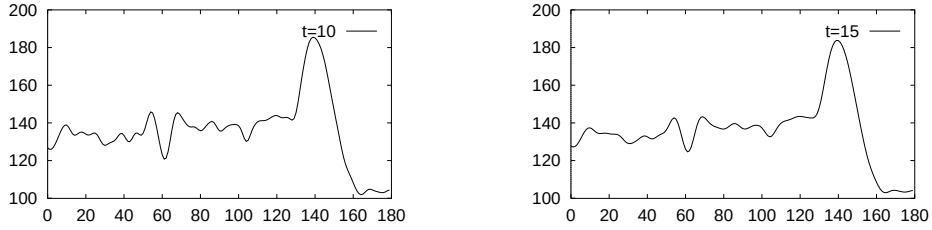
with

$$\sum_{n=-\infty}^{\infty} g(n,t) \ = \sum_{n=-\infty}^{\infty} \frac{1}{t\sqrt{2\pi}} \ e^{-\frac{n^2}{2t^2}} \ \approx 1. \tag{4.6}$$

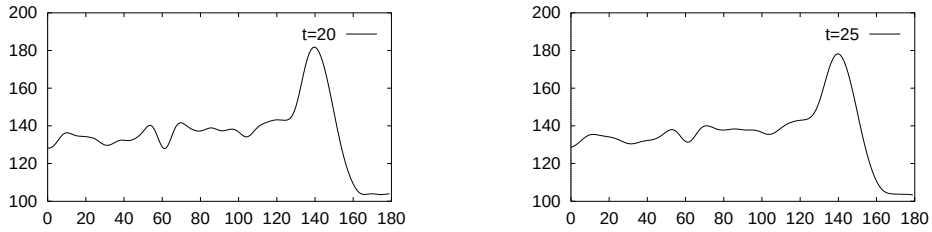
Using this approximation the transformation from the zero level $I(x;0)$ to a higher level never increases the number of local extrema [Wit83]. The transformation from a certain low level $I(x;t_1)$ to a higher level $I(x;t_2)$ in general is not a scale space transformation [Lin94](ch 4). Figure 4.4 shows the results of smoothing one line $I(x)$ in an image with the one-dimensional Gaussian kernel, thus the signal becomes successively smoother by increasing parameter t. It can be seen that the signal becomes successively smoother in higher scale.



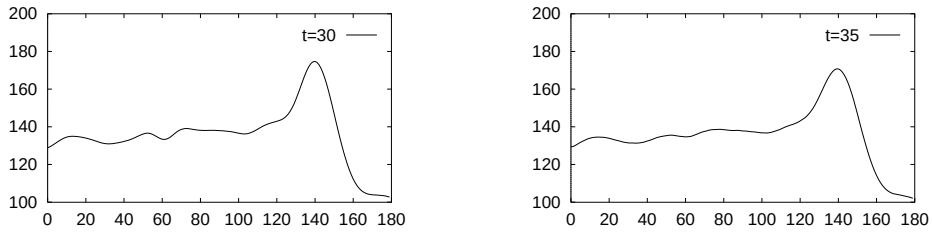
(a) $I(x;0)$ (b) $I(x;5)$



(c) $I(x;10)$ (d) $I(x;15)$



(e) $I(x;20)$ (f) $I(x;25)$



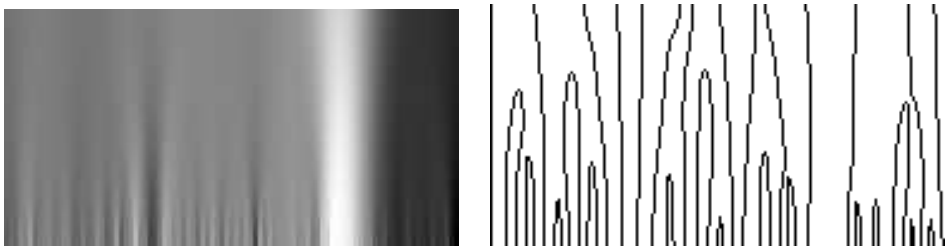
(g) $I(x;30)$ (h) $I(x;35)$

Figure 4.4: Signal $I(x)$ is successively smoothed using Gaussian kernels of increasing width t.

In order to construct a scale-space representation it is very important that

- **fine-scale features disappear** monotonically **with increasing scale** and that
- **no new** artificial **structures** are created at coarser scales due to the use of the smoothing method.

A feature at a coarser level should be defined as a simplification of the same feature in the original signal. Witkin observed that the number of zero-crossings in the second derivative decreases monotonically with scale (Figure 4.5). This fact



(a) scale-space of the signal $I(x)$ (b) zero crossings of the second derivative

Figure 4.5: Scale-space representation of a one-dimensional signal $I(x)$.

is very important and means that the number of local extrema in any derivative of the signal cannot increase with scale.

The extension of the one-dimensional scale-space theory to two and higher dimensions is not obvious, since it is possible to show that there are no non-trivial kernels with the property that they never introduce new local extrema [Lin94]. An illuminating counter-example was proposed by Lifshitz and Pizer [LP87]. A two-dimensional image contains two hills, one being higher than the other. Connect the two tops by a narrow sloping ridge without any local extrema, so that the top point of the lower hill no longer is a local maximum. This configuration is the input "image". Smoothing erodes the ridge much faster than the hills. After a while, the lower hill becomes a local maximum, thus a new maximum has been created by smoothing. The reason is that the narrow ridge is a fine-scale phenomenon and should therefore disappear before the coarse-scale peaks [Lin94](ch. 4). The property that new local extrema can be created by linear smoothing is inherent in two and higher dimensions. In one dimension, the number of local extrema is a natural measure of structure, on which a theory can be founded, but not in higher dimensions.

Koenderink (1984) [Koe84] extended the scale-space concept to two-dimensional signals. He introduced the notion of **causality**, which means that new structures must not be created when the scale parameter is increased. Causality is combined with the notions of **homogeneity** and **isotropy**, which means that all spatial points and all scale levels must be treated in a similar manner. The main idea is that it should be possible to trace every gray-level at a coarse scale to a corresponding gray-level at a finer scale, meaning that no new structures should be created when the scale parameter increases. Related work was given by Yuille and Poggio [YP86], concerning the zero-crossings of the Laplacian of the Gaussian, and also by Babaud et al. [BWBD86] and Hummel [Hum87].

Discrete scale-space approximation for two-dimensional signals:

The discrete scale-space approximation for two-dimensional signals can be written as:

$$I(x;t) \;=\; \sum_{m=-\infty}^{\infty} g(m,t) \sum_{n=-\infty}^{\infty} g(n,t)\; f(x-m, y-n)\;, \tag{4.7}$$

where

$$g(m,t)g(n,t) = \frac{1}{2\pi t^2} e^{-\frac{m^2+n^2}{2t^2}} \tag{4.8}$$

Figure 4.6 illustrates two sampled Gaussian kernels at different scale levels t,

(a) t=128, image size 255 × 255 (b) t=2, zoomed region of 11 × 11

Figure 4.6: Gray-level illustrations of sampled Gaussian kernels at different scale levels t.

represented by gray-level values. Figure 4.6 (b) depicts a zoomed region. The extension from the one-dimensional case to the two-dimensional case is explained in more detail in [Lin94]. In the two-dimensional scale-space the digital image is smoothed with a continuously increasing scale parameter t. Figure 4.7 shows 12 different scales of an image containing one archaeological sherd placed on a newspaper.

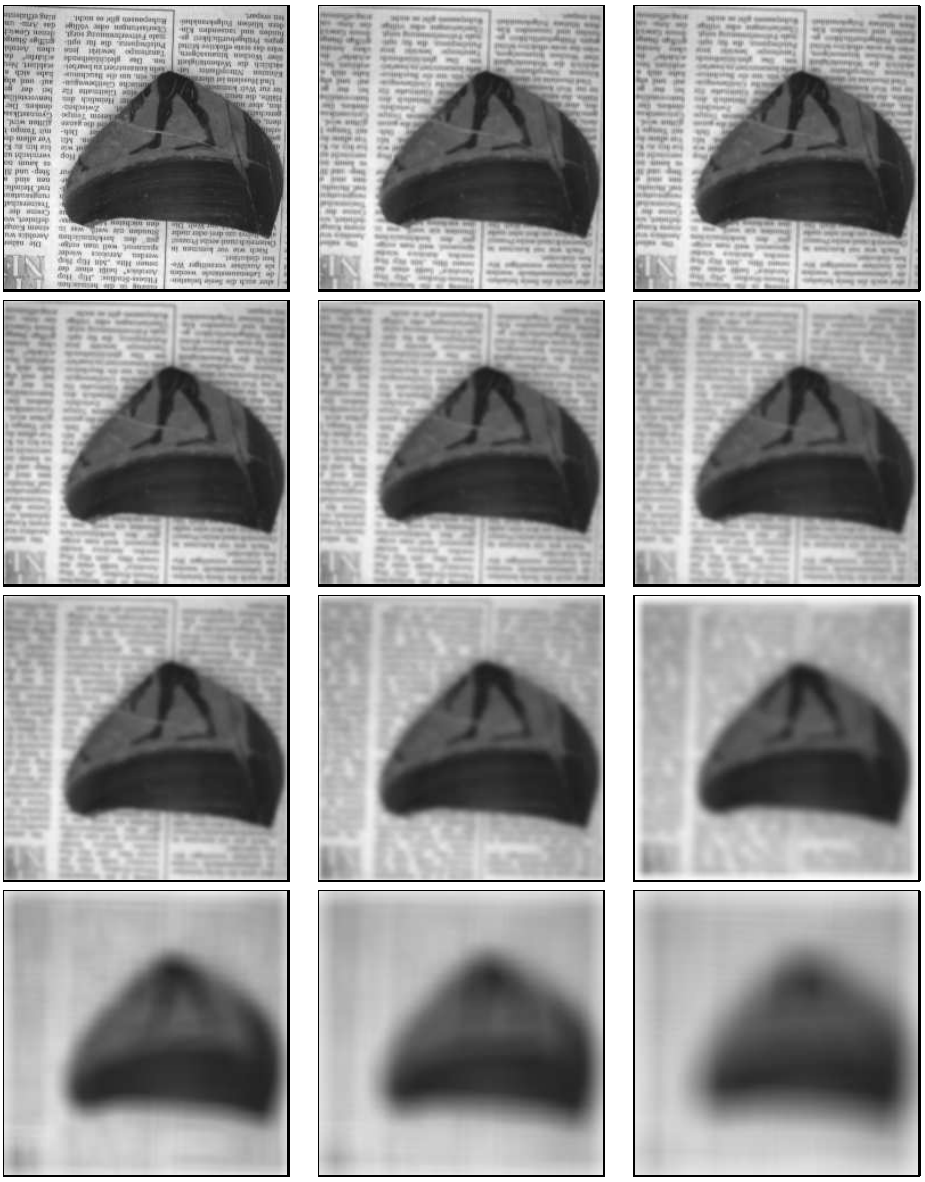


Figure 4.7: Gray-level images of an archaeological sherd at scale levels t=0, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024 (from top left to bottom right).

4.1.3 Multi-Scale vs. Pyramid Representation

The main differences between scale-space representation and pyramid representation are depticted in Table 4.1. The scale-space representation preserves the same resolution (spatial sampling) at all scales. Whereas the pyramid represen-

Pyramid	**Scale space**
decreasing image size	same resolution at all scales
(−) fixed sampling step	(+) continuous scale parameter
(+) efficient → decreasing image size	(−) redundant → same resolution at all scales
(−) not translation invariant	(+) translation invariant
(−) tracking of features difficult	(+) simple tracking of features through scale

Table 4.1: Differences between scale-space and pyramid representations.

tation reduces the number of grid points from one level to the next. A pyramid representation is efficient through the fact that it decreases the image size, while a scale-space representation becomes more redundant with increasing scale parameter. The decreasing image size in pyramids reduces the computational time both in the actual computation and in the subsequent processing step. The memory requirements are small and there are commercially available implementations of pyramids in hardware. On the other hand, in a scale-space representation, the representations of all levels of scale are immediately accessible without any need for further computations. Accessing the data is simplified, since features existing at a coarse scale will correspond to a larger number of grid points than features at finer scale, whereas in pyramid representation there is a fixed relation between the scale parameter and the resolution. Moreover, in contrast to pyramids, scale-space representation is invariant to translations [Lin94](*Lemma 4.5*). Another important property of scale-space is that the behavior of structure across scales can be described analytically with a simple formalism, which means that features at different scales can be related to each other in a precise manner [Lin94](ch. 8).

Pyramid representation implies a fixed sampling step in scale or resolution that cannot be decreased, whereas the scale-space concept provides a continuous scale parameter. Therefore the task for tracking features across scales is easier in a scale-space representation than in a pyramid representation, since refinements of the scale sampling can be performed whenever required.

4.2 Hierarchical Matching Using Image Pyramids

In addition to the used constraints and consistency checks several control strategies have been proposed by many researchers to reduce ambiguous correspondences and enhance stereo matching. These include for instance hierarchical matching strategies [LB88, MT89, MB95]. In hierarchical stereo systems matching takes place between more than one level of image description. In the hierarchical approach the reduction of information is achieved by resampling of the original signal.

In this section an area-based stereo algorithm using image pyramids is presented. For each point in the left stereo image the matching method determines the corresponding point in the right stereo image. This method can also be applied from the right stereo image to the left one. The matching method, which applies both strategies (left $\rightarrow$ right and right $\rightarrow$ left) is called bidirectional matching. The advantage of this matching method is that occluded regions can be detected for both cameras.

For a reconstruction of a surface of an object it is sometimes essential to compute dense disparity maps defined for every pixel in the entire image. In order to get a dense disparity map and to increase the efficiency of the algorithm, $5 \times 5/4$ Gaussian image pyramids are used to solve the correspondence problem in a hierarchical manner [Men91, Kro91, RK82]. The disparity maps $D^n(x_L, y_L)$ for each pyramid level n are computed as follows:

$$D^{n-1}(x_L, y_L) = \begin{cases} |x_L - x_R| & if\ max\{C^{n-1}(x_L, y_L, x_R, y_R, w)\} > T \\ 2 * D^n(x_L, y_L) & else \end{cases}, \tag{4.9}$$

where T defines the threshold accepting a corresponding point and C is defined by equation (2.32). Figure 4.8 illustrates the principle of this approach. Fast stereo evaluation is carried out at the top level n (due to low resolution) and yields a disparity map the information of which is considered in the evaluation process of level $n-1$. Again, the disparity map computed for level $n-1$ contributes to the stereo evaluation for level $n-2$ and this process is iterated until the disparity map for the lowest level is reached. If no corresponding point can be found for a candidate in the left image, the information of the pyramid level above is used to get an average disparity information for that point. The algorithm for this hierarchical approach using for instance $5 \times 5/4$ Gaussian image pyramids is defined as follows:

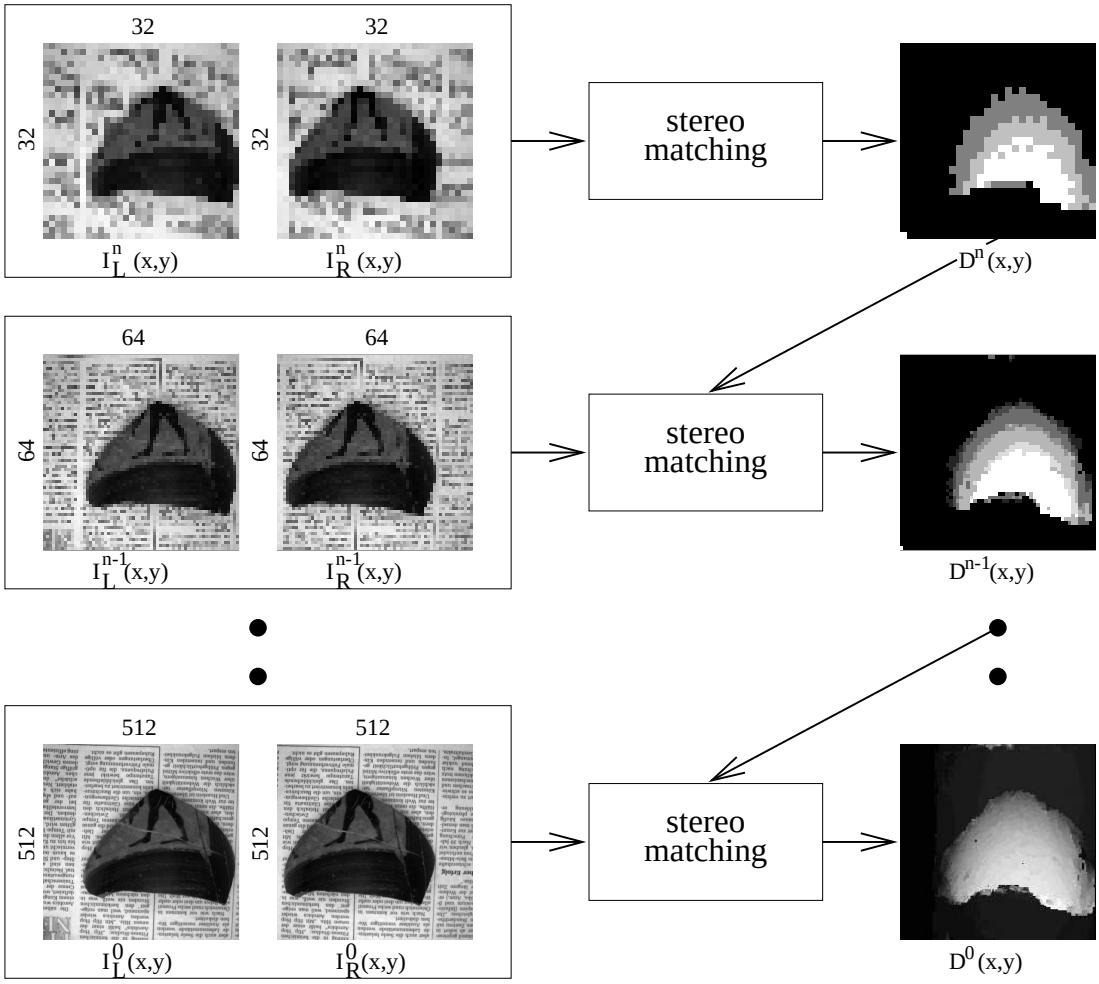


Figure 4.8: Hierarchical matching algorithm.

1. Generate Gaussian image pyramids for I_L and I_R, and start the standard stereo matching algorithm at the top level n using a fixed window size w. $\Rightarrow D^n(x,y)$

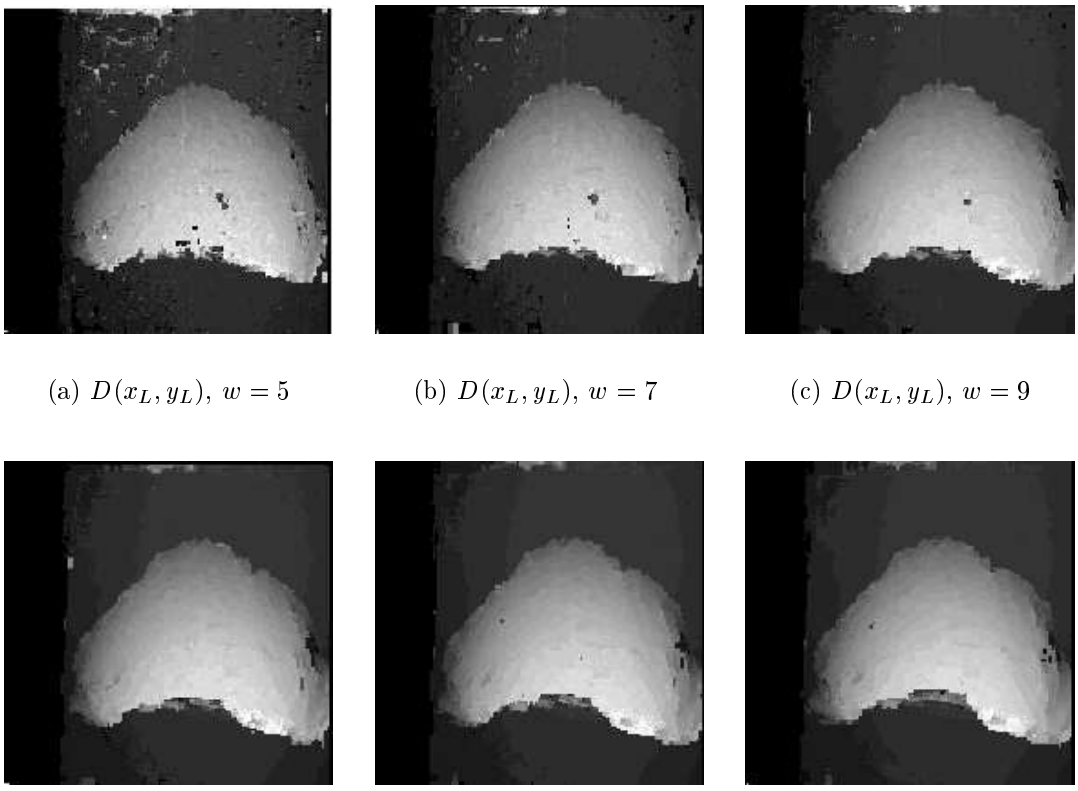
2. Compute the disparity map $D^{n-1}(x,y)$ using the disparity information of level $D^n(x,y)$ to limit the search space (see equation (4.9)).

3. If no corresponding point for $D^{n-1}(x,y)$ can be found, the value of $D^n(x,y)$ is down-projected with $D^{n-1}(x,y) = 2 * D^n(x,y)$.

4. Iterate steps 2 and 3 until the bottom levels of the pyramids are reached.

This coarse to fine matching method using image pyramids has the advantage of being relatively stable against regions with low gray-level variations in the stereo pair. For these regions an average disparity information can be computed at

a higher pyramid level using the same window size w and is down-projected in the further iteration process to the bottom level, thus getting a dense disparity map of the observed object. Furthermore the algorithm is speeded up by using this technique, because the disparity limit in the pyramid levels can be reduced to very small regions. The disadvantage of this technique is that errors which are produced in the higher levels of the pyramids are also down-projected to the bottom level.



(a) $D(x_L, y_L)$, $w = 5$ (b) $D(x_L, y_L)$, $w = 7$ (c) $D(x_L, y_L)$, $w = 9$

(d) $D(x_L, y_L)$, $w = 11$ (e) $D(x_L, y_L)$, $w = 13$ (f) $D(x_L, y_L)$, $w = 15$

Figure 4.9: Disparity maps computed with different window sizes w.

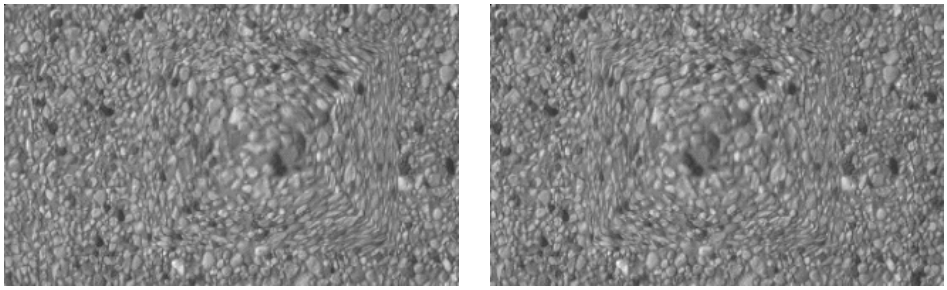
The algorithm is tested for different window sizes. The disparity maps are depicted in Figure 4.9. It can be seen that for small w there exist regions in the disparity map with no (black gaps) and incorrect disparity information. The reason for these gaps without disparity information is that these regions contain low or no gray-level variations, thus no unique maximum in the correlation function can be determined. Using larger window sizes the gaps with no disparity information can be eliminated and mismatches can also be reduced. But using larger

search windows has the disadvantage that depth discontinuities are smoothed, thus only a global disparity information can be achieved. Hence the window size should be changed depending on the local information in the stereo pair.

4.3 Coarse to Fine Matching

The coarse to fine strategy differs from the hierarchical approach in an important way. Although the matching in lower levels is constrained by the result from the higher levels in both methods, the reliability of the results at the higher levels is achieved through different means. In a coarse to fine approach it is obtained by a change of scale, whereas in a hierarchical approach using image pyramids there is a fixed relation between the scale parameter and the resolution.

In coarse to fine analysis, disparity information obtained at coarser scale is used to guide and limit the search space for the matching process [Nis84, Gri85]. In this approach the initial matching begins at a coarse scale, similar to the hierarchical approach, where the feature density is low due to the scale change. This reduction of information makes the matching process easier, but not necessarily more accurate, because the localization of the correlation maximum at coarse scales is less accurate. A way to compute disparity information from a stereo pair using a multi scale approach is to change the size of the search window w during the matching process for the same region, by keeping the resolution of the stereo images. In Figure 4.10 a synthetic stereo pair is depicted which consists of



(a) left image (b) right image

Figure 4.10: Synthetic stereo pair: Pyramid on a flat ground with natural texture added on the surface.

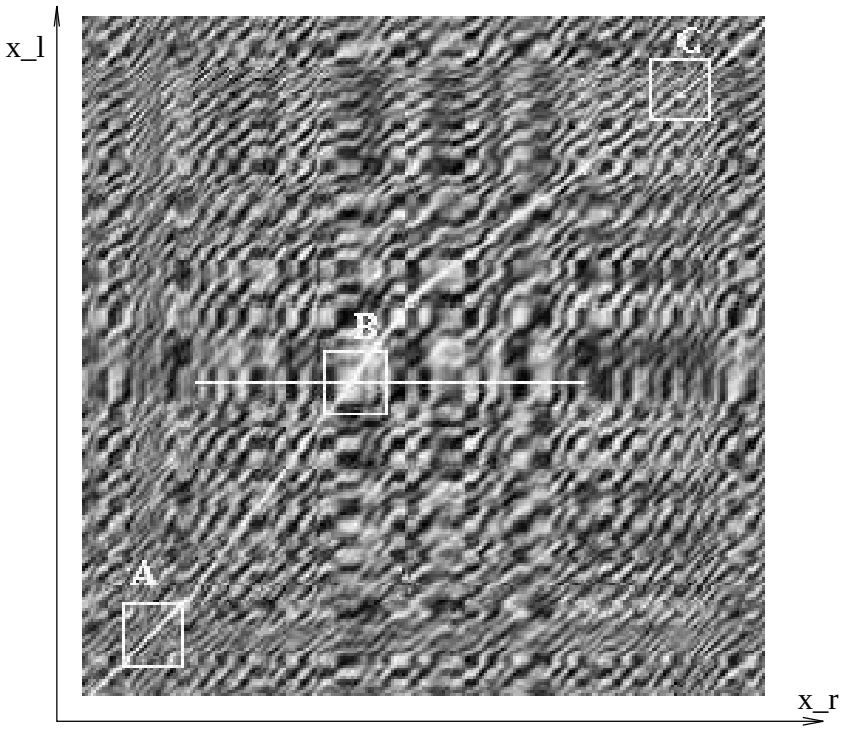
a pyramid on a plane ground with natural texture added on the surface.

Since the size of the search window for the matching process depends on the local information in the stereo pair, the window size should be changed for certain regions. In order to show that, for each pixel x_L on a scanline in the left stereo image I_L the correlation values for all pixels x_R along the scanline (epipolar line)

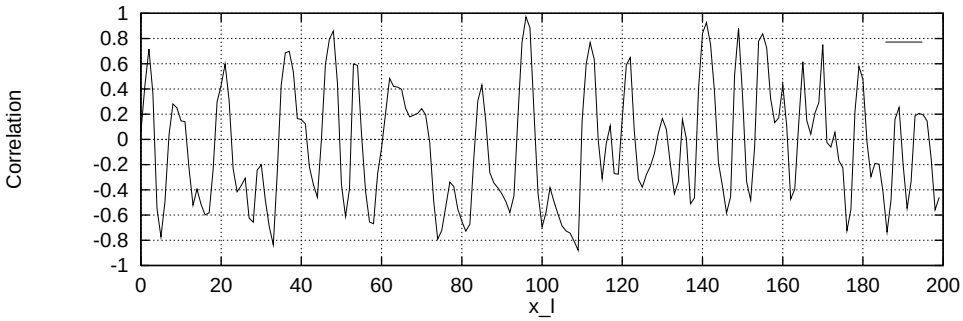
in the right image using a fixed window size w are computed. The result is a two-dimensional correlation function, which is visualized in Figure 4.11. The higher the correlation values are, the brighter the image. The path of the correlation function C from the left bottom to right top (correlation maxima) defines the corresponding points along the scanline in the right image. There are gaps in regions along the path, where no exact correlation maximum can be detected or regions where the path is smoothed. The gaps occur if the size of the search windows is too small. For the smoothed regions in the path the window is too big. In Figures 4.12 (A)-(C) parts of the correlation function of Figure 4.11 are zoomed out for a better visualization. In these figures the correlation maxima vary depending on the local information in the stereo pair. Each row in these zoomed images defines a correlation function for a point on the scanline in the left stereo image. The goal is to find a unique correlation maximum for each row in these images. For some regions the size of the search window is correct, which is depicted in Figure 4.12 (A). In these regions a unique maximum can be determined. In some regions the window is too big as shown in Figure 4.12 (B), where the region near the correlation maximum is smoothed, thus determining the exact position of the maximum is not possible for this region. In regions where the size of the window is too small as in Figure 4.12 (C) no correlation maximum can be determined, because the window does not cover enough gray-level variations. If the size of the search window is too big or too small, no unique correlation maximum can be detected so that there is always a trade-off between decreasing and increasing the size of the search windows for the matching process:

- **Small search windows**
  contain only a small number of data points, and thus are very sensitive to noise and therefore result in false matches. Furthermore the determination of the correct correlation maximum is difficult, because the correlation function may have multiple maxima if local regions around other points are similar to that which is currently under consideration.

- **Large search windows**
  contain data from two or more different objects or surfaces, thus the computed disparity value is wrong. Furthermore the determination of the exact position of the maximum is difficult, since the correlation function is smoothed. In general, large search windows are used to get global range information.

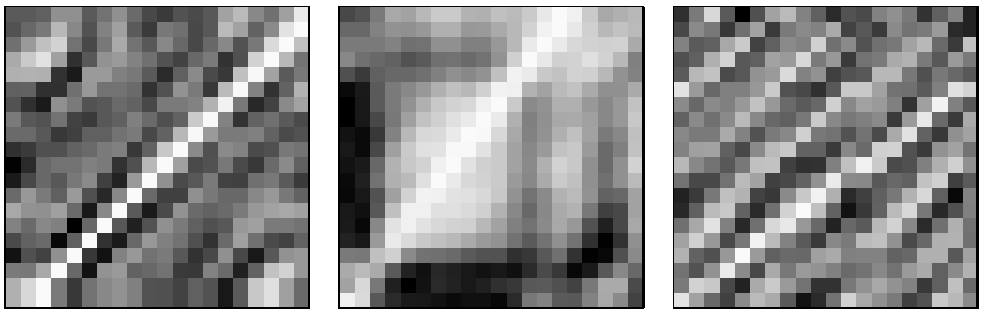


(a) C function for a line in I_L



(b) vertical profile of (a)

Figure 4.11: (a) Two-dimensional correlation function for a complete scanline in the left stereo image by using a fixed window size w=7 and in (b) a profile through this function defined by the white horizontal line in (a). Marked windows A,B,C are zoomed in Figure 4.12



(A) exact window (B) w too big (C) w too small

Figure 4.12: Zoomed regions: It can be seen that the size of the search window used is not optimal for certain regions. (A) For this region the size of w is exact for (B) too big, and for (C) the size of the search window should be increased.

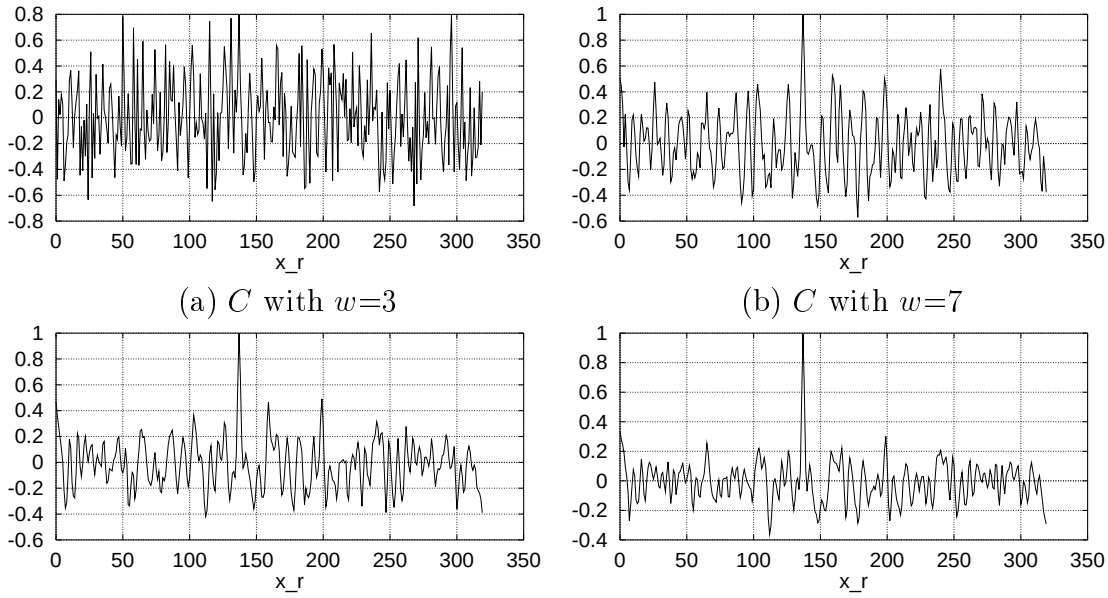
If for one point in the left stereo image the correlation function C is computed along the right scanline for different window sizes w, the function becomes smoother for larger w. In Figure 4.13 the correlation function is computed at the same position for four different window sizes w. It can be seen that for a small size w=3 it is not possible to determine a unique maximum. By increasing the window size the correlation function becomes smoother and the correct maximum survives. Using larger search windows has the advantage of getting a global disparity information of the scene but the disadvantage that the disparity information does not represent the accurate matching result due to different projective distortions in the left and the right stereo images.

The strategy is to find a corresponding point with the smallest window size w. In order to get a good initial estimate for the disparity, C is computed for large windows w to get a global range information. Using this estimate the disparity value is refined by reducing the window size. The estimate can be used to reduce the search space along the epipolar line, thus the algorithm can be outlined as follows:

1. Compute the correlation function C using a large search window w_{max} and determine the disparity, which is defined by the position of the maximum. This value is used as initial estimate. If no unique maximum can be determined there is no corresponding point.

2. Decrease the window size $w = w - \Delta w$, where $\Delta w = 2$.

3. Refine the position of the maximum by recomputing the correlation function C using the new window size w for the region of interest using the previously estimated position of the maximum.



(a) C with w=3 (b) C with w=7

(c) C with w=11 (d) C with w=15

Figure 4.13: The correlation function gets smoother with larger window sizes w. The global maximum can be located more easily in functions where a larger window is used.

4. Iterate steps 2 and 3 until no unique maximum can be found or w_{min} is reached.

The problem to find the optimal window size for certain regions in stereo images cannot be solved sufficiently with this method, since the scale can only be changed in a discrete way.

There exist various works dealing with this problem; Levine et al. use an adaptive correlation window, the size of which varies inversely with the variance of the region which is currently considered [LOY73]. In another adaptive approach Kanade and Okutomi proposed that the size and shape of the matching window is chosen adaptively on the basis of local evaluation of the variation in both the intensity and the disparity [KO94]. In all these works the search window is rectangularly shaped and the size is changed in a discrete way.

In this thesis a new method is proposed by changing the size of the window in a continuous way, thus making it possible to determine an optimal size for a given region.

4.4 Adaptive Matching in Correlation Scale-Space

The problem of changing the size of the search window during the matching process depends on the objects which are considered. If a window contains data from two or more objects or surfaces the correlation for that region cannot be maximized, unless the window is decreased and contains only data points from one single object. Another problem occurs when a search window contains occluded regions. In this case the computed disparity value is not correct. In order to find an optimal size of the search window the function C has to be modified in such way that the scale can be changed in a continuous way.

4.4.1 Correlation with a Weighted Function

In order to find an optimal window size for a certain region the window is changed in a continuous way by defining the correlation with a weighted function. The correlation function C (Eq. 2.29) is modified in such a way that the function $\delta_{1/w}$ is substituted by the one-dimensional Gaussian kernel. In the following,

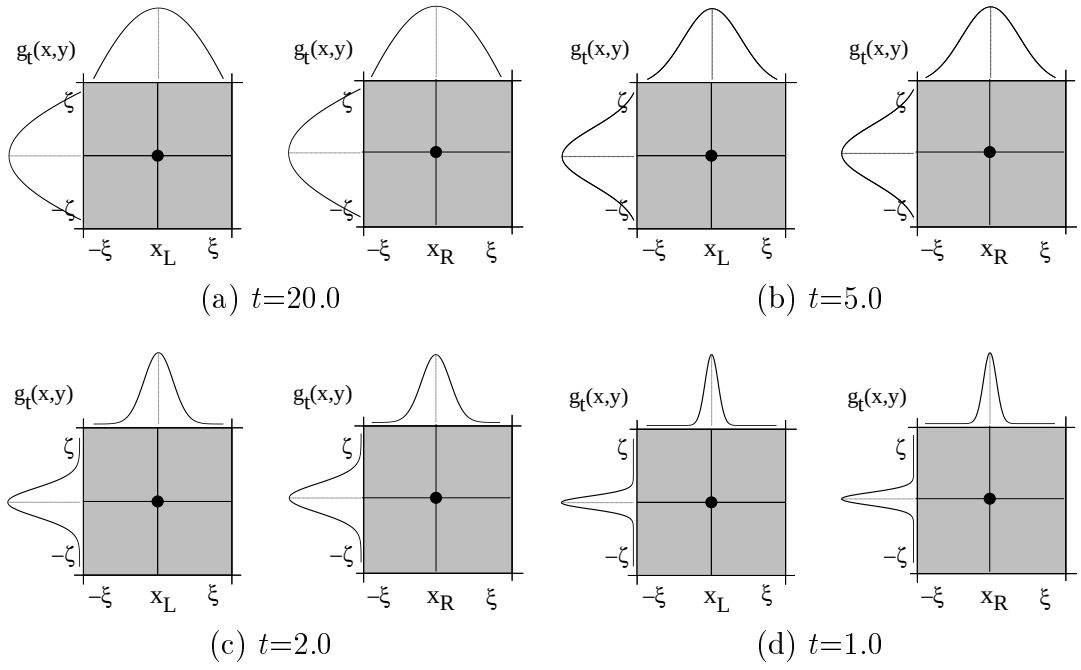


Figure 4.14: Correlation at different scales: The correlation is weighted with the Gaussian kernel (a) - (d) for values t=20,10,5 and 1.

the scale-space notation[1] is used, where the scale is defined by t. For $I : \mathbb{R} \mapsto [0,1]$ and $\mu : \mathbb{R} \times \mathbb{R} \mapsto [0,1]$

$$\mu(x;t) \;=\; \int_{-\infty}^{\infty} I(x-\xi)g(x+\xi,t)\,d\xi = I(x) * g(x;t)\;, \tag{4.10}$$

with

$$g(x;t) \;=\; \frac{1}{t\sqrt{2\pi}} e^{\frac{-x^2}{2t^2}}. \tag{4.11}$$

The standard deviations are

$$\sigma_k^2(x;t) \;=\; I_k^2(x) * g(\bullet\,;t) - \mu_k^2(x;t) \qquad k = L,R\;, \tag{4.12}$$

and the covariance can be written as

$$\sigma_{LR}^2(x_L,x_R;t) \;=\; [I_L(x_L)I_R(x_R)] * g(\bullet\,;t) - \mu_L(x_L;t)\mu_R(x_R;t) \quad k = L,R\;. \tag{4.13}$$

The correlation can be written as convolution with the Gaussian kernel in the one-dimensional case (4.11) as follows:

$$C_\Gamma(\bullet\,;t) \;=\; \frac{[I_L(x_L)I_R(x_R)] * g(\bullet\,;t) - \mu_L(x_L;t)\mu_R(x_R;t)}{\sqrt{[I_L^2(x_L) * g(\bullet\,;t) - \mu_L^2(x_L;t)][I_R^2(x_R) * g(\bullet\,;t) - \mu_R^2(x_R;t)]}}. \tag{4.14}$$

For the two-dimensional case the Gaussian kernel

$$g(x,y;t) \;=\; \frac{1}{2\pi t^2} e^{-\frac{x^2+y^2}{2t^2}} \tag{4.15}$$

is used and the weighted correlation function can be written as

$$C_\Gamma(\bullet,\bullet\,;t) \;= \tag{4.16}$$
$$\frac{[I_L(x_L,y_L)I_R(x_R,y_R)] * g(\bullet,\bullet\,;t) - \mu_L(x_L,y_L;t)\mu_R(x_R,y_R;t)}{\sqrt{[I_L^2(x_L,y_L) * g(\bullet,\bullet\,;t) - \mu_L^2(x_L,y_L;t)][I_R^2(x_R,y_R) * g(\bullet,\bullet\,;t) - \mu_R^2(x_R,y_R;t)]}},$$

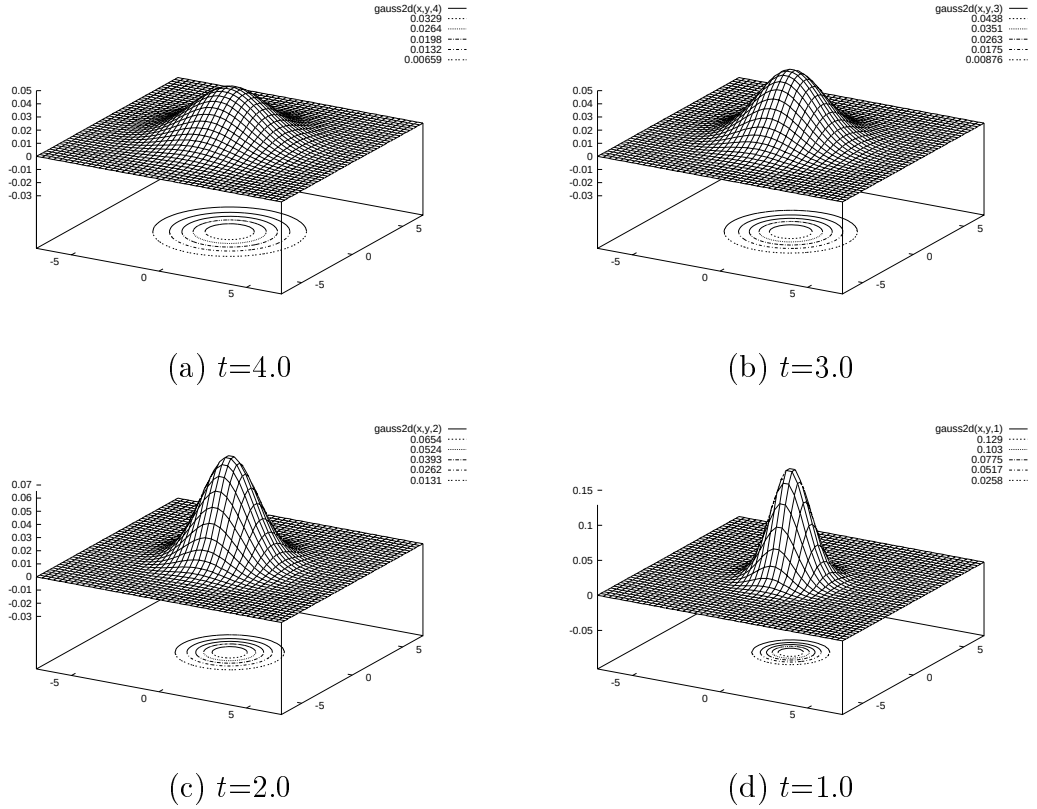
where

$$\mu_k(x,y;t) \;=\; I_k(x,y) * g(\bullet,\bullet\,;t) \qquad k = L,R \tag{4.17}$$

are the mean values.

In Figure 4.14 this principle is visualized schematically. Instead of using all the data points in the correlation window equivalently weighted, the window is weighted with the Gaussian kernel. In Figure 4.15 the two-dimensional Gaussian kernel $g(\bullet,\bullet\,;t)$ for different values t is visualized. The size of the search window can be controlled by the scale parameter t. Furthermore the shape of the search window has changed from rectangular to circular, which can be seen in

[1] The notation $C_\Gamma(\bullet\,;t)$ stands for $C_\Gamma(x_L,x_R;t)$ and $C_\Gamma(\bullet,\bullet\,;t)$ for $C_\Gamma(x_L,y_L,x_R,y_R;t)$.



(a) t=4.0 (b) t=3.0

(c) t=2.0 (d) t=1.0

Figure 4.15: Two-dimensional scale-space kernels for different values $t = 4, 3, 2, 1$. The shape of the window using these kernels is circular.

Figure 4.15. The function C_Γ defines a **Correlation Scale-Space** (*CSS*) for one point $I_L(x, y)$ in the left stereo image. In the *CSS* the similarity value is available at different scales driven by the parameter t of the scale-space kernel. The main advantages of the *CSS* compared to standard methods, such as the hierarchical approach, is that the scale can be changed in a continuous way. Furthermore in this representation all levels of scale are immediately accessible.

If this function is applied for one position in the left stereo image (see Figure 4.10) a correlation scale-space is generated for the complete right image for which the scale can be changed by the parameter t in a continuous way. The result is depicted in Figure 4.16, where the correlation function C_Γ is computed at different scales t. The cross-sections of the *CSS* at different scales are visualized as images where the correlation values are represented by gray-values.

It can be seen that at a lower scale no unique maximum can be determined. By increasing the scale parameter t the correct maximum centered in the image survives, whereas all other maxima disappear successively in higher scales.

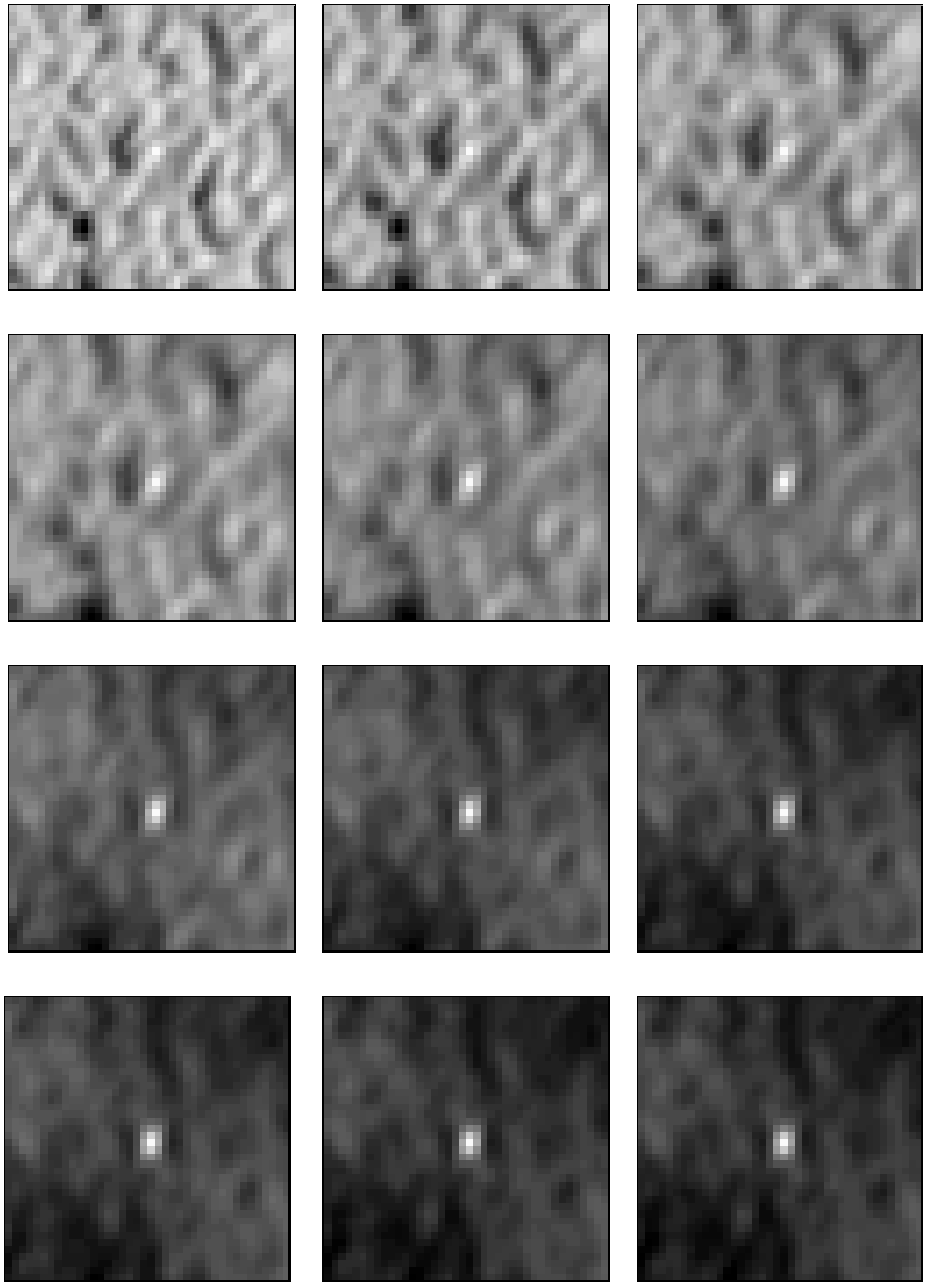


Figure 4.16: Correlation Scale-Space: C_Γ is applied at one position in the left image on the complete right stereo image at different scales $t = 2^n$ for $n = 0, \ldots, 11$. It can be seen that the correlation maximum in the center of the images survives, whereas all other maxima become successively smaller by increasing the scale parameter t.

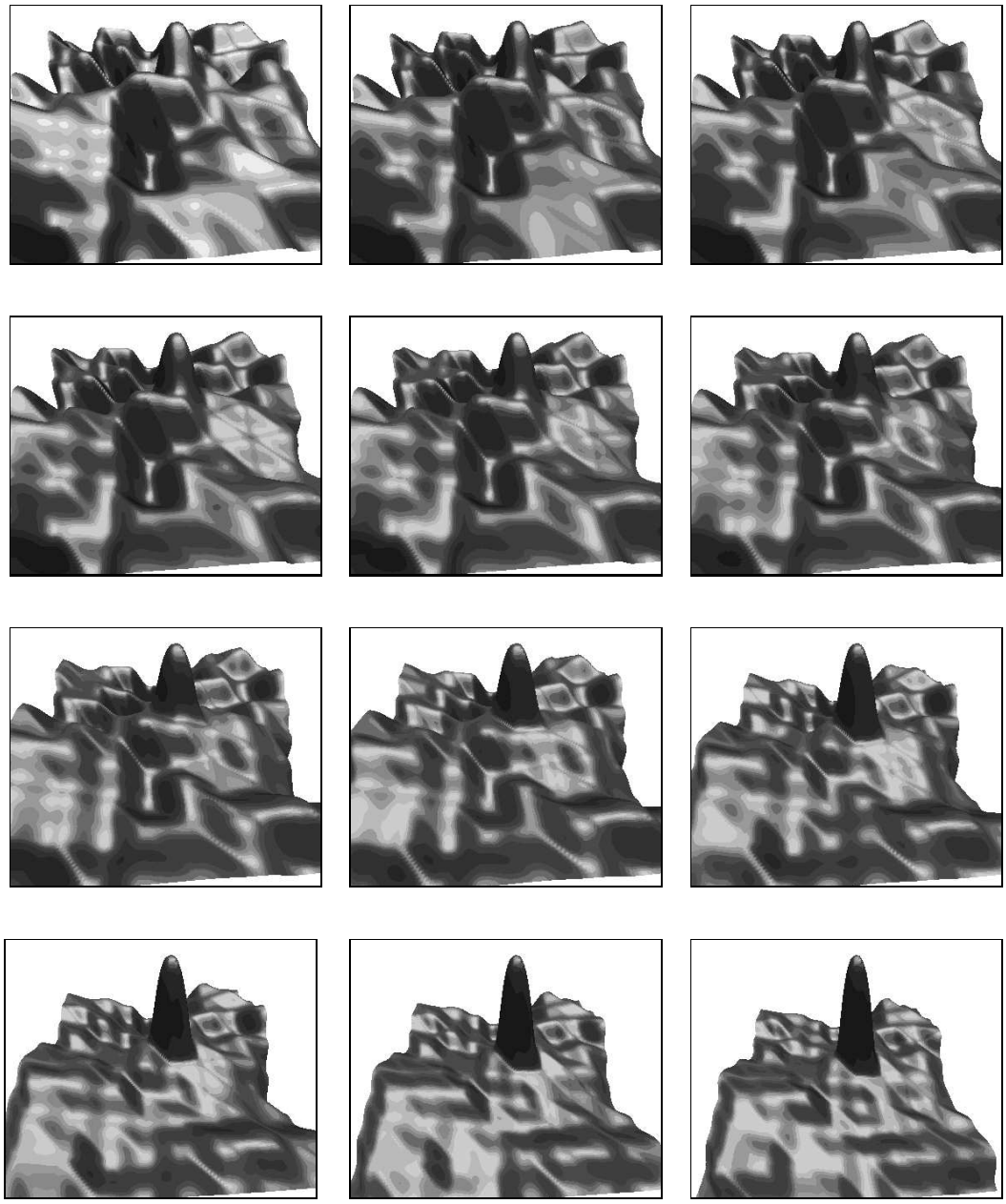
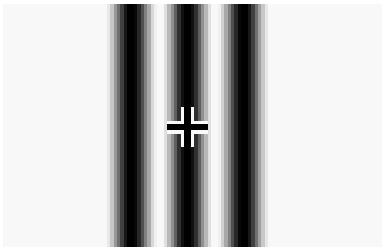


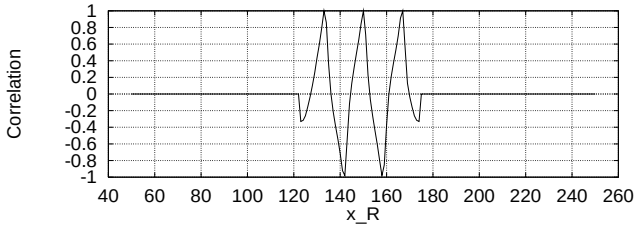
Figure 4.17: 3D plots of the Correlation Scale-Space: C_Γ is applied at one position in the left image on the complete right stereo image at different scales $t = 2^n$ for $n = 0, \ldots, 11$.

In Figure 4.17 the three-dimensional plots of Figure 4.16 are depicted. By using the epipolar constraint the search space can be reduced. So for one point in the left stereo image the correlation function C_Γ along the right scanline is computed at different scales t. Lindeberg stated that in the scale-space for one-dimensional signals local extrema are successively suppressed whereas for two-dimensional signals new maxima may be created. An example is given in section 4.1.2. The reason for these new maxima is called fine-scale phenomenon.

New maxima can also be created in the *CSS* by increasing the scale parameter t. In order to give an example a cosine function is used. The left stereo image I_L is depicted in Figure 4.18 (a). For one point in the image the corresponding



(a) horizontal cosine function

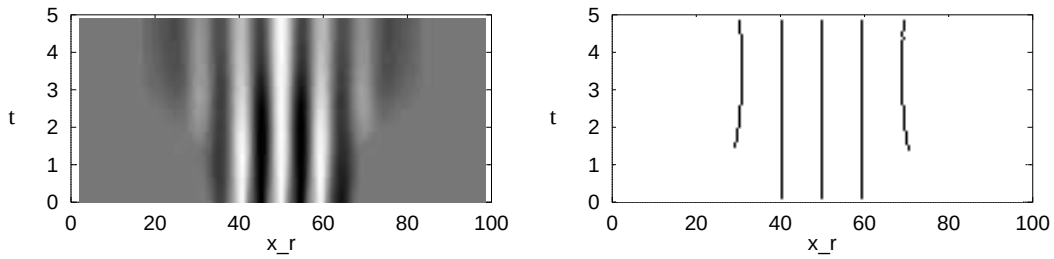


(b) C_Γ with $t = \pi/2$

Figure 4.18: Ambiguous correspondence: The standard correlation method is tested on a horizontal cosine function. It can be seen that no unique maximum can be determined.

point is searched along the scanline using the function C_Γ with scale $t = \pi/2$. The result is visualized in Figure 4.18 (b). It can be seen that for the given point in the left image, which is marked with a cross, three correlation maxima are created with exactly the same value. In this case it is impossible to decide which of the three candidate points corresponds to the point in the left image. For this example it is important to change the scale to get a unique solution. The same example is tested for different scales t. In Figure 4.19 (a) the *CSS* for different scales is depicted and in Figure 4.19 (b) the zero crossings of the first derivative are shown (only maxima). It can be seen that new maxima are created from low to high scale. But the correct maximum survives, whereas the other two maxima are suppressed successively by enlarging the scale parameter t.

If the search window contains only data from one period (of the cosine function) three maxima are created. If the scale $t > \pi/2$ information of the left and the right period is included into the computation of the correlation. Thus the region, which is considered is getting unique.



(a) *CSS* for one point

(b) Zero crossings of the first derivative (only maxima)

Figure 4.19: (a) *CSS* for one point in the valley of the cosine function for different scales, and in (b) the zero crossings of the first derivative. It can be seen that by increasing the scale new maxima are created, but the correct maximum survives whereas the other maxima are suppressed. Thus for this example a unique maximum can be determined.

4.4.2 Scale Selection

In general situations it is not possible to know in advance at what scales interesting structures can be expected to appear. Size variations of image structures in a stereo pair can occur for several reasons:

- objects in the scene have different physical size;
- surface textures contain structures at different scales; and
- scale variations appear due to perspective distortions during image formation.

If it is not clear what scale to choose, the correlation function C_Γ is computed for all scales. For a first stage this is a logical strategy, for later stages there must be some scale selection to find a size for a given operator, which has maximum response. There are many ways to select the best scale for a given problem. A very interesting work in this field was presented by Lindeberg in which he describes the "scale-space primal sketch" [Lin94]. An operator gives maximal output if its size is best tuned to the object. Other approaches study the variation of the information content over scale [Jae95]. For the correspondence establishment it is possible so far to change the scale parameter t in a continuous way using the correlation function C_Γ. But the problem to be solved is to find the "best scale(s)" t for certain regions in a stereo pair. Basically, the scale at which a maximum over scales is attained will be assumed to give information about the

window size for that region. The maximum over scale for a region defines the optimal scale.

In the next step the *CSS* for different placements of a point is analyzed:

- on a plane parallel to the image plane;
- on a roofed surface;
- near a depth edge; and
- in an occluded area

For these situations the correlation values at the corresponding position in the right stereo image are tested by tracing along the local maxima in the direction of the scanline from high to low scale. In Figure 4.20 the correlation function at

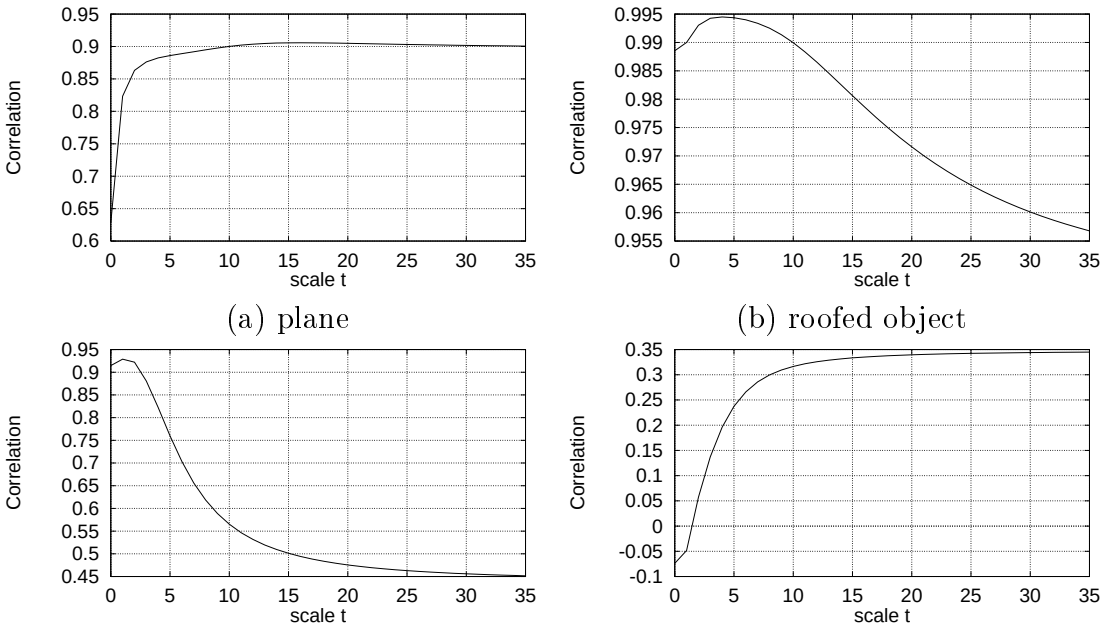


Figure 4.20: Correlation as a function of scale for a point on (a) a plane parallel to the camera, (b) a roofed object, (c) near a depth edge and (d) in an occluded region.

the corresponding point along scale t is visualized for these different situations.

Plane: For the plane lying parallel to the image plane there are only small variations along the correlation function. The correlation value decreases if no or not enough gray-level information is available in the correlation window. The

scale-space maximum for this example lies at t=15, with $C_\Gamma(\bullet,\bullet\,;t) \in [0.6..0.9]$, which means that for this scale the correlation value can be maximized.

Roofed surface: In the case of a roofed surface, by tracing from large to smaller scale the value of the correlation increases successively until it decreases because of too little information in the correlation window. For this example a small scale obtains the highest correlation value at scale t=3, with $C_\Gamma(\bullet,\bullet\,;t) \in [0.95..0.99]$.

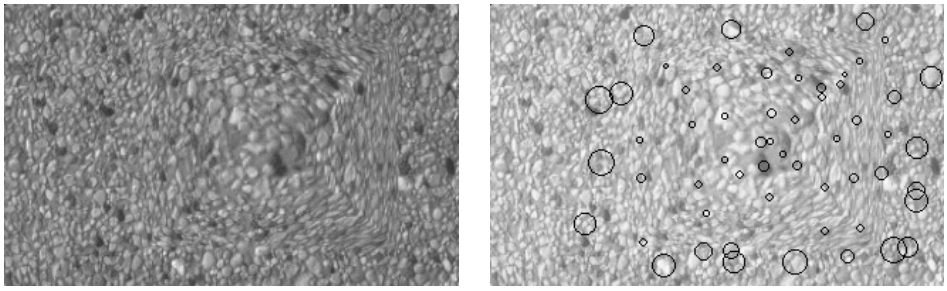
Depth Edge: Near a depth edge it is similar. At larger scales the correlation value is low because the window contains data from two objects which have different disparity values. By decreasing the scale parameter t the correlation value is maximized at a very low scale, since the window contains only data from one single object. The obtained scale for this situation lies at $t = 1.5$, with $C_\Gamma(\bullet,\bullet\,;t) \in [0.45..0.93]$

Occluded area: In the last test the correlation value over scale for an occluded region is determined. It can be seen in Figure 4.20 (d) that the maximum of the correlation function along scale is at t =35 with $C_\Gamma(\bullet,\bullet\,;t) \in [-0.1..0.35]$. This represents the highest scale and the correlation value decreases successively to lower scale. But the correlation value is very low anyway, thus the position of this maximum represents an average depth information of the surrounding regions, which contain the occluded area. In Table 4.2 the optimal scale values, obtained for the different placements, are summarized.

Placements	C_Γ	optimal t
plane	0.915	15.0
roofed surface	0.994	3.0
near a depth edge	0.925	1.5
occluded area	0.35	35.0

Table 4.2: Optimal scales t for different placements.

Under the assumption that there is enough gray-level variation in the search window it can be stated that for surfaces, that are nearly parallel to the camera the plot depicted in Figure 4.20 (a) is typical, since the data points in the search window have nearly the same disparity, even under different scales. If a search window contains data points from two or more surfaces (roofed surface or depth edge) the correlation value cannot be maximized. For these regions a small scale should be used. For occluded regions an average disparity value is determined at high scale with a low correlation maximum, which decreases successively to lower scale. This type of plot represents an occluded region, thus the corresponding point is not accepted.



(a) original left image

(b) best scale computed for some regions of interest

Figure 4.21: *CSS* maxima: The correlation maxima for some points of interest. The radius of the circles depicts the scale used for these regions which maximizes the correlation function C_Γ over scale. The circles are superimposed on a bright copy of (a).

In order to establish correspondences, the function C_Γ is tested on the pyramid (Figure 4.10). In the left image some points of interest are chosen for which the corresponding points are determined in the right stereo image. Every scale-space maximum is graphically illustrated by a circle centered at the point in the left stereo image for which the correspondence is established. The size of the circle is determined in such a way that the area (measured in pixels) is equal to the scale at which the maximum is assumed. The circles are superimposed on a bright copy of the left stereo image. The result is visualized in Figure 4.21. For points on the surface of the pyramid the optimal scale is small, since the plane lies at a certain angle to the image plane, thus not all points in the search window have the same disparity value, whereas for regions on the flat ground the selected scale value is larger. One way to determine the optimal scale for the correspondence establishment for one point is to determine the scale-space maximum by tracing down the scale along the correlation maxima at each scale level. By tracing down a change in the direction is important, because this means a variation in the disparity value. If there is no change in the direction x_r the steps can be enlarged, otherwise refined. The algorithm can be outlined:

1. Compute the initial disparity value using the correlation function C_Γ starting at a large scale t. If no unique maximum can be determined there is no corresponding point.

2. Find the best scale t by tracing down the *CCS* using the previously estimated disparity value.

3. The global maximum along this path defines the optimal scale.

4.5 Experimental Results

The adaptive matching algorithm using function C_Γ is applied on a number of synthetic and real images and is compared to the standard stereo approach.

4.5.1 Results on Synthetic Images

As synthetic stereo pairs a **pyramid** and a **sphere** on a flat ground are used in each case with natural texture added onto the surface. For these synthetic sets of stereo pairs the disparity values are known to compare the accuracy of the matching method. For each test the accuracy is compared to the results computed with the standard matching method by using the *MSE*[2].

Figure 4.22 the synthetic stereo pair of the pyramid and in Figure 4.23 the sphere is depicted. The disparity range for the pyramid is 15 to 28 and for the sphere 15 to 22 pixels. These stereo pairs are tested first for the standard stereo method using fixed window sizes $w = 3$, $w = 7$ and $w = 15$ and with the adaptive matching method using the function C_Γ. In Figures 4.22 (c) and (d) the results for the pyramid and in Figures 4.23 (c) and (d) for the sphere are shown for fixed window sizes. The results for the adaptive approach are depicted for the pyramid in Figure 4.22 (e) and for the sphere in Figure 4.23 (e). Three-dimensional models are created using the disparity map computed with the adaptive approach and the left stereo image, which are visualized in Figures 4.22 (f) for the pyramid and in Figure 4.23 (f) for the sphere. For flat surfaces a large search window obtains good results, whereas on depth edges the disparity values are blurred. The disparity values along depth edges are more accurate using small search windows, but the disparity over the complete image is very noisy. The adaptive approach obtains good results both on depth discontinuities and on flat surfaces.

Table 4.3 illustrates the $MSE(C(w \in [3, 7, 15]) - ideal)$ and the $MSE(C_\Gamma - ideal)$.

MSE	$MSE(C(w=3)-ideal)$	$MSE(C(w=7)-ideal)$	$MSE(C(w=15)-ideal)$	$MSE(C_\Gamma-ideal)$
Pyramid	280	203	266	50
Sphere	320	280	303	65

Table 4.3: Difference between true disparity and the computed disparity.

[2]The notation $MSE(< method > (< w >) - ideal)$ is used for the comparison between the ideal and the computed disparity maps. The *MSE* is only computed for the region which is visible in both stereo images

The adaptive matching method has the smallest *MSE*. This approach reduces two types of errors,

- large random errors all over the image caused by a small search window and
- systematic errors along depth discontinuities which occur when using large search windows.

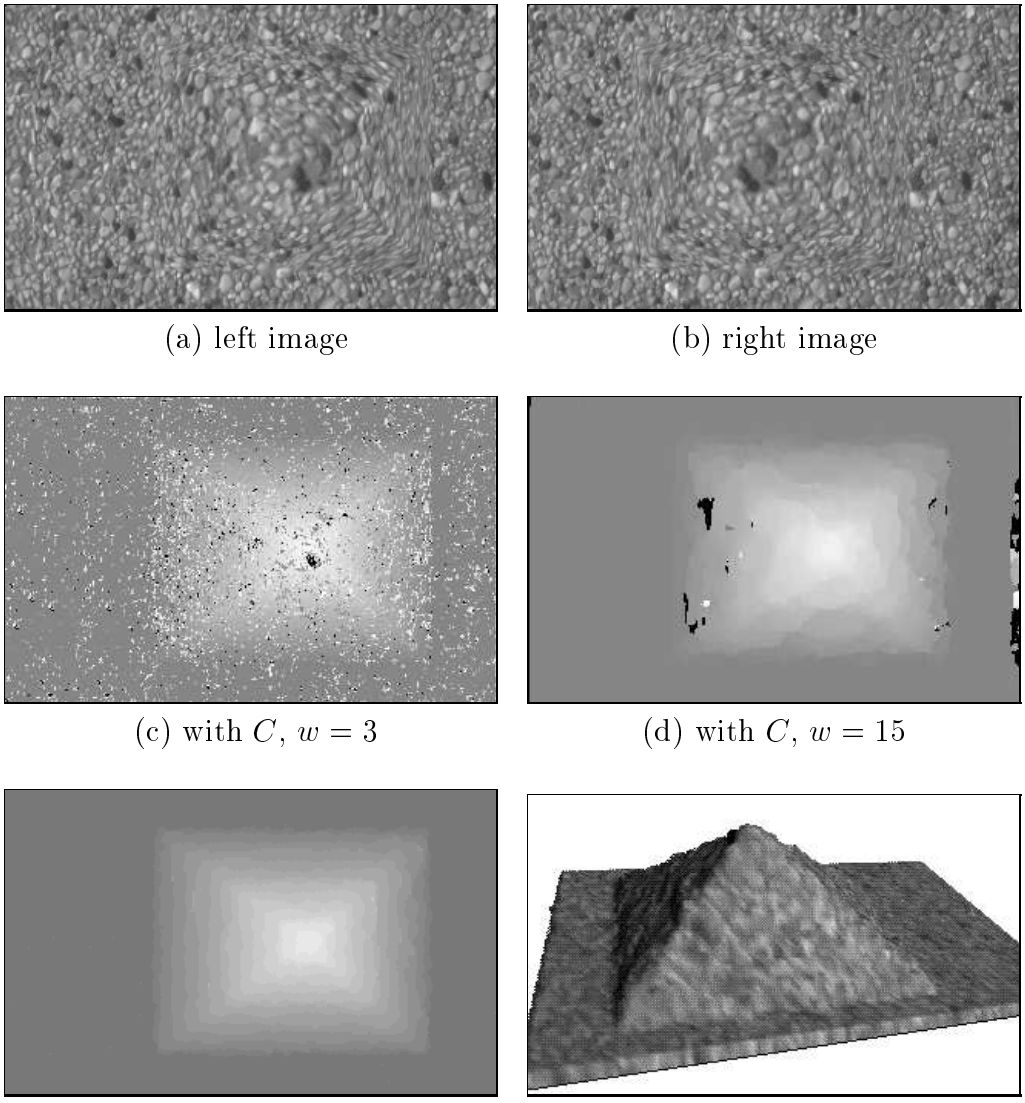
4.5.2 Results on Real Images

Archaeological sherds are used as real objects. These sherds are fragments of pots and are archived by archaeologists. In order to perform such an archivation of sherds a lot of tasks have to be performed [SMD91]. The classification process is divided into classification of shape [SM92, SMP93, MS96, MB96] and material [MT96]. Our department is involved in this project[3] [SMD94, SM95] and therefore these sherds are used as a test set.

The experimental setup consists of two 8Bit-CCD cameras with a resolution of 768×568 pixels and a 133-Pentium running OS Linux. The stereo pairs of two different sherds are depicted in Figures 4.24 and Figures 4.25 (a) and (b). The same test is applied for these real objects. In Figures 4.24 (c)-(d) and 4.25 (c)-(d) the disparity maps for the standard stereo method using fixed window sizes $w = 3$ and $w = 15$ and for the adaptive approach are depicted for both sherds.

It can be seen that the surface of both sherds is recovered. The sherds have a smoothed surface without any false matches and the edge of the sherds is also recovered. For comparison the disparity values computed with the standard method using fixed window sizes are shown in Figures 4.24 (c)-(e) and Figures 4.25 (c)-(e). A small window size produces several mismatches on the surface of both sherds whereas large search windows smooth the disparity values. Together with both gray-level and disparity information three-dimensional models of the objects with the texture added on the surface can be created, which are illustrated in the Figures 4.24 (f) and 4.25 (f).

[3]Austrian National Fonds zur Förderung der wissenschaftlichen Forschung (No. P9110-SPR)

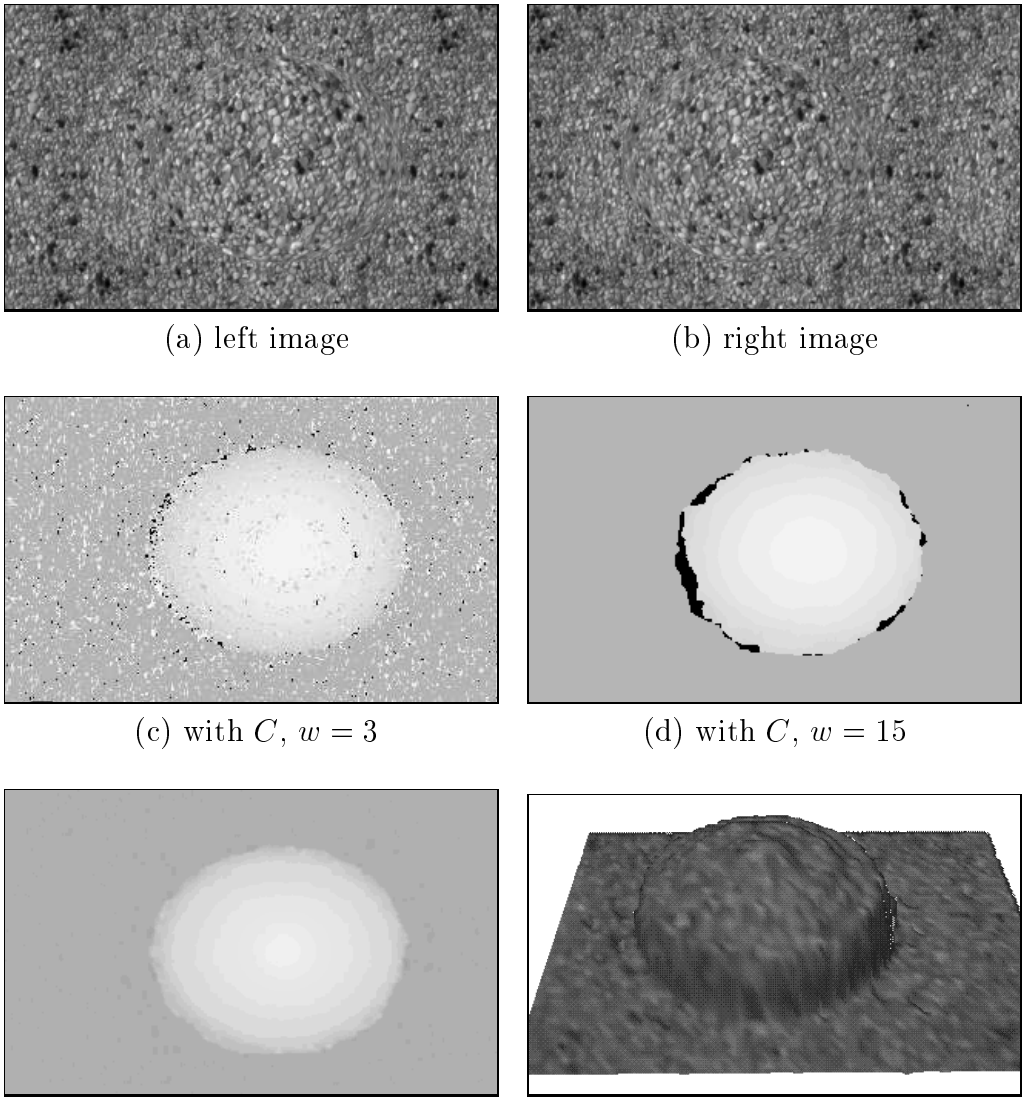


(a) left image (b) right image

(c) with C, $w = 3$ (d) with C, $w = 15$

(e) disparity map with C_Γ (f) 3D model

Figure 4.22: Pyramid on a plane ground with a natural texture added onto the surface: (a) and (b) left and right stereo image, (c)(d) disparity maps computed with the standard stereo method using fixed window sizes $w = 3$ and $w = 15$ and (e) with the adaptive approach and (f) 3D model of the object.

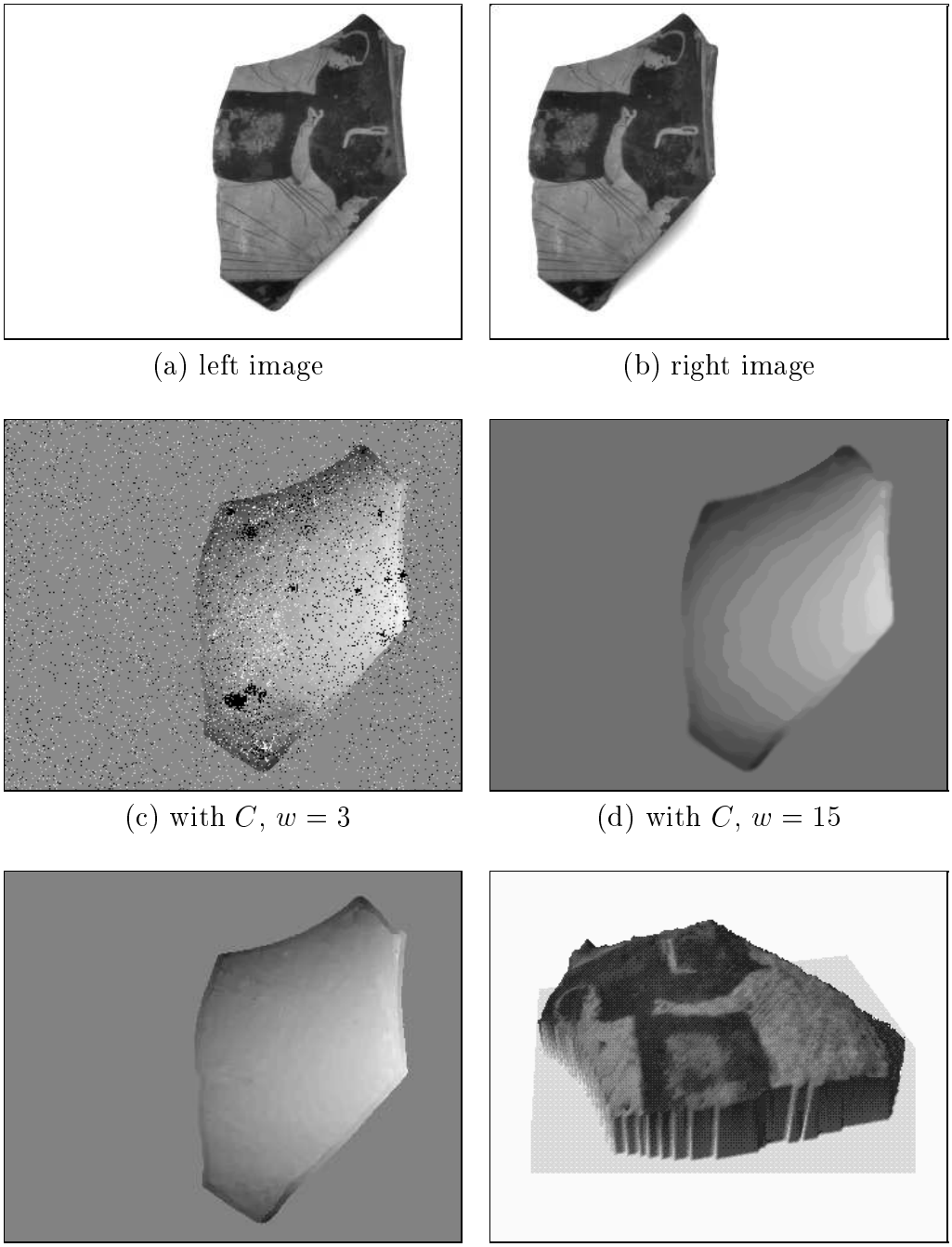


(a) left image (b) right image

(c) with C, $w = 3$ (d) with C, $w = 15$

(e) disparity map with C_Γ (f) 3D model

Figure 4.23: Sphere on a plane ground with a natural texture added onto the surface: (a) and (b) left and right stereo image, (c)(d) disparity maps computed with the standard stereo method using fixed window sizes $w = 3$ and $w = 15$ and (e) with the adaptive approach and (f) 3D model of the object.

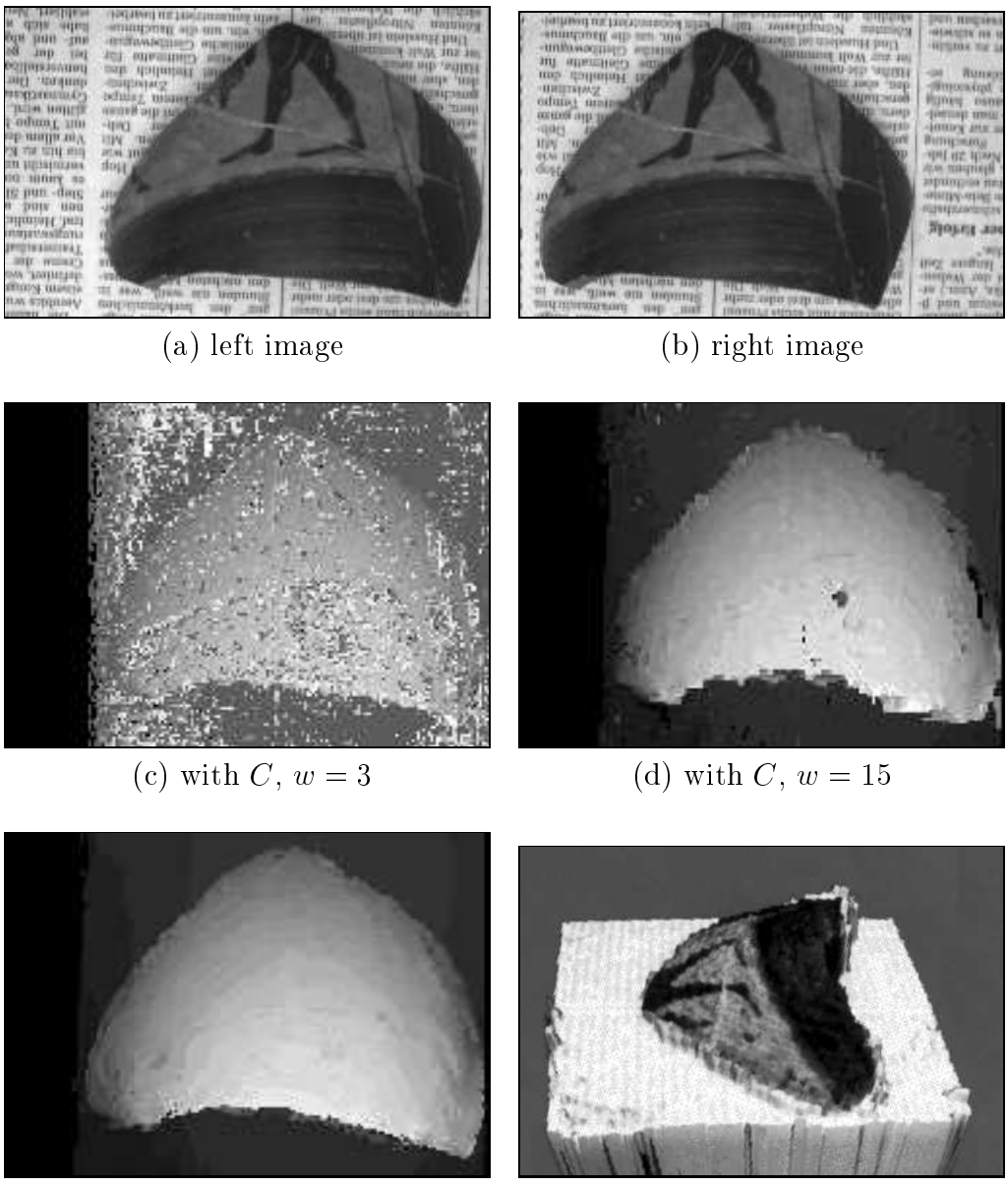


(a) left image (b) right image

(c) with C, $w = 3$ (d) with C, $w = 15$

(e) adaptive matching using C_Γ (f) 3D model

Figure 4.24: Archaeological sherd 1: (a) and (b) left and right stereo image, (c)(d) disparity maps computed with the standard stereo method using fixed window sizes $w = 3$ and $w = 15$ and (e) with the adaptive approach and (f) 3D model of the object.



(a) left image (b) right image

(c) with C, $w = 3$ (d) with C, $w = 15$

(e) adaptive matching using C_Γ (f) 3D model

Figure 4.25: Archaeological sherd 2: (a) and (b) left and right stereo image (c)(d) disparity maps computed with the standard stereo method using fixed window sizes $w = 3$ and $w = 15$ and (e) with the adaptive approach and (f) 3D model of the object.

4.6 Chapter Summary

In this chapter a new method was presented to detect the optimal scale for each region in a given stereo pair. A correlation scale-space was defined in which the scale is defined in a continuous way. For each region in the stereo pair, depending on the gray-level and disparity information, the size of the search window can be changed adaptively in a continuous way by changing the scale parameter t of the correlation scale-space. Furthermore the shape of the search window has changed from rectangular to circular. It can be stated that for regions which have different disparity values a small scale value is better, whereas for regions with low gray-level variations a larger scale value is used to maximize the correlation value over scale. The global scale-space maximum for a certain region, which maximizes the correlation value is defined as the optimal size of the search window.

Chapter 5

Conclusion and Outlook

In this work two complementary original contributions dealing with stereo have been presented, the first combines stereo techniques with robust statistics to form a robust version of the correlation; the second solves the correspondence problem adaptively in a multi-scale framework using a correlation scale-space.

The standard area-based correlation approach was modified so that it can tolerate a significant number of outliers. The approach exhibited a robust behavior not only in the presence of mismatches but also in the case of depth discontinuities. The confidence measure of the correlation and the number of outliers provided two complementary sources of information, which when implemented, resulted in a robust and efficient method. The results have been tested on synthetic images under different noise conditions and were compared to the results computed with the standard stereo method. For the robust technique the tests were also performed for some specific weighting functions. It can be noted that the robust approach obtained better results than the standard stereo method in all cases, because these results were taken as an initial estimate for the robust approach. Especially in the case of replacement noise good results could be achieved.

A central problem in stereo matching using correlation techniques lies in selecting an appropriate size for the search window. The size of the search window must be large enough to contain enough data points to have a large gray-level variation, but small enough to avoid the effect of perspective distortion. If the size of the search window is too small and does not cover enough gray-level variation, the signal to noise ratio is low, thus giving only a poor estimate for the disparity value. If the window is too large it contains data points belonging to different objects or surfaces, thus the estimated depth value is not accurate due to different projective distortions in the left and the right image. So it is all-important to provide an adaptive matching method. An new adaptive matching method was proposed in which the size of the search window could be changed in a continuous way, thus an optimal size for a given region could be estimated. This approach was tested on a set of synthetic and real images. It was shown exper-

imentally that the adaptive approach obtained better results than the standard stereo method, especially in regions with depth discontinuities.

Future work will be directed towards improving the performance of the stereo algorithm by using other robust methods. In addition a stereo algorithm is planned to incorporate both the robust and the adaptive matching technique.

Bibliography

[AAK71] Y.I. Abdel-Aziz and H.M. Karara. Direct linear transformation into object space coordinates in close-range photogrammetry. In *Symposium on Close-Range Photogrammetry*, pages 1–18, Urbana, Illinois, January 26-29 1971.

[ABH+72] D. Andrews, P. Bickel, F. Hampel, P. Huber, W. Rogers, and J. Tukey. *Robust Estimates of Location: Survey and Advances.* Princeton University Press, 1972.

[BA83] P.J. Burt and E. Adelson. The laplacian pyramid as a compact image code. *IEEE Trans. Communications*, 9(4):532–540, 1983.

[BASv86] Peter J. Burt, C. H. Anderson, J. O. Sinniger, and G. van der Wal. A pipelined pyramid machine. In S. Levialdi and V. Cantoni, editors, *Pyramidal Systems for Image Processing and Computer Vision*, volume F25 of *NATO ASI Series*, pages 133–152. Springer-Verlag Berlin, Heidelberg, 1986.

[BBW88] P. J. Besl, J. B. Birch, and L. T. Watson. Robust window operators. In *Proceedings of the 2nd ICCV*, pages 591–600. IEEE, Dec 1988.

[BF82a] S.T. Barnard and M.A. Fischler. Computational stereo. In *ACM Computing Surveys*, volume 14, pages 553–572, Dec. 1982.

[BF82b] S.T. Barnard and M.A. Fischler. Computational stereo, comput. surveys. In *Science Volume 194*, pages 283–287, 1982.

[Bra97] N. Braendle. Rektifizierung oder direkte Tiefenberechnung? Untersuchungen zur digitalen Objektvermessung mit dem Stereoverfahren. Master's thesis, TU Wien, Institut für Automation, Wien, 1997. to appear.

[Bur81] P.J. Burt. Fast filter transforms for image processing. *Computer Vision, Graphics, and Image Processing*, 16:20–51, 1981.

[BWBD86] J. Babaud, A.P. Witkin, M. Baudin, and R.O. Duda. Uniqueness of the gaussian kernel for scale-space filtering. *IEEE Trans. on PAMI*, 8(1):26–33, 1986.

[BWS91] K.L. Boyer, D.M. Wuenscher, and S. Sarka. Dynamic edge warping: An experimental system for recovering disparity maps in weakly constrained systems. In *Trans. on SMC 21(1)*, pages 143–158. IEEE, 1991.

[CF75] E.C. Carterette and M.P. Friedman. *Handbook of Perception Vol. V Seeing.* Academic Press, New York, USA, 1975.

[CN91] R. Chung and R. Nevatia. Use of monocular groupings and occlusion analysis in a hierarchical stereo system. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 50–56, Hawaii, 1991.

[CS87] J.L. Crowley and A.C. Sanderson. Multiple resolution representation and probabilistic matching of 2-d gray-scale shape. *IEEE Trans. Pattern Analysis and Machine Intell.*, 9(1):113–121, 1987.

[CVSG89] L. Cohen, L. Vinet, P.T. Sander, and A. Gagalowicz. Hierarchical region based stereo matching. In *Proceedings of the CVPR*, pages 416–421, 1989.

[DH73] R.O. Duda and P.E. Hart. *Pattern Recognition and Scene Analysis.* New York, Wiley, 1973.

[DH85] D.L. Donoho and J. Huber. The notion of breakdown point. In P. Bickel, K. Doksum, J.L. Hodges, and Belmont Jr., Wadsworth, editors, *A Festschrift for Erich L. Lehmann*, pages 157–184, 1985.

[DW76] J.E. Dennis and R.E. Welsch. Techniques for nonlinear least-squares and robust regression. In *Amer. Statist. Assoc. Comp. Section*, pages 83–87, 1976.

[Fai74] R.C. Fair. On the robust estimation of economic models. *Ann. Econ. Social Measurement*, 3:667–678, 1974.

[Fai75] W. Faig. Calibration of close-range photogrammetry systems: Mathematical formulation. *Photogrammetric Engineering and Remote Sensing*, 41(12):1479–1486, 1975.

[Fau93] Olivier Faugeras. *Three Dimensional Computer Vision.* MIT Press, 1993.

[FB86] M.A. Fischler and R.C. Bolles. Perceptual organization and curve partitioning. *PAMI*, 8(1):100–105, January 1986.

[Fle92] M.M. Fleck. A topological stereo matcher. *Int. J. of Comp. Vision*, 6(3):197–226, 1992.

[Fri80] J.P. Frisby. *Seeing: Illusion, Brain and Mind.* Oxford University Press, U.K., 1980.

[GGC85] O.-J. Gruesser and U. Gruesser-Cornehls. Physiologie des Sehens. In R.F. Schmidt, editor, *Grundriss der Sinnesphysiologie*, pages 174–241, Berlin, 1985.

[GLA92] D. Geiger, B. Ladendorf, and A.Yuille. Occlusions and binocular stereo. In *Proc. 2nd ECCV92*, Santa Margharita Ligure, Italy, 1992.

[GR76] W.L. Gulick and B. Lawson R. *Human Stereoopsis.* Oxford University Press, New York, 1976.

[Gre78] R.L. Gregory. *Eye and Brain: The Psychology of Seeing.* McGraw-Hill Book Co., New York, third edition, 1978.

[Gri85] W.E.L. Grimson. Computational experiments with a feature-based stereo algorithm. *IEEE Trans. on PAMI*, 7:17–34, Jan. 1985.

[Ham71] F.R. Hampel. *A general qualitative definition of robustness*, volume 42. Ann. Math. Stat., 1971. ch 1.2,1.7.

[Hel24] H. Helmholtz. *Handbuch der Physiologischen Optik*, volume 1. Verlag von Leopold Voss, Hamburg, Germany, 3 edition, 1924.

[HJL+89] R.M. Haralick, H. Joo, C. Lee, X. Zhuang, V.G. Vaidya, and M.B. Kim. Pose estimation from corresponding point data. In H. Freeman, editor, *Machine Vision for Inspection and Measurement*, pages 1–84, London, 1989. Academic Press Inc.

[HJS90] H. Helmke, R. Janssen, and G. Saur. Automatische Erzeugung dreidimensionaler Kantenmodelle aus mehreren zweidimensionalen Objektansichten. In Grabkopf R.E., editor, *Mustererkennung 1990, 12. DAGM Symposium*, pages 617–624, 1990.

[HRRS86] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics - The Approach Based on Influence Functions.* John Wiley & Sons, 1986.

[HS93] Robert M. Haralick and Linda G. Shapiro. *Computer and Robot Vision*, volume II. Addison Wesley, 1993.

[HT75] M.J. Hinich and P.P. Talwar. A simple method for robust regression. *J. American Statist. Association*, 70(1):113–119, 1975.

[Hub81] P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[Hum87] R.A. Hummel. The scale-space formulation of pyramid data structures. In Academic Press, editor, *Parallel Computer Vision (L. Uhr, ed.)*, pages 187–223, New York, Oct 1987.

[IB92] J.R. Jordan III and A.C. Bovik. Using chromatic information in dense stereo correspondence. *Pattern Recognition*, 25(4):367–383, 1992.

[IPS85] A. Isaguirre, P. Pu, and J. Summers. A new development in camera calibration: Calibrating a pair of mobile cameras. In *Int. Conf. on Robotics and Automation*, pages 74–79, 1985.

[Jae95] M. Jaegersand. Saliency maps and attention selection in scale and spatial coordinates: An information theoretic approach. In *Proceedings Fifth Intern. Conf. on Computer Vision*, pages 195–202, Cambridge, MA, June 20-23 1995. MIT, IEEE. Catalogue no 95CB35744.

[JJT91] M.R.M. Jenkin, A.D. Jepson, and J.K. Tsotsos. Techniques for disparity measurements. *CVGIP*, 53:14–30, 1991.

[J.L67] Jr. Hodges J.L. Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Fifth Berkeley Symp. Math. Sta. Probab.*, volume 1, pages 163–168, 1967.

[JM92] D.G. Jones and J. Malik. A computational framework for determining stereo correspondences from a set of linear spatial filters. In *Proc. of 2nd Europe Conf. on Computer Vision*, pages 395–410, Italy, 1992.

[JR94] J. Jolion and A. Rosenfeld. *A Pyramid Framework for Early Vision*. Kluwer, 1994.

[Jul71] B. Julesz. *Foundations of Cyclopean Perception*. University of Chicago Press, Chicago, 1971.

[KKM+89] D. Y. Kim, J. J. Kim, P. Meer, D. Mintz, and A. Rosenfeld. Robust computer vision: A least-median of squares based approach. In *Proceedings of the Image Understanding Workshop*, pages 1117–1134, Palo Alto, CA, May 1989. DARPA.

[Kli71] A. Klinger. Pattern and search statistics. In J.S. Rustagi, editor, *Optimizing Methods in Statistics*, New York, 1971. Academic Press.

[KO94] T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: Theory and experiment. *PAMI*, 16(9):920–932, September 1994.

[Koe84] J.J. Koenderink. The structures of images. *Biological Cybernetics*, 50:363–370, 1984.

[Kro91] Walter G. Kropatsch. Image Pyramids and Curves - An Overview. Technical Report PRIP-TR-002, PRIP, TU Wien, 1991.

[KY96] W.G. Kropatsch and S. Ben Yacoub. A revision of pyramid segmentation. In *Proceedings of the 13th ICPR*, volume B, pages 477–481, Vienna, 25.-29. Aug. 1996.

[LB88] H.S. Lim and T.O Binford. Structural correspondence in stereo vision. In *Proc. Image Understanding Workshop*, volume 2, pages 794–808, April 1988.

[LB95] A. Luo and H. Burkhardt. An intensity-based cooperative bidirectional stereo matching with simultaneous detection of discontinuities and occlusions. *IJCV*, 15(3):171–188, July 1995.

[LCK93] C.Y. Lee, D.B. Cooper, and D. Keren. Computing Correspondence based on region and invariants without feature extraction and segmentation. In *Proceedings of the CVPR*, pages 655–656, 1993.

[Lin94] L. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

[LM90] W. Luo and H. Maitre. Using surface model to correct and fit disparity data in stereo vision. In *Tenth International Conference on Pattern Recognition*, pages 60–64, Atlantic City, NJ, June 16-21 1990.

[LOY73] M.D. Levine, D.A. O'Handley, and G.M. Yagi. Computer determination of depth maps. *Comput. Graphics Image Processing*, 2:131–150, September 1973.

[LP87] L.M. Lifshitz and S.M. Pizer. A multiresolution hierarchical approach to image segmentation based on intensity extrema. Technical report, Departments of Computer Science and Radiology, University of North Carolina, Chapel Hill, N.C., U.S.A., 1987.

[LS91] M.W. Levine and J.M. Shefner. *Fundamentals of Sensation and Perception*. Brooks/Cole Publishing Co., Pacific Grove, California, second edition, 1991.

[LT86] R.K. Lenz and R.Y. Tsai. Techniques for calibration of the scale factor and image center for high accuracy 3d machine vision metrology. Technical Report RC 54867, IBM Research Report, Oct. 8. 1986.

[Mar82] D. Marr. *Vision - A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman & Co., New York, USA, 1982.

[MB91] M.J. Mirza and K.L. Boyer. Performance evaluation of a class of m-estimators for surface parameter estimation in noisy range data. *IEEE Trans. on Robotics and Automation*, 1991.

[MB95] C. Menard and N. Brändle. Hierarchical Area-Based Stereo Algorithm for 3D Acquisition. In *Proceedings International Workshop on Stereoscopic and Three Dimensional Imaging*, pages 195–201, Santorini, Greece, September 6-8 1995.

[MB96] Christian Menard and Norbert Brändle. Computer based Acquisition of Archaeological Finds. Poster Presentation, DARV Berlin, February 1996.

[MBK81] H.A. Martins, J.R. Birk, and R.B Kelley. Camera models based on data from two calibration planes. *Computer Graphics and Image Processing*, 17:173–180, 1981.

[MBR87] P. Meer, E. S. Baugher, and A. Rosenfeld. Frequency domain analysis and synthesis of image pyramid generating kernels. *IEEE Trans. Pattern Analysis and Machine Intell.*, 9:512–522, 1987.

[Men91] C. Menard. Das Stereoverfahren, ein Verfahren zur bildhaften Erfassung archäologischer Fundgegenstände. Master's thesis, TU Wien, Institut für Automation, Wien, 1991.

[ML96] C. Menard and A. Leonardis. Robust stereo on multiple resolutions. In *Proceedings of the 13th ICPR*, volume A, pages 910–914, Vienna, Aug 1996.

[MMRK91] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim. Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, 6(1):59–70, 1991.

[MP76] D. Marr and T. Poggio. Cooperative computation of stereo disparity. In *Science Volume 194*, pages 283–287, 1976.

[MP79] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proc. R. Soc. Lond. B.*, 204:301–328, 1979.

[MS96] C. Menard and R. Sablatnig. Computer based Acquisition of Archaeological Finds: The First Step towards Automatic Classification. In Hans Kamermans and Kelly Fennema, editors, *Interfacing the Past. Computer Applications and Quantitative Methods in Archaeology. CAA95*, number 28, pages 413–424, Leiden, March 1996. Analecta Praehistorica Leidensia.

[MT89] S.B. Marapane and M.M. Trivedi. Region-based stereo analysis for robotic applications. *IEEE Trans. Syst., Man, Cybernetics*, 19:1447–1464, Nov./Dec 1989.

[MT96] C. Menard and I. Tastl. Automated color determination for archaeological objects. In *Is&T Fourth Color Imaging Conference*, Scottsdale, Arizona, November 1996. in press.

[Nac95] P.F.M. Nacken. Image segmentation by connectivity preserving relinking in hierarchical graph structures. *Pattern Recognition*, 28(6):907–920, June 1995.

[New86] S. Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *Am. J. Math.*, 8:343–366, 1886.

[Nis84] H.K. Nishihara. Practical real-time imaging stereo matcher. *Opt. Eng.*, 23:536–545, Sept. 1984.

[PS36] E.S. Pearson and C. Chandra Sekar. The efficiency of statistical tools and a criterion for the rejection of outlying observations. *Biometrica*, 28:308–320, 1936.

[RK82] Azriel Rosenfeld and Avinash C. Kak. *Digital Picture Processing Volume 2.* Academic Press, Inc., 1982.

[RL87] P. J. Rousseuw and A. M. Leroy. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.

[Ros84] A. Rosenfeld, editor. *Multiresolution Image Processing and Analysis.* Springer Verlag, 1984.

[RT71] A. Rosenfeld and M. Thurston. Edge and curve detection for visual scene analysis. *IEEE Trans. Computers*, 20(5):562–569, 1971.

[Sch90] B. G. Schunck. Robust computational vision. In *Proc. of the IWRCV*, Seattle, WA, Oct 1990.

[SHC90] J. Subrahmonia, J. Hung, and D.B. Cooper. Model-based segmentation and estimation of 3d surfaces from two or more intensity images using markov random fields. *IEEE Conf. on Pattern Recognition*, 1:390–397, 1990.

[SM92] R. Sablatnig and C. Menard. Stereo and Structured Light as Acquisition Methods in the Field of Archaeology. In New York Springer Verlag Berlin, Heidelberg, editor, *Mustererkennung 92 14. DAGM Symposium Dresden*, pages 398–404. Fuchs S., 1992.

[SM95] Robert Sablatnig and Christian Menard. Computer based acquisition of archaeological finds: The first step towards automatic classification. In *3rd International Symposium on Computing in Archaeology, Rome in press*, 1995.

[SMD91] Robert Sablatnig, Christian Menard, and Petros Dintsis. A Preliminary Study on Methods for a Pictorial Acquisition of Archaeological Finds. Technical Report PRIP-TR-010, PRIP, TU Wien, 1991.

[SMD94] Robert Sablatnig, Christian Menard, and P. Dintsis. Bildhafte, dreidimensionale Erfassung von archäologischen Fundgegenständen als Grundlage für die automatisierte Klassifikation. In O. Stoll, editor, *Computer & Antike*, volume 3, pages 59–84. Scripta Mercaturae Verlag, 1994.

[SMP93] R. Sablatnig, C. Menard, and Disntsis P. A Preliminary Study on Methods for a Pictorial Acquisition of Archaeological Finds. In Gothenburg Paul Astroems Foerlag, editor, *Archaeology and Natural Science (ANS)*, pages 143–151. Fischer P. M., 1993. Volume 1.

[Sob73] I. Sobel. On calibration computer controlled cameras for perceiving 3-d scenes. *Artificial Intelligence*, 5:185–198, 1973.

[SS92] S. S. Sinha and B. G. Schunck. A two-stage algorithm for discontinuity-preserving surface reconstruction. *IEEE Trans. on PAMI*, 14(1):36–55, Jan 1992.

[Tan86] S.L. Tanimoto. Paradigms for pyramid machine algorithms. In S. Levialdi and Berlin V. Cantoni, Springer Verlag, editors, *Pyramidal Systems for Image Processing and Computer Vision*, pages 173–194, Heidelberg, 1986.

[Tsa85] R. Y. Tsai. A versatile camera calibration technique for high accuracy 3d machine vision metrology using off-the shelf tv cameras. Technical Report RC 51342, IBM Research Report RC 51342, May 8 1985.

[Tuk81] J.W. Tukey. Some advanced thoughts on the data analysis involved in configural polysampling directed toward high performance estimates. Technical Report 189, Series 2, Department of Statistics, Princeton University, Princeton,N.J., 1981.

[TWK87] D. Terzopoulos, A. Witkin, and M. Kass. Stereo matching as constrained optimization using scale continuation methods. In *Opt. Dig. PR/ SPIE 754*, pages 92–99, 1987.

[Uhr72] L. Uhr. Layered 'recognition cone' networks that preprocess, classify and describe. *IEEE Trans. Comput.*, pages 759–768, 1972.

[VB91] N.M. Vaidya and K.L. Boyer. Stereoopsis and image registration from extended edge features in the absence of camera pose information. In *Proc. CVPR91*, pages 76–81, Lahaina, Maui, Hawaii, USA, 1991.

[vGSJ85] G. van der Wal, S. Gooitzen, Sinniger, and O. Joseph. Real time pyramid transform architecture. In *Intelligent Robots and Computer Vision*, pages 300–305, 1985. SPIE Vol.579.

[Wen92] J. Weng. Camera calibration with distortion models and accuracy evaluation. *IEEE Trans. Patt. Anal. Machine Intell.*, 14:965–980, 1992.

[Wit83] A.P. Witkin. Scale-space filtering. In *Proc. 8th Int. Joint Conf. Art. Intell.*, pages 1019–1022, Karlsruhe, West Germany, Aug. 1983.

[YP86] A.L. Yuille and T.A. Poggio. Scaling theorems for zero-crossings. *IEEE Trans. on PAMI*, 8:15–25, 1986.